Attention Shifting to Pursue Optimal Representation for Adapting Multi-granularity Tasks

Gairui Bai, **Wei Xi***, **Yihan Zhao**, **Xinhui Liu** and **Jizhong Zhao**
School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
{grbai2018, zhaoyihan, liuxinhui}@stu.xjtu.edu.cn , {xiwei, zjz}@xjtu.edu.cn

Abstract

Object recognition in open environments, *e.g.*, video surveillance, poses significant challenges due to the inclusion of unknown and multi-granularity tasks (MGT). However, recent methods exhibit limitations as they struggle to capture subtle differences between different parts within an object and adaptively handle MGT. To address this limitation, this paper proposes a Class-semantic Guided Attention Shift (SegAS) method. SegAS transforms adaptive MGT into dynamic combinations of invariant discriminant representations across different levels to effectively enhance adaptability to multi-granularity downstream tasks. Specifically, SegAS incorporates a hardness-based Attention Part Filtering Strategy (ApFS) to dynamically decompose objects into complementary parts based on the object structure and relevance to the instance. Then, SegAS shifts attention to the optimal discriminant region of each part under the guidance of hierarchical class semantics. Finally, a diversity loss is employed to emphasize the importance and distinction of different partial features. Extensive experiments validate SegAS' effectiveness in multi-granularity recognition of three tasks.

1 Introduction

Semantic understanding of images stands as a highly regarded problem in computer vision, with the primary objective of precisely capturing objects' semantics without relying on manual annotations. In real-world scenarios, this problem becomes even more challenging, especially when dealing with recognition tasks that involve multiple and unknown granularity [Guo *et al.*, 2023; Wang *et al.*, 2020]. Such challenges are prominent in various fields [Liu *et al.*, 2023] such as autonomous driving or video surveillance. For example, the challenges in video surveillance extend beyond instance recognition and involve fine-grained tasks like identity recognition and occlusion recognition, as shown in Figure 1.

Self-supervised learning methods have garnered significant attention due to their outstanding generalization capabilities.

(a) Video surveillance **(b) Autonomous driving**

Figure 1: Multi-granularity object recognition tasks across two real-world scenarios: (a) video surveillance and (b) autonomous driving. (a) From left to right are instance recognition, fine-grained recognition, and occlusion recognition, respectively. (b) From top to bottom are occlusion and fine-grained recognition.

Instance-based discriminant methods [Chen *et al.*, 2020c; Caron *et al.*, 2021] are considered representative in this realm, focusing on learning consistent representations from different random augmentations of the same sample. Clustering-based methods [Guo *et al.*, 2022; Xu *et al.*, 2022] have incorporated hierarchical clustering to learn multi-granularity representations by deriving compact image representations that gather around corresponding granularity cluster centers. Building upon these foundational designs, recent researches [Choudhury *et al.*, 2021; Amir *et al.*, 2021] have explored parsing the object region to learn fine-grained representation, employing fixed hyperparameters for clustering internal features of objects. They have successfully enhanced performance in handling multi-granularity recognition tasks.

However, these previous methods have limitations in fine-grained discriminability and collaboration when faced with multi-granularity tasks. The limitation arising from these methods only relies on the consistency constraint, which contradicts the objective of achieving semantic differentiation among parts within an object. This contradiction results in a model that lacks the ability to capture subtle differences between different parts within the object and to handle tasks adaptively across granularities. Furthermore, parsing samples using fixed hyperparameters does not always provide the most reasonable way, leading to suboptimal learning of fine-grained representations.

*Corresponding author.

To address the problems, we propose a methodology that converts adaptive multiple-granularity representation learning into the acquisition of discriminative representations invariant across various levels. These representations are then dynamically combined to effectively handle tasks spanning multiple granularities. This idea is inspired by the recognition of the nested and complementary relationship between coarse- and fine-grained representations. Consequently, we propose the class-Semantic Guided Attention Shift (SegAS) method. SegAS considers the discriminative potential of different parts within an object by shifting the model's attention to different parts. Nonetheless, this approach encounters two noteworthy challenges. Firstly, the inherent variations in distribution and semantics among object parts cannot be ignored. Sole reliance on instance-level semantic constraints risks introducing representation bias. Secondly, focusing on different parts and learning important information involves trade-offs with computational resources.

SegAS introduces prototypes (*i.e.,* the class-wise cluster centers) to assist in addressing these challenges. SegAS comprises three components. Firstly, we propose Dual Siamese Networks to reconcile the contradiction between instance consistent and partial differential, to reduce confusion and mitigate representation bias. Secondly, we present Prototype-based Consistency Regularization (ProCR) to supervise representation learning for discriminative feature acquisition. This regularization ensures the alignment of the distribution of instance-prototype relationships while relaxing the constraints. Moreover, we calibrate the distribution by considering the hardness of each sample and its relationship with the prototype. Finally, we propose a hardness adaptive Attention-part Filtering Strategy (ApFS) to generate views that possess independent and complementary features relative to the original image. This strategy is equipped with a diversity loss to emphasize the importance and dissimilarity of different part representations. This strategy restricts the information input to the object, forcing the model's attention to shift to key regions under the supervision of semantic consistency while reducing the computational burden.

