Masked Auto-Encoders Meet Generative Adversarial Networks and Beyond

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang*
Xiaoming Wei, Xiaolin Wei
Meituan
{feizhengcong, fanmingyuan, zhuli09, huangjunshi}@meituan.com
{weixiaoming, weixiaolin02}@meituan.com

Abstract

Masked Auto-Encoder (MAE) pretraining methods randomly mask image patches and then train a vision Transformer to reconstruct the original pixels based on the unmasked patches. While they demonstrates impressive performance for downstream vision tasks, it generally requires a large amount of training resource. In this paper, we introduce a novel Generative Adversarial Networks alike framework, referred to as GAN-MAE, where a generator is used to generate the masked patches according to the remaining visible patches, and a discriminator is employed to predict whether the patch is synthesized by the generator. We believe this capacity of distinguishing whether the image patch is predicted or original is benefit to representation learning. Another key point lies in that the parameters of the vision Transformer backbone in the generator and discriminator are shared. Extensive experiments demonstrate that adversarial training of GAN-MAE framework is more efficient and accordingly outperforms the standard MAE given the same model size, training data, and computation resource. The gains are substantially robust for different model sizes and datasets, in particular, a ViT-B model trained with GAN-MAE for 200 epochs outperforms the MAE with 1600 epochs on fine-tuning top-1 accuracy of ImageNet-1k with much less FLOPs. Besides, our approach also works well at transferring downstream tasks.

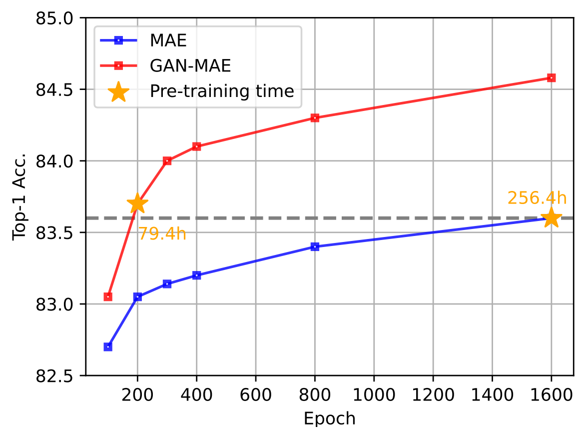

Figure 1. **Performance comparison in different pre-training epochs for ImageNet-1K Fine-tuning top-1 accuracy.** Compared to MAE trained for 1600 epochs, GAN-MAE achieves comparable accuracy with much less training time at 200 epochs.

1. Introduction

In recent years, Transformer [62] has become the *de facto* standard architecture in computer vision, and has surpassed state-of-the-art Convolutional Neural Network (CNN) [31, 58] feature extractors in vision tasks through models such as the Vision Transformer [21]. Meanwhile, self-supervised learning (SSL) algorithms [12, 14, 27, 29] aims to learn transferable representation from unlabeled data by performing instance-level pretext tasks, and has been a long-standing target in the vision community. Particularly, masked image modeling (MIM) in SSL for vision transformers has shown remarkably impressive downstream performance in a wide variety of computer vision tasks [3, 28], attracting increasing attention.

MIM is a simple pretext task that first randomly masks some patches of an image, and then predicts the contents of the masked patches according to the remaining, using various reconstruction targets, *e.g.*, visual tokens [3, 19], semantic features [1, 77] and raw pixels [28, 70]. Essentially, it learns the transferable representation by modeling the image structure itself as content prediction. While more effective than conventional pre-training, masked autoencoder modeling approaches still exist some issues: (**i**) reconstruction optimization with MSE loss leads to blurrier output images than the raw input, it would be better to use a more perceptual loss over pixels to guide the fine-grained seman-

*The corresponding author.

tic understanding and representation learning, leading to more plausible synthesized patches; (**ii**) inner dependency between masked patches is lacked [71], *i.e.*, generation of masked image patches may lack the surrounding information. This situation becomes more serious when the image patch masking ratio is large. We alleviate this problem by introducing confident synthesized patches as complementary information during training; (**iii**) mask-reconstruction methods incur a substantial computation cost because the network only learns from part of the visible patches and misses the information of masked patches.

In this paper, we propose a Generative Adversarial Networks-based pre-training framework, referred to as GAN-MAE, which contains two components: a generator model learns to reconstruct the masked patches according to visible patches in the encoder-decoder architecture and a discriminator model learns to distinguish real image patches from plausible but synthesized remains. Generally, given an image from training dataset, our method first randomly masks parts of patches and reconstructs them using the rest visible patches with a generator, which serves as a standard MAE model. Then we build the corrupt image as the combination of visible and synthesis patches, which is then fed into the discriminator to predict whether each patch is from raw image or synthesized results. In this manner, the discriminator provides a valid guiding for more delicate image patch modeling. Then, with the development of generator capacity, a key advantage of discriminative task is that it integrates the synthesized patches into corrupt images as complementary information, which fills the missing inner relationship between patches during pre-training. Moreover, we shared the parameters of vision transformer backbone in the generator and discriminator to promote memory reduction, training efficiency, as well as performance enhancement.

