Mixed Autoencoder for Self-supervised Visual Representation Learning

Kai Chen[1]   Zhili Liu[1,2]   Lanqing Hong[2]   Hang Xu[2]   Zhenguo Li[2]   Dit-Yan Yeung[1]

[1]Hong Kong University of Science and Technology   [2]Huawei Noah's Ark Lab

{kai.chen,zhili.liu}@connect.ust.hk   {honglanqing,xu.hang,li.zhenguo}@huawei.com

dyyeung@cse.ust.hk

Abstract

Masked Autoencoder (MAE) has demonstrated superior performance on various vision tasks via randomly masking image patches and reconstruction. However, effective data augmentation strategies for MAE still remain open questions, different from those in contrastive learning that serve as the most important part. This paper studies the prevailing mixing augmentation for MAE. We first demonstrate that naïve mixing will in contrast degenerate model performance due to the increase of mutual information (MI). To address, we propose homologous recognition, an auxiliary pretext task, not only to alleviate the MI increasement by explicitly requiring each patch to recognize homologous patches, but also to perform object-aware self-supervised pre-training for better downstream dense perception performance. With extensive experiments, we demonstrate that our proposed Mixed Autoencoder (MixedAE) achieves the state-of-the-art transfer results among masked image modeling (MIM) augmentations on different downstream tasks with significant efficiency. Specifically, our MixedAE outperforms MAE by +0.3% accuracy, +1.7 mIoU and +0.9 AP on ImageNet-1K, ADE20K and COCO respectively with a standard ViT-Base. Moreover, MixedAE surpasses iBOT, a strong MIM method combined with instance discrimination, while accelerating training by 2×. To our best knowledge, this is the very first work to consider mixing for MIM from the perspective of pretext task design. Code will be made available.

1. Introduction

Self-supervised learning (SSL) has become one of the most popular pre-training paradigm due to its independence of human annotation. Previous literature mainly focuses on the handcrafted pretext task design [13, 19, 36] and instance discrimination [6, 10], while with the development of Vision Transformer [15], masked image modeling (MIM), deeply motivated by masked language modeling [12], has started to demonstrate more superior effectiveness by firstly **masking** some patches of the input images and then **reconstructing** the masked patches from visible ones by predicting certain
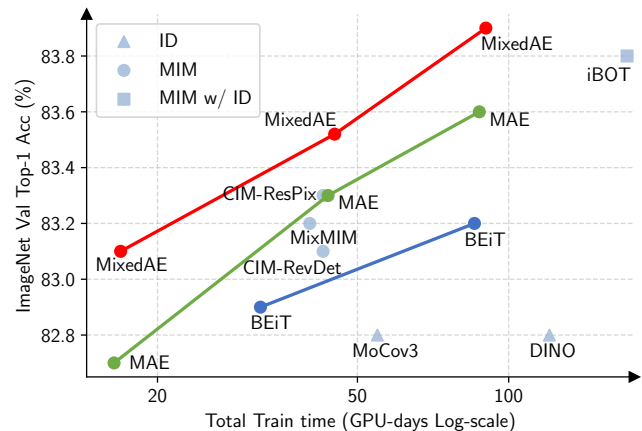

Figure 1. **Fine-tuning accuracy on ImageNet-1K.** Our *MixedAE* achieves the best trade-off between pre-training overhead and transfer performance. Specifically, *MixedAE* surpasses MAE [22] consistently with only 3% extra overhead, while outperforms the strong iBOT [49] with only 53.4% of its computation overhead. See more detailed comparisons in Tab. 1. ID stands for instance discrimination, while MIM represents masked image modeling.

targets generated by masked patches. In order to complete reconstruction, the encoder is expected to generate highly semantic representation which can be better transferred to downstream tasks [21, 29, 30, 48] for superior performance.

Existing MIM works mainly concentrate on the design of the reconstruction targets (*e.g.*, visual tokenizers [3, 14], pixels [22, 44], graphical features [41] and instance discrimination [2, 16, 49]) and masking strategies (*e.g.*, random [3, 22], attention-guide [25] and sample-dependent [39]). See more detailed discussions in Sec. 2. Despite the superior performance, we observe that the input augmentations for MIM have been seldom explored. Specifically, adding *color jittering*, an essential augmentation technique of contrastive learning [8], with MAE [22] even *degrades* the transfer results, suggesting MIM might have a different preference for data augmentations, and the effective data augmentation strategies for MIM are still an open question.

In this paper, we explore the usage of image mixing, a commonly used technique in both supervised [46, 47] and contrastive learning [38, 45], with MAE [22]. We start by

constructing a simple baseline to adopt mixing with MAE directly, which, different from in supervised and contrastive learning, would instead ease the reconstruction pretext by increasing the *mutual information* between the model input and reconstruction target due to the usage of image mixing with global self-attention as proved in Sec. 3.1. To address this issue, we propose *homologous recognition*, an auxiliary pretext task to enforce each patch to recognize homologous patches explicitly according to attention distributions before reconstruction, and build our Mixed Autoencoder network (*MixedAE*) in Sec. 3.3. Moreover, we demonstrate that our simple yet effective method can not only achieve significant performance improvement, but also conduct object-aware SSL pre-training without any specifically designed modules for better downstream dense perception results in Sec. 3.4.

Concurrently, MixMIM [31] also considers mixing with MAE, but different from ours, 1) **Purpose**: MixMIM uses mixing to recover the 2D structure after random masking for an efficient implementation to conduct MAE-style pre-training on hierarchy Vision Transformers [34], while ours utilizes mixing to conduct object-aware SSL pre-training for better representation learning. 2) **Method**: MixMIM uses masked self-attention to only perform attention within patches from the same images given the mixing masks as **input**, sharing the exactly same pretext task with MAE, while ours requires explicit homologous recognition given mixing masks as **target**, actively emerging mixing into the pretext design. 3) **Formulation**: The mixing ratio r is limited to 0.5 in MixMIM, which instead can be flexibly selected from $(0, 0.5]$ in our formulation. See more details in Sec. 3.

