The Impact of Semantic Pairs on Self-Supervised Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Instance discrimination learns visual representations by treating different augmented views of the same image as positive pairs. While this encourages invariance to handcrafted transformations, same-image positives can preserve nuisance correlations such as background, texture, illumination, and object-specific details. Semantic positive pairs, i.e., different same-class instances, may reduce these correlations by presenting objects across diverse contexts. However, previous studies often combine semantic pairs with augmented positives or false neighbors (i.e., incorrectly mapped semantic pairs), making it difficult to isolate the effect of semantic pairing. We present a controlled empirical study of semantic positive pairs for self-supervised representation learning. From ImageNet-1K, we construct two matched subsets: an augmented-pair baseline and a manually curated semantic-pair dataset with the same class composition and training-pair count. We use these datasets to compare representative contrastive and non-contrastive SSL methods under matched training conditions. Across transfer learning and object detection evaluations, semantic-pair pretraining consistently improves generalisation over augmented-pair pretraining. Additional ablations show that semantic pairs induce invariances beyond the standard transformation pipeline. Among the evaluated methods, contrastive learning benefits most strongly from semantic pairs, with SimCLR showing the largest relative improvement. These results clarify the role of semantic positive pairs in SSL and provide guidance for selecting and designing frameworks that can exploit semantic pair information effectively.

1 Introduction

Instance discrimination is a self-supervised learning (SSL) approach that allows the model to learn useful object representations without manual data annotation (Zhang et al., 2025). These approaches treat each instance in the dataset as a class and discriminate it from the other instances. To achieve this, they attract the two augmented views (i.e., positive pairs) of the same instance in the embedding space and use different techniques to avoid collapse to the trivial representation (Bardes et al., 2021; Chen et al., 2020a;b; Chen & He, 2021). Although data transformations, such as geometric (e.g., random cropping, rotation) and appearance-based (e.g., colour jitter, blur), help the model learn invariant representations against nuisances, relying solely on transformations introduces limitations that may constrain generalisation to unseen data (Alkhalefi et al., 2024a; Tian et al., 2020; Xiao et al., 2020). Recent work on data curation for visual contrastive learning similarly argues that single-instance positive-pair construction can limit data diversity, since stochastic augmentations may fail to capture important variations such as viewpoint changes, object deformations, and semantically similar same-class instances (Desai et al., 2025). To better understand the limitations of data transformation, we must realise how instance discrimination approaches learn object representations from positive views. In Figure 1(a), the learnt representations will capture the shared information between the two views (x^1 and x^2) (Alkhalefi et al., 2024b; Tian et al., 2020; Zhang et al., 2022; Meehan et al., 2023). As shown in the Figure, the two views still have similar backgrounds and similar door stickers on the truck. This might lead to inappropriate relationships between the background and foreground, as well as other details, such as the 'sticker', which reduces the model's ability to generalise to the unseen data (Meehan et al., 2023; Zeng et al., 2024). The solution to this problem is to reduce the

shared information between the two views while keeping the relevant task information intact (i.e., truck semantic features) (Tian et al., 2020). Semantic pairs (i.e., two different instances belonging to the same class) place the object in different contexts, allowing the model to focus on the relevant shared information and downweight the nuisance information. In Figure 1(b), the aim is to learn a model that captures the general representation of the tow truck by exposing the model to two views containing 'tow trucks' but in different contexts. Since the captured representation relies on shared information across the two views (i.e., the same information consistently present in both views), placing the truck in different contexts encourages the model to focus on the relevant task information (i.e., learn the truck representation) and discard the irrelevant information needed to recognize the truck (i.e., background and sticker on the door). This leads the model to create a general representation of the tow truck based on the semantic features of the truck (i.e., cab, lights, tow part, etc.) while reducing the associative relationship between the 'tow truck' and other irrelevant information present in the image. In other words, the model creates a general representation of the truck; this representation captures the semantic features of the truck and is invariant to nuisance information, enabling it to recognise the truck in different contexts during downstream tasks.

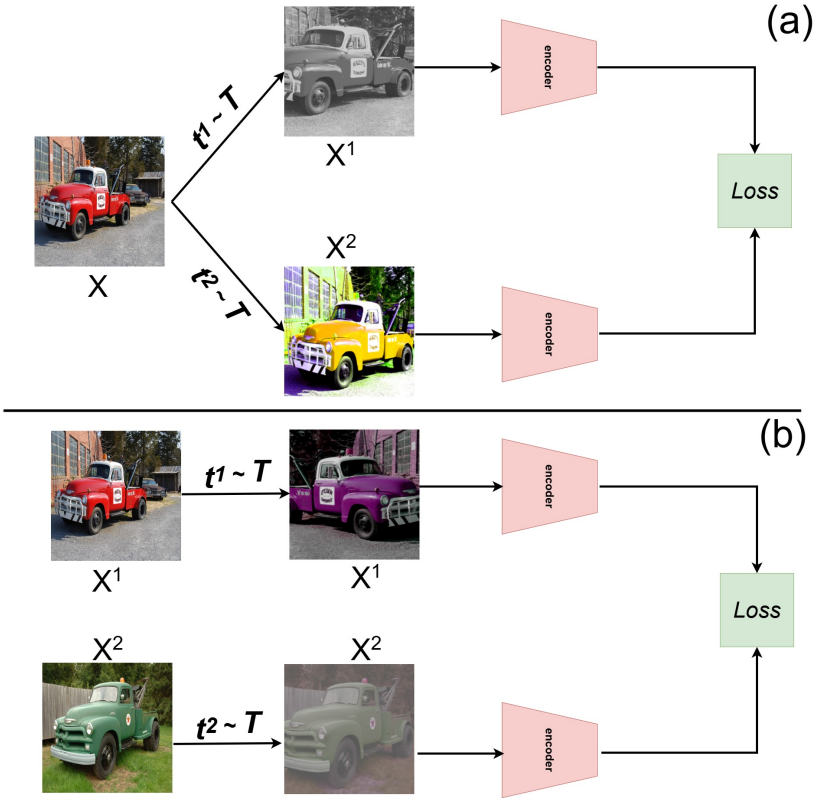

Figure 1: This diagram illustrates how instance discrimination approaches treat (a) augmented pairs (i.e., two views for the same instance) and (b) semantic pairs (i.e., two instances belonging to the same category) during representation learning.