Our experiments follow the same architecture, settings, and pre-training recipe as MAE [28], and we find that the simple incorporation of a discriminator consistently outperforms MAE in variant models, *e.g.*, ViT-S, ViT-B, and ViT-L, when fine-tuning for top-1 accuracy of ImageNet classification. We also conduct extensive ablation studies to validate the effectiveness of our core designs in backbone parameter sharing and advarisal training. As pre-training with more epochs usually results in a better downstream performance, we argue that an important consideration for pre-training methods should be computation efficiency as well as absolute downstream performance. From this viewpoint, we also demonstrate that discrimination of pseudo-image patches forces GAN-MAE to train more efficiently than standard MAE. We further provide a comprehensive comparison with MAE in various epochs and various models and show our framework achieves consistently better performance. In particular, as presented in Figure 1, for the ViT-B model structure, our GAN-MAE achieves comparable classification performance with only 200 pre-training epochs vs. standard MAE 1600 pre-training epochs. Furthermore, the GAN-MAE achieves 0.7 points improvement when pre-training 1600 epochs. Finally, we summarize our contribution as follows:

- We propose a new and effective GAN-alike framework for visual representation self-supervised learning, which to our best knowledge is the first trial of integrating GAN idea into MAE framework. As a generic approach, we suggest that this framework can be easily applied on many other MIM-based tasks.

- We introduce two core designs: shared weight for the main backbones of generator and discriminator, and an adversarial training process, both of which cost fewer amounts of computing resources while obtaining appreciable performance improvements.

- Extensive experiments demonstrate that compared with the original MAE, our method is more compute-efficient and results in better transfer representation learning on downstream tasks.

2. Related Works

Autoencoding. Autoencoder [6, 7, 35] is an unsupervised learning technique for neural networks that learns efficient representations by training the network to ignore signal noise. It includes an encoder that maps the original data to a low-dimensional latent embedding and a decoder that recovers the data from the latent embedding, with the goal of learning a compressed knowledge representation. The denoising autoencoder [63, 64] learns to reconstruct clean data points from a noisy version. Numerous efforts have been devoted for image denoising, such as masking pixels [11, 55, 64], inpainting [69], removing color channels [39,72], and shuffling image patches [22,54]. For a broader overview of denoising autoencoder, we refer to [5,61].

Masked Image Modeling. Masked language modeling [37,49,57], which generalizes well on language understanding and generation tasks, is the domain self-supervised approach in the field of NLP. Similarly, vision transformer [21, 41, 50] based masked image modeling (MIM) approaches [1,3,28,70,77] for computer vision tasks have also been developed. Generally, these MIM approaches first apply a mask to patches of an image, and then the masked patches are predicted given the visible patches. Feature representation learned through such within-image context prediction demonstrate strong transfer performance in downstream tasks. Recently, lots of works exploring MIM have been concurrently developed from different perspectives.

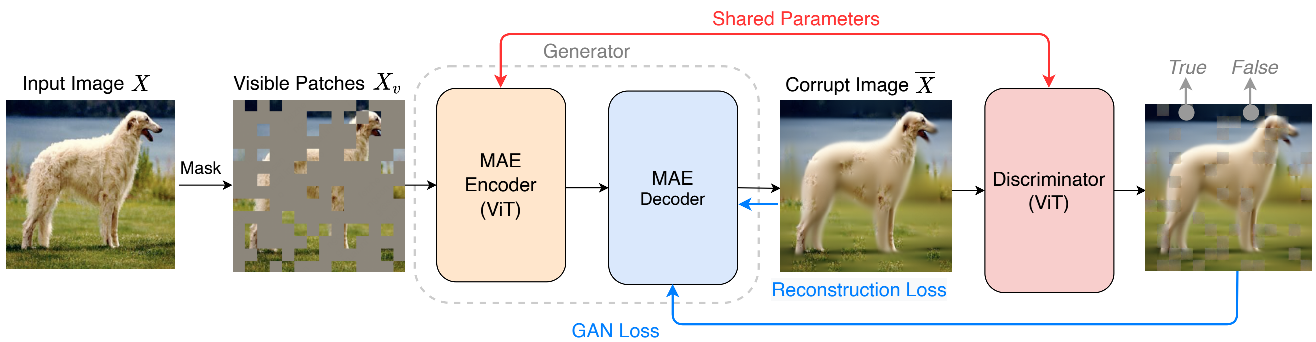

Figure 2. **Overview of GAN-MAE framework**, where a generator is used to predict the masked patches and a discriminator is employed to classify whether the patches are selected from the raw image or synthesized results. In particular, both of MAE encoder and discriminator are based on the vision Transformer backbone and share parameters for memory reduction and training efficiency.

