Object Adaptive Self-Supervised Dense Visual Pre-Training

Yu Zhang, Tao Zhang, Hongyuan Zhu, *Member, IEEE*, Zihan Chen, Siya Mi, Xi Peng, *Senior Member, IEEE*, and Xin Geng, *Senior Member, IEEE*

***Abstract*—Self-supervised visual pre-training models have achieved significant success without employing expensive annotations. Nevertheless, most of these models focus on iconic single-instance datasets (e.g. ImageNet), ignoring the insufficient discriminative representation for non-iconic multi-instance datasets (e.g. COCO). In this paper, we propose a novel Object Adaptive Dense Pre-training (OADP) method to learn the visual representation directly on the multi-instance datasets (e.g., PASCAL VOC and COCO) for dense prediction tasks (e.g., object detection and instance segmentation). We present a novel object-aware and learning-adaptive random view augmentation to focus the contrastive learning to enhance the discrimination of object presentations from large to small scale during different learning stages. Furthermore, the representations across different scale and resolutions are integrated so that the method can learn diverse representations. In the experiment, we evaluated OADP pre-trained on PASCAL VOC and COCO. Results show that our method has better performances than most existing state-of-the-art methods when transferring to various downstream tasks, including image classification, object detection, instance segmentation and semantic segmentation.**

***Index Terms*—Dense visual pre-training, contrastive learning, multi-scale representation.**

I. Introduction

RECENTLY, self-supervised learning [1], [2], [3], [4] shows promising performance to pre-train models [5] without using labels. In self-supervised learning, the critical step is to maximize the consistency of representations from different augmentations of the same instance. These models can be later transferred to a series of downstream tasks, e.g., image classification [6], object detection [7] and semantic segmentation [8].

Received 3 October 2023; revised 9 October 2024; accepted 16 March 2025. Date of publication 1 April 2025; date of current version 7 April 2025. The work of Hongyuan Zhu was supported by the Economic Development Board (EDB) Space Technology Development Program under Project S22-19016-STDP. The associate editor coordinating the review of this article and approving it for publication was Dr. Sebastian Bosse. *(Yu Zhang and Hongyuan Zhu contributed equally to this work.) (Corresponding author: Yu Zhang.)*

Yu Zhang, Tao Zhang, Zihan Chen, and Xin Geng are with the School of Computer Science and Engineering and the Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China (e-mail: zhang_yu@seu.edu.cn).

Hongyuan Zhu is with the Institute for Infocomm Research (I2R), A*STAR, Singapore 138632.

Siya Mi is with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China, and also with Purple Mountain Laboratories, Nanjing 211111, China.

Xi Peng is with the School of Computer Science, Sichuan University, Chengdu 610065, China.

Digital Object Identifier 10.1109/TIP.2025.3555073

Despite their success on iconic single-instance datasets like ImageNet [9], it is still a question whether these methods can be directly applied to non-iconic multi-instance datasets, such as PASCAL VOC [10] and COCO [11], which are commonly used for dense prediction tasks. For iconic datasets, a single instance often occupies a large proportion of the image, making it easier to enforce feature consistency of the same instance across different views. However, non-iconic datasets contain multiple instances at different scales, making it more challenging than single-instance cases. In many downstream tasks like Object Detection and Instance Segmentation, scenarios involving multiple instances in a single image are more complex. These complexities arise from multiple factors, such as the presence of multiple objects, varying scales, and different resolutions, which can influence the effect of methods originally designed for single-instance scenarios. Specifically, multiple-object situations can lead to overlapping regions or ambiguities between similar objects, while varying scales and resolutions introduce challenges in maintaining consistent representations across views. These issues may result in the inability of existing single-instance methods to effectively model complex object relationships or to accurately capture fine-scale or localized object instances. Some self-supervised models [3], [4] employ contrastive learning of local representations during the pre-training process. Nonetheless, they conduct learning at a fixed scale, which ignores contexts at different scales and resolutions, limiting their applicability and robustness across diverse downstream tasks.

In this paper, we propose a novel Object-Adaptive Dense Pre-training (OADP) method that can deal with multiple instances. Specifically, we propose a novel and efficient augmentation method with random resized cropping whose window is object-aware and dynamically decreased to form a prominent object pool. Then a novel cross-scale and cross-resolution constratisve learning is introduced to make the model more sensitive to the small and medium objects. Compared to single-object single-scale methods, our method is specifically designed to handle multiple instances and scales, making it more broadly applicable and effective in complex, real-world scenarios.

The experiments pre-training on non-iconic multi-instance datasets demonstrate that our method can achieve better performance than state-of-the-art self-supervised models. For COCO pre-training model, our proposed OADP outperforms CCOP and baseline BYOL by 0.9% AP and 3.0% AP in

PASCAL VOC object detection task. Meanwhile, OADP improves ORL and BYOL by 0.2% AP and 1.1% AP for COCO object detection tasks and 0.1% AP and 0.8% AP for COCO instance segmentation tasks. For PASCAL VOC and Cityscapes semantic segmentation tasks, OADP is also significantly better than BYOL by 1.9% and 1.8% mIoU. In the experiments of PASCAL VOC pre-training model, we reach a 66.8% top-1 accuracy for PASCAL VOC image classification task, surpassing by 2.1% top-1 accuracy over BYOL. These results strongly demonstrate that the proposed paradigm can boost various downstream tasks.

Overall, our main contribution can be summarized as follows:

1. We propose a object-aware and learning-adaptive self-supervised pre-training framework called OADP for pre-training on multi-instance datasets.
2. We present a novel local and cross-scale and resolution contrastive learning with object-aware and learning-adaptive augmentation to improve feature discrimination at small and medium scale objects.
3. Extensive results on multi-instance datasets validate the effect of OADP on various downstream tasks.

II. Related Works

A. Self-Supervised Learning on Iconic Single-Instance Datasets

To extract a surrogate supervision signal in the form of a pretext task, traditional self-supervised learning (SSL) models utilize the input data on single-centric-object datasets, such as ImageNet [9]. These models are later transferred for a series of downstream tasks, e.g., image classification, object detection and semantic segmentation. The learning objective is to learn the visual representations by distinguishing instance discrimination from input augmentations. Typical augmentation methods consist of rotation prediction [12], image colorization [13], solving Jiasaw Puzzles [14], image inpainting [15] and so on. They consider two augmented views of the same image as the positive pair. Correspondingly, the negative pair is regarded as augmented views from different images, which is not essential for all the SSL models.

MoCo [16] employs a momentum encoder instead of the trained network to maintain consistent representations of negative pairs from a memory bank. SwAV [17] proposes a multi-crop paradigm, learning the visual representations from the clusters of augmented views rather than directly learning the features from the discriminations of augmentations. Meanwhile, SimCLR [1] introduces a learnable nonlinear transformation between the representation and the contrastive loss and replaces the usage of the memory bank. SimSiam [18] put forward a method namely "stop-gradient" to address the issue of the model Collapsing in self-supervised learning. CIM [19] proposes a simple framework for self-supervised learning. It randomly crops the input image (context), then applies transformations such as scaling, reshaping, and rotation to create query and input pairs, which are modeled through a simple cross-attention block. Research results show that its performance on self-supervised and transfer learning benchmarks is comparable to, or even better than, the current SOTA methods. Furthermore, BYOL [20] iteratively bootstraps the outputs of a network and regards them as targets for an enhanced representation to learn the discrimination between the representation of another augmented view of the same image. Nevertheless, the BYOL is built for the image classification with global view, ignoring the localized multi-instance objects that can be useful for dense prediction tasks, and the data augmentation in CIM is overly simple and monotonous. Our OADP is inspired by BYOL [20], which employs two networks with one network to predict the output of another network to avoid the training collapse and high negative sampling cost in contrastive learning.

