Semantic-Aware Auto-Encoders for Self-supervised Representation Learning

Guangrun Wang[1] Yansong Tang[2,1] Liang Lin[3] Philip H.S. Torr[1]

[1] University of Oxford [2] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University [3] Sun Yat-sen University

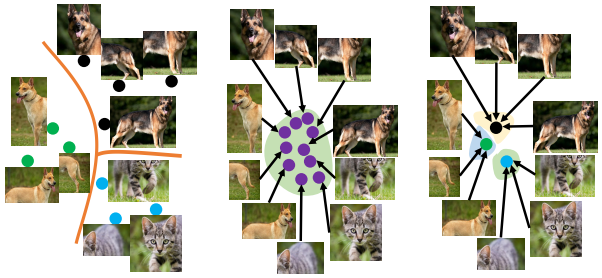
{guangrun.wang,philip.torr}@eng.ox.ac.uk, tang.yansong@sz.tsinghua.edu.cn, linliang@ieee.org

Abstract

The resurgence of unsupervised learning can be attributed to the remarkable progress of self-supervised learning, which includes generative ($\mathcal{G}$) and discriminative ($\mathcal{D}$) models. In computer vision, the mainstream self-supervised learning algorithms are $\mathcal{D}$ models. However, designing a $\mathcal{D}$ model could be over-complicated; also, some studies hinted that a $\mathcal{D}$ model might not be as general and interpretable as a $\mathcal{G}$ model. In this paper, we switch from $\mathcal{D}$ models to $\mathcal{G}$ models using the classical auto-encoder (AE). Note that a vanilla $\mathcal{G}$ model was far less efficient than a $\mathcal{D}$ model in self-supervised computer vision tasks, as it wastes model capability on overfitting semantic-agnostic high-frequency details. Inspired by perceptual learning that could use cross-view learning to perceive concepts and semantics[1]*, we propose a novel AE that could learn semantic-aware representation via cross-view image reconstruction. We use one view of an image as the input and another view of the same image as the reconstruction target. This kind of AE has rarely been studied before, and the optimization is very difficult. To enhance learning ability and find a feasible solution, we propose a semantic aligner that uses geometric transformation knowledge to align the hidden code of AE to help optimization. These techniques significantly improve the representation learning ability of AE and make self-supervised learning with $\mathcal{G}$ models possible. Extensive experiments on many large-scale benchmarks (*e.g., *ImageNet, COCO 2017, and SYSU-30k) demonstrate the effectiveness of our methods. Code is available at* https://github.com/wanggrun/Semantic-Aware-AE.

1. Introduction

Learning representations without human annotations is a long-standing vision full of expectations [5]. Although experiencing a downturn, it has gained a renaissance. Recently, the resurgence of unsupervised learning is attributed to the remarkable process of self-supervised learning (SSL), which can be divided into two groups, *i.e.*, generative models ($\mathcal{G}$ models) and discriminative models ($\mathcal{D}$ models).



(a) $\mathcal{D}$ model (b) semantic-agnostic $\mathcal{G}$ model (c) semantic-aware $\mathcal{G}$ model

Figure 1. A comparison among a discriminative model ($\mathcal{D}$ model), an existing semantic-agnostic generative model ($\mathcal{G}$ model), and our semantic-aware generative model.

In computer vision, the mainstream SSL algorithms belong to $\mathcal{D}$ models that learn representations via agent tasks, e.g., patch ordering [18], solving jigsaw puzzles [43], and rotation prediction [23]. Of all the agent tasks, contrastive learning [12–14, 25, 27] and metric learning [61] are currently the most successful, which randomly augments each image into different views and compares the (dis)similarity between different views (see Figure 1 (a)). But as pointed out by [4,27,61], without careful design, a contrastive learning algorithm would collapse. Special regularizations (*e.g., losses* [4], *normalizations* [27], *centering* [10]), unusual optimizations (*e.g., gradient stopping* [14], *mean teacher* [51]), and non-trival architectures (*e.g., additional predictors* [25]) that are difficult to explain are often needed. Besides, some studies also suggested that a $\mathcal{D}$ model might hold some disadvantages compared to a $\mathcal{G}$ model in generalization and interpretability [3, 6, 26]. Specifically, $\mathcal{G}$ models might be more effective in pretraining foundation models [6] for fine-tuning tasks or downstream tasks, and the development of $\mathcal{G}$ models helps unify the pretraining paradigms in the CV and NLP domains [3, 17]. Moreover, with $\mathcal{G}$ models, one can further conduct a counterfactual intervention for explainability [1].

In this paper, we switch from $\mathcal{D}$ models to $\mathcal{G}$ models

[1]Following [26], we refer to semantics as visual concepts, *e.g.*, a semantic-ware model indicates the model can perceive visual concepts, and the learned features are efficient in object recognition, detection, etc.

