One Leaf Reveals the Season: Occlusion-Based Contrastive Learning with Semantic-Aware Views for Efficient Visual Representation

Xiaoyu Yang [1] **Lijian Xu** [1 2] **Hongsheng Li** [3] **Shaoting Zhang** [2 4]

Abstract

This paper proposes a scalable and straightforward pre-training paradigm for efficient visual conceptual representation called occluded image contrastive learning (OCL). Our OCL approach is simple: we randomly mask patches to generate different views within an image and contrast them among a mini-batch of images. The core idea behind OCL consists of two designs. First, masked tokens have the potential to significantly diminish the conceptual redundancy inherent in images, and create distinct views with substantial fine-grained differences on the semantic concept level instead of the instance level. Second, contrastive learning is adept at extracting high-level semantic conceptual features during the pre-training, circumventing the high-frequency interference and additional costs associated with image reconstruction. Importantly, OCL learns highly semantic conceptual representations efficiently without relying on hand-crafted data augmentations or additional auxiliary modules. Empirically, OCL demonstrates high scalability with Vision Transformers, as the ViT-L/16 can complete pre-training in 133 hours using only 4 A100 GPUs, achieving 85.8% accuracy in downstream fine-tuning tasks. Code is available at https://github.com/XiaoyuYoung/OCL.

Xiaoyu Yang conducted this work during an internship at Shenzhen University of Advanced Technology. [1]Shenzhen University of Advanced Technology, Shenzhen, China [2] Centre for Perceptual and Interactive Intelligence, the Chinese University of Hong Kong, China [3]Department of Electronic Engineering, the Chinese University of Hong Kong, Hongkong, China [4]SenseTime Research, Shanghai, China. Correspondence to: Lijian Xu <bmexlj@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Self-supervised learning (SSL) is a key approach for building world models, especially for pre-training vision models (Chen et al., 2020a; He et al., 2022; Bao et al., 2021; Radford et al., 2021; Wang et al., 2023; Yang et al., 2024b). Its strength lies in learning versatile visual representations without relying on human annotations. Currently, the two main paradigms in visual SSL are Masked Image Modeling (MIM) (He et al., 2022; Wang et al., 2023; Kong et al., 2023; Zhang et al., 2022; Gupta et al., 2023) and Contrastive Learning (CL) (Chen et al., 2020a). Both have shown strong scalability, particularly for Vision Transformers (ViTs) (Dosovitskiy et al., 2021).

Despite the success of existing methods, both MIM and CL struggle to achieve efficient visual representation. MIM focuses heavily on pixel-level reconstruction, which often prioritizes local details over high-level semantic concepts. Similarly, CL suffers from conceptual redundancy, where transformed images may lack meaningful differences. This raises a critical question: *How can we bridge the gap between efficient visual representation and effective conceptual pre-training, overcoming the limitations of current MIM and CL approaches?*

To address this question, we first revisit the two main pre-training paradigms: MIM and CL. MIM learns visual representations by reconstructing masked image patches (see Fig.1b). Notable examples include BEiT (Bao et al., 2021) and MAE (He et al., 2022). MAE highlights that images contain significant semantic redundancy, meaning only a basic understanding of objects and scenes is needed to predict missing patches from their surroundings. However, pixel-level reconstruction is too detailed for pre-training vision models. It focuses excessively on high-frequency details and local features, conflicting with the goal of pre-training: learning high-level semantic concepts. While this detailed task helps models learn visual representations, it comes at the cost of pre-training efficiency.

In CL, the core idea is simple: maximize agreement between different views of the same image (see Fig.1a). Popular methods like SimCLR (Chen et al., 2020a;b), MoCo v3 (Chen et al.), and DINO (Caron et al., 2021) use com-

plex pre-processing and auxiliary networks to create distinct views of an image. The main challenge lies in optimizing the agreement between these views. For CL to work well, distinct views are essential. However, this is difficult because images often have conceptual redundancy. Additionally, the dependency on large-scale batches to generate sufficiently diverse negative samples imposes stringent computational constraints, notably escalating computational and time consumption overhead during training.

Driven by this analysis, we found that these two paradigms can complement each other: masked tokens have the potential to significantly diminish the conceptual redundancy inherent in images, whereas contrastive learning is adept at extracting high-level semantic features during the pre-training phase. Thus, we present a novel and straightforward paradigm for self-supervised visual representation learning: occluded image contrastive learning (OCL). OCL tackles the above issues systematically: I) Masked image tokens offer diverse views of a single image with substantial fine-grained conceptual differences. II) Contrastive learning enables pre-training to concentrate exclusively on the high-level semantic information contained within images while disregarding high-frequency redundancies. III) The proposed paradigm obviates the need for auxiliary modules and expedites the efficient extraction of model features.

OCL has a particularly simple and straightforward workflow, as presented in Fig.1c. Here is how it works: Firstly, we mask a batch of images with a high rate, dividing visible patches within one image into two non-overlapping groups. In succession, the pre-train model extracts the features of these two groups of batch image tokens, respectively. Subsequently, contrastive learning is employed to predict the correct pairings for a batch of visible image tokens. Positive samples are different visible tokens in the same image, while negative samples are from different images of the mini-batch. Finally, inspired by T-distributed classifier (Yang et al.; 2024a), we use the T-distributed spherical loss to constrain the inter-class margins in the pre-training. Comprehensive experiments demonstrate the scalability and efficacy of our approaches, where ViT-L/16 can complete pre-training in 133 hours using only 4 A100 GPUs and attain an 85.8% top-1 accuracy in fine-tuning classification. In particular, our model stands out from other pre-training methods as it operates without the need for auxiliary modules or hand-crafted data augmentation to generate diverse views.

In summary, our paper mainly makes the following contributions:

1. We endeavour to explore an alternative of using masked images to create diverse views with fine-grained conceptual differences for contrastive learning. By forgoing the conventional approach of employing instance-level hand-crafted data augmentation to generate distinct views, OCL diminishes the conceptual redundancy inherent in images and improves efficiency.

2. Our approach eschews the reconstruction of masked images in favour of leveraging contrastive loss to steer the entire model. Independently of additional auxiliary modules, OCL is adept at extracting high-level semantic concept features from images more efficiently.

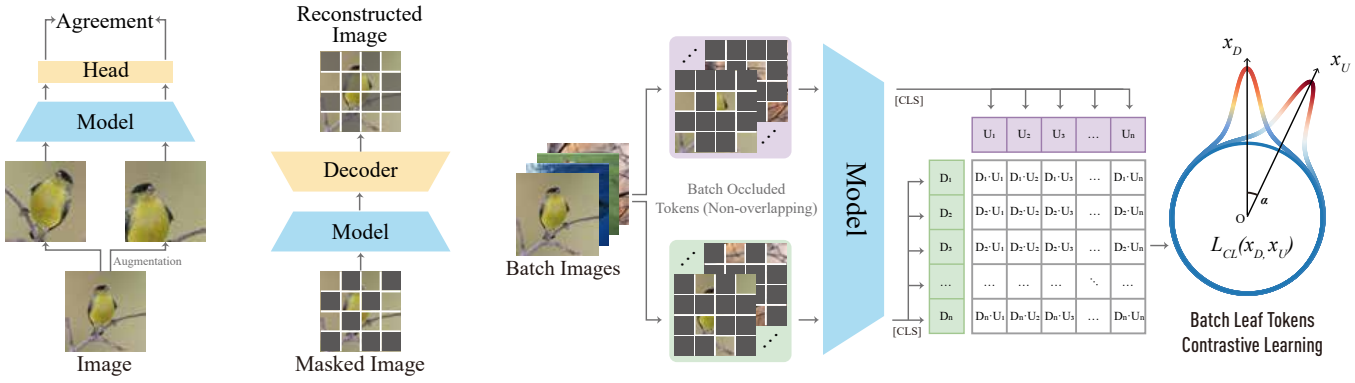
3. Extensive experiments are conducted to verify the efficiency and scaling capability of our method. ViT-L/16 can complete pre-training in 133 hours using only 4 A100 GPUs with 85.8% accuracy in fine-tuning. Additionally, we have structured ablation experiments to delve into the implications of different configurations within OCL, with a particular focus on the need of the MLP head in contrastive learning.