Training the model on semantic pairs using self-supervised instance discrimination approaches exposes it to various contextual instantiations of an object, which enhances the learning of the object's high-level features (i.e., semantic features). Consequently, it improves model generalization and enhances the model's performance on unseen data. This study demonstrates how semantic pairs enhance the representation learning of instance discrimination approaches and conducts experiments to validate the theory behind using semantic pairs. Pretraining on semantic pairs guides the learning process of visual representations towards greater invariance to factors such as occlusion, background, pattern, and illumination, supporting the goal of learning general representations.

Occlusion Invariance. Occlusion invariance means the model can recognize objects even when parts of them are hidden or obscured. This is a crucial property for robust object recognition, as objects in real-world scenes are often partially occluded by other objects. The semantic pairs provide the model with views for the same category but with different types of occlusion. This allows the model to create representations for the shared information of the object between the two views, treat the other irrelevant information as noise, and discard it (Purushwalkam & Gupta, 2020; Alkhalefi et al., 2024b; Zeng et al., 2024; Wang & Qi, 2022). For example, in Figure 2 (occlusion '2'), the first view shows a bird partially blocked by tree branches, while another view shows a complete bird. When the objective function forces the model to develop a similar representation for both views, it captures the remaining semantic features of the birds (i.e., head, body, and tail) present in both views and dismisses the tree branches that appear in only one view. Exposing the model to various instances of natural occlusions (different objects hiding different parts of the bird) can encourage learning a more general representation of "bird" that is not tied to specific viewpoints or occlusion patterns.

Background invariance. Background invariance means the model can recognize objects in different contexts or backgrounds. In SSL instance discrimination, the model treats each instance as a class of its own and learns the representation of the object based on the shared information between the two augmented views. This might lead to memorizing the images and associating a relation between the background and foreground object despite no relation between them (Meehan et al., 2023; Zeng et al., 2024). In this case, the generalization of representations is diminished, which limits the model's ability to recognize objects with different backgrounds in downstream tasks. Semantic pairs represent the same object from two views but within different contexts. By varying the background between paired views, the instance discrimination model focuses on the semantic features of the object depicted in the two views, treating the background as less relevant for object recognition (Zeng et al., 2024). This breaks irrelevant relations between foreground and background and allows the model to capture better semantic features for the object, which leads to recognizing the object in different contexts and improving the performance of the models on different downstream tasks. For example, in Figure 2 (Background '1'), the model is presented with two views of the birdhouse, each with a distinct background. This encourages the model to focus on the semantic features of the birdhouse and reduce the influence of the background, as it is less relevant for recognizing the birdhouse. Thus, the model learns to be invariant to variations in background (i.e., it can recognize the object in the downstream task, regardless of the background).

Abstract representation. In real-world scenarios, identical objects, such as a brand's logo and decorative paint designs, can exhibit varying surface patterns. Training a model on semantically paired data creates a generalized representation for the object by leveraging the common information found in these two distinct views. This allows the model to recognize the object in various downstream tasks, regardless of its pattern. For example, in Figure 2 (pattern '2'), the model is given two aeroplane images, each featuring a different brand's logo (i.e., the Airways brand). The model will capture the semantic features of the aeroplane (i.e., tail, wings, fuselage, etc.) that exist in both views to create a general representation of the aeroplane, treating the logos as noise and discarding them. This enables the model to recognise the aeroplane with a different pattern in downstream tasks, known as pattern invariance.

Illumination invariance. Since lighting conditions change in the real world (e.g., dim indoor lights, outdoor lights, shadows, reflections), we want models to recognize objects regardless of illumination. This is called illumination invariance. To achieve that, we need to train the model on the same object with different illumination (Xiao et al., 2024). Semantic pairs enable the model to train on the same object under varying illumination conditions. For example, if a model is trained on two views of the same object e.g., train, each with different lighting, as shown in Figure 2 (illumination '3'), it is encouraged to focus on the train's semantic features rather than the illumination. Because the illumination varies between the two views, the model learns to be less sensitive to it, effectively treating the illumination differences as less important for object recognition. This helps create a representation of the train that is more invariant to variations in illumination.

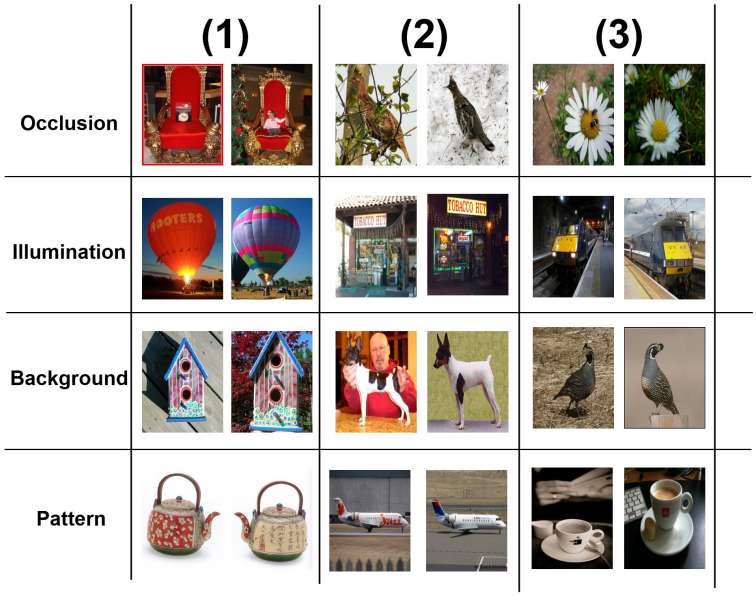

Figure 2: Examples from the semantic pairs dataset illustrate placing objects in diverse real-world context scenes.

Our contributions are as follows:

- We elucidate the mechanism by which semantic pairing induces representation invariance that leads to enhanced generalization capabilities.
- We conduct a controlled empirical study using a custom-built dataset, allowing us to uniquely isolate the impact of semantic pairs from standard positive pairs and false neighbors during self-supervised learning.
- We compare representative contrastive and non-contrastive SSL methods and show that semantic pairs consistently improve downstream transfer performance, with contrastive learning benefiting most strongly.
- We conduct extensive ablation studies to evaluate the sensitivity of semantic pairing to various data transformations, architectural choices, dataset sizes, and occlusion robustness.

2 Related Work

SSL has emerged as a powerful paradigm for learning representations from unlabeled data. SSL approaches are divided into two broad categories: contrastive and non-contrastive learning. These approaches often use pretext tasks to learn useful representations from unlabeled data. This section provides a brief overview of some of these approaches and benchmark datasets, such as ImageNet or CIFAR-10, that are used with SSL. We encourage readers to refer to the respective papers for more details.

