Self-Supervised Visual Representation Learning for Medical Image Analysis: A Comprehensive Survey

Anonymous authors
Paper under double-blind review

Abstract

Deep Learning has developed as a great tool to achieve satisfactory performance in many computer vision or natural language processing tasks. But supervised deep learning algorithms require a large amount of labeled data to achieve satisfactory performance. Self-supervised learning, a subcategory of unsupervised learning, circumvents the issue of the requirement of a large amount of data by learning representations from the data without labeled examples. Over the past few years, Self-Supervised Learning has been applied to various tasks to achieve performance at par with or even surpass the supervised counterparts in several tasks. However, the progress has been so rapid, that any proper account of those works is unavailable. In this study, we attempt to present a review of those methods and show how the Self-Supervised Learning paradigm evolved over the years. Along with the aforementioned objective, we also present an exhaustive review of the Self-Supervised methods applied to Medical Image Analysis. Furthermore, we also present an extensive compilation of the details of the datasets used in the different works, along with a compilation of the performance metrics of some notable works on image and video datasets.

1 Introduction

The advent of Machine Learning has boosted the development of different fields of study like Artificial Intelligence, Computer Vision, and Natural Language Processing. Although there are classical methods, machine learning algorithms have outperformed almost all classical algorithms in various applications. With the invention of artificial neural networks and the availability of increasing computational capabilities at the disposal of researchers, the upscaling of artificial neural networks was made possible. This led to the advent of the Deep Learning paradigm.

Deep learning has played a crucial role in facilitating researchers to make substantial progress in numerous fields such as signal processing, computer vision (CV), natural language processing (NLP), time series analysis and others. Over the years different architectures with parameters ranging from millions to billions have been proposed and used to achieve almost human-level performance in various tasks like object detection, segmentation, and classification, or machine translation, chatbots, or even multi-modal applications like captioning, visual question answering, etc. Deep learning has also found its way to applications in medical image analysis. Applications in brain MRI, knee MRI, colonoscopy videos, chest X-ray images, mammograms, etc. are plentiful.

Nevertheless, supervised deep learning has its share of pros and cons. One of the cons of supervised deep learning methods is the requirement for large amounts of labeled data. Without that, supervised deep learning models tend to overfit and fail to generalize. Even if data scarcity is not an issue, supervised deep learning models require long training periods to achieve satisfactory performance. To prevent overfitting problems, researchers often use transfer learning techniques to train supervised deep learning models on small-scale datasets, based on knowledge learnt from training on large-scale datasets. Deep learning models trained on large-scale datasets like ImageNet or MS-COCO are often used as pre-trained models in many applications, even if there is a domain mismatch between the pre-training dataset and the target dataset. In real-life applications such as medical image analysis, labeled data is limited or hard to obtain. Medical

scans from Mammography or Magnetic Resonance Imaging require expert domain knowledge for efficient and reliable annotation. This proves to be labor-intensive and time-consuming.

To deal with the issues in Supervised deep learning, many machine learning paradigms like semi-supervised learning, self-supervised learning, etc. have emerged. In this study, we are going to focus primarily on self-supervised learning algorithms. In later sections, we will also discuss the applications of self-supervised learning strategies in different medical image modalities for representation learning.

Although several previous surveys have also reviewed the work on SSL, none has provided a detailed and minutely tailored discussion of each work like ours. While there have been reviews on self-supervised learning like Gui et al. (2023) which provides a comprehensive overview of the domain and also cited works on its applications in different sub-domains of computer vision, like point clouds, recommender systems, depth estimation, etc. but lacks a detailed overview of those works. Some older surveys like Jing & Tian (2021) and Mao (2020) do not discuss research work done after 2020. Other surveys such as Wu et al. (2023b) and Liu et al. (2023d) discuss only techniques related to graph neural networks, while Yu et al. (2024a) and Qi & Shah (2022) discuss Contrastive and adversarial techniques, respectively. From the literature, we find three previous surveys on medical image analysis such as Zhang et al. (2023a), Wang et al. (2023d) and Huang et al. (2023). Zhang et al. (2023a) provides a benchmark of popular SSL frameworks and also provides a detailed analysis of the effect of data imbalance in different frameworks. Wang et al. (2023d) comes close to our work but differs in the categorization strategy followed in our work. We believe that the categorization we followed in our work is fine-grained, providing a better resolution of the general overview for a survey, while Wang et al. (2023d) follows a coarser stratification. Huang et al. (2023) provides an implementation-oriented survey rather than going deep into the concepts of individual frameworks. In this survey, our aim is to systematically discuss and categorize the work done in the domain of self-supervised learning. We investigate the roots of self-supervised learning and find the first work Bridle et al. (1991) to propose the use of pseudo-labels and the concepts of 'fairness' and 'firmness', currently widely known as uniformity and alignment, respectively and re-invented in Wang & Isola (2020). We also discuss the work DeSa (1993) which can be considered to be the first work to coin the term "self-supervised learning" Bridle et al. (1991) published three decades ago. Most importantly, we have attempted to exhaustively cover the important and substantially contributing works in the domain of SSL, and discuss the contribution in most of them individually. In addition to that, we have also exhaustively covered works on medical image analysis using SSL and also done two-fold categorization of the works, firstly based on modality, and secondly on the SSL strategy. Furthermore, our survey is more updated than the previous ones.

We also systematically compile the details of different datasets used in the works discussed in our survey, and tabulate them according to their respective modality. We also present a compilation of the performance metrics of the notable works in each category of the SSL frameworks on the natural image and video datasets, which gives a broad overview of the evolution of the SSL domain over the years.

1.1 What is Self-Supervised Learning?

In the (to the best of our knowledge) first paper on Self-Supervised Learning (SSL) Bridle et al. (1991), the authors address the problem of classifying data without prior domain knowledge or labeled examples. This leads us to characterize SSL as being primarily concerned with the learning of representation from unlabeled data. Hence, we can say that SSL is a subcategory of Unsupervised Learning. The representations or features learned are then transferred to perform various tasks. Hence, Self-Supervised Learning consists of two phases: (a) Pretext or Pre-training or Surrogate tasks, and (b) Downstream or Target tasks.
Pretext or Pre-training tasks are used in self-supervised learning on unlabeled data to learn representations without utilizing any human annotations. A common approach is to generate pseudo-labels from the data itself to facilitate learning representations. Since the pretext tasks' sole purpose is to learn representations, a host of algorithms or methods have been used for this purpose. These tasks can be categorized into various types, such as generative, context-based, paired-embedding-based, clustering, or grouping-based methods. We review and discuss these methods individually in the later sections of this study. The pretext tasks aid to obtain pre-trained weights, similar to ImageNet pre-trained weights, which are then used in the Downstream or Target Tasks. The downstream task can be object detection, classification, segmentation, machine translation, etc. depending on the data, on which pre-training was conducted. In the downstream

task, labeled data is used for further fine-tuning and evaluation. Carefully observing the definition of pretext tasks, we can say that (unconditional) GANs and AEs also fall under the purview of SSL.

However, it is not always possible to learn effective and useful representations from the target data itself using self-supervised learning algorithms only. Lately, there has been a development in algorithms, where models or architectures trained on datasets different from those of the target dataset have been used as prior information in self-supervised learning, such as DINO (Caron et al., 2021) in UP-DETR (Dai et al., 2021a), DETReg (Bar et al., 2022). In MoCo-CXR (Sowrirajan et al., 2021), the authors pre-train MoCo on Chest X-Ray images using ImageNet pre-trained weights as initialization. As the domain of the data on which the pre-trained architectures are not the same as the target dataset, we can still consider the entire learning process to be self-supervised learning. Using a pretrained network for extracting information from the data in the pretraining phase does not violate the unwritten rule of self-supervised learning; that is, no information pertaining to the target label information is used during the pre-training phase.

1.2 Self-supervised Learning and Human Psychology

In Orhan et al. (2020), the authors aim to answer the question behind the origin of the ability of infants to learn shapes or animals and also to discriminate between them, using ego-centric videos and modern self-supervised learning architectures. The authors intend to determine how much of this knowledge learned by infants can be learned by generic learning architectures receiving sensory data through the eyes of a developing child, and how much of it requires more substantive inductive biases. The use of self-supervised learning algorithms in this study links this paradigm to the field of psychology.

In fact, from a study by Raymond B. Cattell on the Theory of Fluid and Crystallized Intelligence (Cattell, 1963) and later extended by John Leonard Horn in his Doctoral Dissertation Horn (1971) and also in Horn & Cattell (1966), we can speculate that the learning paradigm of self-supervised learning is in fact similar to the development of fluid intelligence in humans. As defined in Cattell (1963), fluid general ability refers to the ability to adapt to new situations, whereas crystallized general ability refers to those cognitive abilities in which skilled judgment has become crystallized. In Horn (1971), we learn that fluid intelligence and crystallized intelligence in fact are co-dependent. In light of this theory, we can formulate self-supervised learning as the fluid ability to learn novel knowledge bases or representations based on self-designed cues or, in other words, without cues from external agents. When the self-supervised pre-trained weights are transferred to other downstream tasks, it resembles the utilization of fluid intelligence to help in building up crystallized intelligence, as described in Horn (1971).

In fact, the problem statement in Bridle et al. (1991) of learning to classify samples without prior knowledge or labeled examples, fits the analogy of infants having the ability to discriminate between animal classes, which in turn, points to the similarity of this learning paradigm with fluid intelligence or ability in the field of psychology.

1.3 Motivation of the Survey

Although previous surveys have touched on SSL, none has delved into each work as comprehensively as ours. While there are reviews on self-supervised learning, they lack a detailed examination of individual works. Our work differs in the categorization approach too. The categorization we employ is more detailed, offering a finer resolution for surveying. Our survey aims to systematically discuss and classify the work in the realm of self-supervised learning. We trace the origins of self-supervised learning back to the first work by Bridle et al. (1991), and cover the most recent works too. Importantly, we strive to comprehensively cover the significant and impactful works in the field of SSL, discussing their contributions individually. Additionally, we extensively cover works on medical image analysis using SSL and categorize them based on modality and SSL strategy.

Organization of the Survey:

In the following subsections, we will discuss the different approaches used for representation learning in the self-supervised learning paradigm. We will start with the classical approaches, such as context-based pretext

tasks (Sec. 2.1), wherein we will discuss spatial and temporal context-based pretext tasks for representation learning on both images and videos. Following that, we will discuss the clustering-based frameworks in Sec. 2.2. After that, we dive into the discussion of paired embedding-based methods (Sec. 2.3), which includes both contrastive methods (Sec. 2.3.1) and non-contrastive (Sec. 2.3.2) methods as well. In addition to that, we discuss different works in the medical image analysis domain with applications of SSL. We divide the discussion in terms of imaging modality, such as, MRI & CT (Sec. 3.1), Ultrasound (Sec. 3.2), Endoscopy (Sec. 3.3), Radiographs (Sec. 3.4), Retinal images (Sec. 3.5), histopathology (Sec. 3.6), echocardiogram (Sec. 3.7) and skin images (Sec. 3.8). We also discuss two benchmarking datasets in the domain of SSL and also summarize the different datasets used in all the works documented in this survey in Sec. 4. In addition to that, we also present a compilation of the performance metrics of some notable works on natural image and video datasets in Sec. 5. Finally, we end this survey with a summary and conclusion in Sec. 6.

2 Self-Supervised Algorithms and Frameworks

Before delving into the discussion of different categories of SSL frameworks, we can get a taxonomical overview of the same from Fig. 1 and 2. For the convenience of the readers, the taxonomy tree has been divided into two parts in the two figures. In these figures (Fig. 1 and 2), we can observe the different levels of hierarchy that the SSL frameworks can be categorized into based on the approach adopted in those works. Fig. 1 and 2 also serve as a summary of the first part of the survey where we discuss the notable SSL frameworks on natural image and video data. In the following subsections, we delve into a detailed discussion of the different categories of SSL frameworks and understand the basic differences between them as well.

2.1 Context Based Pretext Tasks

In this section, we primarily discuss the different context-based pretext tasks used in self-supervised pre-training. We can categorize them primarily into seven types, namely (Spatial) Context Encoding, geometrical transformation prediction, Jigsaw Puzzle Solving, Colorization, Counting, Spatio-Temporal Context Prediction (used primarily for videos), and Masked Image Modeling. Works that cannot be categorized specifically into any of the above categories have been included in the Miscellaneous category. We will start our discussion with geometrical transformation prediction-based pretext tasks and then continue with the others, as mentioned.

2.1.1 Context Encoding

Relative Patch Prediction: Context encoding for unsupervised feature learning was first introduced in Doersch et al. (2014). In Doersch et al. (2014), the authors use the prediction of the context of a single patch as a supervisory task to learn object clusters for unsupervised object discovery. In a later work, Doersch et al. (2015), the authors employed AlexNet (Krizhevsky et al., 2012) to classify the position of a patch with one reference patch sampled a priori as context.

Image Inpainting: Generative Context encoding via Image Inpainting was first introduced in Pathak et al. (2016), where the authors adopted a DCGAN (Radford et al., 2016) based generative pipeline, with a joint loss, consisting of l_2 and adversarial loss. To prevent a trivial solution, the authors condition the generator on the masked region only. The proposed method showed considerable improvement over the nearest-neighbor based image inpainting method.

2.1.2 Geometrical Transformation Prediction

Camera Transformation Prediction: The task of using geometric transformation as a supervisory signal was first introduced in Agrawal et al. (2015). The authors used the prediction of the transformation of the camera from pair images as a pretext task. The supervisory signal was obtained from the odometry data in the KITTI (Geiger et al., 2012) and SF (Chen et al., 2011) datasets. Another work along similar lines was presented in Jayaraman & Grauman (2017), where the objective is to learn the ego-motion equivariance

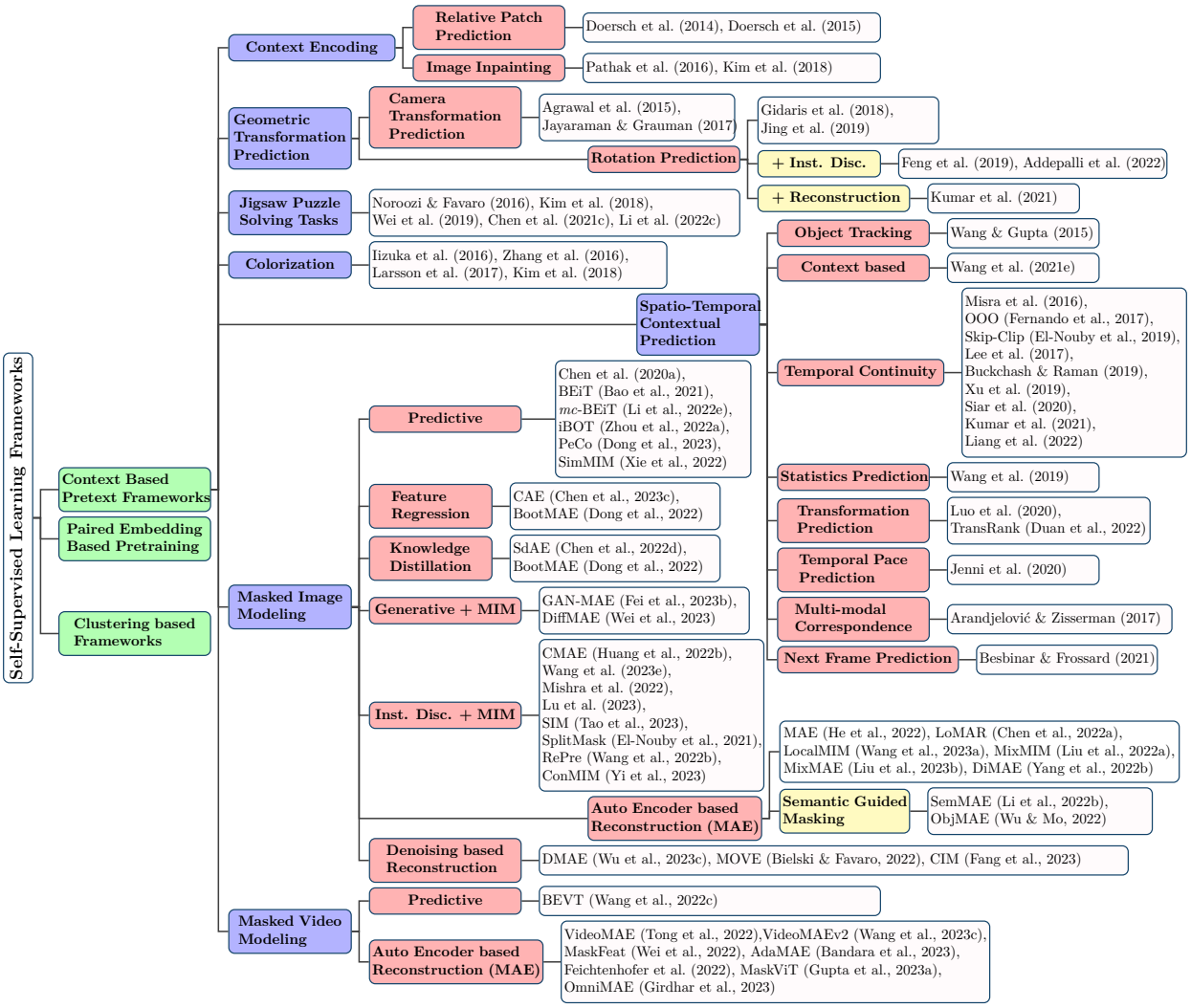

Figure 1: Taxonomy and Summary of SSL Frameworks (Part 1).

from image pairs selected from ego-motion videos, and the sensory signals were used to create pseudo-labels in the pretext task. Jayaraman & Grauman (2017) also combined contrastive loss to enforce equivariance between image pairs with supervised classification loss.

Rotation Prediction: In the work RotNet (Gidaris et al., 2018), the authors claimed that training the network to predict the rotation of the images forces the network to learn to locate salient objects in the image and learn the semantic features in the objects to effectively classify the orientation of the dominant features in the objects.

Combining with Instance Discrimination: To improve representation learning, Feng et al. (2019) combined rotation prediction with instance discrimination task to learn both equivariant and rotation invariant representations.

3D Rotation Prediction: The work done in RotNet was transferred to videos in Jing et al. (2019), where the authors adopted a 3DCNN architecture (3DRotNet) to account for both spatial and temporal information in videos.

Combining with Future Frame Prediction: Deriving from the above work and formulating a multi-task learning scheme by using a 3D Convolutional Auto-Encoder to predict future frames, in addition to

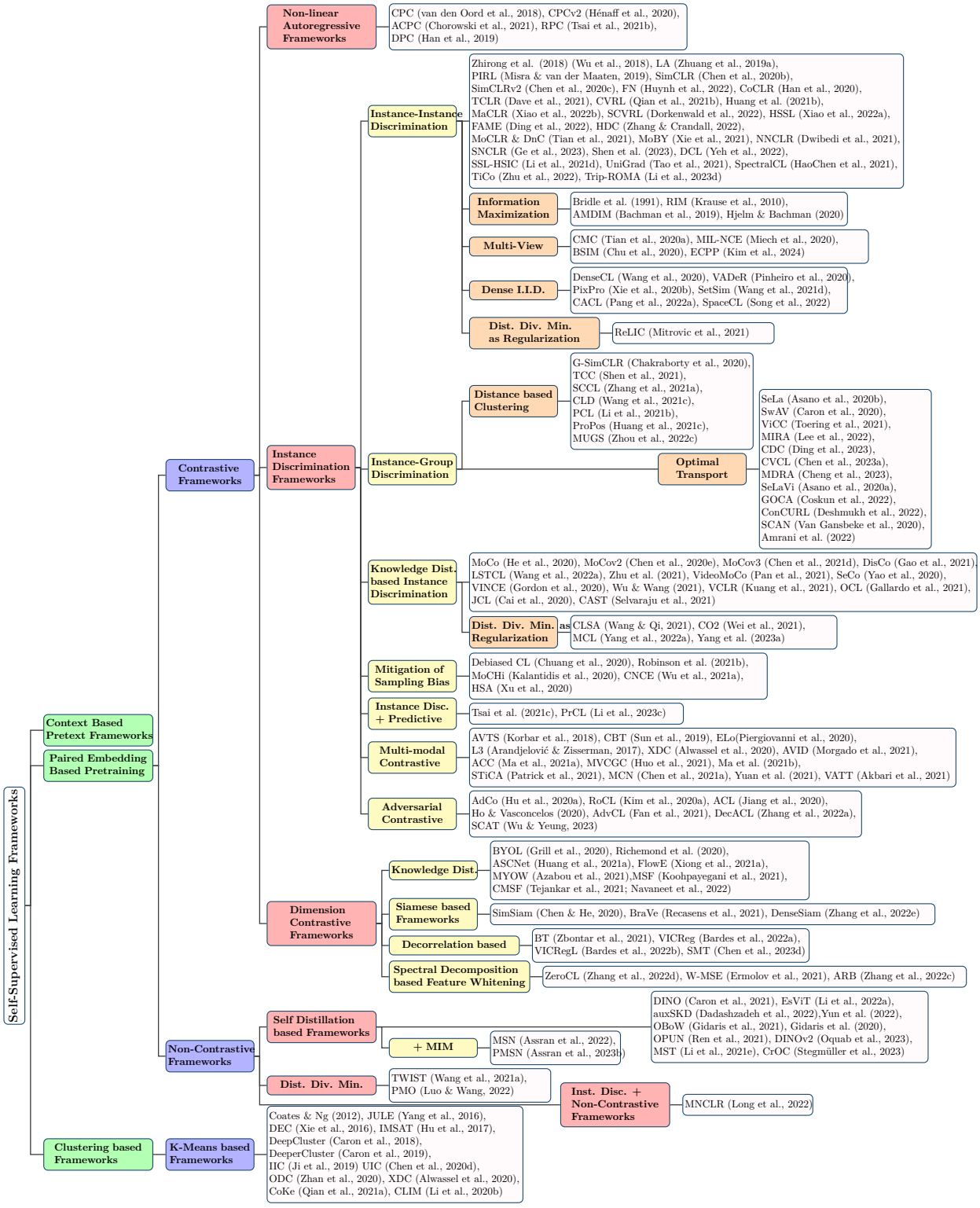

Figure 2: Taxonomy and Summary of SSL Frameworks (Part 2).

classifying rotation applied on the frames, Kumar et al. (2021) outperforms the vanilla 3DRotNet by a considerable margin on video retrieval benchmark tasks.

2.1.3 Jigsaw Puzzle Solving

The primary idea behind using jigsaw puzzle solving is to learn spatially invariant contextual information by learning the relative arrangement of patches with each other. The common approach involves dividing the image into several patches and numbering them in order. Then the position of the patches is rearranged, and a deep learning model is trained to predict the arrangement of the patches or generate the original input from the rearranged input.

To the best of our knowledge, the first such work that used jigsaw puzzle solving as a pretext task was presented in Noroozi & Favaro (2016), where the authors presented a context-free network by keeping the computation of features from individual patches remains independent until the first fully-connected layer. This measure adopted by the authors prevents the learning of low-level artifacts by the backbone network. The authors also use a Hamming distance-based selection of a set of arrangements of the patches. Another issue that the authors talk about in their work is the learning of *shortcuts* in self-supervised learning. This phenomenon results in the learning of features suitable for solving the pretext task but not the downstream task. In jigsaw puzzle solving, it happens when the network learns to associate each patch to an absolute position, but not on the basis of their textural or structural semantics.

In Kim et al. (2018), where the authors combined jigsaw puzzle solving, image colorization, and image inpainting in a multi-task learning problem for self-supervised learning of representations from images. In Wei et al. (2019), the authors take an iterative reorganization approach by using a strategy of probabilistically assigning patches to a particular position, as well as, optimizing the relativistic position assignment of any two patches. The iterative process is repeated until convergence when the patches are all assigned to their optimal positions. In Mundhenk et al. (2017), the authors propose a bag of different methods that are applied for self-supervised learning of representations and also study the effect of each method on performance on the ImageNet benchmark. In another innovative approach Chen et al. (2021c), the authors clustered the patches and predicted the cluster for each patch of an image as a pretext task. JigsawGAN (Li et al., 2022c) combined the flow information from the prediction of the position of the jigsaw patches with the GAN-based generative task for representation learning.

Following a similar strategy to Noroozi & Favaro (2016), jigsaw puzzle solving was also applied to video data in Ahsan et al. (2019). Each frame is divided into 2×2 grid and all the patches over all the timesteps are rearranged. A network is trained to put the jumbled patches in place by learning both spatial context and temporal order of the events in the frames. The authors also adopt a curriculum learning-based approach by first training the network on an easier jigsaw puzzle-solving task, followed by a harder task. In Kim et al. (2019), the basic idea revolves around classifying the order in which the 16 pieces obtained from selected 4 frames are shuffled using a 3D CNN with separated encoding of each 3D space-time piece of the puzzle to prevent learning of low-level cues or trivial solutions.

2.1.4 Colorization

The principle of image colorization as a pretext task is based upon the fact that to color a grayscale image, the model needs to know the semantic information of the image or scene and the location too. This fundamental principle makes end-to-end image colorization algorithms suitable for learning representations from images without using human-annotated labels.

Generative: In the first known work which used image colorization as a pretext task Iizuka et al. (2016), the authors used a combination of self-supervised pre-training and supervised classification for representation learning.

Predictive: The next work that utilized colorization as a pretext task was Zhang et al. (2016), where the authors treat the colorization problem as a classification problem. The *ab* color space is quantized into bins with a grid size of 10, resulting in 313 classes. The authors also took care of class imbalance by using a weighted multi-nomial classification task as the pretext task for class rebalancing. Next, the work Larsson et al. (2017) is heavily inspired by Larsson et al. (2016). The framework proposed in Larsson et al. (2016)

predicts a color histogram at each location of the pixels. Furthermore, the framework uses the hypercolumn strategy Hariharan et al. (2015) as per-pixel descriptors.

2.1.5 Video Spatiotemporal Contextual Prediction

The objective of video representation learning is to learn both spatial and temporal features. We will discuss the works done in this sub-domain and also categorize them according to the pretext strategy used.

Context-based strategy: SSCAP (Wang et al., 2021e) uses context-based SSL representation learning frameworks for feature extraction from video frames for subsequent co-occurrence action parsing for action segmentation.

Transformation Prediction: In Luo et al. (2020), the strategy adopted consists of generating a number of separate clips from a video in order and removing one clip randomly. Several spatial and temporal operations are applied to the removed clip, and the network is trained to predict the option which has been used to alter the removed clip. Recently, another pretext task was proposed in Duan et al. (2022), where a ranking-based framework was used to learn semantic and temporal information from unlabeled videos by scoring the transformations relative to one another.

Object Tracking: In one of the first papers on self-supervised video representation learning Wang & Gupta (2015), two different instances of the same object are obtained by tracking a patch containing an object over the frames of the video, using improved density trajectory on SURF feature points. The primary objective of the learning process was to map the features of the object in two different patches from two different time stamps close to each other. However, to prevent the collapse of representations, hard negative mining for sampling negative samples was used to optimize a ranking loss. Building upon the work done in Doersch et al. (2015) and Wang & Gupta (2015), Wang et al. (2017b) utilized transitive relation invariance, constituted of both inter-instance relations between different object instances of similar appearance and intra-instance relations between identical objects at different timesteps for visual representation learning.

Using Temporal Continuity: Using the sequential order of frames in a video can also provide useful information to learn unsupervised representation from videos, as presented in Misra et al. (2016). In this work, the representations are learned by classifying if a tuple consisting of 3 frames sampled from a high-motion window in the video, is in the correct order or not. Similar to Misra et al. (2016), sequential variation in visual features has been used to learn representations in OOO (Fernando et al., 2017), Lee et al. (2017), Skip-Clip (El-Nouby et al., 2019) as well. In OOO (Fernando et al., 2017), several clips from a video are used as input. All but one clip are in the correct order. The network is tasked with predicting the index of the sample with the frames in the incorrect order. Whereas in Lee et al. (2017), the network is tasked to predict the order in which 4 frames in the input are arranged.

The concept of temporal coherence is perfectly utilized in Skip-Clip (El-Nouby et al., 2019), where the objective of ranking clips based on a given context clip as plausible future clips of the given context is used as self-supervision. A detailed analysis of the method presented in Misra et al. (2016) was conducted in Buckchash & Raman (2019) and improved using by using a different sampling technique. A different take on learning representations from videos was proposed in Xu et al. (2019), where clip-based order ranking strategy with 3D CNN as the backbone was used. A similar approach is also used in Siar et al. (2020) by using 3D CNN to predict the order of the frames selected randomly from non-overlapping clips from a video.

The effect of combining multiple pretext tasks in video representation learning is shown in Liang et al. (2022), where the authors jointly optimize three types of losses to learn representations from videos. Firstly, a binary classification loss to classify clips from the same video in the same batch as continuous or discontinuous. Secondly, another classification loss to predict the location of discontinuity in the clips as well. Finally, a contrastive loss is used to learn the feature representation of the missing section in the discontinuous clips.

Multi-modal correspondence: In Arandjelović & Zisserman (2017), by utilizing audio-video correspondence learning as a pretext task, the authors were able to achieve performance at par with the contemporary state-of-the-art SSL methods on image classification benchmark on the ImageNet dataset.

Temporal Pace Prediction: Jenni et al. (2020) uses speed prediction and temporal context prediction for learning video representations.

Statistic Prediction: In Wang et al. (2019), both appearance and motion statistics were used for representation learning. Motion boundary, spatial-aware motion statistics, spatio-temporal color diversity statistics, and dominant color labels are some of the motion and appearance statistics used in the work.

2.1.6 Masked Image Modeling

Masked Image Modeling (MIM) is a relatively new direction of research in self-supervised learning. Although the terminology seems different, the principle is the same as contextual information learning, as in Doersch et al. (2015) and Pathak et al. (2016), where the primary task is predicting the masked portion of the image from the unmasked regions. The concept is adopted from the BERT masked language modeling framework Devlin et al. (2019) in the domain of Natural Language Processing (NLP).

