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.

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 [Aasran et al., 2022] leverages the idea of mask-denoising while avoiding pixel and token-level reconstruction with siamese structure.
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:

\[
\mathcal{L}_1 = \sum_{i=1}^{n} \log \frac{\exp(z_i \cdot z_i^{q_1}/\tau)}{\sum_{j=0}^{\mathcal{M}} \exp(z_i \cdot z_j^{q_1}/\tau)},
\]

where \( z_j \) includes one same instance’ embedding and \( \mathcal{M} \) for other instances. \( \tau \) is a temperature hyper-parameter. 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 \( x_i^k \), its corresponding embedding is \( z_i^k \). We then calculate the similarities between the feature and prototype, which can be evaluated using the \( S(z_i^k, C) \). \( S \) is the metric function and we utilize cosine similarity. A softmax layer can be applied to process the calculated similarities:

\[
\rho_i^k = \frac{\exp(S(z_i^k \cdot c_i)/\tau_k)}{\sum_{i'}\exp(S(z_i^k \cdot c_{i'}/\tau_k))},
\]

For the other two perspective images \( x_i^{q_1} \) and \( x_i^{q_2} \), the distribution relationship between their embeddings \( z_i^{q_1} \) and \( z_i^{q_2} \) and the prototype set can be expressed as:

\[
\rho_i^{q_j} = \frac{\exp(S(z_i^{q_j} \cdot c_i)/\tau_j)}{\sum_{i'}\exp(S(z_i^{q_j} \cdot c_{i'}/\tau_j))}, j \in \{1, 2\}
\]

where \( \tau_j \) 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:

\[
\mathcal{L}^{\text{ProCR}}_i = \mathcal{L}_\text{kl}(\rho_i^{q_1}, \rho_i^k), j \in \{1, 2\}.
\]

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: \( C = \{ \{ c_i \}_{i=1}^{M_l} \}_{l=1}^{L} \), 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:

\[
\mathcal{L}^{j}_{j}^{\text{ProCR}} = \frac{1}{T} \sum_{l=1}^{L} \sum_{i=1}^{n} (\mathcal{L}^{j}_{j}^{\text{ProCR}})^{l}, j \in \{1, 2\}.
\]

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 \( \mu_s \) of the similarity as a proxy measure.

\[
\mu_s = \frac{1}{M} \sum_{i=1}^{M} \max (S(z^k, c_i)), \text{ where } M \text{ is the number of prototyes and } B_{U} \text{ is the batchsize}. \]

We consider the similarity score less than \( \mu_s \) as negative prototypes and the corresponding distance is set as \(-1\):

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

So, the target probability distribution can be rephrased as:

\[
p_i^k = \frac{\exp(S_{re}(z^k \cdot c_i) / \tau_k)}{\sum_{c'} \exp(S_{re}(z^k \cdot c_i') / \tau_k)},
\]

where \( \tau_k \) 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 \( x^k \) 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 a dispersed probability distribution and unnecessary entropy. 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, \( \rho_i \), which indicates the level of confidence the current model has in its prediction for the \( i \)-th instance.

\[
\rho_{i}^{p1} = \max \left( p_i^{p1} \right) \left( 1 - \frac{-\sum_{j=0}^{M-1} p_i^{p1}(j) \log p_i^{p1}(j)}{\log M} \right),
\]

where \( p_i \) is the diversity loss to measure the contribution of different part representations. The diversity loss is given by:

\[
\mathcal{L}_{\text{diverse}} = \sum_{i=1}^{n} \max \{0, p_i^{q1}(c) - p_i^{q2}(c) + m \} + z_i^{q1} \cdot z_i^{q2}.
\]
where \( p^{(c)}_i \) and \( p^{(c)}_j \) 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:

\[
L_{\text{total}} = L_I + \alpha L_{\text{HProCR}} + \beta L_{\text{HProCR}}^2 + \lambda L_{\text{diver}},
\]

where \( \alpha, \beta \) and \( \lambda \) 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 \( 1 \times 10^{-4} \) and momentum of 0.9. The temperature parameter \( \tau \) was always set to 0.2. The total epoch was set to 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.

#### 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-

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>Accuracy complete</th>
<th>occlusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoCo V2</td>
<td>200</td>
<td>78.0</td>
<td>70.76</td>
</tr>
<tr>
<td>w/o ApFS</td>
<td>200</td>
<td>82.23</td>
<td>78.94</td>
</tr>
<tr>
<td>SegAS</td>
<td>200</td>
<td><strong>83.94</strong></td>
<td><strong>83.21</strong></td>
</tr>
</tbody>
</table>

Table 1: Performance comparison on the occluded image. We report Top-1 accuracy on ImageNet-100 with random occlusion.
Table 2: Quantitative evaluation results (%) on CUB-200-2011 and ImageNet-100.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUB-200-2011</th>
<th>ImageNet-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 Class</td>
<td>Top-1 Loc</td>
</tr>
<tr>
<td></td>
<td>GT-Known Acc</td>
<td></td>
</tr>
<tr>
<td>MoCo V2</td>
<td>19.1</td>
<td>12.36</td>
</tr>
<tr>
<td>PCL V2</td>
<td>20.73</td>
<td>15.86</td>
</tr>
<tr>
<td>HCSC</td>
<td>20.28</td>
<td>15.53</td>
</tr>
<tr>
<td>SegAS</td>
<td>29.70</td>
<td>23.92</td>
</tr>
</tbody>
</table>

Table 3: Performance comparison on Transfer Learning.

<table>
<thead>
<tr>
<th>Method</th>
<th>Object Classification</th>
<th>Object Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VOC07 mAP</td>
<td>Place205 Top1 Acc</td>
</tr>
<tr>
<td>NPID++</td>
<td>64.6</td>
<td>38.7</td>
</tr>
<tr>
<td>SimCLR</td>
<td>86.4</td>
<td>-</td>
</tr>
<tr>
<td>MoCo V1</td>
<td>79.2</td>
<td>48.9</td>
</tr>
<tr>
<td>MoCo V2</td>
<td>84.0</td>
<td>50.1</td>
</tr>
<tr>
<td>PCL V2</td>
<td>85.4</td>
<td>50.3</td>
</tr>
<tr>
<td>AdCo</td>
<td>92.0</td>
<td>51.1</td>
</tr>
<tr>
<td>BYOL</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>InsLoc</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HCSC</td>
<td>92.8</td>
<td>52.2</td>
</tr>
<tr>
<td>SegAS</td>
<td>93.1</td>
<td>53.0</td>
</tr>
</tbody>
</table>

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.
To enhance representation performance, SegAS proposes ProCR instead of directly bringing the sample closer to the prototype. This strategy has been implemented in two distinctive phases. The results presented in Table 4, the $L_{\text{protoNCE}}$ 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 $L_{\text{protoNCE}}$. Moreover, introducing the attention-part filtering strategy further enhances 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 $\alpha$, $\beta$, and $\lambda$ in Equation (13). Firstly, we explore the effects of fixing $\lambda = 0.1$ and tuning $\alpha$ from 0.5 to 1. Additionally, we set $\beta = \gamma \cdot \alpha$ 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 $\alpha$ and $\beta$ at 0.5 and analyze the results when $\lambda$ 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 $L_{\text{diver}}$ 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.

![Table 4: Ablation study of each proposed contrastive method 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.](image)

![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: $\alpha = 1$.](image)
Acknowledgements

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

References


[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.


[Li et al., 2021] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu,