Dataset: Datasets like STL-10 (Coates et al., 2011), CIFAR (Krizhevsky et al., 2009), Tiny-ImageNet (Le et al., 2015), ImageNet-100, and ImageNet-1K (Russakovsky et al., 2015), all organise images into labelled classes, which is a fundamental structure. However, they vary significantly in size and complexity, making them valuable benchmark datasets for various machine learning tasks. Notably, none of these datasets explicitly contains semantic pairs, requiring SSL models to identify these relationships directly from the data during training. In this study, we create a dataset containing curated semantic pairs, which reduces the issues of mapping inaccurate semantic pairs during model training (i.e., false neighbors). Thus, this allows us to focus on studying the effect of semantic pairs in the SSL context.

Contrastive Learning: Instance discrimination methods, such as SimCLR, MoCo, and PIRL (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Misra & Maaten, 2020) employ a similar idea. They attract the positive pairs together and push the negative pairs apart in the embedding space, albeit through a different mechanism. SimCLR (Chen et al., 2020a) employs an end-to-end approach where a large batch size is used for the negative examples and both encoders' parameters in the Siamese network are updated together. PIRL (Misra & Maaten, 2020) uses a memory bank for negative examples, and both encoders' parameters are updated together. MoCo (Chen et al., 2020b; He et al., 2020) employs a momentum contrastive approach, where the query encoder is updated during backpropagation and, in turn, updates the key encoder. Negative examples are stored in a separate dictionary, allowing for the use of large batch sizes.

Non-Contrastive Learning: Non-contrastive approaches use only positive pairs to learn visual representations employing different methods to avoid representation collapse. The first category is clustering-based methods, where samples with similar features are assigned to the same cluster. DeepCluster (Caron et al., 2018) uses pseudo-labels from the previous iteration, which makes it computationally expensive and hard to scale. SWAV (Caron et al., 2020) addresses this issue by using online clustering, though it requires determining the correct number of prototypes. The second category involves knowledge distillation. Methods like BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) draw on knowledge distillation techniques, where a Siamese network comprises an online encoder and a target encoder. The target network parameters are not updated during backpropagation. Instead, only the online network parameters are updated while being encouraged to predict the representation of the target network. Despite their promising results, the mechanisms by which these methods avoid collapse are not yet fully understood. Inspired by BYOL, Self-distillation with no labels (DINO) (Caron et al., 2021) uses centering and sharpening, along with a different backbone (ViT), which allows it to outperform other self-supervised methods while being more computationally efficient. Another approach, Bag of visual words (Gidaris et al., 2020; 2021), utilises a teacher-student framework inspired by natural language processing (NLP) to avoid representation collapse. The student network predicts a histogram of the features for augmented images, similar to the teacher network's histogram. The last category is information maximisation. Methods like Barlow twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021) do not require negative examples, stop gradient or clustering. Instead, they use regularisation to avoid representation collapse. The objective function of these methods tries to eliminate the redundant information in the embeddings by making the correlation of the embedding vectors closer to the identity matrix. While these methods show promising results, they have limitations, such as the representation learning being sensitive to regularisation and decreased effectiveness if certain statistical properties are missing in the data.

Enhanced Instance Discrimination: Efforts to enhance contrastive learning have significantly focused on refining negative examples to improve representation quality. For instance, (Kalantidis et al., 2020) and (Robinson et al., 2020) emphasized mining "hard" negative samples close to the positive anchor, while (Wu et al., 2020) proposed a percentile-based strategy for negative sampling. Another approach by (Chuang et al., 2020) involved assigning weights to positive and negative terms to mitigate the impact of irrelevant negatives. (Dwibedi et al., 2021; Koohpayegani et al., 2021; Huynh et al., 2022; Auh et al., 2023; Alkhalefi et al., 2024a; Zeng et al., 2024) enhance representation learning by utilising semantic pairs to increase data diversity and learn invariant representations to intra-class variation.

Although the aforementioned approaches utilise semantic pairs to enhance representation learning in instance discrimination, accurately discerning the specific contribution of semantic pairs to the observed improvements in model performance requires careful and rigorous analysis. A potential confounding factor arises from the integration of semantic pairs with either augmented pairs or false neighbours in existing approaches, complicating the isolation of the unique impact of semantic pairs on self-supervised learning. In addition, the mentioned approaches employed semantic pairs with either a contrastive approach, such as NNCLR (Dwibedi et al., 2021), or a non-contrastive approach, such as MSF (Koohpayegani et al., 2021), without providing clear justification for their choice of approach with semantic pairs. Therefore, this study aims to conduct a rigorous investigation to precisely determine the impact of semantic pairs on the performance and generalization capabilities of the SSL approaches. Additionally, it seeks to identify the most effective approach for leveraging semantic pairs to learn invariant representations, thereby enhancing the model's generalization on unseen datasets. To achieve this, a curated dataset containing semantic pairs is constructed

to eliminate confounding factors present in previous approaches. Subsequently, a comparative analysis is performed by training one model on this meticulously curated semantic pairs dataset and another on standard augmented pairs.

3 Methodology

This study investigates whether incorporating semantically paired images can enhance the generalization capability of SSL models. To assess this, we compare two pretrained models: one using standard augmented image pairs (our baseline) and another pre-trained on semantically related pairs, both derived from the ImageNet-1K dataset (Russakovsky et al., 2015). As illustrated in Figure 3, our comparative methodology comprises four key stages: (1) dataset construction, (2) image transformations, (3) model training, and (4) performance evaluation.