The contributions of this paper are concluded as follows:

1. We propose a Class-semantic Guided Attention Shift (SegAS) method, which uses dual siamese networks to address adaptability to open-granularity downstream tasks. SegAS achieves this by incorporating an attention-part filtering strategy, which directs the model's attention towards the multi-key parts within objects while minimizing computational costs.

2. We propose a prototype-based consistency regularization to eliminate representation bias caused by partial semantic differences within objects. This regularization approach encourages the model to learn the optimal discriminant representation by regulating the distribution of relationships with the prototype set.

3. Experiments demonstrate that SegAS exhibits significant improvements in some tasks, such as occluded image recognition, fine-grained classification, and object localization. Moreover, our method is shown to enhance the quality of representations through common downstream tasks that are used to verify the effectiveness of self-supervised learning.

2 Related Work

2.1 Self-Supervised Representation Learning

Contrastive learning (CL) is an effective self-supervised representation learning (SSL) method. NPID++[Wu *et al.*, 2018], SimCLR [Chen *et al.*, 2020a] and SimCLR v2 [Chen *et al.*, 2020b] are successful end-to-end models that provide simple frameworks for contrastive learning of visual representations. With the potential issue of having larger batch sizes, one solution is to maintain a separate dictionary called Memory Bank, such as PIRL [Misra and Maaten, 2020]. MoCo [He *et al.*, 2020] and MoCo v2 [Chen *et al.*, 2020c] are the representatives of the kind of method that uses the Momentum Encoder to solve the problem. Other methods improve the performance of the model from different perspectives, such as InfoMin [Tian *et al.*, 2020]. Debiased [Chuang *et al.*, 2020], AdCo[Hu *et al.*, 2021] and InsLoc[Yang *et al.*, 2021]. Some methods have invariant mapping but do not use negative samples *e.g.*, BYOL [Grill *et al.*, 2020] and SimSiam [Chen and He, 2021].

Some previous works discover parts by using image reconstruction [Choudhury *et al.*, 2021], which propose an unsupervised approach to object part discovery and segmentation. Pre-trained Vision Transformer [Amir *et al.*, 2021] is typically able to find the parts of the most relevant object in a semantically consistent manner. PDiscoNet [van der Klis *et al.*, 2023] needed to leverage the class labels to learn the part representation. However, these methods are designed to solve single-grained tasks.

2.2 Representation Learning with Masked Images

Masking, as one of the simplest data transformation methods, is widely used in various data types. Image inpainting [Pathak *et al.*, 2016] is used as a pretext task in SSL. In recent years, inspired by the success of masking on the transformer in NLP [Devlin *et al.*, 2018], some transformer-based masking methods [He *et al.*, 2022; Caron *et al.*, 2021] have achieved success. Because of the high redundancy of images, masking some patches can greatly reduce the redundant information. This approach creates a challenging self-supervised task that improves the overall understanding of the image and representation performance. MAE [He *et al.*, 2022] and SimMIM [Xie *et al.*, 2022] use random masking to assist representation learning by predicting RGB values of raw pixels by direct regression performs. MaskFeat [Wei *et al.*, 2022] proposes to regress HOG features of the masked content and it uses manual features as supervised signals. MST [Li *et al.*, 2021] and AttMask [Kakogeorgiou *et al.*, 2022] use the attention maps to generate the masking. MSN [Assran *et al.*, 2022] leverages the idea of mask-denoising while avoiding pixel and token-level reconstruction with siamese structure.

3 The Proposed Method

This section introduces our proposed method, Class-semantic Guided Attention Shift (SegAS). SegAS is a self-supervised

