NEURAL INFORMATION PROCESSING SYSTEMS

北京大学 PEKING UNIVERSITY

# How Mask Matters:
## Towards Theoretical Understandings of Masked Autoencoders

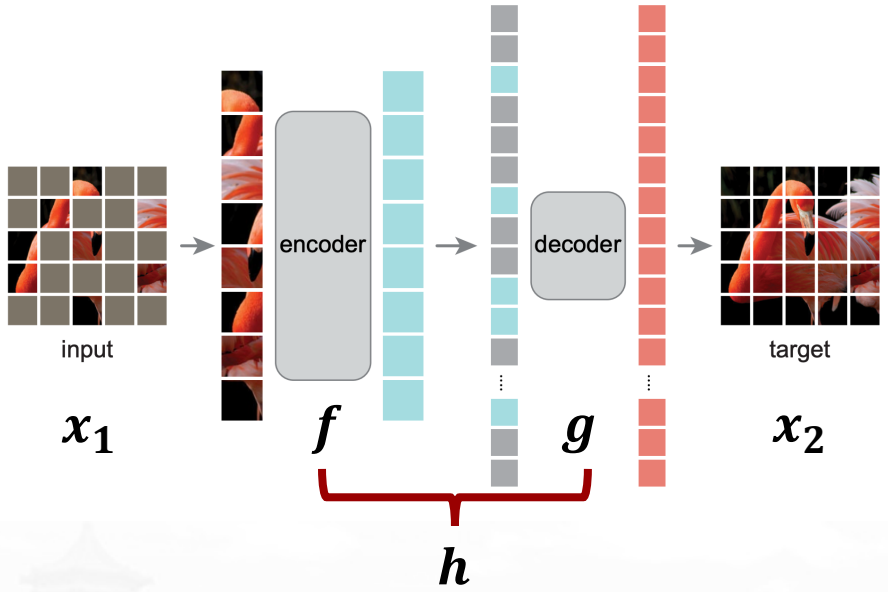
NeurIPS 2022

Presented by Yifei Wang* (PKU)

Joint work with Qi Zhang* (PKU), Yisen Wang (PKU)

# Masked AutoEncoders

- **Development of Self-Supervised Learning (SSL) Paradigms**
  - 2019-2021: Contrastive Learning (SimCLR, MoCo, BYOL, Barlow Twins, …)
  - 2021-now: **Masked Autoencoder (MAE) (He et al., 2021)**

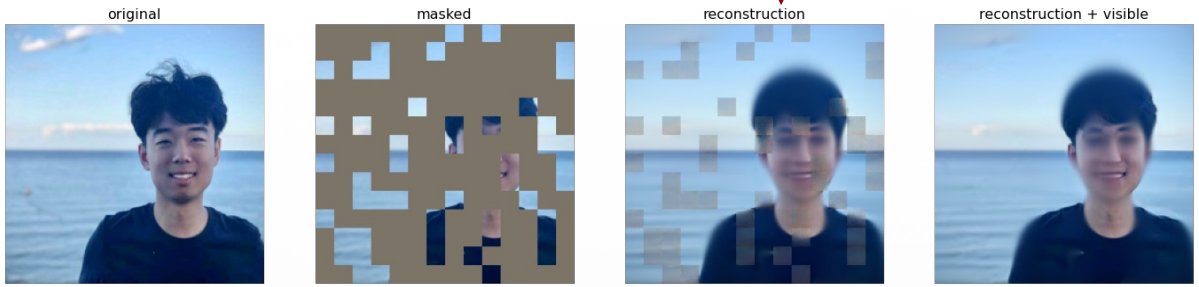



- **MAE Objective (encoder $f$, decoder $g$, whole model $\mathrm{h} = g \circ f$)**

$$\mathcal{L}_{\mathrm{MAE}}(h) = \mathbb{E}_{\bar{x}} \mathbb{E}_{x_1, x_2 | \bar{x}} \left\| g(f(x_1)) - x_2 \right\|^2$$

# Mysteries of MAE

- **MAE = autoencoder + masking**
- **Autoencoder**
  - As a generative model, MAE has poor reconstruction quality from masked inputs



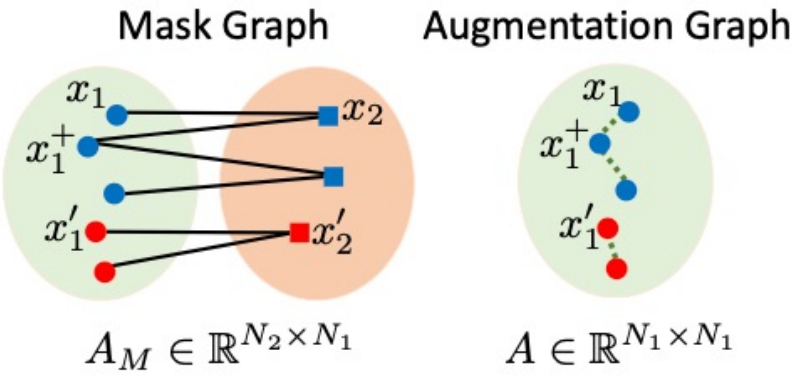

- **Masking**
  - MAE adopts a very large mask ratio (75%)
  - Most image semantics are lost

