# Exploring Localization for Self-supervised Fine-grained Contrastive Learning

Di Wu[1,2,⋆]
wudi@westlake.edu.cn
Siyuan Li[1,2,⋆]
lisiyuan@westlake.edu.cn
Zelin Zang[1,2]
zangzelin@westlake.edu.cn
Stan Z. Li[1,†]
Stan.ZQ.Li@westlake.edu.cn

[1] AI Lab, School of Engineering,
Westlake University,
Hangzhou, Zhejiang, China

[2] Zhejiang University,
Hangzhou, Zhejiang, China

**Abstract**

Self-supervised contrastive learning has demonstrated great potential in learning visual representations. Despite their success in various downstream tasks such as image classification and object detection, self-supervised pre-training for fine-grained scenarios is not fully explored. We point out that current contrastive methods are prone to memorizing background/foreground texture and therefore have a limitation in localizing the foreground object. Analysis suggests that learning to extract discriminative texture information and localization are equally crucial for fine-grained self-supervised pre-training. Based on our findings, we introduce cross-view saliency alignment (CVSA), a contrastive learning framework that first crops and swaps saliency regions of images as a novel view generation and then guides the model to localize on foreground objects via a cross-view alignment loss. Extensive experiments on both small- and large-scale fine-grained classification benchmarks show that CVSA significantly improves the learned representation.

# 1 Introduction

Learning visual representations without supervision by leveraging pretext tasks has become increasingly popular. Various learning approaches such as colorization [56], Rel-Loc [31], Rot-Pred [13] have been proposed to learn such representations. The objective of these pretext tasks is to capture invariant features through predicting transformations applied to the same image. More recently, self-supervised representation learning has witnessed significant progress by the use of contrastive loss [4, 16, 17, 18, 28]. Despite that contrastive-based methods have even outperformed supervised methods under some circumstances, their success has largely been confined to large-scale general-purpose datasets (coarse-grained) such as ImageNet [22]. We argue that current contrastive learning methods only work on coarse-grained iconic images with large foreground objects residing in the background with infor-


⋆The first two authors contribute equally † Corresponding author: Stan Z. Li

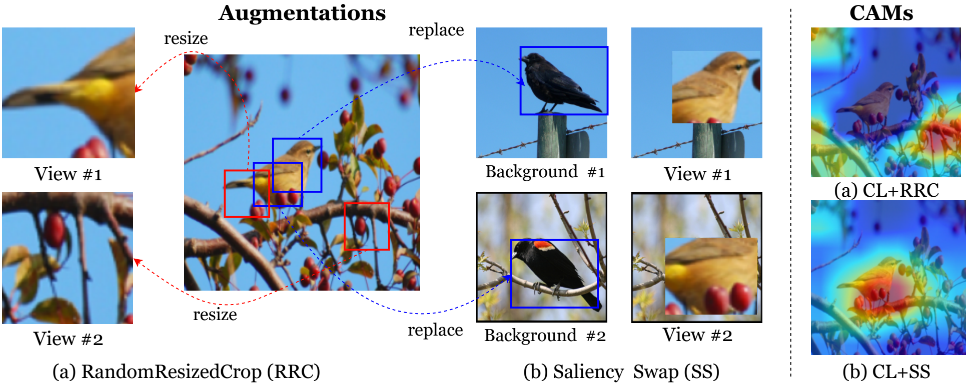


Figure 1: Comparison of *RandomResizedCrop* (RRC) and the proposed Saliency Swap (SS). We visualize Grad-CAM [35] of linear classifiers for pre-trained models. (a) shows the most commonly adopted RRC in contrastive learning (CL) methods. The random nature may cause views to contain mainly the backgrounds of the image leading to semantic inconsistency across views. (b) shows our proposed SS, which crops from regions of interest of the reference image and replaces the saliency regions of two randomly selected background images to guarantee semantic consistency.

mative discriminative texture (*e.g.,* ImageNet) but perform poorly when background texture provides little clue (*e.g.,* CUB-200-2011 [42]) for fine-grained separation.

To bridge the substantial gap between self-supervised and supervised representation learning on fine-grained object recognition, we first analyze and compare knowledge learned by various self-supervised methods and supervised methods during pre-training. We find that current self-supervised contrastive learning methods tend to learn low-level texture information and lack the localization ability of the foreground object. In contrast, the supervised method shows better localization ability. Specifically, we show that the incompetence of localization of current contrastive learning is primarily due to the commonly adopted *RandomResizedCrop* (RRC) augmentation, where a random size patch at a random location is cropped and resized to the original size. The model then might learn a semantic representation of the bird by contrasting the tree and the wing of the bird, as illustrated in Figure 1. This practice may be reasonable for coarse-grained recognition if background cues are more associated with the class than the foreground cues (e.g. $p(bird|tree) > p(car|tree)$). However, the background of the image being a tree is not as informative when distinguishing bird species. Consequently, the model learns by cheating on picking low-level texture clues (usually from the background) instead of learning by localizing the foreground. This phenomenon is mutual for existing contrastive methods such as MoCo.v2[5], BYOL[15] despite different contrastive mechanisms.