2. Methodology

2.1. Architecture

With an input image $x_i \in \mathbb{R}^{H\times W\times C}$, it is reshaped into a sequence of 2D patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H, W) denotes the original image resolution, C is the number of channels, P represents the patch size, and $N = HW/P^2$ indicates the number of patches. Subsequently, a linear projection is employed on x_p to transform it into D dimensions, yielding patch embeddings $x \in \mathbb{R}^{N\times D}$. Following the MoCo v3 (Chen et al.), the linear projection is initialized using the Xavier uniform method and remains fixed throughout pre-training to mitigate potential instability in ViT caused by the large batch size. Thereafter, fixed position embeddings $p \in \mathbb{R}^{(N+1)\times D}$ are incorporated into the patch embeddings to preserve positional information, employing sinusoidal positional encoding.

After random masking patches, visible patches that retain the original image position information are divided into two non-overleaping groups: $x_U \in \mathbb{R}^{n\times D}$ and $x_L \in \mathbb{R}^{n\times D}$, where n denotes the number of visible patches. Each group adds an independent **[CLS]** token $x_{cls} \in \mathbb{R}^D$ to aggregate the information of each group. Moreover, **[CLS]** tokens of each group will add the position embeddings p_0. In succession, ViT (Dosovitskiy et al., 2021) is utilized as our encoder. $z_U = [x^U_{cls}; x_U] \oplus p$ and $z_L = [x^L_{cls}; x_L] \oplus p$ are the input of pre-training ViT, where $\oplus$ denotes element-wise plus, and $f(\cdot)$ represents the ViT. Consequently, image feature tokens extracted by ViT are symbolized as $y = f(z)$, where $y \in \mathbb{R}^{(n+1)\times D}$. Furthermore, we only preserve **[CLS]** tokens $y_{cls} \in \mathbf{R}^D$ of both groups for contrastive learning, as it aggregates the high-level semantic information of each image. It is noteworthy that, unlike existing methodologies, we abstain from employing auxiliary modules. This allows our model to more easily extract image features and reduce



(a) CL (b) MIM (c) Our occluded image contrastive learning

Figure 1. **Comparison between different pre-training paradigms.** The Model in blue is the pre-training model, and the orange modules indicate auxiliary modules. (a) Contrastive Learning (CL) endeavours to maximize the agreement between different views of an image. (b) Masked Image Modeling (MIM) aims to restore masked image patches. (c) Our occluded image contrastive learning: Through non-overlapping occluding, distinct tokens within an image are categorized as intraclass, while across-image tokens within a batch are viewed as interclass. Our objective is to enhance intraclass compactness and interclass separability through a contrastive learning approach. Just as a single leaf can tell the coming of autumn, we believe that a small area of the image contains the majority of the meaning of the entire image.

the consumption of computing resources.

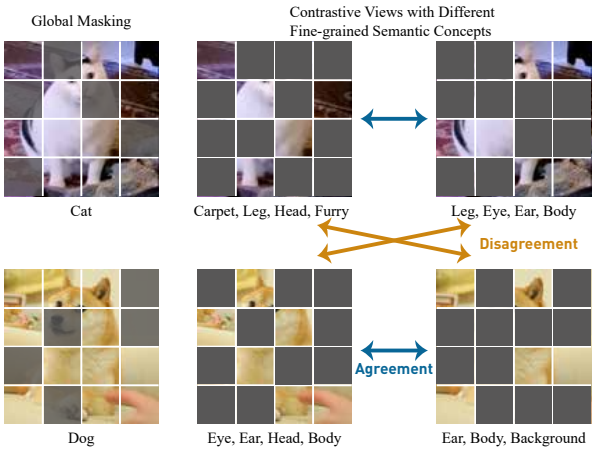


Figure 2. **A toy example of masked images for conceptual contrastive learning.** The low global masking ratio aids the model in capturing comprehensive information from the image and understanding the interconnectedness of various concepts within a mini-batch. Beyond that, each contrastive branch has a higher masking ratio, generating diverse views with different semantic concepts for contrastive learning and diminishing conceptual redundancy within the image.

2.2. Occluded Image for Conceptual Pre-training

Despite the strides made in instance-wise contrastive learning (Chen et al., 2020a; Caron et al., 2021), the differences between various views primarily manifest at the pixel level, instead of semantic concept disparities. In this context, a patch is abstracted as a concept containing fine-grained semantics. The concepts within an image exhibit distinct conceptual characteristics, yet they are all interconnected with the overall meaning of the image, albeit to varying extents. Thus, masked images are used to create views that contain semantic distinctions of concepts, and mitigate the conceptual redundancy present in images, as presented in Fig. 2.

We apply the masking strategy of random sampling, the same as the MAE, which samples random patches without replacement following the uniform distribution. Beyond that, a low masking ratio (e.g. 30%) is implemented overall, but leads to a higher masking ratio (e.g. 65%) for individual contrastive branches, as illustrated in Fig. 2. The low global masking rate helps the model capture the overall image structure and the relationships between different concepts within a mini-batch. During each forward pass, the model only sees a small portion of the visible tokens. This creates diverse views with fine-grained semantic differences, making the task more challenging and reducing the risk of the model relying on simple patterns or redundant information. The high masking ratio for each contrastive branch ensures that the task cannot be solved by simply extrapolating from basic image transformations. This reduces conceptual redundancy and forces the model to focus on localized features. Additionally, by training on only a few parts of the image, the masking strategy also improves scalability and efficiency, making it easier to handle the large batch sizes required for contrastive learning.

To divide visible patches into two separate, non-overlapping groups, we use a random sampling strategy similar to the masking approach described earlier. This helps avoid poten-

tial biases and ensures that the central positions of the visible patches are consistent across both groups. Meanwhile, non-overlapping patches present a challenging scenario, impeding the model from relying solely on analogous patches for inference.

2.3. Efficient Contrastive Learning

We randomly select a mini-batch of B instances and establish the contrastive prediction task on the visible token pairs extracted from this mini-batch, yielding a total of $2 \times B$ data points. Feature tokens from different groups within the same image are treated as positive pairs, while tokens from different images in the same mini-batch are treated as negative pairs. We concurrently maximise the similarity of B positive pairs while minimizing the similarity of $B^2 - B$ negative examples to drive the network.