The works include framework design [13, 23, 28, 70], prediction targets [2, 19, 66, 77], and integration with vision-language representation learning [42, 43, 75]. Our work belongs to the first group and introduces a novel GAN framework that discriminates the reconstructed image patches in a standard MAE to learn deeper semantics.

Generative Adversarial Networks. GANs [26] are effective at generating high-quality synthetic data. Usually, the generator generates an image, and the discriminator determines whether the input image is a real image or a generated image. Subsequently, many improvements based on the original GAN focused on speeding up the training of the network and improving the quality of the generated images [8, 52, 74]. These improvements also help GAN achieving a wider range of applications [45, 48, 67]. Methods based on GAN are also widely used in image-to-image translation [36], super-resolution [4, 40], style transfer [15], text generation [10, 24, 73], and representation learning [17, 25], to name a few. Particularly, [56] brings some designs of CNN architecture to stabilize the training of GAN framework, after that the discriminator can be directly used as feature extractor in downstream tasks. In contrast, our method proposes to integrate the GAN as assistant for the MIM task, which focuses on the study of better pattern for masking based self-supervised learning. Besides, our strategy of shared parameters provides a unified backbone for better vision representation.

3. Approach

In this section, we introduce the GAN-MAE framework in details. At first, we briefly review the conventional masked autoencoder model in Sec. 3.1, and then describe the proposed generator-discriminator pre-training framework in Sec. 3.2. The architecture of our framework is presented in Figure 2. Finally, we suggest an adversarial training processes under the proposed method and discuss our framework in Sec. 3.3 and Sec. 3.4, respectively.

3.1. Preliminaries

Masked Autoencoder (MAE) [28] is a self-supervised approach with a vision transformer encoder and a small transformer decoder, which randomly masks a large portion of input patches, and then reconstructs the masked patches according to the visible patches. Specifically, provided with an image $X \in \mathbb{R}^{C \times H \times W}$, where C, H and W are the channel number, image height and image width respectively, MAE partitions X into $N = \frac{H \times W}{P^2}$ non-overlapping patches with patch size P. In this way, the image is transformed into a sequence of patches $X = \{x^1, \dots, x^N\}$ with each element $x^k \in \mathbb{R}^{P^2 \times C}$. Then, we sample a random set of patch index M in uniform distribution, and split the image patches X into masked patch set $X_m = \{x^k | k \in M\}$ and visible patch set $X_v = \{x^k | k \notin M\}$. During training, the MAE encoder inputs X_v to achieve the latent representations H_v. Then, the MAE decoder attempts to reconstruct X_m with the input of interpolating $[mask]$ token embedding into the sequence of latent representations H_v according to the index set M, and outputs the reconstructed patches $\tilde{X}_m$. Finally, MAE optimizes the mean-squared error reconstruction loss on the masked patches as:

$$L_{mae}(X, M, \theta_{mae}) = \sum_{k \in M} ||\tilde{x}^k - x^k||_2^2, \tag{1}$$

where θ_{mae} represents the parameters of MAE model.

3.2. When MAE meet GAN

In this section, we describe the proposed GAN-MAE framework, as presented in Figure 2. Generally, our framework consists of two parts, a generator G to reconstruct the masked image patches and a discriminator D to predict the realness of image patches.

Image Patch Generator. Identical to the standard MAE, the generator follows the encoder-decoder paradigm and is trained to perform masked-image reconstruction task. Given the partitioned image patches X, the generator randomly masks some image patches and encodes the remaining visible patches X_v into a sequence of contextualized vector representations H_v, based on which the masked patches are reconstructed as $\tilde{X}_m$. Please refer to Sec. 3.1 for more details. In general, the generation process can be formulated as:

$$M \sim \text{Uniform}(1, N), \tag{2}$$

$$H_v = f_e(X_v, M), \tag{3}$$

$$\tilde{X}_m = f_d(H_v, M), \tag{4}$$

where N is the number of image patches. $f_e(\cdot)$ and $f_d(\cdot)$ denote the encoder and decoder in a conventional MAE.

