Deconstructing Denoising Diffusion Models for Self-Supervised Learning

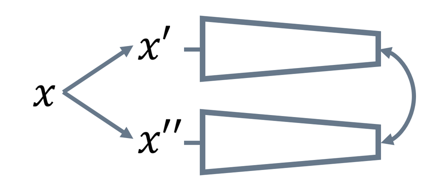
Xinlei Chen, Zhuang Liu, Saining Xie, Kaiming He

ICLR 2025

Visual Self-Supervised Learning (SSL) Paradigms

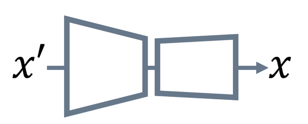
SSL: Pre-train representations from images without human labels

Joint-encoder



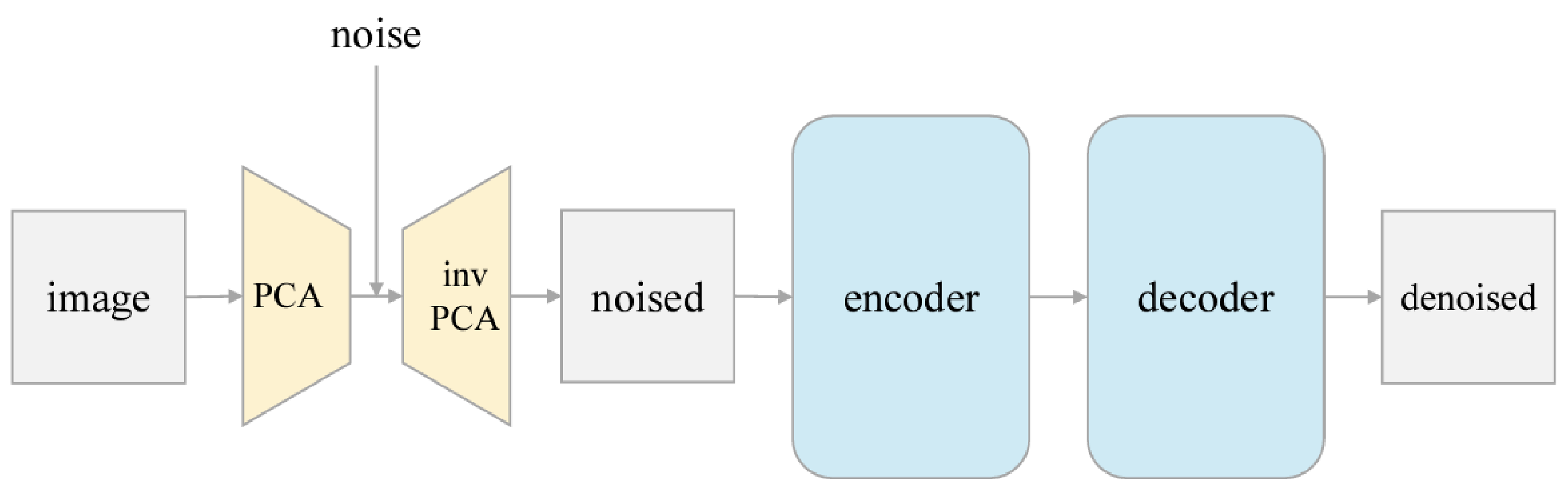
- SimCLR, MoCo
- BYOL, SimSiam
- DINO, JEPA