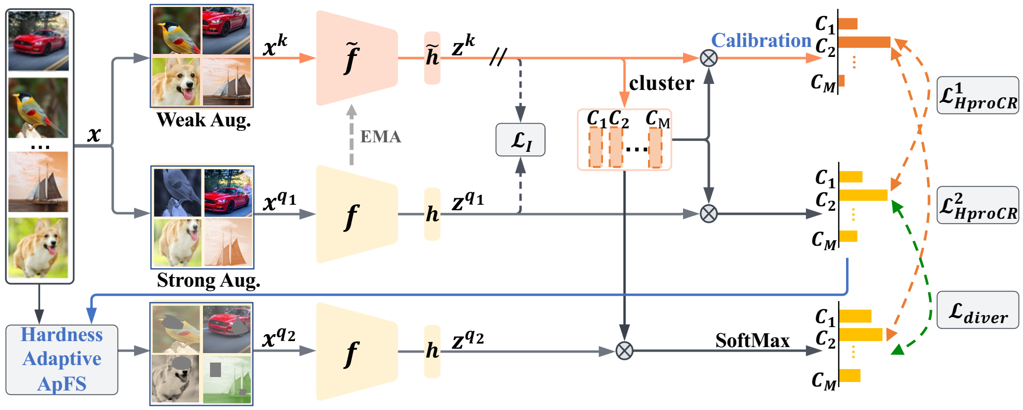

Figure 2: **The overall framework of our proposed SegAS.** SegAS comprises a feature extractor encoder f and a semantic projection head h. $\tilde{f}$ and $\tilde{h}$ are updated with an exponential moving average (EMA). Initially, the image x is transformed into three correlated views x^k, x^{q_1} and x^{q_2}, via three grained augmentations. x^{q_2} is the supplementary view after filtering out the attention part. They are employed as inputs to SegAS, producing embeddings z^k, z^{q_1} and z^{q_2}. Subsequently, z^k is used for clustering to generate the prototype set $\{c_1, c_2, \cdots, c_M\}$. Lastly, we compute the similarity between the embeddings and the prototypes separately, deriving the corresponding distribution using SoftMax. For model training, we employ the loss functions $\mathcal{L}_I, \mathcal{L}^1_{HproCR}, \mathcal{L}^2_{HproCR}$ and $\mathcal{L}_{diver}$. '//' indicates stop gradient. Please see more training details in the Proposed Method section.

approach designed to enhance representation generalization by learning discriminative representations of objects in multiple parts. The overall framework of SegAS is illustrated in Figure 2. The specific process is described in detail below.

3.1 Overview

Given a sample x_i, three data augmentation strategies are employed to produce distinct versions: x_i^k, $x_i^{q_1}$, and $x_i^{q_2}$. Among these, x_i^k can only be used as a target. $x_i^{q_1}$ and $x_i^{q_2}$ are complementary images generated through different strategies.

Initially, x_i^k, $x_i^{q_1}$ serve as inputs to the model to generate distinct embeddings z_i and $z_i^{q_1}$. The model learns how to represent objects by pulling different augmented versions of the same instance closer in an embedding space while pushing away different instances' augmentations. The process is achieved by optimizing a contrastive loss function, such as InfoNCE, defined as:

$$\mathcal{L}_{\mathrm{I}} = \sum_{i=1}^{n} -\log \frac{\exp\left(z_i \cdot z_i^{q_1}/\tau\right)}{\sum_{j=0}^{r} \exp\left(z_i \cdot z_j^{q_1}/\tau\right)}, \tag{1}$$

where z_j includes one same instance' embedding and r embedding for other instances. τ is a temperature hyperparameter. This method learns the instance-discriminant representation.

The above approach is based on Instance-discrimination. However, relying solely on the instance-level semantic constraints is prone to introducing representation bias. This is because the different parts within an object display noticeable distribution and semantic differences. Moreover, this method will inevitably induce a class collision problem [Saunshi *et al.*, 2019; Li *et al.*, 2020]. To address these challenges, we propose utilizing the distribution relationship between samples and class prototypes as a comprehensive and dependable constraint, rather than solely considering the similarity between instances. Consequently, we introduce a prototype-based distribution consistency regularization.

Specifically, the prototype is represented by the average of a set of sample features that exhibit similar semantic features. In this context, we denote the prototype set as $C = \{c_i\}_{i=1}^k$, where each c_i represents an individual prototype and k denotes the total number of prototypes. For a given target view x^k, its corresponding embedding is z^k. We then calculate the similarities between the feature and prototype, which can be evaluated using the $S(z^k, C)$. S is the metric function and we utilize cosine similarity. A softmax layer can be applied to process the calculated similarities:

$$p_i^k = \frac{\exp(S(z^k \cdot c_i)/\tau_k)}{\sum_{i'} \exp(S(z^k \cdot c_{i'})/\tau_k)}, \tag{2}$$

For the other two perspective images x^{q_1} and x^{q_2}, the distribution relationship between their embeddings z^{q_1} and z^{q_2} and the prototype set can be expressed as::

$$p_i^{q_j} = \frac{\exp(S(z^{q_j} \cdot c_i)/\tau_t)}{\sum_{i'} \exp(S(z^{q_j} \cdot c_{i'})/\tau_t)}, j \in \{1, 2\} \tag{3}$$

where τ_t is a different temperature parameter.

We suggest using prototype-based consistency regularization (ProCR) as the loss function. ProCR aims to minimize the Kullback-Leibler (KL) divergence between the prototypical assignments in two different views:

$$\mathcal{L}^j_{\text{ProCR}} = \mathcal{L}_{kl}(p_i^{q_j}, p_i^k), j \in \{1, 2\}. \tag{4}$$

To address the issue of recognizing multi-granularity in an open scene, we employ a hierarchical clustering approach. They serve as the self-supervised signals that enable the model to learn representations from coarse-grained to fine-grained, inspired by HCSC [Guo *et al.*, 2022] and HIRL [Xu *et al.*, 2022]. Hierarchical prototypes are used to guide the learning of image hierarchical semantics representations, facilitating the representation of different levels of granularity. It implements the K-means algorithm in a bottom-up way.

For a detailed description of the specific process, please refer to Appendix A.