Predictive Masked Image Modelling: Some initial works like Chen et al. (2020a), BEiT (Bao et al., 2021) introduced the concept of MIM in SSL. In BEiT (Bao et al., 2021), a ViT-based encoder is used to predict the visual tokens of the masked image patches. The tokenization is done using a discrete variational auto-encoder (dVAE) (Rolfe, 2017). *mc*-BEiT (Li et al., 2022e) uses an off-the-shelf tokenizer to generate soft probabilties for tokens to incorporate the possibility that semantically similar patches may be allocated with discrepant token ids and semantically dissimilar patches may be allocated with the same token id due to their low-level similarities. Unlike BEiT, which uses a pre-trained tokenizer separately, iBOT (Zhou et al., 2022a) uses a momentum updated encoder as an online tokenizer, and uses it to train a target encoder using a self-distillation-based masked image modeling framework. PeCo (Dong et al., 2023) attempts to improve BEiT by replacing the dVAE-based tokenizer with a VQ-VAE (van den Oord et al., 2017) based tokenizer as a learnable perceptual visual codebook to be used in BERT-like pre-training for visual representation learning.

SimMIM (Xie et al., 2022) introduces a simple framework for Masked Image Modeling, by using raw pixel value regression as the objective function when reconstructing the original image from the masked input image. Following BERT (Devlin et al., 2019), SimMIM (Xie et al., 2022) uses a learnable mask token vector to replace each masked patch.

Auto Encoder based MIM: Masked Auto Encoders (MAE) (He et al., 2022) are another prime example of instance based SSL frameworks which utilise masked image modelling. These frameworks also come under the purview of Masked Image Modeling and primarily reconstructive approaches to representation learning. Unlike BERT, MAE uses an encoder-decoder architecture with a high masking ratio of 70-80% for optimal performance. The encoder only processes a small portion of the patches, whereas the decoder processes both the input latent representations and the mask tokens to learn generalized representations.

To make MAE more efficient, LoMAR (Chen et al., 2022a) uses local masked reconstruction by sampling windows of patches from images and performs predictive reconstruction from embeddings of both masked and unmasked patches. LocalMIM (Wang et al., 2023a) uses signals from different layers in the ViT encoder to reconstruct input at multiple scales, to learn both fine- and coarse-level representations.

In MixMAE (Liu et al., 2023b) the visible tokens from two unlabeled images are mixed by using non-overlapping masking on each of the images. The encoder processes the mixed input to reconstruct both the input images, thereby saving compute on processing less informative mask tokens.

Semantic Guided MAE: SemMAE (Li et al., 2022b) uses semantic guided masking as a strategy for learning representation. However, to obtain the semantic information, it uses a separate iBOT (Zhou et al., 2022a) pre-trained encoder and a Style-GAN (Karras et al., 2021) based decoder to obtain the attention maps for different semantic regions. Similarly to SemMAE, ObjMAE (Wu & Mo, 2022) uses segmentation or CAM (Zhou et al., 2016) to select object-wise patches to learn object-wise decoupled representations.

Feature Regression based MIM: CAE (Chen et al., 2023c) involves combining masked representation regression and masked patch reconstruction. Another such work BootMAE (Dong et al., 2022) uses the output of a momentum-updated target encoder as a target for feature prediction in addition to pixel regression from multiple scales/levels of the online encoder.

Knowledge distillation based MIM: SdAE (Chen et al., 2022d) uses a similar self-distillation framework. It also uses a multifold mask strategy to reduce computational overhead in the teacher network pipeline. Another such work BootMAE (Dong et al., 2022) also uses knowledge distillation as an auxiliary task.

Denoising MIM: DMAE (Wu et al., 2023c) aims to learn robust representations by adding noise to the input image itself to produce corrupted images. In CIM (Fang et al., 2023), to generate the corrupted image, a dVAE (Rolfe, 2017) and a small pre-trained BEiT (Bao et al., 2021) are used as the frozen tokenizer and generator, respectively. Following BEiT, DALL-E (Ramesh et al., 2021) tokenizer provides token targets for the small BEiT generator. The image generated by BEiT serves as input to the enhancer network.

Combining Generative algorithms in MIM: In GAN-MAE (Fei et al., 2023b), MAE serves as the corrupt image generator and a GAN-like discriminator is used to classify the output image from MAE as fake or real. To reduce parameters, the discriminator and MAE encoder use shared parameters which improve efficiency. MOVE (Bielski & Favaro, 2022) also combines differentiable image inpainting using MAE and adversarial training to learn representations for object segmentation and detection.

In DiffMAE (Wei et al., 2023), denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) is combined with MAE (He et al., 2022) for visual representation learning. In the forward diffusion process, the masked image is used as input, and at the end of the diffusion process, the masked patches approximate the standard Gaussian distribution. Denoised masked regions are generated in the reverse process by the decoder from the encoded visible tokens, timestep, and noisy masked regions.

Domain Generalizing MIM: DiMAE (Yang et al., 2022b) attempts to tackle multi-domain data by learning domain-invariant representations using an FFT-based style-mix algorithm to convert data from multiple domains into a single input and reconstructing image from each domain using a domain-specific decoder.

Combining Instance based learning with MIM: Several works have attempted to combine instance-based learning with masked image modeling. Contrastive MAE (CMAE) (Huang et al., 2022b), Simple CMAE (Mishra et al., 2022), Wang et al. (2023e), and Lu et al. (2023) uses both masked patch/frame reconstruction to learn locally sensitive semantic representations and contrastive loss optimization to maximize similarity between different representations and also to learn the discriminative relation between different images. Similarly to CMAE, RePre (Wang et al., 2022b) also uses a contrastive objective to maximize representational similarity between different augmented videos. However, it uses a specialized reconstruction decoder to reconstruct the masked patches from multiple hierarchy features obtained from the ViT encoder.

SplitMask (El-Nouby et al., 2021) uses a unique variation of MAE by splitting an image into non-overlapping sets of patches and reconstructing the whole image from both sets separately. SplitMask also employs contrastive loss to maximize the representational similarity between the descriptors of the two sets of patches from the same image.

Siamese image modeling (SIM) (Tao et al., 2023) minimizes pairwise similarity with all the negative tokens obtained from the target encoder as the instance discrimination task. Furthermore, SIM shows that color augmentation, generally used in instance discrimination methods, and believed to not be beneficial for MIM as reported by MAE (He et al., 2022), can be made beneficial is used under proper training settings. SIM uses different views of the same image to effectively use color augmentations for representation learning, as the color variation information is leaked when used with the same view.

However, ConMIM (Yi et al., 2023) uses only the contrastive learning strategy, where a positive pair of images is obtained by using a masked and unmasked version of the same image. The unmasked image is passed through the momentum-updated encoder to form a patch-level feature look-up dictionary.

Architectural Modifications of MAE: ConvMAE (Gao et al., 2022) introduces a hybrid convolution-transformer encoder architecture following Co-AtNet (Dai et al., 2021b), UniFormer (Li et al., 2023b), Early Conv (Xiao et al., 2021). The convolution layers are primarily used for the high-resolution embeddings, while the transformer layers are used for the low-resolution embeddings.

ConvNeXtv2 (Woo et al., 2023) uses a sparse convolutional encoder and a convolutional block decoder as a fully convolutional MAE, inspired by the ConvNeXt (Liu et al., 2022c) architecture. Li et al. (2022d) use a two-stage masking strategy on image inputs to a pyramid-based ViT encoder for learning image representations. HiViT-MIM (Zhang et al., 2022f) uses a hierarchical ViT (Zhang et al., 2023e) architecture as an encoder for MIM, instead of vanilla ViT.

2.1.7 Masked Video Modeling

Predictive strategy: While the previous work discusses masked image modeling, applying the same on videos is tricky as it involves an additional temporal dimension. BEVT (Wang et al., 2022c) uses a VideoSwin (Liu et al., 2022b) transformer as shared image and video encoder, but a separate decoder for image and video. Both the image and video network are jointly trained for spatial representation and temporal dynamics learning. MaskFeat (Wei et al., 2022) is another similar framework that uses masked feature prediction to learn visual features for video understanding, but uses a dVAE codebook such as BEiT for tokenization.

Autoencoder based MVM: VideoMAE (Tong et al., 2022) is one of the foundational works that employs a method called tube masking to handle the two factors, temporal redundancy and correlation, and to learn representations from videos. VideoMAEv2 (Wang et al., 2023c) further scales VideoMAE by using a dual masking strategy to make video understanding more efficient. In addition to an encoder mask, VideoMAEv2 also uses a decoder mask following MAR (Qing et al., 2023).

Feichtenhofer et al. (2022) presents a simple extension of MAE (He et al., 2022) to video understanding, by dividing a video into a regular grid of patches in spacetime. The authors also oberseved that a high masking ratio of space-time agnostic masking reduces the computational complexity in the encoder.

AdaMAE (Bandara et al., 2023) uses an adaptive token sampler to select tokens from high activity regions with higher probability compared to tokens from background or low activity regions in videos. MaskViT (Gupta et al., 2023a) uses VQ-GAN (Esser et al., 2021) to encode video frames into discrete latent codes. Given a frame, the model learns to predict masked tokens in the future frames, using spatial and spatio-temporal attention blocks in each layer, which also reduces the memory requirements. OmniMAE (Girdhar et al., 2023) aims to learn a single model for both image and video using a omnivorous network (Girdhar et al., 2022), and treating images as a single frame video.

2.2 Clustering-based Pretext Tasks

K-Means: Coates & Ng (2012) presented one of the first pre-training approaches using K-Means algorithms to learn patchwise feature dictionary. DeepCluster (Caron et al., 2018) utilizes k-means clustering algorithm to generate pseudo-labels from features extracted by the convolutional neural networks. These pseudolabels are then used for the cross-entropy loss-based classification task for representation learning. DeeperCluster Caron et al. (2019) combines context based pretext task with k-means based clustering step in DeepCluster for better pre-training. UIC (Chen et al., 2020d) improves DeepCluster by using the 1-iteration embedding clustering step, making it scalable for larger datasets. An online version of DeepCluster was presented in ODC (Zhan et al., 2020) where pseudo-labels evolve along with the parameters, preventing rapid change to the pseudo-labels, with two separate memory bank for samples and centroids and without the requirement for an extra feature extraction step. The principle of DeepCluster (Caron et al., 2018) was also applied to multimodal data in XDC (Alwassel et al., 2020).

Agglomerative Clustering: JULE (Yang et al., 2016) proposed a recurrent network-based unsupervised framework that combines agglomerative clustering for pseudolabel generation and subsequent classification.

Using Online K-Means: CLIM (Li et al., 2020b) combines K-means clustering with kNN for positive sample selection. CoKe (Qian et al., 2021a) utilises an online constrained K-means algorithm to compute pseudo-labels and cluster centers to capture the global distribution of the data.

Autoencoder based clustering: DEC (Xie et al., 2016) instead uses a autoencoder for parameter initialization and a KL-divergence based clustering and parameter optimization step. IDEC (Guo et al., 2017) incorporates the auto-encoder in the DEC framework itself to preserve the local embedding structure. SDMVC (Xu et al., 2023) further improves on IDEC to multiple views by using an autoencoder for each view and clustering all views for global discriminative feature learning.

Mutual Information maximization: RIM (Krause et al., 2010) improves Bridle et al. (1991) by using a regularizer to prevent cluster fragmentation and complex cluster boundaries. IMSAT (Hu et al., 2017) uses RIM (Krause et al., 2010) for the clustering step and then uses an information maximization step based on cross-entropy.

2.3 Paired Embedding Based Pretext Tasks

While initial self-supervised learning frameworks based on context encoding, transformation prediction, jigsaw puzzle solving laid the foundation of SSL in the early years of its development, researchers soon realized the limits of such methods. It was the need of the time which led to the emergence of a new class of frameworks. These frameworks use information from paired embeddings to optimize a specific objective or loss function, and learn optimal parameters in the process. We can categorize these frameworks primarily into two categories: (1) Contrastive and (2) Non-contrastive. We will discuss several foundational and current state-of-the-art frameworks in both categories below.

2.3.1 Contrastive Learning Frameworks

Contrastive learning can be considered as learning by comparing different samples. The primary objective of contrastive learning frameworks is to discriminate between dissimilar samples in pairs termed as negative pairs, and closely map similar samples in pairs termed as positive pairs. Triplet loss based contrastive learning frameworks have been around for a long time, and used in works like face detection (Chopra et al., 2005; Schroff et al., 2015), metric learning (Weinberger & Saul, 2009), etc. However, the use of contrastive loss in an unsupervised setting was observed in the early works like Sermanet et al. (2017); Hyvärinen & Morioka (2016). In TCN (Sermanet et al., 2017), the authors used a triplet loss-based contrastive loss for self-supervised representation learning from multiview videos. Different views of an action or event at the same time step constituted the positive pair, and the embeddings of the samples in the positive pair were trained to be located close to the embedding space. On the other hand, clips or images from different time step formed the negative pairs, and were trained to be distant from each other. The representations were transferred for imitation learning in self-supervised robotics.

Non-linear Autoregressive Frameworks The paradigm of contrastive self-supervised learning (SSCL) frameworks received a huge boost in performance with the advent of CPC (van den Oord et al., 2018), primarily based on the principle of noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2012). For vision tasks, CPC uses an autoregressive style predictive framework. Images are first divided into patches and encoded. To predict each patch, the patches preceding it were used and the encoder was optimized using InfoNCE loss. The primary function of optimizing the InfoNCE loss was shown to maximize the mutual information between the input and the encoded representations. CPC described the InfoNCE loss as the categorical cross-entropy of classifying the positive sample correctly.

The principle of CPC (van den Oord et al., 2018) was later used again in several works such as CPCv2 (Hénaff et al., 2020) and ACPC (Chorowski et al., 2021). CPCv2 applies several modifications to CPC, which includes increasing model capacity, applying layer normalization, using more context information for predicting patch embeddings, and patch based augmentations. Harley et al. (2020) also adopt a non-probabilistic

version of CPC for both top-down and bottom-up representations in ego-motion stabilized videos. CPC was further improved in RPC (Tsai et al., 2021b), where the authors aimed to solve three primary issues in contrastive learning, training stability, sensitivity to minibatch, and downstream performance by eliminating the logarithm of contrastive loss and using an additional l_2-regularization. RPC can be associated with Chi-squared divergence. DPC (Han et al., 2019) used the principle of CPC for learning video representation.

Instance Discrimination Frameworks

The principle of instance discrimination frameworks is primarily based on the idea of contrasting samples. Representations of similar samples are drawn closer, while representations from dissimilar samples are pushed away. This principle prevents the collapse of representations. Instance discrimination-based frameworks can be further divided into several sub-categories (as described below), depending on the approach used for contrasting the sample instances. In the following subsections, we explore the evolution of instance discrimination-based SSL frameworks, along with the characteristics of each type of such frameworks.

Instance-Instance Discrimination: A different perspective on instance-instance discrimination was presented much before CPC in the work Dosovitskiy et al. (2014), where sampled patches from a randomly sampled number of images and applied random transformations on the images. A class label is assigned to the set of the transformed version of each image. The pretext task is simply a multiclass classification task, which involves learning to classify the images of the set of images.

Dawn of Instance-Instance Discrimination: Concurrently with CPC (van den Oord et al., 2018), Wu et al. (2018), proposed a novel instance discrimination framework based on the same principle of NCE. Wu et al. (2018) treated each instance as a separate class of its own, and maintained a memory bank to construct a non-parametric softmax classifier for self-supervised representation learning. This work laid the foundation for several SSCL frameworks for vision tasks.

Similarly to Wu et al. (2018), the basic framework of LA (Zhuang et al., 2019a) is based primarily on a clustering step to identify close and background neighbors from a momentum-updated memory bank and then apply a local aggregation metric based on contrast loss. PIRL (Misra & van der Maaten, 2019) uses a different formulation of contrastive loss following Hadsell et al. (2006). The final objective function in PIRL is a convex combination of two contrastive losses, one that maximizes similarity between an image and its augmented or transformed self, and the other that maximizes the dissimilarity between an image and the other images in the dataset. To implement the second part, PIRL follows a similar strategy of using a memory bank as Wu et al. (2018) and LA (Zhuang et al., 2019a). In fact, the formulation of Wu et al. (2018) is a special case of PIRL.

Large Batch Instance Discrimination: The first major SSL framework to not use a memory bank for the formation of negative pairs was SimCLR (Chen et al., 2020b). SimCLR mitigated the requirement for a memory bank by using a large batch size, and showed that increasing the number of negative pairs improves performance. It also emphasized the role of augmentation and non-linear projector in the quality of representations in SSL. SimCLRv2 (Chen et al., 2020c) further found that increasing the number of parameters results in better representation learning for semi-supervised and fine-tuning performance, if the labels are fewer. The popularity of SimCLR is evident from application of SimCLR as the foundation models for several works like sound classification (Fonseca et al., 2021), online distillation (Bhat et al., 2021), learning with noisy labels (Zheltonozhskii et al., 2022), etc. FN (Huynh et al., 2022) attempts to improve the preformance of SimCLR baseline by eliminating the influence of false negative pairs. The basic framework of SimCLR has also been adopted in contrastive video representation learning frameworks such as CoCLR (Han et al., 2020), TCLR (Dave et al., 2021), CVRL (Qian et al., 2021b), Huang et al. (2021b), MaCLR (Xiao et al., 2022b), SCVRL (Dorkenwald et al., 2022), HSSL (Xiao et al., 2022a), FAME (Ding et al., 2022) and HDC (Zhang & Crandall, 2022).

Another work presents MoCLR and DnC (Tian et al., 2021), which presents an improved version of SimCLR. It uses a mixture of experts trained on each superset and a base model trained on the whole data, for knowledge distillation to the final model. It yields considerable improvement in performance, but at the cost of large training datasets, pretraining encoder, and large number of pretraining epochs. The strategy used in MoCLR is also used in MoBY (Xie et al., 2021).

InfoMin (Tian et al., 2020b) empirically inferred that the augmentations used in contrastive learning are downstream task dependent and found a U-shaped relationship between an estimate of mutual information and the quality of the representation in a variety of settings.

Positive Samples Augmented Instance Discrimination: NNCLR (Dwibedi et al., 2021) further incorporates multiple positive instances in SimCLR (Chen et al., 2020b) by sampling various positive samples from the neighbourhood of each instance in the sample manifold represented by a support set queue as in MoCov1 (He et al., 2020) or MoCov2 (Chen et al., 2020e). Similar to NNCLR, SNCLR (Ge et al., 2023) also samples positive samples from a neighbour support set using the nearest neighbour algorithm. However, SNCLR employs an additional weight calculation step to compute the correlation of the sampled neighbours with the instances which is then used in the N-pair contrastive loss.

Positive-Negative Decoupling: In DCL (Yeh et al., 2022), the authors identified a negative-positive-coupling effect in InfoNCE loss which proved harmful to learning efficiency in contrastive frameworks. DCL proposed to remove this effect by eliminating the positive term from the denominator in InfoNCE loss, resulting in a significant improvement in performance without the requirement of a large batch size, such as in SimCLR (Chen et al., 2020b) or momentum encoding in MoCo (He et al., 2020).

Contrastive Learning without InfoNCE: SSL-HSIC (Li et al., 2021d) examines contrastive learning from a statistical dependence point of view. This work proves that InfoNCE approximates SSL-HSIC with a variance-based regularization and proposes a loss inspired by the HSIC (Hilbert-Schmidt Independence Criterion) bottleneck. For computing HSIC, SSL-HSIC uses an estimator provided by (Gretton et al., 2005).

UniGrad (Tao et al., 2021) presents an unifying framework using gradient analysis shows that different SSL frameworks optimize in a similar mechanism. UniGrad proposes a gradient-based framework, which has the primary aim of maximizing the cosine similarity between positive samples and the similarity between negative samples close to zero, and obtains considerably comparative performance with the contemporary state-of-the-art SSL frameworks. TiCo (Zhu et al., 2022) introduced a novel contrastive framework by using a squared contrastive loss instead of the widely used InfoNCE loss. The primary difference between the loss used in TiCo and InfoNCE is the uniformity (Wang & Isola, 2020) or denominator term. Optimizing the TiCo loss also makes contrastive learning easier, as it encourages the representations of negative samples to be orthogonal.

A concurrent work by HaoChen et al. (2021) explores spectral contrastive learning , where the authors use the population augmentation graph to effectively partition the same into sub-graphs, which are representative of fine-grained sub-classes of the actual classes in the downstream task. In reality, this work uses a learnable spectral decomposition component of the embeddings to learn the most important eigenvectors to maximise linear probe performance.

Triplet loss based Instance discrimination: Wang et al. (2021b) uses a truncated triplet loss to deal with the under-clustering and over-clustering issues in contrastive learning frameworks.

Trip-ROMA (Li et al., 2023d) presents a novel approach by combining triplet loss with binary cross-entropy for contrastive SSL. It also presents a randomness based similarity measure which can be used with foundation models like SimCLR to prevent overfitting or shortcuts as studied in Robinson et al. (2021a).

Metrics in Contrastive SSL: Although there have been many improvements in downstream performance with the advent of contrastive learning algorithms, there was no metric in the literature except linear probing or kNN accuracy, which could be used to measure the quality of representations. Wang & Isola (2020) analyzed self-supervised representation learning using two metrics, alignment and uniformity, and studied the relationship of the two metrics with the downstream performance. Furthermore, it also presented several illustrations to understand representation learning on the hyper-sphere. An intriguingly similar work was also presented in Ye et al. (2019). Moon et al. (2022) empirically discovered that the alignment and uniformity are directly correlated with the downstream performance of both instance-level and dense-level downstream tasks.

Dense Contrastive Learning: DenseCL (Wang et al., 2020) introduces a pixel-wise dense contrastive learning framework for dense representation learning, specifically for tasks like semantic or instance segmen-

tation, depth estimation, etc. It uses cosine similarity-based correspondence matching for the formation of pixel-level pairs. The final loss is a convex combination of the loss at the image level and the loss at the pixel level of InfoNCE. VADeR (Pinheiro et al., 2020) also presents a dense representation learning framework using an encoder-decoder architecture for computing pixel-level repsentations. Around the same time PixPro (Xie et al., 2020b) proposed a pixel-based contrastive learning framework. PixPro used a threshold to identify similar feature map pixels and then used a N-pair contrastive loss to maximize the similarity between their representations. SetSim (Wang et al., 2021d) tries to improve DenseCL by finding the correspondence set of features for the query feature vectors from the queue, as well as, pixel-level features from the feature attention maps. Similar to DenseCL, CACL (Pang et al., 2022a) uses a cycle loss to maximize the probability of getting a sample back as the positive sample of its neighbour, in addition to a global and pixel contrastive loss.

Generally, vision-based contrastive learning algorithms are trained on images with single instances. To scale the same for images with multiple instances as in the PASCAL VOC (Everingham et al., 2010) or COCO (Lin et al., 2014) datasets, SpaceCL (Song et al., 2022) proposes a space correlation module to effectively model local semantic similarity and space-correlated similarity in overlapping crops from the same image.

Recent works like Shen et al. (2023) use asymmetric masking to generate positive samples. The authors use stop-gradient like SimSiam (Chen & He, 2020) to prevent collapse of representations with InfoNCE as the loss function.

Multi-view frameworks: CMC (Tian et al., 2020a) extends this to more than two views using a self-supervised version of N-pair loss objective (Sohn, 2016). In addition to the modified objective, CMC also uses a memory bank similar to the above frameworks. Another major contribution of CMC is the introduction of the ImageNet100 dataset in the SSL domain, which is a subset of the 100 class of the original ImageNet1K dataset (Deng et al., 2009). Similar to CMC, MIL-NCE (Miech et al., 2020) uses multiple positive pairs for multi-modal representation learning. However, the objective function differs from CMC. The generation of multiple spurious positive samples was explored by using CutMix augmentation in BSIM (Chu et al., 2020). ECPP (Kim et al., 2024) improves CMC by increasing the number of augmentations using crop-only transforms of smaller dimensions like SwAV (Caron et al., 2020), and also discarding the augmented versions from the list of negatives in the denominator of the contrastive loss term.

In a recent work Hu et al. (2024), the authors use a combination of different modules for multi-view self-supervised learning. The proposed framework uses a pseudo-label guided positive pair sampling step to inhibit the effect of false negative pairs in contrastive learning, both feature and cluster correlation maximization to improve feature alignment, and finally a divergence minimization based clustering loss to further enforce compactness and separability.

Information Maximization: One of the oldest foundational works on SSL is Bridle et al. (1991). In this work, the output is used as the probability distribution over the class label, a discrete random variable. The objective is to maximize the difference between entropy of the average of the outputs (referred to as fairness) and the average of the entropy of the outputs (referred to as firmness). Maximizing the entropy of the average of the outputs prevents the dimensional collapse, while minimizing the average of the entropy of the outputs prevents collapse of the representation to a single point in the latent space. This principle is the backbone of all self-supervised learning frameworks.

AMDIM (Bachman et al., 2019) is one significant work contemporary to Wu et al. (2018), LA (Zhuang et al., 2019a), PIRL (Misra & van der Maaten, 2019) and MoCo (He et al., 2020) but did not use a memory bank. In AMDIM (Bachman et al., 2019), the authors presented a self-supervised version of Deep InfoMax (Hjelm et al., 2019) and expanded the architecture to incorporate multiscale features. This innovation allowed them to maximize the mutual information between samples in positive pairs efficiently by increasing the set of samples as a whole. The framework of AMDIM was also used for video representation learning in Hjelm & Bachman (2020).

Tschannen et al. (2020) explored the role of mutual information (MI) estimators in contemporary self-supervised learning frameworks such as CPC, CMC, and AMDIM. Firstly, it showed that MI and downstream performance are loosely connected, and maximizing MI is not necessary to learn good representations.

Distribution Divergence Minimization as Regularization: ReLIC (Mitrovic et al., 2021) explores SSL from a causal perspective with content and style as latent variables. This work argues that the conditional distribution of the class representations given the content should remain invariant under style changes. For this purpose, the authors used the KL divergence as a regularizer along with contrastive loss. von Kügelgen et al. (2021) tries to understand the success of data augmentations in enhancing pre-training performance from a theoretical perspective by treating content and style as latent variables.

Adversarial Contrastive Learning: AdCo (Hu et al., 2020a) uses learnable adversarial negatives to optimize the parameters using contrastive loss. AdCo formulates the problem as a minimax game between optimizing the parameters and negative adversaries. RoCL (Kim et al., 2020a), ACL (Jiang et al., 2020), Ho & Vasconcelos (2020), AdvCL (Fan et al., 2021), DecACL (Zhang et al., 2022a), SCAT (Wu & Yeung, 2023) are other works that aim to improve representation consistency by using adversarial training with the SimCLR framework as the baseline.

Knowlege Distillation based Instance Discrimination Frameworks: Another significant work using a memory bank for negative sample mining was presented in MoCo (He et al., 2020). However, the queueing procedure in MoCo differs from the contrastive frameworks mentioned above. Instead of storing a representation for each sample in the dataset, MoCo uses a momentum-updated encoder to extract representations to store in the fixed size memory bank. All the representations in the memory bank acts as negative samples, which positive pair is obtained by pairing two differently augmented versions of a sample. This extends the idea of knowledge distillation but did not use the predicted labels by a teacher model to supervise a student model. Instead, it used the distribution of the likelihoods of an augmented version of the query to supervise the retrieval of another version of the query from a pool of representations. LSTCL (Wang et al., 2022a) also uses MoCo as a baseline framework for video representation learning in addition to BYOL and SimSiam too. DisCo (Gao et al., 2021) distills knowledge from a large self-supervised pre-trained model to a student encoder. The student encoder is used as the target encoder in MoCov2 framework to train a mean student encoder, used for the downstream tasks. MiCE (Tsai et al., 2021a) uses MoCo as the baseline framework. Using a mixture of experts comprised of a teacher and student network as in MoCo, and a gating function to channelize the input to the appropriate experts, MiCE improves downstream performance by a considerable margin.

MoCov2 (Chen et al., 2020e) attempted to further improve MoCo (He et al., 2020) by incorporating two design properties of SimCLR, that is, non-linear projection head, and stronger data augmentation. Another attempt to further scale up MoCo (He et al., 2020) was presented in MoCov3 (Chen et al., 2021d). MoCov3 ditched the memory bank as it showed minimal gain when used with a large batch size. In MoCov3, an additional prediction head was also used in the online encoder, following BYOL (Grill et al., 2020). In addition to the aforementioned contributions, MoCov3 presented an exhaustive analysis to train vision transformers (Dosovitskiy et al., 2021) self-supervised. Zhu et al. (2021) proposes a simple yet effective feature transformation, which creates both hard positives and diversified negatives to improve training with MoCov2 as the baseline framework. VideoMoCo (Pan et al., 2021) has presented an extension of the MoCo framework to videos, where the authors used temporal adversarial learning to augment videos. Similar other works which uses MoCo as the baseline framework for video representation learning are SeCo (Yao et al., 2020), VINCE (Gordon et al., 2020), Wu & Wang (2021), VCLR (Kuang et al., 2021). VCLR uses SeCo as the baseline framework, and scales it to perform contrastive learning at video-level. Using MoCov2 as the baseline framework, OCL (Gallardo et al., 2021) found that self-supervised pre-training improves online continual training due to the generalization ability of SSL methods.

In one concurrent work JCL (Cai et al., 2020), instead of penalizing the individual positive sample when paired with the query, multiple positives are simultaneously paired and penalized by using a Gaussian approximation of the neighbourhood of the query. In another work, Zhao et al. (2022) used only MoCo with Swin transformers for representation learning.

CAST (Selvaraju et al., 2021) uses Deep-USPS (Nguyen et al., 2019) to identify salient regions and a constrained cropping augmentation method to avoid the inclusion of noisy background regions. Using MoCo (He et al., 2020) as the baseline, CAST uses an additional attention loss to base predictions on correct regions, as contrastive methods often use wrong regions to match query and key images.