The main contributions of this work contain three parts:

1. We propose the Mixed Autoencoder (*MixedAE*), a simple yet effective approach to conduct object-aware pre-training without introducing any specifically designed modules. With extensive experiments, we demonstrate that *MixedAE* can achieve the state-of-the-art transfer performance on various downstream tasks including image classification, semantic segmentation and object detection, while maintaining significant efficiency.

2. We theoretically demonstrate the underlying design differences between MIM and previous supervision with mixing (*e.g.*, supervised and contrastive learning).

3. To our best knowledge, this is the first work to consider mixing as an effective data augmentation strategy for MIM from the perspective of pretext design with a pure autoencoder-based architecture.

2. Related Work

Self-supervised learning aims at learning a transferable representation without human annotation. Previous works mainly focus on handcrafted pretext design [13, 19, 36] and instance discrimination [8, 9, 20]. Mask image modeling (MIM), inspired by the mask language modeling [12], has achieved significant performance with superior pre-training efficiency by firstly masking portion of an image, and then reconstructing the masked part based on the visible one.

Reconstruction target. BEiT [3] pioneeringly proposes to predict visual tokens generated by a pre-trained visual tokenizer [37], which is simplified by SimMIM [44] to use pixel values as the reconstruction target directly. MAE [22] proposes an asymmetric encoder-decoder architecture for better efficiency. Besides pixels as the target, MaskFeat [41] utilizes HOG features, while PeCo [14] enhances the BEiT tokenizer with an additional perceptual loss. Recent works combine the idea of instance discrimination [8] with MIM. iBOT [49] considers MIM as a self-distillation process with a Siamese architecture, and data2vec [2] proposes a unified framework to conduct masked reconstruction pre-training for speech, images and languages. SplitMask [16] divides an image into two equal partitions and performs contrastive learning and MIM in the multi-task manner. In this paper, we build *MixedAE* based on MAE due to its efficiency and effectiveness, while the improvement brought by *MixedAE* is complementary to more advanced reconstruction targets.

Masking strategy. Instead of random masking [3, 22], AttMask [25] proposes a novel attention-guided masking strategy by masking according to the attention map of the final Transformer layer, while ADIOS [39] introduces an adversarial objective between masking and reconstruction to generate learnable masks for MIM pre-training. In this paper, we utilize random masking for *MixedAE* following MAE due to its simplicity and effectiveness.

Input augmentation instead has been less explored for MIM. Instead of masking, CIM [18] adopts a small BEiT as the generator to corrupt an image, which is further taken as input to an enhancer to reconstruct the corrupted patches or distinguish the corrupted patches from the uncorrupted ones. Concurrently, MixMIM [31] considers mixing with MAE also, but different from ours, MixMIM uses *masked self-attention* to only perform attention within patches from the same images given the mixing masks as *input*, which is exactly the same with MAE from the perspective of pretext design, while ours utilizes mixing as part of the pretext task actively to conduct object-aware SSL pre-training.

3. Method

In this section, we start by adopting mixing in MAE [22] with a simple baseline in Sec. 3.1, which, as we can prove, will instead ease the reconstruction pretext task. Then, we propose a novel auxiliary pretext task to formulate our final *MixedAE*, which can not only alleviate the ease of reconstruction, but also achieve object-aware SSL pre-training without specifically designed modules in Secs. 3.2 to 3.4.
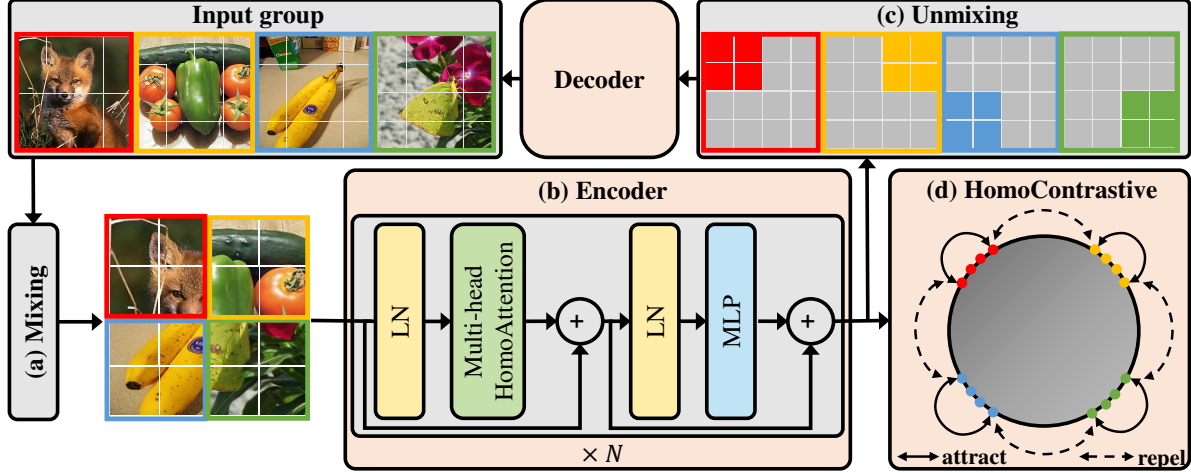
Figure 2. **Model architecture of Mixed Autoencoder (*MixedAE*).** (a) The input images are first separated into groups to generate mixed samples independently, which are further taken as input to the encoder for feature extraction. (b) The self-attention operations are replaced with our homologous attention, enforcing each patch to only attend to patches with the highest attention mass. (c) The encoder features will be "unmixed" and fed into the decoder for pixel reconstruction. (d) Meanwhile, the homologous contrastive loss is adopted to verify the sampling accuracy by encouraging features of homologous patches to be similar, while heterologous ones to be dissimilar.