B. Self-Supervised Learning On Non-Iconic Multi-Instance Datasets

Considering that global image-level representations are unable to capture the inherent differences of local instances, a group of works [3], [21] proposes global-and-local level contrastive learning to remedy the issue of directly pre-training on non-iconic multi-instance datasets. DenseCL [3] grids the feature map and performs a grid matching mechanism between different images to enhance the ability to perceive local instances representations for SSL models. Nevertheless, it employs all the local parts in contrastive learning, ignoring the harm of the non-object background. Self-EMD [22] keeps the convolutional feature maps as the image embedding to preserve spatial structures and utilizes Earth Mover's Distance to compute the similarity between two embeddings. CCOP [23] employs selective search [24] to find rough object regions and put forward a curriculum learning mechanism to allow the model to consistently acquire a valid learning signal. ORL [25] also applies the selective search and presents k-nearest-neighbor(KNN) to catch the objects for local contrastive learning. However, the three-stage process of this method produces incredible complexity. These aforementioned methods only utilize the coarse resolution local representations during self-supervised pretraining. Compared with these models, our OADP performs multi-scale multi-resolution contrastive learning with novel object-aware and learning-adaptive augmentation to learn more discriminative features.

C. Self-Supervised Learning Application

At present, self-supervised learning frameworks have been widely implemented in various applications. In point cloud upsampling tasks, [26] put forward a self-supervised upsampling network to perceptive context information inside and among local regions without large numbers of paired sparse-dense point sets as supervision from real-scanned sparse data. To alleviate the issues of time and expense of annotated labels in multimodal retinal image registration tasks, [27] propose a self-supervised model to register color fundus images with infrared reflectance and fluorescein angiography. In video quality assessment (VQA) tasks, to employ the plentiful unlabeled video data and learn feature representation in a simple-yet-effective way, [28] implement self-supervised

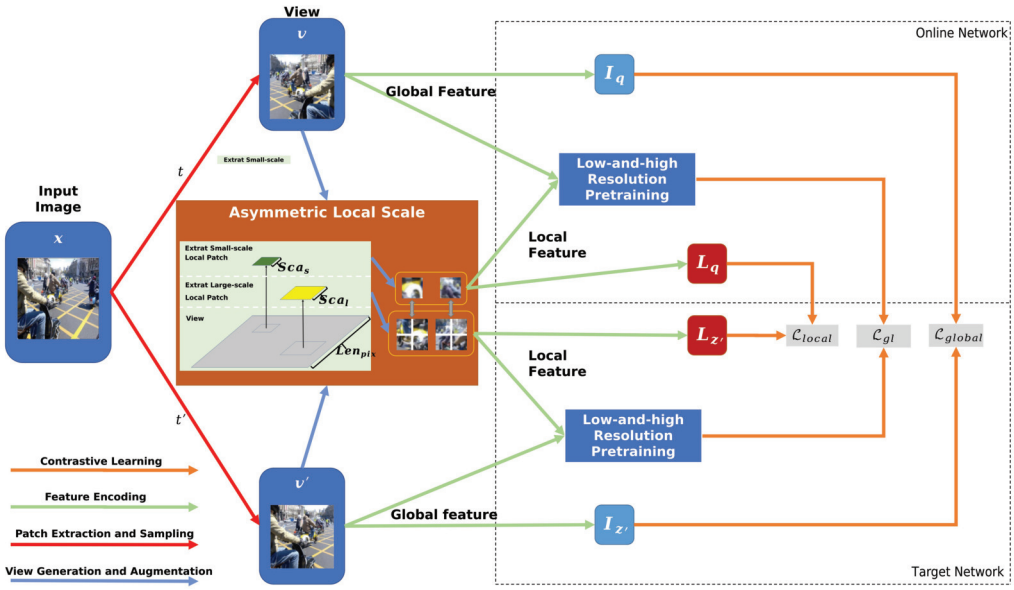

Fig. 1. The overall pipeline of OADP. Given two random augmented views of an input image, the asymmetric local patches L_q and $L_{z'}$ are selected with shape $Sca_l \times Sca_l$ and $Sca_s \times Sca_s$. I represents the global representation of the view. We propose the concatenation of multi-resolution and multi-view for the global-to-local loss $\mathcal{L}_{gl}$ that integrates the high-resolution low-level features of global and asymmetric local views into pre-training.

contrastive learning to capture quality-aware information by maximizing the agreement on feature representations of future frames. In image denoising tasks, [29] proposes a Noisy-As-Clean (NAC) framework to train self-supervised denoising networks only with the corrupted image. In text recognition, self-supervised pre-training is an effective solution that reduces reliance on large amounts of annotated real data, SSM [30] adds the original image to its flipped view, creating a symmetrical overlay input. This approach has led to improvements across various text recognition benchmarks. DegAE [31] proposes a novel self-supervised pre-training paradigm based on contrastive loss learning. By constructing a content reconstruction loss, perceptual loss, generative adversarial loss, and latent vector loss, this approach adapts to the training requirements of low-level vision tasks. It has achieved improvements in tasks such as image denoising and dehazing. Since these aforementioned methods are not for the representative dense prediction task (e.g. object detection, image segmentation), OADP holds the latent capacity to provide enhancement to these dense prediction tasks.

III. Our Method

In this section, we propose OADP (Object Adaptive Self-Supervised Dense Visual Pre-Training), which introduces an object-aware and learning-adaptive augmentation with multi-scale and multi-resolution to learn more discriminative features for multiple object instances without negative sampling, as illustrated in Fig. 1.

Given an input image x in our model, two global views v and v' of x are transformed by random augmentation (For details, please refer to Sec. IV-A). We also design an object-aware and learning-adaptive process to generate a pool of prominent objects regions from small to large scales that allows our method to learn discriminative features for localized object regions with contrastive learning in Sec. III-A.

Our goal is to learn an encoder network that can extract discriminative and dense representations suitable for dense prediction tasks on multi-instance images. The OADP framework leverages object-aware and learning-adaptive augmentations along with multi-scale and multi-resolution feature integration. The main modules of OADP are listed as follows:

Input: an image $x \in \mathcal{X}$;
Output: a dense feature representation $z = f_\theta(x) \in \mathcal{Z}$, where f_θ is the encoder network parameterized by θ;
Encoder network: f_θ maps the input image x to feature representations at multiple scales and resolutions;
Projection head: g_θ transforms the encoder outputs to a representation space suitable for contrastive learning;
Predictor: q_θ used in the online network to predict target representations.

The pre-training step contains:

1. Data Augmentation: For each input image x, generate two augmented global views $v, v' \in \mathcal{T}(x)$ using object-aware and learning-adaptive random resized cropping and other standard augmentations (e.g., color jittering, flipping).
2. Design an object-aware and learning-adaptive process to generate a pool of prominent object regions $\mathcal{P}$ from small to large scales, enhancing the focus on localized object regions (cf. Sec. III-A).
3. Feature Extraction: Pass the views through the networks to extract multi-scale and multi-resolution features from both global views and local patches (cf. Sec. III-B).
4. Parameter Update: Learn the online network parameters θ by minimizing the total loss $\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{local} + \mathcal{L}_{gl}$. Learn the target network parameters θ' as an exponential moving average (EMA) of θ (cf. Sec. III-C).

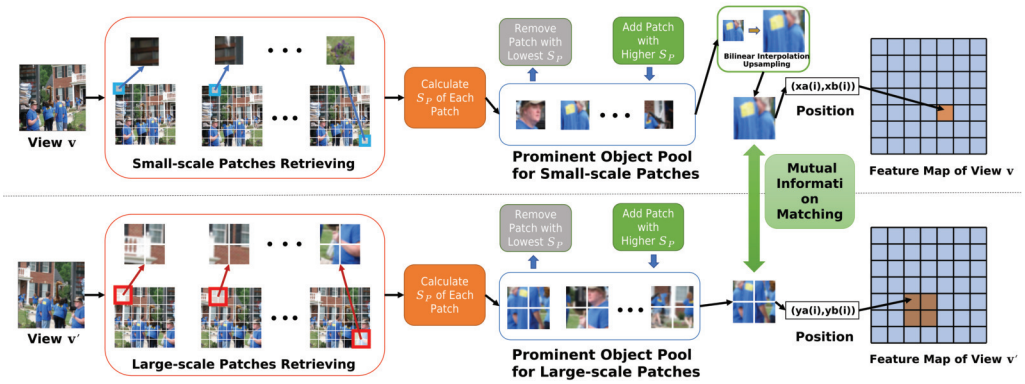

Fig. 2. The asymmetric structure of local contrastive scales. We crop small-scale patches from the augmentation view v and crop large-scale patches from the augmentation view v'. The discrimination of local contrastive scale asymmetry enhances the robustness of local contrastive learning.