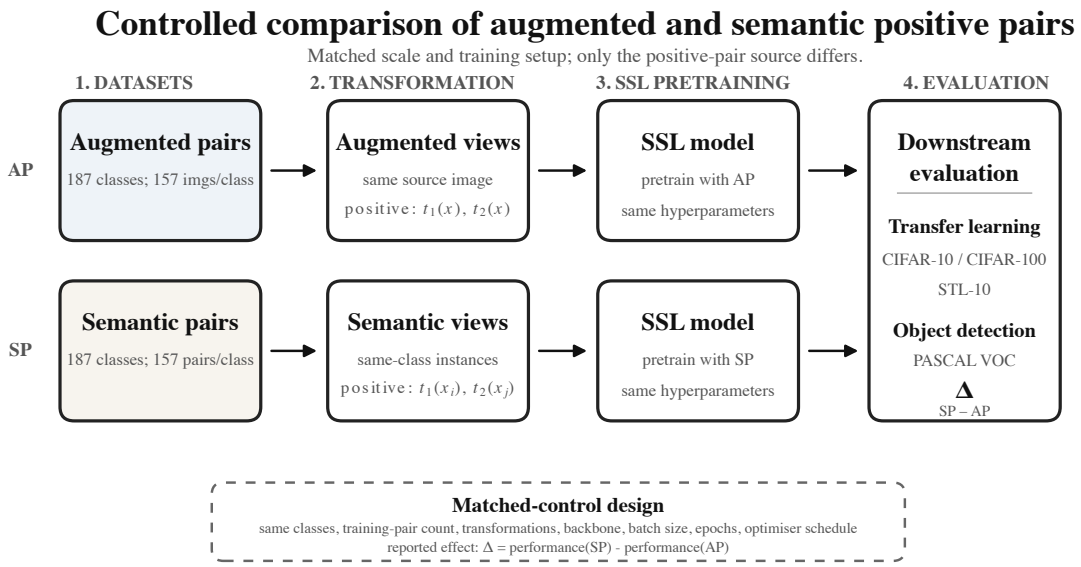

Figure 3: Controlled experimental framework for comparing augmented and semantic positive pairs. The augmented-pair baseline uses two stochastic views of the same image, whereas the semantic-pair dataset uses two different same-class instances. Both branches are matched in class composition, training-pair count, backbone, training schedule, and transformation pipeline. Downstream performance is compared using transfer learning and object detection, with relative gain computed as $\Delta = SP - AP$.

3.1 Dataset Construction

We developed two distinct datasets derived from the ImageNet-1K training split: one comprising augmented pairs (serving as the baseline) and the other consisting of semantic pairs. Both datasets were balanced with respect to the number of classes and images per class to ensure a fair comparison.

3.1.1 Semantic Pairs Dataset

We construct a manually curated dataset of semantic pairs, comprising 187 classes, each containing 157 pairs, resulting in a total of 29,359 semantic pairs (see Appendix A.4). Each pair is manually annotated to ensure precise semantic alignment, mitigating potential misalignments inherent in automated matching methods. For every semantic pair, we identify instances that exhibit semantic correspondence (i.e., the same object) while varying in contextual attributes (See Appendix A (Figures 6 and 7) for illustrated examples). Since the primary objective of this study is to evaluate model generalization on unseen data, we adopted a structured

strategy for class selection rather than sampling random categories from ImageNet-1K. Specifically, we prioritized classes that exhibit semantic overlap with downstream evaluation datasets such as STL-10 and CIFAR. For example, STL-10 consists of ten categories (airplane, bird, car, cat, deer, dog, horse, monkey, ship, and truck). Corresponding categories were identified within ImageNet and integrated into the semantic pairs dataset. This facilitates the evaluation of which model learns more effective representations. For instance, if a model is pretrained on two classes (e.g., dog and cat) but the downstream task involves distinct classes (e.g., car and airplane), both models may exhibit comparably low performance due to the discrepancy between the data distributions of the pretraining and downstream datasets.

3.1.2 Augmented Pairs Dataset (baseline)

The augmented pairs dataset comprises 187 classes, each containing 157 images, resulting in a total of 29,359 images, which maintains consistency with the Semantic pair dataset in both class composition and scale.

3.2 Transformation

During model training, a standard stochastic transformation pipeline is applied, including random cropping, rotation, random horizontal flipping, color jittering, and Gaussian blur, consistent with the approaches included in this study. For the augmented dataset (29,359 images), this process generates synthetic pairs by creating multiple augmented views of the same instance, resulting in 29,359 pairs. In contrast, the semantic dataset consists of predefined pairs, so transformations are applied individually to each image within these pairs without generating additional synthetic pairs. Figure 1 illustrates examples of both transformation approaches, where Figure 1(a–b) presents data transformations for augmented and semantic pairs, respectively.

3.3 Training SSL Models

In this step, we train multiple SOTA self-supervised approaches on both the augmented pair dataset and the semantic pair dataset. First, we compare models trained on augmented pairs against those trained on semantic pairs to evaluate whether semantic pairs help the model generalize better to unseen data. Next, we measure the relative improvement each SSL method achieves when using semantic pairs versus augmented pairs, quantifying the extent to which each model benefits from semantic relationships. This comparison reveals which SSL approach leverages semantic pairs most effectively to learn useful representations.

3.4 Evaluation

After creating both datasets, we eliminate the confounder present in previous approaches, where semantic pairs were combined with either positive pairs or false neighbors. This design enables a direct and fair assessment of their influence on SSL. As is well established, SSL methods are typically evaluated using downstream tasks to assess the quality of learned representations. In this study, model performance is evaluated using two primary downstream tasks: transfer learning and object detection. These tasks assess model generalization on unseen datasets, thereby demonstrating the ability to learn discriminative and transferable features that enable accurate predictions on new data. The evaluation process involves comparing each approach both with itself and with other approaches. To ensure a fair self-comparison, certain hyperparameters were fixed within each approach, including warm-up epochs, learning rate schedules, and optimizer configurations. Additionally, to maintain consistency across all approaches, the number of epochs, batch size, input dimension, backbone, and data transformations were kept identical. Since all these variables were controlled, any improvement in performance between the baseline and the semantic pairs model can be attributed solely to the incorporation of semantic information. Subsequently, we compute the performance difference (Δ) between approaches to identify which method most effectively leverages semantic information to learn high-quality representations. The approach exhibiting the greatest difference (Δ) between the baseline and the semantic pairs model is considered to learn more effective representations from semantic pairs.

4 Experiments

Training Setup. We pre-trained several SOTA approaches on two distinct datasets: the Augmented Pairs Dataset and the Semantic Pairs Dataset. The experiments utilized a ResNet50 backbone, with a batch size of 256, and a training duration of 200 epochs. All remaining hyperparameters and configurations corresponding to each method were adopted from the original publications.

Evaluation. Following the training phase, two downstream tasks are employed for evaluation: transfer learning and object detection. In transfer learning, the backbone is frozen, and a classification head is appended to it. Transfer learning experiments are conducted on three benchmark datasets: STL-10 (Coates et al., 2011), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009). In addition, object detection fine-tuning is performed on PASCAL VOC (Everingham et al., 2010), where the backbone and detection head are jointly trained. These evaluations aim to determine whether the incorporation of semantic information enhances the generalization capability of SSL approaches and, if so, which approach performs best.