3.1. Mixing: A Simple Baseline

Given a unlabeled dataset, we randomly sample a *clean* data batch with size B, which are later divided into non-overlapping patch sequences $\{x_i\}_{i=1}^{B}$ ($x_i \in \mathcal{R}^{L \times (P^2 \cdot C)}$) following ViT [15], where L is the sequence length, P is the patch size, and C is the image channel dimension.

Mixing. The data batch is further separated into *groups* $\{\{x_i^j\}_{i=1}^{1/r}\}_{j=1}^{Br}$, and each group will generate a single *mixed* image, where $r \in (0, 0.5]$ is the *mixing ratio* representing the ratio of patches each clean image contributes to within a single mixed sample. Different from MixMIM [31], r is not restricted to 0.5 in our formulation. Therefore, the mixing process for the j-th group can be represented as,

$$\hat{x}^j = \sigma_{mix}(\{x_i^j\}, M^j) = \sum_{i=1}^{1/r} \mathbb{1}(M^j = i)x_i^j, \quad (1)$$

where $\mathbb{1}(\cdot)$ is indicator function and $M^j \in \{1, 2, ..., 1/r\}^L$ represents a random mixing mask independently generated for the j-th group, which satisfies,

$$\sum_{l=1}^{L} \mathbb{1}(M_l^j = i) = rL, \ \forall i \in \{1, 2, ..., 1/r\}. \quad (2)$$

So, M^j determines the source patch in each position of $\hat{x}^j$, while maintaining the mixing ratio for each clean image equal with r (*i.e.* symmetric mixing). The mixed images $\hat{x}^j$ are further fed into the encoder for feature extraction, which can be represented as $\hat{z}^j = \text{ENC}(\hat{x}^j)$.

Unmixing. Following MAE [22], $\hat{z}^j$ is then "unmixed" to recover the input batch before mixing by inserting a special [MASK] token with M^j. For $\forall i \in \{1, 2..., 1/r\}$, we have,

$$z_i^j = \mathbb{1}(M^j = i)\hat{z}^j + [1 - \mathbb{1}(M^j = i)][\text{MASK}]. \quad (3)$$

The "unmixed" group $\{z_i^j\}_{i=1}^{1/r}$ is then taken as input to the decoder for pixel reconstruction, as $y_i^j = \text{DEC}(z_i^j)$. Finally, the reconstruction loss can be formulated as,

$$\mathcal{L}_{recon} = \sum_{i=1}^{1/r} \sum_{l=1}^{L} [1 - \mathbb{1}(M_l^j = i)](y_{i,l}^j - x_{i,l}^j)^2. \quad (4)$$

So far, we have built a simple baseline to adopt mixing for MAE, which, however, performs even worse than MAE, as demonstrated in Tab. 3f. In the following, we provide a theoretical explanation to prove that this naïve incorporation will actually **ease** the reconstruction pretext task.

Mutual information analysis. Without loss of generality, we take $r = 0.5$ as an example. Denote X_1, X_2 as two random variables representing two input images, and X_1 is further considered as the reconstruction target (symmetric for X_2). Then, we can prove that the mutual information (MI) between the mixed input $\sigma_{mix}(\{X_1, X_2\}, M)$ and the target X_1 is no smaller than that between the MAE input $\sigma_{MAE}(X_1, M)$ and X_1 as (see proofs in Appendix A),

$$I(\sigma_{mix}(\{X_1, X_2\}, M); X_1) \geq I(\sigma_{MAE}(X_1, M); X_1). \quad (5)$$

Therefore, different from masking, which is introduced to **decrease** the mutual information between the model input

(a) Uncertainty of generative modeling.

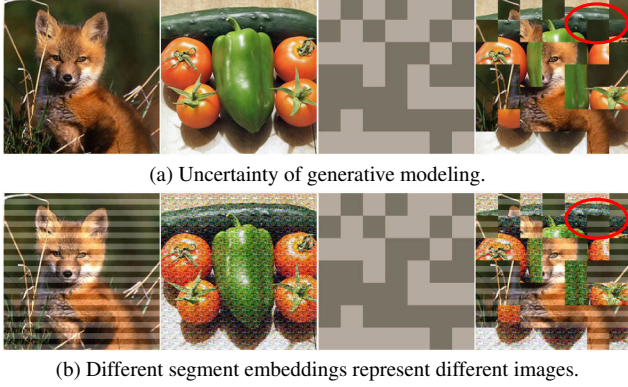(b) Different segment embeddings represent different images.

Figure 3. **Visualization of segment embeddings**. (a) Due to the uncertainty of generative modeling, green colors of the cucumber and the forest are both reasonable for patches in the red ellipse. (b) We adopt different segment embeddings for different images to provide necessary information for homologous recognition.

and the reconstruction target due to the redundancy of image signals [22], naïve mixing will instead **increase** the MI, and thus, ease the reconstruction pretext task. Verification experiment is conducted in Appendix C.

Also note that the MI increasement brought by mixing is target-invariant, suggesting that Eq. (5) also holds when the target is *semantic labels* for supervised learning or *positive samples* for contrastive learning, for which MI increasement is appealing. This might explain why naïve mixing without auxiliary supervision is beneficial for supervised [46, 47] and contrastive learning [38, 45], but not for MAE.

3.2. Recognition: Homologous Recognition

Another indispensable factor to achieve MI increasement is the usage of *global self-attention* in ViT, with which each query patch will inevitably attend to heterologous patches from other images. Due to the uncertainty of generative modeling, heterologous patches might provide a shortcut to complete reconstruction (*e.g.*, the green color of cucumbers is a "valuable" cue to reconstruct the forest behind the fox in Fig. 3a). To address, we propose a novel auxiliary pretext task called *homologous recognition* to enforce each query to explicitly recognize and only attend to homologous patches.