Given the number of semantic levels is denoted as L, the prototype structure can be expressed as: $C = \left\{ \left\{ c_i^l \right\}_{i=1}^{M_l} \right\}_{l=1}^{L}$, where the M_l is the number of prototypes in the l-th hierarchy. Based on hierarchical prototypes, the hierarchical ProCR in the training process can be expressed as:

$$\mathcal{L}_{\text{HProCR}}^{j} = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{n} (\mathcal{L}_{\text{ProCR}}^{j})^{l}, j \in \{1, 2\}. \tag{5}$$

3.2 Distribution Calibration

The relationship between samples and prototypes is reflected in the relative probabilities assigned to different prototype classes. We aim to calibrate this relative probability to obtain a supervisory signal that contains more valid information.

Firstly, excessive prototype assignments for a sample may result in a dispersed probability distribution and unnecessary redundancy. It is advisable to decrease the distribution's entropy to enhance the precision. In simpler terms, by assigning a probability of 0 to easy negative prototypes, the overall probability distribution becomes more focused and concentrated. Specifically, the similarity between the sample and prototype set is $S(z^k, C)$. To estimate the pos/neg prototype, we use the mean μ_s of the similarity as a proxy measure. $\mu_s = \frac{1}{M} \sum_{i=1}^{M} \max\left(S({z^k}_j, c_i)\right)$, where M is the number of prototypes and B_U is the batchsize. We consider the similarity score less than μ_s as negative prototypes and the corresponding distance is set as -1:

$$S_{re}(z^k, c_i) = \begin{cases} -1, & \text{if } S(z^k, c_i) < \mu_s, \\ (z^k, c_i), & \text{otherwise.} \end{cases} \tag{6}$$

So, the target probability distribution can be rephrased as:

$$p_i^k = \frac{\exp(S_{re}(z^k \cdot c_i)/\tau_k)}{\sum_{i'} \exp(S_{re}(z^k \cdot c_{i'})/\tau_k)}, \tag{7}$$

where τ_k is the temperature parameter. This distribution serves to indicate the degree of similarity or match between the feature and each prototype within the set.

Secondly, relying solely on this soft distribution is not dependable, especially when x_i^k contains no or very few objects. During the random-sized crop process, various types of samples are generated, including easy foreground samples, hard foreground samples, and background samples. Treating all samples equally and applying a soft distribution can result in errors. To address this issue, we employ the concept of entropy to assess the difficulty of the samples. By utilizing the aforementioned distribution, we calculate a trade-off factor:

$$\rho_i^k = \max\left(p_i^k\right) \left(1 - \frac{-\sum_{j=0}^{M-1} p_i^{k(j)} \log p_i^k(j)}{\log M}\right), \tag{8}$$

where M is the number of prototypes. When the value of ρ_i is below the specified threshold, we simply utilize the one-hot labels assigned by image clustering. In all other scenarios, we employ the weighted sum of both approaches,

$$p_i^k = \begin{cases} y & \text{if } \rho_i < \tau \\ \rho_i \cdot p_i^k + (1 - \rho_i) \cdot y & \text{otherwise} \end{cases} \tag{9}$$

3.3 Attention Shifting

SegAS employs an attention-part filter to effectively filter out the previously most interesting local discriminative regions and the remaining regions are referred to as x^{q_2}. By applying the supervision of the ProCR, the model's attention is shifted towards the most discriminant region in x^{q_2}, facilitating the learning of an optimal representation.

Attention-part Filter Strategy. We design an Attention-part Filter Strategy(ApFS) to generate new augmentation of images, which is crucial for siamese representation learning. Firstly, we compute the heatmap A^c by aggregating the features of the last convolutional layer across the channel dimension and normalizing it to $[0, 1]$. Secondly, we apply a sigmoid function to filter out the most interesting region. Mathematically, this can be expressed as follows:

$$T\left(A^c\right) = 1 - \frac{1}{1 + \exp\left(-\omega\left(A^c - \boldsymbol{\sigma}\right)\right)}, \tag{10}$$

where $\boldsymbol{\sigma}$ is the threshold whose elements all equal to σ. ω is the scale parameter ensuring $T\left(A^c\right)_{i,j}$ approximately equals to 0 when $A_{i,j}^c$ is larger than σ, or to 1 otherwise.

The operator $T(A^c)$ is used on input x to generate an augmented view x_A. $x_A = T\left(A^c\right) \odot x$, where $\odot$ represents the element-wise multiplication. Additionally, SegAS incorporates Random Filtering to generate augmented samples x^{q_2}, ensuring the sampling is equiprobable.