SOTA Methods on Semantic and Augmented Pairs. In our experimental evaluation, we trained a set of eight SOTA methods across two datasets: one comprising semantically related pairs and another consisting of augmented pairs. The selection included both contrastive and non-contrastive representation learning approaches. While individual methods within each approach differ in their implementation specifics, they generally operate based on a shared underlying principle, resulting in comparable overall learning behavior. Accordingly, our analysis is conducted at the approach level rather than focusing on individual methods. Specifically, SimCLR (Chen et al., 2020a) was selected as a representative of contrastive learning. For non-contrastive learning, three representative methods were chosen to capture different subfamilies: VICReg (Bardes et al., 2021), representing the information maximization family (e.g., VICReg and Barlow Twins); BYOL (Grill et al., 2020), exemplifying self-distillation frameworks that employ architectural constraints to prevent representational collapse (e.g., BYOL and SimSiam); and DINO (Caron et al., 2021), a distinct self-distillation approach incorporating centering and sharpening mechanisms to mitigate collapse. The comparative transfer learning performance of these approaches is presented in Figure 4.

4.1 Transfer Learning

Figure 4 shows that models pretrained on semantic pairs consistently outperform those pretrained on augmented pairs when evaluated on unseen data. This suggests that the semantic pairs dataset provides richer semantic information, thereby enhancing the model's ability to generalize to unseen datasets. These results align with established findings in transfer learning, which demonstrate that models pretrained on semantically rich data transfer more effectively across datasets and achieve superior generalisation (Yosinski et al., 2014). Figure 4 also shows that the contrastive learning approach consistently outperforms competing approaches in learning useful representations from semantic pairs. Among the evaluated methods, SimCLR achieves the highest performance across all downstream datasets. For instance, SimCLR (SP), which is pretrained on a semantic pairs dataset, achieves a 3.83% improvement in transfer learning to the STL-10 dataset compared to its counterpart, augmentation-based SimCLR (AP).

To evaluate which approach learns semantic features more effectively from semantic pairs, we compare the performance differences (Δ) between approaches rather than their overall accuracy values. For each approach, such as SimCLR, the performance difference (Δ) is obtained by subtracting the performance of the model pretrained on augmented pairs from that of the model pretrained on semantic pairs. This provides a more accurate indication of which approach learns semantic features more effectively from the input data. For example, consider the case where SimCLR (SP), trained on semantic pairs, achieves 85.59% accuracy when transferred to the STL-10 dataset, while VICReg (SP), trained on semantic pairs, reports 87%. A direct comparison of these results does not necessarily imply that VICReg learns semantic features more effectively; the difference may be due to VICReg's method performing better than SimCLR, rather than being attributed to semantic pairs. To address this issue, we compare the relative improvements of each method. For SimCLR, accuracy increases from 81.76% on the augmented pairs dataset to 85.59% on the semantic pairs dataset, corresponding to an improvement of +3.83%. In contrast, VICReg increases from 85.5% on augmented pairs to 87.0% on semantic pairs, yielding a smaller improvement of +1.5%. Although VICReg (SP) achieves higher absolute accuracy, SimCLR demonstrates a larger improvement when trained on semantic pairs, indicating that it benefits more from semantic supervision. In other words, the VICReg might be better in absolute accuracy, but SimCLR is better in learning from semantic pairs because the (Δ) between SimCLR (AP) and SimCLR (SP) is greater than the (Δ) of VICReg. This comparison enables researchers to identify the most suitable SSL framework for leveraging semantic pairs, rather than selecting a framework arbitrarily. Additionally, it motivates further investigation into why contrastive learning methods (e.g., SimCLR) capture semantic features more effectively than non-contrastive approaches (e.g., VICReg)

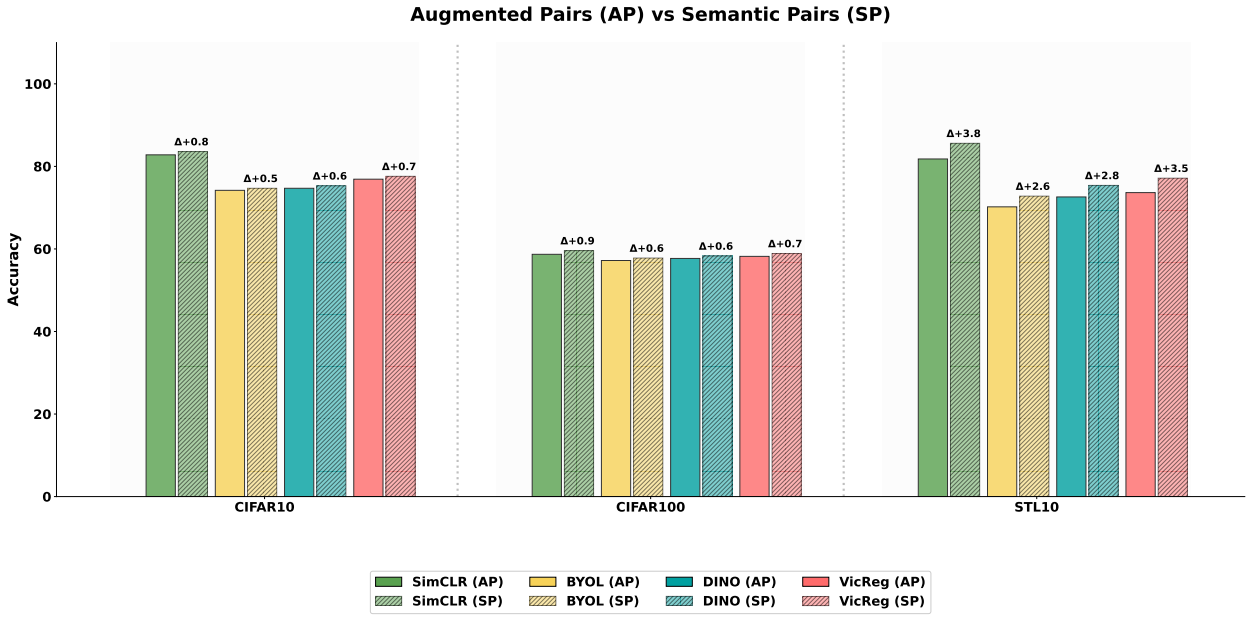

Figure 4: Transfer learning performance of SOTA approaches pretrained on Semantic Pairs (SP, striped) and Augmented Pairs (AP, solid), evaluated on CIFAR10, CIFAR100, and STL10. (SP) Pretraining consistently improves downstream accuracy over(AP). (Δ) values denote (SP-AP) differences.

from semantic pairs. Understanding this dynamic will help optimize the exploitation of semantic pairs within non-contrastive models, thereby enhancing representation learning and generalization across SSL approaches.