Homologous attention recognizes homologous patches on-the-fly by enforcing each query patch to only attend to key patches with the highest attention mass using a $\mathrm{TopK}(\cdot)$ sampling operation. Specifically, the homologous attention can be formulated as,

$$A_{HomoAtt} = \mathrm{softmax}(\mathrm{TopK}(\boldsymbol{q}\boldsymbol{k}^T/\sqrt{D_h})), \quad (6)$$

where $\boldsymbol{q}$ is the query patch, $\boldsymbol{k}$ are the key patches and D_h is the feature dimension. By default, all the self-attention operations in ViT are replaced with homologous attention except the very first layer. See comparisons in Tab. 3e.
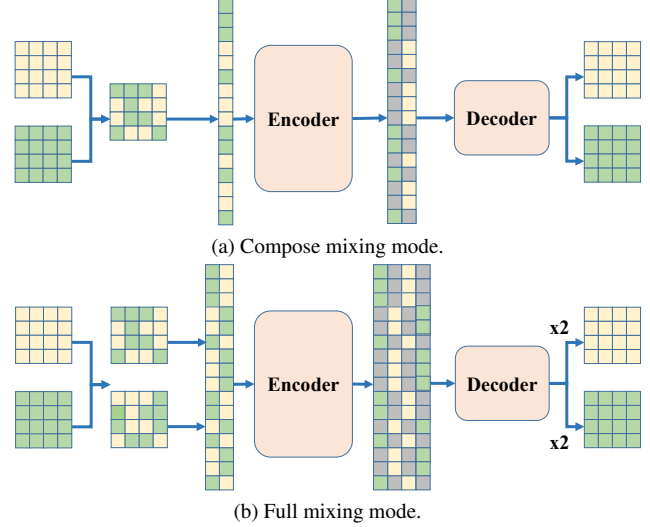

(a) Compose mixing mode.

(b) Full mixing mode.

Figure 4. **Visualization of two mixing modes when** $r = 0.5$. (a) Each group generates a single mixed sample for the compose mixing mode, (b) while $1/r$ mixed samples are generated for the full mixing mode to maintain the effective batch size unchanged.

Homologous contrastive aims at verifying the $\mathrm{TopK}(\cdot)$ sampling accuracy by encouraging the encoder features of homologous patches to be similar, while heterologous ones to be dissimilar in the supervised contrastive manner [26]. The homologous contrastive loss can be formulated as,

$$\mathcal{L}_{HomoCon} = -\sum_{l=1}^{L}\sum_{l^+} \log \frac{exp(cos(\hat{\boldsymbol{z}}_l^j, \hat{\boldsymbol{z}}_{l+}^j)/\tau)}{\sum_{l'\neq l}^{L} exp(cos(\hat{\boldsymbol{z}}_l^j, \hat{\boldsymbol{z}}_{l'}^j)/\tau)}, \quad (7)$$

where τ is the temperature and $cos(\cdot, \cdot)$ is the cosine similarity. As demonstrated in Fig. 6, the $\mathrm{TopK}(\cdot)$ sampling accuracy is significantly improved and stabilized with the usage of the homologous contrastive loss $\mathcal{L}_{HomoCon}$.

Segment embedding. Beside the positional embeddings, we add another segment embedding to the mixed sequence $\hat{\boldsymbol{x}}^j$ following BERT [12] to provide necessary information to complete homologous recognition, due to the uncertainty of generative modeling. The segment embedding is shared for patches from the same image, while different for patches from different images, as demonstrated in Fig. 3b.

Mixing mode. For a fair comparison under different training overheads, two mixing modes are adopted for *MixedAE*, 1) **Compose**: each group generates a single mixed sample following Eq. (1), and the effective encoder batch size is Br; 2) **Full**: each group generates $1/r$ mixed samples by sampling $\boldsymbol{M}^j$ for $1/r$ times independently, and the effective encoder batch size is B. An example is provided in Fig. 4 when $r = 0.5$. As shown in Tab. 1, *MixedAE* achieves the SoTA performance under different training overheads. If no otherwise specified, compose mixing is adopted by default.
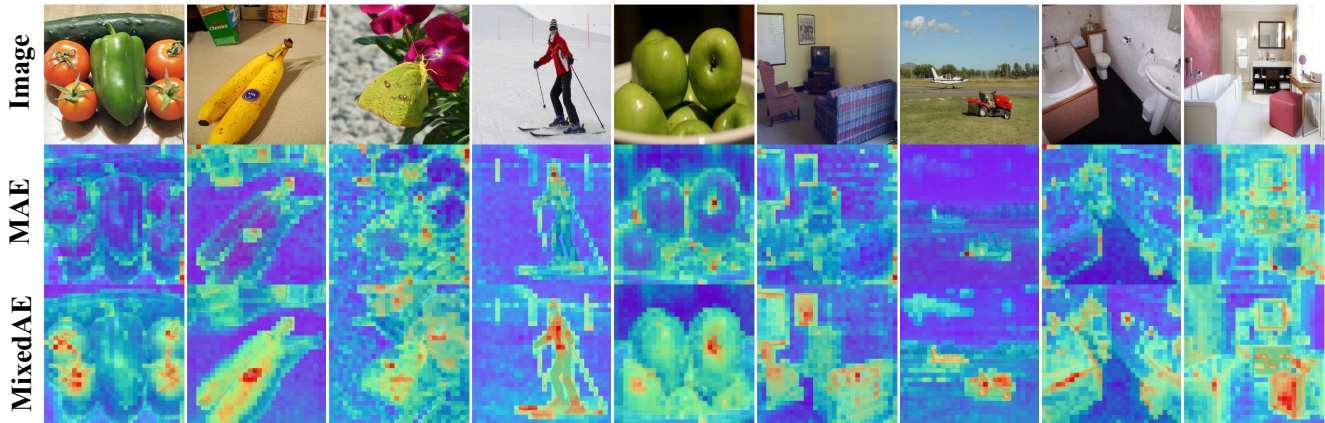
Figure 5. **Visualizations of attention maps** on images from ImageNet-1K [11] (1st-3rd columns), Microsoft COCO [30] (4th-6th columns) and ADE20K [48] datasets (7th-9th columns). Both MAE and *MixedAE* are pre-trained on ImageNet-1K for 300 epochs. Compared with MAE which mainly focuses on the most discriminative patches, (*e.g.*, boundaries (1st, 2nd & 5th) and edges (6th & 8th)), *MixedAE* discovers foreground object patches more precisely (3rd & 9th) and completely (4th & 7th). See more attention maps in Appendix E.