Image Patch Discriminator. For a patch index k and corrupted image sequence $\overline{X} = \{X_v, \tilde{X}_m\}$, the discriminator predicts whether the patch token x^k is real or synthesized as binary classification task. Specifically, we create the corrupted image $\overline{X}$ by maintaining the visible patches X_v in raw image X and replacing the masked patches with generator predicted result $\tilde{X}_m$. Note that X_v and $\tilde{X}_m$ are absolute complement of set to each other, *i.e.*, $X_v \cup \tilde{X}_m = \overline{X}$ and $X_v \cap \tilde{X}_m = \varnothing$. Formally, the task of discriminator model can be formulated as:

$$D(\overline{X}, k) = p_{disc}(y^k|\overline{X}, k), \text{ for } k \in [1, N]. \tag{5}$$

Let the ground-truth classification label sequence $Y = \{y^1, \ldots, y^N\}$ with each element $y_k \in \{0, 1\}$, where 0 and 1 denote the corresponding image patch is reconstructed from generator or comes from the original image, respectively. The training objective of the discriminator can be formulated as:

$$\begin{aligned} L_{disc}(\overline{X}, \theta_{disc}) = \sum_{k=1}^{N} & -y^k \log D(\overline{X}, k) \\ & - (1 - y^k) \log (1 - D(\overline{X}, k)). \end{aligned} \tag{6}$$

Particularly, though the corrupted image is partially unreal, we suggest that this discriminative task can still be benefit to the learning of feature representation with the plausible corrupted image from a well-trained generator, as presented in experiments. Hitherto, due to the GAN-alike strategy for MAE task, we name our framework as GAN-MAE.

3.3. Training Scheme

We explore the training strategy of our proposed GAN-MAE in this section. Particularly, to improve synthesis result, we augment the L-2 reconstruction loss with a perceptual loss that aims to differentiate the real patches and reconstructed patches as:

$$L_{adv}(X, \theta_{mae}) = \log D(X_v) + \log(1 - D(\tilde{X}_m)). \tag{7}$$

Please note that $\tilde{X}_m$ is the reconstructed patches by MAE. Therefore, the final objective for the generator comes to minimize the combined loss as:

$$L_{gen}(X, \theta_{mae}) = L_{mae}(X, \theta_{mae}) + \gamma L_{adv}(X, \theta_{mae}), \tag{8}$$

where we compute the adaptive weight γ according to:

$$\gamma = \frac{\nabla[L_{mae}]}{\nabla[L_{adv}] + \delta}, \tag{9}$$

$\nabla[\cdot]$ denotes the gradients of different loss function w.r.t. the parameters of last layer in network, and $\delta = 1e-6$ is used for numerical stability. Intuitively, the integration of γ adaptively balances the contributions of two loss functions to the gradients of parameters.

Based on the aforementioned strategy, the final training algorithm can be formulated in Algorithm 1. Specifically, at each epoch, we conduct the iteration in two steps: **(i)** train only the generator with L_{gen}; **(ii)** Train the discriminator with L_{disc}. During training, we observe that shared weights of generator and discriminator backbones can usually stabilize the training procedure and bring extra-gain in down-stream tasks.

Algorithm 1: Adversarial training for GAN-MAE

Data: Training data $\mathcal{D}_{train}$, total epoch number N_e, GAN-MAE model with generator parameters θ_{mae} and discriminator parameters θ_{disc};

share weights between generator and discriminator backbones;
while $n_e < N_e$ **do**
  for $x^i \in \mathcal{D}_{train}$ **do**
    ▷ generator training;
    sample masking set M^i and mask image x^i;
    predict masked image patches $\tilde{x}^i_m$;
    compute loss L_{gen};
    loss backward for updating θ_{mae};
    ▷ dicriminator training;
    construct $\overline{x}^i$ based on x^i and $\tilde{x}^i_m$;
    comput loss L_{disc};
    loss backward for updating θ_{disc};
  end
  $n_e + = 1$;
end

3.4. Discussion

Theoretically, the integration of discriminator can be considered as a high-level perceptual loss, which forces the generator to learn better feature representation for the plausible synthesis of masked patches. As the quality of synthesized patches improved, we claim that the introduction of corrupted images, serving as *complementary information*, provides plausible inner dependency between full image patches for representation learning. Furthermore, with the shared parameters of backbones in generator and discriminator, the GAN-alike framework can be considered as a type of *multi-task learning*, leading to even better result as presented in experiments.

4. Experiments

4.1. Datasets and Settings

In the experiments, we pre-train our GAN-MAE model on the ImageNet-1k [16] and evaluate the performance in end-to-end fine-tuning (FT) pattern for the task of classification, semantic segmentation, object detection and instance segmentation. The evaluation metric of classification is the top-1 validation accuracy on 224×224 cropped input images. The input image is partitioned into 14×14 patches and each patch is of size 16×16. Following the setting of MAE, we only use the standard random cropping and horizontal flipping for data augmentation. To validate the effectiveness of GAN-MAE framework, the used ViT architecture and most hyper-parameters are exactly the same to [28,60], *i.e.*, ViT-S (12 transformer blocks with dimension 384), ViT-B (12 transformer blocks with dimension 768), and ViT-L (24 transformer blocks with dimension 1024). All version of ViT models are trained with 4096 batch size on 8 V-100 32GB GPUs. We adopt dynamic token masking with the masked positions decided on-the-fly. We use AdamW [38] optimizer and cosine schedule [51] with warm up for model training. The learning rate is annealed according to the cosine schedule. Unless stated otherwise, results are evaluated on the dev set. Please refer to appendix A for more training implementation and hyperparameter values for different backbone in details.