We explain the details of the components of OADP in separate subsections in the following parts.

A. Object-Aware Cross-Scale Learning

Most existing self-supervised methods [1], [2], [20], [23] employ random augmentations that are for image classification tasks and ignore the multiple object instances for dense prediction tasks. Therefore we propose a novel object-aware cropping method to learn discriminative object features.

Firstly, we perform the random resized cropping in view v and v' of our model. We set the scale size of random resized crop augmentation as follows:

$$\begin{aligned} Scale(X) &= [0, Side_{up} - \tau ln(epoch - beg)], \\ Scale(Y) &= [Side_{low} - \tau ln(epoch - beg), \\ &\quad 1 - \tau ln(epoch - beg)], epoch > beg. \end{aligned} \tag{1}$$

where $Scale(X), Scale(Y)$ are the scale of random resized crop in view v and v'. The scale for view v is set to a smaller number than the scale size for view v' to improve the compatibility with large-scale local features. We set the upper limit of random resized crop augmentation to $Side_{up}$ for small crops in view v, while the lower limit for large crops in view v' is set to be $Side_{low}$. The dynamically decreasing process will begin after beg epochs. $epoch$ denotes the number of model training epochs. τ represents the magnitude of the learning adaptive process as we consider that the learning small and medium crop's feature is more sensitive in the early stage, hence we perform the random resized crop with a larger scale first with richer surrounding context information to stablize training [32].

As illustrated in Fig. 2, we extract small-scale patches from the augmentation view v and large-scale patches from the augmentation view v'. Considering multiple objects only occupy a small part of an image in non-iconic multi-instance datasets, we design a prominent object pool that collects the asymmetric local patches with a high proportion of object information. For a augmentation view, we divide it into $N \times N$ grids. The small-scale patch is considered as each $Sca \times Sca$ grid ($Sca < N$). The number of local patches in the prominent object pool is fixed as k ($k < (N + 1 - Sca) \times (N + 1 - Sca)$). We only employ the patches in this pool in the subsequent contrastive learning processing. We utilize the image information entropy and LC saliency value [33] to measure the objectness score S_P in the local patches. The image information entropy can remove most uninformative background regions, and the LC value is to identify the regions with the most prominent pixel colors. We define S_P as:

$$S_P = \sum_{i=0}^{255} p(i) log_2 p(i) + \gamma \sum_{I_k \in P} \sum_{n=0}^{255} f_n Dist(g(I_k), n), \tag{2}$$

where $p(i)$ represents the proportion of a specific value i in the pixels of the global patch. f_n denotes the frequency of gray value n in the global augmented view. We perform $g(I_k)$ as the gray value of pixel I_k. $Dist(.)$ is Euclidean distance between two gray values. The left and right terms of Eqn. 2 represent the calculation of image information entropy and LC saliency value, respectively. γ is set to balance the weight between two terms.

We maintain the prominent object pool within the patch number of k after calculating S_P of each local patch. We retrieve the local patches from the non-iconic views using the sliding window. During this process, we record the minimum S_P of patches in the prominent object pool. Each retrieved patch is added to the pool directly when the patch number has not exceeded k. Otherwise, if S_P of the retrieved patch is greater than the recorded minimum S_P in the pool, we replace it with the patch with smallest S_P. After all patches in the views have been checked, there are k local patches with representative object information in the prominent object pool.

After obtaining prominent object pools for both small-scale patches and large-scale patches, we design a scale-aligned operation for subsequent matching process. For each selected small-scale patch $P^{X_i}_{[Sg \times Sca_s, Sg \times Sca_s]}$ in augmentation view v, we extend them to the same scale as large-scale patches via the upsampling method of the bilinear interpolation. Hence, the selected small-scale patches are enlarged to $P^{X_i}_{[Sg \times Sca_l, Sg \times Sca_l]}$.

Afterwards, we employ mutual information for asymmetric patch matching on the basis of the image information entropy. For upsampled small-scale patches $P^{X_i}_{[Sg \times Sca_l, Sg \times Sca_l]}$ and large-scale patches $P^{Y_j}_{[Sg \times Sca_l, Sg \times Sca_l]}$, the mutual information M of

two patches is formulated as:

$$H(P^X, P^Y) = -\sum_{a,b} p_{P^X P^Y}(a,b) log_2 p_{P^X P^Y}(a,b), \tag{3}$$

$$M(P^X, P^Y) = H_{P^X} + H_{P^Y} - H(P^X, P^Y), \tag{4}$$

where P^X, P^Y represents the abbreviations of $P^{X_i}_{[Sg\times Sca_s, Sg\times Sca_s]}, P^{Y_j}_{[Sg\times Sca_l, Sg\times Sca_l]}$. We define $p_{P^X P^Y}(a,b)$ as the joint probability distribution of two pixel values a, b in two patches P^X, P^Y. Compared to the measurement of cosine similarity [3], mutual information is able to measure the interrelationship of patches more accurately.

We consider the pair of the small-scale patch and the large-scale patch with the highest mutual information M to be the most relevant match. For each selected small-scale patches $P^{X_i}_{[Sg\times Sca_s, Sg\times Sca_s]}$, the most relevant large-scale patch $P^{Y_j}_{[Sg\times Sca_l, Sg\times Sca_l]}$ will be calculated based on the mutual information. We record the position of top-left grid of $P^{Y_j}_{[Sg\times Sca_l, Sg\times Sca_l]}$ in the $N \times N$ grids as $(ya(j), yb(j))$, as the position of $P^{X_i}_{[Sg\times Sca_s, Sg\times Sca_s]}$ is $(xa(i), xb(i))$.

During the training process, we extract the feature of each selected small-scale patch $P^{X_i}_{[Sg\times Sca_s, Sg\times Sca_s]}$ and its most relevant large-scale patch $P^{Y_j}_{[Sg\times Sca_l, Sg\times Sca_l]}$ directly from the last feature map of backbone Resnet50 [34]. Corresponding to the grid division paradigm we designed, the feature distribution of the $N \times N$ feature map is 7×7. Each small-scale patch can match the $Sca_s \times Sca_s$ feature in space, while large-scale patches can spatially match the sub-feature map with shape $Sca_l \times Sca_l$. Hence, we extract the asymmetric local feature via the average pooling of feature points in corresponding local areas:

$$L_y = avgpool(f[a,b] \mid xa(i) \le a \le xa(i) + Sca_s - 1, \\ xb(i) \le b \le xb(i) + Sca_s - 1), \tag{5}$$

$$L_{y'} = avgpool(f[a,b] \mid ya(j) \le a \le ya(j) + Sca_l - 1, \\ yb(j) \le b \le yb(j) + Sca_l - 1), \tag{6}$$

where $L_{y'}, L_y$ denotes the embedded feature of the matched small-scale patches and large-scale patches, respectively. $f[a,b]$ represents the feature at coordinate (a, b) from the last feature map of Resnet-50. $avgpool$ is the operation of average pooling. The local embedded feature $L_{y'}, L_y$ will be fed into the local contrastive learning module $\mathcal{L}_{local}$:

$$\mathcal{L}_{\text{local}} \triangleq \left\| \frac{L_y}{||L_y||_2} - \frac{L_{y'}}{||L_{y'}||_2} \right\|_2^2 = 2 - 2 \cdot \frac{\langle L_y, L_{y'} \rangle}{||L_y||_2 ||L_{y'}||_2}, \tag{7}$$

B. Object Centric Cross-Resolution Learning

Existing self-supervised models [3], [21], [23] only employ the late layers' features from the backbone, ignoring that early layer's features possess finer resolution and detailed information. To address this issue, we propose to integrate both fine and coarse features in cropped views of augmented images.