The devil lies in semantically discriminative fine-grained feature extraction for a successful contrastive pre-training. To remedy the inadequacy of fine-grained feature capturing due to failure in localizing to discriminative regions, we propose to empower contrastive learning with localization ability by aligning fine-grained semantic features across augmented views. In particular, we come up with a pre-training framework called Cross-View Saliency Alignment (CVSA). CVSA consists of two algorithmic components: (a) A general plug-and-play data augmentation strategy called SaliencySwap, which swaps the saliency region of the reference image with the saliency region of a randomly selected background image. SaliencySwap ensures semantic consistency between augmented views while introducing background variation. A demonstration of SaliencySwap in comparison with RRC is shown in Figure 1. An alignment loss that provides an explicit localization supervision signal by forcing the model to give the highest correspondence response intensity of the foreground object across views.

On top of the proposed CVSA, to further bridge the performance gap between self-supervised and supervised representation learning on the fine-grained recognition problem, we offer a *dual-stage* pre-training setting, which utilizes coarse-grained datasets for low-level feature extraction and fine-grained datasets for high-level target discrimination and localization. In short, this paper makes the following contributions:

- We delve deep into the knowledge learned by various self-supervised methods compared to supervised methods during the fine-grained pre-training phase and point out the cause of limitations.
- We develop a novel contrastive learning framework for fine-grained recognition, which contains a data augmentation technique called *SaliencySwap* to guarantee semantic consistency between views and an *alignment objective* which enables the model to localize.
- Extensive experiments show consistent performance gain of CVSA under various pre-training stage settings on small- and large-scale fine-grained benchmarks.

# 2 Cross-view Saliency Alignment

We explore the capabilities learned out of three classes of pre-training mechanisms, namely self-supervised contrastive, non-contrastive, and supervised methods. In particular, we focus on discriminative feature extraction and object localization ability. Without loss of generality, we select MoCo.v2, BYOL, Rot-Pred, and supervised classification for comparison. We discover that compared to supervised methods, self-supervised methods show **worse object localization ability** and **discriminative feature extraction ability** is also crucial for fine-grained categorization. The details of the experiments and corresponding analysis are provided in Appendix A. Based on our observations, given a fine-grained classification problem, similar to [1], we assume $\mathcal{X}$ to be a set of all samples with an underlying set of discrete latent classes $\mathcal{C}$ that represent semantic content, we obtain the joint distribution between each sample $x$ and its class $c$:

$$p(c,x) = p(c|x_{fore}) \cdot p(x_{fore}|x), \tag{1}$$

where $x_{fore}$ stands for the foreground object. The intuition behind this factorization suggests that given an image of a fine-grained object, the model should localize the foreground object ($p(x_{fore}|x)$) to discriminate the species of the foreground object ($p(c|x_{fore})$). Following this formulation, we propose a *dual-stage* pre-training pipeline for self-supervised fine-grained recognition with the *first-stage* learning discriminative texture extraction ability and *second-stage* learning localization capability.

## 2.1 SaliencySwap

SaliencySwap maximally utilizes the saliency information for foreground semantic consistency across views while introducing background variation. SaliencySwap guarantees that each view at least contains part of the foreground object and thus prevents the encoder from learning irrelevant feature representation through pure background information.

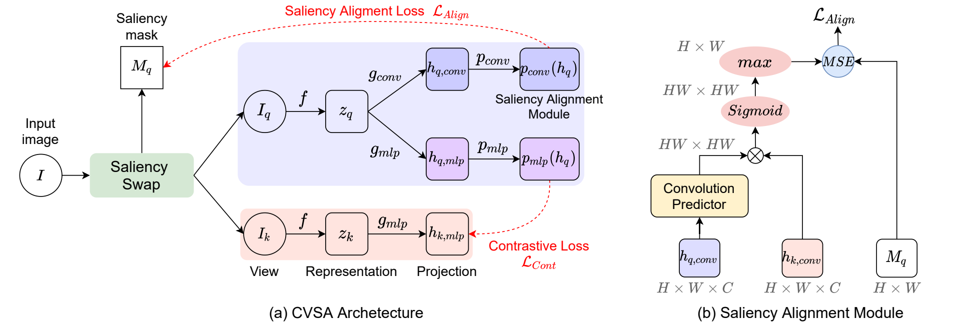


Figure 2: Learning paradigm of Cross-view Saliency Alignment. (a) The network parameters in red are an exponential moving average of the purple part, and the contrastive loss $\mathcal{L}_{Cont}$ is calculated between $p_{q,mlp}$ and $h_{k,mlp}$ (stop-gradient) as BYOL. (b) Saliency alignment module calculates $\mathcal{L}_{Align}$ between predicted attention maps and input saliency masks $M$.

### 2.1.1 Source Saliency Detection

A saliency detection algorithm generates a saliency map that indicates the objects of interest (primarily foreground). Let $I \in \mathbb{R}^{W\times H\times C}$ be an image in the training set, define $\psi$ to be a saliency detection algorithm, then the output saliency map $S_{i,j} = \psi(I_{i,j}) \in \mathbb{R}^{W\times H}$ indicates the saliency intensity value at pixel $I_{i,j}$. The saliency information can be noisy. Therefore, we seek to find a bounding box $B = (l,t,W_b,H_b)$ of the foreground object with the highest averaged saliency information satisfying the following objective function:

$$\underset{W_b,H_b,l,t}{\text{argmax}} \sum_{i=l}^{i=l+W_b} \sum_{j=t}^{j=t+H_b} \frac{S_{i,j}}{W_b \times H_b}. \tag{2}$$

A corresponding binary saliency mask $M \in \mathbb{R}^{W\times H}$ is defined by filling with 1 within the bounding box $B$, otherwise 0. Then we crop a random patch within the bounding box $B$. Similar to RRC, the size of the patch is determined based on an area ratio (to the area of the bounding box), which is sampled from a uniform distribution $U(\lambda, 1)$.