3.3. Reconstruction: Loss Function

Loss function. We formulate *MixedAE* in the multi-task learning manner and the final loss function is a weighted sum of the reconstruction loss $\mathcal{L}_{recon}$ and the homologous contrastive loss $\mathcal{L}_{HomoCon}$ as,

$$\mathcal{L}_{MixedAE} = \mathcal{L}_{recon} + \lambda\mathcal{L}_{HomoCon}, \quad (8)$$

where the balanced weight λ is set to be 0.1 by default.

3.4. Discussion: Object-aware Pre-training

With the usage of mixing, we observe that *MixedAE* can achieve object-aware self-supervised pre-training without any specifically designed components, such as K-means [7], selective search [42] and object discovery network [24], because homologous recognition requires each query patch to recognize all homologous patches within a mixed image. Due to the *single-centric-object* guarantee [7] of ImageNet, that most images are pre-processed to guarantee only one object in the center part of them, the mixed image can be considered as a "pseudo" multi-instance image, and given a query patch, the process of recognizing all patches from the same image within a mixed sample is exactly recognizing all patches from the same object within a given "pseudo" multi-instance image. Therefore, the awareness of object existence and completeness is enhanced in the leant representation of our *MixedAE*.

In Fig. 5, we visualize the attention maps of MAE and *MixedAE* by averaging all attention heads of the last layer, taken the [CLS] token as query and patch tokens as keys. Compared with MAE which mainly focuses on the most discriminative patches (*e.g.*, boundaries and corners), our *MixedAE* can successfully discover the foreground object patches more precisely and completely, which might also

explain why *MixedAE* improves more significantly when transferred to dense perception tasks, such as semantic segmentation [48] and object detection [30], as shown in Tab. 1.

4. Experiments

4.1. Implementation Details

Architecture. We mainly use the standard ViT-Base [15] as the backbone architecture, and further provide ViT-Large experiments in Appendix C. The input images are resized to 224×224, resulting in a total sequence length $L = 196$ with the patch size being 16×16. Following MAE [22], the decoder consists of 8 Transformer layers with the hidden dimension as 512 by default. For a fair comparison with BEiT [3], we additionally build a *MixedAE* in full mixing mode with a lightweight decoder made up of 2 Transformer layers and the hidden dimension as 256, as shown in Tab. 1.

The mixing ratio r is set to be 0.25 (*i.e.*, corresponding to the 0.75 masking ratio in MAE [22]) by default, and the threshold K in the TopK$(\cdot)$ operation is therefore set to be $0.25L$. Following common practices [6, 10], we adopt a linear projector with the output dimension as 128 right before the homologous contrastive loss, and the temperature coefficient τ is set to be 0.25.

Optimization. We pre-train *MixedAE* on the ImageNet-1K [11] training set with the AdamW [35] optimizer and a cosine learning rate schedule with a linear warm-up of 40 epochs. The batch size is set to be 4096 for the compose mixing, while 1024 for the full mode. The base learning rate is set to be $7.5e^{-5}$, which will scale linearly with the batch size ($lr = lr_{base} \times bs/256$). Only standard random cropping and flipping are utilized for data augmentation. The remaining hyperparameters are all maintained the same with MAE for a fair comparison (see Appendix B for more details).

Method	Pre-train Epochs	Pre-train† GPU-days	ImageNet Top-1 Acc.	ADE20K mIoU	COCO					
					AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
DeiT [40]	300	19.6	81.8	46.9	48.8	68.7	52.7	42.5	65.9	45.5
MoCov3 [10]	$600^{\dagger\dagger}$	54.8	82.8	46.8	47.2	66.9	50.8	41.1	63.6	44.1
DINO [6]	$1600^{\dagger\dagger}$	120.5	82.8	46.9	49.5	69.1	53.6	42.9	66.0	46.3
BEiT [3]	300	32.1	82.9	44.7	39.3	57.7	42.4	34.8	55.2	36.8
MAE [22]	300	16.4	82.7	46.1	47.2	65.8	51.3	41.1	62.9	44.4
MixMIM [31]	300	40.2	83.2	-	-	-	-	-	-	-
CIM-RevDet [18]	300	42.7	83.1	-	-	-	-	-	-	-
CIM-ResPix [18]	300	42.7	83.3	-	-	-	-	-	-	-
MixedAE	300	16.9	$83.1^{+0.4}$	$47.0^{+0.9}$	$47.8^{+0.6}$	$66.6^{+0.8}$	$52.0^{+0.7}$	$41.6^{+0.5}$	$63.6^{+0.7}$	$45.0^{+0.6}$
MixedAE-Full*	300	30.8	$83.7^{+1.0}$	$47.4^{+1.3}$	$48.9^{+1.7}$	$67.6^{+1.8}$	$53.3^{+2.0}$	$42.5^{+1.4}$	$64.8^{+1.9}$	$45.9^{+1.5}$
MixedAE-Full	300	62.3	$\mathbf{83.8}^{+1.1}$	$\mathbf{48.9}^{+2.8}$	$\mathbf{51.0}^{+3.8}$	$\mathbf{69.7}^{+3.9}$	$\mathbf{55.2}^{+3.9}$	$\mathbf{44.1}^{+3.0}$	$\mathbf{67.0}^{+4.1}$	$\mathbf{47.9}^{+3.5}$
BEiT [3]	800	85.5	83.2	45.6	40.8	59.4	44.1	36.0	56.8	38.2
MAE [22]	800	43.7	83.3	47.2	49.4	68.1	53.9	42.9	65.5	46.6
MixedAE	800	45.0	$\mathbf{83.5}^{+0.2}$	$\mathbf{48.7}^{+1.5}$	$\mathbf{50.3}^{+0.9}$	$\mathbf{69.1}^{+1.0}$	$\mathbf{54.8}^{+0.9}$	$\mathbf{43.5}^{+0.6}$	$\mathbf{66.2}^{+0.7}$	$\mathbf{47.4}^{+0.8}$
MAE [22]	1600	87.4	83.6	48.1	50.6	69.4	55.0	43.8	66.6	47.5
iBOT [49]	$1600^{\dagger\dagger}$	172.1	83.8	49.6	51.2	70.1	55.2	44.3	67.4	48.0
MixedAE	1600	90.1	$\mathbf{83.9}^{+0.3}$	$\mathbf{49.8}^{+1.7}$	$\mathbf{51.5}^{+0.9}$	$\mathbf{70.2}^{+0.8}$	$\mathbf{55.9}^{+0.9}$	$\mathbf{44.5}^{+0.7}$	$\mathbf{67.5}^{+0.9}$	$\mathbf{48.2}^{+0.7}$