Distribution Divergence Minimization as Regularization: CLSA (Wang & Qi, 2021) uses distribution divergence minimization using representations from a memory bank, in addition to contrastive loss for better representation learning with MoCov2 as baseline. CO2 (Wei et al., 2021) also uses the MoCov2 framework as a baseline and uses a KL divergence-based consistency regularization loss in addition to contrastive loss and also improves MoCov2 by a considerable margin. Similar approach is also explored in MCL (Yang et al., 2022a) and Yang et al. (2023a).

Mitigating Sampling Bias: One noteworthy characteristic of contrastive SSL is that samples with the same label were paired into negative pairs. Debiased CL (Chuang et al., 2020) attempted to remove the negative sampling bias from the available positive sample pairs only. A similar approach to reduce the influence of sampling bias in contrastive learning for sentence representations was also presented in Zhou et al. (2022b). In a similar theoretical approach to Chuang et al. (2020), Robinson et al. (2021b) proposed to utilise hard negative samples to improve the generalization of self-supervised learning frameworks and improve downstream performance. MoCHi (Kalantidis et al., 2020) synthesizes hard negatives by a convex combination of negative samples during the start of training, and harder negatives by mixing the query with negative samples as the samples begin to separate from each other. A similar approach for use on medical images was presented in Zhao & Zhou (2022). Another works which uses negative sampling for contrastive learning of visual representations is CNCE (Wu et al., 2021a). HSA (Xu et al., 2020) uses positive sample mining using k-nearest neighbour algorithm from hierarchical features extracted from the encoder.

Instance-Group Discrimination: Clustering based pretext tasks aim to adapt clustering algorithms for generation of pseudo-labels for end-to-end training of visual features.

Distance based clustering: In G-SimCLR (Chakraborty et al., 2020), the authors used an autoencoder to generate representations for a k-means clustering step to obtain pseudo-labels from each batch before applying contrastive loss. TCC (Shen et al., 2021) uses both cluster and instance-level contrastive learning like most of its concurrent counterparts. SCCL (Zhang et al., 2021a) is another work which uses both contrastive and KL-divergence based cluster assignment loss for representation learning.

IIC (Ji et al., 2019) simply utilises mutual information to learn representations from data in an unsupervised manner, discarding instance-specific details and also avoiding degenerate solutions due to the individual cluster assignment entropy maximization which prevents assignment of all samples to a single cluster.

In CLD (Wang et al., 2021c), the authors add a spherical K-Means based feature grouping step to apply instance-group discrimination, to reduce the effect of instance-instance discrimination between similar samples. Concurrently to CLD, PCL (Li et al., 2021b) formulated the proposed framework as an Expectation-Maximization algorithm. After clustering the features from the momentum encoder, negative sample prototypes are sampled, and finally the NCE loss is optimized. ProPos (Huang et al., 2021c) optimizes the MSE loss between a sample and its gaussian distributed positive neighbours, in addition to the objective proposed in PCL.

In another recent work, MUGS (Zhou et al., 2022c) uses three levels of granularity for representation learning, instance level, local group level, and global group level. For local-group-level discrimination it uses averaged out historical tokens stored in memory queue as neighbors. For the global-level, it builds a set of learnable group prototypes for instance-group discrimination.

Optimal Transport based pre-training: Simply stated, clustering requires assigning samples to each of the clusters. When a uniformity condition is applied to the assignment problem, it can be treated as an optimal transport problem.

In SeLa (Asano et al., 2020b), the first step is the same as the previous clustering-based pretext task, that is, cluster-based pseudo-label assignment and cross-entropy-based optimization. The second step involves *Sinkhorn-Knopp* algorithm based transport polytope computation. SwAV (Caron et al., 2020) uses clustering to compute prototypes which are used to predict codes of one view from the another using *Sinkhorn-Knopp* algorithm. SwAV is the first SSL framework to surpass ImageNet supervised features on COCO, Places205, and VOC07 datasets (Caron et al., 2020). A similar approach was also adopted in ViCC (Toering et al., 2021) for self-supervised video representation learning. SMoG (Pang et al., 2022b) further improves SwAV by adding a group-level discrimination branch to it. MIRA (Lee et al., 2022) also improves SwAV by not using

the equipartition constraint, rather it constraints the marginal entropy by mutual information regularization. Another recent work CDC (Ding et al., 2023) basically combines MoCo with SwAV, but the cosine similarities are computed between feature vectors expressed probabilities over the clusters. CVCL (Chen et al., 2023a) presents a SwAV-like framework (discussed below) for learning representations by using multi-view samples.

SEER (Goyal et al., 2022) explores the challenges of scaling the pre-training architectures using SwAV as the baseline framework and also address some of the engineering challenges and complexity of training at this scale.

Instead of using K-means for generating pseudo-labels like in DeepCluster (Caron et al., 2018), ODC (Zhan et al., 2020) or CoKe (Qian et al., 2021a), SCAN (Van Gansbeke et al., 2020) first uses a pretext task to learn representations and then obtains the clusters using neighbor sampling with the prior knowledge of the number of classes in the dataset. The parameters are then fine-tuned again using cross-entropy loss in a similar fashion to SeLa (Asano et al., 2020b). Whereas Amrani et al. (2022) uses the number of classes similar to SCAN to directly train an end-to-end classifier without labels by optimizing for same-class prediction of two augmented views of the same sample, under the condition of uniform distribution over the classes similar to SeLa or SwAV.

Similar to SeLa and SwAV, MDRA (Cheng et al., 2023) also uses optimal transport for relationship alignment, which is another term used to assign samples to prototypes. However, each feature vector is decomposed into subgroups along the feature dimension, and the procedure of relationship alignment is applied on each subgroup.

One fundamental shortcoming of SeLa, SwAV, and SMoG is the assumption of uniform distribution over the prototypes. However, using a uniform distribution ensures the possibility of converging to degenerate cases, but ignores that real-life datasets are skewed. SeLaVi (Asano et al., 2020a) presents a solution to this using a permutation matrix in the energy equation of the Sinkhorn-Knopp algorithm, which sorts the prototype entropies.

Later, in GOCA (Coskun et al., 2022), the optimal assignment of one mode is used as the optimal assignment prior for the other mode, and vice versa to combine the two information sources, for multimodal video representation learning.

ConCURL (Deshmukh et al., 2022) combines SwAV (Caron et al., 2020) with instance discrimination Wu et al. (2018) for consensus based clustering.

Contrastive + Predictive: Tsai et al. (2021c) provides an information-theoretic perspective to understand the properties of SSL, and also provides a combined framework of contrastive and predictive learning objectives. In addition to that, it also uses a inverse predictive learning step to discard task irrelevant information. In PrCL (Li et al., 2023c) also, the authors combined image inpainting, a context-based predictive task, with contrastive learning to improve the quality of representations in pre-training.

Multi-modal contrastive learning:

AVTS (Korbar et al., 2018) uses triplet contrastive loss instead of InfoNCE loss, with distance based self-supervised synchronization between video and audio modalities. CBT (Sun et al., 2019) uses the principle of masked language modeling on videos and paired textual information separately, as well as cross-modal contrastive learning to maximize mutual information between visual and textual modes of information. Although ELo (Piergiovanni et al., 2020) does not use a single loss, we can categorize it under this subsection, as it definitely uses a multimodal contrastive loss for video representation learning.

Multi-modal contrastive works discussed already in the previous sections are L3 (Arandjelović & Zisserman, 2017), XDC (Alwassel et al., 2020). AVID (Morgado et al., 2021) uses cross-modal contrastive learning and within-modal positive discrimination using a sampled positive and negative set. ACC (Ma et al., 2021a) uses cross-modal contrastive learning with MoCo as the baseline, but the output from the momentum updated encoder of the other modality is used as the target representations. MVCGC (Huo et al., 2021) uses positives sampled from the pool of samples of the other modality as hard positives and also incorporate cross-modal information by optimizing N-pair contrastive loss. Ma et al. (2021b) also uses the same loss. However, the authors factorize the feature space into a spatially local/temporally global (S-local/T-global) subspace and a

spatially global/temporally local (S-global/T-local) subspace and define two N-pair cross-modal contrastive objectives in each of the subspaces. Unlike AVID, STiCA (Patrick et al., 2021) uses cross-modal and within-modal contrastive loss for multi-modal representation learning.

MCN (Chen et al., 2021a) combines contrastive learning with reconstruction and clustering loss for multi-modal (video, audio, text) representation learning. Similarly to MCN, Yuan et al. (2021) explore SSL for multimodal representation learning with MoCov2 as the baseline framework. VATT (Akbari et al., 2021) uses MIL-NCE (Miech et al., 2020) to learn representation from video and text, and an additional NCE loss for video and audio.

FNACL (Sun et al., 2023) uses false negative suppression and true negative enhancement in the contrastive learning framework for sound source localization task. One of the most recent work Xuan et al. (2024) also utilises multi-modal contrastive learning for sound source localization in videos.

Dimension Contrastive Framework

The requirement of large number of negative samples in instance discrimination based contrastive learning led researchers to invent negative-free contrastive learning methods. However, to keep negative samples away from each other, it was necessary to have access to the statistics of the same. Therefore, instead of peeking into the batch dimension, researchers used information available along the embedding dimensions. This allowed these new frameworks to use the information from negative samples without explicitly contrasting with the same. However, by normalizing along the embedding dimensions, allows the information to be distributed over the embedding dimensions, preventing dimensional collapse.

Knowledge Distillation-based Frameworks:

In recent years, one of the groundbreaking and foundational work in the negative-free contrastive or dimension contrastive learning literature has been presented in BYOL (Grill et al., 2020). BYOL used a student (online) - teacher (target) network architecture as in knowledge distillation, but the teacher (target) network learns from the past iterations of student (online) network. The objective was to maximize the similarity between the representations predicted by the online network and the representations of the target network. The predictor MLP is essential to prevent the collapse of representations in BYOL. Initially, it was hypothesized that the collapse of representations was prevented because of the batch normalization (BN) layers used in BYOL and that the BN induced an implicit contrastive effect on the embedding representations. However, these hypotheses were rejected in Richemond et al. (2020). Although BYOL does not satisfy the criteria of dimension contrastive frameworks exactly, however by enabling the information to be distributed uniformly over the embedding dimensions using the BN layers in BYOL projector MLP, the prevention of dimensional collapse is achieved, which is the primary principle of dimension contrastive frameworks.

ASCNet (Huang et al., 2021a) uses BYOL as the baseline framework for video representation learning using both appearance and speed consistency as the objective. FlowE (Xiong et al., 2021a) also use BYOL for predicting the representations of another frame from one frame after applying flow transformation.

In another recent work, MYOW (Azabou et al., 2021) combines a distance loss between a sample x and mined samples from the latent neighbourhood of the sample x, with the objective used in BYOL (Grill et al., 2020). MSF (Koohpayegani et al., 2021) added a positive sampling step from a large memory bank to BYOL to improve consistency regularization in BYOL. CMSF (Tejankar et al., 2021; Navaneet et al., 2022) improves MSF by utilizing different sources of knowledge like multi-modal embeddings to constrain the nearest neighbour search space.

Siamese Frameworks: SimSiam (Chen & He, 2020) uses an architecture similar to BYOL with an encoder, projector and predictor, but it does not use a momentum updated encoder. Instead, SimSiam uses a stop gradient to prevent collapse. This modifies the objective into an alternating optimization problem. However, the authors show that without stop-gradient, the collapse of representations occurs, which is also observed in BYOL without momentum encoder. However, as mentioned in SimSiam, BYOL arXiv v3 included a version of BYOL without momentum encoder, but the learning rate of the predictor was increased 10×. BraVe (Recasens et al., 2021) also uses SimSiam as the baseline framework but for learning multimodal

representation from videos. DenseSiam (Zhang et al., 2022e) adds a dense pixel-wise similarity loss and a region based contrastive loss to SimSiam for dense representation learning.

Joint Embedding Predictive Architectures (JEPAs):

I-JEPA (Assran et al., 2023a) bridges two sub-domains, MAEs and negative-free contrastive learning or dimension contrastive learning. The primary principle of I-JEPA is based on the principle of predicting the embedding of the masked regions, where the target representations are obtained from a momentum updated target network. In principle, I-JEPA is similar to BYOL in all but one aspect. I-JEPA uses multi-block masking strategy instead of morphological transformations. M-JEPA (Bardes et al., 2023) proposes a JEPA for learning a self-supervised optical flow estimator. Combining M-JEPA with VCReg (Bardes et al., 2022a), results in MC-JEPA. A-JEPA (Fei et al., 2023a), and V-JEPA (Bardes et al., 2024) are extensions of I-JEPA to audio and videos, respectively.

Decorrelation based Frameworks:

Barlow Twins (BT) (Zbontar et al., 2021) presents an innovative approach without using any similarity-based loss. The framework proposed in BT uses an objective which can be understood as an instantiation of information bottleneck, maximizing the variability of the representations over the dimensions, thereby preventing dimensional collapse, and also discarding redundant information arising from applied distortions or augmentations. The advantage of BT over InfoNCE-based frameworks is that it does not require a large batch size and benefits from large-dimensional embeddings. Barlow Twins was adopted as the baseline framework in several applications like MohammadAmini et al. (2022), Graph BT (Bielak et al., 2022), (Gomez-Villa et al., 2021), etc. He & Ozay (2022) argue that Barlow Twins output whitened features, and explore the relation between collapse of representations and whitening of features, and the exponent of eigenspectrum which follows the power law decides the gap. The authors also propose a post-hoc method to scale the eigenspectrum of the pre-trained encoder, which eliminates the need to train a linear classifier on top of the encoder for downstream tasks. Hua et al. (2021) explores the concept of collapse in BT and WMSE as baseline frameworks, and claims to discover dimensional collapse in SSL as well. This work explores the role of feature decorrelation, and proposes Shuffled Decorrelated BN for improved representation learning and prevention of dimensional collapse in SSL.

VICReg (Bardes et al., 2022a) also uses the decorrelation principle like BT, in addition to an invariance term to minimize the distance between positive samples, and a variance term to maintain the variance of each term above a predefined threshold, thus enforcing the embeddings to be different and preventing collapse. VICRegL (Bardes et al., 2022b) improves VICReg by adding location-based and feature-based matching of embeddings across both views of positive samples. Shwartz-Ziv et al. (2023) present an information-theoretic perspective of SSL with VICReg as the baseline framework. SMT (Chen et al., 2023d) is a simplest form of VICReg, but uses a more restrictive linearity criterion for similarities, where local, temporal, or spatial neighbor linear interpolation is used to define similarity on the manifold.

Spectral Decomposition based Feature Whitening:

W-MSE (Ermolov et al., 2021) does not use a separate predictor like BYOL and also avoids collapse while using the same loss as BYOL. W-MSE achieves this by using a Cholesky decomposition step to whiten the features along the batch dimension. Whitening prevents the pre-representations from collapsing by having a scattering effect on the samples, and forces the vectors thereby obtained to be uniformly distributed on the unit sphere.

ZeroCL (Zhang et al., 2022d) proposes a novel approach to self-supervised representation learning using independent invariance minimization along both batch and feature dimensions after instance-wise and feature-wise ZCA whitening, respectively. ZeroCL eliminates the redundancy reduction term in BT (Zbontar et al., 2021) by feature-wise whitening of the representations before calculating the loss. However, like W-MSE, ZeroCL involves a whitening step using Cholesky decomposition that is computationally expensive and is of the order of $\mathcal{O}(n^3)$.

ARB (Zhang et al., 2022c) proposes another new approach using orthonormal bases of one view of a sample as a target for the feature representations of the other view. However, ARB divides the representation

in subgroups similar to MDRA (Cheng et al., 2023) before computing the orthonormal bases to reduce computational cost. To account for non-full rank cases, ARB computes the pseudo-bases by using a spectral decomposition of the correlation matrix of the output representations, which involves a computationally expensive step as well.

2.3.2 Non-Contrastive Frameworks

Non-contrastive frameworks can be defined simply as those frameworks which do not explicitly use contrastive loss for self-supervised pre-training. Primarily, these frameworks discard the negative pairs, and only use the positive pairs in the self-supervised pretraining phase while using a symmetric or asymmetric network architecture. One of the first non-contrastive frameworks can be traced back to DeSa (1993) and DeSa (1994), where the primary objective is to minimize the disagreement between the information from two different modalities.

An innovative yet simple approach of using uniformly sampled noise from l_2 unit sphere as fixed target representations to avoid collapse in self-supervised learning was presented in NAT (Bojanowski & Joulin, 2017).

Distilation-based Frameworks: Self-distillation is similar to knowledge distillation (Hinton et al., 2015) in supervised learning, but without *a priori* teacher network. In DINO (Caron et al., 2021) the parameters of the teacher network are generally obtained from the momentum encoding of the parameters of the student network over training iterations. The student network is trained to learn local-to-global correspondences by matching the probability distribution of both networks. EsViT (Li et al., 2022a) improves MoCov3 (Chen et al., 2021d) / DINO (Caron et al., 2021) by studying the properties of ViT in SSL. EsViT observes that ViTs are able to automatically discover semantic correspondence between local regions, but the use of multi-stage ViT causes a loss of property. EsViT proposes a novel non-contrastive region matching pretext task to capture the local region dependency in the features. ReSSL (Zheng et al., 2021b) presents a novel framework based on relational consistency between instances instead of explicitly repelling instances in negative pairs and pulling augmented views of same instance. It uses a teacher-student framework to obtain representations of two views of an instance and computes the similarity distribution of each instance with the representations in the memory bank. Finally, the KL-divergence is minimized to enforce the relation consistency between the two augmented views of an instance. A similar approach to ReSSL is also applied in ISD (Tejankar et al., 2020), however, instead of a memory queue, it uses a collection of random samples to approximate the neighborhood of the sample. In another work, SCE (Denize et al., 2022) combines MoCov2, ReSSL, and N-pair contrastive loss for video representation learning. Whereas, auxSKD (Dadashzadeh et al., 2022) applies a framework similar to DINO for spatio-temporal representation learning as an auxiliary task on top of a predictive primary pretext task. Yun et al. (2022) improves DINO by mining positive patch from the neighbouring patches and using the aggregated representation as the target.

The bag-of-words approach has been used in classical computer vision in the past. Gidaris et al. (2020) uses self-distillation type approach, where a visual words vocabulary is used to quantize / encode the representations and generate a probabilistic softmax output. OBoW (Gidaris et al., 2021) differs from SwAV only in the clustering part. Similarly to Gidaris et al. (2020), OBoW uses a bag-of-words or a queue of features as a visual-words vocabulary, which is synonymous with the cluster prototypes in SwAV.

OPUN (Ren et al., 2021) uses a self-distillation approach similar to DINO, but uses an online clustering algorithm for prototypes, where the authors use an additional loss for new cluster formation. In a more recent work, DINOv2 (Oquab et al., 2023) scales self-supervised pretraining in terms of data and model size. It combines DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022a) with the centering of SwAV (Caron et al., 2020) and KoLeo regularization (Sablayrolles et al., 2019).

MST (Li et al., 2021e) also adopts a similar approach, using a self-distillation architectural framework such as DINO (Caron et al., 2021), and also optimizes a reconstruction loss from the student network. Visual tokens in student network are masked using an attention-guided mask strategy, conditioned on the teacher network encoder output, to mask out low response patches.

CompRess (Koohpayegani et al., 2020) learns a small student network from a large self-supervised teacher network each with a separate memory bank using knowledge distillation (Hinton et al., 2015). Similar to CompRess, SEED (Fang et al., 2021) emphasized that small models with fewer parameters cannot learn instance discrimination effectively. However, the similarity distribution is computed by randomly sampling instances from a dynamically maintained queue, and it fails to effectively model similarity of those highly related samples. To solve this issue, BINGO (Xu et al., 2021a) proposes a new self-supervised distillation method consisting of two components: inter-sample distillation for pushing two augmentations of the same instance together and intra-sample distillation for pushing all instances in one bag to be more similar with the anchor one.

Combining Masked Image Modelling: MSN (Assran et al., 2022) combines masked image modeling with the self-distillation framework. However, unlike DINO (Caron et al., 2021), MSN uses a set of prototypes to calculate the softmax probabilities. PMSN (Assran et al., 2023b) relaxes the condition of uniform clustering in MSN by minimizing KL divergence with a power law distribution instead of negative entropy term in MSN.

Combining Clustering: CrOC (Stegmüller et al., 2023) uses DINO as the baseline framework and adds a representation centroid-based self-distillation pipeline with it.

Distribution Divergence Minimization: TWIST (Wang et al., 2021a) presents an interesting approach by minimizing the divergence between the probability distributions of two augmented samples, along with the entropy of each sample. TWIST also uses a diversity term to ensure that representations of different samples are different to prevent collapse.

PMO (Luo & Wang, 2022) uses a matching operator to match the representations of the input to a prior distribution, without the need to contrast the positive and negative samples, thus preventing collapse.

Non-Contrastive + Instance Disc. Frameworks: MNCLR (Long et al., 2022) combines SimSiam with MoCo (He et al., 2020) to build a multi-network framework for SSL.

2.4 Miscellaneous

In this subsection, we include the works that cannot be explicitly categorized into the above-mentioned categories. In general, these works combine multiple frameworks into one single one. In addition to that, we have also included works which discuss metrics for evaluating SSL frameworks and conduct analysis on different aspects of SSL, in this subsection as well.

The design of the object counting problem as a pretext task in SSL can be seen in Noroozi et al. (2017), where the authors use a contrastive loss, instead of a regression loss, to prevent trivial solutions, a common problem in SSL.

Kolesnikov et al. (2019) studied the effect of CNN architectures with different SSL frameworks and found that: a) architecture choices which negligibly affect performance in the fully labeled setting may significantly affect performance in the self-supervised setting, b) the quality of learned representations in CNN architectures with skip connections does not degrade toward the end of the model, c) increasing the number of filters in a CNN model and consequently, the size of the representation significantly and consistently increases the quality of the learned visual representations, and d) linear probing performance is sensitive to learning rate.

VFS (Xu & Wang, 2021) uses both MoCo (He et al., 2020) and SimSiam (Chen & He, 2020) for two different versions of their proposed approach for correspondence learning from videos. Besbinar & Frossard (2021) uses next frame prediction as the pretext task for reconstruction-based self-supervised representation learning.

Addepalli et al. (2022) combines handcrafted pretext task rotation prediction with foundation models such as SimCLR (Chen et al., 2020b), MoCov2 (Chen et al., 2020e), BYOL (Grill et al., 2020), SwAV (Caron et al., 2020) in a multitask learning environment.

Although SSL has advanced rapidly, there have been dearth of metrics for benchmarking those frameworks. Gwilliam & Shrivastava (2022) proposed several metrics for benchmarking and analyzing self-supervised frameworks in their work.

Islam et al. (2021) show that combining supervised loss with self-supervised contrastive loss improves transfer learning performance. Zhang et al. (2021c) also used contrastive loss as a regularizer along with cross-entropy loss in the downstream finetuning stage. Around the same time, Cole et al. (2021) asks some important questions about the impact of data quality and quantity, task granularity, and pretraining domain on the quality of representations learned in contrastive learning, and also answers them through empirical analysis.

As stated in Ryali et al. (2021), in self-supervised learning, commonly used augmentation pipelines treat images holistically, ignoring the semantic relevance of parts of an image leading to the learning of spurious correlations. This work addresses this problem by investigating a class of simple, yet highly effective background augmentations, which encourage models to focus on semantically-relevant content by discouraging them from focusing on image backgrounds. Basaj et al. (2021) proposes several visual probing tasks previously used in NLP to evaluate SSL frameworks.

UnMix (Shen et al., 2020) employs image mixing methods like CutMix (Yun et al., 2019) and MixUp (Zhang et al., 2018) to implement an unsupervised counterpart of label smoothing in supervised learning to improve representation learning. MixCo (Kim et al., 2020b) and i-Mix (Lee et al., 2021) are other concurrent works exploring the same image mixing strategy for contrastive learning algorithms.

ScoreCL (Kim et al., 2023) adds a score-based weighting mechanism to each term in both contrastive and non-contrastive frameworks to improve representation learning. Li & Liu (2023) uses a two-stage pretraining for video representation learning, the first stage being a contrastive learning stage, and the second stage a combination of distillation and reconstruction tasks.

Allen-Zhu & Li (2023) study the effect of the ensemble in the testing phase and whether the ensemble can be distilled into a single model. Ruan et al. (2023) also explores ensembling of teacher-student networks using different weighting schemes for cross-entropy objectives.

3 Application of SSL in Medical Image Analysis

3.1 MRI & CT

3.1.1 Application of Context based frameworks

Models Genesis (Zhou et al., 2021b) uses an image restoration-based task (from corrupted images) to learn image representations. Semantic Genesis (Haghighi et al., 2020) builds on Models Genesis by adding another stage of image reconstruction pretraining before image restoration pipeline. Jana et al. (2021b) also uses image restoration as pretext task to learn representations from CT images for liver fibrosis diagnosis. CaiD (Taher et al., 2022) reconstructs the original image from the corrupted version of the same image to learn context-aware representations in addition to an instance discrimination task.

Chen et al. (2019) uses restoration of corrupted images as a pretext task using a reconstruction-based framework. Sli2Vol (Yeung et al., 2021) uses a slice reconstruction-based strategy to learn representations to segment regions in 3D CT or MRI volume. Lu et al. (2020) and TractSeg (Lu et al., 2021) use pseudo-labels obtained from tractography for reconstruction to learn representations from fMRI data for segmentation.

Demirel et al. (2021) uses a reconstruction-based framework proposed by Yaman et al. (2020) for simultaneous multi-slice image reconstruction task itself. Jana et al. (2021a) and Zhang et al. (2021b) uses a combination of GAN-based and AE-based reconstruction for unsupervised representation learning.

SSL-LNE (Ouyang et al., 2021) also uses a reconstruction based framework to learn disease progression trajectory of individuals. Akçakaya et al. (2022) gives an overview of the different unsupervised methods used for biomedical image reconstruction.

Sun et al. (2021b) utilizes simulated artifacts obtained from downsampling of MR scans to incorporate cortical thickness as anatomical guidance. In the testing phase, an iterative training stage is used to learn a site-specific segmentation network.

Dong et al. (2021) reconstruct a fixed number of slices preceding and following the input slice of CT scans to learn representations. This work also uses an auxiliary task similar to BYOL (Grill et al., 2020) in addition to the reconstruction-based task. Similarly to Dong et al. (2021), Alice (Jiang et al., 2023) uses a combination of masked image modeling and maximization of similarity between semantically aligned crops obtained using SAM (Yan et al., 2020) for representation learning.

Chen et al. (2022e) show how masked image modeling outperforms traditional contrastive learning by speeding up convergence and greatly improving downstream task performance and can be utilized to advance 3D medical image modeling in a variety of situations.

Matzkin et al. (2020) uses 3D reconstruction from postoperative CT scans to estimate missing bone flap. Another work Zhang et al. (2023b) also uses 3D reconstruction based framework based on UNETR or Swin-UNETR for representation learning. SSPT-bpMRI (Yuan et al., 2023) uses a 3D UNet for reconstruction of 3D volume from an augmented sub-volume. The representations are then used for the detection and diagnosis of csPCa (prostate cancer).

OneSeg (Wu et al., 2022b) uses a reconstruction framework to learn the semantic correspondence between two different 2D slices from 3D CT scans. In the inference phase, the annotated data is propagated using this pre-trained reconstruction framework from a randomly selected representative slice.

Huang et al. (2022a) uses symmetric positional encoding for Brain MR slices and 3D VHOG as targets for reconstruction from masked 3D voxels. VectorPose (Zhang et al., 2023f) uses boundary and voxel reconstruction, as well as spatial vector prediction, to learn spatial and anatomy-sensitive representations of 3D volumes.

Mazher et al. (2024) uses a style transfer-based approach to learn representations from private datasets without compromising the data privacy of clients in a federated learning setting. The transferred models weights are used for subsequent downstream tasks such as segmentation. Several other works like Zhao et al. (2023), M^3AE (Liu et al., 2023a), Tajbakhsh et al. (2019) also use a reconstruction-based framework.

Several works like Jog et al. (2016), Zhao et al. (2018), Xu et al. (2021b), Zhao et al. (2021) use super-resolution as a pretext task. This allows the networks to learn contextual representations specific to the data, and also deal with the scarcity of high resolution medical data.

PrimeGeoSeg (Tadokoro et al., 2023b) and Tadokoro et al. (2023a) synthesis 3D volumetric data using geometric shapes to emulate 3D MR scans and trains a segmentation network using the same pretext task. In another work, Zhang et al. (2023d) uses a synthetic tumor data generation pipeline for learning to segment brain tumors. To et al. (2021) uses a reconstruction-based generative framework to generate augmented samples to deal with data scarcity issues in self-supervised learning.

DualHierNet (Xue et al., 2020) uses low-level features-based adversarial training between the two domains, along with semantic level and edge-level adversarial networks. Tomar et al. (2021) uses a generative style-transfer framework to learning learn volumetric representations for one-shot segmentation of brain MR scans. Other works using generative adversarial training but discussed in other subsections are Jana et al. (2021a), Zhang et al. (2021b), Tao et al. (2020), Yang et al. (2020).

TransMorph (Chen et al., 2022b) also uses a reconstruction-based approach for unsupervised image registration by predicting the deformation between fixed and moving images.

Spitzer et al. (2018) uses a siamese architecture for representation learning from differently cropped versions of the input by maximizing the similarity between the two encoded inputs and also predicting the transformations applied on both the inputs. Yang et al. (2020) uses the rotation and elastic prediction task as the source of self-supervisory signals in their framework, where the downstream segmentation module is also jointly trained with the self-supervision module. Furthermore, for disentanglement of appearance and

content codes, the proposed framework also uses a modality-transfer generative module to learn cross-modal content-aware representations in an adversarial training way, and a self-reconstruction module.

Zhuang et al. (2019b) propose an interesting approach by using Rubik's cube solving as a pretext task for representation learning from both MR and CT volumes. Zhu et al. (2020) further improves it by adding more augmentations for better representation learning. Taleb et al. (2020) also explores several context-based tasks for representation learning from MRI and CT images. Tao et al. (2020) uses an adversarial strategy like GAN to learn representations by training the discriminator to predict the correct arrangement as real and the rest as fake.