In terms of similarity computation, we introduce T-distributed spherical (T-SP) metric (Yang et al.; Kobayashi) to significantly promote the intraclass compactness and interclass separability of features. Given **[CLS]** tokens $y_i^U \in \mathbf{R}^D$ and $y_j^L \in \mathbf{R}^D$ of both non-overlapping groups, the cosine distances between y_i^U and y_j^L are:

$$cos_{LU}(y_i^U, y_j^L) = \frac{{y_i^U}^T y_j^L}{|y_i^U||y_j^L|} \tag{1}$$

and the T-SP similarity is defined as follows:

$$\text{sim}_{tsp}(y_i^U, y_j^L) = 0.5 \times \frac{1 + cos_{LU}}{1 + (1 - cos_{LU}) * \kappa} \tag{2}$$

where $\kappa \geq 0$ denotes the concentration hyperparameter of T-SP metric. As κ decreases, the similarity function becomes more condensed, where only two tokens in close proximity are deemed positive examples. Besides, we add a trainable temputare parameter τ to effectively scale the different samples. Thus, the loss function for a positive pair is:

$$\begin{aligned} &\mathcal{L}(y_i^U, y_j^L) \\ &= -\log \frac{\exp\left(\text{sim}_{tsp}(y_i^U, y_j^L) \times \tau\right)}{\sum_{k=1}^{2B} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}_{tsp}(y_i^U, y_k^L) \times \tau\right)} \end{aligned} \tag{3}$$

where $\mathbb{1}_{[k \neq j]} \in \{0, 1\}$ is an indicator function evaluating to 1 if $k \neq i$. Finally, inspired by CLIP (Radford et al., 2021), we optimize a symmetric loss over these similarity scores within a mini-batch exhibited in Algorithm 1.

2.4. Simple Implementation

The implementation of our OCL pre-training is efficient and involves minimal specialized operations. As pseudo code depicted in Algorithm 1, we make only minor modifications based on the MAE code, mainly involving the process subsequent to the acquisition of image feature embeddings from the encoder. First, we randomly mask a subset of embedded patch tokens with a low masking ratio. In succession, listed tokens are shuffled randomly and divided into two non-overlapping groups. Following MAE (He et al., 2022), within positional and [CLS] embeddings, lists of tokens are encoded by the ViT. It is noteworthy that we obtain the encoded [CLS] token from the ViT directly, without the incorporation of additional auxiliary modules, even a linear head or lightweight decoder.

Algorithm 1 Pytorch-like pseudo-code for the core of an implement of OCL

```
# x[B,N,D] - patch embeddings

# mask image and split into two
non-overleaping groups
# x [B×2, n, L]
x = masking(x, ratio = 0.3)
# extract [CLS] token of both groups
using pre-training model
x = model(x) # [B×2, 1, L]
x = x.reshape(-1, 2, x.shape[-1])
# L2 normalize the [CLS] token of each
group
m, n = x[:,0], x[:,1] # [B, L]
m = m/m.norm(dim=-1,keepdim=True)
n = n/n.norm(dim=-1,keepdim=True)
# compute the scaled pairwise T-SP
similarities
sim_mn = compute_tSP(m @ n.T)
sim_nm = compute_tSP(n @ m.T)
# symmetric loss function
labels = torch.arange(B)
loss=(F.cross_entropy(sim_mn,labels)
+F.cross_entropy(sim_nm,labels))/2
```

Subsequently, the obtained [CLS] tokens within each group are L2 normalized, and then the T-SP metric is applied to calculate the similarity between tokens from each group. Finally, a simple cross-entropy loss is calculated symmetrically to drive model training, enhancing the intraclass conceptual compactness within an image and the interclass semantic separability across images.

3. Experiments

We perform self-supervised pre-training on the ImageNet-1K (Russakovsky et al., 2015) dataset with the resolution of 224×224. By default, ViT-B/16 and ViT-L/16 (Dosovitskiy et al., 2021) are leveraged as the backbone architecture with 800 epochs for pre-training and 40 epochs for warm-up.

Model	Blocks	Dim	Heads	Params
ViT-B/16	12	768	12	86M
ViT-L/16	24	1024	16	304M
ViT-h/16	32	1280	16	632M

Table 1. **Configurations of Vision Transformer (Dosovitskiy et al., 2021) models in our experiments.** Block denotes the number of transformer blocks, with dim representing the input/output channel dimension of all blocks. Heads are the number of heads in multi-head attention modules. We also provide the parameter sizes of different models.

ViT-B/16 applies the overall masked ratio of 0.3, while ViT-L/16 is set to 0.4. The initial base learning rate is 1.5×10^{-4}. Like other contrastive learning (Chen et al., 2020a; Radford et al., 2021), our method relies on a large effective batch size: 9,600 for ViT-B/16 and 2,048 for ViT-L/16. Besides, κ is set to 64 in the T-SP metric for computing similarity by default. Our implementation is based on MAE (He et al., 2022), with further details provided in the Supplementary Material.

In terms of supervised validation, OCL is evaluated through end-to-end fine-tuning and linear probing on the ImageNet-1k dataset for classification with 100 epochs for ViT-B/16 and 50 for ViT-L/16, following common practices (Caron et al., 2021; He et al., 2022; Wang et al., 2023). Top-1 accuracy is utilized to verify the performance of different methods.

3.1. Scalability

To demonstrate the scalability of our OCL for efficient conceptual pre-training, we access the efficiency and scaling of our model in Fig.3. It illustrates the pre-training hours related to various model sizes for different methods, with linear probing accuracy. OCL is highly scalable compared to previous methods, requiring less computational resources while achieving strong results and without relying on handcrafted data augmentations. Unlike reconstruction-based approaches like MAE and I-JEPA, OCL needs fewer training epochs and avoids the need for pixel-level reconstruction, significantly improving training speed. In contrast to contrastive learning methods such as MoCo v3, which depend on handcrafted augmentations and complex architectures to generate and process multiple image views, OCL eliminates the need for auxiliary modules like momentum encoders. This simplicity makes OCL's framework more efficient and accelerates pre-training. For example, when scaling up from ViT-B/16 to ViT-L/16, OCL requires far less additional pre-training time compared to MoCo v3.

Scaling model size. Moreover, our model leverages a scalable model size, resulting in more substantial performance enhancements with the larger model as illustrated in Fig.3.

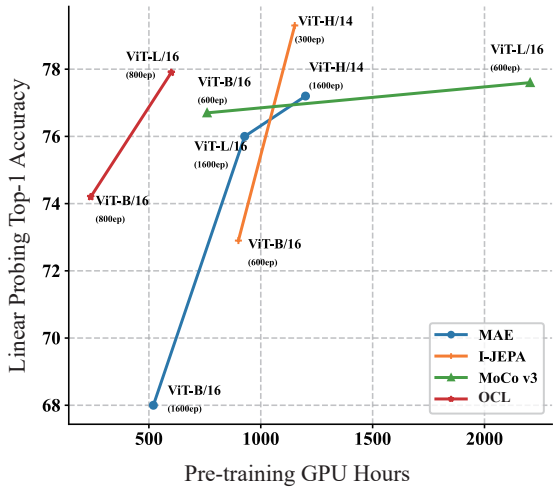


Figure 3. **Efficiency and Scaling.** MAE (He et al., 2022), I-JEPA (Assran et al., 2023) and MoCo v3 (Chen et al.) are opted for comparison. All methods are evaluated by linear probing with Top-1 accuracy (Acc) as the metric, and the pre-training GPU time with A100 hour as the indicator. The pre-training epochs (denoted as ep) and model architecture are also exhibited.

Compared to ViT-B/16, OCL achieves nearly 4% improvement in linear probing with ViT-L/16, surpassing MoCo v3. It implies that we can efficiently train larger models to achieve better performance, within an acceptable timeframe.