The main pipeline of cross-resolution integration is illustrated in Fig. 3. We set the δ layer from the end of backbone Resnet50 [34] as the high-resolution low-level feature maps,

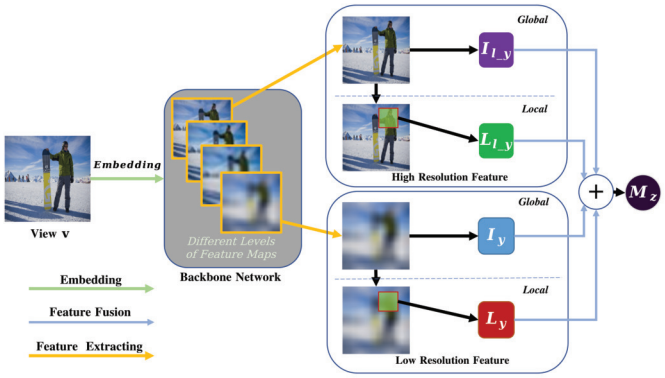

Fig. 3. The pipeline of low and high resolution pre-training. We extract the global and local high resolution features I_{l_y} and L_{l_y} from the low-level feature map of the backbone network. The low resolution features are extracted from the last feature map.

while the low-resolution feature is the adaptive average pooling of the last $N \times N$ feature maps with depth dp.

For augmented global view, we define its fine-resolution feature I_{gf} as:

$$I_{gf} = maxpool\{fp \mid fp \in layer_\delta\}, \tag{8}$$

where $maxpool$ denotes adaptive max pooling. $layer_\delta$ denotes the δ layer of Resnet50, whose shape is $(N \cdot 2^\delta) \times (N \cdot 2^\delta)$. Correspondingly, the augmented global view's coarse-scale feature is defined as $I_{gc} = avgpool\{fp \mid fp \in layer\delta(N \times N)\}(\delta = 0)$. Similarly, we would like to extract the fine-resolution and coarse-resolution features of cropped patches I_{pf} and I_{pc}. After obtaining these features, we concatenate them together:

$$M_z = cat(I_{gf}, I_{gc}, I_{pf}, I_{pc}). \tag{9}$$

We get M_z for view v. Similarly, we can compute the concatenated feature M'_z for view v'. We feed M_z into MLP heads and obtain the projection feature M_q and optimize with the cross-resolution loss $\mathcal{L}_{gl}$ as:

$$\mathcal{L}_{gl} \triangleq \left\| \frac{M_q}{||M_q||_2} - \frac{M_{z'}}{||M_{z'}||_2} \right\|_2^2 = 2 - 2 \cdot \frac{\langle M_q, M_{z'} \rangle}{||M_q||_2 ||M_{z'}||_2}. \tag{10}$$

C. Overall Loss Function

Given the $\mathcal{L}_{local}$ and $\mathcal{L}_{gl}$ introduced in earlier sections. The whole loss $\mathcal{L}$ of OADP is given as:

$$\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{local} + \mathcal{L}_{gl}, \tag{11}$$

and $\mathcal{L}_{global}$ is the conventional contrastive learning loss that with:

$$\mathcal{L}_{global} \triangleq \left\| \frac{I_{gf}}{||I_{gf}||_2} - \frac{I'_{gf}}{||I'_{gf}||_2} \right\|_2^2 = 2 - 2 \cdot \frac{\langle I_{gf}, I'_{gf} \rangle}{||I_{gf}||_2 ||I'_{gf}||_2}, \tag{12}$$

where I_{gf} and I'_{gf} are the global features for view v and v', respectively.

IV. Experiments

In this section, we start by describing the non-iconic multi-instance datasets and implementation details utilized in our experimentation. Then, we report the performance of our proposed OADP on a suite of downstream tasks compared with existing state-of-the-art methods. Afterwards, extensive ablation experiments are implemented to investigate the effect of critical components. Finally, we visualize the results of our OADP to confirm the superiority of our proposed model intuitively.

A. Experimental Settings

1) Datasets: OADP utilizes two widely-used non-iconic multi-instance datasets during pre-training: PASCAL VOC [10] and MS COCO [11]. PASCAL VOC contains 12k images and 28k annotated instances from daily lives, covering 20 classes that consist of four themes of person, animal, vehicle and furniture. As a large-scale dataset, MS COCO contains about 120k images and 890k annotated instances of 80 classes. The images in COCO have 7.3 objects on average. Note that we do not employ any annotations of these two datasets during pre-training.

2) Model Settings: In OADP, Resnet50 [34] is applied as the default backbone network. We utilize the SGD optimizer with a weight decay of 0.0001 and a momentum of 0.9. In the design of asymmetry of local contrastive scale, the sizes of the small-scale patch and large-scale patch are configurated as 1×1 and 2×2 in 7×7 feature maps to achieve the optimal training efficiency. The pre-training processes are implemented on 8 V100 GPUs with the default batchsize 256. In low and high resolution pretraining for global and asymmetric local representations, the high-resolution low-level feature maps are configured as the fourth layer from the end of backbone Resnet50. We employ the cosine learning rate decay schedule with a base learning rate of 0.2. To keep consistent with other self-supervised models, the number of training epochs for both multi-instance datasets PASCAL VOC [10] and MS COCO [11] is set to 800 with a warm-up period of 4 epochs. We implement the exponential moving average parameter χ from 0.99 to 1 during training.

3) Augmentation Methods: OADP requires capturing both global and local features at multiple scales and resolutions. We have designed a detailed and robust data augmentation process to support this. During the first 400 epochs of training, we applied the following data augmentation operations:

a) Transformations for View1:

- RandomResizedCrop: the scale range is set to (0.1, 0.6), i.e., a random crop of 10% to 60% of the original image area is selected and resized to 224×224 pixels.
- RandomApply of ColorJitter: color jitter is applied with a probability of 30%, with maximum changes of 0.8 for brightness, contrast, and saturation, and 0.2 for hue.
- RandomGrayscale: an image is converted to gray scale with a 20% probability.
- RandAugment: five random augmentations are applied with a strength of 10.
- RandomHorizontalFlip: a 50% chance of horizontally flipping the image.
- RandomApply of GaussianBlur: the Gaussian blur is applied with a 20% probability, with a kernel size of $(3, 3)$ and a standard deviation randomly chosen between 1.0 and 2.0.

b) Transformations for View2: Similar to *view1*, but with the following difference:

- RandomResizedCrop: the scale range is set to (0.3, 1), meaning a random crop of 30% to 100% of the original image area is selected and resized to 224×224 pixels. All other augmentation operations (e.g., ColorJitter, Grayscale, RandAugment, Horizontal Flip, Gaussian Blur) are as the same as in *view1*.

After 400 epoches, we dynamically adjusted the parameters of data augmentation, by reducing the scale range of **RandomResizedCrop** as the training progressed. The specific adjustments were as follows:

c) Transformations for View1:

- RandomResizedCrop: the scale range is adjusted to

$$\left(0.1,\ 0.6 - 0.04 \times \ln(\text{epoch} - 400)\right),$$

where ln represents the natural logarithm, and epoch is the current epoch number. As the epoch increases, the upper limit of the scale range gradually decreases, resulting in smaller crop areas. All other augmentation operations (e.g., ColorJitter, Grayscale, RandAugment, Horizontal Flip, Gaussian Blur) remain unchanged.

d) Transformations for View2:

- RandomResizedCrop: the scale range is adjusted to:

$$\begin{pmatrix} 0.3 - 0.04 \times \ln(\text{epoch} - 400), \\ 1 - 0.04 \times \ln(\text{epoch} - 400) \end{pmatrix}.$$

Similarly, as the epoch increases, both the upper and lower limits of the scale range gradually decrease. All other augmentation operations are the same as in *view1*.

