IEEE CVPR 2023

THU-PM

Masked Auto-Encoders Meet Generative Adversarial Networks and Beyond

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang*,

Xiaoming Wei, Xiaolin Wei

2023年5月

美团

Short Summary

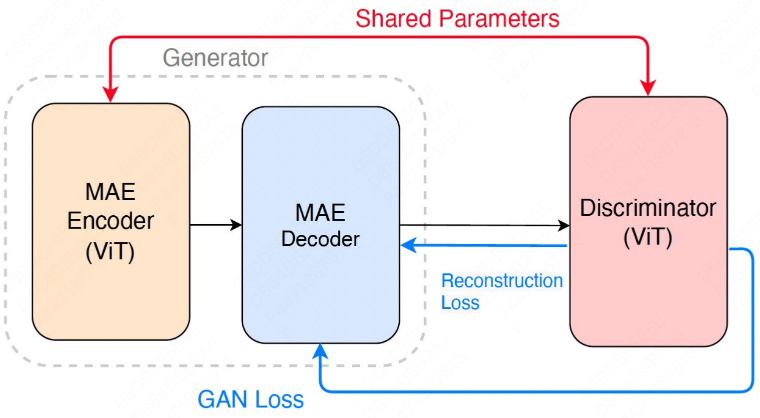


Figure 1. Fast overview of **GAN-MAE** framework.

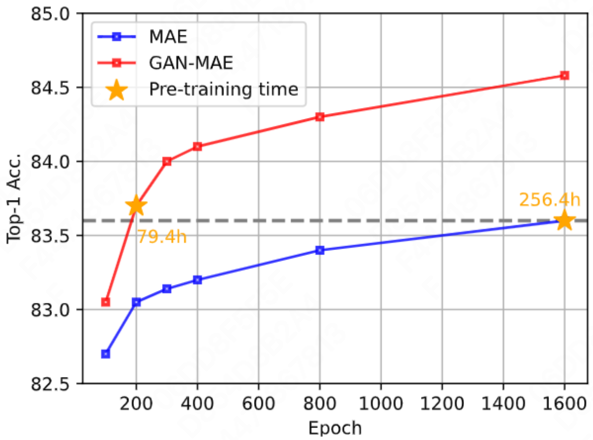
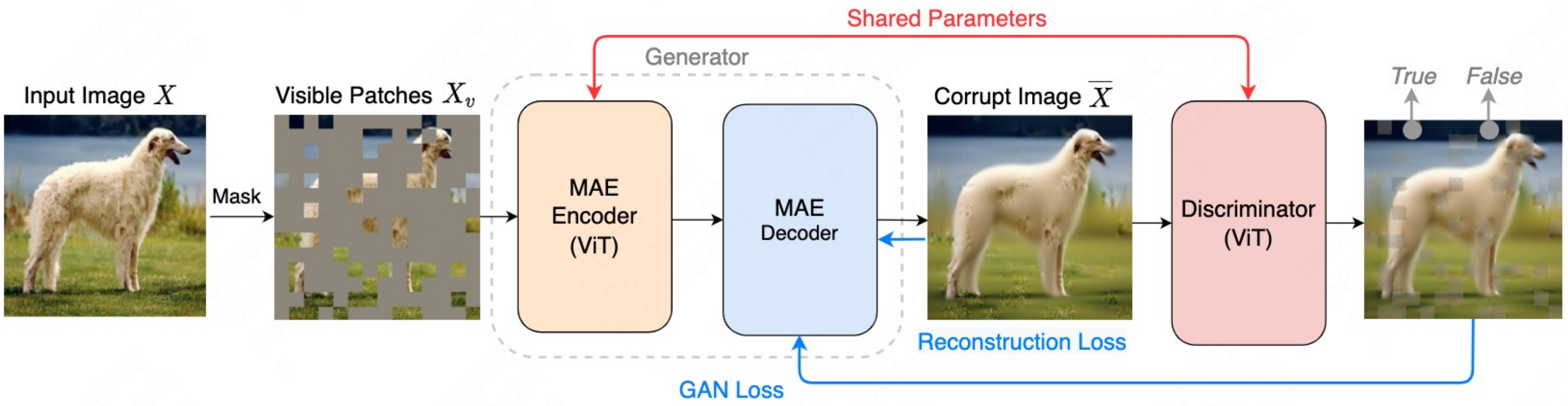


Figure 2. **Performance comparison** in different pre-training epochs for ImageNet-1K Fine-tuning top-1 accuracy.

GAN-MAE Framework



Main components:

1. Image patch generator
2. Image patch discriminator
3. Adversarial training process

*parameter sharing in backbone

GAN-MAE Framework

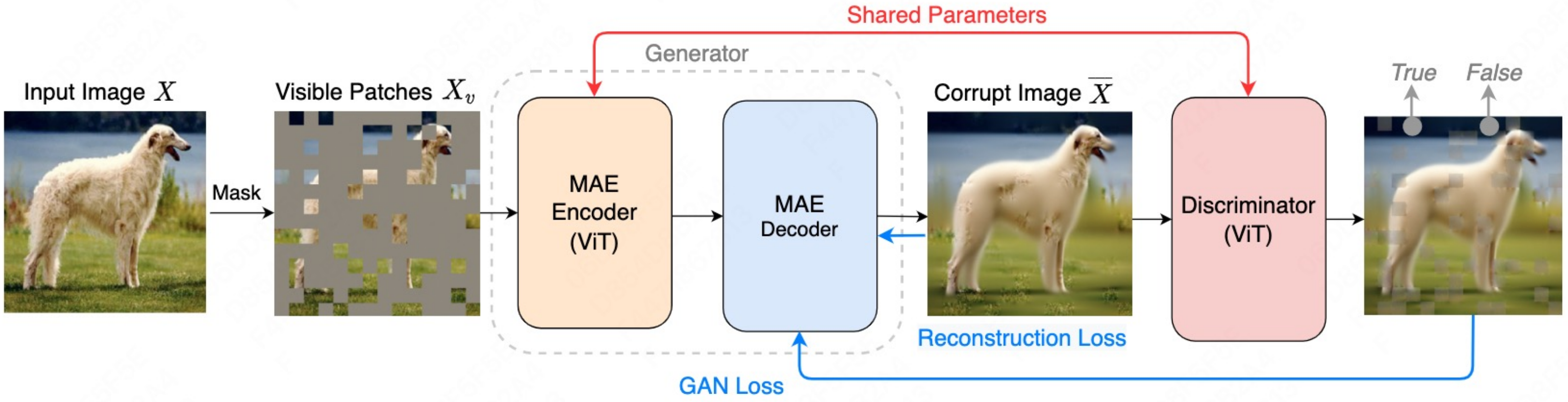


Image Patch generator

- Identical to a standard MAE;
- Overall, the generator randomly masks some image patches M and encodes the remaining visible patches X_v in to hidden states H_v, and the masked patches are then reconstructed as $\tilde{X}_m$:

$$H_v = f_e(X_v, M)$$

$$\tilde{X}_m = \mathrm{f}_d(H_v, M)$$

GAN-MAE Framework

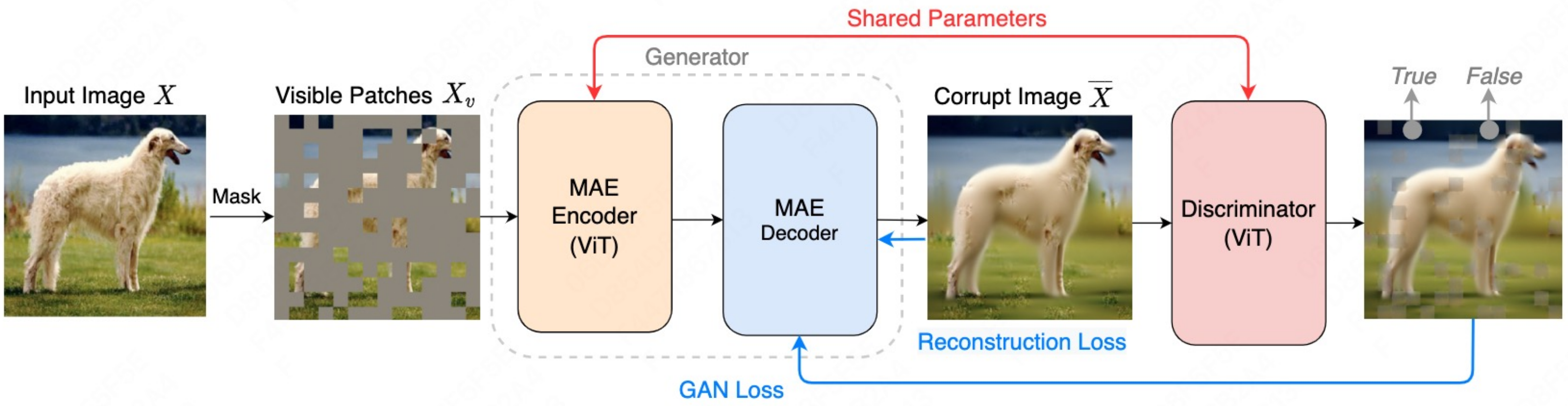


Image Patch Discriminator

- Identical to a ViT for classification;
- For a patch index k and corrupted image sequence $\bar{X} = \{X_v, \tilde{X}_m\}$, the discriminator predicts whether the patch token x^k is real or synthesized as binary classification task:

  $\mathrm{D}(\bar{X}, k) = p_{disc}(\mathrm{y}^k | \bar{X}, k)$

Adversarial Training Process

- At each epoch, iteration is conducted in two steps;
- Train only the generator with L_{gen}:

$$\mathrm{L}_{gen}(X,\theta_{mae}) = L_{mae}(X,\theta_{mae}) + \gamma L_{adv}(X,\theta_{mae})$$

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

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

In between, γ is a adaptive factor

- Train the discriminator with L_{disc}:

$$L_{disc}(\bar{X},\theta_{disc}) = \sum_{k=1}^{N} -y^k logD(\bar{X},k) - (1-y^k)\log(1 - D(\bar{X},k))$$

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

General Comparisons

美团 美团

Table 1. **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**

Case Study

美团

Table 2. **Robustness Evaluation** on the four ImageNet-variants: ImageNet-C, ImageNet-A, ImageNet-R, and ImageNet-Sketch.

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**

Figure 3. Figure 2. **Qualitative analysis for patch reconstruction.**

Model Analysis

Table 3. Effect of **parameter sharing** in GAN-MAE framework. Results demonstrate that shared parameters for backbone benefits both memory cost and performance improvement.

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**

Table 4. 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

Downstream Tasks

美团 美团

Table 3. **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 4. **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**