3.2. Ablation Studies

3.2.1. Masked Ratio

Firstly, we conduct ablation experiments to discuss the impact of the overall masked ratio on the performance of conceptual pre-training, revealed in Table 2. To ensure clarity in presentation, we offer the visible ratio of each contrastive branch within a single forward pass of the mini-batch, which is a crucial factor for the cross-similarity within our OCL framework. Likewise, the effective batch size is intertwined with the scalable masked ratios, with larger batch sizes showcasing performance enhancements for the model, as verified in Section 3.2.2. Thus, we attribute this improvement to the masked ratio.

Regarding large models such as ViT-L/16 in Table 2, the increment of the overall masked ratio from 0.2 to 0.4 could significantly diminish conceptual redundancy and increase the effective batch size, thereby resulting in performance improvement. Nevertheless, with a continued rise in the masked ratio, the visible patches of images in one forward process of the model diminish incrementally. Despite increases in batch size, the precise extraction of conceptual information from the images becomes compromised, resulting in degraded performance. Furthermore, as displayed

Overall Masked Ratio	Forward Visible Ratio	Bsz.	LIN	FT
ViT-L/16				
0.2	0.4	1,024	74.2	84.4
0.4	0.3	2,048	77.9	85.8
0.8	0.1	7,200	69.1	83.0
ViT-B/16				
0.3	0.35	9,600	74.2	83.4
0.6	0.2	9,600	71.0	82.8
0.8	0.1	12,800	65.7	81.6

Table 2. **Ablation evaluation experiments on masked ratio.** The results are based on the ImageNet-1K with the Vit-B/16 and ViT-L/16. All methods are evaluated by linear probing (LIN) and fine-tuning (FT). We provide the visible ratio of each branch for contrastive learning according to the masked ratio. Correspondingly, we give different effective batch sizes (Bsz.) related to the overall masked ratio. The resolution of images is fixed to 224×224. Top-1 accuracy is used as the metric.

in ViT-B/16 of Table 2, once the batch size surpasses its threshold, an excessively high masked ratio can impair the model performance, transitioning from reducing conceptual redundancy to damaging essential semantic information.

3.2.2. Large Batch Size for Effective Contrastive Representation

Effective Batch Size	LIN	FT	Pre-training Hour
2,048	77.9	85.8	533
4,096	77.7	85.7	533
1,800	77.4	85.7	559
1,024	74.4	84.9	586
512	71.6	84.6	613

Table 3. **Ablation evaluation experiments on batch size.** The results are based on the ImageNet-1K with the ViT-L/16. All methods are evaluated by pre-training hours, linear probing (LIN) and fine-tuning (FT). Meanwhile, pre-training hours on A100 are provided. The resolution of images is fixed to 224×224. Top-1 accuracy is used as the metric.

ViT models (Dosovitskiy et al., 2021) are inherently computationally intensive, and training with large batches is a preferred strategy for handling large ViT models. Moreover, a sizable batch size is advantageous for achieving accuracy in contemporary self-supervised learning techniques. In particular, concerning contrastive learning methodologies that heavily lean on large batch sizes, ablation experiments are conducted to ascertain the influence of batch size on our occluded image contrastive learning approach, as shown in Table 3. The utilization of the ViT-L/16 model for validation reveals that a larger effective batch size correlates with improved model performance and efficiency, aligning with the consensus within the community (Goyal et al., 2018; You et al., 2019). We attribute this to the statistical advantage of expanded negative sample pools, which achieve broader coverage of the latent feature space. Such comprehensive sampling sharpens inter-class separation by refining decision boundaries during contrastive optimization. However, limited by computational resources, we are unaware of the maximum capacity of the effective batch size.

3.2.3. MLP Head is Not You Need

MLP Head	FLOPs	Param.	Hours	Bsz.	LIN	FT
w/o	42.3	303	533	2,048	77.9	85.8
w/o	42.3	303	559	1,800	77.4	85.7
2-layer	43.6	305	600	1,800	76.7	85.6
3-layer	43.7	306	611	1,800	76.6	85.6

Table 4. **Ablation experiments on MLP head.** The results are based on the ImageNet-1K with the ViT-L/16. All methods are evaluated by FLOPs/G, Parameters (Param.)/M, pre-training hours (Hours), effective batch size (Bsz.), linear probing (LIN) and fine-tuning (FT). Correspondingly, we give different effective batch sizes related to the MLP head within the pre-trained model. Besides, pre-training hours on A100 are provided. The resolution of images is fixed to 224×224. Top-1 accuracy is used as the metric.

In many contrastive learning methods (Chen et al., 2020a; Chen et al.), the ViT model is typically paired with an MLP head (Taud & Mas, 2018) to assist with the pretext task. The ViT learns semantic features from the image, while the MLP head handles the pretext classification task, enhancing the ViT's learning capabilities, as shown in earlier work. However, traditional contrastive learning methods rely on handcrafted data augmentations to generate and process multiple image views at the instance level. In contrast, OCL explores an alternative approach: using masked images to create diverse views with fine-grained conceptual differences at the token level. As a result, OCL does not require an MLP head for instance-level classification tasks.

The ablation experiments on the MLP head involve testing different configurations, including a 2-layer MLP, a 3-layer MLP, and no MLP head (see Table 4). The design of the MLPs follows MoCo v3 (Chen et al.), where the hidden layers of both the 2-layer and 3-layer MLPs are 1024-dimensional and employ the GELU activation function (Hendrycks & Gimpel, 2023). The output layers of both MLPs are 512-dimensional and do not use GELU. Additionally, all layers in the MLPs incorporate Batch Normalization (BN) (Ioffe, 2015), consistent with the methodology in SimCLR (Chen et al., 2020b). Due to GPU memory constraints, the batch size was adjusted to 1,800 for validation. From

the fine-tuning results presented in Table 4, OCL is capable of operating effectively without an MLP head. Unlike traditional contrastive learning approaches, the absence of an MLP head does not degrade the model's performance. On the other hand, the inclusion of additional auxiliary modules, such as deeper MLP heads, increases both the cost and time required for pre-training. However, the benefits of these modules do not outweigh the performance gains achieved by simply increasing the batch size.

T-SP Loss	MLP Head	LIN	FT
✓	✓	77.0	85.5
×	✓	72.4	85.1
✓	×	77.9	85.8
×	×	61.3	82.6

Table 5. **Ablation experiments on the relationship between MLP head and T-SP loss.** The results are based on the ImageNet-1K with the ViT-L/16. All methods are evaluated by linear probing (LIN) and fine-tuning (FT). The resolution of images is fixed to 224×224. Top-1 accuracy is used as the metric.

Besides, Table 5 provides the ablation experiments about the relationship between the MLP head and T-SP loss. It is demonstrated that T-SP loss is utilized to constrain the similarity distributions between two views, enforcing more elaborate feature extraction and representations through explicit probabilistic margin compression. In contrast, the MLP head operates through implicit feature projection into a lower-dimensional space. While their mechanisms differ, both components contribute to efficient representations in pre-trained models.

3.2.4. Concentration of Similarity Metric κ

kappa	4	16	32	64	128
FT	84.8	85.1	85.6	85.8	85.5

Table 6. **Ablation experiments on concentrate parameter of kappa.** The results are based on the ImageNet-1K with the ViT-L/16. Fine-tuning (FT) results are provided. The resolution of images is fixed to 224×224. Top-1 accuracy is used as the metric.

Following the approach of T-SP (Yang et al.), we conduct ablation experiments to evaluate how different concentrations of the T-distributed adapter affect the pre-training model's performance. The results are shown in Table 6. This study involves five specific degrees of concentration, namely $\kappa = 4$, 16, 32, 64, and 128. From the Table 6, we can infer that as kappa increases, the pre-training performance of the model improves progressively until reaching $\kappa = 64$ with the fine-tuning result of 85.8 under ViT-L/16. We explain that as kappa increases, the model must extract more precise semantic concepts from positive samples, creating a more challenging pretext task for pre-training. However, if kappa becomes too large, the model converges too slowly, which negatively impacts the overall effectiveness of the pre-training process.

