北京大学
PEKING UNIVERSITY

Learning by Reconstruction Produces Uninformative Features For Perception

Randall Balestriero Yann LeCun

PRESENTER: LILANG LIN

2024/03/17

Outline

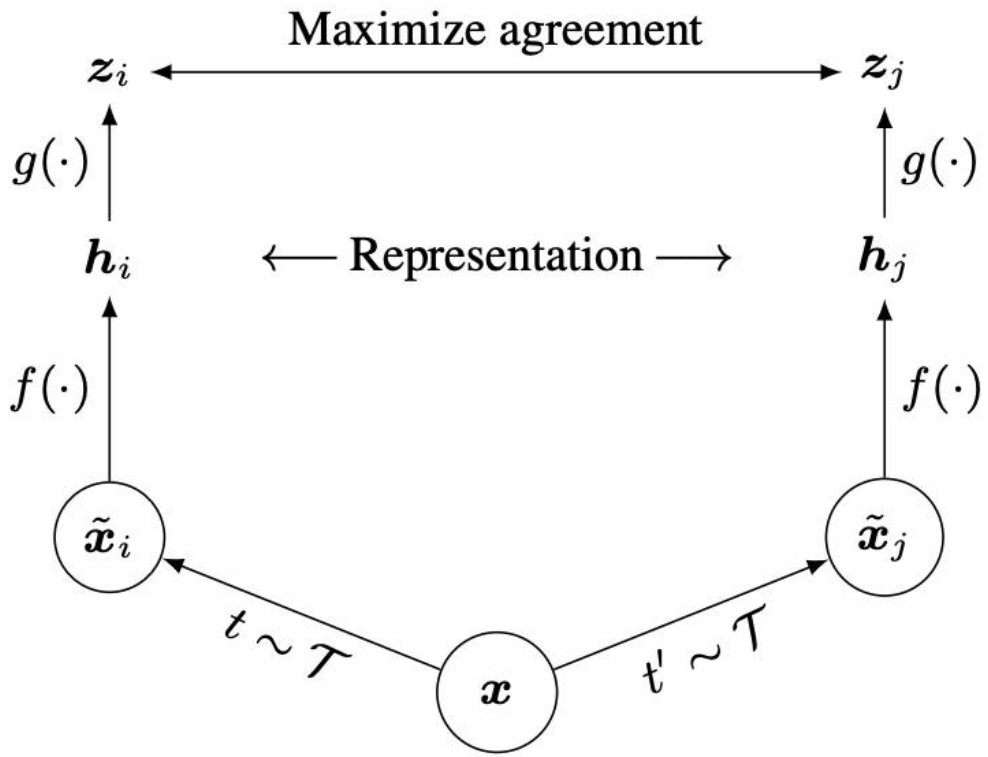
Outline

Background

- **SimCLR (ICML 2020)**

A Simple Framework for Contrastive Learning of Visual Representations

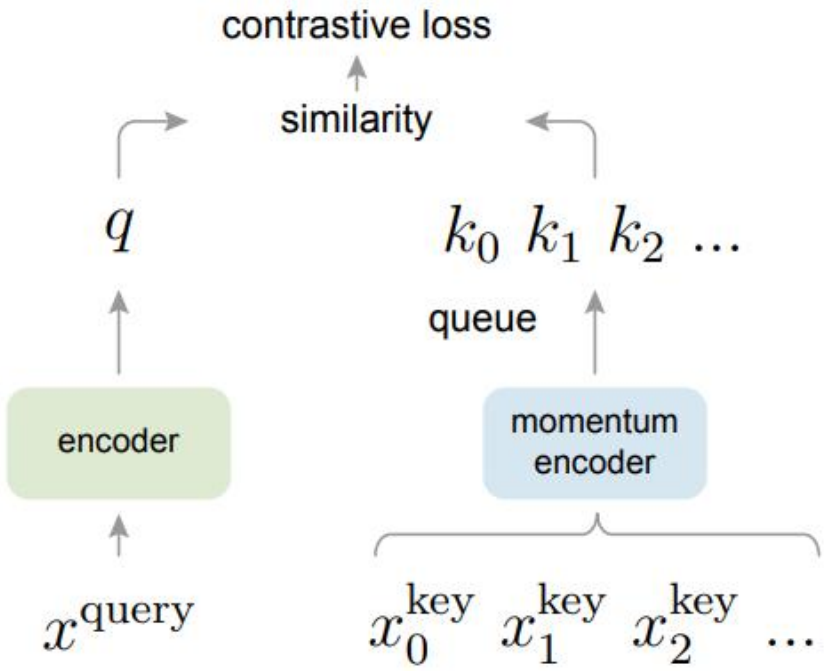
Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]



$$\mathcal{L}_{infoNCE} \triangleq \underbrace{-\sum_b \frac{\langle z_b^A, z_b^B \rangle_i}{\tau \left\| z_b^A \right\|_2 \left\| z_b^B \right\|_2}}_{\text{similarity term}} \underbrace{+\sum_b \log \left(\sum_{b' \neq b} \exp \left(\frac{\langle z_b^A, z_{b'}^B \rangle_i}{\tau \left\| z_b^A \right\|_2 \left\| z_{b'}^B \right\|_2} \right) \right)}_{\text{contrastive term}}$$

Background

- **MoCo (CVPR 2020)**



Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

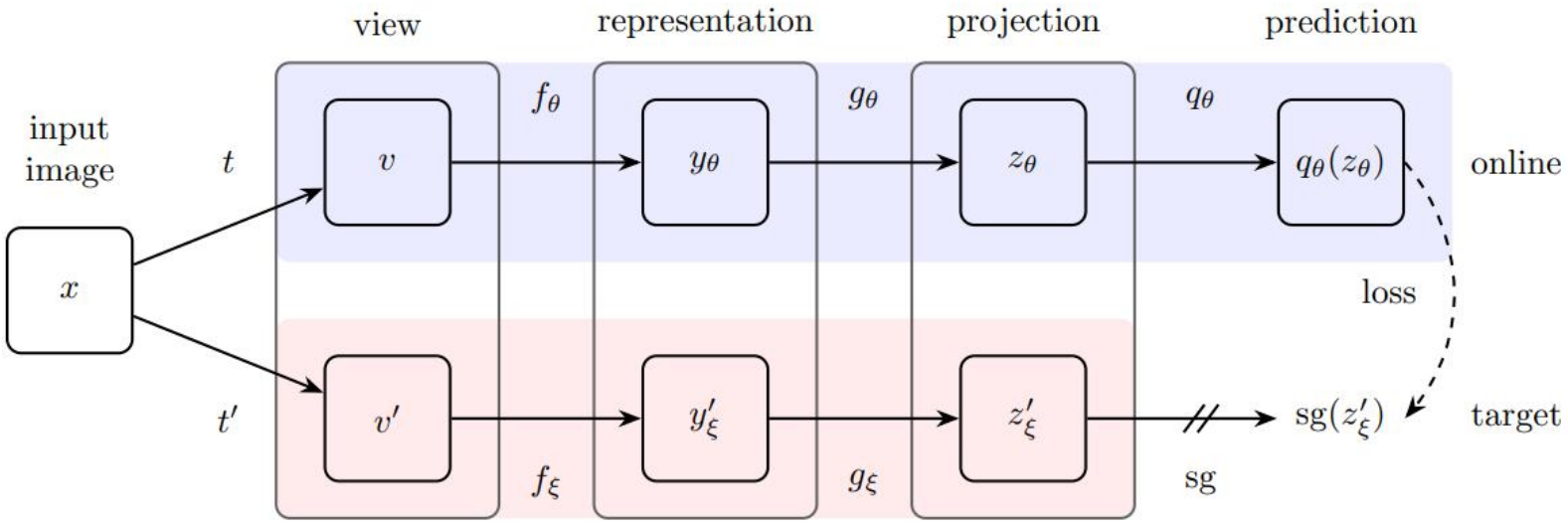
$$\mathcal{L}_{infoNCE} \triangleq \underbrace{-\sum_b \frac{\langle z_b^A, z_b^B \rangle_i}{\tau \left\| z_b^A \right\|_2 \left\| z_b^B \right\|_2}}_{\text{similarity term}} + \underbrace{\sum_b \log \left(\sum_{b' \neq b} \exp \left(\frac{\langle z_b^A, z_{b'}^B \rangle_i}{\tau \left\| z_b^A \right\|_2 \left\| z_{b'}^B \right\|_2} \right) \right)}_{\text{contrastive term}}$$

Background

- **BYOL (NIPS 2020)**

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill[*,1], Florian Strub[*,1], Florent Altché[*,1], Corentin Tallec[*,1], Pierre H. Richemond[*,1,2]
Elena Buchatskaya[1], Carl Doersch[1], Bernardo Avila Pires[1], Zhaohan Daniel Guo[1]
Mohammad Gheshlaghi Azar[1], Bilal Piot[1], Koray Kavukcuoglu[1], Rémi Munos[1], Michal Valko[1]
[1]DeepMind [2]Imperial College
[jbgrill,fstrub,altche,corentint,richemond]@google.com



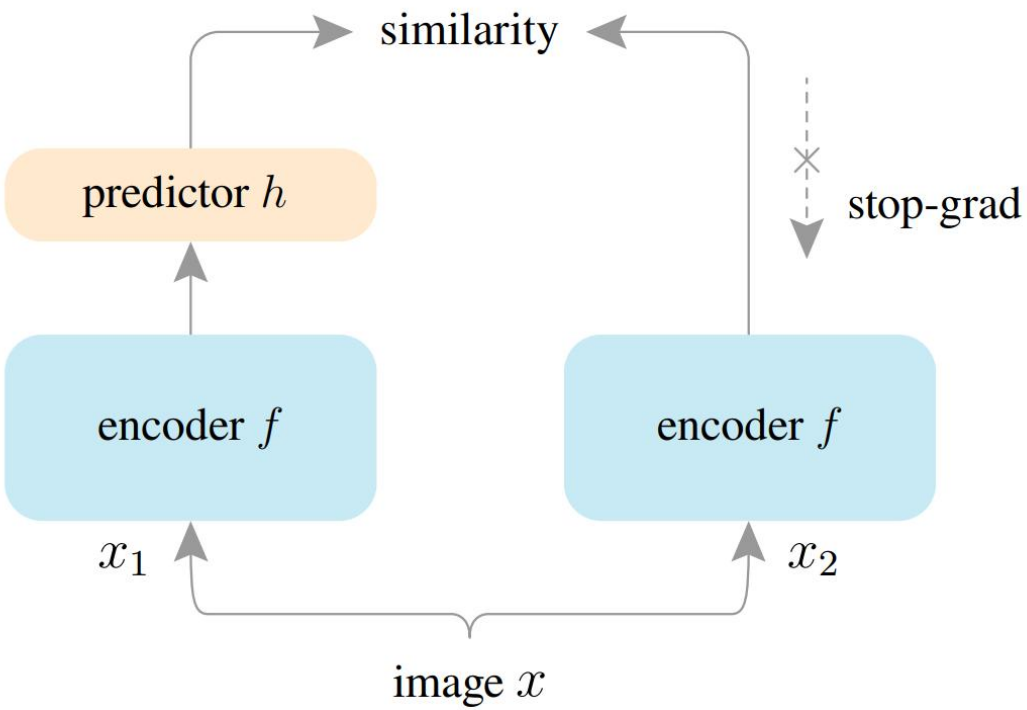
$$\mathcal{L}_{cosine} = -\sum_b \frac{\langle z_b^A, z_b^B \rangle_i}{\left\| z_b^A \right\|_2 \left\| z_b^B \right\|_2}$$

