Implicit Maximum Likelihood Estimation Improves Masked Autoencoders

Tristan Engst · **Alireza Moazeni** · **Ke Li**

Simon Fraser University · Burnaby, BC | tme3@sfu.ca

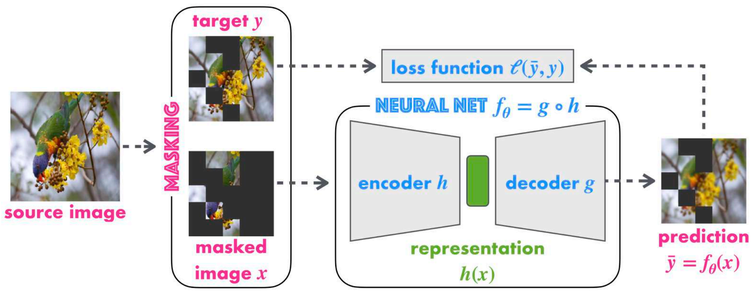
AI/CRV 2026

Abstract

Standard MAEs use **least-squares** pretraining, which limits the masking schemes they can learn from. We propose **conditional implicit maximum likelihood estimation (cIMLE)** pretraining as a superior alternative.

- Even *perfect* least-squares pretraining can fail to learn effective representations—even when they are in principle learnable
- cIMLE-pretrained representations *surpass the state-of-the-art*

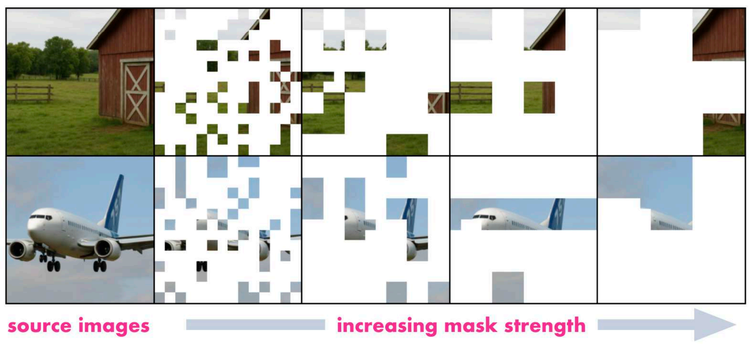
Background



- Masked autoencoding is a self-supervised learning paradigm where pretraining seeks to reconstruct masked visual content
- This incentivizes the encoder to learn representations that signal to the decoder how it should reconstruct what was masked

Least-squares pretraining: $$\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim X\times Y} \|f_\theta(x) - y\|^2$$

Masking Strength



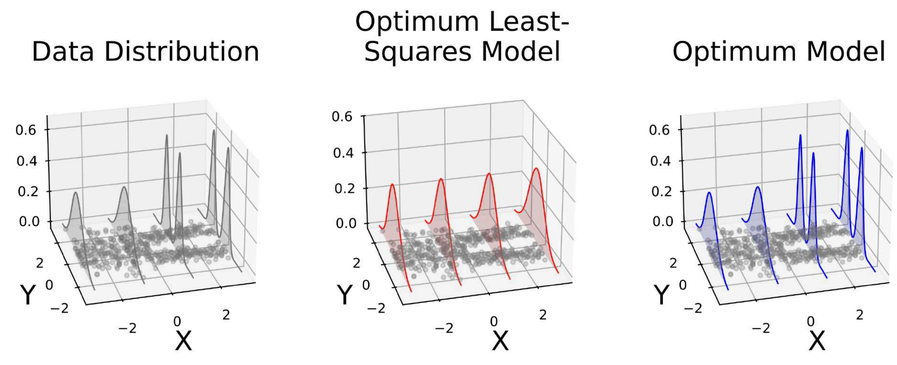
- MAEs benefit from **stronger masking**—but only up to a point!
- Stronger masking schemes increase the **diversity of plausible reconstructions**, but this progressively *collapses* differences between masked images' *expected* reconstructions

Limitations of Least-Squares Pretraining

Least-squares loss is minimized by **regression to the mean**:

$$f_{\theta^\star}(x) = \mathop{\mathbb{E}}_{y\sim p_Y(\cdot|x)}[y]$$

- **Key insight:** representations only need to capture the **expectations** of masked images' distributions of reconstructions—*even if **other properties** of these distributions provide the structure to make representations meaningful!*



Left: Data whose conditionals $p_Y(\cdot|x)$ differ in multimodality but not expectation. **Center:** The best predictive distribution for least-squares pretraining is constant. **Right:** Correctly fitting the data requires learning both distinct conditionals—and therefore representations that distinguish inputs by them.