3.3. Comparisons with previous results

	Aug.	Ep.	FLOPs	Param.	LIN	FT
Masked Image Modeling						
BEiT †	w/o	800	17.6	87	-	83.2
MAE †	w/o	1,600	17.5	86	68.0	83.6
CAE †	w/o	1,600	17.5	86	70.4	83.9
I-JEPA	w/o	600	17.5	86	72.9	-
Contrastive Learning						
DINO	w/	1,600	74.7	171	78.2	82.8
MoCo v3	w/	600	74.7	171	76.7	83.2
Masked Image Modeling with Contrastive Learning						
SiameseIM	w/	1,600	16.3	88	78.0	84.1
ccMIM	w/o	800	39.1	86	68.9	84.2
ConMIM	w/	800	17.5	86	-	85.3
MixMAE †	w/o	600	15.6	88	61.2	84.6
iBOT †	w/	1,600	17.5	86	79.5	-
OCL	w/o	800	12.0	86	74.2	83.4

Table 7. **Comparison with previous methods on ImageNet-1K classification with ViT-B model.** All methods are evaluated by Aug., Epoch (Ep.), FLOPs/G, Parameters (Param.)/M, linear probing (LIN) and fine-tuning (FT). The resolution of images is fixed to 224×224. Aug. indicates the utilization of handcrafted view data augmentation during pre-training. FLOPs/G is utilized to show the runtime and computational resources of the pre-training. Param./M is calculated for the encoders of the pre-training model, following the MixMAE. † denotes the results are copied from the MixMAE. Top-1 accuracy (Acc) is used as the metric.

To demonstrate OCL's ability to learn high-level conceptual representations without relying on handcrafted data augmentations, we compare its performance on linear probing and fine-tuning tasks under pre-training on ImageNet-1k. Importantly, while masked images are used to reduce semantic redundancy, OCL avoids reconstructing masked images. Instead, it uses contrastive loss to guide the network's learning process. As a result, OCL is categorized as a contrastive learning method.

Table.7 exhibits the performance of our method under the fine-tuning and linear probing for ImageNet-1k classification. Of greater significance, our FLOPs of 12.0 G exceed those of the second-best method MixMAE (Liu et al., 2023a) with 15.6 G, improving efficiency by about 23%. Concerning DINO (Caron et al., 2021) and MoCo v3 (Chen et al.), they leverage student-teacher dual networks to pre-train, leading to higher FLOPs and Parameters of the pre-training encoder. Our OCL method has demonstrated outstanding performance in contrastive learning, notably obviating the

need for intricate augmentations across multiple views. It validates the practicality and efficacy of masked images for generating diverse views encapsulating distinct fine-grained semantic concepts, which diminishes conceptual redundancy and expedites the conceptual pre-training process. Beyond that, it also stands out the significant competency of contrastive learning in extracting high-level semantic concepts.

Methods	Epochs	COCO AP^b	COCO AP^m	ADE20k mIoU
ViT-B/16				
Supervised (Tao et al., 2023)	300	47.9	42.9	47.4
DINO (Caron et al., 2021)	800	50.1	43.4	46.8
iBOT (Zhou et al., 2022)	1600	51.2	44.2	50.0
DenseCL (Wang et al., 2021)	400	46.6	41.6	44.5
MoCo-v3 (Chen et al., 2021b)	600	47.9	42.7	47.3
BEiT (Bao et al., 2021)	800	49.8	44.4	47.1
MAE (He et al., 2022)	400	50.6	45.1	45.0
MAE (He et al., 2022)	1600	51.6	45.9	48.1
GreenMIM (Huang et al., 2022)	800	50.0	44.1	-
EsViT (Li et al., 2021a)	300	-	-	47.3
MixMAE (Liu et al., 2023a)	300	52.3	46.4	49.9
MixMAE (Liu et al., 2023a)	600	52.7	47.0	51.1
SiameseIM (Tao et al., 2023)	400	50.7	44.9	49.6
SiameseIM (Tao et al., 2023)	1600	52.1	46.2	51.1
SimMIM (Xie et al., 2022)	300	51.1	45.4	48.9
OCL	800	51.5	45.5	46.1
ViT-L/16				
MoCo-v3 (Chen et al., 2021b)				
BEiT (Bao et al., 2021)	800	53.3	47.1	53.3
MAE (He et al., 2022)	1600‘	53.3	47.2	53.6
SimMIM (Xie et al., 2022)	800	53.8	-	53.6
MixMAE (Liu et al., 2023a)	600	54.3	48.2	53.8
OCL	800	53.2	47.0	53.2

Table 8. **Comparison with previous methods on various downstream tasks, including object detection and segmentation on COCO and ADK20K.** We report AP^{box} (AP^b) and AP^{mask} (AP^m) on COCO, and mIoU on ADE20K. Arch. represents the model architecture, where ViT-B/16 and ViT-L/16 are utilized to validate the performance of various methods.

Compared to MIM methods, due to the absence of pixel-level image reconstruction in OCL, the training process is finished in 800 epochs, in contrast to the 1600 epochs required for MAE (He et al., 2022), CAE (Chen et al., 2024). It corroborates the contribution of our methodology from the perspective of efficiency and high-level semantic concept extraction. Moreover, our OCL method is superior to BEiT (Bao et al., 2021) regarding the fine-tuning and linear probing results, also confirming the competency and robustness of our method.

Concerning the fusion of MIM and CL, significant endeavours (Tao et al., 2023; Yi et al., 2023; Zhou et al., 2022) are made to enhance pre-training performance on downstream tasks. However, they result in inefficiencies due to increased training iterations, augmented data, auxiliary modules, and diverse loss combinations. For instance, SiameseIM (Tao et al., 2023), ConMIM (Yi et al., 2023) and iBOT (Zhou et al., 2022) leverage both hand-crafted data augmentation and more training epochs. Despite our model also integrating masking strategy with contrastive learning, it does not rely on hand-crafted view data augmentations and additional auxiliary modules, reconciling the divergence between efficient visual representation and effective conceptual pre-training. Besides, our model OCL also achieves competitive results on downstream tasks of linear probing and fine-tuning, and more importantly, shows promising scaling behavior.

Beyond that, we conducted supervised fine-tuning on the COCO dataset for object detection and instance segmentation using the Mask RCNN (He et al.) framework, with our pre-trained encoder serving as the backbone. We follow the setup of MixMAE (Liu et al., 2023a), with the window size to 16×16 to align with the 1024×1024 input image resolution. In terms of semantic segmentation on the ADE20k dataset, We use the UperNet (Xiao et al.) framework with our pre-trained encoder as its backbone, with the changed window size as mentioned above. The results are shown in Table 8, that our model achieves competitive results compared to other models with efficient pre-training progress. Compared to the contrastive learning method of MoCo-v3, our method obtains 3.6% improvement of AP^{box} under ViT-B/16, especially without view data augmentations. Besides, the consistent performance gains on ADE20K semantic segmentation, achieving +7.1 mIoU improvements from ViT-B to ViT-L, empirically validate our method's scalability across model capacities while maintaining parameter efficiency

Moreover, we conduct experiments in the domain generalization setting following the SiameseIM (Tao et al., 2023), as shown in Table 9. Experiments are conducted on ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019), with ImageNet (Russakovsky et al., 2015) as the source dataset. Our method achieves results with a competitive edge, demonstrating the advantages of our model in generalization and robustness. Notably, our framework achieves competitive performance with significantly fewer training epochs than conventional contrastive approaches, demonstrating remarkable training efficiency while maintaining stable convergence—a critical advantage for scaling to large-scale datasets and complex architectures.