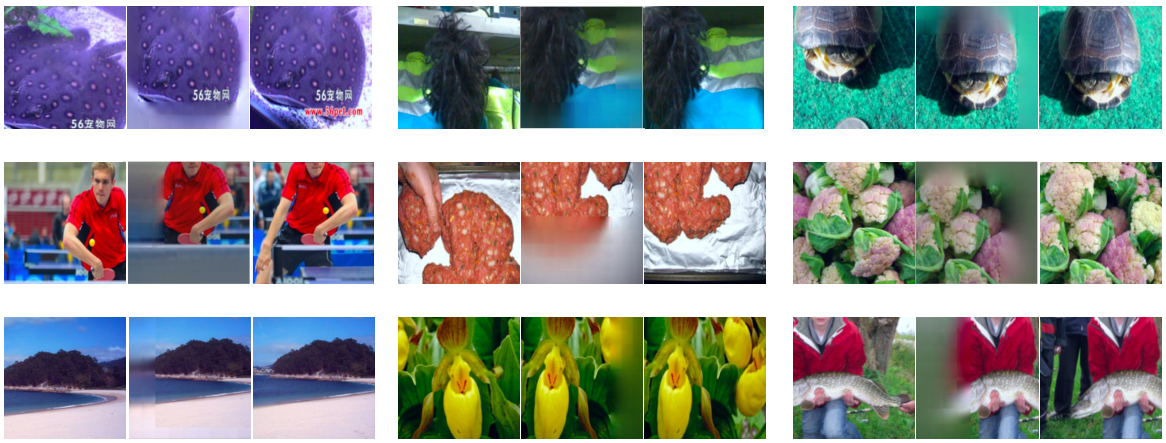


Figure 2. Examples of the cross-view image generation by our semantic-aware AE on the *validation set*. In each triplet, the left is the input, the middle is our generated result, and the right is the reconstruction target. The generated image is similar to the reconstruction target. Although some may be slightly different from the reconstruction target, the generated images are reasonable (semantically plausible).

using classical auto-encoders (AEs)[2]. Note that previous works seldom used a $\mathcal{G}$ model because it was not as efficient as a $\mathcal{D}$ model. For example, a typical $\mathcal{G}$ model BigBiGAN [19] with ResNet-50 [30] achieved a 55.4% top-1 accuracy on the task of linear evaluation on ImageNet [48], which is 20.5 points lower than the triplet loss model [61], one of the best-performing $\mathcal{D}$ models. Similarly, the *latest* $\mathcal{G}$ model BEiT [3] with basic DeiT [52] only obtained 56.7% top-1 accuracy, still holding a 19.2% disadvantage compared to the $\mathcal{D}$ model. Actually, BEiT, iGPT [11], and MAE [26] can only perform well in pretraining tasks but fail to do well in direct discriminative representation learning tasks, *e.g.*, a linear evaluation on ImageNet. A $\mathcal{G}$ model's inefficiency is caused by the waste of capability on overfitting semantic-agnostic local high-frequency details [3, 11, 19] and the ignorance to high-level semantics. For instance, a traditional AE uses an image as an input and the same image as the regression target, making the model overly focused on semantic-agnostic information compression rather than visual concepts (see Figure 1 (b)).

To make $\mathcal{G}$ models feasible in SSL, we need to tackle the above semantic agnosticism problem. Fortunately, Becker & Hinton (1992) found that cross-view learning could enable models to perceive concepts and semantics, and they proposed perceptual learning [5]. Inspired by this prior art, we propose a novel AE that could learn semantic-aware representation via cross-view image reconstruction. We take a view of the image as input and force the AE to reconstruct another view of the image (see Figure 1 (c)). However, this rarely-explored AE model is hard to optimize in practice. To solve this problem, we further propose a novel semantic alignment technology. Using the geometric transformation knowledge, we can adjust the hidden code of AE to ensure that the code semantic is aligned with the reconstruction target, thereby improving the learning and optimization capabilities. These techniques significantly improve the representation learning ability of AE and make SSL with $\mathcal{G}$ models possible in computer vision, leading to a state-of-the-art performance in feature learning, generalizability, and explainability. Figure 2 shows some results of our cross-view image generation, which are promising.

In summary, our contributions are three-fold.

- We seek the possibility of replacing $\mathcal{D}$ models with $\mathcal{G}$ models in SSL in computer vision. We rethink the inefficiency of $\mathcal{G}$ models from the perspective of overfitting semantic-agnostic local high-frequency details and propose a novel semantic-aware AE inspired by perceptual learning. Our AE uses one image view as input and another view of the same image as the reconstruction target, which is rarely explored before.

- To help semantic-aware AE optimization, we propose a novel semantic alignment technique that uses the geometric transformation knowledge to semantically align the hidden AE codes to the reconstruction target. These technologies significantly improve the representation learning ability of AE and make the SSL with $\mathcal{G}$ model possible in computer vision.

- Extensive experiments show state-of-the-art performance of our method on several large-scale benchmarks (*e.g.*, ImageNet [48], SYSU-30k [59], and COCO 2017 [40]) and varieties of tasks, demonstrating the effectiveness (e.g., feature learning, generalizability, and interpretability) of our method.

[2]Sometimes, there is a minor controversy about whether a vanilla AE counts as a $\mathcal{G}$ model. However, the community reached a consensus that AE new varieties like denoising AE [53], masked AE [3, 26], and variational AE [35] are $\mathcal{G}$ models, because they can generate things that are not included in an input, *e.g.*, our semantic-aware AE can generate a new image with a different angle from the input. These AE generators are similar conditional generators in GAN [24], *e.g.*, Conditional GAN [33], CycleGAN [66], and StyleGAN [34], with inputs being conditions.

2. Related work

SSL. The idea of unsupervised representation learning dates back to many years ago, *e.g.*, classical clustering [42]. It is highly anticipated at first, and then disappoints people due to unsatisfactory performance. Although experiencing a downturn, it has recently gained a renaissance. The resurgence of unsupervised learning is attributed to the enormous process of SSL, which achieved massive success in both NLP [7, 17] and computer vision [12–14, 25, 27, 61]. Generally, SSL can be divided into two groups, *i.e.*, generative ($\mathcal{G}$) and discriminative ($\mathcal{D}$) models.