Background

- **SimSiam (CVPR 2021)**

Exploring Simple Siamese Representation Learning

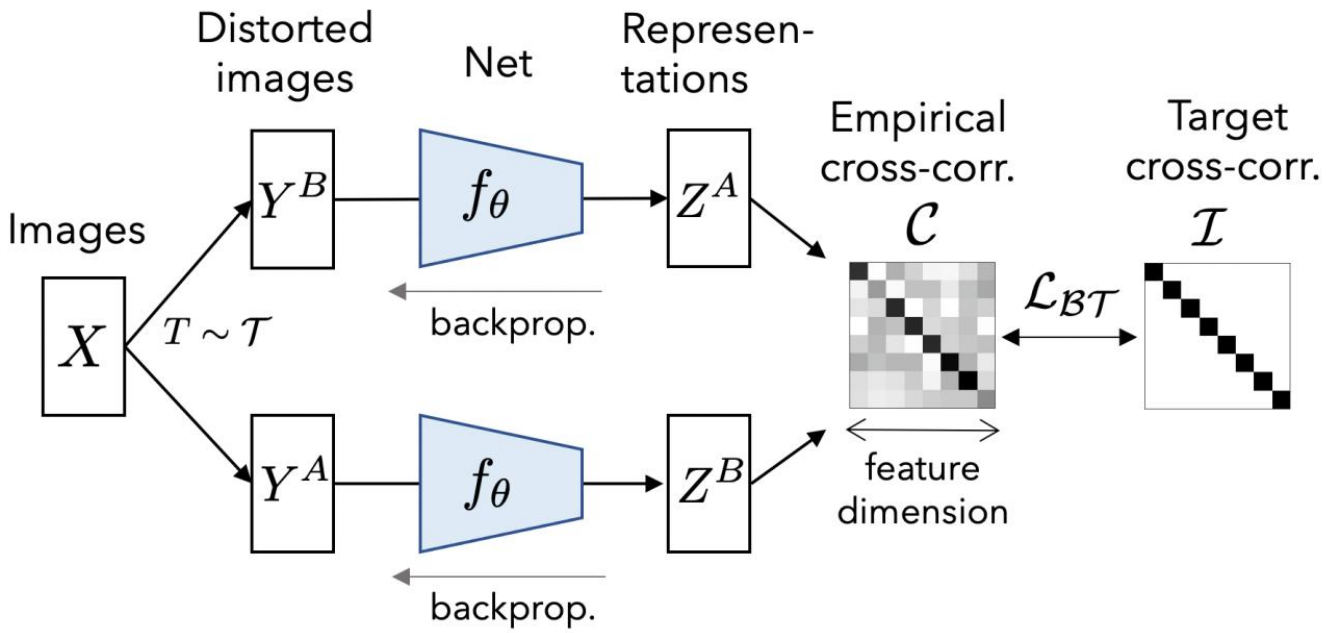
Xinlei Chen Kaiming He
Facebook AI Research (FAIR)



$$\mathcal{L}_{cosine} = -\sum_b \frac{\langle z_b^A, z_b^B \rangle_i}{\|z_b^A\|_2 \|z_b^B\|_2}$$

Background

- **Barlow Twins (ICML 2021)**



$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}}$$

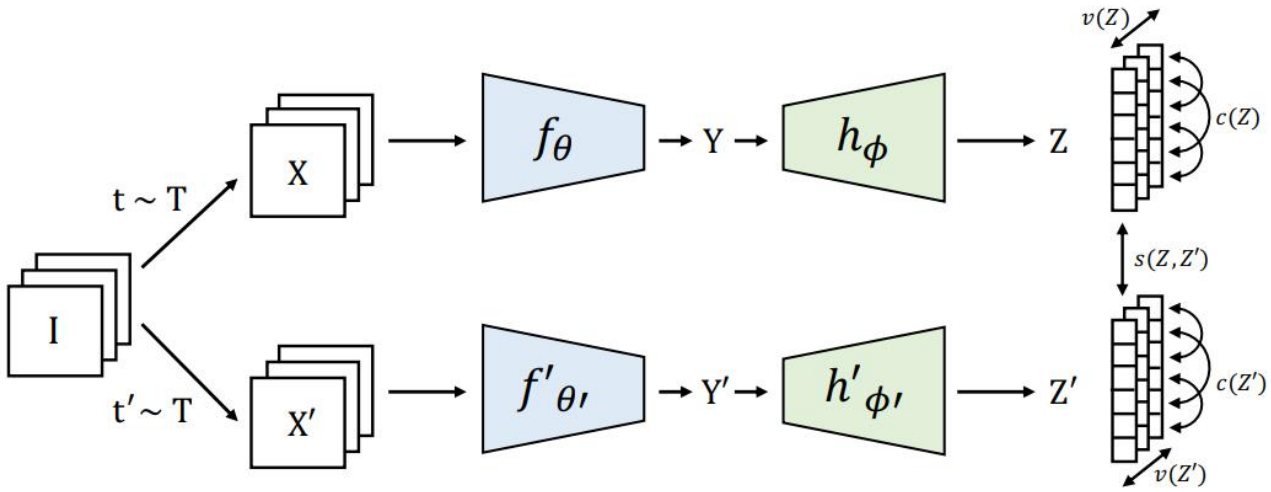
$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2} \sqrt{\sum_b (z^B_{b,j})^2}}$$

Background

- **VICReg (ICLR 2022)**

VICREG: VARIANCE-INVARIANCE-COVARIANCE REGULARIZATION FOR SELF-SUPERVISED LEARNING

Adrien Bardes[1,2] **Jean Ponce**[2,4] **Yann LeCun**[1,3,4]
[1]Facebook AI Research
[2]Inria, École normale supérieure, CNRS, PSL Research University
[3]Courant Institute, New York University
[4]Center for Data Science, New York University



$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - S(z^j, \epsilon)),$$

$$S(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon},$$

$$C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T, \quad \text{where} \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i.$$

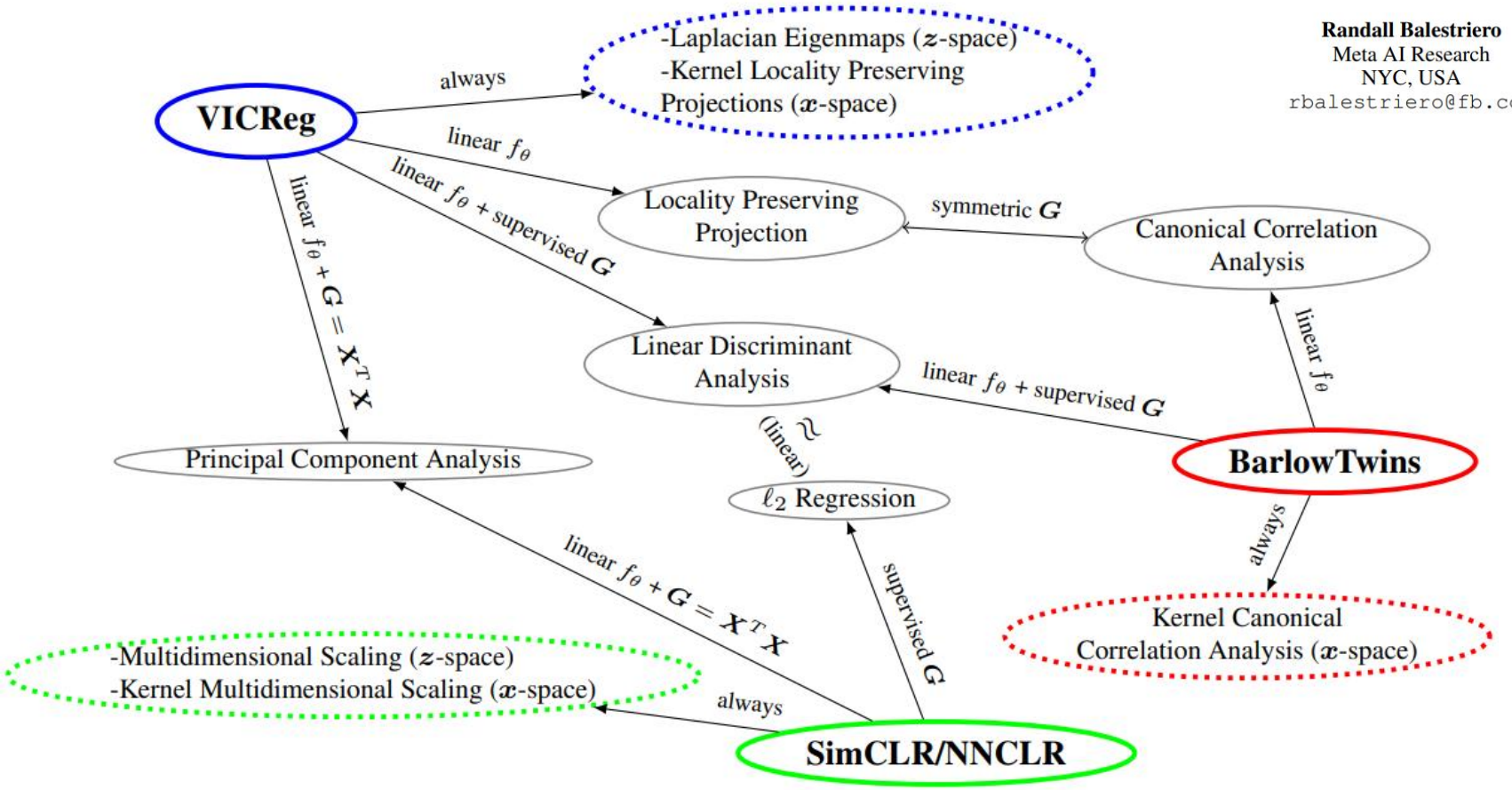
$$s(Z, Z') = \frac{1}{n}\sum_{i} \|z_i - z'_i\|_2^2.$$

Background

Contrastive and Non-Contrastive Self-Supervised Learning Recover Global and Local Spectral Embedding Methods