### 2.1.2 Foreground Background Fusion

We then combine the cropped foreground patch from the source image (foreground image) with another randomly selected image (background image). To avoid saliency ambiguity, we restrict each augmented view to contain the saliency information only of one semantic object. We consider two ways of merging: (I) The background dataset is the same as the foreground dataset. The saliency information of the background needs to be eliminated. We first calculate the bounding box $B_f$, $B_b$ of the foreground and background images, respectively, using Eqn. 2. For scenic images, we use the bounding box with maximum area. Then select a random patch from the foreground $B_f$ and resize it to the shape of $B_b$. Finally, we replace $B_b$ with the resized foreground patch. (II) The background dataset is different from the foreground dataset. We choose a dataset like COCO rather than the iconic dataset like IN such that the background dataset is rich in environments. Again, we calculate the bounding box $B_f$ of the foreground and select a random patch. Then we resize the selected patch based on an area ratio (to the area of the background) which is sampled from a uniform distribution $(\beta, 1)$. Finally, we 'paste' the resized patch to a random location in the background.

## 2.2 Cross-view Saliency Alignment

Given two views $I_q$ and $I_k$ of Image $I$ augmented by a pipeline containing SaliencySwap and other augmentation operations such as random flipping and color jittering. Define $M_k$ and $M_q$ to be their saliency masks, respectively. Let $z_q^l$ and $z_k^l \in \mathbb{R}^{W^l \times H^l \times C^l}$ be the $C^l$ dimensional $H^l \times W^l$ feature maps encoded by an encoder network $z = f(I)$ (*e.g.,* ResNet) truncated at stage $l$. We adopt two types of non-linear projector necks $g(\cdot)$ on top of the encoder to form a $d$-dim projection. The two-layer MLP projector generates the following projection $h_{mlp} = g_{mlp}(z)$ as proposed by SimCLR. Similarly, a convolutional projection $h_{conv}$ is given by a convolution projector $g_{conv}$, which consists of two $1 \times 1$ convolution layers with a batch normalization layer and a ReLU layer in between. Following BYOL, two predictor heads $p_{mlp}(\cdot)$ and $p_{conv}(\cdot)$, which have the same network structures except for different input dimensions, are adopted to match the output of one view to the other. Specially, $p_{conv}(\cdot)$ is designed for saliency alignment. Figure 2 (a) shows the overall framework.

**Cross-view Attention.** We seek to capitalize on the pixel-level foreground semantic interactions between the feature maps of two different augmented views. We first build a cross-view attention map:

$$A_{q,k}^l = p_{conv}(h_{q,conv}^l) \otimes {h_{k,conv}^l}^T, \tag{3}$$

where $A_{q,k}^l$ denotes the attention map is of view $k$ w.r.t. view $q$, $T$ denotes matrix transposition, and $\otimes$ denotes matrix multiplication. The location-aware attention map $A^l \in \mathbb{R}^{W^l H^l \times H^l W^l}$ indicates a pair-wise spatial correspondence between any pixel from $p_{conv}(h_{q,conv}^l)$ and any pixel from $h_{k,conv}^l$. Symetrically, we get the attention map $A_{k,q}^l$ of view $q$ w.r.t. view $k$ by interchanging $q$ and $k$ of Eqn. 3.

**Joint Saliency Alignment.** To enhance the encoder's ability to identify the location of the foreground object, we propose to align the saliency mask with a correspondence intensity matrix that captures the pixel-level correlation from the feature map of one view to the other, as shown in Figure 2 (b). The correspondence intensity matrix $C$ is formulated as follows:

$$C_{q,k}^l = \max(\sigma(A_{q,k}^l)), \tag{4}$$

where $\sigma(\cdot)$ denotes sigmoid activation. Note that the shape of $C_{q,k}^l$ is $W^l \times H^l$ and the $max(\cdot)$ operation is performed over the second axis of $A_{q,k}^l$. We then define a symmetrized alignment loss between the saliency mask $M$ and the correspondence intensity matrix $C$:

$$\mathcal{L}_{Align} = \left\| \delta^l(M_q) - C_{q,k}^l \right\|^2 + \left\| \delta^l(M_k) - C_{k,q}^l \right\|^2, \tag{5}$$

where $\delta^l(\cdot) : \mathbb{R}^{W \times H} \to \mathbb{R}^{W^l \times H^l}$ denotes the bilinear downsampling operation. The proposed alignment loss restricts the most cross-view correlated pixels to the saliency region and thus gives the model localization ability. In addition, leveraging cross-layer semantics also enhances the representation of multi-scale learning.