Table 1: Comparison of Transfer Learning Performance on STL-10 after 800 epochs of pre-training for SimCLR with Augmented Pairs (AP) and Semantic Pairs (SP).

Approach	**200 Epochs**	**400 Epochs**	**800 Epochs**
SimCLR (SP)	85.59%	86.50%	86.56%
SimCLR (AP)	81.76%	82.23%	82.41%
Improvement	+3.83%	+4.27%	+3.75%

SimCLR is selected for further analysis, as it demonstrates superior performance among the evaluated approaches leveraging semantic pairs. To assess whether semantic pairs consistently enhance model generalisation, the training schedule is extended to 800 epochs. As presented in Table 1, SimCLR (SP) consistently outperforms SimCLR (AP) across all epochs. Both models converge after 800 epochs; however, SimCLR (SP) maintains a performance margin of 3.75% over SimCLR (AP).

4.2 Object Detection

This evaluation assesses the generalization capability of representations learned by semantic pairs, with a particular focus on their effectiveness beyond standard classification tasks. The model was fine-tuned for object detection using the PASCAL VOC dataset. Following the experimental protocol of MoCo-v2(Chen et al., 2020b), fine-tuning was conducted on the combined VOC07+12 trainval set employing a Faster R-CNN architecture with a R50-C4 backbone. Evaluation was performed on the VOC07 test set. The fine-tuning process comprised 24K iterations, corresponding to approximately 23 epochs. Table 2 presents a comparative analysis of performance between SimCLR pretrained on semantic pairs and SimCLR pretrained on augmented pairs. The evaluation metrics considered include AP_{50}, AP, and AP_{75} . The results indicate that SimCLR (SP) pretrained on semantic pairs consistently outperforms the model pretrained on augmented

pairs across all metrics, thereby suggesting that pretraining on semantic pairs substantially improves both object localization and classification performance.

Table 2: Results (Average Precision) for PASCAL VOC object detection using Faster R-CNN with ResNet50-C4.

Approach	AP_{50}	AP	AP_{75}
SimCLR (SP)	75.02%	50.30%	55.22%
SimCLR (AP)	73.82%	48.9%	53.72%

SimCLR (SP) achieves the highest gain in AP_{75}, with 55.22 % versus 53.72 %. This metric, which requires a (75%) overlap between predicted and true bounding boxes, highlights the model's ability to perform precise localisation. The model also improves by 1.2 % in AP_{50}, and 1.4 % in overall AP. These results indicate that pretraining on semantic pairs exposes the model to objects across diverse contexts, enabling it to focus more effectively on object features rather than background cues. This exposure fosters greater invariance to background variations, thereby improving both recognition robustness and localization precision (Zhao et al., 2022). Consequently, semantic pair pretraining enhances representation quality and strengthens the model's ability to generalise to downstream object detection tasks more effectively than transformation-based augmentation alone.

5 Ablation Studies

In the ablation studies, a series of experiments is conducted to further investigate the impact of semantic pairs on SSL. First, data transformation variants are removed from the training pipeline, and models are trained on both datasets to determine if semantic pairs alone can impose invariance and help SSL models learn more robust semantic feature representations. Next, the ResNet-50 backbone is replaced with a Vision Transformer (ViT), followed by transfer learning and fine-tuning, to assess whether the generalization improvement from semantic pairs is consistent across different architectures and downstream tasks. Subsequently, a random crop test is introduced to simulate occlusion scenarios, presenting only partial object views instead of full ones obtained through center cropping. This step examines whether incorporating semantic information improves the model's robustness to occlusion. Finally, models are trained on datasets of varying sizes to investigate how performance responds to different quantities of semantic pairs. This design helps analyze whether smaller datasets limit generalization and lead to learning patterns similar to models pretrained with augmented pairs, offering insights into semantic pair sensitivity and data efficiency.

5.1 Evaluating Model Performance with Removed Transformations

In these experiments, we systematically ablate variant transformation techniques from the pipeline and subsequently retrain the models on both datasets. By doing this, we can see if the semantic pairs impose invariance by placing the object in different contexts, which helps the model learn a generalized representation for the input object.

As illustrated in Figure 5, the model pretrained on the augmented dataset exhibits a substantial decline in performance, reaching 24.25%, when all transformation techniques are removed from the pipeline except random cropping (Only_C). This degradation occurs because the model becomes increasingly sensitive to illumination, background colour, and superficial object features such as logos and decorations. Conversely, the model pretrained on the semantic pairs dataset demonstrates a notably higher robustness, achieving 64.23% under the same conditions. This result empirically indicates that semantic pairs induce invariances during representation learning, allowing the model to learn a more invariant representation for the input data. In essence, semantic pairs facilitate more effective representation learning by exposing the model to diverse contextual variations, thereby enhancing its ability to generalize beyond predefined transformations.

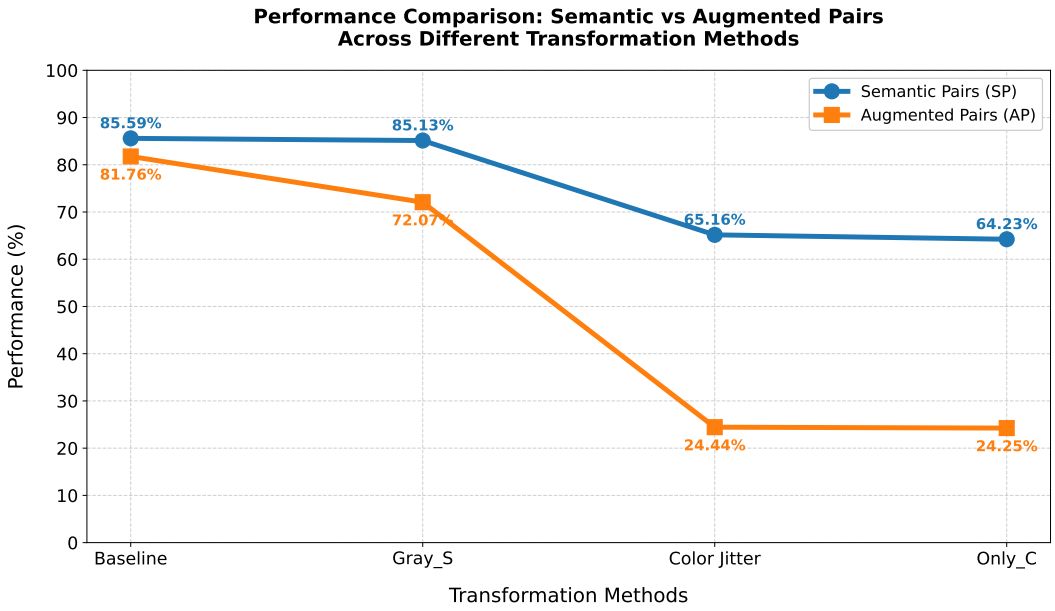

Figure 5: Performance comparison of SimCLR (SP) and SimCLR (AP) pre-trained for 200 epochs with transformation ablations: (1) Gray_s (grayscale removed), (2) color_jitter (both color jitter and grayscale removed), and (3) only_c (only random crop retained).