Randall Balestriero
Meta AI Research
NYC, USA
rbalestriero@fb.com

Yann LeCun
Meta AI Research & NYU
NYC, USA
ylecun@fb.com



Background

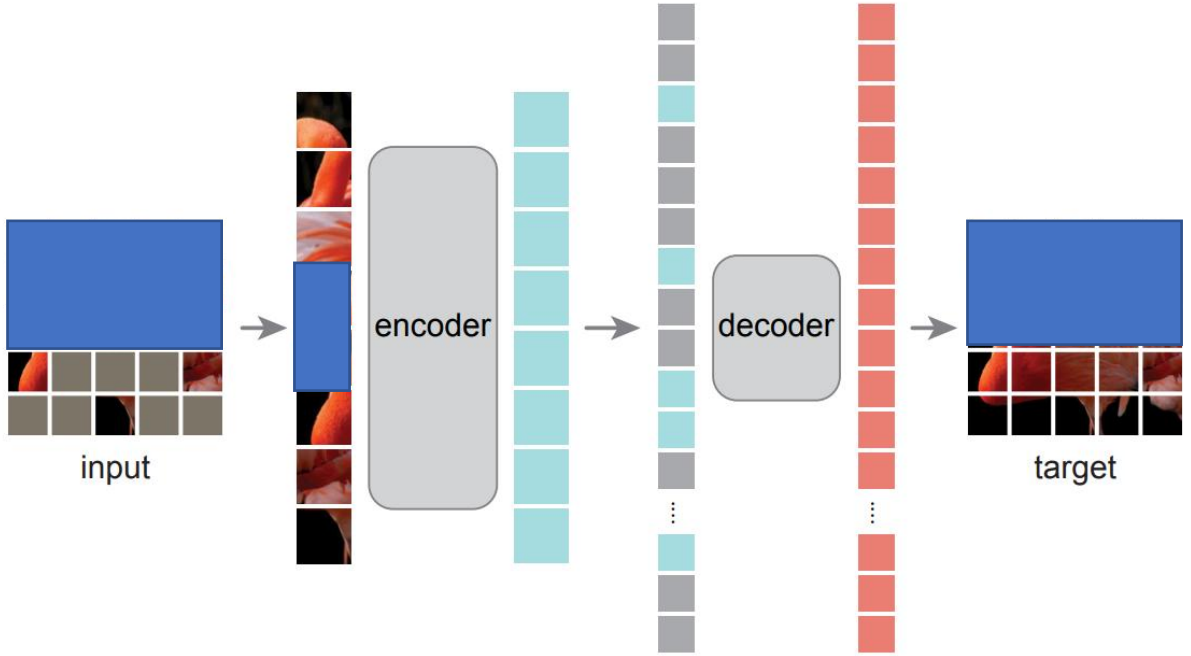
- MAE (CVPR 2022)

Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

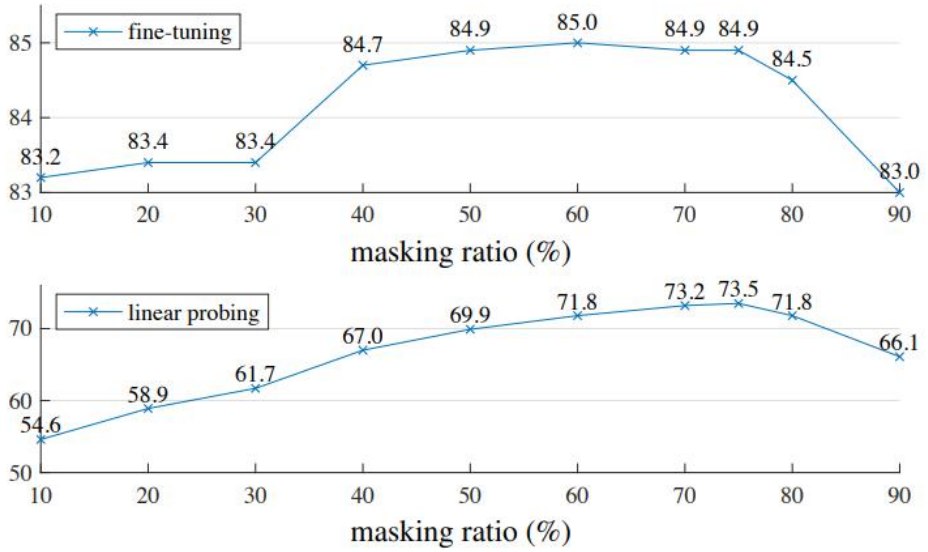
*equal technical contribution †project lead

Facebook AI Research (FAIR)



Background

MAE (CVPR 2022)

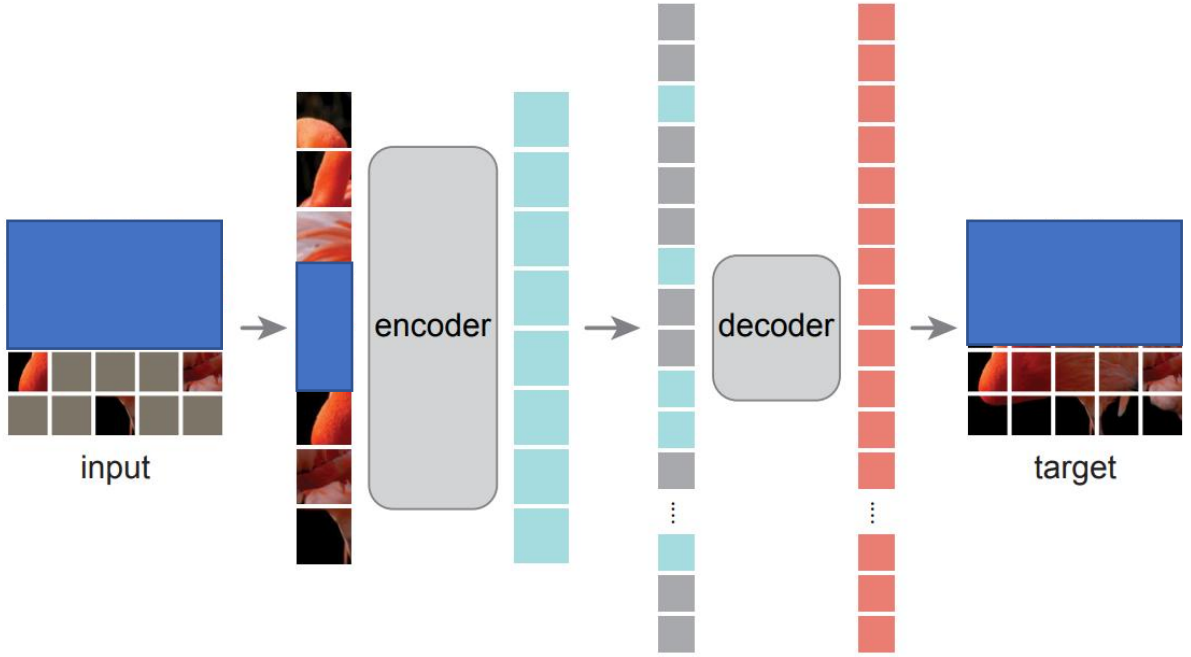


Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)



Background

MAE (CVPR 2022)

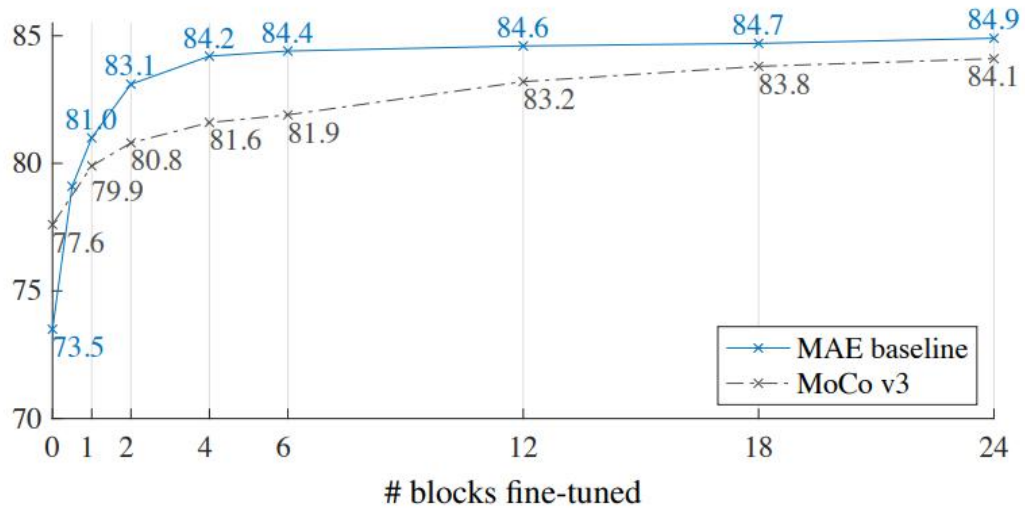
Masked Autoencoders Are Scalable Vision Learners

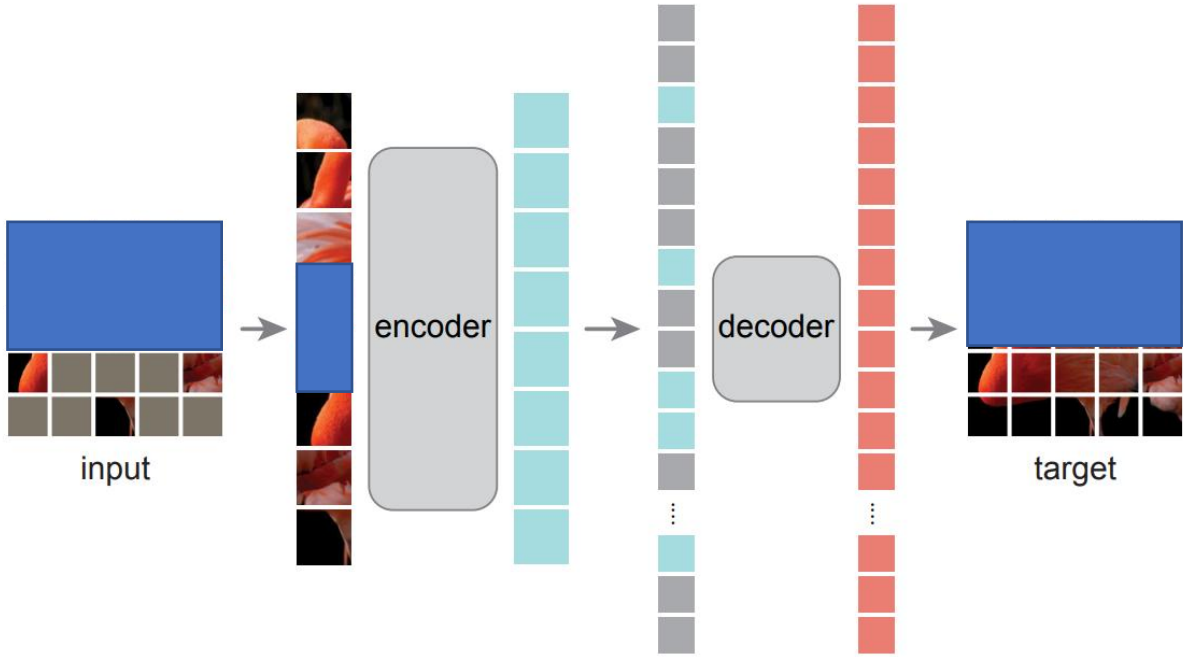
Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**