The application of jigsaw puzzle solving to medical image data was first done in Manna et al. (2022), where the authors used a semi-parallel architecture. However, the authors took a slice-based approach to learn representations from MRI scans. Each slice of the MR scan was divided into 9 patches, similar to Noroozi & Favaro (2016). For each patch, the authors used separate convolutional branches. The outputs were later merged and passed through custom convolutional blocks, to finally predict the arrangement of the patches. Similarly to Noroozi & Favaro (2016) and Ahsan et al. (2019), a hamming distance-based selection strategy was used for the set of patch arrangements. The authors showed the robustness of the jigsaw puzzle solving strategy to class imbalance in the data. This work was further improved in SKID (Manna et al., 2023), where the authors mainly used two different convolutional blocks to further improve representation learning. To deal with the 3D nature of MR scans, the authors used a ConvLSTM-based (Shi et al., 2015) classifier in the downstream task, and kept the encoder parameters frozen. However, in this work, the authors did not use a Hamming distance-based selection strategy. Instead, the authors used a randomly selected set of patch arrangements. The authors achieved an AUC score almost on par with the supervised baseline MRNet (Bien et al., 2018) on the MRNet dataset.

Taleb et al. (2021) uses a multimodal jigsaw puzzle solving task, where each patch is from a different modality and the objective is to minimise the reconstruction loss between the input and the output, that is obtained by rearranging the input patches according to the predicted permutation matrix using a differentiable sinkhorn operator.

PCRL (Zhou et al., 2021a) and PCRLv2 (Zhou et al., 2023) use a combination of three pretext tasks, rotation prediction (Gidaris et al., 2018), context prediction (Pathak et al., 2016), and instance discrimination (Chen et al., 2020b) for learning representation from MR scans. PCLRv2 extends PCRL to multi-scale resolutions for better performance along with other architectural changes. $SSL2$ (Wang et al., 2023b), CSwin (Li et al., 2023e) also uses a similar framework for sclerosis segmentation and prostate cancer detection and segmentation, respectively.

In a recent work by Monsefi et al. (2024), the slices are clustered to encode different features, and a classification task is used as the pretext task, where the network predicts the suitable cluster for a collection of slices from MR scans.

Li et al. (2020d) proposes a 2D Slice Order Prediction Based Framework from 3D MR or CT volumes for self-supervised representation learning. Rivail et al. (2019) learns spatio-temporal representation from optical coherence tomography images by predicting the time gap between two input B-scans.

Spitzer et al. (2018) uses a siamese architecture for representation learning from differently cropped versions of the input by maximizing the similarity between the two encoded inputs and also predicting the transformations applied on both the inputs. Yang et al. (2020) uses the rotation and elastic prediction task as the source of self-supervisory signals in their framework, where the downstream segmentation module is also jointly trained with the self-supervision module. Furthermore, for disentanglement of appearance and content codes, the proposed framework also uses a modality transfer generative module to learn cross-modal content-aware representations in an adversarial training way, and a self-reconstruction module.

In another work, Bai et al. (2019) predicts anatomical positions from cardiac MR data for representation learning. Blendowski et al. (2019) uses the prediction of the relative patch offset as the pretext task. The pretext task is a regression task for predicting the offset as a pair of parameters along all the 3 axes.

Tajbakhsh et al. (2019) does not propose any novel framework, it attempts to find an answer to the question if self-supervised pre-training on limited data provides more effective weight initialization than random initialization or initial weights transferred from an unrelated domain, by analyzing the performance of rotation-based pretext tasks on lung CT scans.

3.1.2 Application of Instance Discrimination based Frameworks

The work Jamaludin et al. (2017) can be considered as one of the first applications of SSL to medical image analysis. This work uses patient discrimination task using contrastive loss with vertebate level prediction as an auxiliary task for representation learning. CADx (Chen et al., 2022c) uses InfoNCE loss in texture information extracted from cervical optical CT images to learn representations to detect high-risk diseases, including high-grade squamous intraepithelial lesion and cervical cancer. Santilli et al. (2021) employ contrastive learning on basal cell carcinoma data for pretraining and transfers the weights to differentiate between cancer and normal breast tissue.

You et al. (2021) uses a momentum-encoded volumetric instance discrimination loss (Chen et al., 2020e), dimension contrastive loss (contrasting representations of different dimensions, treating each dimension as a sample) and a consistency loss between the teacher and student network, along with the supervised loss to learn representations from CT scans for volumetric segmentation.

Wu et al. (2021b) and Wu et al. (2022a) constructs local positives (same partition of different scans from a single patient), and negatives (different partition of different scans from a single patient) from partitioned scan volumes, and also takes scan partitions from different remote patients as negatives. With these samples in a federated environment, momentum-based contrastive learning (Chen et al., 2020e) is applied to representations for volumetric segmentation in the downstream task.

Inglese et al. (2022) also uses volumetric contrastive learning for classification of neuropsychiatric systemic lupus erythematosus patients. Tang et al. (2022b) used a combination of SimCLR, image inpainting, and rotation prediction to learn representation from 3D medical scans.

Fischer et al. (2023) uses a framework similar to that of Jabri et al. (2020) that used contrastive random walks for self-supervised semantic representation learning.

DrasCLR (Yu et al., 2024b) uses N-pair contrastive loss by sampling positive and negative samples from the neighborhood in addition to InfoNCE loss to learn representations of 3D lung CT images.

Chaitanya et al. (2020) used a dense contrastive learning task using global and local pixel-level discrimination for representation learning to segment MR images. Yan et al. (2020) also uses a similar framework. OS2 (Yang et al., 2023c) uses a contrastive learning framework with a novel interactive embedding module for support query (SQIE), equipped with channel-wise co-attention, spatial-wise co-attention, and spatial bias transformation blocks to extract interactive information between slices. Vox2Vec (Goncharov et al., 2023) also uses a contrastive learning framework on multi-scale representations to capture both global semantics and local semantics.

Zheng et al. (2021a) uses representations from multiple hierarchical levels of the encoder. Hierarchical features are aggregated and then used in an instance discrimination task to learn representations from MRI and CT scans. In addition to instance discrimination, the proposed framework also uses other context-based pretext tasks and an auxiliary reconstruction-based task as well.

Nguyen et al. (2023) uses SwAV (Caron et al., 2020) as the baseline framework for clustering semantic representations. To learn the dependence between 2D slices in 3D volumes, the aggregated embedding from all the slices is also trained to map close to embeddings of individual slices. Masked embedding predictions are also used as an auxiliary task.

Windsor et al. (2021) uses contrastive learning-based dense correspondence matching between DXA and magnetic resonance imaging, along with unsupervised image registration, to transfer segmentation annotations between the two modalities. A similar multi-modal contrastive learning is also used in Fedorov et al. (2021a), Fedorov et al. (2021b) and Fedorov et al. (2024) for mutual information maximization using different combinations of local and global representations.

Other works like Dong et al. (2021), CaiD (Taher et al., 2022), CISFA (Hu et al., 2022), CSwin (Li et al., 2023e), Liu et al. (2023c) also use instance discrimination as a pretext task in their proposed framework.

MsVRL (Zheng et al., 2022) uses BYOL as the baseline framework and extends it to multiscale representations of MR scans.

BT-UNet (Punn & Agarwal, 2022) uses Barlow Twins (Zbontar et al., 2021) to train the encoder, which is later fine-tuned for segmentation tasks on MR scans. This work also presents performance data on histopathological and skin lesions.

3.1.3 Application of Non-Contrastive Frameworks

Ouyang et al. (2020), Ouyang et al. (2022) uses superpixel-based semantic segments as pseudolabels for few-shot segmentation. In the downstream phase, the pretrained model can segment organs from MRI or CT data without fine-tuning.

Li et al. (2021a) used a pre-trained network for feature extraction and subsequent k-means clustering for sample re-weighting or imbalance-aware selection. The authors use SimSiam (Chen & He, 2020) as a baseline framework for representation learning.

3.1.4 Other Applications

Jiang & Miao (2022) uses a variety of SSL methods to pre-train 3D CNN on MR scans for alzheimer's disease classification. Tang et al. (2022a) learns a max-tree representations from image features to learn structural information from the image to aid in segmenting the medical image in the subsequent task.

3.2 Ultrasound

Jiao et al. (2020b) uses transformation prediction and video frame order prediction as a joint prediction task to learn the representation from ultrasound videos. Hu et al. (2020b) combined context encoding pretext task like Pathak et al. (2016) with adversarial training and DICOM metadata prediction to form the pre-training framework.

Jiao et al. (2020a) attempt to correlate audio with visual features in ultrasound video, along with ensuring that features of audio and video lie close to each other by minimizing a cross-modal contrastive loss.

3.3 Endoscopic Data

3.3.1 Application of Context based pre-training

Ross et al. (2017) uses an adversarial training strategy, in which the generator produces recolorized images using a U-Net architecture, which is used for segmentation in the downstream task. Vats et al. (2021) uses rotation prediction and jigsaw puzzle solving task for self-supervised pre-training from wireless capsule endoscopic images. This work also discusses the primary reasons behind the gaps that occur in the learning of semantic representation due to inadequate self-supervised training. In Hong et al. (2021), a reconstruction-based framework is used on colorectal images to learn representations for polyp segmentation.

3.3.2 Application of Instance Discrimination based framework

Jian et al. (2021) uses instance discrimination (Chen et al., 2020b) in endoscopic images to learn representations for the detection of Helicobacter Pylori infection.

In Intrator et al. (2023), the authors explore primarily two methods, single frame instance discrimination and multiview tracklet discrimination. Following Qian et al. (2021b), the authors choose the pre-trained network for the multiview tracklet discrimination task to apply reidentification approaches in colorectal videos to track polyps over frames, which effectively improves polyp classification and detection performance.

In Colo-SCRL (Chen et al., 2023b), the authors combined VideoMAE (Tong et al., 2022) with VideoMoCo (Pan et al., 2021) for representation learning from paired colonoscopy videos. The downstream task is the retrieval of polyp areas from colonoscopy videos of 2nd screening from a given query video of 1st screening.

3.3.3 Application of Non-contrastive frameworks

FPSiam (Gan et al., 2023) uses SimSiam as the baseline framework to learn representation from frames extracted from colorectal videos. In addition to the baseline frameworks, FPSiam utilizes features from intermediate layers to implement local feature similarity to reduce the aliasing effect of upsampling. Finally, the features are transferred for polyp detection in colorectal videos.

3.4 X-Ray / Radiographs

3.4.1 Application of Context based Frameworks

IDEAL (Mahapatra et al., 2021) takes a saliency map based interpretability-driven sample selection approach. The only self-supervised part in this work is the use of autoencoder for clustering the X-ray images using the latent feature vectors. DiRA (Haghighi et al., 2022) and Haghighi et al. (2024) also fall into this category of frameworks.

3.4.2 Application of Instance discrimination based framework:

Works like Sowrirajan et al. (2021); Chen et al. (2021e) uses MoCo as a baseline framework for self-supervised pre-training. MedAug (Vu et al., 2021) uses an unique approach of using patient metadata to pair scans to construct positive pairs in constrastive learning. Tiu et al. (2022) uses a multimodal contrastive framework to learn representations from chest radiograph images, to predict pathology in the downstream task. A similar approach is also adopted in Liao et al. (2021b). Hu et al. (2021) also uses MoCov2 (Chen et al., 2020e) framework as the baseline for representation learning from panoramic radiographs of the jaw, for subsequent classification and segmentation of tumors or cysts in downstream tasks.

Liu et al. (2021) uses JCL as the baseline framework for pre-training the mean teacher for semi-supervised classification of chest X-rays. Sun et al. (2021a) uses both patch or node based contrastive learning and graph level contrastive learning to learn both global and local representations from chest radiographs.

DiRA (Haghighi et al., 2022) and Haghighi et al. (2024) use a combination of image restoration, adversarial, and instance discrimination framework for learning representation from chest radiographs.

ConVirt (Zhang et al., 2022g) uses paired chest radiographs and text reports for text-guided cross-modal contrastive learning of visual representations. Zhang et al. (2023c) uses disease classifier by distilling knowledge from a network trained using cross-modal contrastive loss using paired image and text information.

3.4.3 Other frameworks

Li et al. (2023a) uses a novel SSL framework based on SimSiam with an additional cross-view MSE loss for gastritis detection from x-ray images. Park et al. (2022) uses DINO as the baseline framework for learning representations from a teacher network pretrained on a small dataset.

3.5 Retinal Images

3.5.1 Application of Context based prediction task

Holmberg et al. (2020) used the macular thickness obtained from the automatic segmentation of the optical coherence tomography volume as pseudo-labels for the pretext task of predicting macular thickness from IR fundus images. The pretrained network is then used for the classification of diabetic retinopathy in color fundus images.

Hervella et al. (2018) uses multimodal reconstruction as a self-supervised pretext task. This work is used in Álvaro S. Hervella et al. (2020) to deal with label scarcity. In Álvaro S. Hervella et al. (2021), the pretext

task of multimodal reconstruction of fluorescense angiography from retinography is approached using aligned retinography-angiography pairs as pretraining data. In Hervella et al. (2020), the same pretext is used for joint optical disc and cup segmentation in images of the eye fundus.

Uni4Eye (Cai et al., 2022) proposes a masked image modeling approach with a novel unified patch embedding module to learn unified representations from 2D color fundus images or Fundus Fluorescein Angiography (FFA)) and 3D optical coherence tomography (OCT) and optical coherence tomography (OCTA) images.

Yang et al. (2023b) uses multi-modal masked relational modeling, to enrich the semantic relationship among diseases. Relational matching is proposed to capture an abundant disease-related relationship by aligning the sample-wise feature relation between intact and masked features at both the self- and cross-modality levels.

3.5.2 Application of Instance Discrimination framework

Mojab et al. (2020) uses data from multiple devices / domain and applies SimCLR (Chen et al., 2020b) as the baseline framework to learn representations and shows that a multidomain self-supervised contrastive learning approach performs better than supervised transfer learning. Gupta et al. (2023b) also uses instance discrimination for representation learning from fundus images.

Li et al. (2020c) uses two stages for self-supervised representation learning. First, it trains a CyCleGAN (Zhu et al., 2017) to synthesize Fundus Fluorescense Angiography (FFA) images from color fundus images and uses this network to synthesize FFA images from fundus images in the target dataset. The different modalities are then used to learn modality-invariant representations using a patient discrimination (contrastive learning) framework. Li et al. (2021c) also uses two different pretext tasks, but in a collaborative learning or multitask setting. This work uses rotation prediction and patient / instance discrimination together for pre-training.

Other Applications Srinivasan et al. (2022) study the effect of self-supervised pretraining and imagenet-pre-trained weights on data sets for diabetic retinopathy.

3.6 Histopathology

3.6.1 Application of Context based Frameworks

Štepec & Skočaj (2020) uses generative image synthesis as a pretext task for anomaly detection in the downstream task. StarDist (Prakash et al., 2020) uses a denoising framework for learning representation from biomedical microscopy images for downstream segmentation tasks. Stacke et al. (2020) uses the framework of CPC (van den Oord et al., 2018) on histopathological images for representation learning. The study found that only low-level CPC features are relevant for tumor classification.

Sahasrabudhe et al. (2020) uses a scale prediction network along with enforcing equivariance of representations under transformations and smoothness regularization for representation learning.

Xie et al. (2020a) uses a count ranking and a scale discrimination loss based on triplet loss to learn representations for nuclei segmentation. The scale discrimination loss is used to learn nuclei shape aware information, and the count ranking loss is simply used to learn context-aware representations by training the network to learn the number of nuclei-shaped objects in the input.

3.6.2 Application of Instance Discrimination Frameworks

Ciga et al. (2022) uses SimCLR (Chen et al., 2020b) as the baseline framework for applying contrastive learning on histopathological images. DSMIL (Li et al., 2020a) uses a pre-trained SimCLR (Chen et al., 2020b) backbone for weakly supervised multi-instance learning on whole slide images. Saillard et al. (2021) uses a pre-trained U-Net to extract background subtracted whole slide images, and then divide them into multiple patches. These patches are then used to learn representations in a self-supervised way using a multiple-instance learning framework. Srinidhi et al. (2022) uses self-supervised learning only for learning representations from histopathology images.

CELLULOSE (Wolf et al., 2023) uses an object-centric contrastive approach by maximizing the distance between the embeddings of patches from different objects and minimizing the distance between the embeddings of patches from the same object, to allow the segmentation of individual cells in microscopy images.

3.6.3 Application of Few-Shot segmentation based approaches

Pseudo-label based few shot segmentation approaches similar to Ouyang et al. (2020) was also adopted in Dawoud et al. (2022a) for cell segmentation in microsopy images. An edge-based reconstruction branch is also used as self-supervision in a semi-supervised framework in Dawoud et al. (2022b).

Miscellaneous Applications Self-Path (Koohbanani et al., 2021) uses a host of self-supervised tasks as auxiliary tasks along with the primary task from pathological images.

3.7 Echo Cardiogram

Echo-SyncNet (Dezaki et al., 2021) uses multiview echocardiogram videos to learn spatio-temporal information by optimizing consecutive frame similarity, correspondence matching, and temporal order of frames. EP (Chen et al., 2021b) synthesizes ECG panorama which allows real-time querying of any ECG views, from one input view using a reconstruction-based framework. Mehari & Strodthoff (2022) uses BYOL (Grill et al., 2020) to learn the representations of the ECG data.

3.8 Skin Images

JIANet (Zhang et al., 2022b) uses a jigsaw shuffled skin lesion image as one sample in a positive pair in a jigsaw invariant instance discrimination task. This work also uses a VAE-based reconstruction branch as part of the proposed collaborative learning framework. The reconstruction branch serves as the means to preserve the important semantic features necessary for melanoma segmentation in the downstream task.

3.9 Miscellaneous

ImageNet pretraining boosts self-supervised pre-training: Azizi et al. (2021) demonstrate that self-supervised learning on ImageNet, followed by additional self-supervised learning on unlabeled domain-specific medical images, significantly improves the accuracy of medical image classifiers. Similar findings were also reported in MoCo-CXR (Sowrirajan et al., 2021), and Manna et al. (2021).

4 Datasets and Benchmarks

Researchers have also proposed some datasets specifically for the purpose of benchmarking the performance of self-supervised learning frameworks. We will discuss two such works which have proposed benchmarking in 11 different domains to test the versatility and adaptability of self-supervised learning frameworks.

4.1 Benchmarking datasets

DABSv1.0 (Tamkin et al., 2021) proposes a mixture of datasets for domain-agnostic benchmarking of self-supervised learning frameworks. It consists of datasets from several domains such as, natural images, speech, monolingual and multilingual text, medical imaging datasets, multi-channel sensor data, and paired images and text data. In Table 1, we document the different datasets used for pre-training and downstream performance evaluation for each domain covered in DABS.

DABSv2.0 (Tamkin et al., 2022) included more domains in addition to the ones used in DABSv1.0 (Tamkin et al., 2021). Bacterial genomics sequence dataset, semiconductor wafers manufacturing database, particle physics tabular dataset, protein sequence dataset, and multispectral satellite images are the new domains added in DABSv2.0. In Table 2, we document the different datasets used for pre-training and downstream performance evaluation for each domain covered in DABS.

Table 1: Summary of datasets used in DABSv1 benchmarking for SSL

Domain	Pre-training	Downstream
Natural Images	ImageNet (Deng et al., 2009)	FGVC-Aircraft dataset (Maji et al., 2013), the Caltech-UCSD Birds dataset (Wah et al., 2011), the German Traffic Sign Recognition Benchmark dataset (Houben et al., 2013), the Describable Textures Dataset (Cimpoi et al., 2014), the VGG Flower Dataset (Nilsback & Zisserman, 2008), and the CIFAR-10 dataset (Krizhevsky, 2009)
Speech	LibriSpeech (Panayotov et al., 2015)	VoxCeleb (Nagrani et al., 2020) and LibriSpeech (Panayotov et al., 2015) speaker recognition datasets, Fluent Speech Commands cls. (Lugosch et al., 2019), Google Speech Commands (Warden, 2018), and AudioMNIST (Becker et al., 2023) utterance classification tasks
Monolingual Text	WikiText-103 (Merity et al., 2017)	GLUE Benchmark (Wang et al., 2018)
Multilingual Text	mC4 dataset (Raffel et al., 2020)	PAWS-X tasks (Yang et al., 2019)
Medical Imaging	CheXpert (Irvin et al., 2019)	Chest-Xray8 (Wang et al., 2017a)
Multi-Channel Sensor data	PAMAP2 (Reiss & Stricker, 2012)	
Paired Image-Text	MS-COCO (Lin et al., 2014)	MS-COCO and Visual-Question Answering (Antol et al., 2015) modelled as binary classification tasks

Table 2: Summary of datasets added to DABSv1 ot form the DABSv2 benchmarking for SSL

Domain	Pre-training	Downstream
Bacterial Genomics	Genomics OOD dataset (Ren et al., 2019)	
Semiconductor Wafer Manufacturing	WM-811K (Wu et al., 2015a)	
Particle Physics	HIGGS dataset (Baldi et al., 2014)	
Protein Sequence dataset	Pfam (El-Gebali et al., 2019)	TAPE benchmark (Rao et al., 2019)
Multispectral Satellite Imagery	EuroSAT (Helber et al., 2019)	

4.2 Natural Image and Video Datasets

In this section, we also summarize the different datasets used in the works discussed in this survey. In Table 3 and 3, we can see the summary of the natural image and video datasets, respectively, or non-medical datasets. We make two columns to identify the task (both pretext and downstream) in which the datasets have been used. We see that most small-scale datasets have been used primarily, for instance discrimination tasks, while context-based pretext tasks have used datasets that have images with large resolutions. This is mainly due to the nature of the tasks, as the context is less identifiable in images with lower resolution. Whereas, for paired embedding-based tasks, learning a global semantic feature aided the downstream objective of image classification. However, for downstream tasks like object detection, and semantic segmentation, image datasets with higher resolution are preferred.

Table 3: Summary of natural image datasets used in Self-supervised pre-training

Dataset	Training Samples	Num. of Classes	Source	PT Task	DS Task
Image Datasets					
CIFAR10	50K	10	Krizhevsky (2009)	Paired Emb. tasks	Image Cls.
CIFAR100	50K	100	Krizhevsky (2009)	Paired Emb. tasks	Image Cls.
FC100		100	Oreshkin et al. (2018)	-	Few Shot Cls.
STL10	100K (unlabeled) + 5K (train)	10	Coates et al. (2011)	Paired Emb. tasks	Image Cls.
Tiny ImageNet	100K	200	Le & Yang (2015)	Paired Emb. tasks	Image Cls.
Aircraft	10200	102	Maji et al. (2013)	-	Image Cls.
DTD	5640	47	Cimpoi et al. (2014)	-	Image Cls.
Oxford Pets	7.5K	37	Parkhi et al. (2012)	-	Image Cls.

Continued on next page

Table 3 – continued from previous page

Dataset	Training Samples	Num. of Classes	Source	PT Task	DS Task
Oxford Flowers	102	>5K	Nilsback & Zisserman (2008)	-	Image Cls.
ImageNet100	130K	100	Tian et al. (2020a)	Inst. disc.	Image Cls.
ImageNet1K	1300K	1000	Deng et al. (2009)	Paired Emb. tasks, MIM, etc.	Image Cls.
Places205	2.4M	205	Zhou et al. (2014)		Scene Classification
PACS	5156	-	Li et al. (2017)	Jigsaw, Reconstruction	-
ADE20K	20K	150	Zhou et al. (2017)		Semantic seg.
PASCAL VOC	3K	20	Everingham et al. (2010)	-	Obj. Det. & Seg.
MS COCO	328K	91	Lin et al. (2014)	-	Obj. Det., Inst. / Semantic Seg.
NYU-depth v2	407,024 (unlab.) + 1449 (lab.)	1000+	Silberman et al. (2012)	-	Depth Estimation
CityScapes	25K	30	Cordts et al. (2016)	-	Semantic Seg.
JFT-300M	300M	18291	Sun et al. (2017)	Inst. disc.	Multilabel cls.
YFCC100M	100M	-	Thomee et al. (2016)	Inst. disc.	-
SUN397	108754	397	Xiao et al. (2014)	-	Scene classification

Table 4: Summary of Video Datasets used in Self-supervised pre-training

Dataset	Training Samples	Num. of Classes	Source	PT Task	DS Task
Video Datasets					
Moment in Time	1M	339	Monfort et al. (2019)	Context / Inst. disc.	Action Cls.
Kinetics600	500K	600	Kay et al. (2017)	Context / Inst. disc.	Action Cls., Video ret.
Kinetics400	240K	400	Kay et al. (2017)	Paired Emb. tasks	Action Cls, Video ret.
UCF101	13K	101	Soomro et al. (2012)	Context	Action Cls., Video ret.
HMDB51	6766	51	Kuehne et al. (2011)	Context	Action Cls., Video ret.
ActivityNet	19995	200	Heilbron et al. (2015)		Action cls.
BreakFast	≈2K	48	Kuehne et al. (2014)	-	Action cls. & seg.
FineGym	32697	530	Shao et al. (2020)	-	Action cls. & Seg.
50Salads	50	19	Stein & McKenna (2013)	-	Action cls. & Seg.
R2V2	2,788,424	-	Gordon et al. (2020)	Inst. disc.	-
OTB	100	-	Wu et al. (2015b)	-	Object Tracking
Something-Something	100K	174	Goyal et al. (2017)	-	Action cls.
FCVID	91223	239	Jiang et al. (2018)	-	Video Categorization
VidSitu	29.2K	-	Sadhu et al. (2021)		Video / Movie Und.
AudioSet	2M	632	Gemmeke et al. (2017)	Multi-modal Inst. disc.	Audio Event rec.
HowTo100M	136M	2K	Miech et al. (2019)	-	Action cls.
YouCook2	2K	89	Zhou et al. (2018)	-	Action rec.
MSR-VTT	10000	20	Xu et al. (2016)	-	Video Retrieval
DAVIS 2017	150	-	Pont-Tuset et al. (2018)	-	Video Obj Seg.
Diving48	16067	48	Li et al. (2018)	-	Diving action Cls.
SoundNet	2M	-	Aytar et al. (2016)	Multi-modal inst. disc.	Visual and Sound Cls.
AVA	427	80	Gu et al. (2018)	Multi-modal inst. disc.	Action loc.

4.3 Medical Datasets

In Table 5, 6, 7, 8, 9, 10, 11, 12, and 13, we present the different medical data sets used in self-supervised learning for magnetic resonance imaging (MRI), coherence tomography (CT), ultrasound (USG), radiographs, electrocardiogram (ECG), retinal fundus image, endoscopic videos, histopathological images, and

skin image datasets, respectively. We have also tried to note the purpose for which the respective datasets were used. Note that not all datasets were used for self-supervised pretraining. The datasets which were only used for downstream evaluation have an empty "PT Task" cell. We have also provided the source for each of the datasets in each table and also the part of the human body from which the respective datasets are taken. However, apart from the documented datasets, there are some datasets that are not open to public access. Some works have used private datasets, or not provided proper citations to the datasets used, and hence could not be documented. We also eliminated duplicate entries from the datasets. Some abbreviations used in the tables, such as, 'cls.', 'seg.', 'det.', 'rec.' denote classification, segmentation, detection, and recognition, respectively. 'Inst. disc.' denotes instance discrimination.

Table 5: Summary of Magnetic Resonance Imaging (MRI) Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
MRI Datasets					
BraTS 2018	∼300	Brain	Menze et al. (2015)	-	Tumor seg.
BraTS 2021	1251	Brain	Menze et al. (2015)	-	Tumor Seg.
KneeMRI	917	Knee (ACL)	Štajduhar et al. (2017)	-	ACL Tear severity cls.
MRNet	1370	Knee	Bien et al. (2018)	Jigsaw	Knee injury diag.
Human Connectome Project	1200	Brain	Elam & Van Essen (2022)	Reconstruction	-
M&Ms	375	Heart	Campello et al. (2021)	Reconstuction	Cardiac seg.
ACDC	150	Heart	Bernard et al. (2018)		Cardiac seg. & heart disease cls.
MSD (heart)	30	Heart	Antonelli et al. (2022)	-	Left-atrium seg.
ADNI	819	Brain	Petersen et al. (2010)	-	Alzheimer's pred.
PI-CAI	1500	Prostate	Saha et al. (2022)	Context, Inst. disc.	Prostate cancer diag.
Prostate158	158	Prostate	Adams et al. (2022)	Context, Inst. disc.	Prostate cancer diag.
CRL Fetal	81	Fetal brain	Gholipour et al. (2017)	Reconstruction	Segmentation and analysis
OASIS	434	Brain	Marcus et al. (2007)	Reconstruction	Segmentation
CANDI	103	Brain	Kennedy et al. (2012)	Reconstruction	Segmentation
BigBrain	7404	Brain	Amunts et al. (2013)	Reconstruction	Modeling
UMCL Multi-rater Consensus	30	Brain	Lesjak et al. (2018)	-	White matter /sclerosis lesion seg.
MICCAI MSSeg 2016 Challenge	53	Brain	Commowick et al. (2016)	-	Sclerosis lesion seg.
Longitudinal MS Lesion Segmentation Challenge (ISBI 2015)	82	Brain	Carass et al. (2017)	-	Sclerosis lesion sg.
MRI-WHS	60	Heart	Gao et al. (2023)	-	Whole Heart seg.
MRBrainS18	7	Brain	Kuijf et al. (2024)		Brain structure seg.
Left Atrium (LA) dataset	100	Heart	Xiong et al. (2021b)	Inst. disc.	Left atrium seg.
OASIS3	2842	Brain	LaMontagne et al. (2019)	Reconstruction	Alzheimer's det

Table 6: Summary of Coherence Tomography (CT) Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
CT Datasets					

Continued on next page

Table 6 – continued from previous page

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge	50	Abdomen	Challenge (2015)	Inst. Disc., Few Shot seg.	Few shot seg., organ seg.
LUNA 2016	888	Lung	Setio et al. (2017)	Context, Inst. disc.	Lung nodule seg.
NIH Pancreas-CT	82	Pancreas	Roth et al. (2016)	-	Pancreas seg.
LIDC-IDRI	7371	Lung	Armato et al. (2011)	Reconstruction, Inst. disc.	Lung nodule seg.
CAD-PE	91	Lung	Gonzalez Serrano (2019)	-	Pulmoray embolism det. & seg.
LiTS 2017	200	Liver	Bilic et al. (2023)	Reconstruction	Liver tumor seg.
TCIA-COVID19	461	Lung	Desai et al. (2020)	Reconstruction	Lung diseae diag.
C4KC-KiTS	621	Kidney	Heller et al. (2019)	Reconstruction	Kidney tumor seg.
NIH Lymph Nodes	352	Abdomen and Medi-astinum	Roth et al. (2015)	Reconstruction	Lymphadenopathy
Sliver07	40	Liver	Heimann et al. (2009)	-	Liver Seg.
CHAOS	40	Abdomen	Valindria et al. (2018)	Inst. disc., Few shot seg.	Few shot seg., Multi organ Seg.
3Dircadb-01,02	20,2	Liver	Soler et al. (2010)		Hepatic tumor seg.
COPDGene	947	Lung	Regan et al. (2010)	Inst. disc.	COPD (Ephysema) det.
MosMed	1110	Lung	Morozov et al. (2020)	Inst. disc.	COVID19 severity cls.
DeepLesion	10594	multiple	Yan et al. (2018)	Instance disc.	Lesion det.
FLARE	511	Abdomen	Ma et al. (2022)	Inst. disc.	Abdominal organ seg.
AMOS	500	Abdomen	Ji et al. (2022)	-	Abdominal organ seg.
NSCLC Radiomics	1265	Lung	Aerts et al. (2019)	Inst. Disc.	Tumor seg.
NLST Lung Cancer	203,099	Lung	National Lung Screening Trial Research Team (2013)	-	Lung cancer det.
MIDRC-RICORD-1A	120	Chest	Tsai et al. (2020)		Thoracic seg.
SDOCT	154	Eye	Tee et al. (2017)	Context based	Retinal disease diag.
GAMMA	300	Eye	Wu et al. (2023a)	-	Glaucoma grading
OCTA500	500	Eye	Li et al. (2024)	-	Retinal seg.