Table 1. **Transfer performance comparison between methods pre-trained on ImageNet-1K.** 1) Effectiveness: *MixedAE* achieves the state-of-the-art performance under different pre-training epochs and overheads. 2) Efficiency: *MixedAE* consistently surpasses the strong iBOT [49] baseline, while only requiring 53.4% of the pre-training overhead. 3) Object-aware pre-training: more significant improvements are achieved when transferred to downstream dense perception tasks (0.3 vs. 1.7 vs. 0.9). *: a lightweight decoder is deployed to maintain similar pre-training overhead with BEiT [3]. †: GPU-days estimated on Tesla V100 GPUs. ††: effective epochs following iBOT [49].

4.2. Transfer Results on ImageNet-1K

Setup. We consider the fully fine-tuning performance on ImageNet-1K for 100 epochs and report the Top-1 accuracy. Following MAE [22], we average all the patch tokens after the final Transformer layer, which is taken as input to a linear head for classification. See more details in Appendix B.

Comparison with MAE. As shown in Tab. 1, *MixedAE* obtains consistent improvements over MAE under different pre-training epochs with only 3% additional overhead. The 300-epoch pre-trained *MixedAE* with full mixing acquires even better accuracy than the 1600-epoch pre-trained MAE, demonstrating the efficiency of data utilization.

Comparison with other MIM augmentations. Our *MixedAE* with the lightweight decoder and the full mixing mode obtains 83.7% Top-1 accuracy, 0.5% and 0.4% higher than MixMIM [31] and CIM [18] respectively, meanwhile saving at least 23.4% computational overhead, revealing the simplicity of our *MixedAE*.

Comparison with other SSL approaches. Our *MixedAE* obtains consistent improvements under various pre-training epochs and overheads, and achieves a better trade-off between pre-training overhead and transfer performance, as shown in Fig. 1. Specifically, *MixedAE* pre-trained for 1600 epochs achieves 83.9% accuracy, constructing a new state-of-the-art result with a pure **autoencoder**-based framework. Requiring only 53.4% of the training overhead, *MixedAE*

surpasses the strong iBOT [49], demonstrating remarkable efficiency. Moreover, the improvement brought by mixing is orthogonal to the usage of more advanced reconstruction targets [2, 14, 41] and masking strategies [25, 39].

4.3. Transfer Results on Downstream Tasks

We further consider three downstream settings to evaluate the learnt representation, and more details about the different transfer procedures are included in Appendix B.

Semantic segmentation. We utilize the UperNet [43] to perform semantic segmentation on ADE20K [48] following BEiT [3]. As reported in Tab. 1, our 800-epoch *MixedAE* achieves 48.7 mIoU, even surpassing the MAE pre-trained for 1600 epochs by 0.6 mIoU, and our 1600-epoch *MixedAE* further improves to 49.8 mIoU, outperforming all baseline methods by a non-trivial margin, which is more significant than the improvement on ImageNet-1K (1.7 vs. 0.3), thanks to the object-aware pre-training, as discussed in Sec. 3.4.

Object detection and instance segmentation. We utilize Cascade Mask R-CNN [4, 23] to produce bounding boxes and instance masks simultaneously on COCO [30]. As demonstrated in Tab. 1, *MixedAE* consistently outperforms MAE under different epochs (0.6/0.9/0.9 & 0.5/0.6/0.7). Similarly with ADE20K, more significant improvements are observed due to the high-quality attention maps learnt by the object-aware pre-training, as demonstrated in Fig. 5.

Method	Aircraft	Caltech	Cars	C10	C100	DTD	Flowers	Food	Pets	SUN	VOC	Avg.
SSL ResNets												
MoCov2 [9]	79.9	84.4	75.2	96.5	71.3	69.5	94.4	76.8	79.8	55.8	71.7	77.7
SimCLR [8]	78.7	82.9	79.8	96.2	79.1	70.2	94.3	82.2	83.2	61.1	78.2	80.5
BYOL [20]	79.5	89.4	84.6	97.0	84.0	73.6	94.5	85.5	89.6	64.0	82.7	84.0
SwAV [5]	**83.1**	89.9	86.8	96.8	84.4	75.2	95.5	87.2	89.1	66.2	84.7	85.3
SDR [33]	82.6	89.0	87.5	97.4	84.4	75.6	97.0	86.1	89.3	66.1	85.3	85.5
SSL Transformers												
MoCov3 [10]	76.6	91.2	86.6	98.3	88.3	72.6	95.5	86.4	92.0	65.6	84.5	85.2
DINO [6]	69.4	91.2	81.3	**98.4**	**88.9**	77.6	96.9	87.3	93.5	64.7	86.3	85.1
BEiT [3]	66.3	80.2	78.6	96.1	80.0	69.9	92.9	83.2	85.3	57.1	76.7	78.7
MAE [22]	78.2	91.2	88.4	97.0	82.5	75.3	96.6	84.7	92.6	65.4	86.0	85.3
MixedAE	82.1	**91.5**	**88.8**	97.9	85.9	**78.7**	**97.1**	**87.4**	**93.6**	66.2	86.4	**86.9**$^{+1.6}$