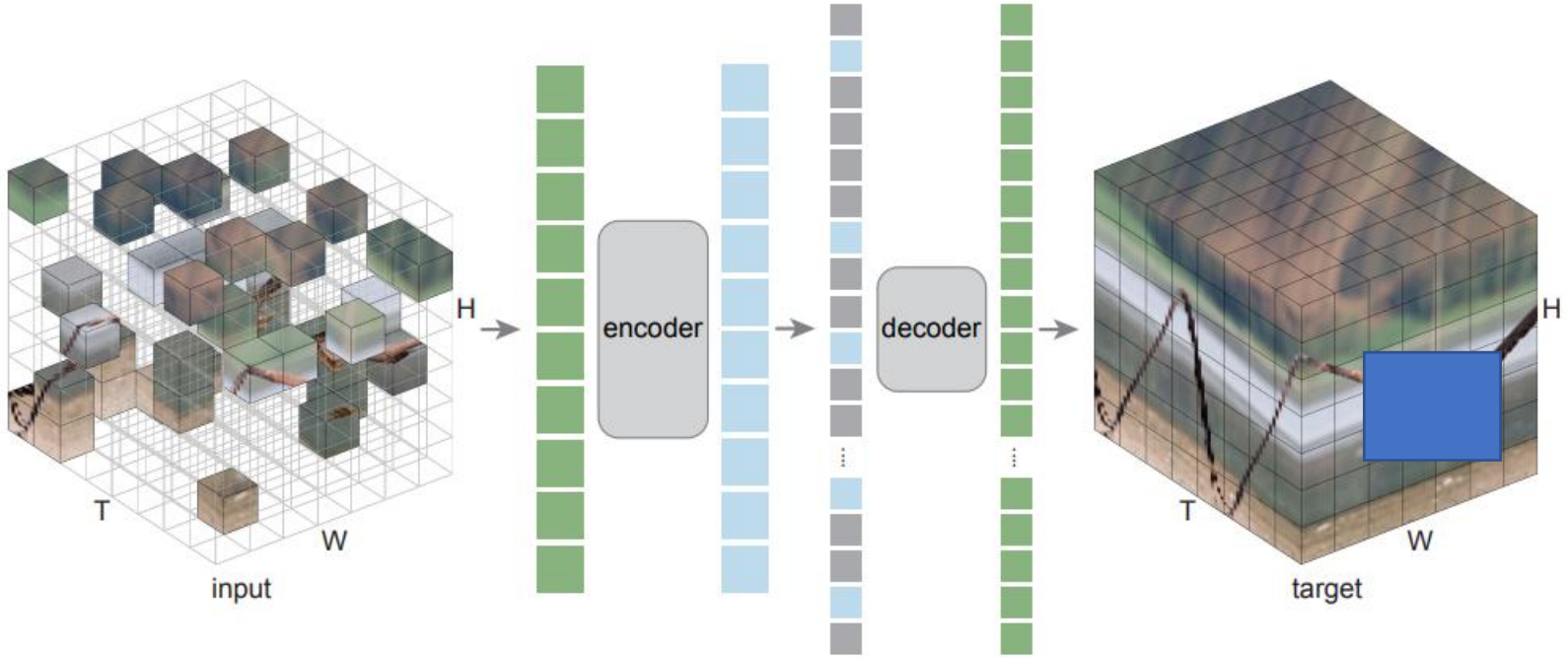
Background

- Video MAE

Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer* Haoqi Fan* Yanghao Li Kaiming He
Meta AI, FAIR

https://github.com/facebookresearch/mae_st



Background

Video MAE

Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer* **Haoqi Fan*** **Yanghao Li** **Kaiming He**
Meta AI, FAIR

https://github.com/facebookresearch/mae_st

pre-train set	# pre-train data	pre-train method	K400	AVA	SSv2
-	-	none (from scratch)	71.4	-	-
IN1K	1.28M	supervised	78.6	17.3	50.2
IN1K	1.28M	MAE	82.3	26.3	65.6
K400	240k	supervised	-	21.6	55.7
K400	240k	MAE	84.8	31.1	72.1
K600	387k	MAE	**84.9**	32.5	73.0
K700	537k	MAE	n/a†	33.1	**73.6**
IG-uncurated	1M	MAE	84.4	**34.2**	**73.6**

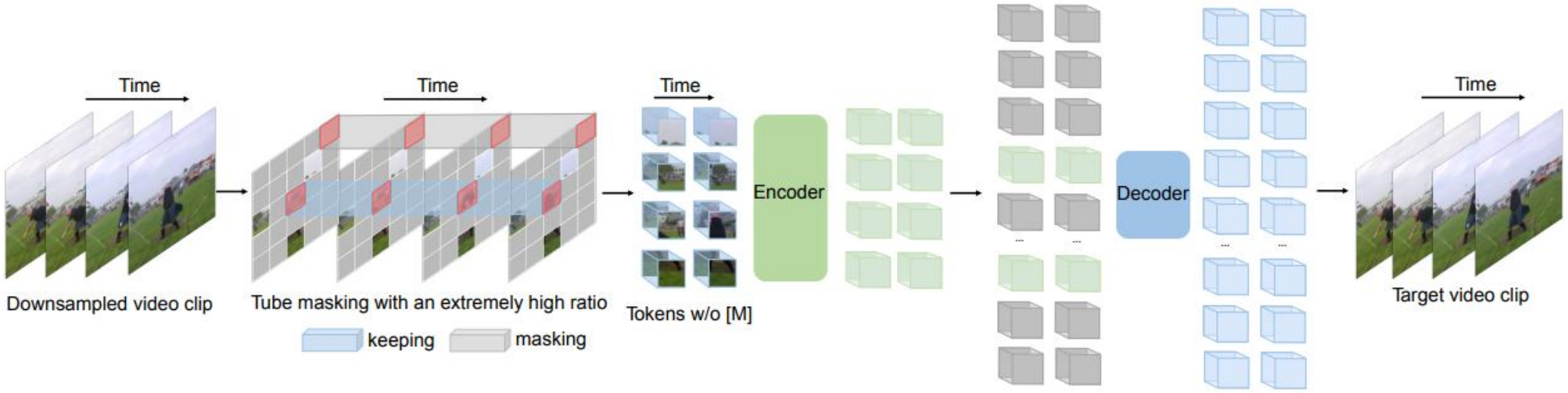
MAE		ViT-B	16×224^2	81.3	94.9	$180 \times 3 \times 7$	87
MAE		ViT-L	16×224^2	84.8	96.2	$598 \times 3 \times 7$	304
MAE		ViT-H	16×224^2	85.1	96.6	$1193 \times 3 \times 7$	632
MAE		ViT-L	40×312^2	85.8	96.9	$4757 \times 3 \times 7$	304
MAE		ViT-H	32×312^2	86.0	97.0	$6382 \times 3 \times 7$	632

Background

- VideoMAE (NeurIPS 2022)

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn



Background

- VideoMAE (NeurIPS 2022)

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn

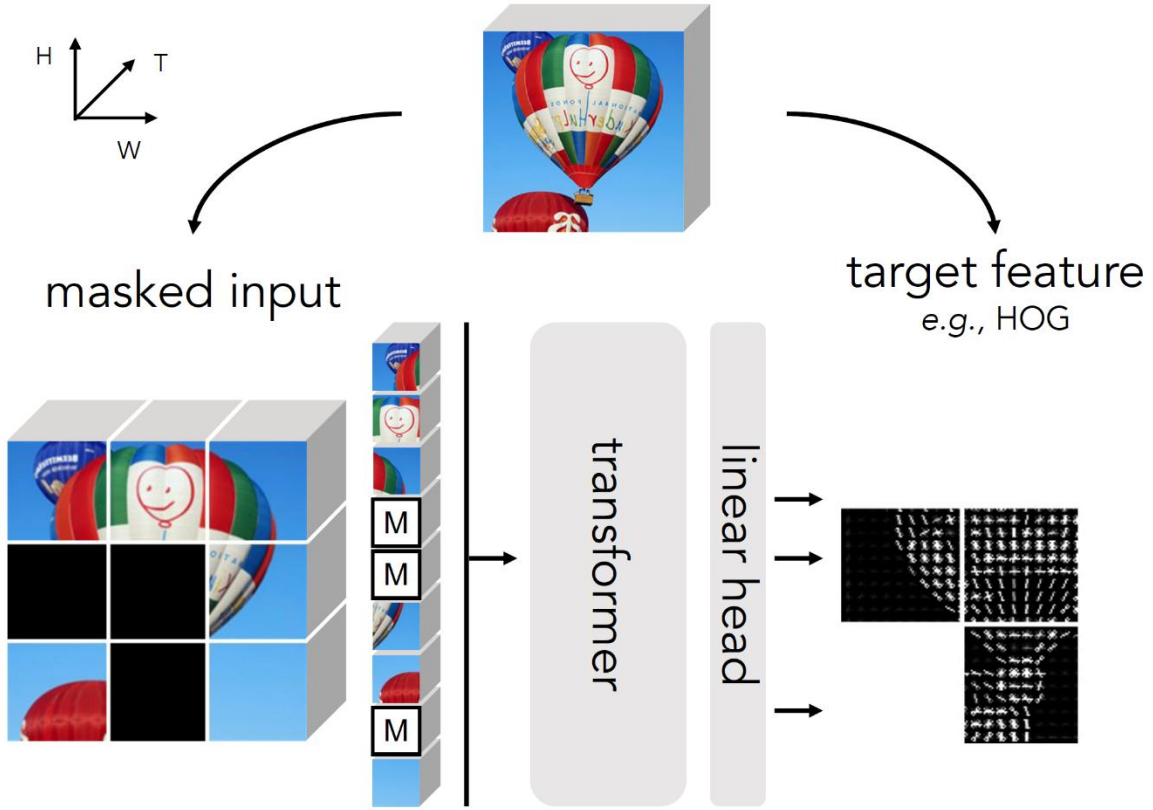
dataset	training data	*from scratch*	MoCo v3	VideoMAE
K400	240k	68.8	74.2	**80.0**
Sth-Sth V2	169k	32.6	54.2	**69.6**
UCF101	9.5k	51.4	81.7	**91.3**
HMDB51	3.5k	18.0	39.2	**62.6**

Background

Masked Feature Prediction for Self-Supervised Visual Pre-Training

Chen Wei [*, 1, 2] Haoqi Fan [1] Saining Xie [1] Chao-Yuan Wu [1] Alan Yuille [2] Christoph Feichtenhofer [*, 1]
[*]equal technical contribution
[1]Facebook AI Research [2]Johns Hopkins University

- **MaskFeat (CVPR 2022)**

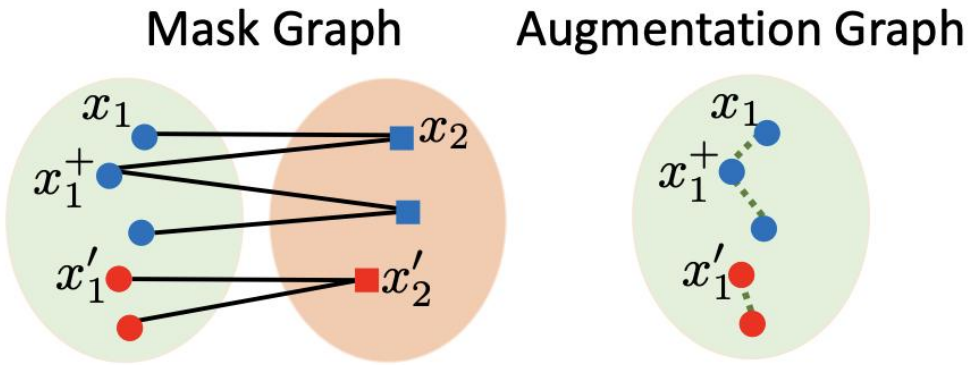


Background

- **U-MAE (ICLR 2023)**

How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders

Qi Zhang[1*] **Yifei Wang**[2*] **Yisen Wang**[1,3†]
[1] Key Lab. of Machine Perception (MoE),
School of Intelligence Science and Technology, Peking University
[2] School of Mathematical Sciences, Peking University
[3] Institute for Artificial Intelligence, Peking University