4.2. Analysis of GAN-MAE

We analyze our GAN-MAE framework by proposing and evaluating several extensions to the model. Unless stated otherwise, all these experiments use the same model size as ViT-B and training dataset ImageNet-1K.

Parameter Sharing of Backbone. In this work, we propose to improve the efficiency of the pre-training by sharing the parameters of the backbone vision transformer between the generator encoder and discriminator. One cause is that the generator and discriminator utilize the same network architecture, and all of the transformer weights can be tied. However, we can release the weight sharing and train the generator and discriminator independently. In between, we can adopt both the parameters of generator and discriminator for downstream learning. The experimental results are shown in Table 1, where we employ the adversarial training with the same training epochs. We can see that, generally, the fine-tuning top-1 accuracy of shared weight outperforms the independent generator and discriminator conspicuously, in particular 0.4 point improvements for the generator when pre-training 800 epochs. We hypothesize that GAN-MAE framework benefits from both mask-then-reconstruct and pseudo-patch classification tasks, which can incorporate the visual semantic and consistency understanding.

Table 1. Effect of **parameter sharing** in GAN-MAE framework. Results demonstrate that shared parameters for backbone benefits both memory cost and performance improvements.

Models	Epoch	Mask ratio	FT
Generator	800	75%	83.9
Discriminator	800	75%	84.2
Shared	800	75%	**84.3**
Generator	1600	75%	84.4
Discriminator	1600	75%	84.4
Shared	1600	75%	**84.6**

It is also surprising that under the GAN process without weight sharing, the discriminator with learning of patch classification can lead to better performance than generator. We suspect that the fine-grained patch classification is more delicate while harder than image-level classification. Thus, we further tried to add an image-level contrastive objective. For this task, we input 50% of the input image unchanged rather than noising them with the generator. We then added a prediction head to the model that predicted if the entire input image was corrupted or not. However, the result didn't improve the final accuracy. In conclusion, we believe that the design of discriminative task, *e.g.*, getting closer to downstream and be more difficult, is an important exploration direction in the future. Another interesting finding is that with the pre-training epochs increase, the independent generator performance boosts, we believe that it is caused by the adversarial training, where the discriminator becomes an effective guiding for generation.

Training Schemes. We analyze the effect of proposed adversarial training scheme on GAN-MAE with shared backbones. Two training variants are considered as following:

- *Two-stage Training.* It is natural to consider to remove the discriminator guiding signal during generator training, which leads to the disentangled optimizing process. Specifically, at each epoch, we do the following steps: train the generator only with L_{mae} and train the discriminator only with L_{disc}. The difference with ad-

Table 2. Effect of different **training schemes**.

Models	Epoch	Mask ratio	FT	GPU Time
Two-stage	300	75%	82.0	94.3h
Combined	300	75%	**82.2**	**90.9h**
Adversarial	300	75%	**82.2**	118.8h
Two-stage	800	75%	84.0	252.2h
Combined	800	75%	84.1	**240.5h**
Adversarial	800	75%	**84.3**	317.5h

Table 3. Effect of **different masking ratio** in GAN-MAE framework with different pre-training epochs.

Models	Epoch	Mask ratio	FT
MAE	800	75%	83.4
GAN-MAE	800	70%	**84.3**
GAN-MAE	800	75%	**84.3**
GAN-MAE	800	80%	84.0
MAE	1600	75%	83.6
GAN-MAE	1600	70%	84.5
GAN-MAE	1600	75%	**84.6**
GAN-MAE	1600	80%	84.4

versarial training lies in that the training loss of the generator in the first step is changed.

- *Combined Training.* Instead of iterative in total dataset, we can jointly trains the generator and discriminator at each step. That is, for each image X in the training dataset $\mathcal{D}_{train}$, we can directly minimize the combined loss as:

$$L_{gen}(\theta_{mae}) + \lambda L_{disc}(\theta_{disc}). \tag{10}$$

  We set λ, the weight for the discriminator objective in the loss to 2.0, as we searched for λ out of [1,2,5,10] in early experiments.

Note that regardless of the form, the nature of ***sequential training***, *i.e.*, the corrupt image built from generator will be fed into discriminator, is not changed.