**Joint Objective.** We define a contrastive loss $\mathcal{L}_{Cont}$ with the prediction vector $p_{q,mlp} \stackrel{\text{def}}{=} p_{mlp}(h_{q,mlp})$ and the projection vector $h_{k,mlp}$ using negative cosine similarity $\mathcal{D}(\cdot)$ as:

$$\mathcal{L}_{Cont} = \frac{1}{2}\mathcal{D}(p_{q,mlp}, h_{k,mlp}) + \frac{1}{2}\mathcal{D}(p_{k,mlp}, h_{q,mlp}), \tag{6}$$

where $\mathcal{D}(p,h) = -\frac{p}{\|p\|_2} \cdot \frac{h}{\|h\|_2}$. The joint objective for the *second-stage* pretext task is:

$$\mathcal{L}_{CVSA} = \mathcal{L}_{Cont} + \mathcal{L}_{Align}. \tag{7}$$

# 3 Experiments

## 3.1 Experimental settings

**Implementation details.** We use ResNet as the encoder $f$ followed by two projectors and two predictor sub-networks. During the *first-stage* pre-training, we follow the exact experimental setup in BYOL. For the *second-stage* in *dual-stage* pre-training, the model is initialized with the pre-trained weight from the *first-stage* and the first two stages of the ResNet backbone are frozen. We use a learning rate of $lr \times BatchSize/256$ with $BatchSize = 1024$ and a base $lr$ selected from $\{0.3, 0.6, 0.9, 1.2\}$. The embedding dimension is set to $d = 256$ as BYOL. As for image augmentations, we follow the settings in MoCo.v2 for all contrastive learning methods. We replace *RandomResizedCrop* (RRC) with the proposed SaliencySwap (SS) and adopt all other augmentations in MoCo.v2 [5], detailed in Appendix A. We grid search cropping scale ratio of SaliencySwap $\lambda \in \{0.08, 0.2, 0.5, 0.8\}$ and set $\lambda = 0.5$ by default. To balance the performance and computational cost, we adopt the stage $l = 4$ for the alignment. For simplicity, we use the ground truth bounding box of fine-grained datasets because most saliency detectors are trained in a supervised manner. An ablation study of the performance using different detection algorithms is given in Appendix B. We use the same pre-training setup as in Appendix A.1 for other unstated setups.

**Dataset.** We assess the performance of the representation pre-trained using *dual-stage*, *first-stage* only and *second-stage only* on four small-scale and one large-scale fine-grained benchmarks: 1) CUB-200-2011 [42] (CUB) contains 11,788 images from 200 wild bird species, 2) Stanford-Cars [21] (Cars) contains 16,185 images of 196 car subcategories, 3) FGVC-Aircraft [29] (Aifcrafts) contains 10,000 images of 100 classes of aircrafts, 4) NA-birds [41] (NAbirds) is a large dataset with 48,562 images for over 555 bird classes and 5) iNaturalist2018 (iNat2018) [19] is an unbalanced long tail dataset which contains 437,513 images of 8,142 taxa coming from 14 super-classes. We follow the standard dataset partition in the original works. For the *first-stage* pre-training, we adopt two popular datasets: 1) ImageNet-1k (IN-1k) contains 1.28 million of training images. 2) MS-COCO (COCO) contains 118k images with more complex scenes of many objects.

**Hyperparameters for CVSA.** We search for the optimal hyperparameters for our proposed CVSA on the validation set of fine-grained datasets using ResNet-18 as the encoder. We set the batch size to 1024 and follow other training settings of BYOL [15]. The base learning rate is set to 0.3 for the Cars dataset and 0.6 for the other three datasets. The cropping scale ratio of SaliencySwap is $\lambda = 0.5$ by default. As for the stage $l$ in CVSA, we analyze the performance and the computational cost of using $l \in \{3, 4\}$ since the first two stages are frozen in the *dual-stage*. Furthermore, we find that using $l = 4$ achieves a balance between the performance and the computation cost.

## 3.2 Comparison with State-of-the-art

We choose two hand-crafted methods (Rel-Loc and Rot-Pred), three commonly used contrastive learning methods (SimCLR, MoCo.v2, and BYOL), and three extension methods (LooC*, DiLo, and InsLoc) for comparison. DiLo and InsLoc are reproduced using official code, with other methods reproduced using OpenMixup [24]. LooC* denotes its rotation version reproduced by us. Since our approach extends BYOL, we choose BYOL as the baseline. Notice that **grey** indicates baseline, **green** denotes improvement over baseline while **red** for degradation performance, **bold** denotes the best performance.

| Methods | CUB | NAbirds | Aircrafts | Cars |
|---|---|---|---|---|
| Random | 58.51 (-6.34) | 65.78 (-6.76) | 70.58 (-2.02) | 70.51 (-5.36) |
| Rel-Loc | 65.89 (+1.04) | 72.60 (+0.06) | 72.24 (-0.36) | 75.27 (-0.60) |
| Rot-Pred | 66.67 (+1.82) | 73.01 (+0.47) | 72.67 (+0.07) | 75.48 (-0.39) |
| SimCLR | 63.43 (-1.42) | 72.05 (-0.49) | 72.42 (-0.18) | 75.12 (-0.75) |
| MoCo.v2 | 63.21 (-1.64) | 71.36 (-1.18) | 71.93 (-0.67) | 74.89 (-0.98) |
| LooC* | 66.42 (+1.57) | 72.84 (+0.30) | 72.49 (-0.11) | 75.69 (-0.18) |
| InsLoc | 64.87 (+0.02) | 72.80 (+0.26) | 73.43 (+0.83) | 76.61 (+0.64) |
| BYOL | 64.85 (+0.00) | 72.54 (+0.00) | 72.60 (+0.00) | 75.87 (+0.00) |
| BYOL+DiLo | 66.16 (+1.31) | 73.12 (+0.58) | 73.52 (+0.92) | 76.36 (+0.49) |
| **BYOL+CVSA** | **66.88** (+2.03) | **73.75** (+1.21) | **74.55** (+1.95) | **77.45** (+1.58) |

