Mixed Autoencoder for Self-supervised Visual Representation Learning

Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung

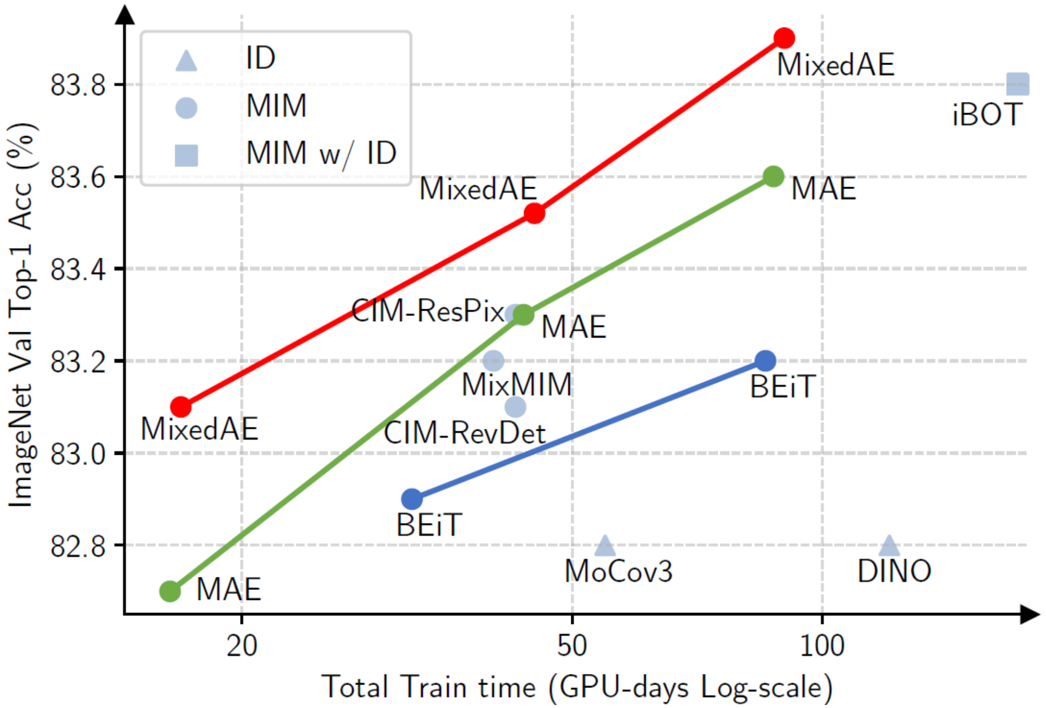
https://kaichen1998.github.io/

Poster session: @THU-PM-204

June 1st, 2023

Overview

- Mixing as effective augmentation for Masked Image Modeling (MIM)



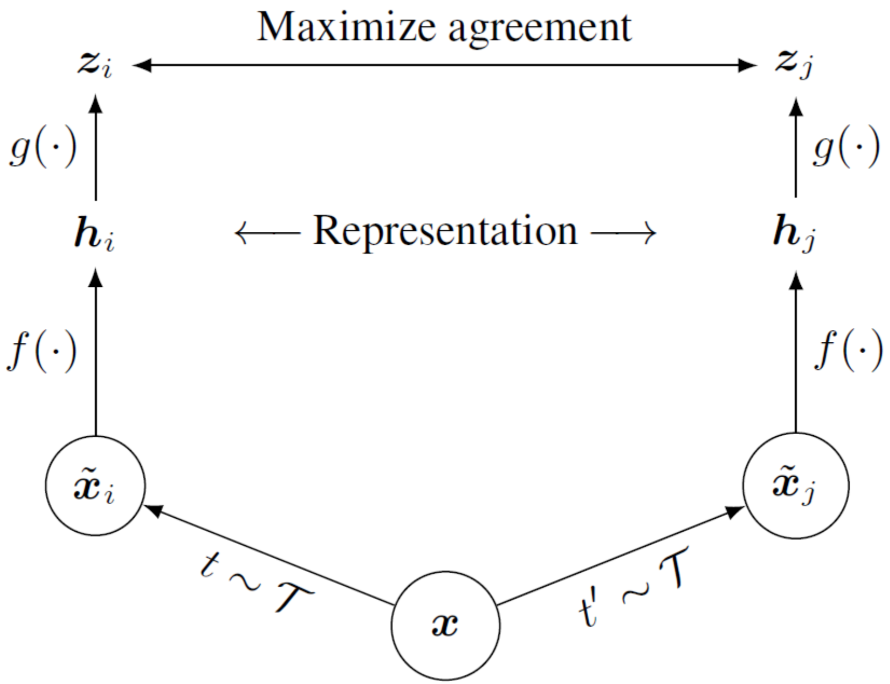
- Theoretical analysis between MIM and previous supervisions.