$\mathcal{G}$ models in SSL. Although $\mathcal{G}$ models achieve good performance in pretraining language models in NLP [7, 17, 20], they are less effective in SSL tasks in computer vision. The **first** group of $\mathcal{G}$ models in computer vision are proxy generative tasks. Typical works include image denoising [53], image inpainting [44], and color jittering [64]. Although they contribute to the renaissance of SSL, their learned representations do not generalize well. The **second** group are pure generative methods. Representative work is BigBiGAN [19]. Its original intention is generative rather than discriminative representation learning; thus, its learned features are not very helpful in image recognition tasks. The **third** group are NLP-inspired generative models. Outstanding works include BEiT [3] and iGPT [11], which are inspired by BERT [20] and GPT [7], respectively. iGPT uses image tokens as inputs and targets, which are relatively low-level codes. Although BEiT combines masked language modeling with DALL·E codes [46], DALL·E codes still contain local dependencies used for image reconstruction. Thereby, these simple regression tasks (*e.g.*, BEiT, iGPT, and MAE [26]) have difficulties in capturing high-level semantics. Hence, they can only perform well in pretraining tasks but fail to do well in direct discriminative representation learning tasks, *e.g.*, a linear evaluation on ImageNet.

Note that most of the above methods are AEs, whether based on CNN [37] or transformers [21]. The idea of AE dates back to several decades ago at an unclear/untraceable starting point [2, 36, 47]. Basically and traditionally, AEs are employed for generative representation learning whose purpose is dimensionality reduction. However, AEs are inefficient in discriminative representation learning because it wastes model capability on overfitting semantic-agnostic local high-frequency details [3, 11, 19] for reconstruction.

$\mathcal{D}$ models in SSL. Similar to $\mathcal{G}$ models, the **first** group of $\mathcal{D}$ models in computer vision is proxy discriminative tasks. Typical works included patch ordering [18], solving jigsaw puzzles [43], and rotation prediction [23]. Since there was a gap between the proxy tasks and the main mission, their learned representations could not generalize well. The **second** group are currently the most effective methods, i.e., contrastive [12–14, 25, 27] and metric learning [61], which could be traced back to perceptual learning [5]. Lacking annotations, perceptual learning uses cross-view agreement to perceive concepts and semantics. Following it, contrastive-metric learning randomly augments each image into different views and compares the similarity between views. However, this way of representation learning often collapses [4, 27, 61]. To stabilize learning, careful designs are required. SimCLR [12] employs multi-node computing to enlarge the batch. MoCo v1/v2 [13,27], triplet loss model [61], BYOL [25], DINO [10], and SimSiam [14] need gradient-free teachers (e.g., mean teachers [51] or gradient-stopping teachers). Triplet loss model, BYOL, and SimSiam [14] need additional predictors. DINO needs centering and sharpening of mean teachers. Most above methods benefit from synchronous batch normalization [32]. Although the recent VICReg [4] needs no normalization or predictor, it requires three special losses (*i.e.*, variance, invariance, and covariance loss) for regularization. Also, the hyper-parameters in VICReg are uneasy to tune, and training VICReg is sometimes unstable. In summary, designing a workable $\mathcal{D}$ model could be over-complicated.

Besides, the literature hinted that $\mathcal{D}$ models might be less general and explainable than $\mathcal{G}$ models [3, 6, 26]. Specifically, $\mathcal{D}$ models are less effective than $\mathcal{G}$ models in pretraining foundation models for fine-tuning tasks or downstream tasks [6], and $\mathcal{D}$ models hold a gap in the pretraining paradigms between computer vision and NLP domains [3, 17]. Moreover, $\mathcal{D}$ models have poorer interpretability than $\mathcal{G}$ models, e.g., in performing causal inference [1].

3. Method

3.1. Vanilla AE (semantic-agnostic AE)

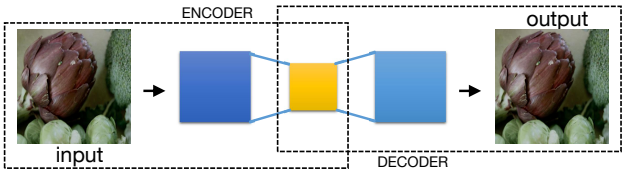
AE is a classic model in the field of representation learning. Basically, a vanilla AE includes two modules, *i.e.*, an encoder and a decoder (see Figure 3 (a)), which can be defined with two mappings g and f respectively, such that:

$$\begin{aligned} &f: z \to h, \qquad g: h \to z, \\ &f^*, g^* = \arg\min_{f,g} \mathcal{L}(z, (g \circ f)(z)), \end{aligned} \tag{1}$$

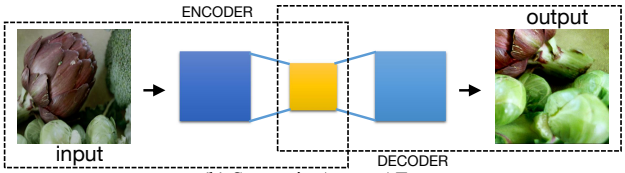
where $\circ$ denotes a composite function, and $\mathcal{L}$ denotes a loss function that can minimize the reconstruction errors (such as squared errors). f^* and g^* are the trained encoder and decoder. This formula shows that AE is to learn two complex mappings to minimize the error between input z and output $(g \circ f)(z)$. Therefore, the essential goal of AE is to learn a representation for information compression, and this representation learning is semantically ignorant.

3.2. Semantic-aware AE

To obtain semantic-aware AE, predecessors have made many efforts and proposed some outstanding works, including variational AE [35] and masked AE [3]. But these works



(a) Vanilla AE (semantic-agnostic AE)



(b) Semantic-Aware AE

Figure 3. A comparison between vanilla AE (*i.e.*, semantic-agnostic AE) that achieves self-reconstruction and our semantic-aware AE achieving cross-view generation.

are still inefficient in capturing semantics because they waste too much capacity on overfitting local high-frequency signals. We ask: *could we learn sematic-aware representations without labels?* Fortunately, this problem has been studied by Becker & Hinton (1992). They found that cross-view learning could enable models to perceive concepts and semantics, and they proposed perceptual learning [5]. Specifically, they perform two independent random augmentations on each image to get two different views. Then, sematic-aware representations are obtained by learning the similarity of two views of the same image and the dissimilarity of different images. Recently, perceptual learning has been the basis of $\mathcal{D}$ models (i.e., contrastive learning).