B. Transferring to Downstream Tasks

Here, we employ our pre-trained model to a suite of downstream tasks, including PASCAL VOC object detection and classification, COCO object detection and instance segmentation, PASCAL VOC and Cityscapes semantic segmentation. Afterwards, we compare the performance of our OADP model with existing self-supervising methods.

1) Object Detection Fine-Tuned on Pascal VOC: We pre-train our OADP model on COCO `train2017` for 800 epochs with Resnet50 used SGD optimizer with an initial learning rate set to 0.02, momentum of 0.9, weight decay of 1×10^{-4}, and 32 batchsize. We utilize the Mask R-CNN with R50-FPN backbone. The image scale ranged from 480 to 800 during training and 800 at inference. Afterwards, we evaluate the object detection performance of the COCO pre-trained model on PASCAL VOC and compare it with other state-of-the-art methods. After freeze the backbone of pre-trained model, we train the detector on PASCAL VOC `trainval2007+2012` and evaluate the object detection performance on PASCAL VOC `test2007`. As illustrated in Table I, our proposed OADP

TABLE I
PASCAL VOC OBJECT DETECTION RESULTS PRE-TRAIN ON COCO FOR 800 EPOCHS

Model	Pre-train data	Epoch	AP	AP50	AP75
MoCo-v2 [2]	COCO	800	52.1	79.0	56.7
BYOL [20]	COCO	800	53.7	80.2	59.9
DenseCL [3]	COCO	800	55.3	80.5	60.8
MaskCo [21]	COCO	800	55.6	81.0	60.5
CCOP [23]	COCO	800	55.8	81.7	60.9
univip [35]	COCO	800	56.5	82.3	62.6
OADP	COCO	800	**56.7**	**82.4**	**62.5**

TABLE II
PASCAL VOC CLASSIFICATION RESULTS PRE-TRAIN ON COCO AND PASCAL VOC

Model	Pre-train Data	Epoch	VOC07 Accuracy
Random init	-	800	9.6
SimCLR [1]	VOC07+12	800	58.3
MoCoV2 [2]	VOC07+12	800	60.6
CLSA [36]	VOC07+12	800	61.8
BYOL [20]	VOC07+12	800	64.7
univip [35]	VOC07+12	800	65.7
OADP(Resnet-50)	VOC07+12	800	**66.8**
OADP(Resnet-50(2x))	VOC07+12	800	**67.7**
OADP(Resnet-50(4x))	VOC07+12	800	**68.5**
SimCLR [1]	COCO	800	74.1
MoCo [16]	COCO	800	76.3
MoCoV2 [2]	COCO	800	77.4
CLSA [36]	COCO	800	78.9
BYOL [20]	COCO	800	80.8
ORL [25]	COCO	800	86.7
OADP(Resnet-50)	COCO	800	**86.9**
PIRL [37]	ImageNet	800	81.1
MoCoV2 [2]	ImageNet	200	84.1

model achieves results of 56.7% AP, 82.4% AP50 and 62.5% AP75 on PASCAL VOC object detection tasks. These results outperform CCOP and univip by 0.9% AP and 0.2% AP. Compared with baseline BYOL, OADP can surpass it by 3.0% AP. It can be observed that we achieve much larger enhancement on AP75 index compared to AP50, indicating that OADP has the significant effects to improve the localization accuracy. These results strongly confirm the efficiency of our model on multi-instance datasets.

2) Pascal VOC Image Classification: In addtion to COCO `train2017`, we also pre-train our OADP model directly on PASCAL VOC `trainval2007+2012` for 800 epochs and the backbone model remains ResNet50, using the SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, weight decay of 1×10^{-4}, and a batch size of 256.Subsequently, we added a fully connected layer for fine-tuning image classification, adjusting the learning rate to 0.03. The backbone was frozen, and the model was further trained for 200 epochs. In Table II, we report the top-1 image classification accuracy on VOC `test2007` with both COCO and PASCAL VOC pretrained model. It can be observed that our OADP reaches 86.9% top-1 accuracy with COCO pre-trained model, yielding 0.2% and 6.1% top-1 accuracy improvements over state-of-the-art method ORL and baseline BYOL. For PASCAL VOC pre-trained model, OADP gain the enhancement of 2.1% top-1 accuracy over BYOL. To further compare whether a wider model structure affects performance, we continued to fine-tune ResNet50(2x) and ResNet50(4x) on the PASCAL VOC dataset. As reported in Table VI, ResNet50(2x) and ResNet50(4x) improved top-1 accuracy by 0.9% and 1.7%, respectively, compared to the standard ResNet50. It is noted that the performance of OADP COCO pre-trained model even yields improvements over the result of MoCoV2 pre-trained on ImageNet. It indicates that our model possesses the competitive performance on image classification downstream tasks.

3) Object Detection and Instance Segmentation Fine-Tuned on COCO: The results of MS COCO object detection and instance discrimination are shown in Table III. The model is pre-trained on COCO `train2017` for 800 epochs. We fine-tune all layers end-to-end on COCO `train2017` using the Mask R-CNN with R50-FPN for 90k iterations, using the SGD optimizer with an initial learning rate of 0.1, momentum of 0.9, weight decay of 1×10^{-4}, and a batch size of 16, and evaluate the performance on COCO `val2017`. It can be observed in Table III that our OADP gains the enhancement of 0.2% AP and 1.1% AP over state-of-the-art method ORL and baseline BYOL for COCO Object detection task. As for COCO instance segmentation tasks, our model can also outperform CCOP and BYOL by 0.1% AP and 0.8% AP. These improvements further demonstrate the superiority of our OADP for learning the discrimination of unsupervised localized representations. We have also fine-tuned the COCO pre-trained model with detection transformer (DETR) [39]. We convert the CNN backbone to MASP and BYOL pre-trained model on COCO and random initialize the parameters of transformers. Following UP-DETR [40], we employ the initial learning rate 1×10^{-4} for transformers and 5×10^{-5} for the CNN backbone. The model is fine-tuned for 300 epochs. As shown in Table IV, our OADP yield 1.1% AP improvements over baseline BYOL for COCO Object detection task with detection transformer. It indicates that our self-supervised model proves beneficial for the transformer architecture.

4) Semantic Segmentation Fine-Tuned on Pascal VOC and Cityscapes: Table V demonstrates PASCAL VOC and Cityscapes semantic segmentation results pre-trained on MS COCO `train2017`. We utilize PASCAL VOC `trainval2012` and Cityscapes `train_fine` for FCN fine-tuning, the backbone model remains ResNet50, using the SGD optimizer with an initial learning rate of 0.2, momentum of 0.9, weight decay

TABLE III
COCO OBJECT DETECTION AND INSTANCE SEGMENTATION RESULTS PRE-TRAIN ON COCO

Task	Model	Pre-train Data	Epoch	AP	AP50	AP75
Object Detection	Random [38]		800	32.8	50.9	35.3
	Supervised [38]	ImageNet	800	39.7	59.5	43.3
	SimCLR [1]	COCO	800	38.0	57.3	41.5
	MoCo-v2 [2]	COCO	800	38.5	58.0	41.9
	Self-EMD [22]	COCO	800	39.3	60.1	42.8
	BYOL [20]	COCO	800	39.4	59.4	43.1
	DenseCL [3]	COCO	800	39.6	59.3	43.3
	CCOP [23]	COCO	800	39.9	59.8	43.9
	ORL [25]	COCO	800	40.3	60.2	**44.4**
	OADP	COCO	800	**40.5**	**60.4**	44.3
Instance Segmentation	Random [38]		800	29.9	47.9	32.0
	Supervised [38]	ImageNet	800	35.9	56.6	38.6
	SimCLR [1]	COCO	800	34.3	54.5	36.6
	MoCo-v2 [2]	COCO	800	34.8	55.3	37.3
	DenseCL [3]	COCO	800	35.7	56.8	38.4
	BYOL [20]	COCO	800	35.6	56.5	38.3
	CCOP [23]	COCO	800	36.2	56.8	38.8
	ORL [25]	COCO	800	36.3	57.3	38.9
	OADP	COCO	800	**36.4**	**57.4**	**38.9**

TABLE IV
COCO OBJECT DETECTION WITH TRANSFORMERS

Model	Pre-train data	Epoch	AP	AP50	AP75
BYOL [20]	COCO	300	32.5	49.8	34.2
OADP	COCO	300	**33.6**	**50.8**	**35.6**

of 5×10^{-4}, and a batch size of 16. The evaluation datasets are set as PASCAL VOC `val2012` and Cityscapes `val2012`. As shown in Table V, our OADP yields 1.9% and 0.9% mIoU gains over baseline model BYOL and state-of-the-art method DenseCL. For Cityscapes semantic segmentation, OADP can outperform BYOL and DenseCL by 1.8% mIoU and 0.6% mIoU. These results strongly confirm that our proposed OADP possesses a significant performance on these dense prediction tasks.