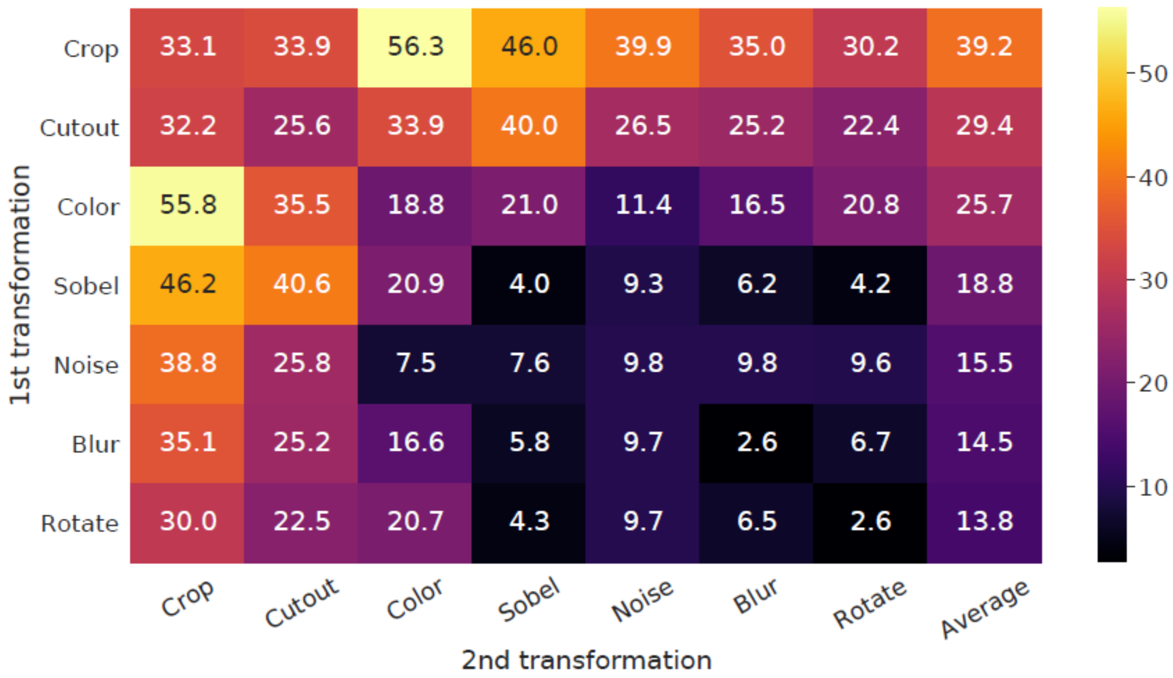
- Mixed Autocoder (MixedAE) achieves SoTA performance with superior efficiency (e.g., surpasses iBOT with 2x acceleration).

Zhou, Jinghao, et al. "iBOT: Image BERT Pre-Training with Online Tokenizer." *ICLR*. 2022.

Contrastive Learning

- Discriminate positive samples from negative ones.
- Rely on strong data augmentation pipelines.

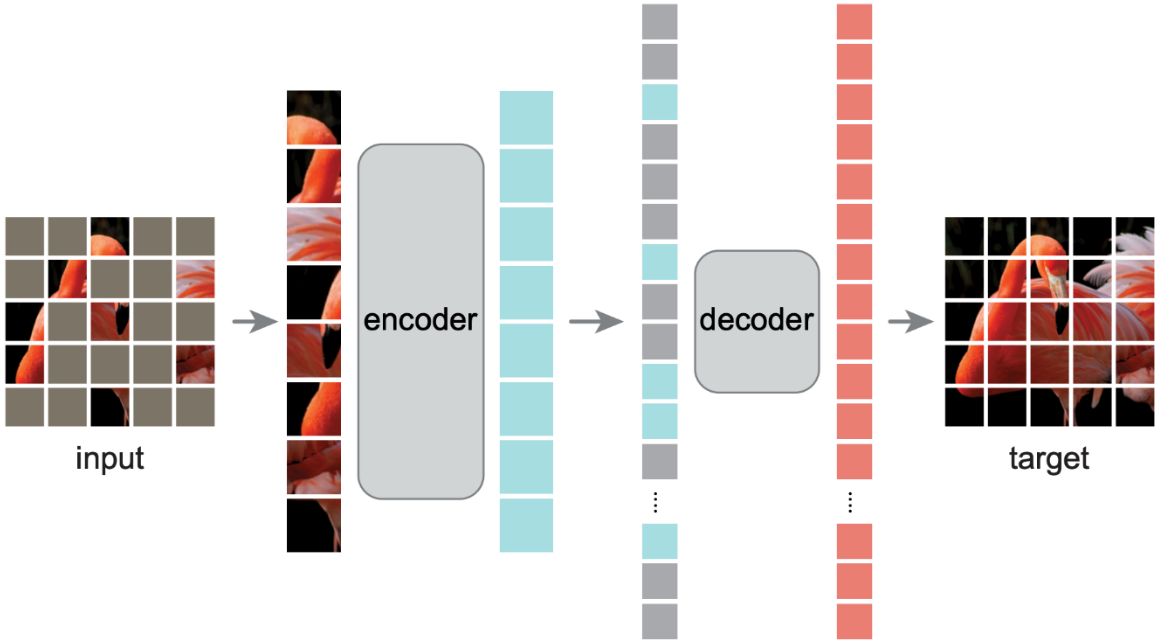




Chen, Ting, et al. "A Simple Framework for Contrastive Learning of Visual Representations." *ICML*. 2020.