Auto-encoder



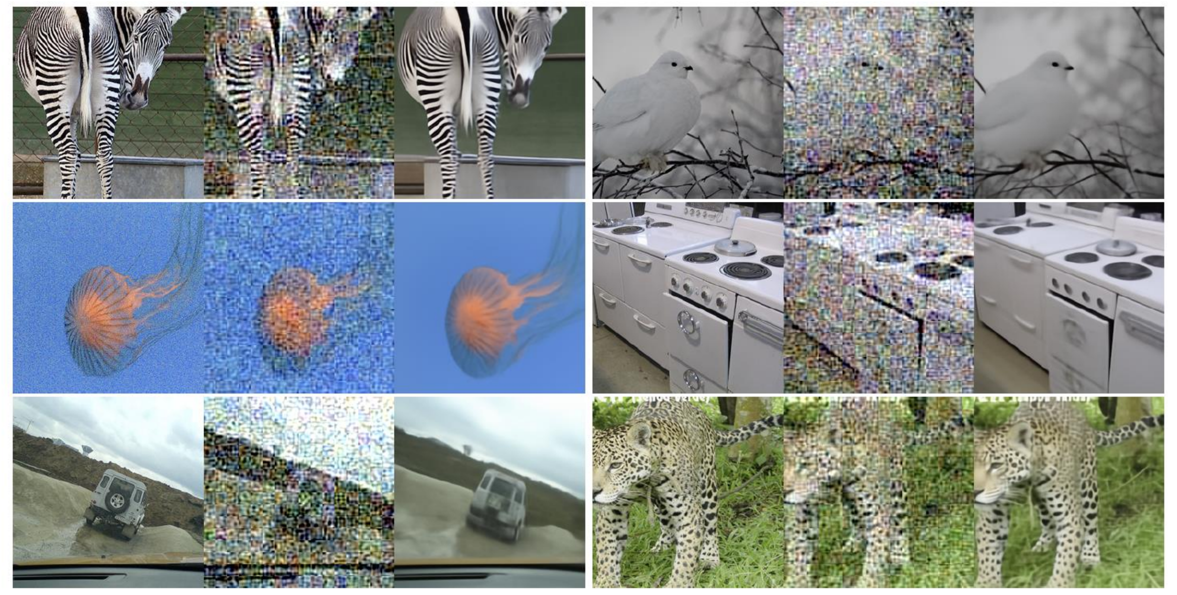
- MAE
- l-DAE (this work)

l-DAE Architecture



Qualitative Visualizations

ImageNet *validation* set



Quantitative Comparisons

ImageNet *validation* set

Linear probing

pre-train	ViT-B	ViT-L
MoCo v3	**76.7**	**77.6**
MAE	68.0	75.8
l-DAE	66.6	75.0

Fine-tuning

pre-train	ViT-B	ViT-L
MoCo v3	83.2	84.1
MAE	83.6	**85.9**
l-DAE	**83.7**	84.7

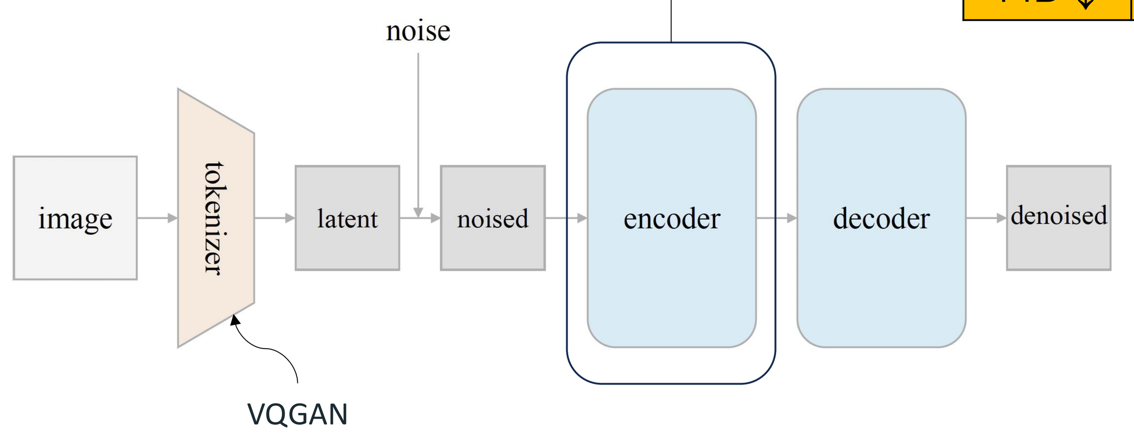
- l-DAE performs decently against MAE
- Joint-encoder methods are generally better for for linear probing
- Auto-encoders are generally better with fine-tuning

the Deconstruction

Init: Diffusion Transformer (L)