Table 7: Summary of Ultrasound Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Ultrasound Datasets					
Thyroid Nodule Segmentation	466	Neck	Pedraza et al. (2015)	Reconstruction	Thyroid lesion, cystic nodules, adenomas seg.

Table 8: Summary of Radiograph Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Radiograph Datasets					
CheXpert	224316	Chest	Irvin et al. (2019)	Inst. disc., Context	Chest disease cls.
ChestX-ray8	108948	Chest	Wang et al. (2017a)	Inst. disc., Context	Chest disease cls.
ChestX-Ray14	112120	Chest	Wang et al. (2017a)	Inst. disc., Context	Chest disease cls.
SIIM-ACR-2019	15000	Chest	Zawacki et al. (2019)	Inst. disc., Context	Pneumothorax Seg.

Continued on next page

Table 8 – continued from previous page

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Montgomery	138	Chest	Jaeger et al. (2014)	-	Pneumothorax Seg.
MIMIC-CXR v2	371920	Chest	Johnson et al. (2019)	Contrastive	Disease cls.
EdemaSeverity	16108	Chest	Liao et al. (2021a)	-	Edema Severity cls.

Table 9: Summary of ECG Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
ECG Datasets					
MIT-BIH	48 $\frac{1}{2}$-hour @ 360Hz	Heart	Moody & Mark (2001)	Reconstruction	Arythmia det.
PTB	549 @ 1KHz	Heart	Bousseljot et al. (1995)	-	Cardiac disease det.
PTB-XL	21837	71	Wagner et al. (2022)	Contrastive	Cardiac anomaly det.
Tianchi ECG	31779 @ 500 Hz	Heart	-	-	Cardiac anomaly det.
CinC2020	43,093	Heart	Perez Alday et al. (2021)	Contrastive	Cardiac anomaly det.
Chapman	10,646	Heart	Zheng et al. (2020)	Contrastive	Cardiovascular condition det.

Table 10: Summary of Retinal Fundus Image Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Retinal Fundus Image Datasets					
EyePACS	35126	Eye	Dugas et al. (2015)	Inst. disc.	Diabetic Retinopathy det.
APTOS 2019	3662	Eye	Karthik (2019)	Inst. disc.	Blindness det.
Messidor	1200	Eye	Decencière et al. (2014)		Diabetic retinopathy & risk of macular edema
IchallengeAMD	1200	Eye	iChallenge (2018)	Inst. disc., Context	age-related macular degeneration
IchallengePM	1200	Eye	iChallenge (2018)	Inst. disc., Context	pathological myopia
Isfahan MISP	59	Eye	Hajeb Mohammad Alipour et al. (2012a)	reconstruction	retinography-angiography registration
DRIVE	40	Eye	Staal et al. (2004)	-	Blood vessel seg. and optic disc loc.
DRIONS	110	Eye	Carmona et al. (2008)	-	optic disc seg.
IDRiD	516	Eye	Porwal et al. (2018)	-	Diabetic retinopathy cls. and fovea loc.
REFUGE	800	Eye	Orlando et al. (2020)	-	glaucoma det.
ADAM	400	Eye	Fang et al. (2022)	Inst. disc.	age-related macular degeneration
DRISHTI-GS	101	Eye	Sivaswamy et al. (2014)	Inst. disc.	glaucoma det.
RFMiD	3200	Eye	Pachade et al. (2021)	Context	fundus disease cls.
PALM	1200	Eye	Fu et al. (2019)	Context	disc and atrophy segmentation
Fundus FFA	70	Eye	Hajeb Mohammad Alipour et al. (2012b)	Context	Diabetic retinopathy cls.

Table 11: Summary of Endoscopic Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Endoscopy Datasets					
CVC-ClinicDB	612	Colon	Bernal et al. (2015)	Context based	Polyp seg.
CVC-ColonDB	300	”	Vázquez et al. (2017)	Context based	Polyp seg.
Kvasir Seg	1000	”	Jha et al. (2020)	Context based	Polyp seg.
ETIS Larib	196	”	Bernal et al. (2017)	Context based	Polyp seg.
LD-PolypDB	160	”	Ma et al. (2021c)	Non-contrastive siamese	Polyp seg. and det.
CVC-VideoClinicDB	40	”	Bernal et al. (2018)	-	Polyp seg. and det.

Table 12: Summary of Histopathological Image Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Histopathology Datasets					
DSB 2018	4470	Nucleus	Goodman et al. (2018)	denoising / context based	Nucleus seg.
BBBC 004	880	Nucleus	Ljosa et al. (2012)	”	Nucleus seg.
BACH	400	Breast	Polónia et al. (2019)		Breast cancer cls. and seg.
BreakHisv1	9109	Breast	BreakHis (2018)		breast cancer cls.
NCT-CRC-HE-100K	100K	Colorectal	Kather et al. (2018)	Inst. disc.	colorectal cancer tissue cls
Gleason2019	333	Prostate	Nir et al. (2018)	-	prostate cancer cls.
DigestPath2019	687	Intestine, colon	Da et al. (2022)	-	early-stage colon tumors seg.
BreastPathQ	3700	Breast	Petrick et al. (2021)	-	cancer cellularity
Camelyon16	271	Breast	Ehteshami Bejnordi et al. (2017)	-	Breast cancer metastasis det.
PAIP	100	Liver	Kim et al. (2021)		Liver cancer seg. & Viable tumor burden estimation
TissueNet	-	Cervix	Greenwald et al. (2022)		Cervical epithelial lesion cls.
Cell Tracking Challenge	-	-	Ulman et al. (2017)	-	Cell tracking and seg.
Electron Microscopy	165	Cell	Lucchi et al. (2013)	-	Mitochondria seg.
TNBC	50	Breast	Naylor et al. (2019)	-	Cell seg.
MoNuSeg	30	-	Kumar et al. (2017)	Context	Nuclei seg.
CoNSeP	41	Colon & rectum	Graham et al. (2019)	Context	Nuclei seg.

Table 13: Summary of Skin Image Datasets used in Self-supervised pre-training

Dataset	Training Samples	Body Part	Source	PT Task	DS Task
Skin Images					
ISIC 2020	33126	Skin	Rotemberg et al. (2021)	Inst. disc., Reconstruction	Melanoma cls.
ISIC 2018	2594	Skin	Codella et al. (2019)	Reconstruction	Melanoma cls. and segmentation
ISIC 2017	2750	Skin	Codella et al. (2018)	”	Melanoma segmentation
HAM10000	10015	Skin	Tschandl et al. (2018)	-	Skin lesion cls. and seg.

5 Comparison of self-supervised frameworks on benchmark datasets

In this survey, we discuss a plethora of frameworks. However, to truly assess the effectiveness of the frameworks, we need to look into the performance of those frameworks on a few benchmark datasets. Although we have tried to provide comparisons on the same downstream or target datasets for all frameworks, it is to be noted that the pre-training conditions may differ in some.

5.1 Comparison of Image-based SSL frameworks

In this subsection, we compare the image based SSL frameworks based on the performance on (1) the ImageNet1K classification task, (2) Classification, detection, and segmentation tasks on the PASCAL VOC dataset, and (3) Object detection and Instance segmentation tasks in MS COCO dataset. In some frameworks, the version of the PASCAL VOC dataset used for fine-tuning in the downstream task varies between VOC2007, VOC2012 (†), or VOC07+12 (∗), and has been discriminatively indicated in the table. The results in Table 14 are either obtained from the original papers or from papers that are mentioned in the table (Table 14) and have compared to those works in their respective manuscripts, that is, the results are cross-verified.

Besides that, we often find that the nature of the backbone encoder, or the number of pre-training epochs do not match for all the frameworks. As the domain has evolved, so has the choice of hyperparameters to obtain better performance on the benchmark datasets. However, we have done our best to document every detail of notable work done right from the advent of SSL with works like Agrawal et al. (2015) or Pathak et al. (2016) to the recent works like SMoG (Pang et al., 2022b), I-JEPA (Assran et al., 2023a), ConvNextv2 (Woo et al., 2023), etc.

All the frameworks mentioned in Table 14, are pre-trained on ImageNet1K (Deng et al., 2009) dataset for varying number of epochs as per their respective needs. We have attempted to present a comparative analysis using both linear classification (using frozen encoder) and fine-tuning Top-1 accuracy on the ImageNet-1K dataset. We also present our findings on the PASCAL VOC dataset for object classification (mAP), detection (AP50) and segmentation (mIOU) tasks. Unless otherwise mentioned, the default dataset for PASCAL VOC tasks is VOC2007. There are a few frameworks that opt for the VOC2012 and VOC07+12 versions of the dataset for finetuning. We also report the bounding box AP (AP_{bb}) and mask AP (AP_{mk}) in the MS COCO dataset.

Table 14: Comparison of performance of a few notable image-based self-supervised frameworks. † and ∗ indicate the use of PASCAL VOC2012 and PASCAL VOC07+12, respectively. ‡ and § indicate that the encoder was pre-trained for 400 and 200 epochs, respectively.

Frameworks	Encoder	ImageNet		PASCAL VOC07			MS-COCO		Eps.
		Lin.	FT	Cls.	Det.	Seg.	Det.	ISeg.	
Context based frameworks									
Agrawal et al. (2015)		-	-	54.2	43.9	-	-	-	
Context (Doersch et al., 2015)		31.7	45.6	65.3	51.1	38.4	-	-	
Pathak et al. (2016)		21.0	-	56.5	44.5	29.7	-	-	
Wang & Gupta (2015)		-	38.8	63.1	47.4	35.4	-	-	
Colorization (Zhang et al., 2016)	AlexNet	31.5	40.7	65.6	46.9	35.6	-	-	
Counting (Noroozi et al., 2017)		34.3	-	67.7	51.4	36.6	-	-	
ColorProxy (Larsson et al., 2017)		-	-	65.9		38.4	-	-	
Jigsaw (Noroozi & Favaro, 2016)		34.0	45.3	67.6	53.2	37.6	-	-	
RotNet (Gidaris et al., 2018)		38.7	50.0	72.97	54.4	39.1	-	-	
MIM based frameworks									
BEiT (Bao et al., 2021)		56.7	83.2	-	-	-	50.1	43.5	
mc-BEiT (Li et al., 2022e)							50.1	43.1	
iBOT (Zhou et al., 2022a)		79.5	84.0	-	-	-	51.2	44.2	
MAE (He et al., 2022)		-	83.6	-	-	-	50.3	44.9	
LocalMIM (Wang et al., 2023a)	ViT-B	-	84.0	-	-	-	50.7	44.9	
SimMIM (Xie et al., 2022)		56.7	83.8				50.4	44.4	
CAE (Chen et al., 2023c)		71.4	83.9				52.9	45.5	

Continued on next page

Table 14 – continued from previous page

Frameworks	Architecture	ImageNet		PASCAL VOC			MS-COCO		Eps.
		Lin.	FT	Cls.	Det.	Seg.	Det.	ISeg.	
BootMAE (Dong et al., 2022)		66.1	84.2				48.5	43.4	
ConvNext v2 (Woo et al., 2023)		-	84.9				52.9	46.6	
Clustering-based frameworks									
DeepCluster (Caron et al., 2018)	AlexNet	41.0	-	73.7	55.4	45.1	-	-	
UIC (Chen et al., 2020d)	AlexNet	41.6	-	75.9	54.9	45.9	-	-	
ODC (Zhan et al., 2020)	AlexNet (ResNet50)	41.4 (55.7)	-	(78.2)	-	-	-	-	
LA (Zhuang et al., 2019a)	ResNet50 (AlexNet)	60.2 (42.4)			53.5				
CLIM (Li et al., 2020b)	ResNet50	75.5	-	82.8	-	-	41.8	37.7	
CoKe (Qian et al., 2021a)	ResNet50	76.4	-	83.2			40.9	37.2	
Instance Discrimination frameworks									
CPC (van den Oord et al., 2018)	ResNetv2 101	48.7							
PIRL (Misra & van der Maaten, 2019)	ResNet50	63.6	-	81.1	80.7*				
MoCo (He et al., 2020)	ResNet50 (RN50w4×)	60.6 (68.6)	-	-	81.5*		40.8	36.9	
SimCLR (Chen et al., 2020b)	ResNet50 (RN50w4×)	69.3 (76.5)	89.0 (93.2)	86.6	79.4		38.5	34.8	
MoCov2 (Chen et al., 2020e)	ResNet50	71.1		82.5	82.4†§		39.8	36.1	
CPCv2 (Hénaff et al., 2020)	ResNet50 (ResNet161*)	63.8 (71.5)	85.3 (90.1)	(76.6)					
InfoMin (Tian et al., 2020b)	ResNet50	73.0	91.1		82.7		42.5	38.4	
SimCLRv2 (Chen et al., 2020c)	ResNet50 (RN152 w3×+SK)	71.7 (79.8)							
DenseCL (Wang et al., 2020)	ResNet50				82.8*	69.4*	40.3	36.4	
MoCov3 (Chen et al., 2021d)	ViT-B	76.7	83.2	-	-	-	47.9	42.7	
SSL-HSIC (Li et al., 2021d)	ResNet50 (RN200w2×)	74.8 (79.6)		84.1	76.0†		41.3	36.8	
PCL (Li et al., 2021b)	ResNet50-MLP	67.6		85.4	71.7 (78.5*)				
MUGS (Zhou et al., 2022c)	ViT-B	80.6	84.3				49.8	43.0	
SeLa (Asano et al., 2020b)	ResNet50 (AlexNet)	61.5		77.2	59.2	45.7			
SwAV (Caron et al., 2020)	ResNet-50	75.3		88.9	82.6*		42.1		
SMoG (Pang et al., 2022b)	ResNet50 (RN50w4×)	76.4 (79.0)		85.01†	76.2†		40.1	36.9	
MDRA (Cheng et al., 2023)	ResNet50	71.9					40.2	36.0	
Dimension-Contrastive frameworks									
BYOL (Grill et al., 2020)	ResNet50 (RN200w2×)	74.3 (79.6)		85.4	77.5†	76.3†	38.4	34.9	
I-JEPA (Assran et al., 2023a)	ViT-B	72.9							
Barlow Twins (Zbontar et al., 2021)	ResNet50	73.2		86.2	82.6*		39.2	34.3	
VICReg (Bardes et al., 2022a)	ResNet50	73.2		86.6	82.4*		39.4	36.4	
W-MSE (Ermolov et al., 2021)	ResNet50	72.56‡							
Zero-CL (Zhang et al., 2022d)	ResNet50	72.6‡							
Non-Contrastive frameworks									
SimSiam (Chen & He, 2020)	ResNet50	71.3			48.5§		39.2§	34.4§	
DINO (Caron et al., 2021)	ViT-B	78.2	83.6	-	-	-	50.1	43.4	
ReSSL (Zheng et al., 2021b)	ResNet50 (+5crops)	69.9§ (74.7)							
OBoW (Gidaris et al., 2021)	ResNet50	73.8§		89.3	82.9*				
MSN (Assran et al., 2022)	ViT-L (ViT-B)	80.7	83.4						
CrOC (Stegmüller et al., 2023)	ViT-S				70.6				

5.2 Comparison of different frameworks on Video benchmark

In this subsection, we present the comparison of several SSL frameworks on benchmark video datasets. Similar to the image-based frameworks, the base encoder architecture choice differs between different frameworks, as does the choice pre-training dataset. For UCF101, HMDB51, and Kinetics400 datasets, we report the action recognition accuracy. For UCF101 and HMDB51, we report the average accuracy over the 3 splits. Unless otherwise mentioned, the values mentioned in Table 15, are obtained after finetuning on the respective datasets.

Table 15: Comparison of performance of a few notable video-based self-supervised frameworks. The results on UCF-101 and HMDB-51 were obtained after finetuning. The results on the Kinetics dataset are obtained by linear evaluation unless otherwise mentioned. *lin* indicates Linear Evaluation. † indicates that the use of Kinetics600 instead of Kinsteics400 for finetuning and evaluation.

Frameworks	Encoder	Pretrain Data	UCF-101	HMDB51	Kinetics 400
Context based frameworks					
Shuffle & Learn (Misra et al., 2016)	CaffeNet	UCF-101	50.9	19.8	
3DRotNet (Jing et al., 2019)	3D RN18	Kinetics600	76.6	47.0	
VCOP (Xu et al., 2019)	R(2+1)D-18	Kinetics	72.4	30.9	
VidCloze (Luo et al., 2020)	C3D		68.5	32.5	
OOO (Fernando et al., 2017)	AlexNet	UCF101, HMDB51	60.0	32.4	
SkipClip (El-Nouby et al., 2019)	3D RN18	UCF-101	64.4		
(Jenni et al., 2020)	3D RN18 [R(2+1)D-18]	Kinetics600 [UCF-101]	79.3 [81.6]	49.8 [46.4]	
CPNet (Liang et al., 2022)	R(2+1)D-18	UCF-101 [Kinetics400]	81.8 [83.8]	51.2 [57.1]	
TransRank (Duan et al., 2022)	R(2+1)D-18	Kinetics200	90.7	64.2	
MIM based frameworks					
BEVT (Wang et al., 2022c)	Video-SWIN	ImageNet1K+ Kinetics400			81.1
VideoMAE (Tong et al., 2022)	ViT-B	Kinetics400	96.1	73.3	81.5
VideoMAEv2 (Wang et al., 2023c)	ViT-H	Kinetics400	99.6	88.1	88.6
AdaMAE (Bandara et al., 2023)	ViT-B	Kinetics400			81.7
OmniMAE (Girdhar et al., 2023)	ViT-B (ViT-H)	ImageNet1K+ SSv2			80.6 (85.4)
Instance discrimination based frameworks					
CoCLR (Han et al., 2020)	S3D	UCF-101 [Kinetics400]	87.1 [90.6]	58.7 [62.9]	
CVRL (Qian et al., 2021b)	R3D-152w2×	Kinetics600	93.9	69.9	72.9
SCVRL (Dorkenwald et al., 2022)	MViT-B	Kinetics400	89.0	62.6	
FAME (Ding et al., 2022)	R(2+1)D [I3D]	Kinetics400	84.8 [88.6]	53.5 [61.1]	
HDC (Zhang & Crandall, 2022)	R(2+1)D-10	Kinetics400	76.8	40.0	
SeCo (Yao et al., 2020)	3D RN18	Kinetics400	88.26	55.55	50.81^{lin}
VINCE (Gordon et al., 2020)	ResNet50	Kinetics400			49.1^{lin}
VCLR (Kuang et al., 2021)	R2D-50	Kinetics400	85.6	54.1	64.1^{lin}
Dimension Contrastive frameworks					
V-JEPA (Bardes et al., 2024)	ViT-L	Kinetics4000 [VideoMix2M]			78.7 [79.1]
Non-contrastive frameworks					
BraVe (Recasens et al., 2021)	R3D50	Kinetics600	95.1	74.3	$68.1^{lin\dagger}$

6 Summary and Conclusion

In this survey, we take a different approach to reviewing the work done in the domain of self-supervised learning. Firstly, we divided the works into several categories according to the approaches taken. We then try to further categorize each into finer subcategories. This allows us to learn the different avenues of research pursued in the past years and also the research directions currently being pursued. The finer discussions in each subcategory allow us to understand the differences between the frameworks or approaches better.

Furthermore, the massive number of work done in the last few years on masked image modelling shows the potential of the approach in visual self-supervised learning. Among the contrastive learning approaches, SimCLR, MoCov2, BYOL, SimSiam are the most popular and were also adopted for several downstream applications.

From the review of the works done in medical image analysis using self-supervised frameworks, it becomes evident that a majority of the work done has used MR or CT scans primarily. This is mainly due to the easy availability of benchmark datasets for the MR or CT modality. We can also observe that self-supervised

learning has been applied to many medical imaging modalities. This definitely indicates the acceptability and adaptability of SSL in the current computer vision domain.

References

Lisa C Adams, Marcus R Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K Bressem. Prostate158 - an expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Comput. Biol. Med.*, 148(105817):105817, September 2022.

Sravanti Addepalli, Kaushal Bhogale, Priyam Dey, and R. Venkatesh Babu. Towards efficient and effective self-supervised learning of visual representations. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 523–538, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Hugo J W L Aerts, Leonard Wee, Emmanuel Rios Velazquez, Ralph T H Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ren Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M Rietbergen, C Ren Leemans, Andre Dekker, John Quackenbush, Robert J Gillies, and Philippe Lambin. Nsclc-radiomics, 2019.

P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 37–45, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.13. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.13.

Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 179–189, 2019. doi: 10.1109/WACV.2019.00025.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24206–24221, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cb3213ada48302953cb0f166464ab356-Abstract.html.

Mehmet Akçakaya, Burhaneddin Yaman, Hyungjin Chung, and Jong Chul Ye. Unsupervised deep learning methods for biological image reconstruction and enhancement: An overview from a signal processing perspective. *IEEE Signal Processing Magazine*, 39(2):28–44, 2022. doi: 10.1109/MSP.2021.3119273.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=Uuf2q9TfXGA.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6f2268bd1d3d3ebaabb04d6b5d099425-Abstract.html.

Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 116–132, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Katrin Amunts, Claude Lepage, Louis Borgeat, Hartmut Mohlberg, Timo Dickscheid, Marc-Étienne Rousseau, Sebastian Bludau, Pierre-Louis Bazin, Lindsay B Lewis, Ana-Maria Oros-Peusquens, Nadim J Shah, Thomas Lippert, Karl Zilles, and Alan C Evans. BigBrain: an ultrahigh-resolution 3D human brain model. *Science*, 340(6139):1472–1475, June 2013.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015. doi: 10.1109/ICCV.2015.279.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F Christ, Richard K G Do, Marc J Gollub, Stephan H Heckers, Henkjan Huisman, William R Jarnagin, Maureen K McHugo, Sandy Napel, Jennifer S Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L Simpson, Lena Maier-Hein, and M Jorge Cardoso. The medical segmentation decathlon. *Nat. Commun.*, 13(1):4128, July 2022.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017.

Samuel G Armato, 3rd, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, Ella A Kazerooni, Heber MacMahon, Edwin J R Van Beeke, David Yankelevitz, Alberto M Biancardi, Peyton H Bland, Matthew S Brown, Roger M Engelmann, Gary E Laderach, Daniel Max, Richard C Pais, David P Y Qing, Rachael Y Roberts, Amanda R Smith, Adam Starkey, Poonam Batrah, Philip Caligiuri, Ali Farooqi, Gregory W Gladish, C Matilda Jude, Reginald F Munden, Iva Petkovska, Leslie E Quint, Lawrence H Schwartz, Baskaran Sundaram, Lori E Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallamm, Michael D Heath, Michael H Kuhn, Ekta Dharaiya, Richard Burns, David S Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, and Barbara Y Croft. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.*, 38(2):915–931, February 2011.

Yuki Markus Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/31fefc0e570cb3860f2a6d4b38c6490d-Abstract.html.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=Hyx-jyBFPr.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael G. Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:248178208.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver,*

BC, Canada, June 17-24, 2023, pp. 15619–15629. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/pdf?id=04K3PMtMckp.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 892–900, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C. Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Keith B. Hengen, William Gray-Roncal, Michal Valko, and Eva L. Dyer. Mine your own view: Self-supervised learning through across-sample prediction. *ArXiv*, abs/2102.10106, 2021. URL https://api.semanticscholar.org/CorpusID:231979539.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3458–3468, 2021. doi: 10.1109/ICCV48922.2021.00346.

Philip Bachman, R Devon Hjelm, and William Buchwalter. *Learning Representations by Maximizing Mutual Information across Views*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 541–549, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8.

P Baldi, P Sadowski, and D Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.*, 5(1):4308, July 2014.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14507–14517, June 2023.

Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. URL https://arxiv.org/abs/2106.08254.

Amir Bar, Xin Wang, Vadim Kantorov, Colorado J. Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14585–14595. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01420. URL https://doi.org/10.1109/CVPR52688.2022.01420.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/39cee562b91611c16ac0b100f0bc1ea1-Abstract-Conference.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *CoRR*, abs/2307.12698, 2023. doi: 10.48550/ARXIV.2307.12698. URL https://doi.org/10.48550/arXiv.2307.12698.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michal Górszczak, Barbara Rychalska, Tomasz Trzcinski, and Bartosz Zielinski. Explaining self-supervised image representations with visual probing. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 592–598. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/82. URL https://doi.org/10.24963/ijcai.2021/82.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark, 2023.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.*, 43:99–111, July 2015.

Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J. Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth R. Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, and Aymeric Histace. Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, 36(6):1231–1249, 2017. doi: 10.1109/TMI.2017.2664042.

Jorge J Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodriguez, Maroua Hammami, Ana Garcia-Rodriguez, Henry Córdova, Olivier Romain, Gloria Fernández-Esparrach, Xavier Dray, and Javier Sanchez. Polyp Detection Benchmark in Colonoscopy Videos using GTCreator: A Novel Fully Configurable Tool for Easy and Fast Annotation of Image Databases. In *Proceedings of 32nd CARS conference*, Berlin, Germany, June 2018. URL https://hal.science/hal-01846141.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Beril Besbinar and Pascal Frossard. Self-supervision by prediction for object discovery in videos. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1509–1513, 2021. doi: 10.1109/ICIP42928.2021.9506062.

P. Bhat, E. Arani, and B. Zonooz. Distill on the go: Online knowledge distillation in self-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2672–2681, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPRW53098.2021.00301. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW53098.2021.00301.

Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V. Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *Knowledge-Based Systems*, 256:109631, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109631. URL https://www.sciencedirect.com/science/article/pii/S095070512200822X.

Adam Bielski and Paolo Favaro. Move: Unsupervised movable object segmentation and detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33371–33386. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d7eb232f196124894f2e65b9010a5c57-Paper-Conference.pdf.

Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F Amanatullah, Christopher F Beaulieu, Geoffrey M Riley, Russell J Stewart, Francis G Blankenberg, David B Larson, Ricky H Jones, Curtis P Langlotz, Andrew Y Ng, and Matthew P Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.*, 15(11):e1002699, November 2018.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró-I-Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The liver tumor segmentation benchmark (LiTS). *Med. Image Anal.*, 84(102680): 102680, February 2023.

Maximilian Blendowski, Hannes Nickisch, and Mattias P. Heinrich. How to learn from unlabeled volume data: Self-supervised 3d context feature learning. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 649–657, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32226-7.

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:2168245.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedical Engineering / Biomedizinische Technik*, 40(s1):317–318, 1995. doi: doi:10.1515/bmte.1995.40.s1.317. URL https://doi.org/10.1515/bmte.1995.40.s1.317.

BreakHis. Breakhis, 2018. URL https://www.kaggle.com/datasets/ambarish/breakhis.

John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and 'phantom targets. In J. Moody, S. Hanson, and R.P. Lippmann (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/a8abb4bb284b5b27aa7cb790dc20f80b-Paper.pdf.

Himanshu Buckchash and Balasubramanian Raman. Sustained self-supervised pretraining for temporal order verification. In *Pattern Recognition and Machine Intelligence*, pp. 140–149, Cham, 2019. Springer International Publishing. ISBN 978-3-030-34869-4.

Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12638–12648. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9523147e5a6707baf674941812ee5c94-Paper.pdf.

Zhiyuan Cai, Li Lin, Huaqing He, and Xiaoying Tang. Uni4eye: Unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 88–98, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16452-1.

Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Vilades, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Cavus, Steffen E. Petersen, Sergio Escalera, Santi Segui, Jose F. Rodriguez-Palomares, and Karim Lekadir. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12): 3543–3554, 2021. doi: 10.1109/TMI.2021.3090082.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, Manuel Jorge Cardoso, Niamh Cawley, Olga Ciccarelli, Claudia A M Wheeler-Kingshott, Sébastien Ourselin, Laurence Catanese, Hrishikesh Deshpande, Pierre Maurel, Olivier Commowick, Christian Barillot, Xavier Tomas-Fernandez, Simon K Warfield, Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, Ganapathy Krishnamurthi, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Leonardo O Iheme, Devrim Unay, Saurabh Jain, Diana M Sima, Dirk Smeets, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Hayit Greenspan, Pierre-Louis Bazin, Peter A Calabresi, Ciprian M Crainiceanu, Lotta M Ellingsen, Daniel S Reich, Jerry L Prince, and Dzung L Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage*, 148:77–102, March 2017.

Enrique J Carmona, Mariano Rincón, Julián García-Feijoó, and José M Martínez-de-la Casa. Identification of the optic nerve head with genetic algorithms. *Artif. Intell. Med.*, 43(3):243–259, July 2008.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 139–156, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01264-9.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2959–2968, 2019. doi: 10.1109/ICCV.2019.00305.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. URL https://api.semanticscholar.org/CorpusID:233444273.

R. B. Cattell. Theory of fluid and crystallized intelligence: A critical experiment. 54(1):1–22, 1963. URL https://doi.org/10.1037/h0046743.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

S. Chakraborty, A. Gosthipaty, and S. Paul. G-simclr: Self-supervised contrastive learning with guided projection via pseudo labelling. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 912–916, Los Alamitos, CA, USA, nov 2020. IEEE Computer Society. doi: 10.1109/ICDMW51313.2020.00131. URL https://doi.ieeecomputersociety.org/10.1109/ICDMW51313.2020.00131.

Multi-Atlas Abdomen Labeling Challenge. Miccai 2015 multi-atlas abdomen labeling challenge, 2015. URL https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, Michael Picheny, and Shih-Fu Chang. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7992–8001, 2021a. doi: 10.1109/ICCV48922.2021.00791.

David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. City-scale landmark identification on mobile devices. In *CVPR 2011*, pp. 737–744, 2011. doi: 10.1109/CVPR.2011.5995610.

Jing Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16706–16715, 2023a. URL https://api.semanticscholar.org/CorpusID:258291429.

Jintai Chen, Xiangshang Zheng, Hongyun Yu, Danny Z. Chen, and Jian Wu. Electrocardio panorama: Synthesizing new ECG views with self-supervision. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 3597–3605. ijcai.org, 2021b. doi: 10.24963/IJCAI.2021/495. URL https://doi.org/10.24963/ijcai.2021/495.

Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. 06 2022a. doi: 10.48550/arXiv.2206.00790.

Junyu Chen, Eric C. Frey, Yufan He, William Paul Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Anal.*, 82:102615, 2022b. doi: 10.1016/J.MEDIA.2022.102615. URL https://doi.org/10.1016/j.media.2022.102615.

Kaiyi Chen, Qingbin Wang, and Yutao Ma. Cervical optical coherence tomography image classification based on contrastive self-supervised texture learning. *Med. Phys.*, 49(6):3638–3653, June 2022c.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101539. URL https://www.sciencedirect.com/science/article/pii/S1361841518304699.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/chen20s.html.

Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11526–11535, June 2021c.

Q. Chen, S. Cai, C. Cai, Z. Yu, D. Qian, and S. Xiang. Colo-scrl: Self-supervised contrastive representation learning for colonoscopic video retrieval. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1056–1061, Los Alamitos, CA, USA, jul 2023b. IEEE Computer Society. doi: 10.1109/ICME55011.2023.00185. URL https://doi.ieeecomputersociety.org/10.1109/ICME55011.2023.00185.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020b.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020c. Curran Associates Inc. ISBN 9781713829546.

Weijie Chen, Shiliang Pu, Di Xie, Shicai Yang, Yilu Guo, and Luojun Lin. Unsupervised image classification for deep representation learning. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, pp. 430–446, Berlin, Heidelberg, 2020d. Springer-Verlag. ISBN 978-3-030-66095-6. doi: 10.1007/978-3-030-66096-3_30. URL https://doi.org/10.1007/978-3-030-66096-3_30.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, Los Alamitos, CA, USA, oct 2021d. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00950. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00950.

Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021e. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.107826. URL https://www.sciencedirect.com/science/article/pii/S0031320321000133.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, Aug 2023c. ISSN 1573-1405. doi: 10.1007/s11263-023-01852-4. URL https://doi.org/10.1007/s11263-023-01852-4.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2020. URL https://api.semanticscholar.org/CorpusID:227118869.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020e. URL https://arxiv.org/abs/2003.04297.

Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pp. 108–124. Springer, 2022d. doi: 10.1007/978-3-031-20056-4_7. URL https://doi.org/10.1007/978-3-031-20056-4_7.

Yubei Chen, Zeyu Yun, Yi Ma, Bruno A. Olshausen, and Yann LeCun. Minimalistic unsupervised representation learning with the sparse manifold transform. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023d. URL https://openreview.net/pdf?id=nN_nBVKAhhD.

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, Venkat S. Sethuraman, and Kevin Brown. Masked image modeling advances 3d medical image analysis. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1969–1979, 2022e. URL https://api.semanticscholar.org/CorpusID:248376910.

Haoyang Cheng, Hongliang Li, Heqian Qiu, Qingbo Wu, Xiaoliang Zhang, Fanman Meng, and King Ngi Ngan. Unsupervised visual representation learning via multi-dimensional relationship alignment. *IEEE Transactions on Image Processing*, 32:1613–1626, 2023. doi: 10.1109/TIP.2023.3246801.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

Jan Chorowski, Grzegorz Ciesielski, Jarosław Dzikowski, Adrian Łańcucki, Ricard Marxer, Mateusz Opala, Piotr Pusz, Paweł Rychlikowski, and Michał Stypułkowski. Aligned Contrastive Predictive Coding. In *Proc. Interspeech 2021*, pp. 976–980, 2021. doi: 10.21437/Interspeech.2021-1544.

Xiangxiang Chu, Xiaohang Zhan, and Bo Zhang. A unified mixture-view framework for unsupervised representation learning. In *British Machine Vision Conference*, 2020. URL https://api.semanticscholar.org/CorpusID:252780465.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2021.100198. URL https://www.sciencedirect.com/science/article/pii/S2666827021000992.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates and Andrew Y. Ng. *Learning Feature Representations with K-Means*, pp. 561–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_30. URL https://doi.org/10.1007/978-3-642-35289-8_30.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

Elijah Cole, Xuan S. Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge J. Belongie. When does contrastive visual representation learning work? *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 01–10, 2021. URL https://api.semanticscholar.org/CorpusID:234469977.

Olivier Commowick, Frédéric Cervenansky, and Roxana Ameli. Msseg challenge proceedings: Multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016. URL https://api.semanticscholar.org/CorpusID:51996766.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.350. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.350.

Huseyin Coskun, Alireza Zareian, Joshua L. Moore, Federico Tombari, and Chen Wang. GOCA: guided online cluster assignment for self-supervised video representation learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pp. 1–22. Springer, 2022. doi: 10.1007/978-3-031-19821-2_1. URL https://doi.org/10.1007/978-3-031-19821-2_1.

Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, Hongmei Yi, Yan Guo, Zhe Wang, Ling Chen, Li Zhang, Xianying He, Xiaofan Zhang, Ke Mei, Chuang Zhu, Weizeng Lu, Linlin Shen, Jun Shi, Jun Li, Sreehari S, Ganapathy Krishnamurthi, Jiangcheng Yang, Tiancheng Lin, Qingyu Song, Xuechen Liu, Simon Graham, Raja Muhammad Saad Bashir, Canqian Yang, Shaofei Qin, Xinmei Tian, Baocai Yin, Jie Zhao, Dimitris N Metaxas, Hongsheng Li, Chaofu Wang, and Shaoting Zhang. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Med. Image Anal.*, 80(102485):102485, August 2022.

A. Dadashzadeh, A. Whone, and M. Mirmehdi. Auxiliary learning for self-supervised video representation via similarity-based knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4230–4239, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPRW56347.2022.00468. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00468.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1601–1610. Computer Vision Foundation / IEEE, 2021a. doi: 10.1109/CVPR46437.2021.00165. URL https://openaccess.thecvf.com/content/CVPR2021/html/Dai_UP-DETR_Unsupervised_Pre-Training_for_Object_Detection_With_Transformers_CVPR_2021_paper.html.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *CoRR*, abs/2106.04803, 2021b. URL https://arxiv.org/abs/2106.04803.

Ishan Rajendrakumar Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Comput. Vis. Image Underst.*, 219:103406, 2021. URL https://api.semanticscholar.org/CorpusID:231648417.

Youssef Dawoud, Arij Bouazizi, Katharina Ernst, G. Carneiro, and Vasileios Belagiannis. Knowing what to label for few shot microscopy image cell segmentation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3557–3566, 2022a. URL https://api.semanticscholar.org/CorpusID:253708381.

Youssef Dawoud, Katharina Ernst, Gustavo Carneiro, and Vasileios Belagiannis. Edge-based self-supervision for semi-supervised few-shot microscopy image cell segmentation. In *Medical Optical Imaging and Virtual Microscopy Image Analysis*, pp. 22–31, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-16961-8.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. Feedback on a publicly distributed image database: The messidor database. *Image Anal. Stereol.*, 33(3):231, August 2014.

Omer Burak Demirel, Burhaneddin Yaman, Logan Dowdle, Steen Moeller, Luca Vizioli, Essa Yacoub, John Strupp, Cheryl A. Olman, Kâmil Uğurbil, and Mehmet Akçakaya. Improved simultaneous multi-slice functional mri using self-supervised deep learning. In *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 890–894, 2021. doi: 10.1109/IEEECONF53345.2021.9723264.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, and Romain H'erault. Similarity contrastive estimation for image and video soft contrastive self-supervised learning. *Machine Vision and Applications*, 34:1–19, 2022. URL https://api.semanticscholar.org/CorpusID:254926500.

Virginia R. DeSa. Learning classification with unlabeled data. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pp. 112–119, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

Virginia R. DeSa. Minimizing disagreement for self-supervised classification. In *Proceedings of the 1993 Connectionist Models Summer School (1st ed.)*. Psychology Press, 1994.

Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri D Blake, Tracy Nolan, Thomas Powell, Kevin Sexton, William Bennett, and Fred Prior. Chest imaging with clinical and genomic correlates representing a rural COVID-19 positive population, 2020.

Aniket Anand Deshmukh, Jayanth Reddy Regatti, Eren Manavoglu, and Ürün Dogan. Representation learning for clustering via building consensus. *Mach. Learn.*, 111(12):4601–4638, 2022. doi: 10.1007/S10994-022-06194-9. URL https://doi.org/10.1007/s10994-022-06194-9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Fatemeh Taheri Dezaki, Christina L. Luong, Tom Ginsberg, Robert N. Rohling, Ken Gin, Purang Abolmaesumi, and Teresa S.M. Tsang. Echo-syncnet: Self-supervised cardiac view synchronization in echocardiography. *IEEE Transactions on Medical Imaging*, 40:2092–2104, 2021. URL https://api.semanticscholar.org/CorpusID:231802422.

Fei Ding, Dan Zhang, Yin Yang, Venkat Krovi, and Feng Luo. Contrastive representation disentanglement for clustering, 2023.

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9726, 2022.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Context as supervisory signal: Discovering objects with predictable context. In *Computer Vision – ECCV 2014*, pp. 362–377, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10578-9.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1422–1430, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.167. URL https://doi.org/10.1109/ICCV.2015.167.

Nanqing Dong, Michael Kampffmeyer, and Irina Voiculescu. Self-supervised multi-task representation learning for sequential medical images. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 779–794, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86523-8.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 247–264, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20056-4.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i1.25130. URL https://doi.org/10.1609/aaai.v37i1.25130.

Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4131–4140, 2022. doi: 10.1109/CVPRW56347.2022.00458.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

H. Duan, N. Zhao, K. Chen, and D. Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2990–3000, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00301. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00301.

Emma Dugas, Jared Jorge, and Will Cukierski. Diabetic retinopathy detection, 2015. URL https://kaggle.com/competitions/diabetic-retinopathy-detection.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9568–9577, 2021. URL https://api.semanticscholar.org/CorpusID:233444011.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A W M van der Laak, the CAMELYON16 Consortium, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory Crf van Dijk, Peter Bult, Francisco Beca, Andrew H Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov,

Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, December 2017.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The pfam protein families database in 2019. *Nucleic Acids Res.*, 47(D1):D427–D432, January 2019.

Alaaeldin El-Nouby, Shuangfei Zhai, Graham W. Taylor, and Joshua M. Susskind. Skip-clip: Self-supervised spatiotemporal representation learning by future clip order ranking. In *ICCV Workshop*, 2019. URL https://arxiv.org/pdf/1910.12770.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jégou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *CoRR*, abs/2112.10740, 2021. URL https://arxiv.org/abs/2112.10740.

Jennifer Stine Elam and David Van Essen. *Human Connectome Project*, pp. 1647–1650. Springer New York, New York, NY, 2022. ISBN 978-1-0716-1006-0. doi: 10.1007/978-1-0716-1006-0_592. URL https://doi.org/10.1007/978-1-0716-1006-0_592.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 2021. URL http://proceedings.mlr.press/v139/ermolov21a.html.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21480–21492, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b36ed8a07e3cd80ee37138524690eca1-Abstract.html.

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, Sharath M. Shankaranarayana, Yi-Ting Chen, Chuen-Heng Wang, Nisarg A. Shah, Chia-Yen Lee, Chih-Chung Hsu, Hai Xie, Baiying Lei, Ujjwal Baid, Shubham Innani, Kang Dang, Wenxiu Shi, Ravi Kamble, Nitin Singhal, Ching-Wei Wang, Shih-Chang Lo, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. doi: 10.1109/TMI.2022.3172773.

Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=09hVcSDkea.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. SEED: self-supervised distillation for visual representation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=AHm3dbp7D1D.

Alex Fedorov, Tristan Sylvain, Eloy Geenjaar, Margaux Luck, Lei Wu, Thomas P. DeRamus, Alex Kirilin, Dmitry Bleklov, Vince D. Calhoun, and Sergey M. Plis. Self-supervised multimodal domino: in search of biomarkers for alzheimer's disease. In *9th IEEE International Conference on Healthcare Informatics, ICHI 2021, Victoria, BC, Canada, August 9-12, 2021*, pp. 23–30. IEEE, 2021a. doi: 10.1109/ICHI52183.2021.00017. URL https://doi.org/10.1109/ICHI52183.2021.00017.

Alex Fedorov, Lei Wu, Tristan Sylvain, Margaux Luck, Thomas P. DeRamus, Dmitry Bleklov, Sergey M. Plis, and Vince D. Calhoun. On self-supervised multimodal representation learning: An application to alzheimer's disease. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1548–1552, 2021b. doi: 10.1109/ISBI48211.2021.9434103.

Alex Fedorov, Eloy Geenjaar, Lei Wu, Tristan Sylvain, Thomas P. DeRamus, Margaux Luck, Maria Misiura, Girish Mittapalle, R. Devon Hjelm, Sergey M. Plis, and Vince D. Calhoun. Self-supervised multimodal learning for group inferences from mri data: Discovering disorder-relevant brain regions and multimodal links. *NeuroImage*, 285:120485, 2024. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2023.120485. URL https://www.sciencedirect.com/science/article/pii/S1053811923006353.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: joint-embedding predictive architecture can listen. *CoRR*, abs/2311.15830, 2023a. doi: 10.48550/ARXIV.2311.15830. URL https://doi.org/10.48550/arXiv.2311.15830.

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Masked auto-encoders meet generative adversarial networks and beyond. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24449–24459. IEEE, 2023b. doi: 10.1109/CVPR52729.2023.02342. URL https://doi.org/10.1109/CVPR52729.2023.02342.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/e97d1081481a4017df96b51be31001d3-Abstract-Conference.html.

Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10364–10374. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01061. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Feng_Self-Supervised_Representation_Learning_by_Rotation_Feature_Decoupling_CVPR_2019_paper.html.

B. Fernando, H. Bilen, E. Gavves, and S. Gould. Self-supervised video representation learning with odd-one-out networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5729–5738, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.607. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.607.

Marc Fischer, Tobias Hepp, Sergios Gatidis, and Bin Yang. Self-supervised contrastive learning with random walks for medical image segmentation with limited annotations. *Computerized Medical Imaging and Graphics*, 104:102174, 2023. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2022.102174. URL https://www.sciencedirect.com/science/article/pii/S0895611122001446.

Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E. O'Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375, 2021. doi: 10.1109/ICASSP39728.2021.9415009.

Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. Palm: Pathologic myopia challenge, 2019. URL https://dx.doi.org/10.21227/55pk-8z03.

Jhair Gallardo, Tyler L. Hayes, and Christopher Kanan. Self-supervised training enhances online continual learning. In *British Machine Vision Conference*, 2021. URL https://api.semanticscholar.org/CorpusID:232352548.

Tianyuan Gan, Ziyi Jin, Liangliang Yu, Xiao Liang, Hong Zhang, and Xuesong Ye. Self-supervised representation learning using feature pyramid siamese networks for colorectal polyp detection. *Sci. Rep.*, 13(1): 21655, December 2023.

Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. MCMAE: masked convolution meets masked autoencoders. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/e7938ede51225b490bb69f7b361a9259-Abstract-Conference.html.

Shangqi Gao, Hangqi Zhou, Yibo Gao, and Xiahai Zhuang. BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Med. Image Anal.*, 89(102889):102889, October 2023.

Yuting Gao, Jia-Xin Zhuang, Ke Li, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Xing Sun. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *CoRR*, abs/2104.09124, 2021. URL https://arxiv.org/abs/2104.09124.

Chongjian Ge, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, and Ping Luo. Soft neighbors are positive supporters in contrastive visual representation learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=l9vM_PaUKz.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Ali Gholipour, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia M Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, Judy A Estroff, and Simon K Warfield. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Sci. Rep.*, 7(1), March 2017.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick P'erez, and Matthieu Cord. Learning representations by predicting bags of visual words. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6926–6936, 2020. URL https://api.semanticscholar.org/CorpusID:211532737.

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6830–6840. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00676. URL https://openaccess.thecvf.com/content/CVPR2021/html/Gidaris_OBoW_Online_Bag-of-Visual-Words_Generation_for_Self-Supervised_Learning_CVPR_2021_paper.html.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16081–16091. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01563. URL https://doi.org/10.1109/CVPR52688.2022.01563.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10406–10417. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01003. URL https://doi.org/10.1109/CVPR52729.2023.01003.

Alex Gomez-Villa, Bartlomiej Twardowski, Lu Yu, Andrew D. Bagdanov, and Joost van de Weijer. Continually learning self-supervised representations with projected functional regularization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3866–3876, 2021. URL https://api.semanticscholar.org/CorpusID:245634408.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 605–614, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43907-0.

German Gonzalez Serrano. Cad-pe, 2019. URL https://dx.doi.org/10.21227/9bw7-6823.

Allen Goodman, Anne Carpenter, Elizabeth Park, jlefman nvidia, Josette BoozAllen, Kyle, Maggie, Nilofer, Peter Sedivec, and Will Cukierski. 2018 data science bowl, 2018. URL https://kaggle.com/competitions/data-science-bowl-2018.

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *ArXiv*, abs/2003.07990, 2020. URL https://api.semanticscholar.org/CorpusID:212747934.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *ArXiv*, abs/2202.08360, 2022. URL https://api.semanticscholar.org/CorpusID:246904713.

R. Goyal, S. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5843–5851, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.622. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.622.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101563. URL https://www.sciencedirect.com/science/article/pii/S1361841519301045.

Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, Cole Pavelchek, Sunny Cui, Isabella Camplisson, Omer Bar-Tal, Jaiveer Singh, Mara Fong, Gautam Chaudhry, Zion Abraham, Jackson Moseley, Shiri Warshawsky, Erin Soon, Shirley Greenbaum, Tyler Risom, Travis Hollmann, Sean C Bendall, Leeat Keren, William Graf, Michael Angelo, and David Van Valen. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.*, 40(4):555–565, April 2022.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, 2005. URL https://api.semanticscholar.org/CorpusID:2179911.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018. doi: 10.1109/CVPR.2018.00633.

Jie Gui, Tuo Chen, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *CoRR*, abs/2301.05712, 2023. doi: 10.48550/ARXIV.2301.05712. URL https://doi.org/10.48550/arXiv.2301.05712.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1753–1759, 2017. doi: 10.24963/ijcai.2017/243. URL https://doi.org/10.24963/ijcai.2017/243.

Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/pdf?id=QAV2CcLEDh.

Ekta Gupta, Varun Gupta, Muskaan Chopra, Prakash Chandra Chhipa, and Marcus Liwicki. Learning self-supervised representations for label efficient cross-domain knowledge transfer on diabetic retinopathy fundus images. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pp. 1–7. IEEE, 2023b. doi: 10.1109/IJCNN54540.2023.10191796. URL https://doi.org/10.1109/IJCNN54540.2023.10191796.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012. URL http://jmlr.org/papers/v13/gutmann12a.html.

M. Gwilliam and Abhinav Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9632–9642, 2022. URL https://api.semanticscholar.org/CorpusID:249712495.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

F. Haghighi, M. Taher, M. B. Gotway, and J. Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20792–20802, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.02016. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.02016.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 137–147, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? *Medical Image Analysis*, pp. 103086, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103086. URL https://www.sciencedirect.com/science/article/pii/S1361841524000112.

Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi. Diabetic retinopathy grading by digital curvelet transform. *Computational and Mathematical Methods in Medicine*, 2012:761901, Sep 2012a. ISSN 1748-670X. doi: 10.1155/2012/761901. URL https://doi.org/10.1155/2012/761901.

Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi. Diabetic retinopathy grading by digital curvelet transform. *Comput. Math. Methods Med.*, 2012:761901, September 2012b.

T. Han, W. Xie, and A. Zisserman. Video representation learning by dense predictive coding. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1483–1492, Los Alamitos, CA, USA, oct 2019. IEEE Computer Society. doi: 10.1109/ICCVW.2019.00186. URL https://doi.ieeecomputersociety.org/10.1109/ICCVW.2019.00186.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235367888.

B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 447–456, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. doi: 10.1109/CVPR.2015.7298642. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298642.

Adam W. Harley, Shrinidhi Kowshika Lakshmikanth, Fangyu Li, Xian Zhou, Hsiao-Yu Fish Tung, and Katerina Fragkiadaki. Learning from unlabelled videos using contrastive predictive neural 3d mapping. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJxt60VtPr.

Bobby He and Mete Ozay. Exploring the gap between collapsed and whitened features in self-supervised learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8613–8634. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/he22c.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

Tobias Heimann, Bram van Ginneken, Martin A. Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, GyÖrgy Bekes, Fernando Bello, Gerd Binnig, Horst Bischof, Alexander Bornik, Peter M. M. Cashman, Ying Chi, AndrÉs Cordova, Benoit M. Dawant, MÁrta Fidrich, Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, Dagmar KainmÜller, Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, Hans-Peter Meinzer, GÁbor Nemeth, Daniela S. Raicu, Anne-Mareike Rau, Eva M. van Rikxoort, MikaËl Rousson, LÁszlÓ Rusko, Kinda A. Saddi, GÜnter Schmidt, Dieter Seghers, Akinobu Shimizu, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. Waite,

Andreas Wimmer, and Ivo Wolf. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009. doi: 10.1109/TMI.2009.2013851.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpaul, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. C4KC KiTS challenge kidney tumor segmentation dataset, 2019.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron Van Den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Retinal image understanding emerges from self-supervised multimodal reconstruction. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 321–328, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00928-1.

Álvaro S. Hervella, Lucía Ramos, J. Rouco, J. Novo, and Marcos Ortega. Multi-modal self-supervised pre-training for joint optic disc and cup segmentation in eye fundus images. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 961–965, 2020. URL https://api.semanticscholar.org/CorpusID:216483440.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

R. Devon Hjelm and Philip Bachman. Representation learning with video deep infomax. *CoRR*, abs/2007.13278, 2020. URL https://arxiv.org/abs/2007.13278.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. *CoRR*, abs/2010.12050, 2020. URL https://arxiv.org/abs/2010.12050.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Olle G. Holmberg, Niklas D. Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, Johannes Schiefelbein, Siegfried Priglinger, Karsten U. Kortuem, and Fabian J. Theis. Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. *Nature Machine Intelligence*, 2(11):719–726, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00247-1. URL https://doi.org/10.1038/s42256-020-00247-1.

Le Thi Thu Hong, Nguyen Chi Thanh, and Tran Quoc Long. Self-supervised visual feature learning for polyp segmentation in colonoscopy images using image reconstruction as pretext task. In *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 254–259, 2021. doi: 10.1109/NICS54270.2021.9701580.

J. L. Horn and R. B. Cattell. Refinement and test of the theory of fluid and crystallized general intelligences. 57(5):253–270, 1966. URL https://doi.org/10.1037/h0023816.

J.L. Horn. *Fluid and Crystallized Intelligence: A Factor Analytic Study of the Structure Among Primary Mental Abilities*. University Microfilms, 1971. URL https://books.google.co.in/books?id=zg0eQgAACAAJ.

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2013. doi: 10.1109/IJCNN.2013.6706807.

Jiacong Hu, Zunlei Feng, Yining Mao, Jie Lei, Dan Yu, and Mingli Song. A location constrained dual-branch network for reliable diagnosis of jaw tumors and cysts. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 723–732, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1074–1083, 2020a. URL https://api.semanticscholar.org/CorpusID:226976086.

Shizhe Hu, Chengkun Zhang, Guoliang Zou, Zhengzheng Lou, and Yangdong Ye. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024. doi: 10.1109/TNNLS.2024.3354731.

Szu-Yen Hu, Shuhang Wang, Wei-Hung Weng, JingChao Wang, XiaoHong Wang, Arinc Ozturk, Quan Li, Viksit Kumar, and Anthony E. Samir. Self-supervised pretraining with dicom metadata in ultrasound imaging. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pp. 732–749. PMLR, 07–08 Aug 2020b. URL https://proceedings.mlr.press/v126/hu20a.html.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1558–1567. JMLR.org, 2017.

Xinrong Hu, Corey Wang, and Yiyu Shi. Contrastive image synthesis and self-supervised feature adaptation for cross-modality biomedical image segmentation. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2329–2338, 2022. URL https://api.semanticscholar.org/CorpusID:251104842.

T. Hua, W. Wang, Z. Xue, S. Ren, Y. Wang, and H. Zhao. On feature decorrelation in self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9578–9588, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00946. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00946.

D. Huang, W. Wu, W. Hu, X. Liu, D. He, Z. Wu, X. Wu, M. Tan, and E. Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8076–8085, Los Alamitos, CA, USA, oct 2021a. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00799. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00799.

Junjia Huang, Haofeng Li, Guanbin Li, and Xiang Wan. Attentive symmetric autoencoder for brain mri segmentation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 203–213, Cham, 2022a. Springer Nature Switzerland. ISBN 978-3-031-16443-9.

Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13881–13890, 2021b. doi: 10.1109/CVPR46437.2021.01367.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):74, Apr 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00811-0. URL https://doi.org/10.1038/s41746-023-00811-0.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners, 2022b.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7509–7524, 2021c. URL https://api.semanticscholar.org/CorpusID:252992558.

Yuqi Huo, Mingyu Ding, Haoyu Lu, Nanyi Fei, Zhiwu Lu, Ji-Rong Wen, and Ping Luo. Compressed video contrastive learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14176–14187, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7647966b7343c29048673252e490f736-Abstract.html.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 986–996, 2022. doi: 10.1109/WACV51458.2022.00106.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3772–3780, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

iChallenge. Baidu Research Open-Access Dataset - Introduction — ai.baidu.com, 2018. URL http://ai.baidu.com/broad/introduction.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.*, 35(4), jul 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925974. URL https://doi.org/10.1145/2897824.2925974.

Francesca Inglese, Minseon Kim, Gerda M Steup-Beekman, Tom W J Huizinga, Mark A van Buchem, Jeroen de Bresser, Dae-Shik Kim, and Itamar Ronen. MRI-based classification of neuropsychiatric systemic lupus erythematosus patients with self-supervised contrastive learning. *Front. Neurosci.*, 16:695888, February 2022.

Yotam Intrator, Natalie Aizenberg, Amir Livne, Ehud Rivlin, and Roman Goldenberg. Self-supervised polyp re-identification in colonoscopy. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pp. 590–600. Springer Nature Switzerland, Cham, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard J. Radke, and Rogério Schmidt Feris. A broad study on the transferability of visual representations with contrastive learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8825–8835, 2021. URL https://api.semanticscholar.org/CorpusID:232352365.

Allan A. Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.*, 4(6):475–477, December 2014.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised learning for spinal mris. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 294–302, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9.

Ananya Jana, Hui Qu, Carlos D. Minacapelli, Carolyn Catalano, Vinod Rustgi, and Dimitris Metaxas. Liver fibrosis and nas scoring from ct images using self-supervised learning and texture encoding. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1553–1557, 2021a. doi: 10.1109/ISBI48211.2021.9433920.

Ananya Jana, Hui Qu, Carlos D Minacapelli, Carolyn Catalano, Vinod K. Rustgi, and Dimitris N. Metaxas. Liver fibrosis and nas scoring from ct images using self-supervised learning and texture encoding. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1553–1557, 2021b. URL https://api.semanticscholar.org/CorpusID:232135290.

Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to egomotion from unlabeled video. *International Journal of Computer Vision*, 125:136–161, 2017.

Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, pp. 425–442. Springer, 2020. doi: 10.1007/978-3-030-58604-1_26. URL https://doi.org/10.1007/978-3-030-58604-1_26.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pp. 451–462. Springer, 2020.

X. Ji, A. Vedaldi, and J. Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9864–9873, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00996. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00996.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ee604e1bedbd069d9fc9328b7b9584be-Abstract-Datasets_and_Benchmarks.html.