The evaluation results for classification are listed in Table 2. GPU time means pre-training time (hours) on 8 V100 32GB GPUs environment. As we can see, the combined training shows a superior training time while adversarial training slows down as the training procedures of generator and discriminator are isolated. On the other hand, the performance of adversarial training is not better than combined training in the early stages; when pre-training epochs come to 800, the benefits of adversarial training appear.

Table 4. Comparison of **computation resource usage** during self-supervied pre-training.

Backbone	Models	FLOPs	Params
ViT-B	MAE	9.4e9	111.654M
ViT-B	GAN-MAE	2.6e10	111.656M
ViT-L	MAE	2.0e10	329.238M
ViT-L	GAN-MAE	8.0e10	329.240M

Masking Ratio. Table 3 shows the influence of masking ratio in MAE-GAN under different pre-training epochs. The optimal ratio for MAE-GAN is identical to 75%, showing a obviously better classification performance compared with same masking ratio in previous work [28]. Meanwhile, we present several reconstructed images from MAE and GAN-MAE. Although the MIM model can infer missing patches as different yet plausible outputs when mask ratio comes to large, the generator of GAN-MAE can predict the images with more realistic and fine-grained details, *e.g.*, the outline of a mountain peak in first case, which we believe is relative to the learning of useful representation.

Computation Resource. In terms of pre-training cost, we conduct a computing resource comparison with the baseline MAE on different vision transformer backbones. The results are shown in Table 4. We first point out that GAN-MAE slows down the training process as the discriminator integration, *e.g.* 317.5h vs. 127.7h for ViT-B model in 800 epochs. Moreover, we also choose to measure computation usage in terms of floating point operations (FLOPs) as it is a metric agnostic to the particular hardware, low-level optimizations, etc. Note that an "operation" is a mathematical operation, not a machine instruction and `thop` package is used to compute FLOPs in practice. As expected, GAN-MAE employs ~ 3 extra theoretical computation cost. Besides, benefits from the weight sharing of backbone, the GAN-MAE incorporates less than 1% addition parameter, which helps to reduce the running memory usage, and provides a convenience to set large batch size. Last but not least, we want to state that although our GAN-MAE method incorporates additional computation resources during training at each epoch, considering the superiority of reduction of epoch number and classification performance improvements, it is fully acceptable and exploration valuable.

4.3. ImageNet Classification Comparison

We compare our GAN-MAE methods with previous state-of-the-art works on the ImageNet-1K classification task. Table 5 reports the top-1 validation accuracy for fine-tuning results. We can find that compared to the supervised models, trained from scratch, all of the self-supervised pre-training methods achieve significant improvement, sug-

Figure 3. **Qualitative analysis for patch reconstruction.** Example results are from ImageNet validation set. For each tuple, we show the raw image, masked image, MAE reconstructed image, and our proposed GAN-MAE reconstructed image from left to right. We can see that the reconstructed images from GAN-MAE are significantly clearer than MAE, which we believe benefits the fine-grained visual semantic understanding.

Table 5. **End-to-end fine-tuning on ImageNet-1K.** We report the fine-tuning top-1 accuracy for classification in different vision transformer architectures and results show that GAN-MAE outperforms previous self-supervised methods.

Model	Pre-train data	Pre-train epochs	ViT-S	ViT-B	ViT-L
Supervised [59]	IN1K w/ labels	300	79.7	81.8	82.6
DINO [9]	IN1K	800	81.5	82.8	-
MoCo v3 [14]	IN1K	300	81.4	83.2	84.1
BEiT [3]	IN1K+DALLE	800	81.7	83.2	85.2
MSN [1]	IN1K	600	-	83.4	-
iBOT [77]	IN1K	800	82.3	84.0	84.8
BootMAE [20]	IN1K	800	-	84.2	85.9
MAE [28]	IN1K	800	-	83.4	85.4
MAE [28]	IN1K	1600	-	83.6	85.9
GAN-MAE	IN1K	300	82.2	84.0	85.6
GAN-MAE	IN1K	800	**82.4**	**84.3**	**86.1**

gesting the effectiveness of pre-training. We further compare our GAN-MAE framework with prior popular self-supervised pre-training models. We can see that the proposed GAN-MAE achieves the best fine-tuning performance either based on the ViT-B or based on the ViT-L architectures. For example, compared with the recent work MAE [28], our GAN-MAE in ViT-L network achieves 86.1% with 0.7 point improvement when pre-trained on 800 epochs. More encouragingly, for all ViT backbone sizes, GAN-MAE mostly outperforms the previous self-supervised methods. These results suggest that the incorporation of a discriminator scheme could have consistently benefits for various scale ViT models.

