Masked AutoEncoders

DongHu Kim

2023. 11. 16

Origin

Masked Image Modeling (MIM)

- Reconstruct(inpaint) masked parts of given image

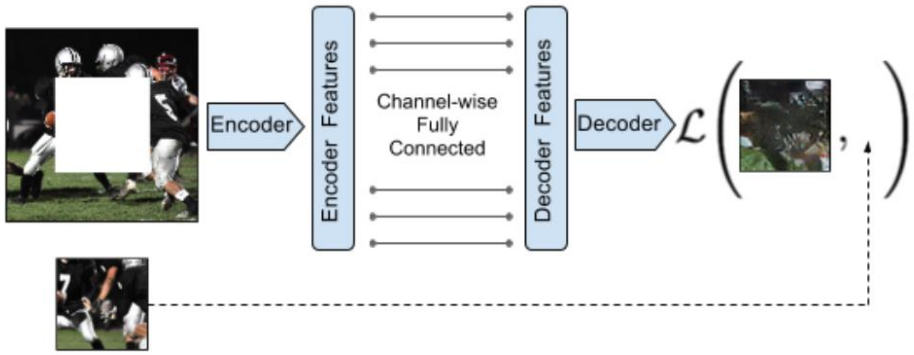
Input

Target

Origin

Early work

| Context Encoders: Using autoencoders to inpaint a large, centered hole



(a) Input context (c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

[1] Context Encoders: Feature Learning by Inpainting, Pathak et al.

Origin

When transformers were invented... (MLM, BERT)

- BEiT: Direct application of BERT.
  Image patches tokenized by DALL-E tokenizer, train BERT on those tokens.

When transformers got extended to images... (ViT)

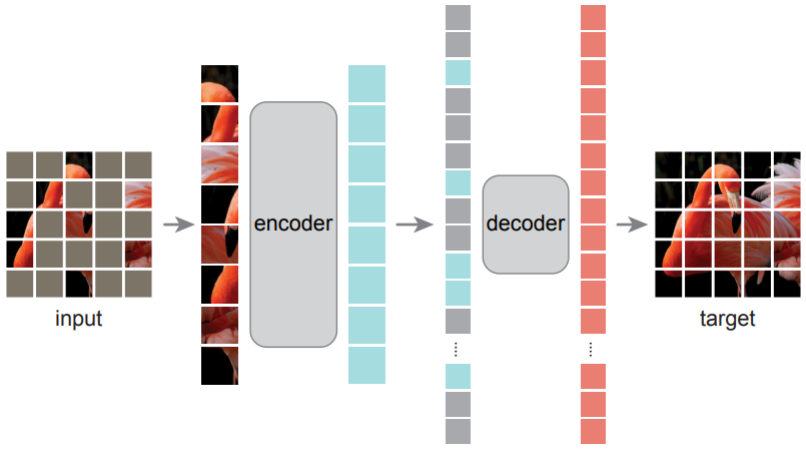
- SimMIM, MAE: ViT encoder + Masking + Decoder
- Recent researches are more focused on MAE's, creating a LOT of variants!
- Theoretical analysis on MAE has just begun.

Contents

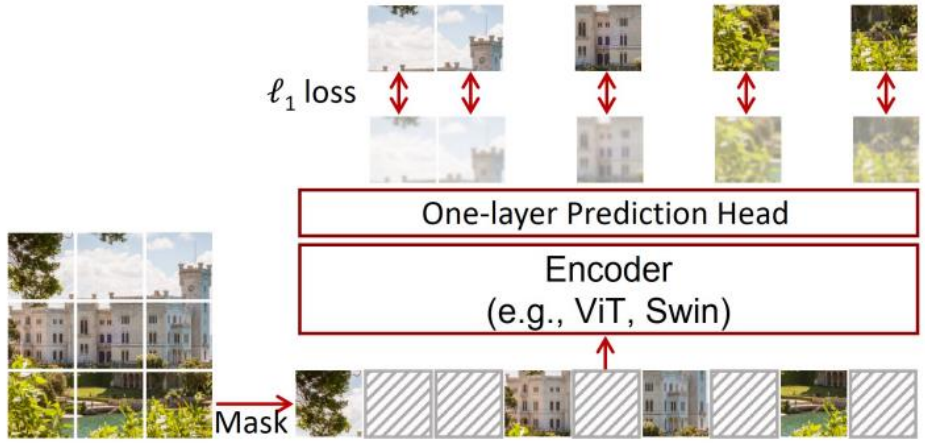
MAE (& SimMIM)

Self-supervised pretraining strategy

- Random patch masking – ViT encoder – Unmasking – Transformer decoder – Prediction
- Use encoder on downstream tasks (without masking, of course).
- SimMIM's only difference is that masked patches are also passed through encoder, while MAE's encoder only gets non-masked patches.