Masked Autoencoder (MAE)

- A representative implementation of MIM.
- Perform even worse with strong data augmentations.



case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

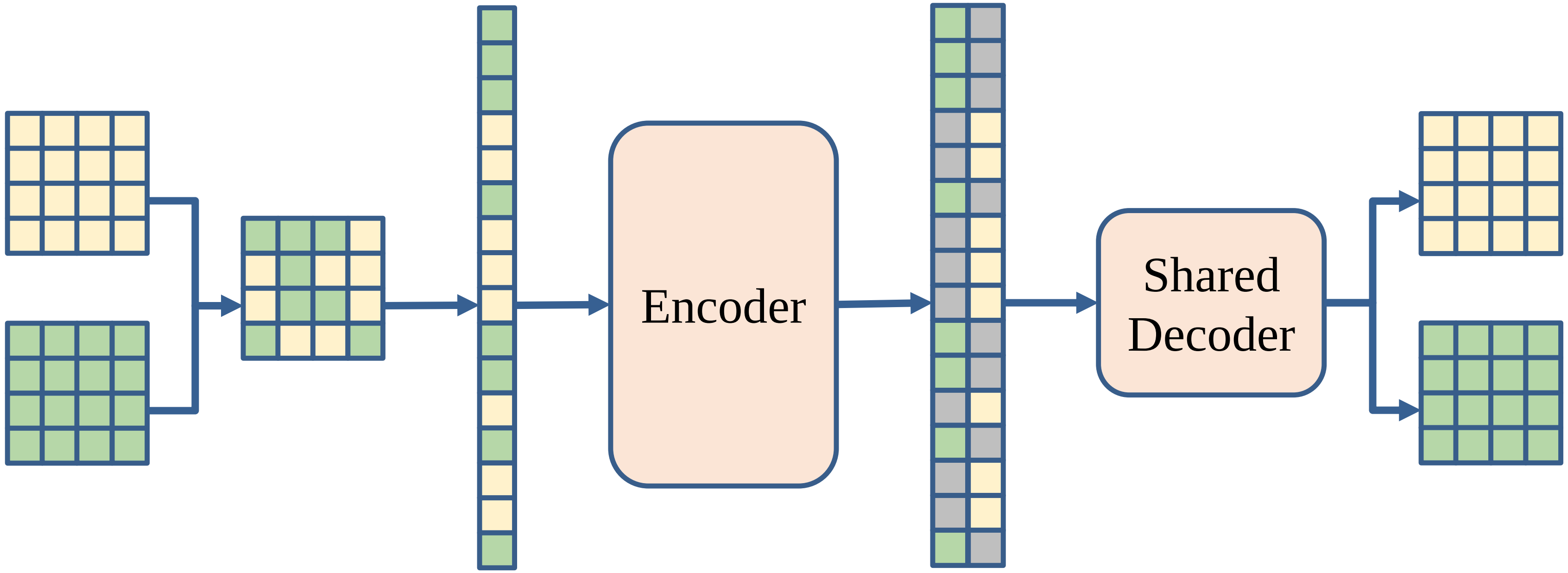
(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *CVPR*. 2022.

A Simple Mixing Baseline

- Mixed image reconstruction: symmetric mixing
  - Given a mixing ratio $r \in (0, 0.5]$ and a random mixing mask M,

$$\hat{x} = \sigma_{mix}\left(\{x_i\}_{i=1}^{1/r}, M\right) = \sum_{i=1}^{1/r} 1(M = i)x_i$$



Mutual Information Analysis

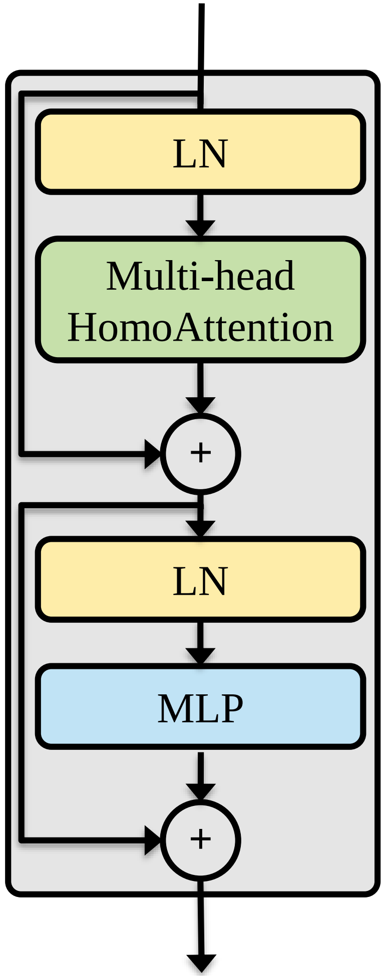
- Image mixing will improve the mutual information (MI),
  - While masking is introduced to decrease MI instead.
- Target-invariant
  - MI increasement is appealing for supervised and contrastive learning

$$
\begin{aligned}
I(\sigma_{mix}(\{\boldsymbol{X}_1, \boldsymbol{X}_2\}, \boldsymbol{M}); \boldsymbol{X}_1) &= I(\mathbb{1}(\boldsymbol{M}=1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M}=2)\boldsymbol{X}_2; \boldsymbol{X}_1) \\
&= H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M}=1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M}=2)\boldsymbol{X}_2) \\
&\geq H(\boldsymbol{X}_1) - H(\boldsymbol{X}_1 | \mathbb{1}(\boldsymbol{M}=1)\boldsymbol{X}_1 + \mathbb{1}(\boldsymbol{M}=2)\overrightarrow{\boldsymbol{0}}) \\
&= I(\sigma_{MAE}(\boldsymbol{X}_1, \boldsymbol{M}); \boldsymbol{X}_1),
\end{aligned}
$$