In addition, we present a comprehensive comparison with MAE under different pre-training epochs for the ViT-B model. We plot the results in Figure 1. We can see that our GAN-MAE approach consistently performs better than MAE. It is worth mentioning that the proposed GAN-MAE at 300 epochs achieves 84.0% accuracy, which is already better than MAE pre-trained at 1600 epochs. This demonstrates that our approach is more efficient to achieve comparable performance. What's more, no additional speed and parameter cost during inference. Besides, we can confirm that, similar to other MIM-based self-supervised training, the accuracy of GAN-MAE also improves steadily as the pre-training steps increase. Particularly, our GAN-MAE on ViT-B achieves 84.6% top-1 accuracy at 1600 epochs, which is almost 1% higher than that of MAE.

Table 6. **Robustness Evaluation** on the four ImageNet-variants: ImageNet-C, ImageNet-A, ImageNet-R, and ImageNet-Sketch. Except for ImageNet-C which is measured in terms of mean Corruption Error (mCE), top-1 accuracy is used as the remaining evaluation metric. For simplicity, we denoted IN-C, IN-A, IN-R, In-Skectch correspondingly.

Model	IN-C (mCE ↓)	IN-A (top-1 ↑)	IN-R (top-1 ↑)	IN-Sketch (top-1 ↑)
Supervised [53]	42.5	35.8	48.7	36.0
MAE [28]	51.7	35.9	48.3	34.5
GAN-MAE	**49.5**	**36.8**	**49.6**	**35.9**

Table 7. **Semantic segmentation** comparison on the ADE20K dataset for mIoU (%) metric with the ViT-B backbone.

Models	Pre-train data	Epochs	mIoU
Supervised [28]	IN1K w/ labels	300	47.4
MoCo v3 [14]	IN1K	300	47.3
BEiT [3]	IN1K+DALLE	800	47.1
MAE [28]	IN1K	800	47.6
MAE [28]	IN1K	1600	48.1
BootMAE [20]	IN1K	800	49.1
GAN-MAE	IN1K	800	**49.5**

Table 8. **COCO object detection and segmentation** using Mask R-CNN framework with ViT-B backbone.

Models	Pre-train data	AP-box	AP-mask
Supervised [28]	IN1K w/ labels	44.1	39.8
MoCo v3 [14]	IN1K	44.9	40.4
BEiT [3]	IN1K+DALLE	46.3	41.1
MSN [1]	IN1K	46.6	41.5
iBOT [77]	IN1K	47.3	42.2
MAE [28]	IN1K	47.2	42.0
BootMAE [20]	IN1K	48.5	43.4
GAN-MAE	IN1K	**49.0**	**43.8**

4.4. Downstream Tasks

Semantic Segmentation. We compare our GAN-MAE with supervised as well as state-of-the-art self-supervised models on the widely used dataset ADE20K [76] for semantic segmentation. Specifically, we use the UperNet framework [68] in the experiments. We train Upernet for 160K iterations with batch size set as 64 and report the results in Table 7. The evaluation metric is mean Intersection of Union (mIoU) averaged over all semantic categories and the single-scale test results are reported. Importantly, we can see that the proposed GAN-MAE gets superior performance than all the other baselines in the same configuration, further validating the effectiveness of adversarial training with a reconstructed patch discriminator.

Object Detection and Segmentation. We also perform object detection and instance segmentation, compared with other popular self-supervised methods and the supervised model, on the COCO dataset [47]. In practice, we choose the Mask R-CNN [30] framework and adopt FPNs [46] to scale the feature map into different sizes as introduced in [44]. The performance is tested on the COCO validation set, following the previous work [18]. The results are listed in Table 8 in terms of box AP metric for object detection and mask AP metric for instance segmentation. Importantly, we can observe that our GAN-MAE model achieves 49.0% for object detection and 43.8% for segmentation, surpassing the previous state-of-the-art BootMAE by 0.5% and 0.4% point, respectively.

Classification Robustness. Similar to [28], we further evaluate the robustness of classification performance on the four ImageNet variants, *i.e.*, ImageNet-C [33], ImageNet-A [34], ImageNet-R [32], and ImageNet-Sketch [65], which are common benchmarks to evaluate robustness for perturbations. Table 6 demonstrates the robustness comparison with GAN-MAE and MAE using the ViT-B backbone, as well as previous supervised SoTA models. The results illustrate that GAN-MAE outperforms the MAE baseline consistently on all robustness datasets, indicating that the promising of adversarial training in representation learning.