MAE



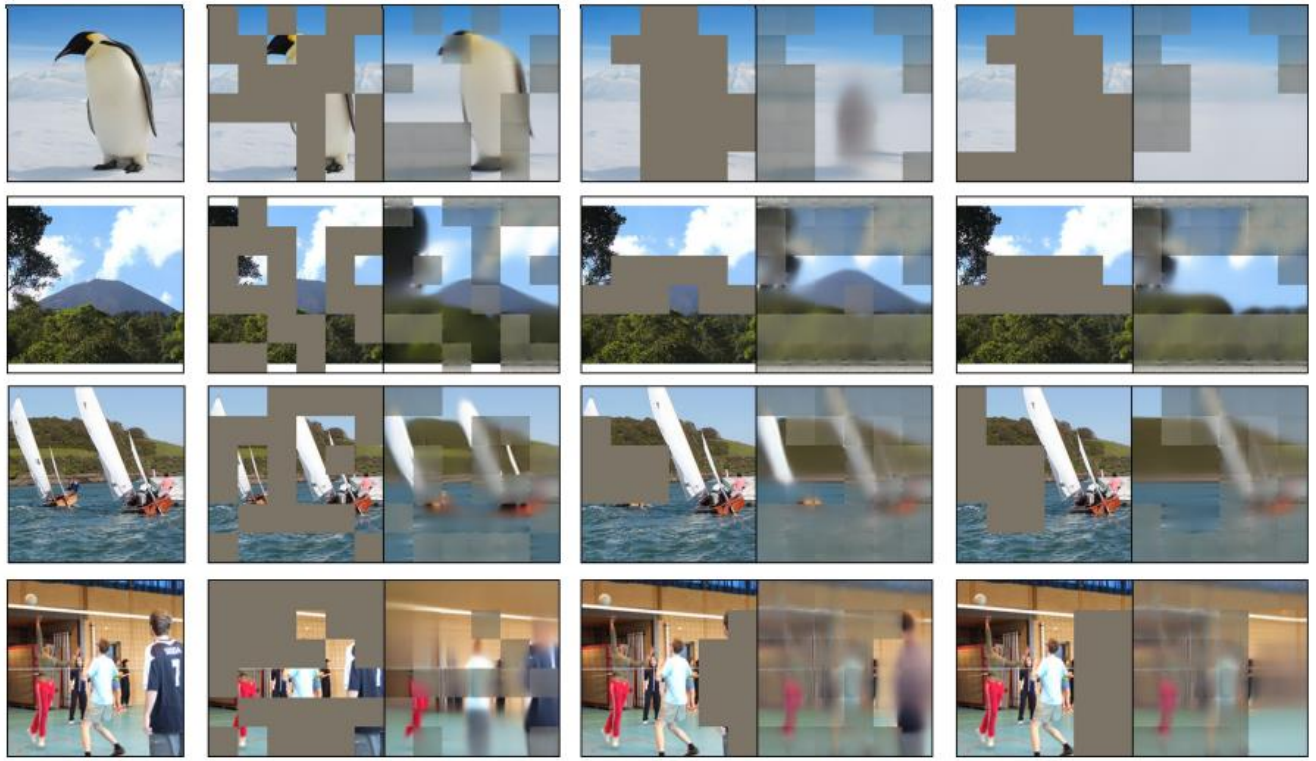
SimMIM

MAE (& SimMIM)

Great job!

MAE

Good job!



SimMIM

MAE

MAE doesn't need heavy, specific augmentations

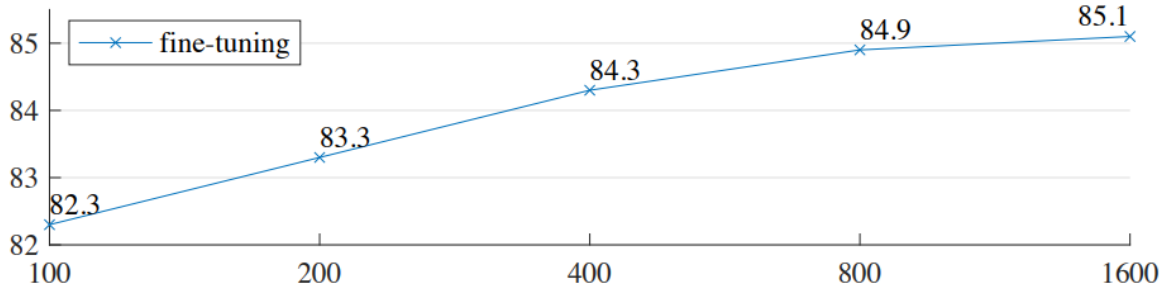
- 50 epochs finetuning after MAE pretraining is better than 200 epochs from scratch with heavy augmentation (ImageNET)

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.