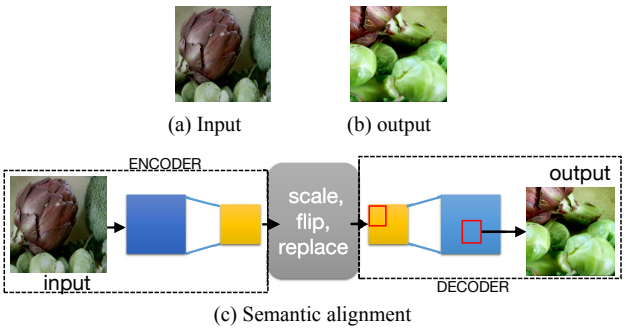
Inspired by perceptual learning, we propose a novel AE that could learn semantic-aware representations via a new cross-view learning, i.e., cross-view image generation. As shown in Figure 3 (b), we perform two independent random data augmentations $\mathcal{T}_1$ and $\mathcal{T}_2$ for each image x to obtain z_1 and z_2, such that: $z_1 = \mathcal{T}_1(x)$ and $z_2 = \mathcal{T}_2(x)$. Then, our goal is to reconstruct z_2 from z_1. Overall, our semantic-aware AE is written as:

$$\begin{aligned} &z_1 = \mathcal{T}_1(x), \qquad z_2 = \mathcal{T}_2(x), \\ &f : z_1 \to h, \qquad g : h \to z_2, \\ &f^*, g^* = \arg\min_{f,g} \mathcal{L}(z_2, (g \circ f)(z_1)). \end{aligned} \tag{2}$$

3.3. Semantic alignment

Empirically, we found that it is extremely difficult to optimize the objective in Eqn. (2) straightforwardly; the training loss cannot converge (see Section 6). Therefore, it is uneasy to learn an effective semantic-aware representation with this formula. To help optimization, we introduce a novel semantic aligner below.

Geometric transformation. To illustrate the semantic alignment process, we first present the difference between the input view (Figure 4 (a)) and the expected output view



(c) Semantic alignment

Figure 4. An illustration of semantic alignment. (a-b) illustrate the geometric transformation between the input and the reconstruction target. (c) illustrates the semantic alignment process.

(*i.e.*, reconstruction target, see Figure 4 (b)). The geometric transformation between them are summarized as follows. (1) Random cropping: The two views are cropped from different locations of the original image, so the regional knowledge of interest is different. (2) Random scaling: The two views have different receptive fields due to different zooming scales. (3) Random flipping: There is a random difference in horizontal flip for the two views, so their mirror perception ability is also different.

Our semantic aligner is placed at the end of the encoder, *i.e.*, on the feature map. Specifically, we send the input z_1 into the encoder and get the intermediate code $f(z_1)$, which is in the form of a feature map. Then, having the geometric transformation between the input and the reconstruction target, we perform the following semantic alignment steps: (1) Zoom in/out ($\mathcal{Z}$, *i.e.*, the "scaling" in Figure 4 (c)): We resize the cropped code into the target size, *i.e.*, $(\mathcal{Z} \circ f)(z_1)$. (2) Flipping ($\mathcal{F}$): We perform horizontal flipping for the cropped code, *i.e.*, $(\mathcal{F} \circ \mathcal{Z} \circ f)(z_1)$. (3) Replacing ($\mathcal{R}$): we replace the transformed code in the appropriate coding position (the position corresponding to the area of interest in the reconstruction target), *i.e.*, $(\mathcal{R} \circ \mathcal{F} \circ \mathcal{Z} \circ f)(z_1)$. Finally, the decoder will decode the latent code into the output space, where a reconstruction image can be obtained via cropping, *i.e.*, $(\mathcal{C} \circ g \circ \mathcal{R} \circ \mathcal{F} \circ \mathcal{Z} \circ f)(z_1)$. An illustration of our semantic alignment process is shown in Figure 4 (c).

3.4. Technical implementation details

Encoder. Our encoder is a standard ViT [21]. We implement ViT following DeiT [52], which is the same as that in [3, 10]. Our ViT architecture is the most widely-used basic one (*i.e.*, ViT-base). For more detail about the architecture protocol, please refer to Section 4.1.

Decoder. Our decoder g is composed of several simple transformer blocks. Having the output $(\mathcal{C} \circ g \circ \mathcal{R} \circ \mathcal{F} \circ \mathcal{Z} \circ f)(z_1)$ in Section 3.3, we can easily calculate the reconstruction error to get the reconstruction loss $\mathcal{L}(z_2, (\mathcal{C} \circ g \circ \mathcal{R} \circ \mathcal{F} \circ \mathcal{Z} \circ f)(z_1))$, and use it to optimize our goal in Formula (2). Please see Section 6 for more details.

Data augmentation. Our data augmentation methods are all standard methods, and the method we use is similar to the standard ImageNet supervised training, or even less. We did not use label smoothing [50] and did not use dropping paths [31]. For more data augmentation information, please see the "Training protocol" part of Section 4.1.

Others. Besides, we have also thoroughly discussed the exclusion of global features in Section 6.

4. Main results

4.1. General protocols

Architecture protocol. As is implied by [3, 10, 16], SSL has more potential in a vision transformer (ViT) [21], and thus a ViT is an ideal architecture to examine the effectiveness of SSL. Thereby, we focus on VITs to test the effectiveness of our method (see footnote[3] for more reasons), similar to [3, 10, 16, 26]. We implement the ViT following DeiT [52], which is the same as that in [3, 10]. Our ViT architecture is the most widely-used basic one (*i.e.*, ViT-base): Specifically, it takes a 224x224 image as input, and then divides it into 14x14 patches, each of which has 16x16 pixels. The encoder embeds each patch with a linear mapping into an embedding (including position codings), and then flattens these embeddings into 196 sequential vectors. These sequential vectors are input into 12 transformer blocks for processing. The hidden size of the transformer block is 768.