Methods	Epochs	IN-A	IN-R	IN-S
MSN (Assran et al., 2022b)	1200	37.5	50.0	36.3
iBOT (Zhou et al., 2022)	1600	42.4	50.9	36.9
DenseCL (Wang et al., 2021)	400	30.8	43.8	29.9
MoCo-v3 (Chen et al., 2021b)	600	32.4	49.8	35.9
MAE (He et al., 2022)	1600	35.9	48.3	34.5
SiameseIM (Tao et al., 2023)	1600	43.8	52.5	38.3
OCL	800	42.2	52.3	37.6

Table 9. **Comparsion with previous methods on generalization capability and robustness on ImageNet-A (IN-A), ImageNet-R (IN-R) and ImageNet-Sketch (IN-S) datasets.** ViT-B/16 is utilized as the backbone to validate the performance of various methods. Top-1 accuracy is used as the metric.

4. Conclusions and Outlooks

In this paper, we introduce OCL, a novel, simple, and effective pre-training paradigm for visual conceptual representation. Our approach uses a masking strategy to generate diverse views with fine-grained semantic differences, enabling contrastive learning to classify and learn agreements within a mini-batch. This design eliminates the need for I) addressing semantic conceptual redundancy within images, II) reconstructing images, III) hand-crafted data augmentations, and IV) additional auxiliary modules, thereby improving efficiency and scalability. Experiments showcase the efficiency and scalability of our method, yielding competitive results compared to previous approaches. Additionally, ablation experiments provide insights that could inspire future pre-training paradigms.

We hope that our work will inspire future advancements in contrastive pre-training paradigms, specifically efficient visual representation. Our further work will focus on enhancing computational efficiency for billion-parameter models without sacrificing representation quality.

Impact Statement

This paper presents work whose goal is to advance the field of machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgment

This research was partially supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Shaoting Zhang and Hongsheng Li are Principal Investigators of the Centre of Perceptual and Interactive Intelligence (CPII) under the InnoHK.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2022a.

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked Siamese Networks for Label-Efficient Learning. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, pp. 456–473. Springer Nature Switzerland, 2022b. ISBN 978-3-031-19821-2. doi: 10.1007/978-3-031-19821-2_26.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pp. 1298–1312. PMLR, 2022.

Baevski, A., Babu, A., Hsu, W.-N., and Auli, M. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pp. 1416–1429. PMLR, 2023.

Bao, H., Dong, L., Piao, S., and Wei, F. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=p-BhZSz59o4.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big Self-Supervised Models are Strong Semi-Supervised Learners, October 2020b.

Chen, X., Xie, S., and He, K. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021a.

Chen, X., Xie, S., and He, K. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021b. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. Context Autoencoder for Self-supervised Representation Learning. *International Journal of Computer Vision*, 132 (1):208–223, January 2024. ISSN 1573-1405. doi: 10.1007/s11263-023-01852-4.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. URL http://arxiv.org/abs/2010.11929.

Gao, P., Lin, Z., Zhang, R., Fang, R., Li, H., Li, H., and Qiao, Y. Mimic before Reconstruct: Enhancing Masked Autoencoders with Feature Mimicking. 132 (5):1546–1556, 2024. ISSN 1573-1405. doi: 10.1007/s11263-023-01898-4. URL https://doi.org/10.1007/s11263-023-01898-4.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, April 2018.

Gupta, A., Wu, J., Deng, J., and Fei-Fei, L. Siamese Masked Autoencoders, 2023. URL http://arxiv.org/abs/2305.14344.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969. URL https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Hendrycks, D. and Gimpel, K. Gaussian Error Linear Units (GELUs), June 2023.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Huang, L., You, S., Zheng, M., Wang, F., Qian, C., and Yamasaki, T. Green Hierarchical Vision Transformer for Masked Image Modeling. 35:19997–20010, 2022.

Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.-M., Fu, D., Shen, X., and Feng, J. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jiang, Z., Chen, Y., Liu, M., Chen, D., Dai, X., Yuan, L., Liu, Z., and Wang, Z. Layer Grafted Pre-training: Bridging Contrastive Learning And Masked Image Modeling

For Label-Efficient Representations. In *The Eleventh International Conference on Learning Representations*, September 2023.

Jing, L., Zhu, J., and LeCun, Y. Masked siamese convnets: Towards an effective masking strategy for general-purpose siamese networks.

Kobayashi, T. T-vMF Similarity For Regularizing Intra-Class Feature Distribution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6621. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00655. URL https://ieeexplore.ieee.org/document/9578219/.

Kong, L., Ma, M. Q., Chen, G., Xing, E. P., Chi, Y., Morency, L.-P., and Zhang, K. Understanding Masked Autoencoders via Hierarchical Latent Variable Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7918–7928, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Kong_Understanding_Masked_Autoencoders_via_Hierarchical_Latent_Variable_Models_CVPR_2023_paper.html.

Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., and Gao, J. Efficient Self-supervised Vision Transformers for Representation Learning. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=fVu3o-YUGQK.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023.

Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., and Wang, J. MST: Masked Self-Supervised Transformer for Visual Representation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13165–13176. Curran Associates, Inc., 2021b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 740–755. Springer International Publishing, 2014. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Liu, J., Huang, X., Zheng, J., Liu, Y., and Li, H. MixMAE: Mixed and Masked Autoencoder for Efficient Pretraining of Hierarchical Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6252–6261, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_MixMAE_Mixed_and_Masked_Autoencoder_for_Efficient_Pretraining_of_Hierarchical_CVPR_2023_paper.html.

Liu, W., Chen, Y., Yue, X., Zhang, C., and Xie, S. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8870–8878, 2023b.

Liu, W., Chen, Y., and Yue, X. Building trust in decision with conformalized multi-view deep classification. In *ACM Multimedia 2024*, 2024.

Liu, W., Chen, Y., and Yue, X. Enhancing multi-view classification reliability with adaptive rejection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18969–18977, 2025.

Mishra, S., Robinson, J., Chang, H., Jacobs, D., Sarna, A., Maschinot, A., and Krishnan, D. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. *arXiv preprint arXiv:2210.16870*, 2022.

Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., and Yi, L. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pp. 28223–28243. PMLR, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., and Bernstein, M. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., and Dai, J. Siamese Image Modeling for Self-Supervised Vision Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2132–2141, 2023.

Taud, H. and Mas, J. Multilayer Perceptron (MLP). In Camacho Olmedo, M. T., Paegelow, M., Mas, J.-F., and Escobar, F. (eds.), *Geomatic Approaches for Modeling Land Change Scenarios*, pp. 451–455. Springer International Publishing, Cham, 2018. ISBN 978-3-319-60801-3. doi: 10.1007/978-3-319-60801-3_27.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Wang, H., Fan, J., Wang, Y., Song, K., Wang, T., and Zhang, Z.-X. DropPos: Pre-Training Vision Transformers by Reconstructing Dropped Positions. 36:46134–46151, 2023.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense Contrastive Learning for Self-Supervised Visual Pre-Training. pp. 3024–3033, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Dense_Contrastive_Learning_for_Self-Supervised_Visual_Pre-Training_CVPR_2021_paper.html.