5) Monocular Depth Estimation on NYU Depth v2 Dataset: To evaluate the generalizability of our OADP framework, we conducted experiments on depth estimation using the NYU Depth v2 dataset. Following the experimental setup of BYOL, we employed a standard ResNet-50 as the backbone network, initialized with our pre-trained weights from MS COCO. The model was fine-tuned on the NYU Depth v2 dataset for approximately 3,000 steps with a batch size of 128. As shown in Table VI, our OADP method achieved better performance compared to models pre-trained with BYOL and SimCLR. We

TABLE V
SEMANTIC SEGMENTATION RESULTS PRE-TRAIN ON PASCAL VOC AND-CITYSCAPES

Validation Data	Model	Pre-train Data	Epoch	mIoU
PASCAL VOC	Random [38]		800	39.3
	SimCLR [1]	COCO	800	48.7
	MoCoV2 [2]	COCO	800	48.9
	BYOL [20]	COCO	800	55.7
	DenseCL [3]	COCO	800	56.7
	OADP	COCO	800	**57.6**
Cityscapes	Random [38]		800	61.3
	SimCLR [1]	COCO	800	67.9
	MoCoV2 [2]	COCO	800	71.3
	CLSA [36]	COCO	800	71.7
	BYOL [20]	COCO	800	74.1
	DenseCL [3]	COCO	800	75.3
	OADP	COCO	800	**75.9**

TABLE VI
DEPTH ESTIMATION RESULTS ON NYU DEPTH V2

	Higher better			Lower better	
Method	pct.< 1.25	pct.$< 1.25^2$	pct.$< 1.25^3$	rms	rel
Supervised-IN [20]	81.1	95.3	98.8	0.573	0.127
SimCLR [1]	83.3	96.5	99.1	0.557	0.134
BYOL [20]	84.6	96.7	99.1	0.541	**0.129**
OADP (ours)	**84.8**	**97.4**	**99.4**	**0.435**	0.130

evaluate our method using the relative (rel) error, root mean squared (rms) error, and the percentage of pixels (pct) where the error $\max\left(\frac{d_{\text{gt}}}{d_p}, \frac{d_p}{d_{\text{gt}}}\right)$ is below 1.25^n thresholds, where d_p is the predicted depth and d_{gt} is the ground truth depth [20]. OADP is better or on par with other methods for each metric. For instance, the challenging pct.<1.25 measure is respectively improved by 3.7% and 1.5% compared to Supervised-IN and SimCLR baselines.

C. Ablation Study

Here, we carry out extensive ablation experiments to further investigate the contribution of critical components and the most effective setting of the model parameter in our proposed OADP.

1) Effect of Low and High Resolution Pretraining: We evaluate the performance of concatenation of low and high resolutions for our localized asymmetric self-supervised pre-training elaborated in Sec III-B on PASCAL VOC object detection tasks. All the models are pre-trained on COCO `train2017`. The results are illustrated in Table VII. Layer1 denotes the first low-level feature map with high-resolution

TABLE VII
EFFECT OF INTEGRATION OF DIFFERENT RESOLUTIONS

Integration members in the backbone	AP	AP50	AP75
layer4	56.5	82.1	62.4
layer1+layer4	**56.7**	**82.4**	**62.5**
layer2+layer4	55.9	82.1	62.1
layer1+layer3	54.1	80.2	61.9
layer1+layer2+layer3+layer4	55.9	81.7	61.9

TABLE VIII
EFFECT OF LOCALIZED ASYMMETRIC PATCH NUMBER

Patch Number	VOC07 Accuracy
$k = 8$	64.9
$k = 16$	66.1
$k = 24$	**66.8**
$k = 32$	64.7

TABLE IX
EFFECT OF ASYMMETRIC MUTATIVE AUGMENTATION

settings of random resized crop augmentation	VOC07 Accuracy
Equivalent scale sizes	66.3
Asymmetric scale sizes with dynamic decreasing	**66.8**

TABLE X
EFFECT OF EACH LOSS COMPONENT ON PASCAL VOC CLASSIFICATION

$\mathcal{L}_{global}$	$\mathcal{L}_{local}$	$\mathcal{L}_{gl}$	VOC07 Accuracy (%)
0	0	1	77.1
0	1	0	74.3
1	0	0	81.7
0	1	1	78.5
1	0	1	82.4
1	1	0	83.1
1	0.5	0.5	84.3
0.5	1	0.5	82.5
0.5	0.5	1	83.5
1	1	1	**86.9**

representations, and layer4 represents the last 7×7 feature map of the backbone. Correspondingly, layer2 and layer3 represent the middle two layers of the backbone. The feature integration paradigms of these layers are all the same, designed in Sec III-B. We can observe that the integration of layer1 and layer4 yields 0.2% AP enhancement over only utilizing the last feature map. It strongly demonstrates the positive effects of the integration of low-level high-resolution features for localized self-supervised pre-trainings. Note that the performance of concatenation of all layers is 0.8% AP lower than the integration of layer1 and layer4. It indicates that integrating too many intermediate low-level features into one contrastive learning object will cause information redundancy during pre-training.

2) Effect of Localized Asymmetric Patch Number: We conduct several experiments to explore the most effective setting of the number k of asymmetric local patches. These experiments are carried out for PASCAL VOC image classification downstream tasks, and models are pre-trained on PASCAL VOC `trainval2007+2012` for 800 epochs. The 1×1 small-scale local patches and 2×2 large-scale local patches have an equal number of k. As shown in Table VIII, the performance of our OADP varies significantly with different asymmetric patch numbers. The overall effect of k presents a trend of descending after ascending. When we utilize about 24 asymmetric local patches, our proposed OADP obtains the highest top-1 classification accuracy on VOC07.

3) Effect of Asymmetric Dynamic Augmentation: We investigate the effects of our proposed asymmetric dynamic augmentation in Sec III-A via conducting experiments on PASCAL VOC image classification tasks. We pre-train our OADP model on PASCAL VOC `trainval2007+2012` for 800 epochs with different settings of the random resized crop augmentation. Table IX ablates the results of these experiments. It can be observed that asymmetric scale sizes of the random resized crop augmentation with dynamic decreasing outperform equivalent scale sizes by 0.5% top-1 accuracy on VOC07. It strongly demonstrates that our proposed asymmetric dynamic augmentation can improve the performance of self-supervised models in OADP.

4) Effect of Each Loss Component: We conducted ablation experiments to evaluate the contribution of each component of our model to OADP. Specifically, OADP introduces the improvement over BYOL, corresponding to the losses $\mathcal{L}_{global}$, $\mathcal{L}_{local}$, and $\mathcal{L}_{gl}$. In this set of experiments, we controlled the participation and contribution of each component to the training process by adjusting the weight parameters of the corresponding losses. All models with different loss combinations were pre-trained on COCO `train2017` for 800 epochs using the same training parameters. Subsequently, each model was fine-tuned on the PASCAL VOC image classification downstream task. As shown in Table X, the results indicate that the model achieves the best performance when all three losses are involved with equal weight parameters. Models trained with any two-loss combinations outperform those using only one loss. Among single-loss models, $\mathcal{L}_{global}$ outperforms both $\mathcal{L}_{local}$ and $\mathcal{L}_{gl}$.