Training protocol. In the unsupervised representation learning stage, we trained the model for 400 epochs, and we used the AdamW optimizer. Our learning rate adjustment scheme is cosine descent, and the basic learning rate is 5e-4. The learning rate is linearly warmed up during the first then epochs to its base. We set the weight decay factor as 4e-2. Our data augmentation is very common, including random scaling, cropping, and horizontal flips. Our model is trained using 1.28M training images on ImageNet, but we did not access its annotation information. Thanks to the simplicity of our method, we don't need complex training techniques as $\mathcal{D}$ models did. See our code for more details.

Evaluation benchmarks. Since ViTs are data-hungry, we verify the effectiveness of our method by comparing it with the best existing methods in three large-scale benchmarks. The tasks include linear evaluation on ImageNet [48] (1.28M images), person re-ID on SYSU-30k [59] (30M images), object segmentation and instance detection evaluation on COCO 2017 [40] (123K images, 900K instances).

[3]We did not use a CNN for three reasons. First, CNNs typically operate on regular grids, and it is not straightforward to integrate 'indicators' such as mask tokens into CNNs, which are required in our framework. Second, cross-view generation is a tough learning task that requires a big model to overfit it. ViTs are very big ($\gg$ResNets) and tends to overfit [15, 26, 52], so we choose ViTs rather than ResNets. Third, some recent leading SSL works also only study ViTs (e.g., MoCov3 [16], BEiT [3], MAE [26], and iBOT [65]), and our choice is in line with them.

Critical discussion on evaluation protocol. At present, linear evaluation is the most mainstream criterion to test the ability of SSL methods. The standard process is first to use an SSL method to train a backbone network and then freeze the trained model parameters. Next, people add a linear classifier to the top of the frozen backbone network and merely train this linear classifier for evaluation. [26] thinks that the occasional inconsistency between linear evaluation and fine-tuning indicates that fine-tuning should be emphasized (see footnote[4]). In contrast, we hold a slightly different opinion and believe that linear evaluation/probing can highly measure representation ability.

- As shown in [10], representations with good linear evaluation performance can even be used to segment the object directly in an unsupervised manner.
- Classic visual matching tasks like face recognition and person re-ID also use linear query methods.
- Using fine-tuning will bring new unfairness because fine-tuning is highly dependent on hyperparameters (*e.g.*, learning rate and training epochs). Fine-tuning will change the original parameters, and it is uneasy to standardize the extent to which parameters are allowed to be changed for evaluation. In fact, fine-tuning itself is still an open question [28, 41]. (see footnote[5])

What causes the occasional inconsistency between fine-tuning and linear evaluation may be complicated and multifold (see footnote[6] for detail). We believe that the widely-used linear evaluation should still be valued by the community. Hence, we adopt this widely-used linear probing on ImageNet and SYSU-30k. Neverthelss, we also report the fine-tuning results on ImageNet. In addition, we follow the standard protocol to adopt fine-tuning evaluation on COCO.

4.2. Linear evaluation on ImageNet

We evaluate linear probes on ImageNet. In the unsupervised learning stage, we perform SSL with the 1.28M pictures on the ImageNet training set (but without their annotations). In the linear evaluation stage, all SSL methods are trained on the same training set and validated with 50k images on the validation set. The standard batch size is 256. A total of 100 epochs of training is adopted. This process does not use weight decay. The top-1 accuracy of the single-scale-center-crop scheme is used as the evaluation metric. Please see our code for more details.

[4]In fact, the linear evaluation accuracy of [26] is not very satisfactory.

[5]In [26], the fine-tuning learning rate is **TEN TIMES** larger than that of supervised training from scratch (*i.e.*, 1e-3 vs. 1e-4). Their parameters are changed to a great extent due to the large fine-tuning learning rate.

[6]**One possible reason** is that the model with good linear evaluation performance might ideally fit the source task, so a different learning rate is needed in the fine-tuning stage [61]. **Another possible reason** is that it is likely that a $\mathcal{G}$ model has learned more local dependencies that are beneficial for middle-level tasks (e.g., segmentation and detection tasks).

Table 1. Top-1 accuracy and training epochs of state-of-the-art methods on ImageNet using linear evaluation.

Method	Top-1	Epochs	Backbone	#param.
$\mathcal{D}$ models				
Random	4.4	0	R50	23M
Ordering [18]	38.8	200	R50	23M
Rotation [23]	47.0	200	R50	23M
DeepCluster [8]	46.9	200	R50	23M
NPID [62]	56.6	200	R50	23M
ODC [63]	53.4	200	R50	23M
SimCLR [12]	60.6	200	R50	23M
SimCLR [12]	69.3	1000	R50	23M
MoCo [27]	61.9	200	R50	23M
MoCo v2 [13]	67.0	200	R50	23M
MoCo v2 [13]	71.1	800	R50	23M
SwAV [9]	72.7	200	R50	23M
BYOL [25]	71.5	200	R50	23M
BYOL [25]	72.5	300	R50	23M
BYOL [25]	74.3	1000	R50	23M
SimSiam [14]	68.1	100	R50	23M
SimSiam [14]	70.0	200	R50	23M
SimSiam [14]	70.8	400	R50	23M
SimSiam [14]	71.3	800	R50	23M
Triplet [61]	75.9	700	R50	23M
DINO [10]	78.2	400	ViT-base	86M
MoCo v3 [16]	76.7	600	ViT-base	86M
$\mathcal{G}$ models				
BigBiGAN [19]	55.4	-	R50	23M
iGPT [11]	65.2	-	ViT-super	1362M
iGPT [11]	65.2	-	ViT-super	1362M
BEiT [3]	56.7	800	ViT-base	86M
MAE [26]	68.0	1600	ViT-base	86M
Ours	**70.1**	400	ViT-base	86M