Hardness Adaptive ApFS. The simulator described above produces an enhanced training sample. However, there are limitations to equally enhancing all samples, especially those that are difficult to train or in the early stages of training. The variation in sample learning difficulty arising from cropping, coupled with the inherent complexity differences among individual samples. Moreover, the initial training step also necessitates the need for easy samples. Therefore, we tend to strongly enhance the easy-to-learn samples in the mini-batch to improve the performance of the model. We first estimate a confidence score, ρ_i, which indicates the level of confidence the current model has in its prediction for the $i-$th instance,

$$\rho_i^{p_1} = \max\left(p_i^{p_1}\right) \left(1 - \frac{-\sum_{j=0}^{M-1} p_i^{p_1}(j) \log p_i^{p_1}(j)}{\log M}\right), \tag{11}$$

where we use the weighted average of the normalized prediction entropy on $p_i^{p_1}$ to estimate the confidence score. We employ ρ_i as a triggering probability to randomly apply the sample and obtain a candidate enhancement.

Diverse and Importance Learning. SegAS uses different augmentations to learn complementary features of the same instances. They learn the discriminant representations of each part under semantic consistent supervision. However, there are differences between these representations and their contributions to overall discriminability. Therefore, we introduce the diversity loss to measure the contribution of different part representations. The diversity loss is given by:

$$\mathcal{L}_{\text{diver}} = \sum_{i=1}^{n} \max\left\{0, p_i^{q_2}(c) - p_i^{q_1}(c) + m\right\} + z_i^{q_1} \cdot z_i^{q_2}, \tag{12}$$

where $p_i^{q_1}(c)$ and $p_i^{q_2}(c)$ denotes the prediction probability on the most relevant prototype c. m is the threshold. Intuitively, this loss leads to the first learning of discriminatory representations being closer to the relevant prototype than filtered representations. It measures the contribution of different representations to discriminant quality. The second part of the loss is to avoid overlapping between different representations.

We train the self-supervised representation learning model with the following total loss:

$$\mathcal{L}_{total} = \mathcal{L}_I + \alpha\mathcal{L}^1_{\text{HProCR}} + \beta\mathcal{L}^2_{\text{HProCR}} + \lambda\mathcal{L}_{\text{diver}}, \quad (13)$$

where α, β and λ are the coefficient to balance these loss. In the experiment, $\alpha = \beta$.

4 Experiment

We evaluated the performance of SegAS in various experiments, including occlusion recognition, object detection, and fine-grained recognition. Specifically, we performed these experiments on ImageNet-100 [Russakovsky *et al.*, 2015], Pascal VOC [Everingham *et al.*, 2010], Place 205 [Zhou *et al.*, 2014], COCO [Lin *et al.*, 2014] datasets, and CUB-200-2011 [Welinder *et al.*, 2010]. For all experiments, the reported results represent the average performance over five runs. In addition, we also conducted ablation experiments on the ImageNet-100 dataset to verify the effectiveness of the proposed components. In this manuscript, we used MoCo v2 as a baseline method to make a fair comparison.

4.1 Implementation Details

For data augmentation, the weak augmentation only consists of random crops and horizontal flips. The contrastive augmentation involves random resized crops, color distortion (strength=0.5), flipping, and Gaussian blur.

In the training process, the backbone used the ResNet-50 [He *et al.*, 2016]. The model was trained using SGD [Robbins and Monro, 1951] optimizer with a weight decay of 1×10^{-4} and momentum of 0.9. The temperature parameter τ was always set to 0.2. The total epoch was set as 200. In ImageNet-1k, the number of semantic levels was defined as $L = 3$ and $(M_1, M_2, M_3) = (30000, 10000, 1000)$, details are in appendix B.

4.2 Representation Performance under Occlusion

To assess the performance of SegAS on occluded objects, we employed different masking strategies on images from the ImageNet-100 dataset to simulate different occlusions and compare the results against the baseline method.

Evaluation Setup. During the evaluation process, the parameters of the feature extractor were kept fixed, and an FC layer was trained as the classifier. The classifier was trained using SGD, with a total of 60 epochs, an initial learning rate of 5.0, and a step learning rate schedule that drops at epochs 30, 40, and 50.

Random Occlusion

During the evaluation process using the filtering strategy, MoCo v2 was selected as the baseline method. All methods were pre-training on ImageNet-100. The comparison results are presented in Table 1, and SegAS achieved a performance of 82.33% on occluded images. Notably, compared to previous methods that did not incorporate filtering, our method demonstrated little change in performance after occlusion. These results suggest that the incorporation of a filtering strategy and alignment of feature distribution can lead to the learning of additional information.

Method	Epochs	Accuracy	
		complete	occlusion
MoCo V2	200	78.0	70.76
w/o ApFS	200	82.23	78.94
SegAS	200	**83.94**	**83.21**

Table 1: **Performance comparison on the occluded image.** We report Top-1 accuracy on ImageNet-100 with random occlusion.

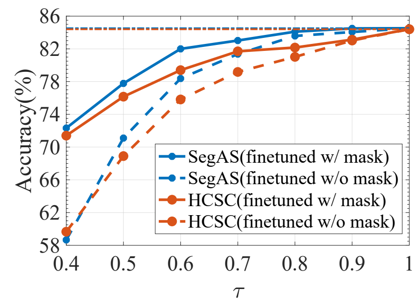

Figure 3: Effect of varying filtering threshold (τ).