- **Our perspective**
  - Reconstruction is only a surrogate task for **representation learning**
  - Question: *What is the role of masking? How does it affect downstream performance?*

The original figure is a personal selfie, and the processed images are produced by MAE following https://github.com/facebookresearch/mae

# MAE Implicitly Performs Contrastive Learning

$$\mathcal{L}_{\mathrm{MAE}}(h) = \mathbb{E}_{\bar{x}} \mathbb{E}_{x_1, x_2 | \bar{x}} \left\| g(f(x_1)) - x_2 \right\|^2,$$




***MAE implicitly defines positive input pairs (as in contrastive learning) as 2-hop neighbors in the mask graph***

## Main Theorem: A small MAE loss implies good alignment of positive input pairs

**Theorem 3.4.** *Under Assumption 3.1, MAE's reconstruction loss (Eq. (2)) can be lower bounded by the alignment loss between positive pairs* $(x_1, x_1^+) \sim \mathcal{A}(x_1, x_1^+)$,

$$\mathcal{L}_{MAE}(h) \geq \frac{1}{2}\mathcal{L}_{align}(h) - \varepsilon + const. \tag{7}$$

$$\mathcal{L}_{\mathrm{align}}(h) = -\mathbb{E}_{x_1, x_1^+} h(x_1)^\top h(x_1^+).$$

# Feature Collapse in MAE: Implicit Regularizations and Limitations

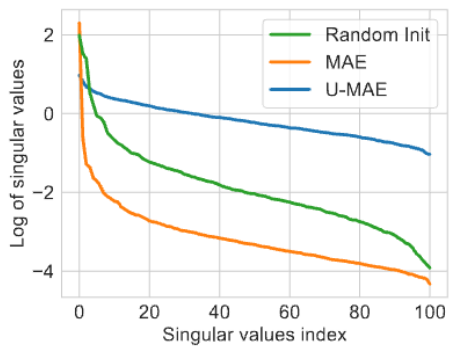
- **MAE can avoid full feature collapse**
  - Fully collapsed encoder suffers from a large MAE loss

**Theorem 3.6.** *When the encoder fully collapses, i.e.,* $\forall x \in \mathcal{X}_1, f(x) = c$*, the MAE loss has a large lower bound:*

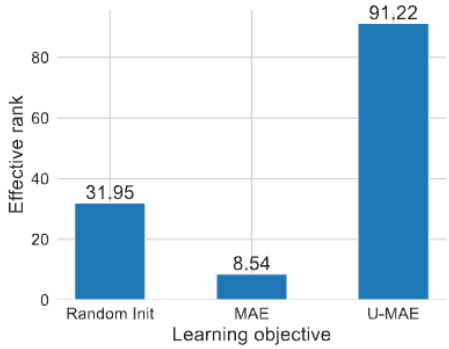
$$\mathcal{L}_{MAE}(h) \geq \mathrm{Var}(x_2), \tag{9}$$

*where* $\mathrm{Var}(x_2)$ *denotes the variance of masked targets computed on the training dataset.*

- **MAE still suffers from Dimensional Collapse**




(b) comparison of singular values




(c) comparison of effective rank

- **Uniformity-enhanced MAE (U-MAE): promote feature diversity with an explicit uniformity loss**
  - minimizes the similarities of randomly drawn **negative input pairs** $(x_1, x_1^-)$

$$\mathcal{L}_{\text{U-MAE}}(h) = \mathcal{L}_{\text{MAE}}(h) + \lambda \cdot \mathcal{L}_{\text{unif}}(f),$$
$$\text{where } \mathcal{L}_{\text{unif}}(f) = \mathbb{E}_{x_1} \mathbb{E}_{x_1^-} (f(x_1)^\top f(x_1^-))^2,$$

# Downstream Generalization of Masked Autoencoders

北京大学 PEKING UNIVERSITY

- **Based on this connection, we provide theoretical guarantees on downstream performance**

**Theorem 4.1.** *Denote the mask-induced label error as* $\alpha = \mathbb{E}_{\bar{x},x_1}\mathbb{1}[y(x_1) \neq y(\bar{x})]$. *Then, for* $\forall\, h \in \mathcal{H}$ *(the hypothesis class) with* $h = g \circ f$, *the downstream classification error of its encoder can be upper bounded by its U-MAE pretraining loss:*

$$\Pr(\bar{y} \neq p_f(\bar{x})) \leq c_1 L \cdot \mathcal{L}_{U\text{-}MAE}(h) + c_2\alpha + c_3 L\varepsilon + c_4, \tag{14}$$

*where* $c_1, \ldots, c_4$ *are constants and* $c_3 > 1$.