Table 2. **Transfer performance comparison on 11 downstream classification tasks**. Our 1600-epoch pre-trained *MixedAE* achieves consistent improvements over MAE on all 11 classification datasets with an average accuracy of 86.9%, surpassing all counterparts.

Downstream classification. Following [17, 32, 33], we study transfer performance on 11 downstream classification datasets, including both fine-grained (*e.g.*, Cars [27]) and coarse-grained ones (*e.g.*, CIFAR100 [28]) in Tab. 2. Our *MixedAE* achieves consistent improvement over MAE on all 11 downstream tasks with an average accuracy of 86.9%, outperforming all counterparts as demonstrated in Tab. 2.

4.4. Ablation Study

Setup. We conduct 300-epoch pre-training with a base learning rate of $1.5e^{-4}$ for all ablation studies on *MixedAE* with compose mixing. By default, we report the fine-tuning accuracy on ImageNet-1K [11] and mIoU on ADE20K [48]. See more detailed settings and results in Appendix C.

Mixing ratio r. Different from MixMIM [31], the mixing ratio r in our formulation can be flexibly selected from the range of $(0, 0.5]$. As shown in Tab. 3a, $r = 0.25$ works better, while requiring less pre-training overhead (since the effective encoder batch size scales linearly with r, as shown in Eq. (1)). Note that $r = 0.25$ is also corresponding to the default 0.75 masking ratio in MAE.

Position of homologous contrastive. We study whether the encoder features before or after the final Layer Normalization (`[LN]`) [1] of ViT [15] achieves better performance as input to the homologous contrastive loss in Tab. 3b. The latter achieves consistent improvements on both ImageNet-1K and ADE20K, suggesting that the features after `[LN]` are more suitable for homologous recognition.

Positives of homologous contrastive. In Eq. (7), given a query patch, all homologous patches are considered as positive samples but taken separately to calculate $\mathcal{L}_{HomoCon}$ in the supervised contrastive manner [26]. We further ablate to utilize the average of both the query and all its homologous patches as its positive, which, however, performs slightly worse than the separate manner (-0.2 mIoU on ADE20K).

Threshold K of homologous attention. We study the threshold number K in $A_{HomoAtt}$ in Tab. 3d, where **all** the global self-attention operations in the ViT are replaced with our homologous attention. Compared with the no sampling baseline, the usage of homologous attention obtains consistent improvement on ImageNet-1K, while the best achieves at $K = 0.25L$, which is consistent with the mixing ratio r in Tab. 3a specifically.

Position of homologous attention. As shown in Fig. 6a, homologous attention cannot achieve promising accuracy in early Transformer layers without sufficient information engagement. Thus, we further explore to maintain the global self-attention in early layers unchanged in Tab. 3e, and observe empirically that utilizing global self-attention in the very first layer only achieves the best performance.

Homologous recognition. In Tab. 3f, we further compare the effectiveness of different components of homologous recognition. Without the homologous contrastive loss for verification, utilizing homologous attention only obtains a significant drop of 0.5 mIoU on ADE20K. Although achieving improvement, utilizing the homologous contrastive only still suffers from the ease of reconstruction brought by MI increasement, as previously discussed in Eq. (5). Finally, the best performance is achieved when using homologous attention and contrastive loss simultaneously.

4.5. Analysis

Effectiveness of TopK sampling. To further observe the effectiveness of TopK sampling in homologous attention, we visualize the sampling accuracy with respect to different layers and pre-training epochs in Fig. 6. As demonstrated in Fig. 6a, the naïve usage of homologous attention only cannot achieve promising sampling accuracy, which, therefore, suffers from a significant performance drop in Tab. 3f. Specifically, neither the sampling accuracy of the first two

mixing ratio r	GPU-days	acc	mIoU
0.25	**16.9**	**82.4**	**45.0**
0.5	21.0	82.3	44.2

(a) **Mixing ratio** r, which is not limited to 0.5, different from the MixMIM [31] formulation.

case	acc	mIoU
pre-[LN]	82.4	44.5
post-[LN]	**82.5**	**45.3**

(b) **Position of homologous contrastive**. The encoder features after [LN] works better.

case	acc	mIoU
all positives separately	**82.5**	**45.3**
mean positive	82.5	45.1

(c) **Positives of homologous contrastive**. It's better to utilize positive patches separately.

threshold K	acc	mIoU
$0.125L$	82.6	45.1
$0.25L$	**82.6**	**45.6**
$0.5L$	82.6	45.5
$1.0L$ (no sampling)	82.5	45.3

(d) **Threshold** K works best when consistent with the mixing ratio r in Tab. 3a.

layer ID	acc	mIoU
none	82.6	45.6
1st	**82.7**	**46.4**
1st & 2nd	82.5	44.8
1st & 2nd & 3rd	82.5	45.6

(e) **Usage of global self-attention** in the very first layer only achieves the best performance.

	HomoAtt	HomoCon	acc	mIoU
			82.4	45.0
MixedAE	✓		82.6	44.5
		✓	82.5	45.3
	✓	✓	**82.7**	**46.4**

(f) **Homologous recognition**. Best to utilize homologous attention and contrastive together.