Wei, Y., Gupta, A., and Morgado, P. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2025.

Wu, Z., Lai, Z., Sun, X., and Lin, S. Extreme masking for learning instance and distributed visual representations. *arXiv preprint arXiv:2206.04667*, 2022.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Tete_Xiao_Unified_Perceptual_Parsing_ECCV_2018_paper.html.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xie_SimMIM_A_Simple_Framework_for_Masked_Image_Modeling_CVPR_2022_paper.html.

Yang, J., Chen, H., Liang, Y., Huang, J., He, L., and Yao, J. ConCL: Concept Contrastive Learning for Dense Prediction Pre-training in Pathology Images. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, pp. 523–539, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19803-8. doi: 10.1007/978-3-031-19803-8_31.

Yang, J., Guo, S., Wu, G., and Wang, L. CoMAE: Single Model Hybrid Pre-training on Small-Scale RGB-D Datasets, February 2023a.

Yang, X., Chen, Y., Yue, X., Xu, S., and Ma, C. T-distributed Spherical Feature Representation for Imbalanced Classification. 37(9):10825–10833. ISSN 2374-3468. doi: 10.1609/aaai.v37i9.26284. URL https://ojs.aaai.org/index.php/AAAI/article/view/26284.

Yang, X., Lu, J., and Yu, E. Adapting multi-modal large language model to concept drift from pre-training onwards. *arXiv preprint arXiv:2405.13459*, 2024a. URL https://arxiv.org/abs/2405.13459.

Yang, X., Xu, L., Sun, H., Li, H., and Zhang, S. Enhancing visual grounding and generalization: A multi-task cycle training approach for vision-language models. *arXiv preprint arXiv:2311.12327*, 2024b. URL https://arxiv.org/abs/2311.12327.

Yang, X., Lu, J., and Yu, E. Adapting multi-modal large language model to concept drift from pre-training onwards. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=b20VK2GnSs.

Yang, X., Lu, J., and Yu, E. Walking the tightrope: Disentangling beneficial and detrimental drifts in non-stationary custom-tuning. *arXiv preprint arXiv:2505.13081*, 2025b. URL https://arxiv.org/abs/2505.13081.

Yang, X., Lu, J., and Yu, E. Causal-informed contrastive learning: Towards bias-resilient pre-training under concept drift. *arXiv preprint arXiv:2502.07620*, 2025c. URL https://arxiv.org/abs/2502.07620.

Yang, Y., Huang, W., Wei, Y., Peng, H., Jiang, X., Jiang, H., Wei, F., Wang, Y., Hu, H., Qiu, L., et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2771–2781, 2023b.

Yi, K., Ge, Y., Li, X., Yang, S., Li, D., Wu, J., Shan, Y., and Qie, X. Masked Image Modeling with Denoising Contrast. In *The Eleventh International Conference on Learning Representations*, September 2023.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *International Conference on Learning Representations*, September 2019.

Zhang, Q., Wang, Y., and Wang, Y. How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders. 35:27127–27139, 2022.

Zhang, S., Zhu, F., Zhao, R., and Yan, J. Contextual Image Masking Modeling via Synergized Contrasting without View Augmentation for Faster and Better Visual Pretraining. In *The Eleventh International Conference on Learning Representations*, September 2023.

Zhao, Y., Wang, G., Luo, C., Zeng, W., and Zha, Z.-J. Self-Supervised Visual Representations Learning by Contrastive Mask Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10160–10169. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Self-Supervised_Visual_Representations_Learning_by_Contrastive_Mask_Prediction_ICCV_2021_paper.html.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene Parsing Through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*, October 2022.

A. Related Works

The remarkable success of large-scale models across diverse domains has catalyzed a paradigm shift in machine learning (Yang et al., 2025b;a), with pretraining emerging as a cornerstone methodology. However, visual representation learning remains constrained by prohibitive computational requirements compared to its NLP counterparts, leading to more research concerning the efficiency of the visual pre-training. In this section, we will review the three main vision pre-training methods, namely masked image modeling, contrastive learning and their combination. Besides, the difference between ours and theirs is also provided.

A.1. Masked Image Modeling

Inspired by masked language modeling in NLP, the core of masked image modeling is to predict the masked part of the input image. Among them, BEiT (Bao et al., 2021) tokenizes image patches through the reconstruction of the individual image using dVAE, and then predicts the tokens of the masked patches to learn visual representation. Similarly, MAE (He et al., 2022) utilizes a high masked ratio (75%) to corrupt the image and directly reconstruct the pixel-level masked image patches. Subsequently, numerous studies have referenced this paradigm for pre-training endeavours, including DropPos (Wang et al., 2023), U-MAE (Zhang et al., 2022) and CAE (Chen et al., 2024). DropPos (Wang et al., 2023) incorporates position reconstruction to bolster the spatial awareness of ViTs. U-MAE (Zhang et al., 2022) introduces a uniformity loss as a regularization to the MAE loss to further encourage the feature consistency of the pre-training, and addresses the dimensional feature collapse. CAE (Chen et al., 2024) decouples the learning processes for image representation and pretext tasks, enabling the pre-trained model to prioritize image representation while disregarding the pretext task. Additionally, following the successful implementation of MAE, its applicability has been extended to diverse disciplines, such as SiamMAE (Gupta et al., 2023) and MR-MAE (Gao et al., 2024). Similarly, I-JEPA (Assran et al., 2023) predicts the feature encoding of the contextual region via MIM, and employs contrastive learning to align the features of neighbouring regions at the feature level.

Besides, latentMIM (Wei et al., 2025) tackles key training challenges in Latent MIM, showcasing its ability to produce spatially diverse, high-level semantic representations. Context Autoencoders (Chen et al., 2021a) employ an encoder-decoder framework optimized with a combined reconstruction loss and alignment constraint, ensuring predictable representations of missing patches. data2vec (Baevski et al., 2022) predicts representations of missing patches using an online target encoder, eliminating the need for handcrafted augmentations and achieving strong performance across vision, text, and speech modalities. Its successor, data2vec-v2 (Baevski et al., 2023), further explores efficient architectures for multimodal learning.

A.2. Contrastive Learning

Among these methods, SimCLR (Chen et al., 2020a) relies on complex pre-processing techniques to create distinct views of an image, aiming to make the task challenging enough to learn effective visual representations during pre-training. However, it requires large batch sizes to generate negative sample pairs, leading to long training times and high computational costs. DINO (Caron et al., 2021) uses a student-teacher architecture to extract visual representations from different views. Similarly, MoCo v3 (Chen et al.) employs momentum updates to optimize its auxiliary network. These methods highlight that the core of contrastive learning lies in creating views with significant differences. However, this is challenging due to the inherent semantic redundancy in images. Additionally, ConCL (Yang et al., 2022) generates distinct concepts by cropping images and applies contrastive learning within a teacher-student framework, specifically for pre-training on pathological images.

DenseCL (Wang et al., 2021) introduces the concept of dense contrastive loss, which calculates the contrastive loss between dense feature vectors generated by the dense projection head at the local feature level, contrasting with traditional contrastive learning methods that operate at the global feature level. Moreover, MaskCo (Zhao et al.) utilizes the teacher-student network for generating image features of query and keys. The key features of different images are used to query certain images to generate positive and negative examples for contrastive learning. Furthermore, MSN (Assran et al., 2022b) leverages data augmentation to create two views known as the anchor view and the target view. Subsequently, a random mask is applied to the anchor view, aligning the representation of the masked anchor view with the clusters of the unmasked target view. Besides, ResilientCL(Yang et al., 2025c) leverages causal inference to analyze bias sources in contrastive pre-training's momentum updates and proposes a causal interventional objective to address distribution shifts. Moreover, building on vision-specific uncertainty principles, multi-stage framework (Liu et al., 2023b; 2024; 2025) employs learnable concept lattices to stabilize pretraining under covariate shift.