(a) mask graph and augmentation graph

Background

ON THE ROLE OF DISCRETE TOKENIZATION IN VISUAL REPRESENTATION LEARNING

- **K-MIM (ICLR 2024)**

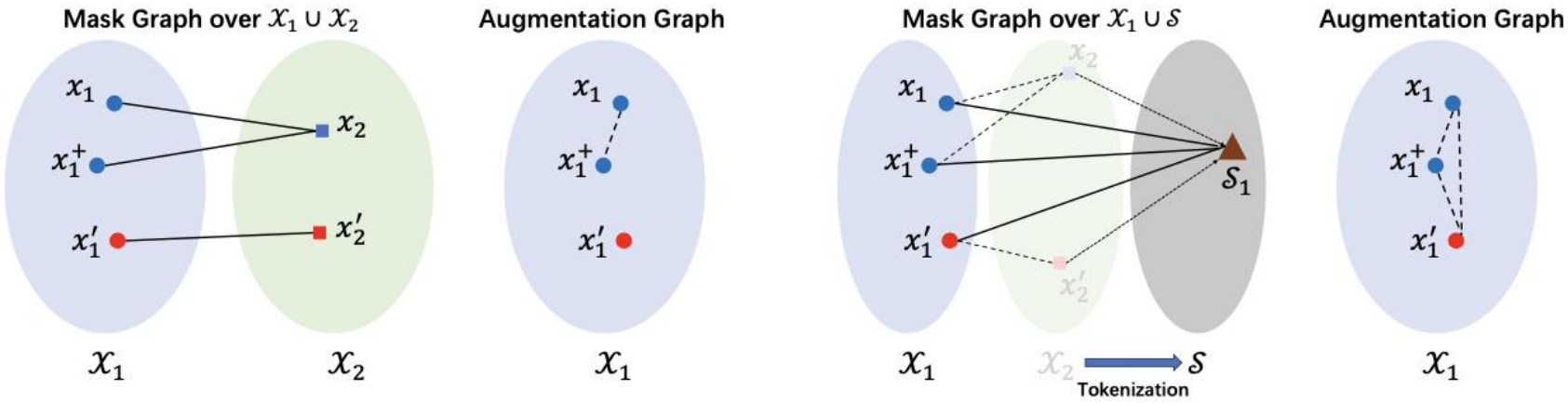
Table 1: Linear probing and fine-tuning accuracies (%) of several MIM methods using different tokenizers pretrained on ImageNet100 for 200 epochs. The architecture is chosen as ViT-small following Touvron et al. (2021)). The comparison baseline is the identity function.

Tokenizer	Linear Probing Acc.	Fine-tuning Acc.
Identity function (MAE)	45.94	81.78
dVAE (Vahdat et al., 2018)	41.24 (−4.70)	80.21 (−1.57)
VQGAN (Esser et al., 2021)	44.25 (−1.69)	80.88 (−0.90)
Perceptual tokenizer (Dong et al., 2022)	50.29 (+5.35)	83.13 (+1.45)
Maskfeat (HOG targets) (Wei et al., 2022)	48.52 (+2.58)	82.91 (+1.13)

Background

ON THE ROLE OF DISCRETE TOKENIZATION IN VISUAL REPRESENTATION LEARNING

- **K-MIM (ICLR 2024)**



(a) Graphs of MAE

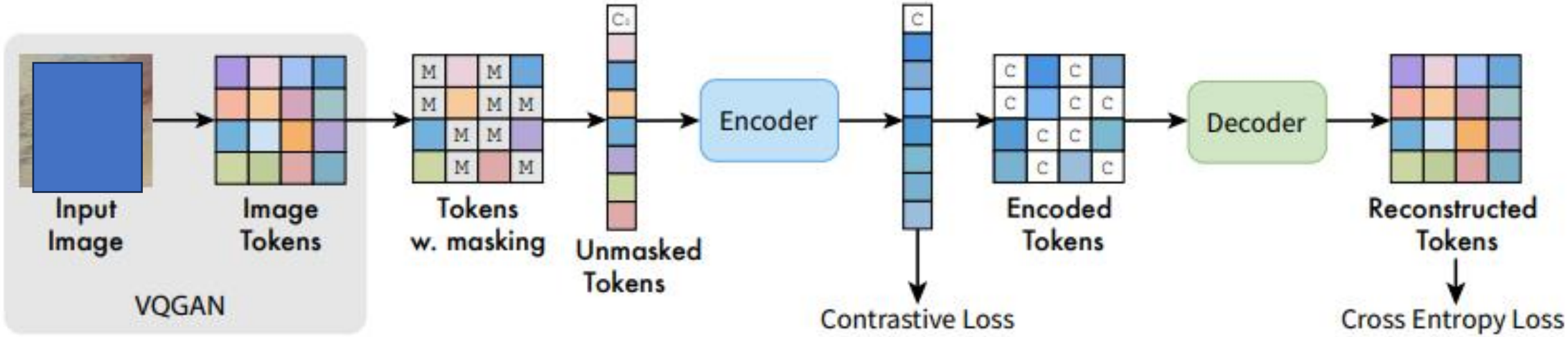
(b) Graphs of MIM with discrete tokenization

Background

- **MAGE (CVPR 2023)**

MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis

Tianhong Li[1*] Huiwen Chang[3] Shlok Kumar Mishra[2] Han Zhang[3]
Dina Katabi[1] Dilip Krishnan[3]
[1]MIT CSAIL [2]University of Maryland [3]Google Research



Background

- **MAGE (CVPR 2023)**

MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis

Tianhong Li[1]* Huiwen Chang[3] Shlok Kumar Mishra[2] Han Zhang[3]
Dina Katabi[1] Dilip Krishnan[3]
[1]MIT CSAIL [2]University of Maryland [3]Google Research

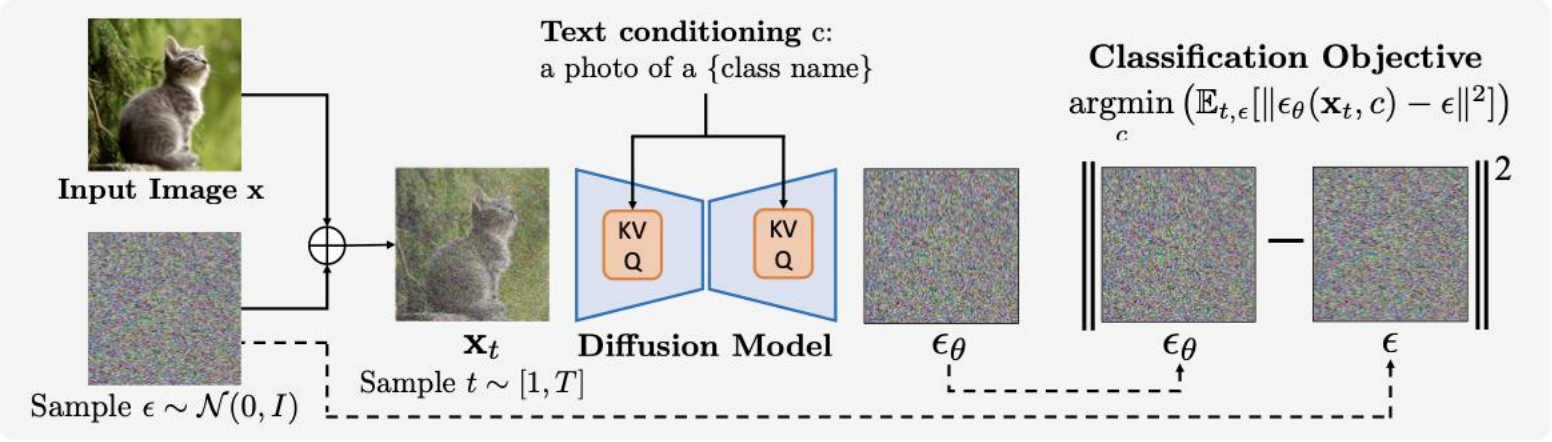
MIM methods			
BEiT [2]	ViT-B	86M	56.7
MAE [26]	ViT-B	86M	68.0
Ge^2-AE [36]	ViT-B	86M	**75.3**
MAGE	ViT-B	24M+86M	74.7
MAE [26]	ViT-L	304M	75.8
MAGE	ViT-L	24M+304M	**78.9**
Contrastive methods			
SimCLRv2 [10]	RN50w2	94M	75.6
BYOL [25]	RN50w2	94M	77.4
CAE [11]	ViT-B	86M	70.4
CMAE [31]	ViT-B	86M	73.9
MoCo v3 [13]	ViT-B	86M	76.7
DINO [59]	ViT-B	86M	72.8
iBOT [59]	ViT-B	86M	76.0
MAGE-C	ViT-B	24M+86M	**78.2**

Background

- **Diff-Classifier (ICCV 2023)**

Your Diffusion Model is Secretly a Zero-Shot Classifier

Alexander C. Li Mihir Prabhudesai Shivam Duggal Ellis Brown Deepak Pathak
Carnegie Mellon University

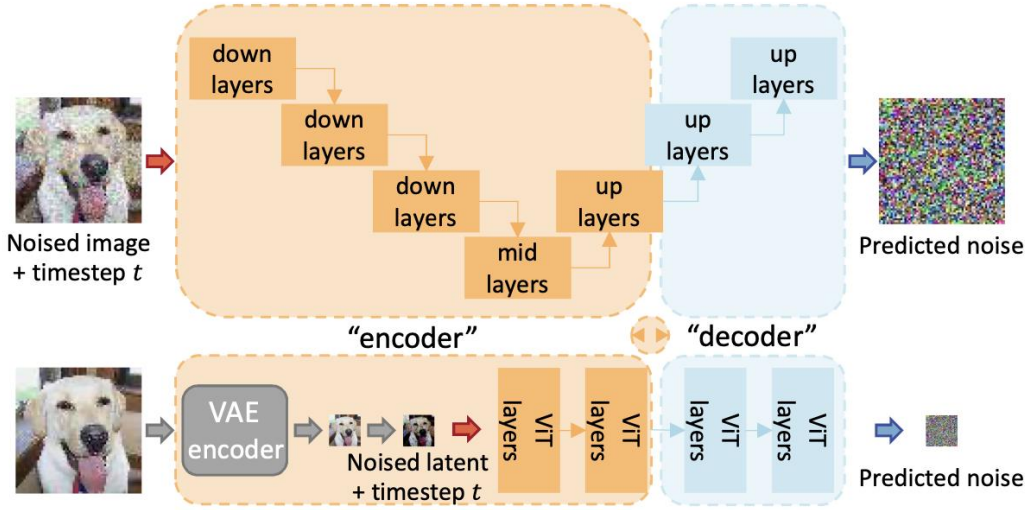


Background

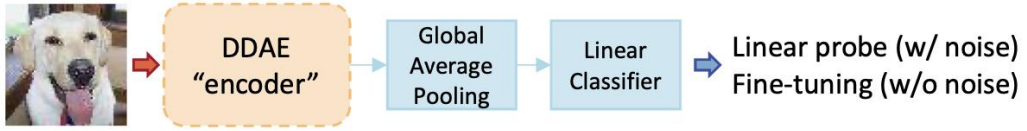
- DDAE (ICCV 2023)