Homologous Recognition

- Homologous attention
  - Explicit recognition of homo patches on-the-fly
  - Adopt a TopK$(\cdot)$ sampling in standard MHSA

$$A_{HomoAtt} = \mathrm{softmax}(\mathrm{TopK}(\boldsymbol{q}\boldsymbol{k}^T/\sqrt{D_h})),$$

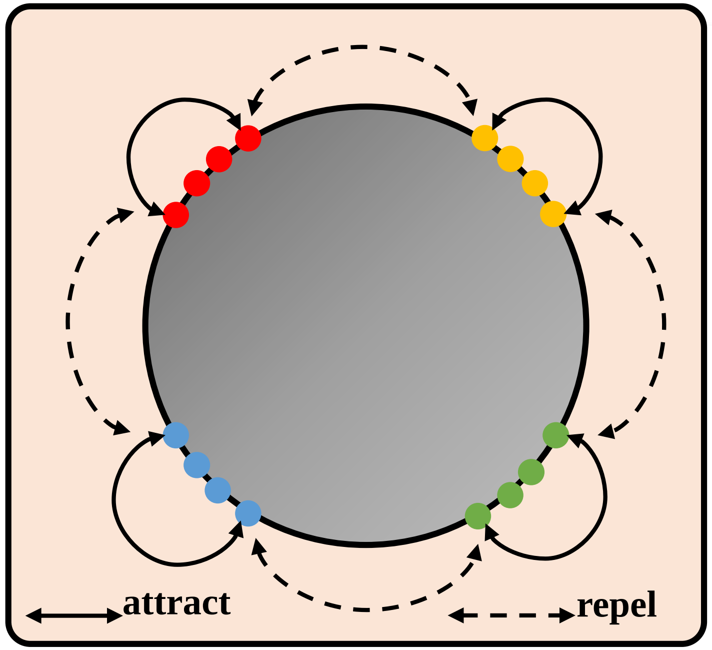


Homologous Recognition

- Homologous contrastive
  - Verify homologous patches by similarity
  - Homologous patches as positive samples

$$\mathcal{L}_{HomoCon} = -\sum_{l=1}^{L}\sum_{l+}\log\frac{exp(cos(\hat{z}_l^j, \hat{z}_{l+}^j)/\tau)}{\sum_{l'\neq l}^{L} exp(cos(\hat{z}_l^j, \hat{z}_{l'}^j)/\tau)},$$