Table 1: **Comparison of *second-stage only* pre-training on fine-grained benchmarks.** Top-1 accuracy (%) under fine-tuned evaluation is reported. Random denotes random initialized models in the *second-stage*. BYOL is regarded as the baseline by default.

| Methods | Stage 2 | CUB | NAbirds | Aircrafts | Cars |
|---|---|---|---|---|---|
| Rel-Loc | ✓ | 67.33 (-1.22) | 73.82 (-2.47) | 81.20 (-0.53) | 85.80 (-0.56) |
| Rot-Pred | ✓ | 67.75 (-0.80) | 74.26 (-2.03) | 81.58 (-0.15) | 85.74 (-0.62) |
| SimCLR | × | 68.30 (-0.30) | 73.51 (-2.78) | 81.18 (-0.55) | 85.93 (-0.43) |
| MoCo.v2 | × | 68.47 (-0.08) | 73.73 (-2.56) | 81.78 (+0.05) | 86.08 (-0.28) |
| MoCo.v2 | ✓ | 67.60 (-1.05) | 73.18 (-3.11) | 81.04 (-0.69) | 85.71 (-0.65) |
| LooC* | ✓ | 68.71 (+0.16) | 74.45 (-1.84) | 81.75 (+0.02) | 85.90 (-0.46) |
| InsLoc | ✓ | 67.94 (-0.20) | 76.36 (+0.07) | 81.54 (-0.23) | 86.38 (+0.02) |
| BYOL | × | 68.55 (+0.00) | 76.29 (+0.00) | 81.73 (+0.00) | 86.36 (+0.00) |
| BYOL | ✓ | 68.01 (-0.54) | 75.82 (-0.47) | 80.53 (-1.20) | 85.49 (-0.87) |
| BYOL+DiLo | ✓ | 68.70 (+0.15) | 76.94 (+0.65) | 82.04 (+0.31) | 86.46 (+0.10) |
| **BYOL+CVSA** | ✓ | **69.14** (+0.59) | **77.57** (+1.28) | **82.77** (+1.04) | **87.13** (+0.77) |

Table 2: **Comparison of *dual-stage* pre-training on fine-grained benchmarks.** Top-1 accuracy (%) under fine-tuned evaluation is reported. The *first-stage* is pre-trained on COCO, and ✓denotes performing the *second-stage* pre-training on fine-grained datasets.

| Methods | Stage 2 | CUB | NAbirds | Aircrafts | Cars |
|---|---|---|---|---|---|
| Supervised | × | 81.02 | 80.09 | 87.25 | 90.61 |
| SimCLR | × | 73.99 (-2.64) | 76.30 (-2.59) | 85.96 (-1.23) | 88.16 (-1.43) |
| MoCo.v2 | × | 73.19 (-3.44) | 75.64 (-3.25) | 85.49 (-1.70) | 87.51 (-2.08) |
| MoCo.v2 | ✓ | 71.77 (-4.86) | 73.96 (-4.93) | 83.25 (-3.94) | 86.88 (-2.71) |
| InsLoc | ✓ | 75.83 (-0.80) | 78.86 (-0.03) | 86.60 (-0.59) | 88.87 (-0.72) |
| BYOL | × | 76.63 (+0.00) | 78.89 (+0.00) | 87.19 (+0.00) | 89.59 (+0.00) |
| BYOL | ✓ | 72.46 (-4.17) | 76.12 (-2.77) | 84.58 (-3.61) | 87.08 (-2.51) |
| BYOL+DiLo | ✓ | 76.60 (-0.03) | 79.04 (+0.15) | 87.03 (-0.16) | 89.26 (-0.33) |
| **BYOL+CVSA** | ✓ | **77.10** (+0.47) | **79.64** (+0.75) | **87.27** (+0.12) | **89.76** (+0.18) |

Table 3: **Comparison of *dual-stage* pre-training on fine-grained benchmarks.** Top-1 accuracy (%) under fine-tuned evaluation is reported. The *first-stage* is pre-trained on IN-1k, and ✓denotes performing the *second-stage* pre-training on corresponding fine-grained datasets. Supervised denotes the supervised pre-training on IN-1k.