5. Conclusion

In this paper, we have proposed a new self-supervised GAN-alike framework for visual representation learning, where a generator is used to predict masked image patches according to the visible patches and a discriminator is employed to predict whether the patch is from raw image or generated by generator. The key idea is adversarial training a shared vision Transformer to distinguish the input patches from high-quality negative samples, which we believe is beneficial for the understanding of visual conception. It works well while incorporating no much addition parameter. More encouragingly, compared to standard masked image modeling, our GAN-MAE is more compute-efficient, as fewer pre-training epochs result in a better performance on downstream tasks.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022. 1, 2, 7, 8

[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. In *Proc. ICML*, pages 1298–1312. PMLR, 2022. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *Proc. ICLR*, 2021. 1, 2, 7, 8

[4] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Proc. NIPS*, 32, 2019. 3

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 2

[6] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009. 2

[7] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Proc. NIPS*, 19, 2006. 2

[8] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE CVPR*, pages 9650–9660, 2021. 7

[10] Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. *Proc. NIPS*, 31, 2018. 3

[11] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proc. ICML*, pages 1691–1703. PMLR, 2020. 2

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pages 1597–1607. PMLR, 2020. 1

[13] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 3

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proc. IEEE ICCV*, pages 9640–9649, 2021. 1, 7, 8

[15] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE CVPR*, pages 8789–8797, 2018. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, pages 248–255, 2009. 5

[17] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 3

[18] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proc. IEEE CVPR*, pages 12124–12134, 2022. 8

[19] 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, 3

[20] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. *arXiv preprint arXiv:2207.07116*, 2022. 7, 8, 12

[21] 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. In *Proc. ICLR*, 2020. 1, 2

[22] Bo Du, Wei Xiong, Jia Wu, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Stacked convolutional denoising autoencoders for feature representation. *IEEE transactions on cybernetics*, 47(4):1017–1027, 2016. 2

[23] 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. 3

[24] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the _. In *Proc. ICLR*, 2018. 3

[25] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 3

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Proc. NIPS*, 33:21271–21284, 2020. 1

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE CVPR*, pages 16000–16009, 2022. 1, 2, 3, 5, 6, 7, 8, 12

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE CVPR*, pages 9729–9738, 2020. 1

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE ICCV*, pages 2961–2969, 2017. 8

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016. 1

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. IEEE CVPR*, pages 8340–8349, 2021. 8

[33] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 8

[34] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proc. IEEE CVPR*, pages 15262–15271, 2021. 8

[35] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE CVPR*, pages 1125–1134, 2017. 3

[37] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186, 2019. 2

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[39] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Proc. ECCV*, pages 577–593. Springer, 2016. 2

[40] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE CVPR*, pages 4681–4690, 2017. 3

[41] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proc. IEEE CVPR*, pages 7287–7296, 2022. 2

[42] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proc. AAAI*, volume 34, pages 11336–11344, 2020. 3

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[44] 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. 8

[45] Jie Liang, Hui Zeng, and Lei Zhang. High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In *Proc. IEEE CVPR*, pages 9392–9400, 2021. 3

[46] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE CVPR*, pages 2117–2125, 2017. 8

[47] 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 *Proc. ECCV*, pages 740–755. Springer, 2014. 8

[48] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proc. IEEE CVPR*, pages 9371–9381, 2021. 3

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[50] 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 *Proc. IEEE ICCV*, pages 10012–10022, 2021. 2

[51] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[52] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proc. IEEE CVPR*, pages 2794–2802, 2017. 3

[53] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proc. IEEE CVPR*, pages 12042–12051, 2022. 8

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. ECCV*, pages 69–84. Springer, 2016. 2

[55] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE CVPR*, pages 2536–2544, 2016. 2

[56] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[57] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. ICML*, pages 6105–6114. PMLR, 2019. 1

[59] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. ICML*, pages 10347–10357. PMLR, 2021. 7

[60] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. *arXiv preprint arXiv:2203.09795*, 2022. 5, 12

[61] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018. 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS*, pages 5998–6008, 2017. 1

[63] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. ICML*, pages 1096–1103, 2008. 2

[64] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 2

[65] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Proc. NIPS*, 32, 2019. 8

[66] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proc. IEEE CVPR*, pages 14668–14678, 2022. 3

[67] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proc. IEEE CVPR*, pages 2256–2265, 2021. 3

[68] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proc. ECCV*, pages 418–434, 2018. 8

[69] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Proc. NIPS*, 25, 2012. 2

[70] 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. In *Proc. IEEE CVPR*, pages 9653–9663, 2022. 1, 2, 3

[71] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Proc. NIPS*, 32, 2019. 2

[72] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. ECCV*, pages 649–666. Springer, 2016. 2

[73] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *Proc. ICML*, pages 4006–4015. PMLR, 2017. 3

[74] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *Proc. ICLR*, 2017. 3

[75] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proc. IEEE CVPR*, pages 18697–18709, 2022. 3

[76] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE CVPR*, pages 633–641, 2017. 8

[77] 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, 3, 7, 8