Comparison with $\mathcal{G}$ models. As a new $\mathcal{G}$ model, we first compare our method with the existing $\mathcal{G}$ models. Due to the low efficiency of $\mathcal{G}$ models in recognition tasks, people were rarely interested in this task and thus, there are only a handful of $\mathcal{G}$ models that can be compared, including BigBiGAN [19], iGPT [11], BEiT [3], and MAE [26]. The comparison results are reported in Table 1. The performance of our method far exceeds the existing $\mathcal{G}$ models. For example, our accuracy is 70.1%, which is significantly higher than the best method before (*e.g.*, 56.7% for BEiT). Note that the performance of the concurrent work MAE is also lower than our method (*i.e.*, 68.0% vs. 70.1%). These comparisons confirmed the superiority of our approach.

Comparison with $\mathcal{D}$ models. In Table 1, we compare our method with the state-of-the-art methods. That is, our $\mathcal{G}$ model is compared with $\mathcal{D}$ models. We can observe that $\mathcal{G}$ models still have a long way to go. They have significantly lower learning efficiency than $\mathcal{D}$ models. But as shown in the table, our method is the closest to the $\mathcal{D}$ models. It is worth noting that DINO uses an extra multi-crop scheme that we didn't adopt, which has a significant gain (without multi-crop augmentation and Sinkhorn-Knopp, DINO only achieves 72.5%; see [10] for detail).

Table 2. Top-1 accuracy of state-of-the-art methods on ImageNet using pretraining and fine-tuning evaluation.

Rand	MoCo v3 [16]	DINO [10]	BEiT [3]	MAE [26]	Ours
81.8	83.2	82.8	83.2	**83.6**	83.4

4.3. Pretraining and fine-tuning on ImageNet

As we have mentioned above, pretraining-finetuning evaluation has several shortcomings. Nevertheless, following [3] and [26], we also report the accuracies under this evaluation metric in Table 2. We have two observations. First, the fine-tuned result of our method is better than $\mathcal{D}$ models. Second, compared to $\mathcal{G}$ models, our result is slightly better than BEiT [3] and comparable to MAE [26].

4.4. Transfer to downstream tasks on COCO 2017

A goal of SSL is to learn general features. Thereby, we need to test the generalization ability of the learned features of our method by transferring them to downstream tasks. The downstream benchmark we adopt is COCO 2017 [40], which is currently one of the largest benchmarks for general object detection and segmentation, containing a total of 119k training images. Specifically, our ViT backbone is first trained using the above-mentioned unsupervised learning. Then these pretraining parameters are used as the initialization parameters of a Cascade Mask-RCNN [29], the widely-adopted framework for ViTs in object detection [65]. Next, we use the COCO 2017 training set to fine-tune all ViT layers. As suggested by [27,61], the distribution of features obtained by unsupervised pretraining is different from that of supervised pretraining. Therefore, in the fine-tuning stage, we use a larger learning rate than supervised pretraining counterparts. We report the accuracy on the COCO 2017 validation set. For the object detection task, we report the standard AP^{box} metric; for the instance segmentation task, we report the standard AP^{mask} metric.

Comparison with $\mathcal{G}$ models. Table 3 shows that, in the COCO 2017 object detection and instance segmentation task, our SSL pretraining approach achieves state-of-the-art performance. Our approach is better than BEiT (*e.g.*, 51.0% vs. 50.1% for the AP^{Box} metric, and 44.1% vs. 43.5% for the AP^{Mask} metric). It is worthy to note that all the $\mathcal{G}$ models outperform both $\mathcal{D}$ models and the supervised counterparts. These comparisons confirm the effectiveness of our method. Besides, we are unable to compare our method with MAE because their training protocol is missing. Note that MAE loads pretrained weights into a *new windowed ViT* [38] for fine-tuning, claiming MAE is superior to existing methods. We are really worried about this claim because loading weights into a new architecture puts existing methods at a disadvantage, as it introduces architecture

Table 3. Object detection and instance segmentation on COCO 2017 for Mask-RCNN.

Method	AP^{Box} (%)	AP^{Mask} (%)
$\mathcal{D}$ models		
DINO [10]	50.1	43.4
$\mathcal{G}$ models		
BEiT [3]	50.1	43.5
Ours	**51.0**	**44.1**
Supervised		
supervised	49.8	43.2

gaps. Moreover, the code of [38] is unavailable.

Comparison with $\mathcal{D}$ models. Different from the linear evaluation in ImageNet, in downstream segmentation and detection tasks, the current best-performing methods are $\mathcal{G}$ models rather than $\mathcal{D}$ models. For example, as Table 3 shows, our method is 0.9% higher in AP^{Box} accuracy than the best $\mathcal{D}$ model DINO. Two reasons might account for this. First, our method learns more general features that are transferable to downstream tasks. Second, as a $\mathcal{G}$ model, our approach pays more attention to local dependencies that are beneficial for dense prediction tasks.

Comparison with supervised learning. As shown in Table 3, our AE method fully surpasses the supervised pre-training method. Our AP^{Box} is 51.0%, and the supervised counterpart is 49.8%. These comparisons confirm the effectiveness of our method. Note that, in summary, $\mathcal{G}$ models are better than $\mathcal{D}$ models, and $\mathcal{D}$ models are close to the performance of the supervised model.