**Small-scale scenarios.** We perform 800 epochs pre-training with ResNet-50 encoder using different pre-training stage settings on four small-scale fine-grained datasets and report the top-1 classification accuracy under the fine-tune protocol. As shown in Table 1, when performing the *second-stage only* setting, namely pre-training with the training set of fine-grained datasets from scratch, our proposed CVSA outperforms BYOL by a large margin on all benchmarks. When applying the *dual-stage* setting using COCO for the *first-stage*, as shown in Table 2, CVSA shows a consistent improvement on representations learned during the *first-stage* while most comparing methods yield worse performance than the *first-stage* pre-trained BYOL baseline. However, when using ImageNet for the *first-stage* pre-training, as shown in Table 3, the improvement of CVSA in the *dual-stage* setting is not as significant on Aircraft and Cars. We hypothesize that this is due to the iconic nature of these two datasets. We find that most images of these two datasets are of similar and straightforward backgrounds, restricting the background variation imposed by our approach. Besides, comparison of Table 1 with Table 2 and 3 suggests that the performance gain is affected by the size of dataset used. This is consistent with our formulation in Appendix A.2 that discriminative feature extraction relies on the size of the pre-training dataset, and localization ability alone is not enough for fine-grained recognition. In a word, CVSA and *dual-stage* pre-training pipeline improve learned representation for the small-scale fine-grained classification.

| Stage 1 | Stage 2 | Sup. | MoCo.v2 | BYOL | BYOL+DiLo | BYOL+CVSA |
|---|---|---|---|---|---|---|
| × | iNat2018 | 5.7 | 41.8(-2.8) | 44.6(+0.0) | 44.3(-0.3) | **45.4**(+0.8) |
| IN-1k | iNat2018 | 46.5 | 45.0(-1.9) | 46.9(+0.0) | 47.0(+0.1) | **47.5**(+0.6) |

Table 4: **Comparison of *dual-stage* pre-training on iNat2018.** Top-1 accuracy (%) under linear evaluation is reported. *Sup.* denotes the supervised pre-training on iNat2018 in the *second-stage*.

**Large-scale scenarios.** Next, we evaluate CVSA on large-scale benchmark iNat2018 with the *dual-stage* pre-training. Following MoCo [17], we adopt the linear evaluation protocol as mentioned in Appendix A.1 training

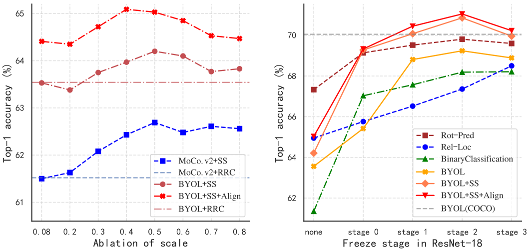


Figure 3: **Ablation of hyperparameters**. Left: the effect of different scaling factors of SS during *second-stage*. Right: the effect of freezing different ResNet stages during *second-stage*. The dotted grey line indicates the performance of BYOL pre-trained on COCO during *first-stage*.

| Methods | Stage 1 | Stage 2 | CUB | Aircrafts | Cars |
|---|---|---|---|---|---|
| BYOL | IN-1k | × | **76.6** | 87.2 | 89.6 |
| **BYOL+CVSA** | IN-1k | × | 75.9 | 86.8 | 89.2 |
| **BYOL+CVSA** | IN-1k | ✓ | 76.5 | **87.3** | **89.7** |
| BYOL | iNat2018 | × | **77.2** | 80.4 | 85.8 |
| **BYOL+CVSA** | iNat2018 | ✓ | 77.0 | **81.2** | **86.5** |
| BYOL | COCO | × | 68.6 | 81.7 | 86.4 |
| **BYOL+CVSA** | COCO | ✓ | **68.9** | **82.3** | **86.7** |

Table 5: **Ablation of *dual-stage* pre-training.** The *first-stage* is pre-trained 200 epochs on IN-1k and iNat2018 and 400 epochs on COCO. The *second-stage* is pre-trained on downstream datasets (denoted by ✓) or not (denoted by ×).

100 epochs with the basic learning rate $lr = 0.025$ and batch size of 256. In Table 4, CVSA outperforms existing methods in both settings indicating its effectiveness in the large-scale datasets, especially improving the baseline by 0.9% with *second-stage-only*. Compared with the results in Table 3, the performance gain is more significant than *second-stage* pre-training on small-scale datasets like Aircraft and Cars, which suggests that CVSA benefits from a larger data size. Meanwhile, using *dual-stage* pre-training with IN-1k for the *first-stage* helps contrastive methods to learn better representations than the *second-stage only* setting.

## 3.3 Ablation Study

We perform 400 epochs pre-training with ResNet-18 on CUB for the first three ablation studies. As for the fourth ablation for the *dual-stage* pre-training, we perform 200 epoch *first-stage* pre-training on large datasets and 400 epochs *second-stage* pre-training on target datasets. The top-1 accuracy under the fine-tune evaluation is reported. Appendix B provides more results.

**Module effectiveness ablation.** We demonstrate the effectiveness of our CVSA by adding modules one by one onto the baseline. We compare the performance of MoCo.v2 and BYOL using **SS** against *RandomResizedCrop* (RRC) while keeping all other augmentations unchanged. As shown in Figure 3 left, **SS** outperforms RRC (using the default scaling factor of 0.08) both for MoCo.v2 and BYOL. We observe a further performance improvement from adding the saliency alignment loss (**Align**) onto BYOL using **SS**. Now, we have shown that both **SS** and **Align** contribute to higher performance.