  - Perform as a regularization term instead of a separate supervision

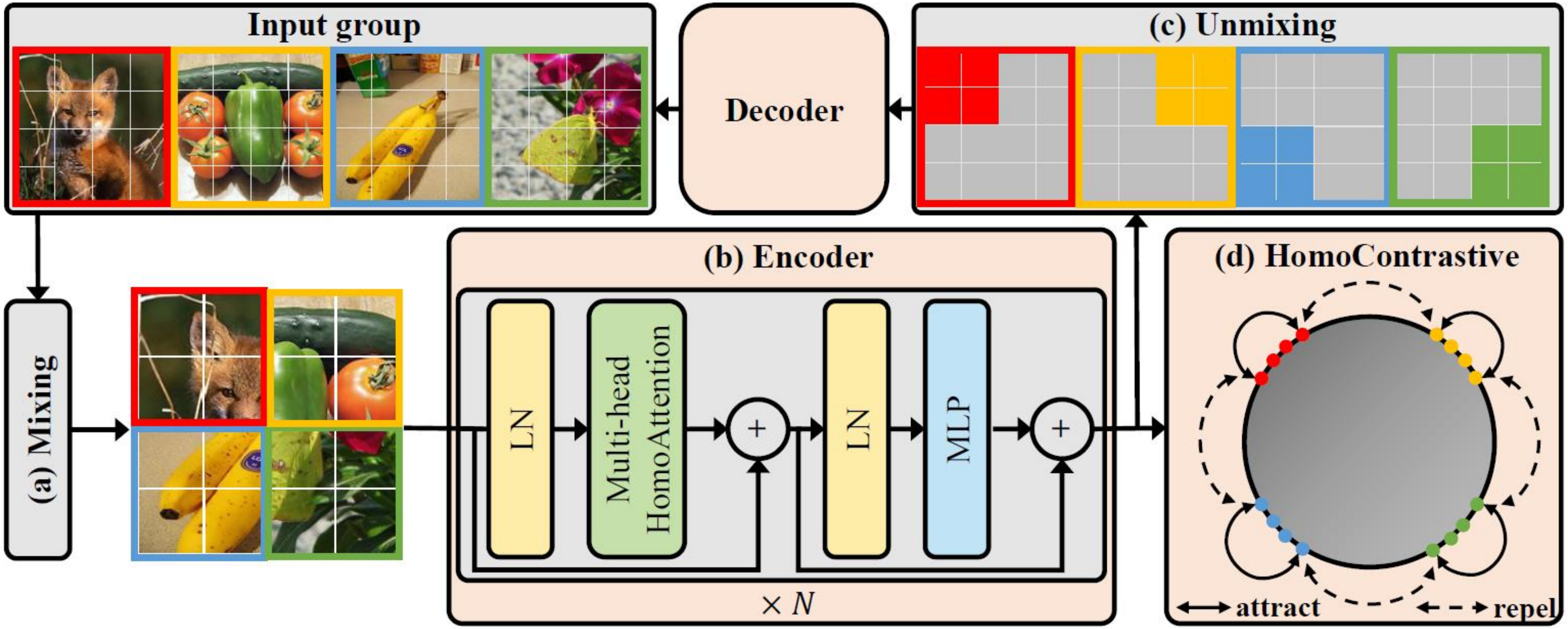


$\mathcal{L}_{recon}$	$\mathcal{L}_{HomoCon}$	acc	mIoU
✓		82.4	45.0
	✓	7.8	8.3
✓	✓	**82.7**	**46.4**

(b) **Functionality of the $\mathcal{L}_{HomoCon}$.** When adopting $\mathcal{L}_{HomoCon}$ alone, *MixedAE* cannot even achieve reasonable transfer performance.

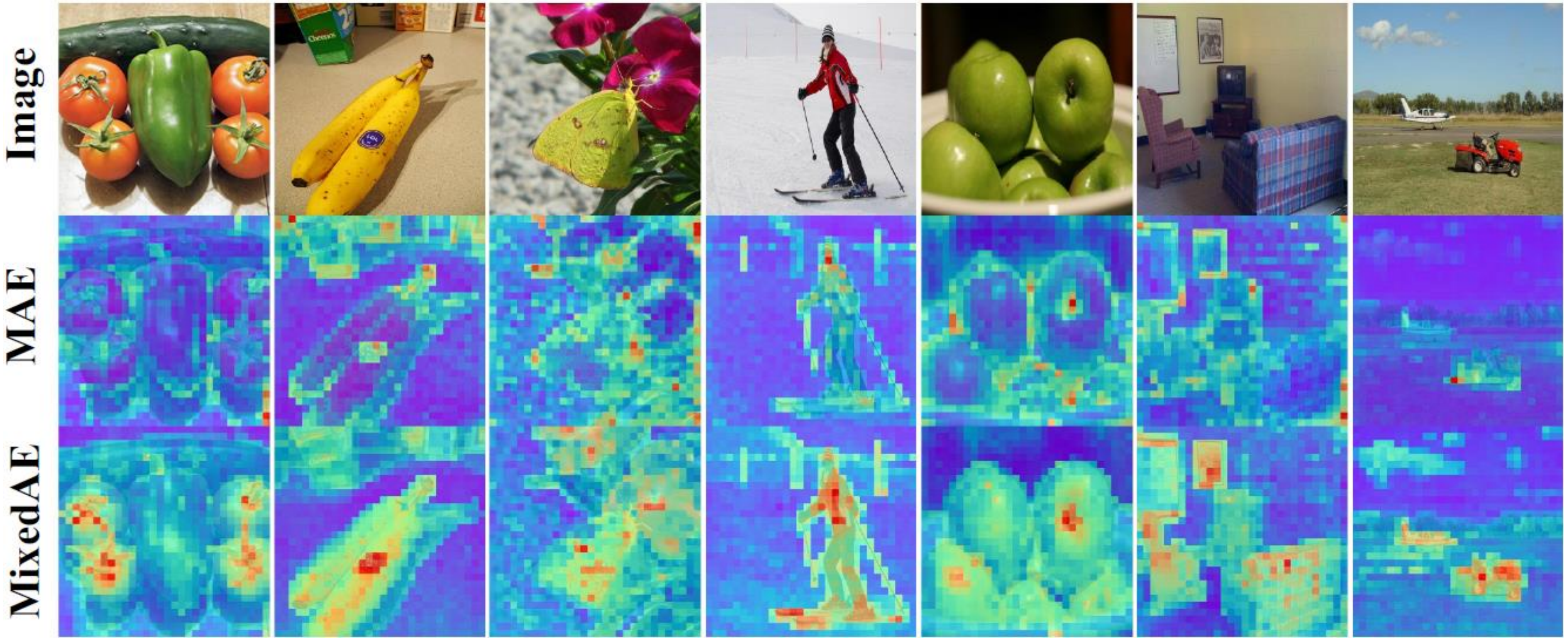
Mixed Autoencoder (MixedAE)

- Formulated in a multi-task learning manner



Object-aware Pre-training

- Object-aware Pre-training
  - Single-centric-object guarantee (Chen et al., 2021)



Chen, et al. "MultiSiam: Self-supervised Multi-instance Siamese Representation Learning for Autonomous Driving." *ICCV* 2021.