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

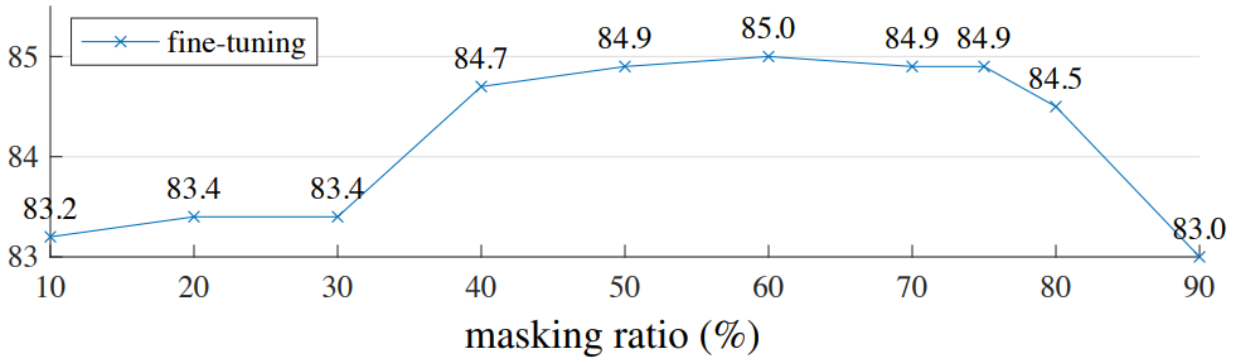
MAE benefits from longer training lengths



MAE

MAE benefits from high masking ratio (0.75)

- Unlike words in sentences, pixels in images are mostly redundant.
- Heavy mask ratio prevents MAE from copy-pasting (extrapolating) nearby pixels.
- No special masking technique is required (random is enough).



case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

MAE

Decoder should be reasonably large, but not too large.

- Aligns with our intuition that later parts of a model learns task-specific features, and earlier parts learn general ones.
- SimMIM also shows that a simple linear decoder works better than more complex ones.

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

Head	#params	Training costs	Top-1 acc (%)
Linear	89.9M	1×	82.8
2-layer MLP	90.9M	1.2×	82.8
inverse Swin-T	115.2M	1.7×	82.4
inverse Swin-B	174.8M	2.3×	82.5

Table 2. Ablation on different prediction heads. A simple linear layer performs the best with lower training costs.

SimMIM

MAE

Giving mask tokens to encoder is BAD.

- 1. Doesn't align with downstream performance - mask tokens are never seen there!
- 2. Encoder pretraining becomes much heavier (4x more tokens)
- Uh-oh moment for SimMIM

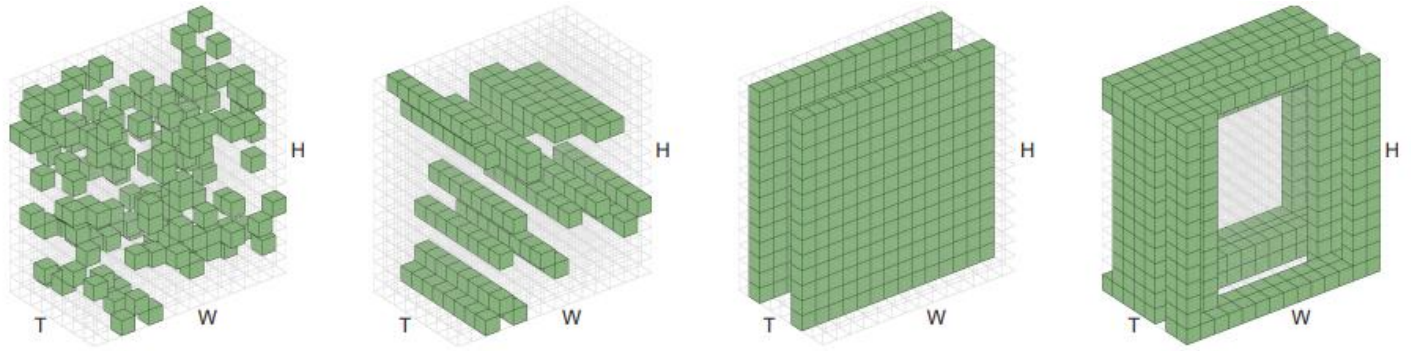
case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

MAE + Temporality

How can we train MAE's on videos?