5) Effect of Mutual Information for Patch Matching: To demonstrate the effectiveness of mutual information in patch patching, we compare it with the cosine similarity.

TABLE XI

EFFECT OF SIMILARITY MEASUREMENT FOR PATCH MATCHING ON PASCAL VOC CLASSIFICATION

Similarity Measurement	Pre-train Data	VOC07 Accuracy (%)
Cosine Similarity	COCO	85.7
Mutual Information	COCO	**86.9**

TABLE XII

EFFECT OF CONTRASTIVE LOSS FUNCTIONS ON PASCAL VOC CLASSIFICATION

Loss Function	Pre-train Data	Epochs	VOC07 Accuracy (%)
MSE Loss	COCO	800	86.6
Cosine Similarity	COCO	800	**86.9**

Both methods are pre-trained with identical training parameters for 800 epochs on the COCO `train2017` dataset. Fine-tuning is then performed on the PASCAL VOC image classification downstream task. As shown in Table XI, the model using mutual information for patch matching achieved 1.2% higher accuracy than the model using cosine similarity. This result aligns with the fact that mutual information is better than cosine similarity for capturing region differences.

6) Effect of Contrastive Loss Functions: To explore the impact of different contrastive loss functions on our model, we conducted ablation experiments comparing cosine similarity and Mean Squared Error (MSE) loss following [1] and [20]. We replaced the cosine similarity in our loss calculations $\mathcal{L}_{global}$, $\mathcal{L}_{local}$, and $\mathcal{L}_{gl}$ with MSE loss. Both models were pre-trained on the COCO dataset for 800 epochs using identical training configurations. Fine-tuning was then performed on the PASCAL VOC image classification downstream task. As shown in Table XII, the model trained with cosine similarity as the loss function outperformed the model trained with MSE loss by 0.3% on the VOC07 image classification validation set. The cosine similarity provides the performance improvement over MSE, which is more suitable for our method.

7) Effect of Date Augmentation: To explore the effect of data augmentation during the pre-training phase, we designed three different combinations of data augmentation experiments, which is shown in Table XIII. All three experiments used ResNet50 as the backbone, pre-trained for 800 epochs on the PASCAL VOC trainval2007 + 2012 dataset, followed by 200 epochs of fine-tuning on the PASCAL VOC07 image classification task. The model using only basic augmentations achieved the lowest accuracy, while the model with more augmentation techniques improved accuracy by 2.4%. The best performance was achieved by incorporating RandomResizedCrop with dynamic adjustment, demonstrating that diverse data augmentation and dynamic RandomResizedCrop benefit OADP pre-training.

TABLE XIII

PASCAL VOC (VOC07 + 12) CLASSIFICATION RESULTS WITH DIFFERENT DATA AUGMENTATION STRATEGIES. AUG1: RANDOMRESIZEDCROP(WITHOUT DYNAMIC ADJUSTMENT); AUG2: RANDOMHORIZONTALFLIP; AUG3: RANDOMAPPLY OF COLORJITTER; AUG4: RANDOMGRAYSCALE; AUG5: RANDOMAPPLY OF GAUSSIANBLUR; AUG6: RAND AUGMENT; AUG7:RANDOMRESIZEDCROP(**WITH DYNAMIC ADJUSTMENT AFTER EPOCH 400**)

Data Augmentation Combination	VOC07 Accuracy (%)
Aug1, Aug2	62.1
Aug1, Aug2, Aug3, Aug4, Aug5, Aug6	64.5
Aug2, Aug3, Aug4, Aug5, Aug6, Aug7	**66.8**

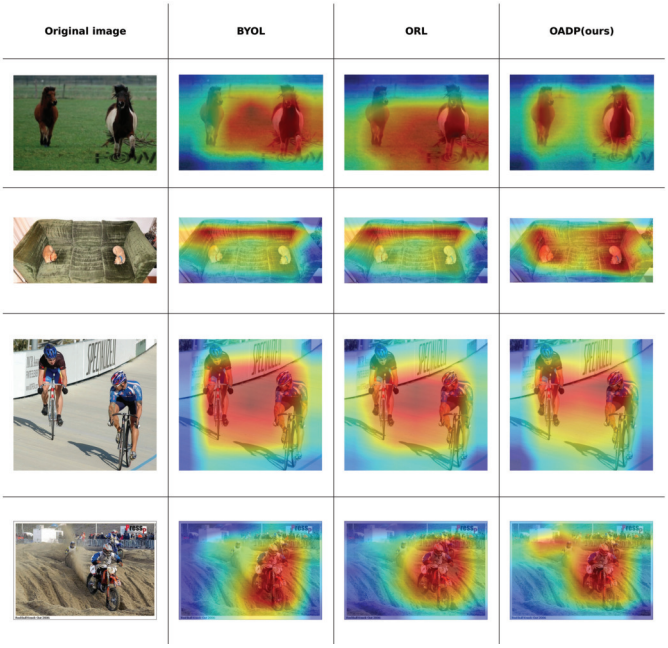

Fig. 4. **Heat maps generated by BYOL, ORL and OADP with CAM tools.** If the color intends to be red, the attention to this area is more strong. Compared with baseline BYOL and ORL, OADP possesses a more accurate activation of instance regions.

D. Visualization Results

To further understand the feature representation intuitively learned by different self-supervised methods, we visualize the heat maps of BYOL, ORL and our proposed OADP pre-trained on MS COCO `train2017` with CAM tools. Fig. 4 illustrates the visualization results of these models. The red areas in the heat maps represent the high-temperature part that the model pays more attention to. Compared with ORL and baseline BYOL, we can observe that our proposed OADP distinguish the boundaries between background noises and objects more accurately in multi-instance images. For example, in the first, BYOL failed to focus on the horse on the left, while ORL paid more attention to the horse on the right. In contrast, our OADP successfully focused on both horses, forming two distinct heatmap centers. These visualized results strongly confirm that OADP obtains a better performance of localizing the objects in multi-instance images than other methods.

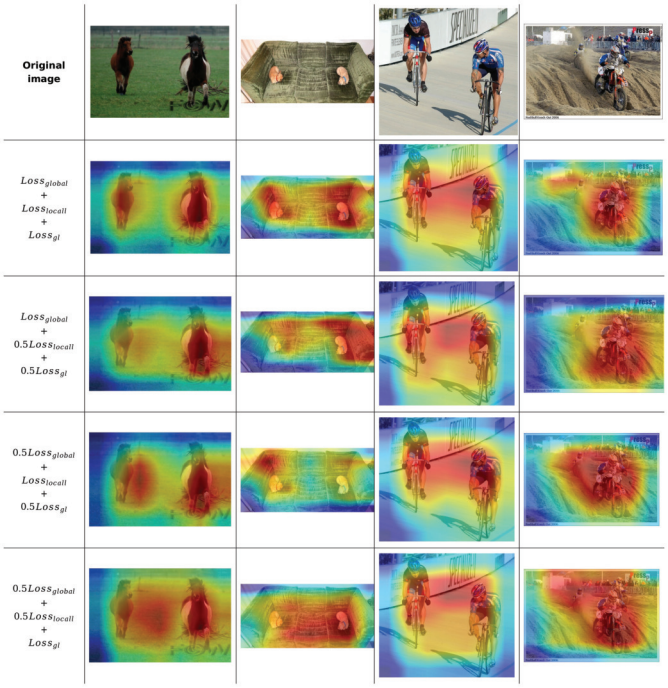

Fig. 5. Heat maps generated by models trained with different loss weight combinations. The red areas indicate regions where the model focuses its attention.

To understand the feature representations learned by different components of our proposed method, we visualize the heat maps of models trained with different loss weight combinations, which is shown in Fig. 5. The red areas in the heat maps represent regions where the model pays more attention. We found that when the three loss weight parameters are equal, the model achieves the best performance. This balanced configuration allows the model to effectively capture the relationship between global and local features, which can identify multiple instances within an image. No matter the weight for which loss is reduced, the heatmap may not accurately locate different instances. These visualization results strongly confirm that each component of OADP provides is useful on locating objects in multi-instance images.