Denoising Diffusion Autoencoders are Unified Self-supervised Learners

Weilai Xiang[1,2] Hongyu Yang[1,3*] Di Huang[2] Yunhong Wang[1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
{xiangweilai, hongyuyang, dhuang, yhwang}@buaa.edu.cn



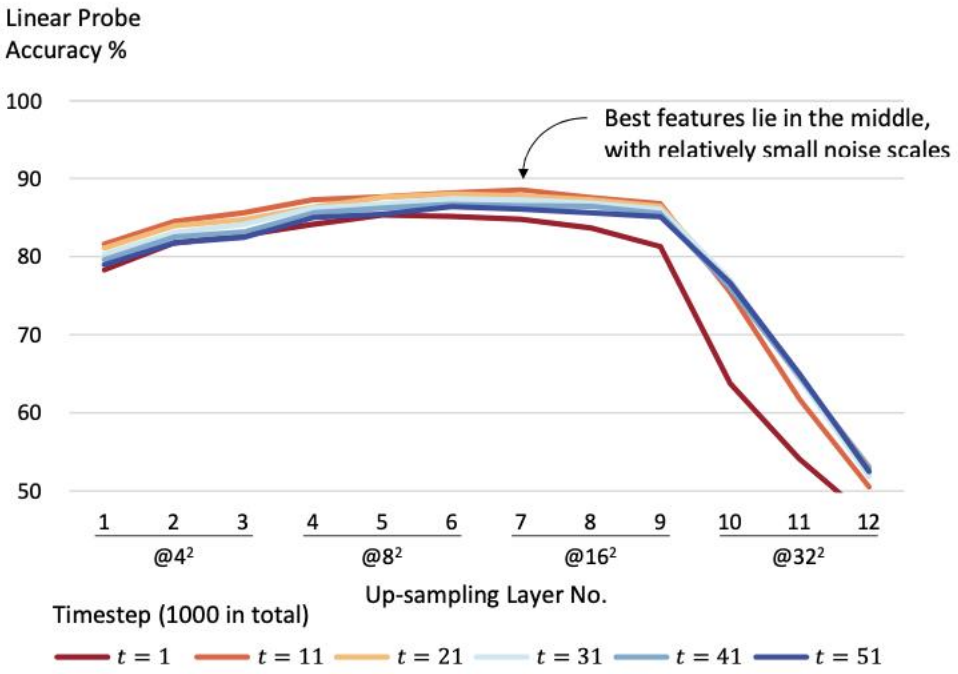
(a) Denoising networks in pixel-space and latent-space diffusion models.



(b) Evaluating DDAEs as self-supervised representation learners.

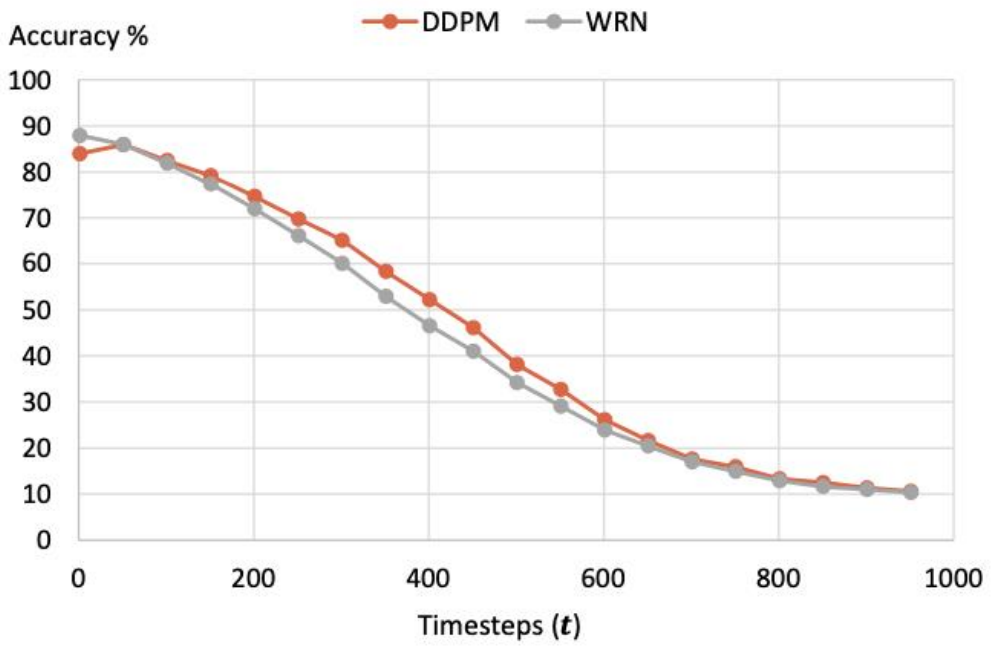
Background

- **DDAE (ICCV 2023)**



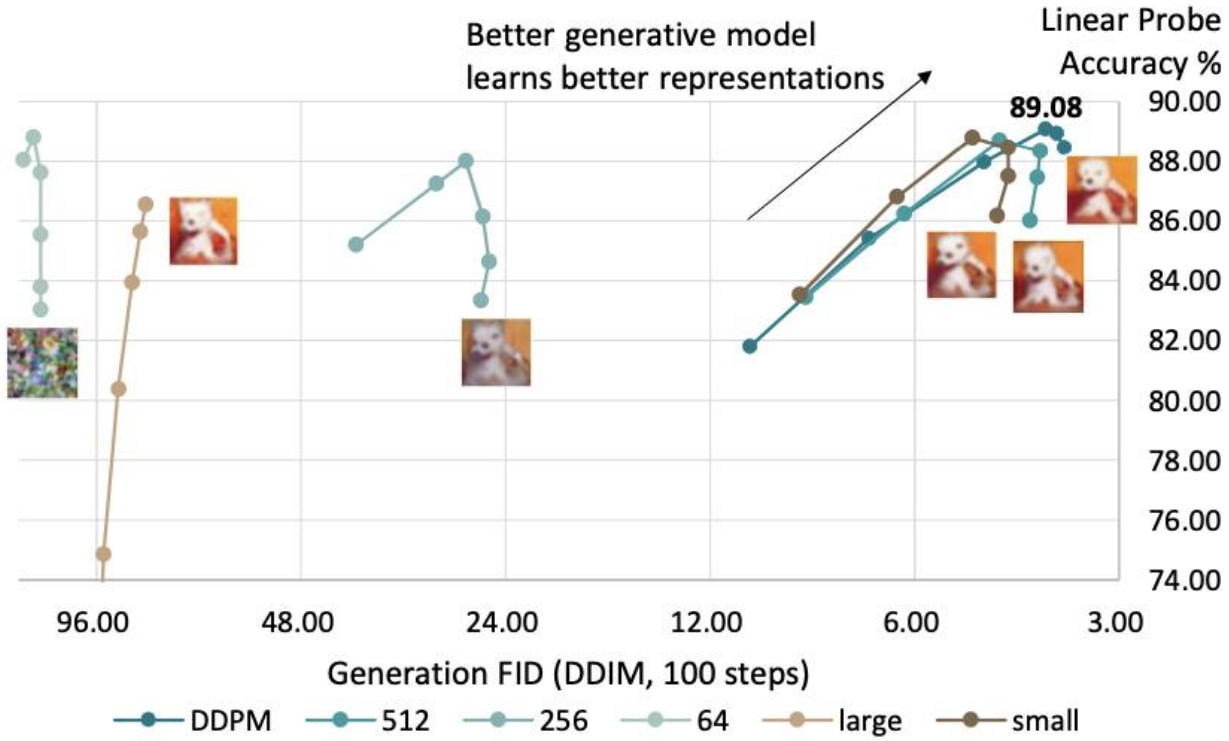
Denoising Diffusion Autoencoders are Unified Self-supervised Learners

Weilai Xiang[1,2] Hongyu Yang[1,3*] Di Huang[2] Yunhong Wang[1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
{xiangweilai, hongyuyang, dhuang, yhwang}@buaa.edu.cn



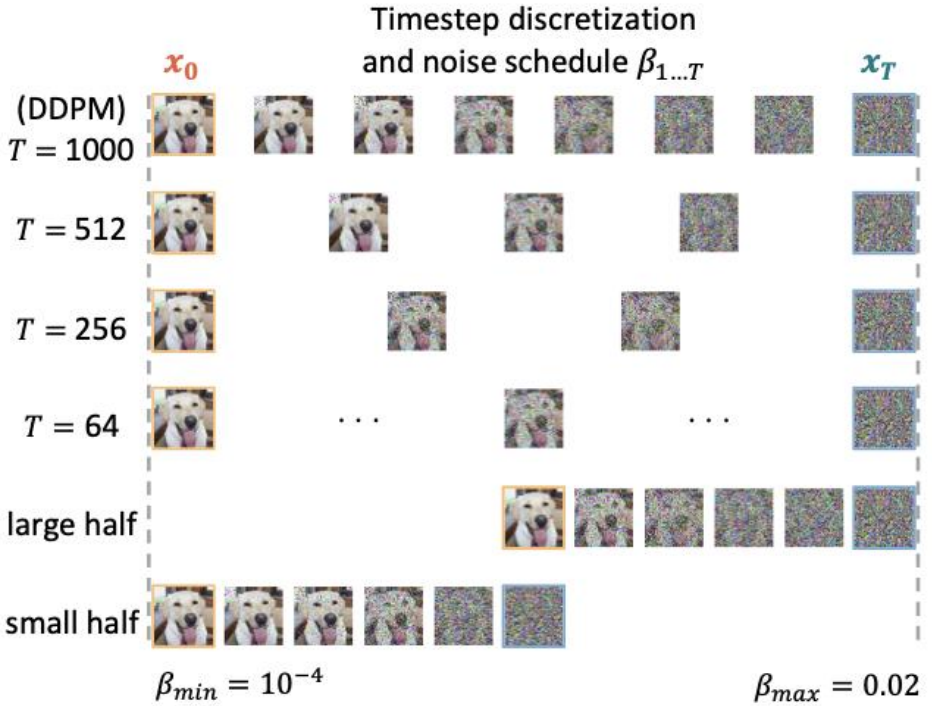
Background

- **DDAE (ICCV 2023)**



Denoising Diffusion Autoencoders are Unified Self-supervised Learners

Weilai Xiang[1,2] Hongyu Yang[1,3*] Di Huang[2] Yunhong Wang[1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
{xiangweilai, hongyuyang, dhuang, yhwang}@buaa.edu.cn