***Perfect* least-squares models can learn *arbitrary* representations**

Conditional Implicit Maximum Likelihood Estimation (cIMLE)

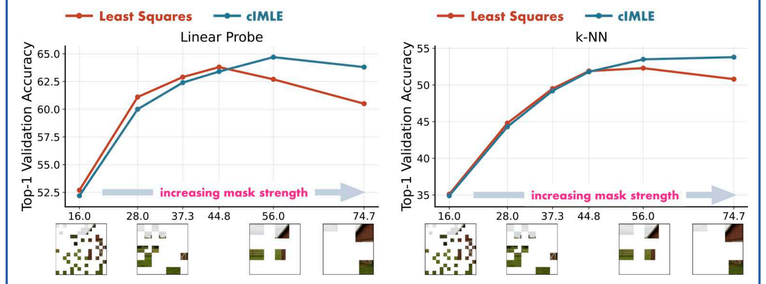
We conduct MAE pretraining, but *swap least squares for **cIMLE***, an objective that seeks to *match the model's predictive distribution to that of the data:*

cIMLE pretraining: $$\min_\theta \mathop{\mathbb{E}}_{(x,y)\sim X\times Y}\left[\mathop{\mathbb{E}}_{z_1\ldots z_k\sim Z}\min_{i\in[k]}\|f_\theta(x, z_i) - y\|^2\right]$$

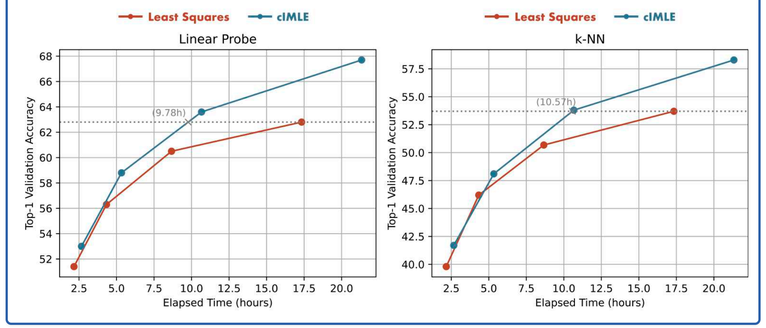
- cIMLE helps *learning from stronger masking schemes*, which can capture differences between masked images' with properties of their distributions of reconstructions *beyond their means*—eg. their **covariances**, **higher moments**, or **multimodalities**

Perfect cIMLE pretraining learns representations that distinguish inputs by *all* properties of their distributions of reconstructions

cIMLE vs. Least Squares across Mask Strength



Pretraining Efficiency under Strong Masking



Results

Method	Linear Probe	k-NN	Finetuning
ViT-T/16 · ImageNet-100 · 800 epochs pretraining			
MAE	52.3	36.8	87.0
ColorMAE	57.8	40.5	**87.6**
Ours	**64.7** (+12.4 vs. MAE)	**53.7** (+16.9 vs. MAE)	87.5
ViT-B/16 · ImageNet-1K · 800 epochs pretraining			
MAE	64.3	30.3	83.2
ColorMAE	68.3	50.6	83.6
PrototypicalMAE	68.9	47.4	83.6
MI-MAE	69.3	—	84.1
GAN-MAE	69.3	—	**84.3**
Ours	**70.9** (+6.6 vs. MAE)	**61.8** (+31.5 vs. MAE)	83.6

Method	Pets	DTD	Flowers
Transfer learning — k-NN · ViT-B/16 · IN-1K · 800 epochs pretraining			
MAE	44.6	48.4	58.8
Ours	**73.0** (+28.4 vs. MAE)	**54.3** (+5.9 vs. MAE)	**70.5** (+11.7 vs. MAE)
Transfer learning — Finetuning · ViT-B/16 · IN-1K · 800 epochs pretraining			
MAE	92.9	71.3	95.2
Ours	**93.6** (+0.7 vs. MAE)	**71.8** (+0.5 vs. MAE)	**95.8** (+0.6 vs. MAE)