Vanilla autoencoding error (no mask)

Label error

- **The minimal U-MAE loss is determined by the connectivity of the augmentation graph**

**Theorem 4.2.** *The U-MAE pretraining loss has the following common lower bound:*

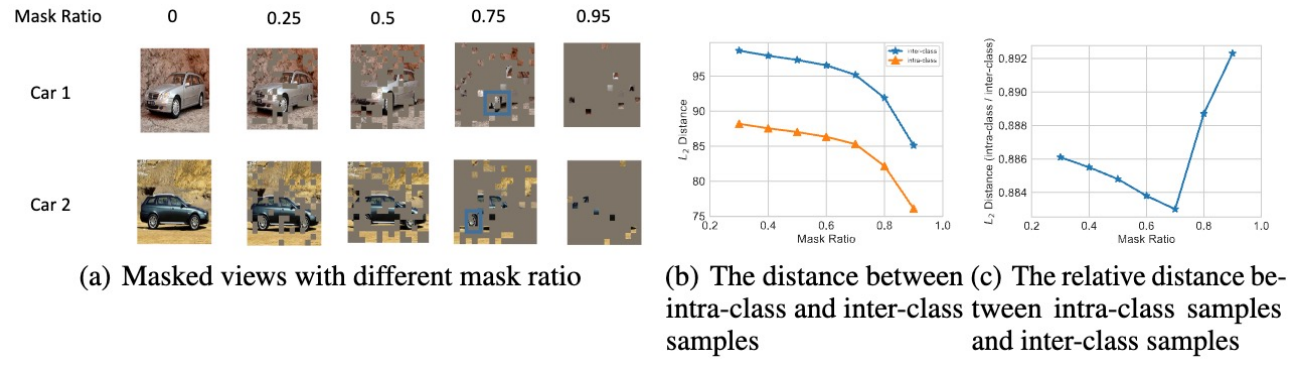
$$\forall\, h \in \mathcal{H}, \quad \mathcal{L}_{U\text{-}MAE}(h) \geq \frac{1}{4L}\sum_{i=k+1}^{N_1}\lambda_i^2 - \varepsilon + const, \tag{15}$$

*where* $\lambda_1 \geq \cdots \geq \lambda_{N_1}$ *denote the eigenvalues of* $\bar{A}$.

Graph connectivity

According to the theory, we need

- **Powerful backbone**
  - Capable of vanilla autoencoding
- **A large mask ratio**
  - Increase intra-class edges
- **Not too large mask ratio**
  - Fewer inter-class edges




(a) Masked views with different mask ratio

(b) The distance between intra-class and inter-class samples

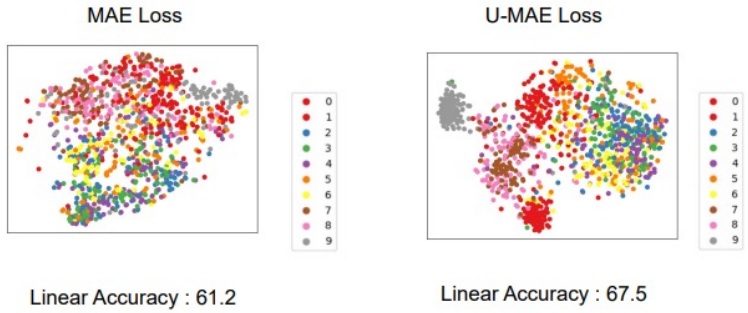
(c) The relative distance between intra-class samples and inter-class samples

***Empirical verification agrees with MAE's choice of mask ratio***

# Experiments

- **U-MAE improves MAE a lot on the linear probing task**
  - 9% ↑ on CIFAR-10, 8% ↑ on ImageNet-100, 3% ↑ on ImageNet-1K
  - no degradation on the fully finetuning task

| Downstream Task | Method | CIFAR-10 | | ImageNet-100 | | ImageNet-1K | |
|---|---|---|---|---|---|---|---|
| | | ViT-Tiny | ViT-Base | ViT-Base | ViT-Large | ViT-Base | ViT-Large |
| Linear Probing | MAE | 59.6 | 61.7 | 61.2 | 64.4 | 55.4 | 62.2 |
| | U-MAE | **68.9** | **70.2** | **67.5** | **72.8** | **58.5** | **65.8** |
| Fine-tuning | MAE | **89.6** | 90.7 | **86.9** | **87.3** | 82.9 | **83.3** |
| | U-MAE | 89.4 | **90.8** | **86.8** | **87.3** | **83.0** | 83.2 |




- **Also effective on other MIM methods, such as SimMIM, named U-SimMIM**
  - 7% ↑ on ImageNet-100

Table 2: Linear probing accuracy (%) of U-SimMIM (ViT-Base) on ImageNet-100.

| SimMIM | U-SimMIM |
|---|---|
| 54.3 | **61.1** |

# Take Home Messages

北京大学 PEKING UNIVERSITY

- **MAE ≈ contrastive learning**
  - masking also induces positive pairs!
- **MAE still suffers from dimensional collapse**
  - can be resolved by U-MAE with uniformity regularization!
- **Theoretical guarantees on downstream performance**
  - which explains the choice of large mask ratio
- **Tips for designing masks**
  - increase intra-class edges (requires a large mask ratio)
  - avoid inter-class edges (not too large to distort belonging classes)


# Thanks for Listening!

Yifei Wang (Peking University)

Contact yifei_wang@pku.edu.cn for further questions