Guo-Zhang Jian, Guo-Shiang Lin, Chuin-Mu Wang, and Sheng-Lei Yan. Helicobacter pylori infection classification based on convolutional neural network and self-supervised learning. In *Proceedings of the 5th International Conference on Graphics and Signal Processing*, ICGSP '21, pp. 60–64, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389419. doi: 10.1145/3474906.3474912. URL https://doi.org/10.1145/3474906.3474912.

Hongchao Jiang and Chunyan Miao. Pre-training 3d convolutional neural networks for prodromal alzheimer's disease classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022. doi: 10.1109/IJCNN55064.2022.9891966.

Yankai Jiang, Ming Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15813–15823, 2023. URL https://api.semanticscholar.org/CorpusID:261030572.

Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):352–364, 2018. doi: 10.1109/TPAMI.2017.2670560.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jianbo Jiao, Yifan Cai, Mohammad Alsharid, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-supervised contrastive video-speech representation learning for ultrasound. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 12263:534–543, 2020a. URL https://api.semanticscholar.org/CorpusID:221139482.

Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Self-supervised representation learning for ultrasound video. *Proc. IEEE Int. Symp. Biomed. Imaging*, 2020:1847–1850, April 2020b.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4037–4058, 2021. doi: 10.1109/TPAMI.2020.2992393. URL https://doi.org/10.1109/TPAMI.2020.2992393.

Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction, 2019.

Amod Jog, Aaron Carass, and Jerry L Prince. Self super-resolution for magnetic resonance images. 9902: 553–560, October 2016.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21798–21809. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and; Machine Intelligence*, 43(12):4217–4228, dec 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2970919.

Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection, 2019. URL https://kaggle.com/competitions/aptos2019-blindness-detection.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, May 2018. URL https://doi.org/10.5281/zenodo.1214456.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

David N Kennedy, Christian Haselgrove, Steven M Hodge, Pallavi S Rane, Nikos Makris, and Jean A Frazier. CANDIShare: a resource for pediatric neuroimaging data. *Neuroinformatics*, 10(3):319–322, July 2012.

Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 793–802, 2018. doi: 10.1109/WACV.2018.00092.

Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33018545. URL https://doi.org/10.1609/aaai.v33i01.33018545.

Jaeill Kim, Duhun Hwang, Eunjung Lee, Jangwon Suh, Jimyeong Kim, and Wonjong Rhee. Enhancing contrastive learning with efficient combinatorial positive pairing, 2024.

Jinyoung Kim, Soon Mo Kwon, Hyojun Go, Yunsung Lee, and Seungtaek Choi. Scorecl: Augmentation-adaptive contrastive learning via score-matching function. *ArXiv*, abs/2306.04175, 2023. URL https://api.semanticscholar.org/CorpusID:259095523.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020b.

Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, Hyun Jung, Yanling Liu, Haran Rajkumar, Mahendra Khened, Ganapathy Krishnamurthi, Sen Yang, Xiyue Wang, Chang Hee Han, Jin Tae Kwak, Jianqiang Ma, Zhe Tang, Bahram Marami, Jack Zeineh, Zixu Zhao, Pheng-Ann Heng, Rüdiger Schmitz, Frederic Madesta, Thomas Rösch, Rene Werner, Jie Tian, Elodie Puybareau, Matteo Bovio, Xiufeng Zhang, Yifeng Zhu, Se Young Chun, Won-Ki Jeong, Peom Park, and Jinwook Choi. PAIP 2019: Liver cancer segmentation challenge. *Med. Image Anal.*, 67(101854):101854, January 2021.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019. URL https://api.semanticscholar.org/CorpusID:59292019.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021. doi: 10.1109/TMI.2021.3056023.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10306–10315, 2021. URL https://api.semanticscholar.org/CorpusID:233992782.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 7774–7785, Red Hook, NY, USA, 2018. Curran Associates Inc.

Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3188–3188, 2021. doi: 10.1109/ICCVW54120.2021.00358.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014. doi: 10.1109/CVPR.2014.105.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool (eds.), *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pp. 2556–2563. IEEE Computer Society, 2011. doi: 10.1109/ICCV.2011.6126543. URL https://doi.org/10.1109/ICCV.2011.6126543.

Hugo J Kuijf, Edwin Bennink, Koen L Vincken, Nick Weaver, Geert Jan Biessels, and Max A Viergever. MR brain segmentation challenge 2018 data, 2024.

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017. doi: 10.1109/TMI.2017.2677499.

Vidit Kumar, Vikas Tripathi, and Bhaskar Pant. Unsupervised learning of visual representations via rotation and future frame prediction for video retrieval. In Mayank Singh, Vipin Tyagi, P. K. Gupta, Jan Flusser, Tuncer Ören, and V. R. Sonawane (eds.), *Advances in Computing and Data Sciences*, pp. 701–710, Cham, 2021. Springer International Publishing. ISBN 978-3-030-81462-5.

Pamela J. LaMontagne, Tammie LS. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902. URL https://www.medrxiv.org/content/early/2019/12/15/2019.12.13.19014902.

G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 840–849, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.96. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.96.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 577–593, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. URL https://api.semanticscholar.org/CorpusID:16664790.

Dong Hoon Lee, Sungik Choi, Hyunwoo J. Kim, and Sae-Young Chung. Unsupervised visual representation learning via mutual information regularized assignment. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/bedc61a9936af18cb51b7c5e8f3b89a3-Abstract-Conference.html.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 667–676. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.79. URL https://doi.org/10.1109/ICCV.2017.79.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=T6AxtOaWydQ.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16(1):51–63, January 2018.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2020a. URL https://api.semanticscholar.org/CorpusID:227013055.

Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=fVu3o-YUGQK.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, 2017. URL https://api.semanticscholar.org/CorpusID:6037691.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 14290–14302. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5c186016d0844767209dc36e9e61441b-Paper-Conference.pdf.

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Self-supervised learning for gastritis detection with gastric x-ray images. *Int. J. Comput. Assist. Radiol. Surg.*, 18(10):1841–1848, 2023a. doi: 10.1007/S11548-023-02891-5. URL https://doi.org/10.1007/s11548-023-02891-5.

Hao Li, Xiaopeng Zhang, Ruoyu Sun, Hongkai Xiong, and Qi Tian. Center-wise local image mixture for contrastive representation learning. In *British Machine Vision Conference*, 2020b. URL https://api.semanticscholar.org/CorpusID:226254277.

Hongwei Li, Fei-Fei Xue, Krishna Chaitanya, Shengda Luo, Ivan Ezhov, Benedikt Wiestler, Jianguo Zhang, and Bjoern Menze. Imbalance-aware self-supervised learning for 3d radiomic representations. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 36–46, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-87196-3.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=KmykpuSrjcq.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12581–12600, 2023b. doi: 10.1109/TPAMI.2023.3282631. URL https://doi.org/10.1109/TPAMI.2023.3282631.

Mingchao Li, Kun Huang, Qiuzhuo Xu, Jiadong Yang, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. OCTA-500: A retinal dataset for optical coherence tomography angiography study. *Med. Image Anal.*, 93(103092):103092, April 2024.

R. Li and D. Liu. Spatial-then-temporal self-supervised learning for video correspondence. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2279–2288, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00226. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00226.

Ru Li, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Jigsawgan: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *Trans. Img. Proc.*, 31:513–524, jan 2022c. ISSN 1057-7149. doi: 10.1109/TIP.2021.3120052. URL https://doi.org/10.1109/TIP.2021.3120052.

T. Li, L. Fan, Y. Yuan, H. He, Y. Tian, R. Feris, P. Indyk, and D. Katabi. Addressing feature suppression in unsupervised visual representations. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1411–1420, Los Alamitos, CA, USA, jan 2023c. IEEE Computer Society. doi: 10.1109/WACV56688.2023.00146. URL https://doi.ieeecomputersociety.org/10.1109/WACV56688.2023.00146.

Wenbin Li, Xuesong Yang, Meihao Kong, Lei Wang, Jing Huo, Yang Gao, and Jiebo Luo. Trip-ROMA: Self-supervised learning with triplets and random mappings. *Transactions on Machine Learning Research*, 2023d. ISSN 2835-8856.

Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling MAE pre-training for pyramid-based vision transformers with locality. *CoRR*, abs/2205.10063, 2022d. doi: 10.48550/ARXIV.2205.10063. URL https://doi.org/10.48550/arXiv.2205.10063.

Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39 (12):4023–4033, 2020c. doi: 10.1109/TMI.2020.3008871.

Xiaomeng Li, Xiaowei Hu, Xiaojuan Qi, Lequan Yu, Wei Zhao, Pheng-Ann Heng, and Lei Xing. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 40:2284–2294, 2021c. URL https://api.semanticscholar.org/CorpusID:233382359.

Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image BERT pre-training. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pp. 231–246. Springer, 2022e. doi: 10.1007/978-3-031-20056-4_14. URL https://doi.org/10.1007/978-3-031-20056-4_14.

Yazhe Li, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. In *Neural Information Processing Systems*, 2021d. URL https://api.semanticscholar.org/CorpusID:235436250.

Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 520–535, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01231-1.

Yuexiang Li, Jiawei Chen, and Yefeng Zheng. A multi-task self-supervised learning framework for scopy images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 2005–2009, 2020d. doi: 10.1109/ISBI45749.2020.9098527.

Yuheng Li, Jacob F. Wynne, Jing Wang, Richard L. J. Qiu, Justin Roper, Shaoyan Pan, Ashesh B. Jani, Tian Liu, Pretesh R. Patel, Hui Mao, and Xiaofeng Yang. Cross-shaped windows transformer with self-supervised pretraining for clinically significant prostate cancer detection in bi-parametric MRI. *CoRR*, abs/2305.00385, 2023e. doi: 10.48550/ARXIV.2305.00385. URL https://doi.org/10.48550/arXiv.2305.00385.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. MST: masked self-supervised transformer for visual representation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 13165–13176, 2021e. URL https://proceedings.neurips.cc/paper/2021/hash/6dbbe6abe5f14af882ff977fc3f35501-Abstract.html.

Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 1564–1573. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20047.

Ruizhi Liao, Geeticka Chauhan, Polina Golland, Seth Berkowitz, and Steven Horng. Pulmonary edema severity grades based on MIMIC-CXR, 2021a.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M. Wells. Multimodal representation learning via maximization of local mutual information. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 273–283, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87196-3.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://dblp.uni-trier.de/db/journals/corr/corr1405.html#LinMBHPRDZ14.

Fengbei Liu, Yu Tian, Filipe R. Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Self-supervised mean teacher for semi-supervised chest x-ray classification. In Chunfeng Lian, Xiaohuan Cao, Islem Rekik, Xuanang Xu, and Pingkun Yan (eds.), *Machine Learning in Medical Imaging*, pp. 426–436, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87589-3.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023a. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25253. URL https://doi.org/10.1609/aaai.v37i2.25253.

J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6252–6261, Los Alamitos, CA, USA, jun 2023b. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00605. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00605.

Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *CoRR*, abs/2205.13137, 2022a. doi: 10.48550/ARXIV.2205.13137. URL https://doi.org/10.48550/arXiv.2205.13137.

Yifei Liu, Qingbin Wang, Ling Zhang, and Kai Zhang. Multi-layer feature refinement extraction with contrastive learning for cervical oct image classification. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2111–2116, 2023c. URL https://api.semanticscholar.org/CorpusID:267044515.

Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip S. Yu. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5879–5900, 2023d. doi: 10.1109/TKDE.2022.3172903.

Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3192–3201, Los Alamitos, CA, USA, jun 2022b. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00320. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00320.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022c. doi: 10.1109/CVPR52688.2022.01167.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nat. Methods*, 9(7):637, June 2012.

Xianzhong Long, Zhiyi Zhang, and Yun Li. Multi-network contrastive learning of visual representations. *Knowledge-Based Systems*, 258:109991, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109991. URL https://www.sciencedirect.com/science/article/pii/S095070512201084X.

Cheng Lu, Xiaojie Jin, Zhicheng Huang, Qibin Hou, Mingg-Ming Cheng, and Jiashi Feng. Cmae-v: Contrastive masked autoencoders for video action recognition. *ArXiv*, abs/2301.06018, 2023. URL https://api.semanticscholar.org/CorpusID:255941594.

Qi Lu, Yuxing Li, and Chuyang Ye. White matter tract segmentation with self-supervised learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 270–279, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59728-3.

Qi Lu, Yuxing Li, and Chuyang Ye. Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. *Medical Image Analysis*, 72:102094, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102094. URL https://www.sciencedirect.com/science/article/pii/S1361841521001407.

Aurélien Lucchi, Yunpeng Li, and Pascal Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1987–1994, 2013. doi: 10.1109/CVPR.2013.259.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding, 2019.

Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11701–11708. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6840.

Dongliang Luo and Jian Wang. Prior matching operator in self-supervised learning. In *2022 7th International Conference on Signal and Image Processing (ICSIP)*, pp. 777–781, 2022. doi: 10.1109/ICSIP55141.2022.9886345.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, Shuiping Gou, Franz Thaler, Christian Payer, Darko Štern, Edward G.A. Henderson, Dónal M. McSweeney, Andrew Green, Price Jackson, Lachlan McIntosh, Quoc-Cuong Nguyen, Abdul Qayyum, Pierre-Henri Conze, Ziyan Huang, Ziqi Zhou, Deng-Ping Fan, Huan Xiong, Guoqiang Dong, Qiongjie Zhu, Jian He, and Xiaoping Yang. Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. *Medical Image Analysis*, 82:102616, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102616. URL https://www.sciencedirect.com/science/article/pii/S1361841522002444.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=OMizHuea_HB.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local video representations. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 7025–7040, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/38ef4b66cb25e92abe4d594acb841471-Abstract.html.

Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. LDPolypVideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture notes in computer science, pp. 387–396. Springer International Publishing, Cham, 2021c.

Dwarikanath Mahapatra, Alexander Poellinger, Ling Shao, and Mauricio Reyes. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2548–2562, 2021. doi: 10.1109/TMI.2021.3061724.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. Interpretive self-supervised pre-training: boosting performance on visual medical data. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450375962. doi: 10.1145/3490035.3490273. URL https://doi.org/10.1145/3490035.3490273.

Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. Self-supervised representation learning for detection of acl tear injury in knee mr videos. *Pattern Recognition Letters*, 154:37–43, 2022. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2022.01.008. URL https://www.sciencedirect.com/science/article/pii/S0167865522000149.

Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. Self-supervised representation learning for knee injury diagnosis from magnetic resonance data. *IEEE Transactions on Artificial Intelligence*, pp. 1–11, 2023. doi: 10.1109/TAI.2023.3299883.

Huanru Henry Mao. A survey on self-supervised pre-training for sequential transfer learning in neural networks. *CoRR*, abs/2007.00800, 2020. URL https://arxiv.org/abs/2007.00800.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.*, 19(9):1498–1507, September 2007.

Franco Matzkin, Virginia F. J. Newcombe, Susan Stevenson, Aneesh Khetani, Tom Newman, Richard Digby, Andrew Stevens, Ben Glocker, and Enzo Ferrante. Self-supervised skull reconstruction in brain CT images with decompressive craniectomy. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part II*, volume 12262 of *Lecture Notes in Computer Science*, pp. 390–399. Springer, 2020. doi: 10.1007/978-3-030-59713-9_38. URL https://doi.org/10.1007/978-3-030-59713-9_38.

Moona Mazher, Imran Razzak, Abdul Qayyum, M. Tanveer, Susann Beier, Tariq Khan, and Steven A Niederer. Self-supervised spatial–temporal transformer fusion based federated framework for 4d cardiovascular image segmentation. *Information Fusion*, 106:102256, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102256. URL https://www.sciencedirect.com/science/article/pii/S1566253524000344.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Comput. Biol. Med.*, 141(C), feb 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.105114. URL https://doi.org/10.1016/j.compbiomed.2021.105114.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

A. Miech, D. Zhukov, J. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2630–2640, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00272. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00272.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9876–9886, 2020. doi: 10.1109/CVPR42600.2020.00990.

Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations, 2022.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2019. URL https://api.semanticscholar.org/CorpusID:208617491.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *Computer Vision – ECCV 2016*, pp. 527–544, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

Jovana Mitrovic, Brian McWilliams, Jacob C. Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=9p2ekP904Rs.

Mohammad MohammadAmini, Driss Matrouf, Jean-François Bonastre, Sandipana Dowerah, Romain Serizel, and Denis Jouvet. Barlow twins self-supervised learning for robust speaker recognition. In Hanseok Ko and John H. L. Hansen (eds.), *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pp. 4033–4037. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-11301. URL https://doi.org/10.21437/Interspeech.2022-11301.

Nooshin Mojab, Vahid Noroozi, Darvin Yi, Manoj Prabhakar Nallabothula, Abdullah Aleem, Philip S. Yu, and Joelle A. Hallak. Real-world multi-domain data applications for generalizations to clinical settings. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 677–684, 2020. URL https://api.semanticscholar.org/CorpusID:220768875.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

Amin Karimi Monsefi, Payam Karisani, Mengxi Zhou, Stacey Choi, Nathan Doble, Heng Ji, Srinivasan Parthasarathy, and Rajiv Ramnath. Masked logonet: Fast and accurate 3d image analysis for medical domain, 2024.

G B Moody and R G Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.*, 20(3):45–50, May 2001.

Jong Hak Moon, Wonjae Kim, and E. Choi. Correlation between alignment-uniformity and performance of dense contrastive representations. In *British Machine Vision Conference*, 2022. URL https://api.semanticscholar.org/CorpusID:252918494.

P. Morgado, N. Vasconcelos, and I. Misra. Audio-visual instance discrimination with cross-modal agreement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12470–12481, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01229. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01229.

S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset, 2020.

Terrell Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9339–9348, 2017.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.*, 60(101027):101027, March 2020.

National Lung Screening Trial Research Team. Data from the national lung screening trial (NLST), 2013.

K. L. Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, Kossar Pourahmadi, Akshayvarun Subramanya, and Hamed Pirsiavash. Constrained mean shift using distant yet related neighbors for representation learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pp. 23–41. Springer, 2022. doi: 10.1007/978-3-031-19821-2_2. URL https://doi.org/10.1007/978-3-031-19821-2_2.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38(2):448–459, 2019. doi: 10.1109/TMI.2018.2865709.

Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 204–214, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/54229abfcfa5649e7003b83dd4755294-Abstract.html.

Duy M. H. Nguyen, Hoang Nguyen, Truong T. N. Mai, Tri Cao, Binh T. Nguyen, Nhat Ho, Paul Swoboda, Shadi Albarqouni, Pengtao Xie, and Daniel Sonntag. Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i12.26687. URL https://doi.org/10.1609/aaai.v37i12.26687.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, Kenneth A Iczkowski, M Scott Lucia, Peter C Black, Purang Abolmaesumi, S Larry Goldenberg, and Septimiu E Salcudean. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Med. Image Anal.*, 50:167–180, December 2018.

M. Noroozi, H. Pirsiavash, and P. Favaro. Representation learning by learning to count. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5899–5907, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.628. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.628`.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 69–84, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46466-4.

Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. URL `https://api.semanticscholar.org/CorpusID:258170077`.

Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: task dependent adaptive metric for improved few-shot learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 719–729, Red Hook, NY, USA, 2018. Curran Associates Inc.

A. Emin Orhan, Vaibhav V. Gupta, and Brenden M. Lake. Self-supervised learning through the eyes of a child. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, Joonho Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R Phaye, Sharath M Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, and Hrvoje Bogunović. REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.*, 59(101570):101570, January 2020.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision – ECCV 2020*, pp. 762–780, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58526-6.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41:1837–1848, 2022. URL `https://api.semanticscholar.org/CorpusID:246700149`.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V. Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M. Pohl. Self-supervised longitudinal neighbourhood embedding. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 80–89, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. *Data (Basel)*, 6(2):14, February 2021.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11200–11209, 2021. URL https://api.semanticscholar.org/CorpusID:232170175.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Bo Pang, Yizhuo Li, Yifan Zhang, Gao Peng, Jiajun Tang, Kaiwen Zha, Jiefeng Li, and Cewu Lu. Unsupervised representation for semantic segmentation by implicit cycle-attention contrastive learning. In *AAAI Conference on Artificial Intelligence*, 2022a. URL https://api.semanticscholar.org/CorpusID:250296749.

Bo Pang, Yifan Zhang, Yaoyi Li, Jia Cai, and Cewu Lu. Unsupervised visual representation learning by synchronous momentum grouping. In *European Conference on Computer Vision*, 2022b. URL https://api.semanticscholar.org/CorpusID:250490993.

Sang Ok Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, Changhyun Park, and Jong-Chul Ye. Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. *Nature Communications*, 13, 2022. URL https://api.semanticscholar.org/CorpusID:246822434.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.278. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.278.

M. Patrick, P. Huang, I. Misra, F. Metze, A. Vedaldi, Y. M. Asano, and J. Henriques. Space-time crop and attend: Improving cross-modal video representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10540–10552, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.01039. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01039.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In Eduardo Romero and Natasha Lepore (eds.), *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, pp. 92870W. International Society for Optics and Photonics, SPIE, 2015. doi: 10.1117/12.2073532. URL https://doi.org/10.1117/12.2073532.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D Clifford, and Matthew A Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in cardiology challenge 2020. *Physiol. Meas.*, 41(12):124003, January 2021.

R C Petersen, P S Aisen, L A Beckett, M C Donohue, A C Gamst, D J Harvey, C R Jack, Jr, W J Jagust, L M Shaw, A W Toga, J Q Trojanowski, and M W Weiner. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*, 74(3):201–209, January 2010.

Nicholas Petrick, Shazia Akbar, Kenny H Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne L Martel, and BreastPathQ Challenge Group. SPIE-AAPM-NCI BreastPathQ challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *J. Med. Imaging (Bellingham)*, 8(3):034501, May 2021.

AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 130–139, 2020. doi: 10.1109/CVPR42600.2020.00021.

Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

António Polónia, Catarina Eloy, and Paulo Aguiar. BACH dataset : Grand challenge on breast cancer histology images, 2019.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data (Basel)*, 3(3):25, July 2018.

Mangal Prakash, Tim-Oliver Buchholz, Manan Lalit, Pavel Tomancak, Florian Jug, and Alexander Krull. Leveraging self-supervised denoising for image segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 428–432, 2020. doi: 10.1109/ISBI45749.2020.9098559.

Narinder Singh Punn and Sonali Agarwal. Bt-unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Mach. Learn.*, 111(12):4585–4600, dec 2022. ISSN 0885-6125. doi: 10.1007/s10994-022-06219-3. URL https://doi.org/10.1007/s10994-022-06219-3.

Guo-Jun Qi and Mubarak Shah. Adversarial pretraining of self-supervised deep networks: Past, present and future. *CoRR*, abs/2210.13463, 2022. doi: 10.48550/ARXIV.2210.13463. URL https://doi.org/10.48550/arXiv.2210.13463.

Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16619–16628, 2021a. URL https://api.semanticscholar.org/CorpusID:235187404.

R. Qian, T. Meng, B. Gong, M. Yang, H. Wang, S. Belongie, and Y. Cui. Spatiotemporal contrastive video representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6960–6970, Los Alamitos, CA, USA, jun 2021b. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00689. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00689.

Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Xiang Wang, Yuehuan Wang, Yiliang Lv, Changxin Gao, and Nong Sang. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, pp. 1–16, 2023. doi: 10.1109/TMM.2023.3263288.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06434.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. *Evaluating protein transfer learning with TAPE*. Curran Associates Inc., Red Hook, NY, USA, 2019.

A. Recasens, P. Luc, J. Alayrac, L. Wang, F. Strub, C. Tallec, M. Malinowski, V. Patraaucean, F. Altche, M. Valko, J. Grill, A. van den Oord, and A. Zisserman. Broaden your views for self-supervised video learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1235–1245, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00129. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00129.

Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7(1):32–43, February 2010.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. *Likelihood ratios for out-of-distribution detection*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Mengye Ren, Tyler R. Scott, Michael L. Iuzzolino, Michael C. Mozer, and Richard S. Zemel. Online unsupervised learning of visual representations and categories. *CoRR*, abs/2109.05675, 2021. URL https://arxiv.org/abs/2109.05675.

Pierre H. Richemond, Jean-Bastien Grill, Florent Altch'e, Corentin Tallec, Florian Strub, Andrew Brock, Samuel L. Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. Byol works even without batch statistics. *ArXiv*, abs/2010.10241, 2020. URL https://api.semanticscholar.org/CorpusID:224802757.

Antoine Rivail, Ursula Margarethe Schmidt-Erfurth, Wolf-Dieter Vogl, Sebastian M. Waldstein, Sophie Riedl, Christoph Grechenig, Zhichao Wu, and Hrvoje Bogunović. Modeling disease progression in retinal octs with longitudinal self-supervised learning. In *PRIME@MICCAI*, 2019. URL https://api.semanticscholar.org/CorpusID:204754406.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 4974–4986, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/27934a1f19d678a1377c257b9a780e80-Abstract.html.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Jason Tyler Rolfe. Discrete variational autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=ryMxXPFex.

Tobias Ross, David Zimmerer, Anant Suraj Vemuri, Fabian Isensee, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Peter Müller-Stich, Hannes Kenngott, Stefanie Speidel, Klaus Maier-Hein, and Lena Maier-Hein. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13:925–933, 2017. URL https://api.semanticscholar.org/CorpusID:13818971.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data*, 8(1):34, January 2021.

Holger Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. A new 2.5 D representation for lymph node detection in CT (CT lymph nodes), 2015.

Holger Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from Pancreas-CT, 2016.

Yangjun Ruan, Saurabh Singh, Warren Richard Morningstar, Alexander A. Alemi, Sergey Ioffe, Ian Fischer, and Joshua V. Dillon. Weighted ensemble self-supervised learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=CL-sVR9pvF.

Chaitanya Ryali, David J. Schwab, and Ari S. Morcos. Learning background invariance improves generalization and robustness in self-supervised learning on imagenet and beyond. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL https://openreview.net/forum?id=zZnOG9ehfoO.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkGuG2R5tm.

A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi. Visual semantic role labeling for video understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5585–5596, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00554. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00554.

Anindo Saha, Jasper Jonathan Twilt, Joeran Sander Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. The PI-CAI challenge: Public training and development dataset, 2022.

Mihir Sahasrabudhe, Stergios Christodoulidis, Roberto Salgado, Stefan Michiels, Sherene Loi, Fabrice André, Nikos Paragios, and Maria Vakalopoulou. Self-supervised nuclei segmentation in histopathological images using attention. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 393–402, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59722-1.

Charlie Saillard, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoît Schmauch, and Simon Jégou. Self supervised learning improves dmmr/msi detection from histology slides across multiple cancers. In *COMPAY@MICCAI*, 2021. URL https://api.semanticscholar.org/CorpusID:237165674.

Alice M. L. Santilli, Amoon Jamzad, Alireza Sedghi, Martin Kaufmann, Kathryn Logan, Julie Wallis, Kevin Yi Mi Ren, Natasja N. Y. Janssen, Shaila J. Merchant, Jay Engel, Doug McKay, Sonal Varma, Ami Wang, Gabor Fichtinger, John F. Rudan, and Parvin Mousavi. Domain adaptation and self-supervised learning for surgical margin detection. *International Journal of Computer Assisted Radiology and Surgery*, 16:861 – 869, 2021. URL https://api.semanticscholar.org/CorpusID:233746729.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. URL https://api.semanticscholar.org/CorpusID:206592766.

R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik. Casting your model: Learning to localize improves self-supervised representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11053–11062, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi:

10.1109/CVPR46437.2021.01091. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01091.

Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 486–487, 2017. doi: 10.1109/CVPRW.2017.69.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S N Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M J de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T M C Manders, Alexander Sóñora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M Schaefer-Prokop, Ernst T Scholten, Luuk Scholten, Miranda M Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C A Zuidhof, Bram van Ginneken, and Colin Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.*, 42:1–13, December 2017.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622, 2020. doi: 10.1109/CVPR42600.2020.00269.

Chengchao Shen, Jianzhong Chen, Shu Wang, Hulin Kuang, Jin Liu, and Jianxin Wang. Asymmetric patch sampling for contrastive learning. *CoRR*, abs/2306.02854, 2023. doi: 10.48550/ARXIV.2306.02854. URL https://doi.org/10.48550/arXiv.2306.02854.

Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip H. S. Torr, and Ling Shao. You never cluster alone. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 27734–27746, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e96ed478dab8595a7dbda4cbcbee168f-Abstract.html.

Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric P. Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:245329909.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: a machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 802–810, Cambridge, MA, USA, 2015. MIT Press.

Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information-theoretic perspective on variance-invariance-covariance regularization, 2023.

Fatemeh Siar, Amin Gheibi, and Ali Mohades. Unsupervised learning of visual representations by solving shuffled long video-frames temporal order prediction. In *ACM SIGGRAPH 2020 Posters*, SIGGRAPH '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379731. doi: 10.1145/3388770.3407409. URL https://doi.org/10.1145/3388770.3407409.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, pp. 746–760, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642337147. doi: 10.1007/978-3-642-33715-4_54. URL https://doi.org/10.1007/978-3-642-33715-4_54.

Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A. Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 53–56, 2014. doi: 10.1109/ISBI.2014.6867807.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Luc Soler, Alexandre Hostettler, Vincent Agnus, Arnaud Charnoz, J Fasquel, Johan Moreau, A Osswald, Mourad Bouhadjar, and Jacques Marescaux. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 1(1), 2010.

Danming Song, Yipeng Gao, Junkai Yan, Wei Sun, and Wei-Shi Zheng. Space-correlated contrastive representation learning with multiple instances. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 4715–4721, 2022. doi: 10.1109/ICPR56361.2022.9956034.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In Mattias P. Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schlaefer, and Floris Ernst (eds.), *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*, volume 143 of *Proceedings of Machine Learning Research*, pp. 728–744. PMLR, 2021. URL https://proceedings.mlr.press/v143/sowrirajan21a.html.

Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 663–671, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00931-1.

Chetan L. Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L. Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical Image Analysis*, 75:102256, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102256. URL https://www.sciencedirect.com/science/article/pii/S1361841521003017.