**Hyperparameter ablation.** We then analyze the performance of **SS** using different crop scaling factors, specifically $\lambda \in \{0.08, 0.2, 0.3, 0.4, 0.5, 0.8\}$ in Figure 3 left. From Figure A1 left, the performance of BYOL fluctuates drastically under the different scaling factors of RRC, with the best result achieved with a scaling factor being 0.08. However, BYOL yields similar performance under different choices of $\lambda$ of **SS**. We argue that **SS**, together with **Align**, helps the model to localize on the foreground object, and thus local texture is no longer the only clue for contrastive pre-training to learn representation. Then, we study the effect when freezing various stages of ResNet-18 in the *second-stage*. We initialize the model of all methods in the *second-stage* with the weights of BYOL baseline pre-trained on COCO in the *first-stage*. In Figure 3 right, the horizontal axis indicates freezing up to different stages of ResNet-18. The best performance is reached when freezing up to the *second-stage* of ResNet. The early stages of ResNet mostly extract low-level texture information, while the later stages are for higher-level features such as discrimination and localization. Intuitively, we wish to enhance the localization ability without jeopardizing the texture extraction ability acquired during the *first-stage*. This explains the reason for freezing the first two stages during the *second-stage*, which is common in object detection [34].

**Background dataset ablation.** Moreover, we compare different foreground and background fusion methods (**SS** and DiLo) based on CUB dataset using *second-stage only* settings. As shown in Table 3.3, replacing the background with other datasets usually results in performance degradation, while fusing with complex backgrounds might improve the performance. We hypothesize that exploring task-specific backgrounds is more critical in fine-grained scenarios. Meanwhile, it indicates that the proposed CVSA (**SS+Align**) can well explore the background of CUB.

| Background | BYOL+DiLo | BYOL+SS | BYOL+SS+Align |
|---|---|---|---|
| CUB | 64.14 (+0.00) | **64.35** (+0.00) | 65.02 (+0.00) |
| COCO | 60.07 (-4.07) | 60.42 (-3.93) | 61.23 (-3.79) |
| NAbirds | 62.51 (-1.81) | 63.06 (-1.29) | 64.80 (-0.22) |
| CUB+COCO | 61.26 (-2.88) | 61.49 (-2.86) | 62.01 (-3.01) |
| CUB+NAbirds | **64.28**(+0.14) | 64.23 (-0.12) | **65.50**(+0.48) |

Table 6: **Evaluation of background datasets extension for the *second-stage* pre-training.** We study the effect of fusing different background datasets on CUB.

**First-stage pre-training ablation.** Additionally, we verify the necessity of the *first-stage* learning discrimination on iconic datasets in the *dual-stage* pre-training scheme. We compare naive BYOL and CVSA for the *first-stage* pre-training on IN-1k, iNat2018, and COCO. As shown in Table 5, using IN-1k (iconic) for the *first-stage* yields better performance that scenic datasets, indicating the necessity of the *first-stage* on iconic datasets. Since CVSA is design for the *second-stage* on target datasets, we find that using CVSA on *second-stage* outperforms naive BYOL while producing degraded performance on the *first-stage*.

**Visualization of localization abilities.** Lastly, we compare the localization abilities by Grad-CAM [35] visualization of SS (BYOL+CVSA) and RRC (BYOL). Figure 4 (a) and (b) shows that SS enables the model to better localize the fine-grained target than RRC. Comparing to Figure 4 (a)(b), Figure 4 (c) shows that *dual-stage* training further improves localization than the *first-* or *second-stage only*.

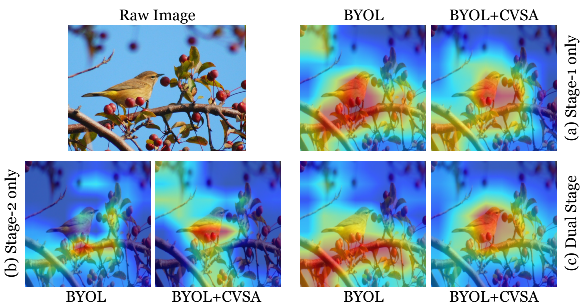


Figure 4: Grad-CAM visualization of BYOL and BYOL+CVSA on CUB200.

# 4 Related Work

Self-supervised methods have largely reduced the performance gap between supervised models on various downstream vision tasks. Most early methods design hand-crafted pretext tasks [10, 13, 56]. For example, Gidaris et al. [13] proposed learning image features by training the network to predict different rotation angles of the same image. Doersch*et al.* [10] proposed learning image features by training the network to predict the correct order of random cut image patches. These pretext tasks rely on somewhat ad-hoc heuristics, which limits the generality of learned representations.

Recently, contrastive learning [3, 4, 5, 17, 23, 51, 52] achieved state-of-the-art performance, which learns instance-level discriminative representations by contrasting positive pairs against negative pairs. In particular, MoCo [17] adopts a memory bank to store negative samples while embedding different views of the same image using an online encoder and a momentum encoder. SimCLR [4] yields comparable results using the sufficiently large batch size in replacement of the memory bank. Getting rid of the notion of negative pairs, BYOL [15] pulls together positive pairs generated by online and momentum encoders with the help of a predictor and the stop gradient mechanism. Another popular form of self-supervised learning is clustering-based methods [2, 3, 54]. Without computing pairwise comparisons, SwAV [3] maps image features to a set of learnable prototype vectors. Various mechanisms are proposed to learn useful representations rather than trivial solutions in contrastive methods such as a constant representation.