Attention Occlusion

To further verify the effectiveness of SegAS, we assessed its capability to learn discriminant representations under disturbances to the local region of interest. Based on the feature map, we masked the region of interest using a pre-determined threshold value. The parameter settings for this experiment align with those used in the random mask experiment. In this experiment, we evaluated the impact of the local mask on our method by varying the threshold and local area masked. We set the threshold within the range $[0.4 - 1]$ and used the HCSC method as the baseline for comparison. This experiment adopted the pre-trained model on ImageNet-1K, with the model parameters of HCSC taken from the original paper.

The results shown in Figure 3 reveal that when the threshold was set to a large value, and the local area masked was small, our method exhibited less performance reduction compared to the baseline method. This suggests that the local mask has little impact on the performance of our method under such conditions. However, when the threshold was set to a value less than or equal to 0.4, it became challenging to identify the object. This is because most of the area is masked, blocking all discriminative regions.

4.3 Transfer Learning

Fine-grained Classification

We evaluated SegAS on the CUB-200-2011 dataset, specifically for the task of fine-grained classification. The model was pre-trained on ImageNet-1K, and the parameters of the feature extractor were kept fixed during fine-tuning. We fine-tuned the model with the training set and evaluated the per-

Method	CUB-200-2011			ImageNet-100		
	Top-1 Clas	GT-Known Acc	Top-1 Loc	Top-1 Clas	GT-Known Acc	Top-1 Loc
MoCo V2	19.1	52.2	12.36	74.64	61.46	47.74
PCL V2	20.73	64.0	15.86	78.26	57.78	46.92
HCSC	20.28	62.69	15.53	84.40	60.6	53
SegAS	**29.70**	**67.1**	**23.92**	**84.52**	**65.72**	**57.3**

Table 2: Quantitative evaluation results (%) on CUB-200-2011 and ImageNet-100.

Method	Object Classification		Object Detection	
	VOC07 mAP	Place205 Top1 Acc	VOC07+12 AP_{50}	COCO AP
NPID++	64.6	38.7	-	-
SimCLR	86.4	-	-	-
MoCo V1	79.2	48.9	81.1	-
MoCo V2	84.0	50.1	82.4	40.6
PCL V2	85.4	50.3	78.5	41.0
AdCo	92.0	51.1	82.6	41.2
BYOL	-	-	81.0	40.3
InsLoc	-	-	82.9	41.4
HCSC	92.8	52.2	82.5	41.4
SegAS	**93.1**	**53.0**	**82.93**	**42.1**

Table 3: Performance comparison on Transfer Learning.

formance with the testing set, using the Top-1 class accuracy as the performance metric.

In Table 2, we present a detailed comparison of the results on the CUB-200-2011 dataset, including methods such as MoCo V2, PCL V2, and HCSC. All three methods are based on the MoCo V2 framework. Our method achieved a classification performance of 29.70%, which is 9.42% higher than the previous best performance of HCSC at 20.28%. This indicates that more discriminant regions generate more robust and generalized representations in different tasks.

Object Location

We evaluated the performance of SegAS for object localization tasks on two datasets: CUB-200-2011 and ImageNet-100. We also kept the parameters of the feature extractor fixed during the fine-tuning process. For ImageNet-100, we fine-tuned the model using the training set and evaluated the performance on the validation set. The settings were the same as the above experiment for CUB-200-2011. We used several metrics to assess the performance. The top-1 class is used to metric the performance for classification. The ground-truth class (GT-known Loc) and Top-1 Loc to metric the performance for unsupervised object detection. GT-known Loc measures the location accuracy by determining the correctness based on the intersection over union (IoU) between the ground truth bounding box and the estimated box for the ground truth class. It is considered correct when the IoU is 50% or higher. Top-1 Loc is considered correct when both GT-known Loc and Top-1 Class are determined to be correct.

We conducted a comprehensive comparison of the proposed SegAS method with several recent contrastive learning techniques. The quantitative evaluation results are presented in Table 2. The results demonstrate that SegAS exhibits strong performance in both classification and object localization tasks on both the ImageNet-100 and the more challenging CUB-200-2011 datasets. In terms of classification accuracy, SegAS achieves an accuracy of 84.52% on the ImageNet-100, which is a slight improvement of 0.12% compared to the previous method. Notably, SegAS outperforms the highest-performing method in terms of localization accuracy by 4.26%.

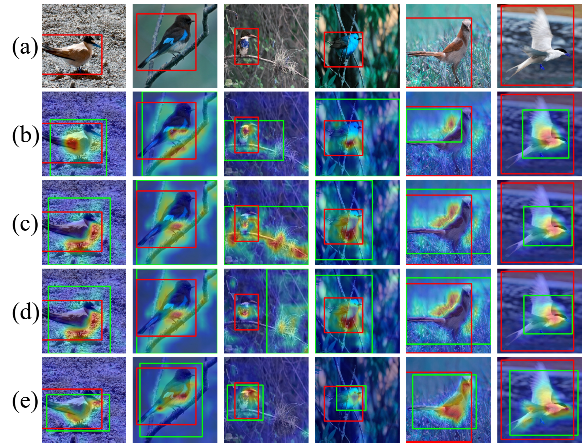

Figure 4: Comparison of localization results from the vanilla method and our method on CUB-200-2011 datasets. Red boxes denote the ground truth bounding boxes and green boxes denote the predicted bounding boxes. From the 1st to the 5th row: (a) Original images, (b) MoCo v2, (c) PCL v2, (d) HCSC, and (e) SegAS.