E. Computational Cost

To demonstrate the practical performance and significance of our method, we compared the resource consumption of our approach with others self-supervised methods during both training and inference. For fair comparison, all methods used a standard ResNet50. Given the nature of contrastive loss learning, we measured resource consumption during training with a batch size of 2 and during inference with a batch size of 1. The experimental results are reported in Table XIV. From the results, we can observe that all methods have the same GFLOPs (4 GFLOPs) during inference, as only the backbone remains after removing the training components. During training, BYOL requires 18 GFLOPs, about twice as much as MoCo and SimCLR, because it generates two views of each image. Our method, due to the need to compute multi-scale, multi-resolution, global, and local contrastive losses, consumes 44 GFLOPs during training. Thus, our method does not introduce excessive overhead but rather performs multiple computations on a single image, allowing the model to learn richer multi-scale and multi-instance information.

TABLE XIV
COMPARISON OF COMPUTATIONAL COST FOR DIFFERENT METHODS (PRE-TRAIN WITH BATCH SIZE 2 AND INFERENCE WITH BATCH SIZE 1)

Method	Backbone	Pre-train	GPU Memory	Inference
SimCLR	ResNet-50	8 GFLOPs	1.1 GB	4 GFLOPs
MoCo	ResNet-50	9 GFLOPs	1.1 GB	4 GFLOPs
BYOL	ResNet-50	18 GFLOPs	2.3 GB	4 GFLOPs
OADP	ResNet-50	44 GFLOPs	3.4 GB	4 GFLOPs

V. CONCLUSION

In this work, we have performed a novel self-supervised learning framework, OADP, for visual representation learning on non-iconic multi-instance images. It is motivated by the analysis of the limitation for feature discrimination of multi-scale objects in these images. To magnify the discrimination of localized representations with the instance information and configure adaptive augmentations, we employ the asymmetric self-supervised design paradigms for local contrastive learning scales and configurations of resized crop augmentation. We implement the asymmetry of local contrastive scale in OADP to learn more discriminative local feature information between small-scale and large-scale perspectives. Furthermore, to leverage the prominent detail features of local objects, we design a low and high resolution pre-training method that integrates the low-level high-resolution features for both global and local views of images into pre-training. Extensive experimental results strongly demonstrate that our model pre-trained on non-iconic multi-instance datasets possesses remarkable performance for transferring to different downstream tasks.

Although we have achieved encouraging results on various tasks, there is still room to improve our method. In out paper, since we use the combination of multi-scale and multi-resolution feature extraction, as well as object-aware cropping, which may increase the consumption of computational resources. More efficient feature representation methods are desired for multi-instance images. In addition, for different tasks, such as scene recognition and event recognition, where foreground objects are not apparent, our method may generate too many object candidate regions, which significantly influence the computation efficiency. For these scenarios, one possible way is to improve the prominent region selection step to contain more contextual information to represent the image regions.

We expect the local asymmetric paradigms in OADP can inspire later works in the field of unsupervised training. For future work, we plan to investigate more effective multi-level feature integration of local representations for contrastive learning.

REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.

[2] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[3] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3023–3032.

[4] E. Xie et al., "DetCo: Unsupervised contrastive learning for object detection," 2021, *arXiv:2102.04803*.

[5] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[6] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[7] S. Jiang, S. Liang, C. Chen, Y. Zhu, and X. Li, "Class agnostic image common object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2836–2846, Jun. 2019.

[8] Y. Tian and S. Zhu, "Partial domain adaptation on semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3798–3809, Jun. 2022.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[11] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[12] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.

[13] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 649–666.

[14] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 69–84.

[15] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[18] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2021, pp. 15750–15758.

[19] W. Li, J. Xie, and C. C. Loy, "Correlational image modeling for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15105–15115.

[20] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[21] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z.-J. Zha, "Self-supervised visual representations learning by contrastive mask prediction," 2021, *arXiv:2108.07954*.

[22] S. Liu, Z. Li, and J. Sun, "Self-EMD: Self-supervised object detection without ImageNet," 2020, *arXiv:2011.13677*.

[23] C. Yang, L. Huang, and E. J. Crowley, "Contrastive object-level pre-training with spatial noise curriculum learning," 2021, *arXiv:2111.13651*.

[24] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[25] J. Xie, X. Zhan, Z. Liu, Y.-S. Ong, and C. C. Loy, "Unsupervised object-level representation learning from scene images," in *Proc. NeurIPS*, Jan. 2021, pp. 28864–28876.

[26] X. Liu, X. Liu, Y.-S. Liu, and Z. Han, "SPU-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization," *IEEE Trans. Image Process.*, vol. 31, pp. 4213–4226, 2022.

[27] C. An, Y. Wang, J. Zhang, and T. Q. Nguyen, "Self-supervised rigid registration for multimodal retinal images," *IEEE Trans. Image Process.*, vol. 31, pp. 5733–5747, 2022.

[28] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Contrastive self-supervised pre-training for video quality assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 458–471, 2022.

[29] J. Xu et al., "Noisy-as-clean: Learning self-supervised denoising from corrupted image," *IEEE Trans. Image Process.*, vol. 29, pp. 9316–9329, 2020.

[30] Z. Gao et al., "Self-supervised pre-training with symmetric superimposition modeling for scene text recognition," 2024, *arXiv:2405.05841*.

[31] Y. Liu, J. He, J. Gu, X. Kong, Y. Qiao, and C. Dong, "DegAE: A new pretraining paradigm for low-level vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23292–23303.

[32] S. Gao, Z.-Y. Li, Q. Han, M.-M. Cheng, and L. Wang, "RF-Next: Efficient receptive field search for convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2984–3002, Mar. 2023.

[33] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 815–824.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] Z. Li et al., "UniVIP: A unified framework for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14607–14616.

[36] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," 2021, *arXiv:2104.07713*.

[37] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6706–6716.

[38] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6390–6399.

[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[40] Z. Dai, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised pre-training for object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1601–1610.

Yu Zhang received the B.S. and M.S. degrees in telecommunications engineering from Xidian University, China, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore. He has been a Postdoctoral Fellow with the Bioinformatics Institute, A*STAR, Singapore. He is currently an Associate Professor with Southeast University. His research interests include computer vision.

Tao Zhang received the M.S. degree in cartography and geographic information systems from Zhejiang Normal University, China, in 2022. He is currently pursuing the Ph.D. degree with the School of Cyber Space Security, Southeast University. His research interests include machine learning and its application to computer vision and multi-modal analysis.

Hongyuan Zhu (Member, IEEE) is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. His research interests include multimedia content analysis and segmentation, specially image segmentation/cosegmentation, object detection, scene recognition, and saliency detection.

Zihan Chen received the B.S. degree in computer science and technology from the University of Electronic Science and Technology of China and the M.S. degree in computer science and technology from the School of Computer Science and Engineering, Southeast University. His research interests include deep learning and computer vision.

Siya Mi received the double B.S. degree from Beijing University of Posts and Telecommunications, Beijing, China, and the University of London, London, U.K., in 2010, and the M.S. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2011 and 2018, respectively. She is currently a Lecturer with Southeast University, Nanjing, China. Her research interests include the data processing and computer vision for cyber security.

Xi Peng (Senior Member, IEEE) received the Ph.D. degree from Sichuan University in 2013. From 2014 to 2017, he was a Research Scientist with the Institute for Infocomm, Research Agency for Science, Technology, and Research (A*STAR), Singapore. He is currently a Professor with Sichuan University. His main current research interests include machine learning and its applications in image processing, computer vision, multi-modal analysis, and natural language processing.

Xin Geng (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Nanjing University, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Deakin University, Australia, in 2008. He is currently a Chair Professor with the School of Computer Science and Engineering, Southeast University, China. His research interests include machine learning, pattern recognition, and computer vision.