Experiments

- Effectiveness: SoTA performance under various overheads

Method	Pre-train	Pre-train†	ImageNet	ADE20K	COCO					
	Epochs	GPU-days	Top-1 Acc.	mIoU	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
DeiT [52]	300	19.6	81.8	46.9	48.8	68.7	52.7	42.5	65.9	45.5
MoCov3 [11]	600††	54.8	82.8	46.8	47.2	66.9	50.8	41.1	63.6	44.1
DINO [7]	1600††	120.5	82.8	46.9	49.5	69.1	53.6	42.9	66.0	46.3
BEiT [3]	300	32.1	82.9	44.7	39.3	57.7	42.4	34.8	55.2	36.8
MAE [27]	300	16.4	82.7	46.1	47.2	65.8	51.3	41.1	62.9	44.4
MixMIM [38]	300	40.2	83.2	-	-	-	-	-	-	-
CIM-RevDet [22]	300	42.7	83.1	-	-	-	-	-	-	-
CIM-ResPix [22]	300	42.7	83.3	-	-	-	-	-	-	-
MixedAE	300	16.9	83.1$^{+0.4}$	47.0$^{+0.9}$	47.8$^{+0.6}$	66.6$^{+0.8}$	52.0$^{+0.7}$	41.6$^{+0.5}$	63.6$^{+0.7}$	45.0$^{+0.6}$
MixedAE-Full*	300	30.8	83.7$^{+1.0}$	47.4$^{+1.3}$	48.9$^{+1.7}$	67.6$^{+1.8}$	53.3$^{+2.0}$	42.5$^{+1.4}$	64.8$^{+1.9}$	45.9$^{+1.5}$
MixedAE-Full	300	62.3	**83.8**$^{+1.1}$	**48.9**$^{+2.8}$	**51.0**$^{+3.8}$	**69.7**$^{+3.9}$	**55.2**$^{+3.9}$	**44.1**$^{+3.0}$	**67.0**$^{+4.1}$	**47.9**$^{+3.5}$
BEiT [3]	800	85.5	83.2	45.6	40.8	59.4	44.1	36.0	56.8	38.2
MAE [27]	800	43.7	83.3	47.2	49.4	68.1	53.9	42.9	65.5	46.6
MixedAE	800	45.0	**83.5**$^{+0.2}$	**48.7**$^{+1.5}$	**50.3**$^{+0.9}$	**69.1**$^{+1.0}$	**54.8**$^{+0.9}$	**43.5**$^{+0.6}$	**66.2**$^{+0.7}$	**47.4**$^{+0.8}$
MAE [27]	1600	87.4	83.6	48.1	50.6	69.4	55.0	43.8	66.6	47.5
iBOT [63]	1600††	172.1	83.8	49.6	51.2	70.1	55.2	44.3	67.4	48.0
MixedAE	1600	90.1	**83.9**$^{+0.3}$	**49.8**$^{+1.7}$	**51.5**$^{+0.9}$	**70.2**$^{+0.8}$	**55.9**$^{+0.9}$	**44.5**$^{+0.7}$	**67.5**$^{+0.9}$	**48.2**$^{+0.7}$

Experiments

- Efficiency: exceed strong iBOT with a 2x acceleration

Method	Pre-train Epochs	Pre-train† GPU-days	ImageNet Top-1 Acc.	ADE20K mIoU	COCO AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
DeiT [52]	300	19.6	81.8	46.9	48.8	68.7	52.7	42.5	65.9	45.5
MoCov3 [11]	600††	54.8	82.8	46.8	47.2	66.9	50.8	41.1	63.6	44.1
DINO [7]	1600††	120.5	82.8	46.9	49.5	69.1	53.6	42.9	66.0	46.3
BEiT [3]	300	32.1	82.9	44.7	39.3	57.7	42.4	34.8	55.2	36.8
MAE [27]	300	16.4	82.7	46.1	47.2	65.8	51.3	41.1	62.9	44.4
MixMIM [38]	300	40.2	83.2	-	-	-	-	-	-	-
CIM-RevDet [22]	300	42.7	83.1	-	-	-	-	-	-	-
CIM-ResPix [22]	300	42.7	83.3	-	-	-	-	-	-	-
MixedAE	300	16.9	83.1$^{+0.4}$	47.0$^{+0.9}$	47.8$^{+0.6}$	66.6$^{+0.8}$	52.0$^{+0.7}$	41.6$^{+0.5}$	63.6$^{+0.7}$	45.0$^{+0.6}$
MixedAE-Full*	300	30.8	83.7$^{+1.0}$	47.4$^{+1.3}$	48.9$^{+1.7}$	67.6$^{+1.8}$	53.3$^{+2.0}$	42.5$^{+1.4}$	64.8$^{+1.9}$	45.9$^{+1.5}$
MixedAE-Full	300	62.3	**83.8**$^{+1.1}$	**48.9**$^{+2.8}$	**51.0**$^{+3.8}$	**69.7**$^{+3.9}$	**55.2**$^{+3.9}$	**44.1**$^{+3.0}$	**67.0**$^{+4.1}$	**47.9**$^{+3.5}$
BEiT [3]	800	85.5	83.2	45.6	40.8	59.4	44.1	36.0	56.8	38.2
MAE [27]	800	43.7	83.3	47.2	49.4	68.1	53.9	42.9	65.5	46.6
MixedAE	800	45.0	**83.5**$^{+0.2}$	**48.7**$^{+1.5}$	**50.3**$^{+0.9}$	**69.1**$^{+1.0}$	**54.8**$^{+0.9}$	**43.5**$^{+0.6}$	**66.2**$^{+0.7}$	**47.4**$^{+0.8}$
MAE [27]	1600	87.4	83.6	48.1	50.6	69.4	55.0	43.8	66.6	47.5
iBOT [63]	1600††	172.1	83.8	49.6	51.2	70.1	55.2	44.3	67.4	48.0
MixedAE	1600	90.1	**83.9**$^{+0.3}$	**49.8**$^{+1.7}$	**51.5**$^{+0.9}$	**70.2**$^{+0.8}$	**55.9**$^{+0.9}$	**44.5**$^{+0.7}$	**67.5**$^{+0.9}$	**48.2**$^{+0.7}$