Furthermore, SegAS also has a localization accuracy of 67.1% on CUB-200-2011. This demonstrates the effectiveness of our method in accurately localizing objects within fine-grained datasets. the experimental results indicate that our method not only focuses on the optimal regions of an object but also considers other representative areas.

Visualization. We visualize the results of the class activation map (CAM) [Zhou *et al.*, 2016] and localization on CUB-200-2011, as shown in Figure 4. Red boxes represent the ground-truth bounding boxes, while green boxes represent the predicted boxes. SegAS performs remarkably well in comparison to existing methods. It closely approximates the ground-truth, even when there is interference nearby, enabling accurate localization of the object. These visualizations provide evidence that SegAS possesses the ability to identify the multiple discriminative regions within the object.

Object Detection

We learned representations on the ImageNet-1k and fixed the parameter of the feature extractor. We evaluated transfer learning performance across 3 natural image datasets (Place 205 [Zhou *et al.*, 2014], PASCAL VOC [Everingham *et al.*, 2010] and COCO [Lin *et al.*, 2014]) in linear classification and object detection. The fine-tuning paradigms on these two types of tasks completely follow those in MoCo, details are in Appendix C.

Table 3 provides a comparison of our method with other

Method	Accur
	w/o mask
BaseLine($\mathcal{L}_I$)	78.0
w/o filtering	
$+\mathcal{L}_{\text{HprotoNCE}}$	80.28
$+\mathcal{L}^1_{\text{HProCR}}$	82.23
w/ filtering	
$+\mathcal{L}^1_{\text{HprotoNCE}} + \mathcal{L}^2_{\text{HprotoNCE}}$	81.50
$+\mathcal{L}^1_{\text{HProCR}} + \mathcal{L}^2_{\text{HProCR}}$	83.52
$+\mathcal{L}^1_{\text{HProCR}} + \mathcal{L}^2_{\text{HProCR}} + \mathcal{L}_{\text{diver}}$	**83.94**

Table 4: Ablation study of each proposed cor

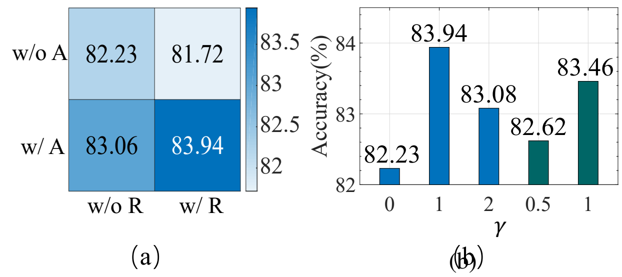

Figure 5: Ablation study. (a) different filtering strategies. 'A' represents Attention-part Filtering, and 'R' is Random. (b) hyper-parameter. The blue bar: $\alpha = 0.5$, green bar:$\alpha = 1$.

contrastive methods in transfer learning. For linear classification, our method achieved a higher 93.1% mAp on PASCAL VOC and 53.0% Top-1 accuracy on Place205 with 200 epochs of pre-training, outperforming the state-of-the-art HCSC model (92.8% and 52.2%). For the object detection task, SegAS improved significantly to 82.93% and 42.1% on PASCAL VOC and COCO datasets, respectively, compared to the previous best performance. These results demonstrate that our method outperforms other models pre-trained on ImageNet-1K and has better generalization ability for different downstream tasks.

4.4 Ablation Study

In this section, more experiments are presented to evaluate the effectiveness of our proposed module on ImageNet-100.

Effectiveness of Prototype-based Consistency Regularization. To enhance representation performance, SegAS proposes ProCR instead of directly bringing the sample closer to the prototype. This strategy has been implemented in two distinct phases. The results presented in Table 4, the $\mathcal{L}_{\text{HprotoNCE}}$ is a loss function optimized to directly assign a fixed prototype to the sample, the same as in HCSC. The ablation study demonstrated that the representation of SegAS has significant improvements in both masked and un-masked images during the first phase using ProCR, achieving 82.23% and 78.94%, respectively, compared to use $\mathcal{L}_{\text{HprotoNCE}}$. Moreover, introducing the attention-part filtering strategy further enhanced the discriminative capability of the representation. Specifically, the distributed alignment method resulted in 83.52% and 82.34% improvements in both unoccluded and occluded images. The experimental findings suggest that SegAS enables the identification of multiple discriminant regions within objects based on prior learning, leading to an overall enhancement of the discriminative representation, regardless of the level of occlusion in the images.