More recently, some research endeavors have been made on top of contrastive methods to enhance pre-training quality for specific downstream tasks, such as object detection and segmentation [25, 26, 45, 47, 48]. Most methods adopts detection or segmentation components to learn pixel-level contrastive representation [8, 11, 37, 49]. DSC [25] proposed to model pixel-level semantic structures within images by taking into consideration the semantic relations of both intra- and inter-image pixels. Self-EMD [26] learns representations by measuring the similarity among all location pairs using the earth mover's distance (EMD). LooC [39] proposed to construct separate embedding sub-spaces for each augmentation instead of a single embedding space. DiLo [57] proposed a copy-paste [12, 33] based augmentation approach that randomly pastes masked foreground onto a variety of backgrounds. CASTing [37] crops views based on a ratio threshold of the area of saliency regions (required mask-level supervision) to the area of the cropped patch. However, existing methods are trained on general-purpose coarse-grained datasets while neglecting fine-grained scenarios where low-level background texture features provide little clue to the category information of the foreground subject. To address this issue, we propose a *dual-stage* pre-training pipeline that utilizes coarse- and fine-grained datasets for better fine-grained representation learning. Appendix C provides a discussion of the relationship between our proposed CVSA and previous methods.

Current efforts [9, 20, 38, 46, 58, 59] in fine-grained recognition are primarily dedicated to fine-tuning model pre-trained on supervised ImageNet either by localizing distinct parts or by learning fine-grained features. However, there exists little exploration in self-supervised pre-training for fine-grained categorization. In the paper, we attempt to bring localization and fine-grained feature representation learning to the pre-training stage, using fine-grained datasets. To address this issue, we propose a *dual-stage* pre-training pipeline that utilizes coarse- and fine-grained datasets for better fine-grained representation learning.

# 5 Conclusion

In this paper, we find that learning to extract discriminative texture information and localization are equally crucial for fine-grained self-supervised pre-training with empirical analysis. We proposed a *dual-stage* pre-training pipeline with the *first-stage* to train feature extraction and the *second-stage* to train localization. To empower the model with localization abilities in the *second-stage*, we propose *cross-view saliency alignment* (CVSA), a new unsupervised contrastive learning framework. Extensive experiments on fine-grained benchmarks demonstrate the effectiveness of our contributions in learning better fine-grained representations.

# Acknowledgement

This work is supported by the Science and Technology Innovation 2030- Major Project (No. 2021ZD0150100) and the National Natural Science Foundation of China (No. U21A20427). We thank Zicheng Liu, Yifan Zhao, and all reviewers for polishing the writing.


# References


[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *International Conference on Machine Learning (ICML)*, 2019.

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

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

[6] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[8] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pretraining for object detection with transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[9] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6599–6608, 2019.

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[11] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[12] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, 2021.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[14] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742. IEEE, 2006.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8662–8672, 2020.

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[23] Siyuan Li, Zelin Zang, Di Wu, Zhiyuan Chen, and Stan Z. Li. Genurl: A general framework for unsupervised representation learning. *ArXiv*, abs/2110.14553, 2021.

[24] Siyuan Li, Zedong Wang, Zicheng Liu, Di Wu, and Stan Z. Li. Openmixup: Open mixup toolbox and benchmark for visual representation learning. https://github.com/Westlake-AI/openmixup, 2022.

[25] Xiaoni Li, Yu Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1368–1376, 2021.

[26] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020.

[27] Zicheng Liu, Siyuan Li, Di Wu, Zhiyuan Chen, Lirong Wu, Jianzhu Guo, and Stan Z Li. Automix: Unveiling the power of mixup. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022.

[28] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *International Conference on Learning Representations*, 2018.

[29] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[30] Sebastian Montabone and Alvaro Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3):391–402, 2010.

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016.

[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Tal Remez, Jonathan Huang, and Matthew Brown. Learning to segment via cut-and-paste. In *Proceedings of the European conference on computer vision (ECCV)*, pages 37–52, 2018.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015.

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[36] Ramprasaath R. Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11058–11067, 2021.

[37] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11058–11067, 2021.

[38] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1143–1151, 2015.

[39] Xiao Tete, Wang Xiaolong, A. Efros Alexei, and Darrell Trevor. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[40] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

[41] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604, 2015.

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[43] Bin Wang and Piotr Dudek. A fast self-tuning background subtraction algorithm. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 401–404, 2014.

[44] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 29–42, 2012.

[45] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. *arXiv preprint arXiv:2103.12902*, 2021.

[46] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.

[47] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[48] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[49] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 6023–6032, 2019.

[51] Zelin Zang, Siyuan Li, Di Wu, Ge Wang, Lei Shang, Baigui Sun, Hao Li, and Stan Z. Li. Dlme: Deep local-flatness manifold embedding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[52] J. Zbontar, J. Li, I. Misra, Y Lecun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[53] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1476–1485, 2019.

[54] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[57] Nanxuan Zhao, Zhirong Wu, Rynson W.H. Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[58] Yifan Zhao, Ke Yan, Feiyue Huang, and Li Jia. Graph-based high-order relation discovery for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[59] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5012–5021, 2019.

[60] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2814–2821, 2014.