5.2 Generalization Across Architectures and Tasks

In this experiment, we conduct further assessment for the robustness of the semantic pairs in supporting SSL generalization. This study extends the analysis to encompass both an alternative backbone architecture and a distinct downstream task. Specifically, the ViT-S/8 model is pretrained on both datasets, followed by transfer learning via fine-tuning to assess downstream performance. Among the methods examined in this study, only the DINO framework natively supports the ViT architecture. Accordingly, DINO is trained with the ViT backbone on both datasets. The training protocol utilizes a batch size of 256 and 300 epochs. In this experiment, we employed only two global views per sample, while all remaining hyperparameters adhere to those defined in the original DINO setup. The results presented in Table 3 indicate that DINO (SP) with semantic pairings consistently outperforms DINO (AP) with augmented pairings across all evaluations. These outcomes provide robust evidence that semantic pairs enhance the generalization capability of SSL irrespective of the architectural framework or downstream task considered.

Table 3: Comparison of DINO (ViT-S/8) fine-tuning results after 300 epochs of pretraining on semantic pairs (SP) and augmented pairs (AP).

Approach	**CIFAR-10**	**CIFAR-100**	**STL-10**
DINO (SP)	81.8%	65.3%	82.1%
DINO (AP)	81.1%	64.5%	79.2%

5.3 Centre and Random Crop Test

In this experiment, both models are evaluated using randomly cropped images from the STL-10 dataset, rather than the centred object regions typically used. This testing approach presents only a portion of the object, positioned variably within the image, thereby increasing the difficulty of accurately predicting the correct label. The experiment is designed to simulate real-world conditions, where objects are not always centrally located or fully visible within an image. Superior performance under these conditions indicates

that the model exhibits stronger occlusion invariance, demonstrating its ability to infer the correct semantic label when objects only partially appear.

Table 4: Comparison of Transfer Learning Performance on STL-10 after 200 epochs of pre-training for SimCLR with Augmented Pairs (AP) and Semantic Pairs (SP).

Approach	**Center Crop**	**Random Crop**
SimCLR (SP)	85.59%	79.29%
SimCLR (AP)	81.76%	74.26%

As shown in Table 4, the performance of the model pretrained with augmented pairs SimCLR (AP) decreases by 7.5%, while that of the semantic pairs SimCLR (SP) declines by 6.3%. This indicates that incorporating semantic pairs promotes occlusion invariance in self-supervised learning, enabling the model to perform more effectively when only a portion of the image is visible, thereby enhancing its generalization capability.

5.3.1 Impact of Training Set Size on Generalisation

In this experiment, we evaluate the models by training them on datasets of varying sizes (50, 100, and 157 images per class) for 200 epochs. After training, we apply transfer learning to the STL-10 dataset to assess its impact on generalisation performance. This design enables a systematic investigation of the relationship between the semantic pairs dataset size and the models' ability to generalise.

Table 5: Comparison of Transfer Learning Performance on STL-10 after 200 epochs of pre-training for SimCLR with Augmented Pairs (AP) and Semantic Pairs (SP).

Approach	**50 Images**	**100 Images**	**157 Images**
SimCLR (SP)	81.68%	83.01%	85.59%
SimCLR (AP)	77.48%	78.97%	81.76%
Improvement	+4.20%	+4.04%	+3.83%

While Table 5 indicates that SimCLR (SP), pretrained on semantic pairs, consistently outperforms SimCLR (AP), which relies solely on the transformation pipeline, across different dataset sizes, an important trend can be observed. Specifically, when the number of training examples is increased to 157 per class, the performance gap between the two models decreases to 3.83%. In contrast, when the number of examples is reduced to 50 per class, the gap widens to 4.20%. This widening occurs because when the number of samples is reduced to 50 per class, the traditional transformation pipeline fails to capture sufficient intra-class variability, leading to weaker invariant representations and, consequently, reduced generalization to unseen data. In other words, increasing the number of images in the augmented pairs dataset allows the model to be exposed to diverse data, which allows it to be more invariant and improve its generalization closer to semantic pairs on the unseen dataset. In contrast, semantic pairs, which comprise images that present the same object in varying contexts, inherently provide greater data diversity. This richer diversity enables the model to become more invariant to intra-class variations, leading to more robust and generalizable results, even in scenarios with limited training data. This is evident in Table 5, where training SimCLR (SP) on 50 images per class yields performance comparable to that of SimCLR (AP) trained on 157 images per class. This observation aligns with the hypothesis of this study, which posits that leveraging semantic positive pairs can enhance representation learning by exposing models to meaningful variations beyond the limited variations of the transformation pipeline.

6 Conclusion

This study examined the role of semantic positive pairs in self-supervised visual representation learning. To isolate their effect from standard same-instance augmented positives and false neighbors, we constructed two matched ImageNet-1K subsets: an augmented-pair baseline and a manually curated semantic-pair dataset with the same class composition and training-pair count. This controlled design enabled a direct comparison between models trained with augmented positive pairs and those trained with semantic positive pairs.

Across transfer learning and object detection evaluations, models pretrained with semantic pairs consistently outperformed their augmented-pair counterparts. These results suggest that semantic pairs provide useful invariance beyond that induced by conventional transformation pipelines, helping models learn representations that transfer more effectively to unseen datasets. The ablation studies further support this interpretation: semantic-pair pretraining remained more robust when transformations were removed, improved performance under random-crop evaluation, generalized across a ViT-based DINO setting, and retained advantages across different dataset sizes.

Among the evaluated SSL methods, contrastive learning benefited most strongly from semantic pairs, with SimCLR showing the largest relative improvement over its augmented-pair baseline. This finding suggests that the interaction between semantic positive pairs and the training objective is important: not all SSL frameworks exploit semantic relationships equally effectively. Understanding why contrastive objectives appear to benefit more from semantic-pair information remains an important direction for future work.