Table 3. *MixedAE* **ablation experiments** with ViT-B/16. We report the ImageNet-1K fine-tuning accuracy (acc) and ADE20K semantic segmentation performance (mIoU). We explore (a) the mixing ratio r, (b,c) the position and positives for homologous contrastive, (d,e) the threshold K and position for homologous attention, and (f) a summary of homologous recognition. Default settings are marked in gray .
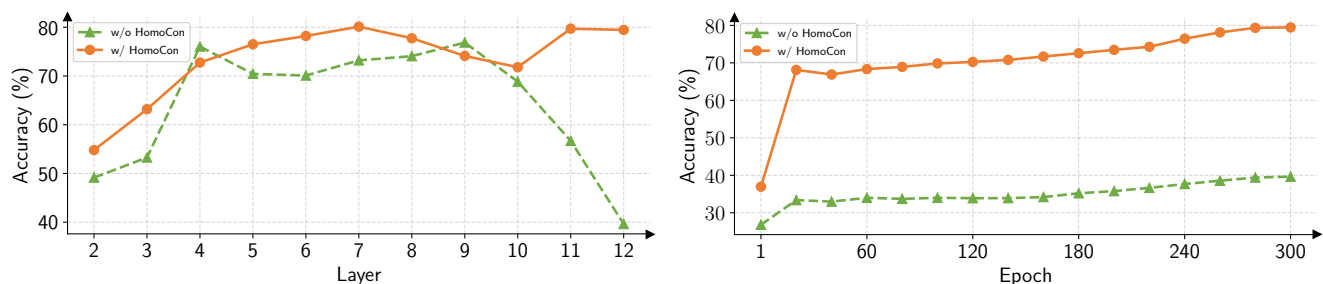
(a) **Accuracy curve with respect to Transformer layers** of our *MixedAE* pre-trained for 300 epochs, where the 1st layer adopts global self-attention.

(b) **Accuracy curve with respect to pre-training epochs** of the very final Transformer layer in our *MixedAE*.

Figure 6. **Analysis of** TopK **sampling accuracy**. Without the homologous contrastive loss for verification (green curve), utilizing the homologous attention only cannot achieve promising accuracy, dramatically varying between different layers. However, with the usage of homologous contrastive loss (orange curve), the sampling accuracy is significantly improved and stabilized mostly around 70% to 80% throughout the whole pre-training process, which is essential to achieve remarkable transfer performance, as demonstrated in Tab. 3f.

nor final two layers exceed 60%, and accuracy of the very final layer even maintains under 40% throughout the whole pre-training process, as shown in Fig. 6b (the green curve).

However, with the usage of the homologous contrastive loss for verification, the sampling accuracy is significantly improved and stabilized around 70% to 80% for all layers except the first two, as in Fig. 6a. The sampling accuracy of the final layer rapidly increases to around 70% when pre-trained only for 20 epochs, maintaining stable throughout the remaining pre-training, as in Fig. 6b (the orange curve).

Comparison with existing MIM methods combined with contrastive. Although also utilizing a "contrastive loss" with reconstruction, *MixedAE* differs from existing MIM works [16,49] combined with contrastive learning from two perspectives, 1) **Purpose**: existing works utilize contrastive loss to perform instance discrimination simultaneously with MIM, while our homologous contrastive is only utilized to guarantee the sampling accuracy. Therefore, homologous contrastive performs more like a regularization term instead of an individual self-supervision in [16,49]. To verify, we pre-train a *MixedAE* with $\mathcal{L}_{HomoCon}$ only without $\mathcal{L}_{recon}$, which cannot achieve reasonable performance, as reported

in Appendix C. 2) **Efficiency**: given a single input, existing works require forward propagation at least twice for several augmented views to conduct instance discrimination, while only once for our homologous contrastive, resulting in the significant efficiency. Specifically, our *MixedAE* surpasses iBOT [49] with only 53.4% of its computational overhead.

5. Conclusion

This paper explores the usage of image mixing for MAE. Different from in supervised and contrastive learning, we first theoretically demonstrate naïve mixing might instead ease the reconstruction pretext task. To address that, our *MixedAE* with the proposed homologous recognition as the auxiliary supervision can not only achieve state-of-the-art performance with a better trade-off between transfer results and pre-training overhead, but also conduct object-aware pre-training without any specifically designed modules. We hope our simple yet effective method can bring researchers' attention to more effective data augmentations for MIM.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 7

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 1, 2, 6

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 5, 6, 7

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, 2018. 6

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 7

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 5, 6, 7

[7] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 7

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 7

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1, 5, 6, 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 4

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 1, 2

[14] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 1, 2, 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 5, 7

[16] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 1, 2, 8

[17] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, 2021. 7

[18] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 2, 6

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2

[20] 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, and others. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2, 7

[21] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 1

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 2, 3, 4, 5, 6, 7

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 6

[24] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022. 5

[25] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 1, 2, 6

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 4, 7

[27] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Workshop on Fine-Grained Visual Categorization*, 2013. 7

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 7

[29] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 1

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5, 6

[31] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 2, 3, 6, 7, 8

[32] Zhili Liu, Kai Chen, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, and James T. Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. 7

[33] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 7

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 1, 2

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[38] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *AAAI*, 2022. 1, 4

[39] Yuge Shi, N Siddharth, Philip HS Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. *arXiv preprint arXiv:2201.13100*, 2022. 1, 2, 6

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 6

[41] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 1, 2, 6

[42] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In *NeurIPS*, 2021. 5

[43] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 6

[44] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021. 1, 2

[45] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Wenrui Dai, Hongkai Xiong, and Qi Tian. Seed the views: Hierarchical semantic alignment for contrastive representation learning. In *PAMI*, 2022. 1, 4

[46] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1, 4

[47] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 4

[48] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. In *IJCV*, 2019. 1, 5, 6, 7

[49] 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. 1, 2, 6, 8