Vignesh Srinivasan, Nils Strodthoff, Jackie Ma, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek. To pretrain or not? a systematic analysis of the benefits of pretraining in diabetic retinopathy. *PLOS ONE*, 17(10):1–18, 10 2022. doi: 10.1371/journal.pone.0274291. URL https://doi.org/10.1371/journal.pone.0274291.

J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. doi: 10.1109/TMI.2004.825627.

Karin Stacke, Claes Lundström, Jonas Unger, and Gabriel Eilertsen. Evaluation of contrastive predictive coding for histopathology applications. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland (eds.), *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pp. 328–340. PMLR, 11 Dec 2020. URL https://proceedings.mlr.press/v136/stacke20a.html.

Ivan Štajduhar, Mihaela Mamula, Damir Miletić, and Gözde Ünal. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput. Methods Programs Biomed.*, 140:151–164, March 2017.

Thomas Stegmüller, Tim Lebailly, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. Croc: Cross-view online clustering for dense visual representation learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7000–7009, 2023. URL https://api.semanticscholar.org/CorpusID:257687708.

Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pp. 729–738, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450317702. doi: 10.1145/2493432.2493482. URL https://doi.org/10.1145/2493432.2493482.

Dejan Štepec and Danijel Skočaj. Image synthesis as a pretext for unsupervised histopathological diagnosis. In Ninon Burgos, David Svoboda, Jelmer M. Wolterink, and Can Zhao (eds.), *Simulation and Synthesis in Medical Imaging*, pp. 174–183, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59520-3.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 2017. doi: 10.1109/ICCV.2017.97.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *CoRR*, abs/1906.05743, 2019. URL http://arxiv.org/abs/1906.05743.

Li Sun, Ke Yu, and Kayhan Batmanghelich. Context matters: Graph-based self-supervised representation learning for medical images. *Proc. Conf. AAAI Artif. Intell.*, 35(6):4874–4882, February 2021a.

W. Sun, J. Zhang, J. Wang, Z. Liu, Y. Zhong, T. Feng, Y. Guo, Y. Zhang, and N. Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6420–6429, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00621. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00621.

Yue Sun, Kun Gao, W. Lin, Gang Li, Sijie Niu, and Li Wang. Multi-scale self-supervised learning for multi-site pediatric brain mr image segmentation with motion/gibbs artifacts. *Machine learning in medical imaging. MLMI*, 12966:171–179, 2021b. URL https://api.semanticscholar.org/CorpusID:238222800.

Ryu Tadokoro, Ryosuke Yamada, and Hirokatsu Kataoka. Pre-training auto-generated volumetric shapes for 3d medical image segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4740–4745, 2023a. URL https://api.semanticscholar.org/CorpusID:260908068.

Ryu Tadokoro, Ryosuke Yamada, Kodai Nakashima, Ryo Nakamura, and Hirokatsu Kataoka. Primitive geometry segment pre-training for 3d medical image segmentation. In *34th British Machine Vision Conference 2022, BMVC 2022, Aberdeen, UK, November 20-24, 2023*, pp. 152–160. BMVA Press, 2023b. URL http://proceedings.bmvc2023.org/152/.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. CAiD: Context-Aware instance discrimination for self-supervised learning in medical imaging. *Proc. Mach. Learn. Res.*, 172:535–551, July 2022.

Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1251–1255, 2019. URL https://api.semanticscholar.org/CorpusID:59291957.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen (eds.), *Information Processing in Medical Imaging*, pp. 661–673, Cham, 2021. Springer International Publishing. ISBN 978-3-030-78191-0.

Alex Tamkin, Vincent Liu, Rongfei Lu, Daniel Fein, Colin Schultz, and Noah D. Goodman. DABS: a domain-agnostic benchmark for self-supervised learning. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/8d5e957f297893487bd98fa830fa6413-Abstract-round1.html.

Alex Tamkin, Gaurab Banerjee, Mohamed Owda, Vincent Liu, Shashank Rammoorthy, and Noah D. Goodman. DABS 2.0: Improved datasets and algorithms for universal self-supervision. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fa73aca7b2af724fafbd4852957cd3e0-Abstract-Datasets_and_Benchmarks.html.

Qian Tang, Bo Du, and Yongchao Xu. Self-supervised learning based on max-tree representation for medical image segmentation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2022a. doi: 10.1109/IJCNN55064.2022.9892853.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett A. Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20698–20708. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.02007. URL https://doi.org/10.1109/CVPR52688.2022.02007.

C. Tao, X. Zhu, W. Su, G. Huang, B. Li, J. Zhou, Y. Qiao, X. Wang, and J. Dai. Siamese image modeling for self-supervised vision representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2132–2141, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00212. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00212.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14411–14420, 2021. URL https://api.semanticscholar.org/CorpusID:245006078.

Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 238–248, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59719-1.

James J L Tee, Joseph Carroll, Andrew R Webster, and Michel Michaelides. Quantitative analysis of retinal structure using spectral-domain optical coherence tomography in RPGR-associated retinopathy. *Am. J. Ophthalmol.*, 178:18–26, June 2017.

Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. Isd: Self-supervised learning by iterative similarity distillation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9589–9598, 2020. URL https://api.semanticscholar.org/CorpusID:229297747.

Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Constrained mean shift for representation learning. *CoRR*, abs/2110.10309, 2021. URL https://arxiv.org/abs/2110.10309.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 776–794, Cham, 2020a. Springer International Publishing. ISBN 978-3-030-58621-8.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Yonglong Tian, Olivier J. Hénaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10043–10054, 2021. doi: 10.1109/ICCV48922.2021.00991.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng.*, 6(12):1399–1406, December 2022.

Minh-Son To, Ian G. Sarno, Chee Chong, Mark Jenkinson, and Gustavo Carneiro. Self-supervised lesion change detection and localisation in longitudinal multiple sclerosis brain imaging. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 670–680, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2.

Martine Toering, Ioannis Gatopoulos, Maarten C. Stol, and Vincent Tao Hu. Self-supervised video representation learning with cross-stream prototypical contrasting. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 846–856, 2021. URL https://api.semanticscholar.org/CorpusID:235485193.

Devavrat Tomar, Behzad Bozorgtabar, Manana Lortkipanidze, Guillaume Vray, Mohammad Saeed Rad, and Jean-Philippe Thiran. Self-supervised generative style transfer for one-shot medical image segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1737–1747, 2021. URL https://api.semanticscholar.org/CorpusID:238354132.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Emily Tsai, Scott Simpson, Matthew P Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, Jody Shen, Mona A F Hafez, Susan John, Prabhakar Rajiah, Brian P Pogatchnik, John Thomas Mongan, Emre Altinmakas, Erik Ranschaert, Felipe Campos Kitamura, Laurens Topff, Linda Moy, Jeffrey P Kanne, and Carol C Wu. Medical imaging data resource center - RSNA international COVID radiology database release 1a - chest CT covid+ (MIDRC-RICORD-1a), 2020.

Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=gV3wdEOGy_V.

Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=068E_JSq9O.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *9th International Conference on Learning Representations, ICLR 2021,*

Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021c. URL https://openreview.net/forum?id=-bdp_8Itjwp.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5(1):180161, August 2018.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Vladimír Ulman, Martin Maška, Klas E G Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, Ihor Smal, Karl Rohr, Joakim Jaldén, Helen M Blau, Oleh Dzyubachyk, Boudewijn Lelieveldt, Pengdong Xiao, Yuexiang Li, Siu-Yeung Cho, Alexandre C Dufour, Jean-Christophe Olivo-Marin, Constantino C Reyes-Aldasoro, Jose A Solis-Lemus, Robert Bensch, Thomas Brox, Johannes Stegmaier, Ralf Mikut, Steffen Wolf, Fred A Hamprecht, Tiago Esteves, Pedro Quelhas, Ömer Demirel, Lars Malmström, Florian Jug, Pavel Tomancak, Erik Meijering, Arrate Muñoz-Barrutia, Michal Kozubek, and Carlos Ortiz-de Solorzano. An objective comparison of cell-tracking algorithms. *Nat. Methods*, 14(12):1141–1152, December 2017.

Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 268–285, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58607-2.

Anuja Vats, Marius Pedersen, and Ahmed Mohammed. A preliminary analysis of self-supervision for wireless capsule endoscopy. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, 2021. doi: 10.1109/EUVIP50544.2021.9484012.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.*, 2017:1–9, 2017.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 16451–16467, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8929c70f8d710e412d38da624b21c3c8-Abstract.html.

Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 755–769. PMLR, 06–07 Aug 2021. URL https://proceedings.mlr.press/v149/vu21a.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset, 2022.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *CoRR*, abs/2110.07402, 2021a. URL https://arxiv.org/abs/2110.07402.

G. Wang, K. Wang, G. Wang, P. S. Torr, and L. Lin. Solving inefficiency of self-supervised representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9485–9495, Los Alamitos, CA, USA, oct 2021b. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00937. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00937.

H. Wang, Y. Tang, Y. Wang, J. Guo, Z. Deng, and K. Han. Masked image modeling with local multi-scale reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2122–2131, Los Alamitos, CA, USA, jun 2023a. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00211. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00211.

J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4001–4010, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00413. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00413.

J. Wang, G. Bertasius, D. Tran, and L. Torresani. Long-short temporal contrastive learning of video transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13990–14000, Los Alamitos, CA, USA, jun 2022a. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01362. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01362.

Jiacheng Wang, Hao Li, Han Liu, Dewei Hu, Daiwei Lu, Keejin Yoon, Kelsey M. Barter, Francesca R Bagnato, and Ipek Oguz. Ssl2: self-supervised learning meets semi-supervised learning: multiple sclerosis segmentation in 7t-mri from large-scale 3t-mri. In *Medical Imaging*, 2023b. URL https://api.semanticscholar.org/CorpusID:257427075.

L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, Los Alamitos, CA, USA, jun 2023c. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01398. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01398.

Luyang Wang, Feng Liang, Yangguang Li, Honggang Zhang, Wanli Ouyang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. In *International Joint Conference on Artificial Intelligence*, 2022b. URL https://api.semanticscholar.org/CorpusID:246035555.

R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y. Jiang, L. Zhou, and L. Yuan. Bevt: Bert pretraining of video transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14713–14723, Los Alamitos, CA, USA, jun 2022c. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01432. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01432.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Wei-Chien Wang, Euijoon Ahn, Dagan Feng, and Jinman Kim. A review of predictive and contrastive self-supervised learning for medical images. *Machine Intelligence Research*, 20(4):483–513, Aug 2023d. ISSN 2731-5398. doi: 10.1007/s11633-022-1406-4. URL https://doi.org/10.1007/s11633-022-1406-4.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, Los Alamitos, CA, USA, jul 2017a. IEEE Computer Society. doi: 10.1109/CVPR.2017.369. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.369.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:5549–5560, 2021. URL https://api.semanticscholar.org/CorpusID:233289707.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2794–2802. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.320. URL https://doi.org/10.1109/ICCV.2015.320.

Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1338–1347, 2017b. doi: 10.1109/ICCV.2017.149.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3023–3032, 2020. URL https://api.semanticscholar.org/CorpusID:227012687.

Xudong Wang, Ziwei Liu, and Stella X. Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12581–12590, 2021c. doi: 10.1109/CVPR46437.2021.01240.

Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. Contrastive masked autoencoders for self-supervised video hashing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023e. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25373. URL https://doi.org/10.1609/aaai.v37i3.25373.

Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, N. Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16569–16578, 2021d. URL https://api.semanticscholar.org/CorpusID:236087216.

Zhe Wang, Hao Chen, Xinyu Li, Chunhui Liu, Yuanjun Xiong, Joseph Tighe, and Charless C. Fowlkes. Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 175–184, 2021e. URL https://api.semanticscholar.org/CorpusID:239768686.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018.

C. Wei, L. Xie, X. Ren, Y. Xia, C. Su, J. Liu, Q. Tian, and A. L. Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1910–1919, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00201. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00201.

Chen Wei, Huiyu Wang, Wei Shen, and Alan L. Yuille. CO2: consistent contrast for unsupervised visual representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=U4XLJhqwNF1.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14648–14658, 2022. doi: 10.1109/CVPR52688.2022.01426.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16284–16294, October 2023.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, jun 2009. ISSN 1532-4435.

Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised multi-modal alignment for whole body medical imaging. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 90–101, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

Steffen Wolf, Manan Lalit, Henry Westmacott, Katie McDole, and Jan Funke. Unsupervised learning of object-centric embeddings for cell instance segmentation in microscopy images. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21206–21215, 2023. URL `https://api.semanticscholar.org/CorpusID:263908915`.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, June 2023.

Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10129–10139, 2021. URL `https://api.semanticscholar.org/CorpusID:234482727`.

Jiantao Wu and Shentong Mo. Object-wise masked autoencoders for fast pre-training, 2022.

Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Yue Huang, Qinji Yu, Sifan Song, Xinxing Xu, Yanyu Xu, Wensai Wang, Lingxiao Wang, Shuai Lu, Huiqi Li, Shihua Huang, Zhichao Lu, Chubin Ou, Xifei Wei, Bingyuan Liu, Riadh Kobbi, Xiaoying Tang, Li Lin, Qiang Zhou, Qiang Hu, Hrvoje Bogunović, José Ignacio Orlando, Xiulan Zhang, and Yanwu Xu. GAMMA challenge: Glaucoma grading from multi-modality images. *Med. Image Anal.*, 90(102938):102938, December 2023a.

Junjie Wu and Dit-Yan Yeung. SCAT: robust self-supervised contrastive learning via adversarial training for text classification. *CoRR*, abs/2307.01488, 2023. doi: 10.48550/ARXIV.2307.01488. URL `https://doi.org/10.48550/arXiv.2307.01488`.

Lirong Wu, Haitao Lin, Cheng Tan, Zhangyang Gao, and Stan Z. Li. Self-supervised learning on graphs: Contrastive, generative, or predictive. *IEEE Trans. on Knowl. and Data Eng.*, 35(4):4216–4235, apr 2023b. ISSN 1041-4347. doi: 10.1109/TKDE.2021.3131584. URL `https://doi.org/10.1109/TKDE.2021.3131584`.

Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah D. Goodman. Conditional negative sampling for contrastive learning of visual representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL `https://openreview.net/forum?id=v8b3e5jN66j`.

Ming-Ju Wu, Jyh-Shing R. Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1):1–12, 2015a. doi: 10.1109/TSM.2014.2364237.

Quanlin Wu, Hang Ye, Yuntian Gu, Huishuai Zhang, Liwei Wang, and Di He. Denoising masked autoencoders help robust classification. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023c. URL https://openreview.net/pdf?id=zDjtZZBZtqK.

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. Federated contrastive learning for volumetric medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 367–377, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87199-4.

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. Distributed contrastive learning for medical image segmentation. *Medical Image Analysis*, 81:102564, 2022a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102564. URL https://www.sciencedirect.com/science/article/pii/S1361841522002079.

Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015b. doi: 10.1109/TPAMI.2014.2388226.

Yixuan Wu, Bo Zheng, Jintai Chen, Danny Z. Chen, and Jian Wu. Self-learning and one-shot learning based single-slice annotation for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 244–254, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-16452-1.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. doi: 10.1109/CVPR.2018.00393.

F. Xiao, K. Kundu, J. Tighe, and D. Modolo. Hierarchical self-supervised representation learning for movie understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9717–9726, Los Alamitos, CA, USA, jun 2022a. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00950. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00950.

Fanyi Xiao, Joseph Tighe, and Davide Modolo. Maclr: Motion-aware contrastive learning of representations for videos. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 353–370, Berlin, Heidelberg, 2022b. Springer-Verlag. ISBN 978-3-031-19832-8. doi: 10.1007/978-3-031-19833-5_21. URL https://doi.org/10.1007/978-3-031-19833-5_21.

Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119:3–22, 2014. URL https://api.semanticscholar.org/CorpusID:10224573.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. Early convolutions help transformers see better. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235658393.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 478–487. JMLR.org, 2016.

Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. Instance-aware self-supervised learning for nuclei segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 341–350, Cham, 2020a. Springer International Publishing. ISBN 978-3-030-59722-1.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16679–16688, 2020b. URL https://api.semanticscholar.org/CorpusID:227054503.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *CoRR*, abs/2105.04553, 2021. URL https://arxiv.org/abs/2105.04553.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9643–9653, 2022. doi: 10.1109/CVPR52688.2022.00943.

Y. Xiong, M. Ren, W. Zeng, and R. Waabi. Self-supervised representation learning from flow equivariance. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10171–10180, Los Alamitos, CA, USA, oct 2021a. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.01003. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01003.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, Pheng-Ann Heng, Dong Ni, Caizi Li, Qianqian Tong, Weixin Si, Elodie Puybareau, Younes Khoudli, Thierry Géraud, Chen Chen, Wenjia Bai, Daniel Rueckert, Lingchao Xu, Xiahai Zhuang, Xinzhe Luo, Shuman Jia, Maxime Sermesant, Yashu Liu, Kuanquan Wang, Davide Borra, Alessandro Masci, Cristiana Corsi, Coen de Vente, Mitko Veta, Rashed Karim, Chandrakanth Jayachandran Preetha, Sandy Engelhardt, Menyun Qiao, Yuanyuan Wang, Qian Tao, Marta Nuñez-Garcia, Oscar Camara, Nicolo Savioli, Pablo Lamata, and Jichao Zhao. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Med. Image Anal.*, 67(101832):101832, January 2021b.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10326–10335, 2019. doi: 10.1109/CVPR.2019.01058.

Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Hongkai Xiong, and Qi Tian. Hierarchical semantic aggregation for contrastive representation learning. *ArXiv*, abs/2012.02733, 2020. URL https://api.semanticscholar.org/CorpusID:227305667.

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations*, 2021a. URL https://api.semanticscholar.org/CorpusID:247519082.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10055–10065, 2021. doi: 10.1109/ICCV48922.2021.00992.

Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2023. doi: 10.1109/TKDE.2022.3193569.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016. doi: 10.1109/CVPR.2016.571.

Junshen Xu, Esra Abaci Turk, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Stress: Super-resolution for dynamic fetal mri using self-supervised learning. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 197–206, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87234-2.

Hanyu Xuan, Zhiliang Wu, Jian Yang, Bo Jiang, Lei Luo, Xavier Alameda-Pineda, and Yan Yan. Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos. *IEEE transactions on pattern analysis and machine intelligence*, PP, 02 2024. doi: 10.1109/TPAMI.2024.3363508.

Yingying Xue, Shixiang Feng, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. Dual-task self-supervision for cross-modality domain adaptation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 408–417, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020. doi: https://doi.org/10.1002/mrm.28378. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.28378.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J. Med. Imaging (Bellingham)*, 5(3): 036501, July 2018.

Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Adam P. Harrison, Dazhou Guo, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, 41:2658–2669, 2020. URL https://api.semanticscholar.org/CorpusID:227305852.

Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 3045–3053. AAAI Press, 2022a. doi: 10.1609/AAAI.V36I3.20211. URL https://doi.org/10.1609/aaai.v36i3.20211.

Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023a. doi: 10.1109/TPAMI.2023.3257878.

Haiyang Yang, Shixiang Tang, Meilin Chen, Yizhou Wang, Feng Zhu, Lei Bai, Rui Zhao, and Wanli Ouyang. Domain invariant masked autoencoders for self-supervised learning from multi-domains. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pp. 151–168. Springer, 2022b. doi: 10.1007/978-3-031-19821-2_9. URL https://doi.org/10.1007/978-3-031-19821-2_9.

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5147–5156, 2016. doi: 10.1109/CVPR.2016.556.

Junlin Yang, Xiaoxiao Li, Daniel Pak, Nicha C. Dvornek, Julius Chapiro, MingDe Lin, and James S. Duncan. Cross-modality segmentation by self-supervised semantic alignment in disentangled content space. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pp. 52–61, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60548-3.

Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. Mrm: Masked relation modeling for medical image pre-training with genetics. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21395–21405, 2023b. URL https://api.semanticscholar.org/CorpusID:267023203.

Yang Yang, Bo Wang, Dingwen Zhang, Yixuan Yuan, Qingsen Yan, Shijie Zhao, Zheng You, and J. Han. Self-supervised interactive embedding for one-shot organ segmentation. *IEEE Transactions on Biomedical Engineering*, 70:2799–2808, 2023c. URL https://api.semanticscholar.org/CorpusID:261696036.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382.

Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:220935968.

M. Ye, X. Zhang, P. C. Yuen, and S. Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6203–6212, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00637. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00637.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 668–684, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19809-0.

Pak-Hei Yeung, Ana I. L. Namburete, and Weidi Xie. Sli2vol: Annotate a 3d volume from a single slice with self-supervised learning. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 69–79, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=1fZd4owfJP6.

Chenyu You, Ruihan Zhao, Lawrence H. Staib, and James S. Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 13434:639–652, 2021. URL https://api.semanticscholar.org/CorpusID:234742038.

J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge & Data Engineering*, 36(01):335–355, jan 2024a. ISSN 1558-2191. doi: 10.1109/TKDE.2023.3282907.

Ke Yu, Li Sun, Junxiang Chen, Maxwell Reynolds, Tigmanshu Chaudhary, and Kayhan Batmanghelich. Drasclr: A self-supervised framework of learning disease-related and anatomy-specific representation for 3d lung ct images. *Medical Image Analysis*, 92:103062, 2024b. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.103062. URL https://www.sciencedirect.com/science/article/pii/S1361841523003225.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 6995–7004. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00692. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yuan_Multimodal_Contrastive_Training_for_Visual_Representation_Learning_CVPR_2021_paper.html.

Yuan Yuan, Euijoon Ahn, Dagan Feng, Mohamad Khadra, and Jinman Kim. SSPT-bpMRI: A self-supervised pre-training scheme for improving prostate cancer detection and diagnosis in bi-parametric MRI. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2023:1–4, July 2023.

S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00612. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00612.

Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8344–8353. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00817. URL https://doi.org/10.1109/CVPR52688.2022.00817.

Anna Zawacki, Carol Wu, George Shih, Julia Ellitt, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. URL https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6687–6696, 2020. URL https://api.semanticscholar.org/CorpusID:219792957.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D. Yoo, and In So Kweon. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 725–742, Cham, 2022a. Springer Nature Switzerland. ISBN 978-3-031-20056-4.

Chuyan Zhang, Hao Zheng, and Yun Gu. Dive into the details of self-supervised learning for medical image analysis. *Medical Image Analysis*, 89:102879, 2023a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102879. URL https://www.sciencedirect.com/science/article/pii/S1361841523001391.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Supporting clustering with contrastive learning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5419–5430, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.427. URL https://aclanthology.org/2021.naacl-main.427.

Hao Zhang, Sheng Xu, Wei Ren, Huping Ye, and Yi Hong. Pretrain once and finetune many times: How pretraining benefits brain mri segmentation. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1724–1731, 2023b. URL https://api.semanticscholar.org/CorpusID:267044772.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Hui Zhang, Jiaxuan Liu, Tianyue Wu, Yurong Chen, Caiping Liu, and Yaonan Wang. Jianet: Jigsaw-invariant self-supervised learning of autoencoder-based reconstruction for melanoma segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–13, 2022b. doi: 10.1109/TIM.2022.3218033.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 649–666, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.

Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16579–16588, 2022c. doi: 10.1109/CVPR52688.2022.01610.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022d. URL https://openreview.net/forum?id=RAW9tCdVxLj.

Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. Dense siamese network for dense unsupervised learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pp. 464–480. Springer, 2022e. doi: 10.1007/978-3-031-20056-4_27. URL https://doi.org/10.1007/978-3-031-20056-4_27.

Xiaoman Zhang, Shixiang Feng, Yuhang Zhou, Ya Zhang, and Yanfeng Wang. Sar: Scale-aware restoration learning for 3d tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 124–133, Cham, 2021b. Springer International Publishing. ISBN 978-3-030-87196-3.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14, 2023c. URL https://api.semanticscholar.org/CorpusID:257220515.

Xiaoman Zhang, Weidi Xie, Chaoqin Huang, Ya Zhang, Xin Chen, Qi Tian, and Yanfeng Wang. Self-supervised tumor segmentation with sim2real adaptation. *IEEE Journal of Biomedical and Health Informatics*, 27:4373–4384, 2023d. URL https://api.semanticscholar.org/CorpusID:256483217.

Xiaosong Zhang, Yunjie Tian, Wei Huang, Qixiang Ye, Qi Dai, Lingxi Xie, and Qi Tian. Hivit: Hierarchical vision transformer meets masked image modeling. *CoRR*, abs/2205.14949, 2022f. doi: 10.48550/ARXIV.2205.14949. URL https://doi.org/10.48550/arXiv.2205.14949.

Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023e. URL https://openreview.net/pdf?id=3F6I-0-57SC.

Yejia Zhang, Pengfei Gu, Nishchal Sapkota, Hao Zheng, Peixian Liang, and Danny Z. Chen. A point in the right direction: Vector prediction for spatially-aware self-supervised volumetric representation learning. In *20th IEEE International Symposium on Biomedical Imaging, ISBI 2023, Cartagena, Colombia, April 18-21, 2023*, pp. 1–5. IEEE, 2023f. doi: 10.1109/ISBI53787.2023.10230378. URL https://doi.org/10.1109/ISBI53787.2023.10230378.

Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29848–29860, 2021c. URL https://proceedings.neurips.cc/paper/2021/hash/fa14d4fe2f19414de3ebd9f63d5c0169-Abstract.html.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2022, 5-6 August 2022, Durham, NC, USA*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022g. URL https://proceedings.mlr.press/v182/zhang22a.html.

Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 975–985, 2022. doi: 10.1109/WACV51458.2022.00105.

Can Zhao, Aaron Carass, Blake E. Dewey, and Jerry L. Prince. Self super-resolution for magnetic resonance images using deep networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 365–368, 2018. doi: 10.1109/ISBI.2018.8363594.

Can Zhao, Blake E. Dewey, Dzung L. Pham, Peter A. Calabresi, Daniel S. Reich, and Jerry L. Prince. Smore: A self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. *IEEE Transactions on Medical Imaging*, 40(3):805–817, 2021. doi: 10.1109/TMI.2020.3037187.

Hang Zhao, Chen Zhang, Belei Zhu, Zejun Ma, and Kejun Zhang. S3T: self-supervised pre-training with swin transformer for music classification. *CoRR*, abs/2202.10139, 2022. URL https://arxiv.org/abs/2202.10139.

Liang Zhao, Chaoran Jia, Jiajun Ma, Yu Shao, Zhuo Liu, and Hong Yuan. Medical image segmentation based on self-supervised hybrid fusion network. *Front. Oncol.*, 13:1109786, April 2023.

Xi Zhao and ShuiSheng Zhou. Fast mixing of hard negative samples for contrastive learning and use for covid-19. In *Proceedings of the 4th International Conference on Big Data Technologies*, ICBDT '21, pp. 6–12, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450385091. doi: 10.1145/3490322.3490324. URL https://doi.org/10.1145/3490322.3490324.

E. Zheltonozhskii, C. Baskin, A. Mendelson, A. M. Bronstein, and O. Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 387–397, Los Alamitos, CA, USA, jan 2022. IEEE Computer Society. doi: 10.1109/WACV51458.2022.00046. URL https://doi.ieeecomputersociety.org/10.1109/WACV51458.2022.00046.

Hao Zheng, Jun Han, Hongxiao Wang, Lin Yang, Zhuo Zhao, Chaoli Wang, and Danny Z. Chen. Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I*, pp. 622–632, Berlin, Heidelberg, 2021a. Springer-Verlag. ISBN 978-3-030-87192-5. doi: 10.1007/978-3-030-87193-2_59. URL https://doi.org/10.1007/978-3-030-87193-2_59.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Sci. Data*, 7(1): 48, February 2020.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In *Neural Information Processing Systems*, 2021b. URL https://api.semanticscholar.org/CorpusID:236134328.

Ruifeng Zheng, Ying Zhong, Senxiang Yan, Hongcheng Sun, Haibin Shen, and Kejie Huang. Msvrl: Self-supervised multiscale visual representation learning via cross-level consistency for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42:91–102, 2022. URL https://api.semanticscholar.org/CorpusID:252088219.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.319. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 487–495, Cambridge, MA, USA, 2014. MIT Press.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017. doi: 10.1109/CVPR.2017.544.

Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3479–3489, 2021a. doi: 10.1109/ICCV48922.2021.00348.

Hong-Yu Zhou, Chi-Ken Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:8020–8035, 2023. URL https://api.semanticscholar.org/CorpusID:255708339.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=ydopy-e6Dg.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. Debiased contrastive learning of unsupervised sentence representations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6120–6130, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.423. URL https://aclanthology.org/2022.acl-long.423.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. Mugs: A multi-granular self-supervised learning framework. *ArXiv*, abs/2203.14415, 2022c. URL https://api.semanticscholar.org/CorpusID:247762920.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021b. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101840. URL https://www.sciencedirect.com/science/article/pii/S1361841520302048.

Jiachen Zhu, Rafael M. Moraes, Serkan Karakulak, Vlad Sobol, Alfredo Canziani, and Yann LeCun. Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. *CoRR*, abs/2206.10698, 2022. doi: 10.48550/ARXIV.2206.10698. URL https://doi.org/10.48550/arXiv.2206.10698.

Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S. Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101746. URL https://www.sciencedirect.com/science/article/pii/S1361841520301109.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10286–10295, 2021. URL https://api.semanticscholar.org/CorpusID:236950631.

C. Zhuang, A. Zhai, and D. Yamins. Local aggregation for unsupervised learning of visual embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6001–6011, Los Alamitos, CA, USA, nov 2019a. IEEE Computer Society. doi: 10.1109/ICCV.2019.00610. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00610.

Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 420–428, Cham, 2019b. Springer International Publishing. ISBN 978-3-030-32251-9.

Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Learning the retinal anatomy from scarce annotated data using self-supervised multimodal reconstruction. *Applied Soft Computing*, 91:106210, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106210. URL https://www.sciencedirect.com/science/article/pii/S1568494620301502.

Álvaro S. Hervella, José Rouco, Jorge Novo, and Marcos Ortega. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications*, 185:115598, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115598. URL https://www.sciencedirect.com/science/article/pii/S0957417421009982.