linear probing

Acc ↑	57.5
FID ↓	11.6



Re-oriente DiT for SSL

1. Remove class-conditioning
   - Otherwise not *legitimate* for SSL

Acc ↑	57.5 → 62.5
FID ↓	11.6 → 30.9

2. Remove LPIPS loss in VQGAN
   - LPIPS is based on VGG trained on labels

Acc ↑	62.5 → 58.4
FID ↓	30.9 → 54.3

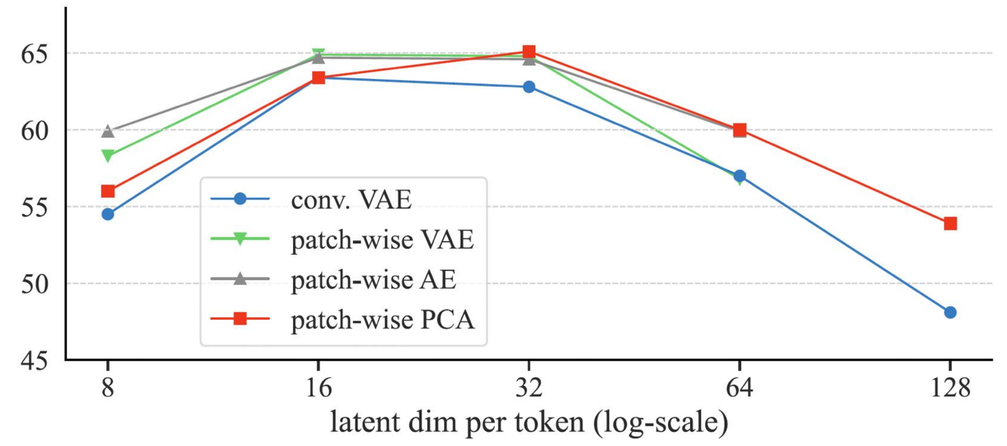
3. Remove GAN loss in VQGAN
   - GAN hurts understanding

Acc ↑	58.4 → 59.0
FID ↓	54.3 → 75.6

4. Noise schedule change
   - High noise levels hurt understanding

Acc ↑	59.0 → 63.4
FID ↓	75.6 → 93.2

Deconstruct the tokenizer

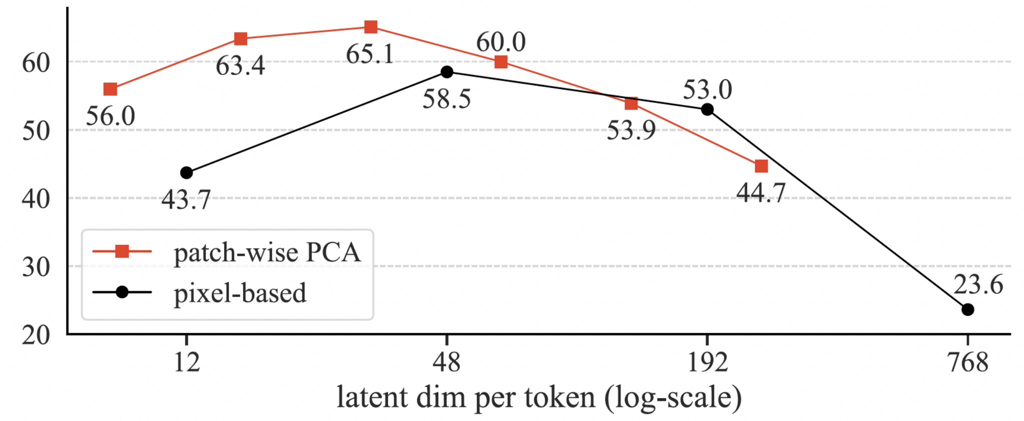


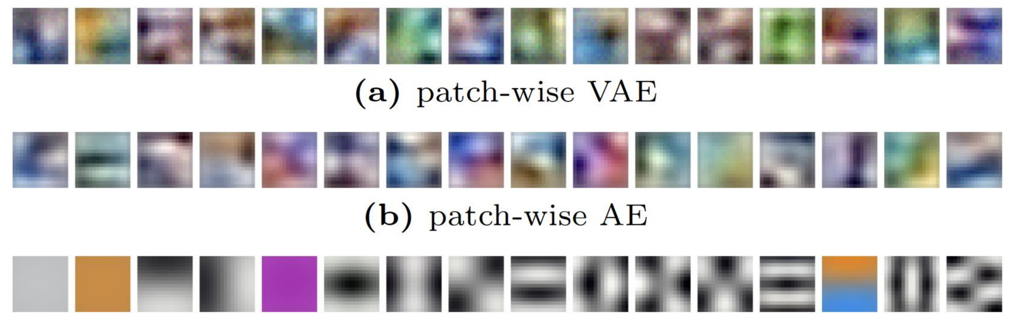
Main finding of the paper

latent dim	8	16	32	64
conv. VAE	54.5	**63.4**	62.8	57.0
patch-wise VAE	58.3	**64.9**	64.8	56.8
patch-wise AE	59.9	**64.7**	64.6	59.9
patch-wise PCA	56.0	63.4	**65.1**	60.0