4.5. Person re-identification on SYSU-30k

Since ViTs are data-hungry models, we next verify the effectiveness of our method on a more extensive data set, SYSU-30k [59], that is 30 times larger than ImageNet both in terms of category number and image number. From a more general perspective, the above tasks (image classification, detection, and segmentation) are all visual classification tasks[7]. In view of this, the effectiveness of our method needs to be verified on more types of tasks. As mentioned in Section 4.1, if we can find a task to directly evaluate the features learned by SSL without fine-tuning network parameters, we will have a more transparent understanding of the effectiveness of SSL. Fortunately, person re-ID [22] is such a satisfying visual matching task[8]. Hence, we adopt the re-ID task for examination. The benchmark SYSU-30k is employed for three reasons. First, SYSU-30k is not only the largest re-ID dataset, but also one of the largest datasets in computer vision, containing 29,606,918 images of 30,508 pedestrians. Second, this dataset does not hold an exact label for each image. Evaluation on SYSU-30k means that we use its training set to perform SSL, and then the learned model is directly used to extract features for matching without any fine-tuning. This linear probing is more challenging than linear evaluation on ImageNet because linear evaluation on ImageNet can learn an extra classifier for recognition, but no extra classifier is allowed here. Third, another challenge in the linear probing is that there are 478,730 mismatching images as the wrong answer in the gallery. Evaluation using the SYSU-30k test set is like searching for a needle in a haystack. Unless being an extraordinary SSL feature learner, it is challenging to excel in this task.

Table 4. Comparison in re-ID tasks on SYSU-30k.

method	rank-1 (%)	Backbone
$\mathcal{D}$ models		
SimCLR [12]	10.9	R50
MoCo v2 [13]	11.6	R50
BYOL [25]	12.7	R50
Triplet [61]	14.8	R50
MoCo v3 [16]	14.96	ViT-base
$\mathcal{G}$ models		
BEiT [3]	8.3	ViT-base
Ours	**11.8**	ViT-base

Comparision with SSL methods. We compare our method with existing SSL methods, including SimCLR [12], MoCo v2 [13], BYOL [25], MoCo v3 [16], and BEiT [3]. Among them, MoCo v3, BEiT, and our method use ViT-base [52] as the backbone, while others use ResNet-50 [30] as the backbone. Both BEiT and our method are $\mathcal{G}$ models; other methods are $\mathcal{D}$ models. The experimental results are shown in Table 4. We can see that our method has achieved a good performance (11.8%), which is comparable to $\mathcal{D}$ models. Overall, the performance of the $\mathcal{D}$ models is better than $\mathcal{G}$ models. Even so, our method is satisfactory. These comparisons prove that our method is an effective SSL visual feature learner.

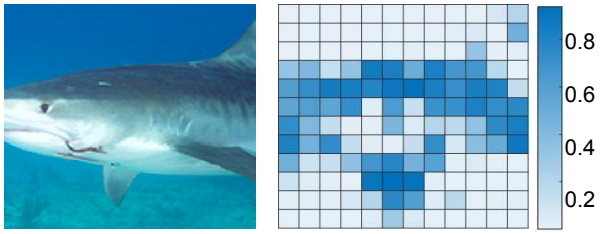
One thing that could not be ignored is that all SSL models perform unsatisfactorily on the challenging benchmark of SYSU-30k, *i.e.*, the rank-1 values are very low. This is attributed to the challenge of the dataset, whereas warning us that SSL still has a long way to go.

5. Visualization

We show our image generation results in Figure 2 (Page 2) in the form of triplets. In each triplet, the left is the input image, the middle is our generated image, and the right is the ground truth. We can see that our generated images are very close to the ground truths. Even if some generated im-

[7]This is because object detection and instance segmentation can be seen as classifying regions and pixels.

[8]Specifically, re-ID is a visual matching problem of recognizing pedestrians across cameras [39, 54–58, 60, 67]. But in recent years, there are some concern about privacy issues of face recognition and re-ID technology, which is beyond the scope of the scientific community. We evaluate the SYSU-30k in this paper for research purposes only. Our source code and model are not allowed to use for any applications like surveillance that might raise ethical concerns.



(a) (b)

Figure 5. Achieving interpretability via counterfactual intervention with the help of a $\mathcal{G}$ model.

age is slightly different from the ground truth, the generated image is semantically plausible, indicating that our model has learned reasonably semantic representations. Besides, there is some grid effect in the visualization, similar to that in MAE [26]. We conjecture that this is attributed to two reasons. First, the loss is applied to the masked region. Second, it might be due to feature distortion.

We have proven the generalizability of our method, e.g., in outperforming $\mathcal{D}$ models in pretraining tasks and dense prediction tasks and towards unifying pretraining methodology between CV and NLP. Next, we show the interpretability of our method. Our method can elegantly perform causal inference that is beyond the ability of $\mathcal{D}$ models. We conduct counterfactual interventions by masking an image to leave only one window at a time to exclude confounding factors for reconstruction. The reconstruction scores are visualized in Figure 5. As shown, the shark is successfully found via causal inference.

6. Necessities, exclusions, and options

The rest of this paper aims not to pursue state-of-the-art results but to gain insight into the role of different components of our method. Thus, we reduce the training epochs to 100 to fast access results. Other rraining protocol in this section is the same as in Section 4. This section only reports the linear evaluation results on ImageNet, because, as we said before, it is a general and reliable evaluation method.

Necessity of semantic-aware generation.As described, our method inputs one view of the image and generates another view of it to perceive semantics. Now we removed this cross-view generation design and let the model perform self-regression, degenerating it back to a vanilla AE. Table 5 shows the result. It can be seen that after removing the semantic-aware generation, the performance of the visual feature learner plummeted from 46.4 to 6.1. This comparison confirms the necessity of semantic-aware generation.

Necessity of semantic alignment. As mentioned, to train our semantic-aware AE, we need to align semantics. Specifically, we need to know the geometric transformation of the input and target and align them. Now we remove the aligner and directly regress the input into the target. This learning goal is different from all previous AEs, causing huge learning difficulties. We observed that the training loss dropped quickly at the start, then it almost no longer declined (see Table 5). This comparison confirms that semantic alignment is necessary in a good feature learner.