Experiments

- Object-aware pre-training: better on dense perception tasks

Method	Pre-train Epochs	Pre-train† GPU-days	ImageNet Top-1 Acc.	ADE20K mIoU	COCO AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
DeiT [52]	300	19.6	81.8	46.9	48.8	68.7	52.7	42.5	65.9	45.5
MoCov3 [11]	600††	54.8	82.8	46.8	47.2	66.9	50.8	41.1	63.6	44.1
DINO [7]	1600††	120.5	82.8	46.9	49.5	69.1	53.6	42.9	66.0	46.3
BEiT [3]	300	32.1	82.9	44.7	39.3	57.7	42.4	34.8	55.2	36.8
MAE [27]	300	16.4	82.7	46.1	47.2	65.8	51.3	41.1	62.9	44.4
MixMIM [38]	300	40.2	83.2	-	-	-	-	-	-	-
CIM-RevDet [22]	300	42.7	83.1	-	-	-	-	-	-	-
CIM-ResPix [22]	300	42.7	83.3	-	-	-	-	-	-	-
MixedAE	300	16.9	$83.1^{+0.4}$	$47.0^{+0.9}$	$47.8^{+0.6}$	$66.6^{+0.8}$	$52.0^{+0.7}$	$41.6^{+0.5}$	$63.6^{+0.7}$	$45.0^{+0.6}$
MixedAE-Full*	300	30.8	$83.7^{+1.0}$	$47.4^{+1.3}$	$48.9^{+1.7}$	$67.6^{+1.8}$	$53.3^{+2.0}$	$42.5^{+1.4}$	$64.8^{+1.9}$	$45.9^{+1.5}$
MixedAE-Full	300	62.3	**83.8**$^{+1.1}$	**48.9**$^{+2.8}$	**51.0**$^{+3.8}$	**69.7**$^{+3.9}$	**55.2**$^{+3.9}$	**44.1**$^{+3.0}$	**67.0**$^{+4.1}$	**47.9**$^{+3.5}$
BEiT [3]	800	85.5	83.2	45.6	40.8	59.4	44.1	36.0	56.8	38.2
MAE [27]	800	43.7	83.3	47.2	49.4	68.1	53.9	42.9	65.5	46.6
MixedAE	800	45.0	**83.5**$^{+0.2}$	**48.7**$^{+1.5}$	**50.3**$^{+0.9}$	**69.1**$^{+1.0}$	**54.8**$^{+0.9}$	**43.5**$^{+0.6}$	**66.2**$^{+0.7}$	**47.4**$^{+0.8}$
MAE [27]	1600	87.4	83.6	48.1	50.6	69.4	55.0	43.8	66.6	47.5
iBOT [63]	1600††	172.1	83.8	49.6	51.2	70.1	55.2	44.3	67.4	48.0
MixedAE	1600	90.1	**83.9**$^{+0.3}$	**49.8**$^{+1.7}$	**51.5**$^{+0.9}$	**70.2**$^{+0.8}$	**55.9**$^{+0.9}$	**44.5**$^{+0.7}$	**67.5**$^{+0.9}$	**48.2**$^{+0.7}$