Overall, the results indicate that semantic positive pairs can serve as a valuable complement to transformation-based augmentation in self-supervised learning. By exposing models to same-class objects across diverse contexts, semantic pairs encourage representations that are less dependent on nuisance factors such as background, illumination, occlusion, and superficial visual details. Future work should investigate larger-scale semantic-pair construction, automated pair selection, stronger controls for the number of unique source images, and objective functions designed specifically to exploit semantic positive relationships.

Acknowledgments

We would like to thank the University of Aberdeen's High Performance Computing facility for enabling this work.

References

Mohammad Alkhalefi, Georgios Leontidis, and Mingjun Zhong. Semantic positive pairs for enhancing visual representation learning of instance discrimination methods. *Transactions on Machine Learning Research*, 2024a.

Mohammad Alkhalefi, Georgios Leontidis, and Mingjun Zhong. Leoclr: Leveraging original images for contrastive learning of visual representations. *Transactions on Machine Learning Research*, 2024b.

Joonsun Auh, Changsik Cho, and Seon-Tae Kim. Contrastive learning for reducing false negatives with global and local views in augmented data. In *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, 2023. doi: 10.1109/ASYU58738.2023.10296635.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Shasvat Desai, Debasmita Ghose, and Deep Chakraborty. A survey on data curation for visual contrastive learning: Why crafting effective positive and negative pairs matters. *arXiv preprint arXiv:2502.08134*, 2025.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6928–6938, 2020.

Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Perez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6840, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2785–2795, 2022.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10326–10335, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ya Le, Xuan Yang, et al. Tiny imagenet challenge. 2015.

Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *Advances in Neural Information Processing Systems*, 36:42775–42798, 2023.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6707–6717, 2020.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022.

Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. Conditional negative sampling for contrastive learning of visual representations. *arXiv preprint arXiv:2010.02037*, 2020.

Jianqiang Xiao, Weiwen Guo, Junfeng Liu, and Mengze Li. Generalization gap in data augmentation: Insights from illumination. In *International Conference on Pattern Recognition*, pp. 190–205. Springer, 2024.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, and Yiyu Shi. Contrastive learning with synthetic positives. In *European Conference on Computer Vision*, pp. 430–447. Springer, 2024.

Jianping Zhang, Lei Yang, Seyed Mahmoud Sajjadi Mohammadabadi, and Feng Yan. A survey on self-supervised learning: Recent advances and open problems. *Neurocomputing*, 655:131409, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.131409. URL `https://www.sciencedirect.com/science/article/pii/S0925231225020818`.

Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16580–16589, 2022.

Bing Zhao, Jun Li, and Hong Zhu. Codo: Contrastive learning with downstream background invariance for detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4196–4201, June 2022.

A

Dataset Details

We introduce a dataset of curated semantic pairs derived from ImageNet-1K, comprising **187** carefully selected classes (from the original 1,000) designed to advance representation learning through semantically meaningful instance pairs. Our class selection was strategically aligned with standard benchmarks, such as STL-10, to ensure optimal transfer learning performance.

A.1 Key Features

- Maintains original ImageNet-1K image resolutions
- Each instance is **paired semantically** with another instance from the same class
- These curated pairs serve as **ground-truth supervision** for learning tasks
- The pairing strategy is designed to:
  1. Capture meaningful **semantic similarities** (e.g., invariant features across different object poses or lighting conditions)
  2. Minimize attention to **irrelevant variations** (e.g., background clutter or texture biases)

A.2 Dataset Composition

- **187 classes** covering diverse categories (animals, objects, landscapes, etc.)
- **157 semantic pairs** per class (29,359 pairs total)
- Classes intentionally selected to mirror standard benchmark datasets

A.3 Construction Details

- Required **6 months** of full-time manual curation (8 hours/day).
- Ensures **high-quality labels** through human verification.

A.4 Dataset Classes

This subsection illustrates the 187 classes selected from ImageNet-1K.

- **Animals**:
  African_grey, albatross, American_alligator, American_coot, American_egret, American_lobster, Afghan_hound, Arabian_camel, bald_eagle, basset, beagle, bee_eater, bittern, black_grouse, black_stork, black_swan, black_widow, Blenheim_spaniel, brambling, brown_bear, bullfrog, bulbul, bustard, centipede, chickadee, Chihuahua, chimpanzee, cock, cockroach, cougar, crane, drake, dowitcher, Egyptian_cat, European_fire_salamander, flamingo, gazelle, goldfinch, goldfish, goose, great_grey_owl, hen, hornbill, house_finch, hummingbird, indigo_bunting, jacamar, Japanese_spaniel, jay, king_penguin, kite, lemon_shark, limpkin, little_blue_heron, lion, macaw, magpie, Maltese_dog, monarch, oystercatcher, peacock, pelican, Pekinese, ptarmigan, quail, red-breasted_merganser, red-backed_sandpiper, redshank, Rhodesian_ridgeback, robin, ruffed_grouse, Shih-Tzu, snail, spoonbill, sulphur_butterfly, sulphur-crested_cockatoo, tabby, tadpole, tailed_frog, tarantula, tiger_shark, toucan, toy_terrier, vulture, water_ouzel, white_stork, birdhouse.
- **Plants & Fungi**:
  banana, bell_pepper, corn, daisy, Granny_Smith, lemon, mushroom, orange, pomegranate, rapeseed, strawberry, yellow_lady's_slipper

- **Vehicles**:
  aircraft_carrier, airliner, airship, ambulance, amphibian, balloon, bullet_train, cab, convertible, electric_locomotive, fire_engine, forklift, freight_car, garbage_truck, go-kart, golfcart, jeep, limousine, minivan, Model_T, moving_van, passenger_car, pickup, police_van, racer, recreational_vehicle, schooner, snowplow, space_shuttle, sports_car, streetcar, trailer_truck, tractor, tow_truck, thresher, trolleybus, unicycle, yawl
- **Architectural & Natural Landmarks**:
  alp, cliff, geyser, lakeside, promontory, sandbar, seashore, triumph_arch, valley, viaduct, volcano
- **Household Items**:
  computer_keyboard, cup, dining_table, eggnog, espresso, hot_pot, lawn_mower, plate, red_wine, table_lamp, teapot, television, throne, torch, tray, typewriter_keyboard, wall_clock, wardrobe, washer, water_bottle, water_jug, wine_bottle, wok
- **Human Activities**:
  groom, scuba_diver, tennis_ball, volleyball

Figure 6: Examples from the semantic pairs dataset

Figure 7: Additional examples from the semantic pairs dataset