A.3. Combination between MIM and CL

There are also a lot of efforts to bridge the gap between mask image modeling and contrastive learning, where the integration of teacher-student models emerges as the prevailing approach. iBOT (Zhou et al., 2022) employs the teacher-student network to independently encode the two augmented views, with the student network processing masked images. The objectives of MIM and CL are jointly trained for self-distillation. Likewise, MST (Li et al., 2021b) introduces a masked token strategy leveraging multi-head self-attention maps, which selectively mask the tokens of the student network based on the output self-attention map of the teacher network, ensuring vital foreground remains intact. Similarly, SiameseIM (Tao et al., 2023) employs a Siamese network featuring two branches. The online branch encodes the initial view and predicts the representation of the second view based on their relative positions. Meanwhile, the target branch generates the target by encoding the second view. MSCN (Jing et al.) generates multiple augmented views from input images and applies random masking. These masked views are encoded using a standard ConvNet, with representations optimized through a joint-embedding loss. RECON (Qi et al., 2023) integrates generative and contrastive learning paradigms via ensemble distillation, where a generative student model guides a contrastive student to unify both approaches. CMAE (Huang et al., 2023) employs a dual-branch architecture: an online branch with an asymmetric encoder-decoder for reconstructing masked images, and a momentum branch with a momentum encoder for contrastive learning on full images. This design enables holistic feature learning and enhanced discrimination.

ccMIM (Zhang et al., 2023) leverages a contrastive loss to aid the reconstruction task as a regularizer, facilitating the extraction of image-wide global information from both masked and unmasked patches. Likewise, ConMIM (Yi et al., 2023) produces simple intra-image inter-patch contrastive constraints as the sole learning objectives for masked patch prediction, and strengthens the denoising mechanism with asymmetric designs to improve the network pre-training. Additionally, CoMAE (Yang et al., 2023a) also applies CL to assist cross-modal MIM tasks. Besides, LGP (Jiang et al., 2023) integrates MIM and CL in a sequential cascade manner: early layers are first trained under one MIM loss, on top of which latter layers continue to be trained under another CL loss.

A.4. Differences of OCL from Existing Methods

Masking has been adopted as an effective data augmentation technique to enhance training efficiency in several studies (Li et al., 2023; Yang et al., 2023b; Mishra et al., 2022; Wu et al., 2022; Assran et al., 2022a). Previous studies have demonstrated its effectiveness in improving model performance and training efficiency (Li et al., 2023; Yang et al., 2023b). ExtreMA (Wu et al., 2022) employs random masking as a computationally efficient augmentation for Siamese representation learning, accelerating learning and enhancing performance on large datasets. MSN (Assran et al., 2022a), the most relevant work to ours, generates two image views: a masked anchor view and an unmasked target view, aiming to cluster their representations. CAN (Mishra et al., 2022) integrates contrastive learning, masked auto-encoding, and diffusion denoising into a unified framework. Our approach distinguishes itself from previous hybrid methods by achieving superior performance with a better performance-efficiency trade-off. Specifically, our masking strategy generates semantically diverse views and leverages contrastive learning to promote classification agreement within mini-batches. This approach eliminates: (I) semantic redundancy, (II) image reconstruction, (III) hand-crafted augmentations, and (IV) additional auxiliary modules, resulting in enhanced efficiency and scalability.

B. Experiments

In this section, we first introduce the utilized datasets for pre-training and various downstream tasks. Subsequently, implementation details are provided. Finally, we present more results on downstream tasks, such as object detection, segmentation and domain generalization.

B.1. Datasets

The ImageNet-1k dataset (Russakovsky et al., 2015) is a widely used image dataset consisting of 1.28M labeled images across 1k categories, with 50K validation images and 100k test images. The dataset has been instrumental in advancing computer vision research by providing a large-scale benchmark for image classification tasks. By leveraging the vast amount of labeled images in ImageNet-1k, self-supervised models can learn rich representations of visual data in an unsupervised manner, which can then be fine-tuned on downstream tasks with smaller labeled datasets. Besides, ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019) are leveraged for validation

of the generalization capability and robustness of the vision model, with the training source of ImageNet.

MSCOCO (Lin et al., 2014) dataset is a large-scale dataset widely used for object detection and instance segmentation tasks created by Microsoft Research Asia. The COCO Detection dataset contains more than 330K images, including more than 1.5M labeled object instances for 80 different categories. Each object instance is labeled with category, bounding box, and segmentation mask information.

The ADE20K (Zhou et al.) semantic segmentation dataset comprises over 20K scene-centric images for 150 semantic categories, meticulously annotated with pixel-level object and object parts labels.

B.2. Implementation Details

We employ ViT-Base and ViT-Large as our visual backbones, respectively. Among them, Vit-Base consists of 12 transformer encoder layers and an FFN intermediate size of 3,072. The hidden dimensions of the ViT-Base are 768, with 12 attention heads. The number of parameters is about 86 million. The input image size is set to 224×224. In terms of ViT-L/16, ViT-L/16 consists of 24 transformer encoder layers and an FFN intermediate size of 4,096. The input image size is set to 224×224, with a patch size of 16×16. The hidden dimensions of the ViT-Large are 1,024, with 16 attention heads. And, the number of parameters is about 307 million.

Table 10. The pre-training hyperparameters.

	ViT-B/16	ViT-L/16
Training Epochs	800	800
Warmup Epochs	40	40
Optimizer	AdamW	AdamW
Base Learning Rate	1.5e-4	1.5e-4
Learning Rate Decay	Cosine	Cosine
Adam β	(0.9, 0.95)	(0.9, 0.95)
Weight Decay	0.05	0.05
Batch Size	9,600	2,400

Table 11. The fine-tuning hyperparameters.

	ViT-B/16	ViT-L/16
Training Epochs	100	50
Warmup Epochs	5	5
Optimizer	AdamW	AdamW
Base Learning Rate	5e-4	1e-3
Learning Rate Decay	Cosine	Cosine
Adam β	(0.9, 0.95)	(0.9, 0.95)
Weight Decay	0.05	0.05
Batch Size	1,024	1,024

In terms of the pre-training progress, the hyperparameters are presented in Table 10. We utilize the AdamW optimizer, which is configured with a cosine annealing schedule as the learning policy. The initial base learning rate is set to 1.5×10^{-4}, and the AdamW optimizer is employed with hyperparameters $\beta = (0.9, 0.95)$. Additionally, we set the weight decay to 0.05 without dropout. We use the strategy of cosine learning rate decay, with 40 warm-up epochs. Unless otherwise specified, the pre-training of our vision language model consists of 800 epochs, executed on 2×2 NVIDIA A100 GPUs.

Concerning the downstream tasks of fine-tuning and linear probing on ImageNet, the hyperparameters are shown in Table 11 and Table 12.

Table 12. The linear probing hyperparameters.

	ViT-B/16	ViT-L/16
Training Epochs	90	50
Warmup Epochs	10	10
Optimizer	LARS	LARS
Base Learning Rate	0.1	0.1
Learning Rate Decay	Cosine	Cosine
Weight Decay	0.0	0.0
Batch Size	16,384	1,024