- Latent dimension of the tokenizer is crucial
- Specific variants of the tokenizer matter much less
- 16 or 32 are about optimal for 16x16 patches

1. Convolutional VAE: $\|x - g(f(x))\|^2 + \mathbb{KL}[f(x)|\mathcal{N}]$
2. Patch-wise VAE: $\|x - U^T Vx\|^2 + \mathbb{KL}[Vx|\mathcal{N}]$
3. Patch-wise AE: $\|x - U^T Vx\|^2$
4. Patch-wise PCA: $\|x - V^T Vx\|^2$ → via eigen decomposition

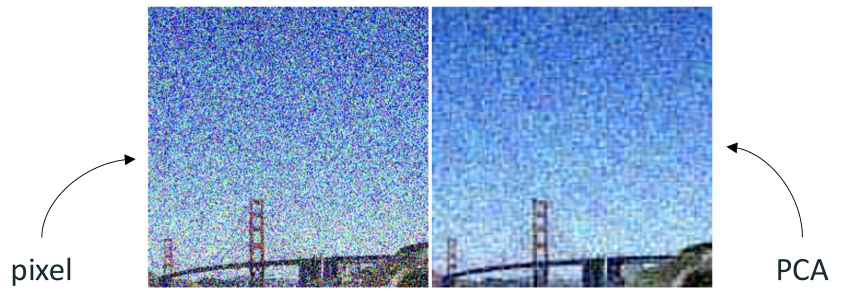




(a) patch-wise VAE

(b) patch-wise AE

(c) patch-wise PCA



- PCA has patch-wise noise, akin-to MAE that drops patches
- high-resolution, pixel-based diffusion are not great for SSL

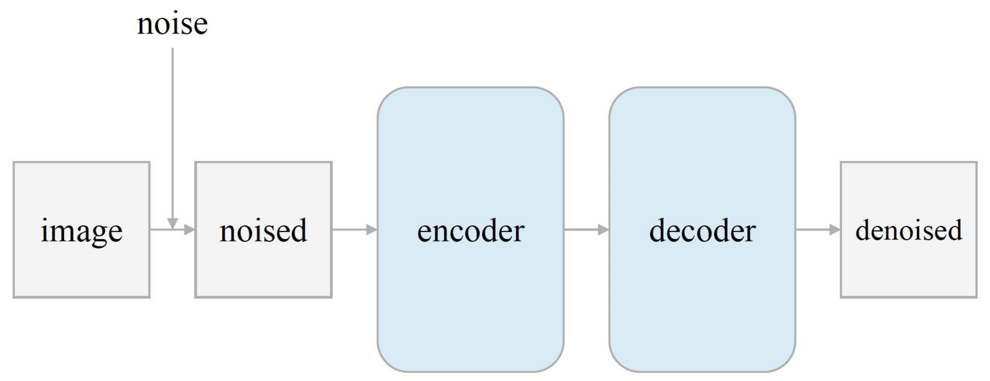
Move toward classical DAE

1. Signal vs. noise simplifications

	input	output	Acc
• DiT default	$\gamma z_0 + \sqrt{1-\gamma^2}\epsilon$	ϵ	65.1
• Predict signal, not noise	$\gamma z_0 + \sqrt{1-\gamma^2}\epsilon$	z_0	62.4
• Remove signal scaling	$z_0 + \sigma\epsilon$	z_0	63.6

2. Operate on pixels

	input	output	Acc
• Inv PCA as input	inv PCA	PCA	63.6
• Inv PCA also as output	inv PCA	inv PCA	63.9
• Original image as output	inv PCA	original	64.5



Analysis

- Time steps
  - key for diffusion, but not so important

	multiple	single
Acc ↑	64.5	61.5

- Data augmentation
  - Helps with more augmentation

	center-crop	multi-crop
Acc ↑	64.5	65.0

- Data scaling
  - Helps with more longer epochs

	400	800	1600
Acc ↑	65.0	67.5	69.6

https://arxiv.org/abs/2401.14404