Table 5. Effectiveness analysis and insight of our method.

Properties	Top-1 (%)
Semantic-aware AE (*full*)	46.4
- Semantic-aware generation	6.1 (-40.3)
Semantic-aware AE (*full*)	46.4
- Semantic alignment	not converge
Semantic-aware AE (*full*)	46.4
+ Global feature	not converge

Exclusion of global feature. Unlike $\mathcal{D}$ models, existing state-of-the-art $\mathcal{G}$ models [3, 11, 26] do not have a global feature vector. As pointed out by [3, 11], lacking globality creates a gap between pretraining and linear evaluation. Here we insert a global pooling module at the end of the encoder and then increase the spatial size of the feature map using transformer decoder [45] and deconvolution [49] respectively to complete image reconstruction. As a result, the training loss cannot converge regardless of whether using a transformer decoder or deconvolution (see Table 5).

7. Conclusion

In NLP/NLU, $\mathcal{G}$ models play a vital role in SSL pretraining. But this role was absent in computer vision before, until [3, 26] came. This paper aims to bridge this gap. To tackle the problem that $\mathcal{G}$ models waste capacity in learning semantic-agnostic local signals, we propose a novel semantic-aware AE. Our AE uses one view of the image as input, but reconstructs another view of the image. In this way, our AE learns semantic representations, achieving good performance in many tasks. We hope that our approach will be inspiring for rethinking the new position of $\mathcal{G}$ models as a feature learner in computer vision, especially for closing the gap between NLP and computer vision.

Broader impacts. This paper uses existing datasets, so the potential negative impact of the existing datasets will also be inherited. For example, the datasets ImageNet and SYSU-30k inevitably contain human photos. One limitation of this paper is that the proposed $\mathcal{G}$ still has a slight gap to the $\mathcal{D}$ models so far. The proposed method can generate non-existent images, which might be uncontrollable.

Acknowledgement

This work is supported by the EPSRC/MURI grant EP/N019474/1, China National Key R&D Program grant 2021ZD0111600, China National Natural Science Foundation grant 61836012, and Guangdong Natural Science Foundation grant 2020A1515010423. We would also like to thank the Royal Academy of Engineering of the UK.

References

[1] Arjun R Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience*, 25(1):103581, 2022. 1, 3

[2] Dana H. Ballard. Modular learning in neural networks. In Kenneth D. Forbus and Howard E. Shrobe, editors, *Proceedings of the 6th National Conference on Artificial Intelligence. Seattle, WA, USA, July 1987*, pages 279–284. Morgan Kaufmann, 1987. 3

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. 1, 3

[5] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992. 1, 2, 3, 4

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 3

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 139–156. Springer, 2018. 6

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020. 6

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. 1, 3, 4, 5, 6, 7

[11] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 2020. 2, 3, 6, 8

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 1, 3, 6, 7

[13] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 1, 3, 6, 7

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020, virtual conference online, June 19-25, 2021*, page , 2021. 1, 3, 6

[15] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 5

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021. 5, 6, 7

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1, 3

[18] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1422–1430. IEEE Computer Society, 2015. 1, 3, 6

[19] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10541–10551, 2019. 2, 3, 6

[20] Chenhe Dong, Guangrun Wang, Hang Xu, Jiefeng Peng, Xiaozhe Ren, and Xiaodan Liang. Efficientbert: Progressively searching multilayer perceptron via warm-up knowledge distillation. *CoRR*, abs/2109.07222, 2021. 3

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3, 4, 5

[22] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2360–2367. IEEE Computer Society, 2010. 7

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 1, 3, 6

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 2

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. 1, 3, 6, 7

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1, 2, 3, 5, 6, 8

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 2020. 1, 3, 6

[28] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4917–4926. IEEE, 2019. 5

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 6

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 7

[31] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 646–661. Springer, 2016. 5

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. 3

[33] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017. 2

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 2

[35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2, 3

[36] Yann Le Cun and Françoise Fogelman-Soulié. Modèles connexionnistes de l'apprentissage. *Intellectica*, 2(1):114–143, 1987. 3

[37] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[38] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 6, 7

[39] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Junyong Zhu. M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. *arXiv preprint arXiv:1811.03768*, 2018. 7

[40] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, 2014. 2, 5, 6

[41] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[42] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982. 3

[43] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016. 1, 3

[44] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544. IEEE Computer Society, 2016. 3

[45] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, October 2021. 8

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. 3

[47] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 3

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 2, 5

[49] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3483–3491, 2015. 8

[50] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. 5

[51] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 1, 3

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021. 2, 4, 5, 7

[53] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM, 2008. 2, 3

[54] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8933–8940. AAAI Press, 2019. 7

[55] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10565–10574. Computer Vision Foundation / IEEE, 2020. 7

[56] Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2777–2787, 2017. 7

[57] Guangrun Wang, Liang Lin, Shengyong Ding, Ya Li, and Qing Wang. DARI: distance metric and representation integration for person verification. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3611–3617. AAAI Press, 2016. 7

[58] Guangrun Wang, Guangcong Wang, Keze Wang, Xiaodan Liang, and Liang Lin. Grammatically recognizing images with tree convolution. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–912, 2020. 7

[59] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. In *IEEE Transactions on Neural Networks and Learning Systems (T-NNLS)*, 2020. 2, 5, 7

[60] Guangrun Wang, Keze Wang, and Liang Lin. Adaptively connected neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1781–1790, 2019. 7

[61] Guangrun Wang, Keze Wang, Guangcong Wang, Philip H.S. Torr, and Liang Lin. Solving inefficiency of self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9505–9515, October 2021. 1, 2, 3, 5, 6, 7

[62] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance

discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. 6

[63] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6687–6696. Computer Vision Foundation / IEEE, 2020. 6

[64] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer, 2016. 3

[65] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 5, 6

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 2

[67] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 7