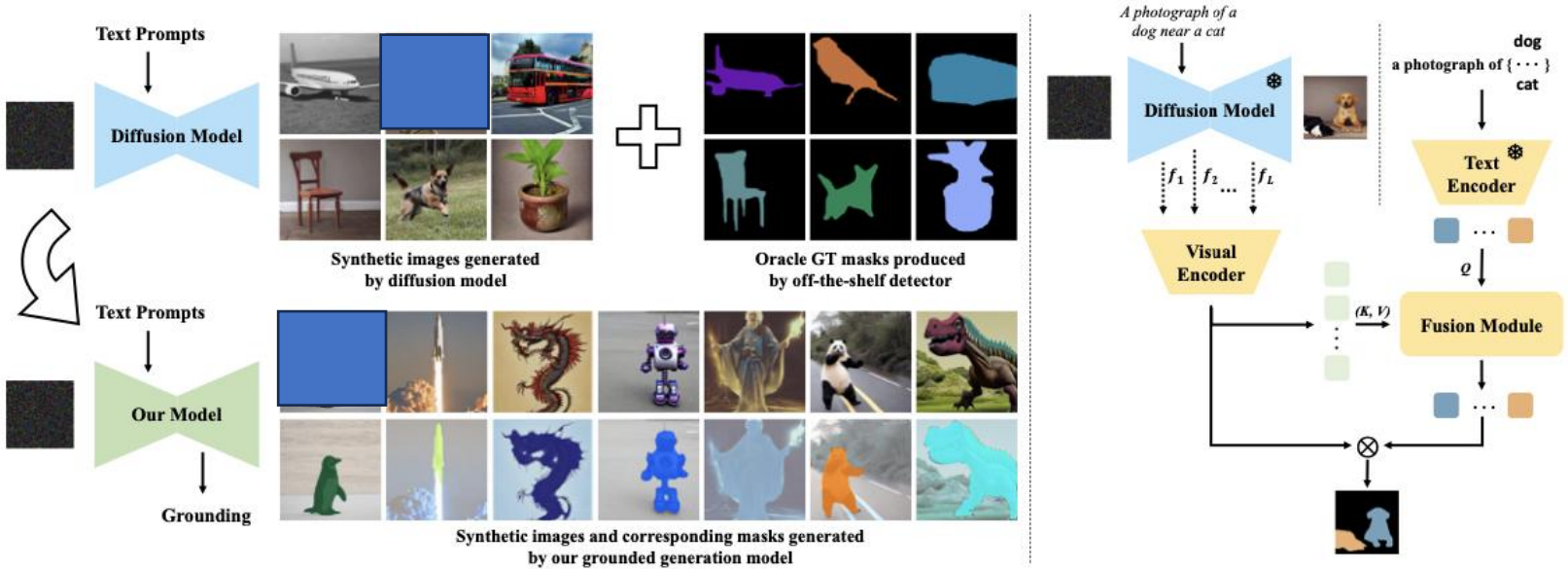


Background

- **OVOS Diffusion (ICCV 2023)**

Open-vocabulary Object Segmentation with Diffusion Models

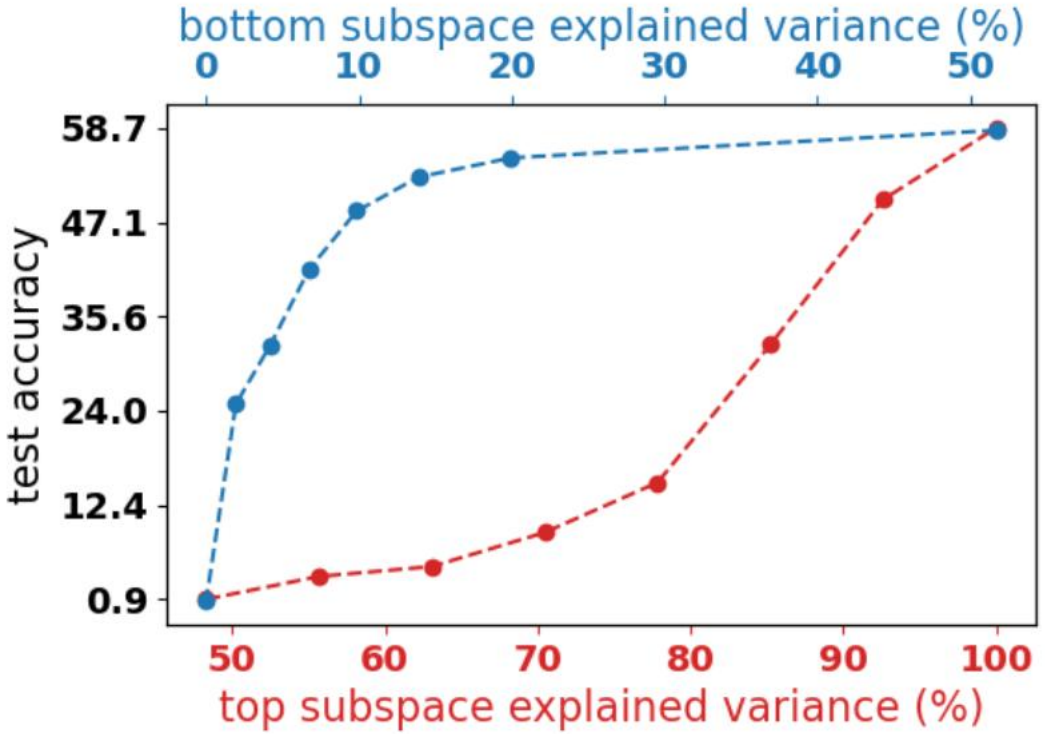
Ziyi Li[1*], Qinye Zhou[1*], Xiaoyun Zhang[1], Ya Zhang[1,2], Yanfeng Wang[1,2], and Weidi Xie[1,2]
[1]Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
[2]Shanghai AI Laboratory, China
https://lipurple.github.io/Grounded_Diffusion/

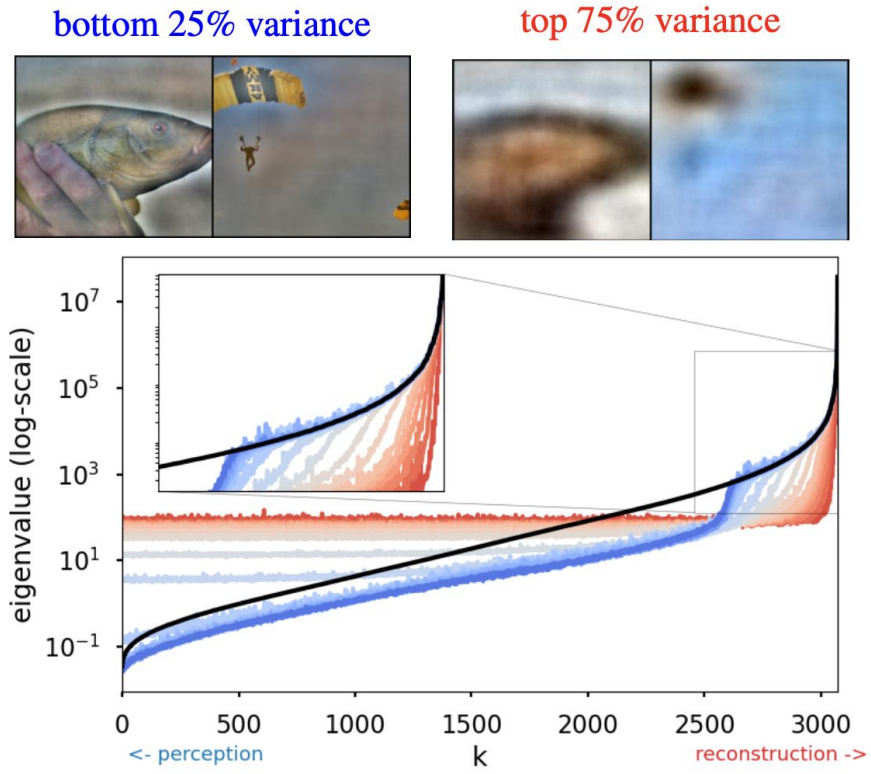


Outline

Method

- **Reconstruction are uninformative for perception**





Method

Loss of Reconstruction and Supervised Tasks

$$\mathcal{L}(V, W, Z) = \|W^\top V^\top X - Y\|_F^2 + \lambda \|Z^\top V^\top X - X\|_F^2,$$

Theorem 1. *The loss function from Eq. (3) is minimized for*

$$V^* \text{ spans } P_{XX^\top} D_{XX^\top}^{-\frac{1}{2}} (P_H)_{.,1:K}, \tag{4}$$

$$W^* = \left(V^{*\top} X X^\top V^*\right)^{-1} V^{*\top} X Y^\top, \tag{5}$$

$$Z^* = \left(V^{*\top} X X^\top V^*\right)^{-1} V^{*\top} X X^\top, \tag{6}$$

where $H \triangleq D_{XX^\top}^{-\frac{1}{2}} P_{XX^\top}^\top A P_{XX^\top} D_{XX^\top}^{-\frac{1}{2}}$. *(Proof in Section 6.1, empirical validation in Fig. 8.)*

$$A \triangleq X \left(Y^\top Y + \lambda X^\top X\right) X^\top.$$

Method

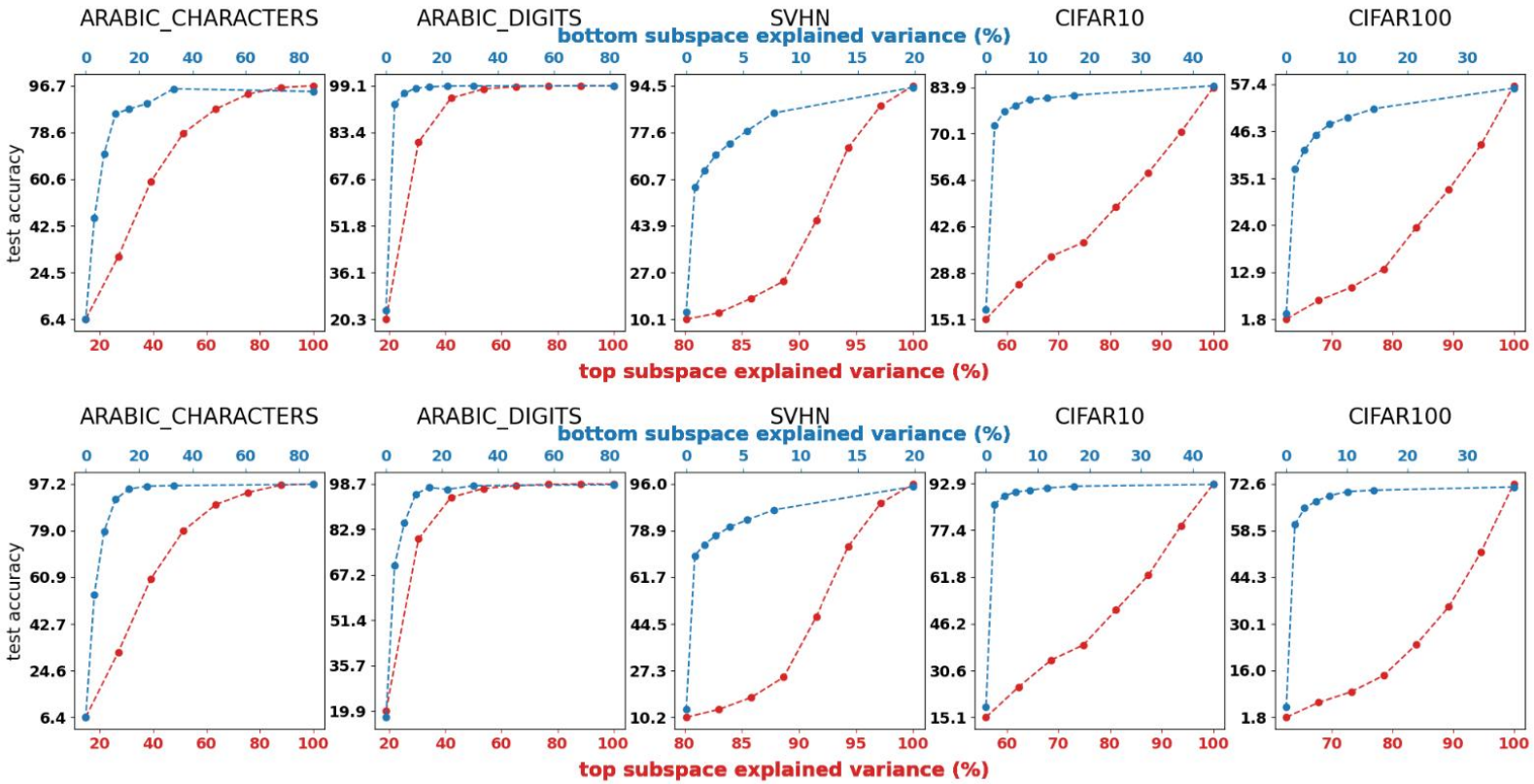
- **Loss of Reconstruction and Supervised Tasks**