Evaluation on Different Filtering Strategies. We conducted an ablation study on various filtering strategies that are crucial in SegAS. The results, as illustrated in Figure 5a, demonstrate that the highest linear evaluation accuracy is achieved when both random and ApFS strategies are employed simultaneously on ImageNet-100. Notably, using only the attention filtering strategy yields better accuracy compared to using only the random filtering strategy. These experimental results indicate that the ApFS can assist the model in selecting the optimal discriminant alternative region, which aligns well with our initial motivation.

Evaluation on Varied Hyper-parameters. In this study, we investigate the impact of adjusting the hyperparameters α, β, and λ in Equation (13). Firstly, we explore the effects of fixing $\lambda = 0.1$ to and tuning α from 0.5 to 1. Additionally, we set $\beta = \gamma * \alpha$ in our experiment, as shown in Figure 5b. Our findings reveal that our method achieves optimal performance when $\alpha = 0.5$, $\beta = 0.5$, and $\lambda = 0.1$ on the ImageNet-100 dataset. Secondly, we fix α and β at 0.5 and analyze the results when λ is either 0 or 0.1. In the final row of Table 4, it is demonstrated that the performance of the system increased by 83.94% at $\lambda = 0.1$. This finding provides evidence of the effectiveness of $\mathcal{L}_{\text{diver}}$ in improving performance.

Evaluation of Efficiency. To verify the efficiency of our proposed method, we conducted experiments on four NVIDIA-GeForce-RTX 3090. On the ImageNet100 dataset, SegAS trained 200 epochs 0.5 days longer than MoCo v2. Although slightly longer than MoCo v2, this is offset by notable gains in efficiency and performance.

5 Conclusion and Future Work

In this paper, we propose SegAS, a novel self-supervised learning method, that aims to learn the discriminative representation of different parts of objects through dynamic attention shifting. SegAS incorporates prototype-based consistency regularization to facilitate semantically consistent alignment of models. Furthermore, SegAS employs a hardness adaptive attention-part filtering strategy to generate a supplementary view and then re-guides the model's attention shift to other discriminant regions via consistency regularization constraints. Extensive experimental evaluations demonstrate that the learned representation of SegAS exhibits strong discriminability and generalization capabilities across various downstream tasks. In future endeavors, SegAS can be adapted to leverage the transformer architecture, enabling the handling of multi-grained tasks in open scenarios with minimal overhead.

Acknowledgements

The project was supported by the National Key R&D Program of China (2022YFF0901800), NSFC (62072367, 62176205).

References

[Amir *et al.*, 2021] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[Assran *et al.*, 2022] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision*, pages 456–473. Springer, 2022.

[Caron *et al.*, 2021] 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*, pages 9650–9660, 2021.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

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

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020.

[Chen *et al.*, 2020c] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[Choudhury *et al.*, 2021] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34:28104–28118, 2021.

[Chuang *et al.*, 2020] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Everingham *et al.*, 2010] 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(2):303–338, 2010.

[Grill *et al.*, 2020] 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.

[Guo *et al.*, 2022] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9706–9715, 2022.

[Guo *et al.*, 2023] Baoshen Guo, Weijian Zuo, Shuai Wang, Xiaolei Zhou, and Tian He. Attention enhanced package pick-up time prediction via heterogeneous behavior modeling. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 189–208. Springer, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[He *et al.*, 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*, pages 9729–9738, 2020.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[Hu *et al.*, 2021] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021.

[Kakogeorgiou *et al.*, 2022] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *Proceedings of the European Conference on Computer Vision*, pages 300–318. Springer, 2022.

[Li *et al.*, 2020] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2021] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2023] Shenghao Liu, Guoyang Wu, Xianjun Deng, Hongwei Lu, Bang Wang, Laurence Yang, and James J Park. Graph sampling based fairness-aware recommendation over sensitive attribute removal. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 428–437. IEEE, 2023.

[Misra and Maaten, 2020] 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*, pages 6707–6717, 2020.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Russakovsky *et al.*, 2015] 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(3):211–252, 2015.

[Saunshi *et al.*, 2019] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

[Tian *et al.*, 2020] 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.

[van der Klis *et al.*, 2023] Robert van der Klis, Stephan Alaniz, Massimiliano Mancini, Cassio F Dantas, Dino Ienco, Zeynep Akata, and Diego Marcos. Pdisconet: Semantically consistent part discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1866–1876, 2023.

[Wang *et al.*, 2020] Fangxin Wang, Jiangchuan Liu, and Wei Gong. Multi-adversarial in-car activity recognition using rfids. *IEEE Transactions on Mobile Computing*, 20(6):2224–2237, 2020.

[Wei *et al.*, 2022] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[Xie *et al.*, 2022] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[Xu *et al.*, 2022] Minghao Xu, Yuanfan Guo, Xuanyu Zhu, Jiawen Li, Zhenbang Sun, Jian Tang, Yi Xu, and Bingbing Ni. Hirl: A general framework for hierarchical image representation learning. *arXiv preprint arXiv:2205.13159*, 2022.

[Yang *et al.*, 2021] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021.

[Zhou *et al.*, 2014] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.