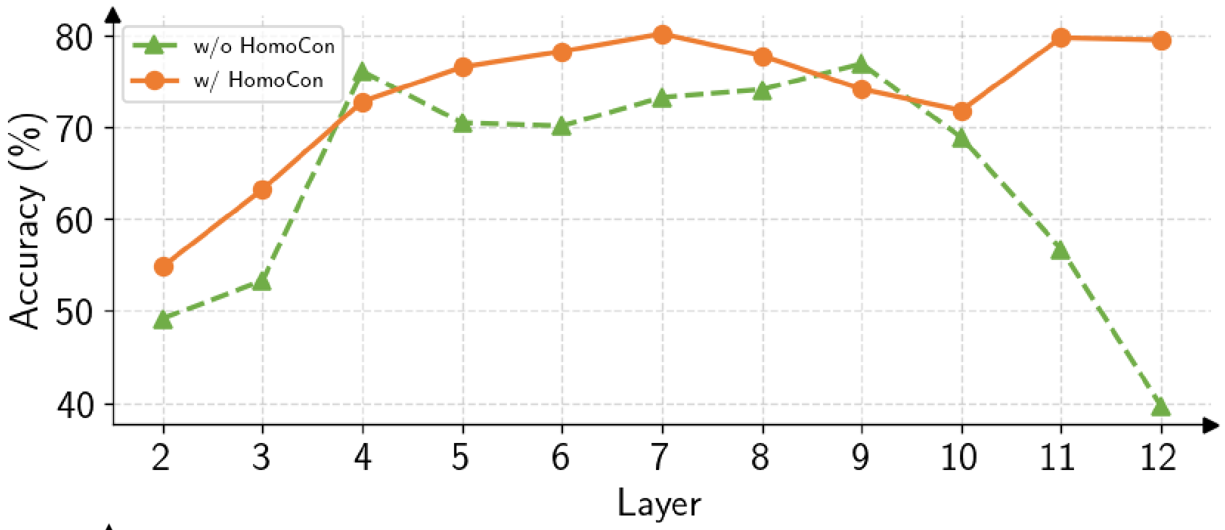
Experiments

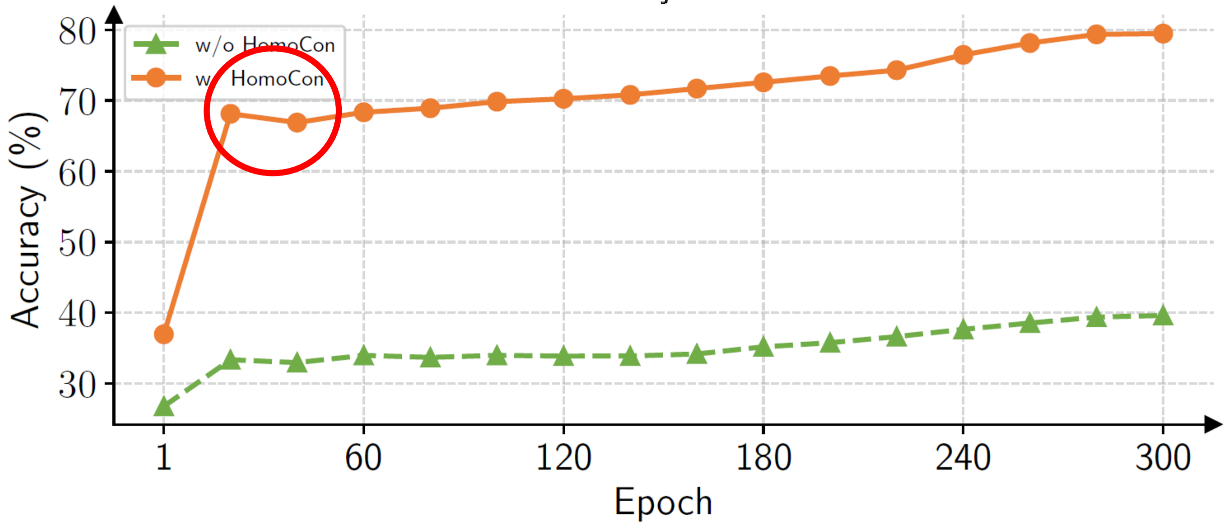
- Transferability

Method	Aircraft	Caltech	Cars	C10	C100	DTD	Flowers	Food	Pets	SUN	VOC	Avg.
SSL ResNets												
MoCov2 [10]	79.9	84.4	75.2	96.5	71.3	69.5	94.4	76.8	79.8	55.8	71.7	77.7
SimCLR [9]	78.7	82.9	79.8	96.2	79.1	70.2	94.3	82.2	83.2	61.1	78.2	80.5
BYOL [25]	79.5	89.4	84.6	97.0	84.0	73.6	94.5	85.5	89.6	64.0	82.7	84.0
SwAV [6]	**83.1**	89.9	86.8	96.8	84.4	75.2	95.5	87.2	89.1	66.2	84.7	85.3
SDR [40]	82.6	89.0	87.5	97.4	84.4	75.6	97.0	86.1	89.3	66.1	85.3	85.5
SSL Transformers												
MoCov3 [11]	76.6	91.2	86.6	98.3	88.3	72.6	95.5	86.4	92.0	65.6	84.5	85.2
DINO [7]	69.4	91.2	81.3	**98.4**	**88.9**	77.6	96.9	87.3	93.5	64.7	86.3	85.1
BEiT [3]	66.3	80.2	78.6	96.1	80.0	69.9	92.9	83.2	85.3	57.1	76.7	78.7
MAE [27]	78.2	91.2	88.4	97.0	82.5	75.3	96.6	84.7	92.6	65.4	86.0	85.3
MixedAE	82.1	**91.5**	**88.8**	97.9	85.9	**78.7**	**97.1**	**87.4**	**93.6**	**66.2**	**86.4**	**86.9**$^{+1.6}$

Analysis

- Effect of Homo Recognition
  - Fluctuate w/o Homo contrastive
  - Stable and fast convergence w/ Homo contrastive
  - However, still far from 100%, suggesting potential future improvement space





Thank you!

Paper Wechat Post

Paper: https://arxiv.org/abs/2303.17152

Also check our series of works on efficient SSL: [link]

Welcome to join our poster session (@THU-PM-204)!