Proposition 1. *The supervised and reconstruction tasks are aligned (the optimal solutions do not depend on λ) iff the intersection of the top-K eigenspaces of $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{Y}^\top \boldsymbol{Y}$ is of dimension K.*

Outline

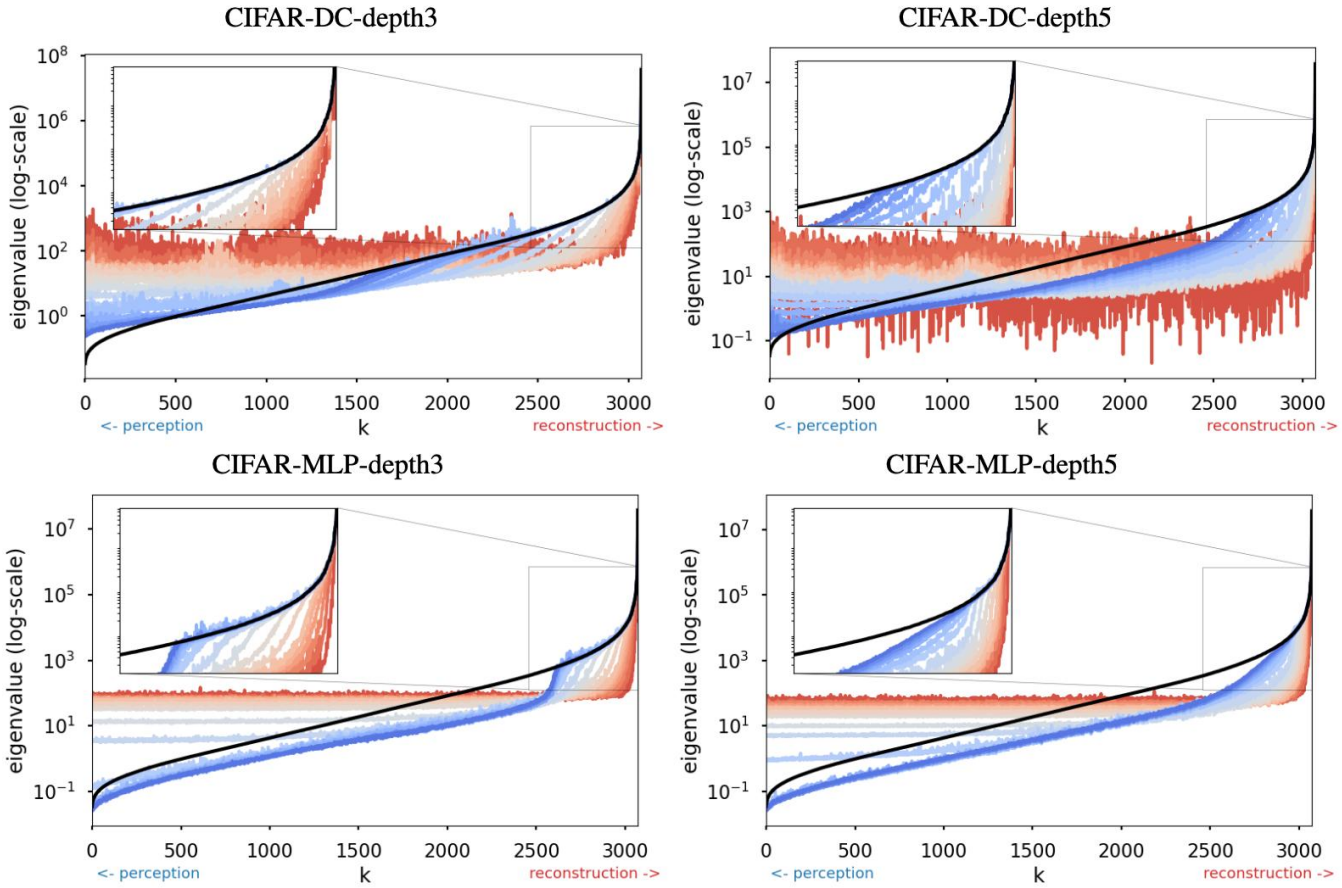
Experiments

- **Reconstruction and Perception Features In Different Subspace**



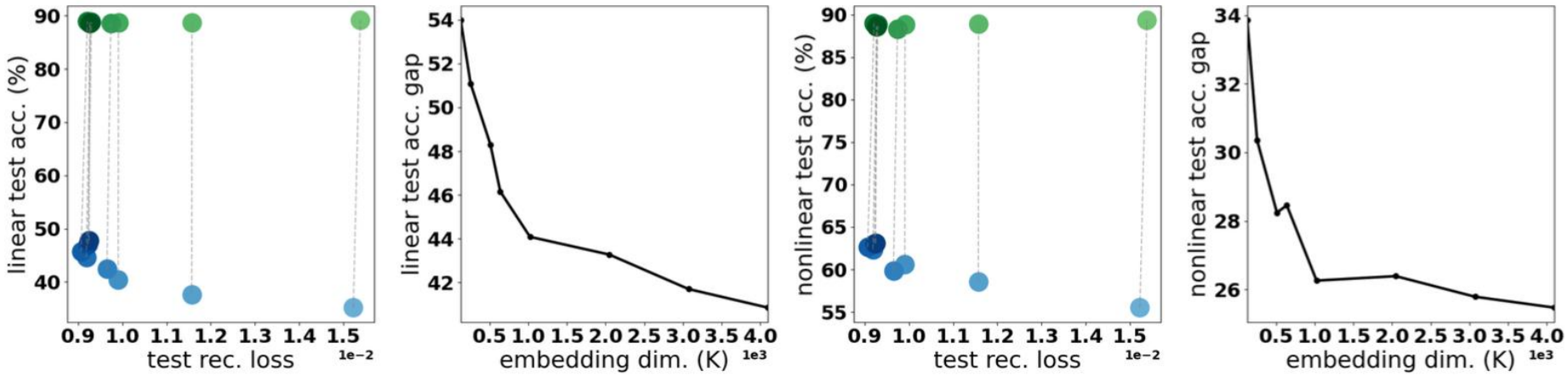
Experiments

- **Useful features for perception are learned last**



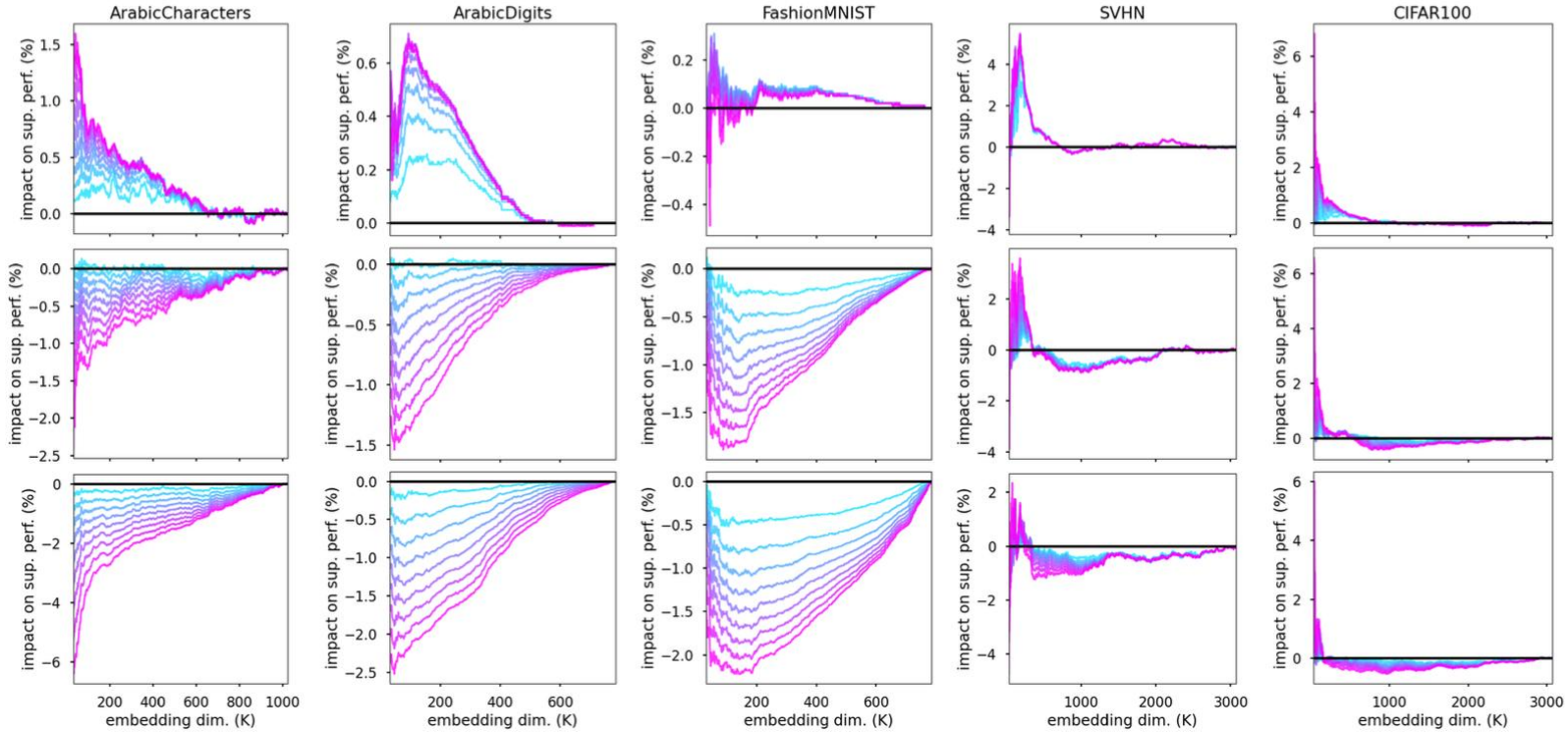
Experiments

- **Learning By Reconstruction Needs Guidance**



Experiments

- **Provable Benefits of Learning by Denoising**



Outline

Discussion

- Perception can not be solved from the principal subspace of the data.

Thanks!