- Videos not only have even more pixel redundancy, but also adds temporal redundancy.
- MAE-ST: Just use higher mask ratio (0.9).
- VideoMAE: Higher mask ratio(0.9) + Tube masking + Spatio-temporal position embedding



(a) agnostic, 90% (b) space-only, 90% (c) time-only, 75% (d) block, 75%

(random) (tube) (frame)

[2] Masked Autoencoders As Spatiotemporal Learners, Feichtenhofer et. Al
[3] VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, Tong et. al

MAE + Temporality

Results

- Two disagree on which masking strategy is better but agree on using 0.9 mask ratio.
- Other results consistent with those from Image MAE (just extended to video datasets). e.g., pretrain length, decoder 4-layer-512-width, minimal augmentation

MAE-ST

	case	ratio	acc.
(random)	agnostic	90	**84.4**
(frame)	space-only	90	83.5
(tube)	time-only	75	79.1
	block	75	83.2

(a) **Mask sampling**. See also Fig. 4. Random sampling that is spacetime-*agnostic* works the best.

VideoMAE

case	ratio	SSV2	K400
tube	75	68.0	79.8
tube	90	**69.6**	**80.0**
random	90	68.3	79.5
frame	87.5*	61.5	76.5

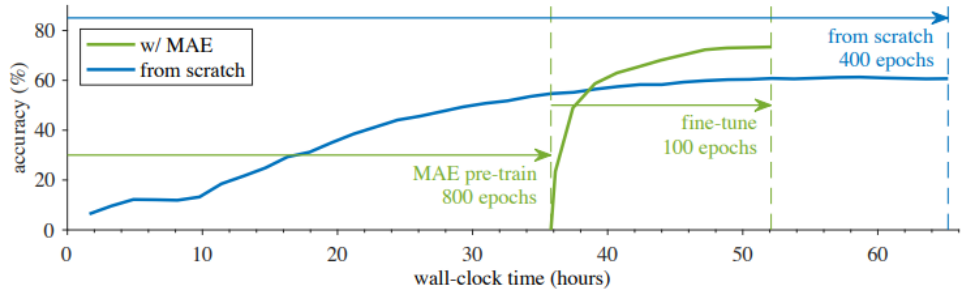
(b) **Mask sampling**. We compare different masking strategies. Our proposed tube masking with an extremely high ratio works the best. ∗"87.5" means masking 14/16 frames.

MAE + Temporality

Results

- MAE-ST got better performance overall.

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**



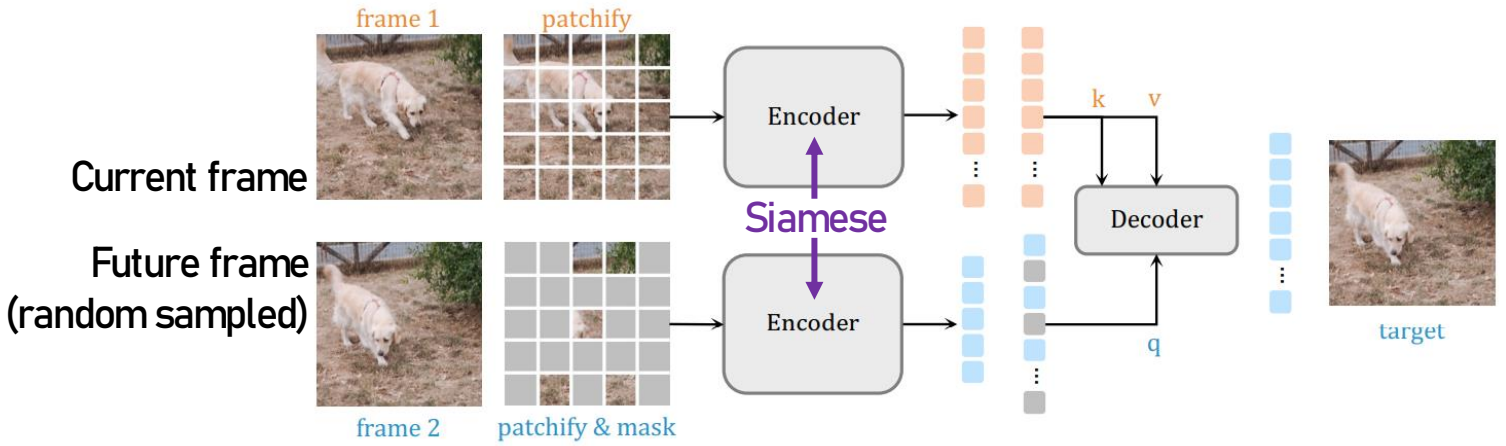
	scratch	MAE
1-view	60.7	**73.4** (+12.7)
multi-view	71.4	**84.4** (+13.0)

MAE-ST
(on Kinetic-400)

MAE + Temporality

SiamMAE: Learning future predictive representation using MAE

- Use Siamese architecture to learn to propagate information through time!
- No masking on current frame, extremely heavy masking (0.95) on future frame.
- To reconstruct future frame, the masked tokens(query) must use current frame's information(key, value) at decoder stage.



[4] Siamese Masked Autoencoders, Gupta et. al

MAE + Temporality

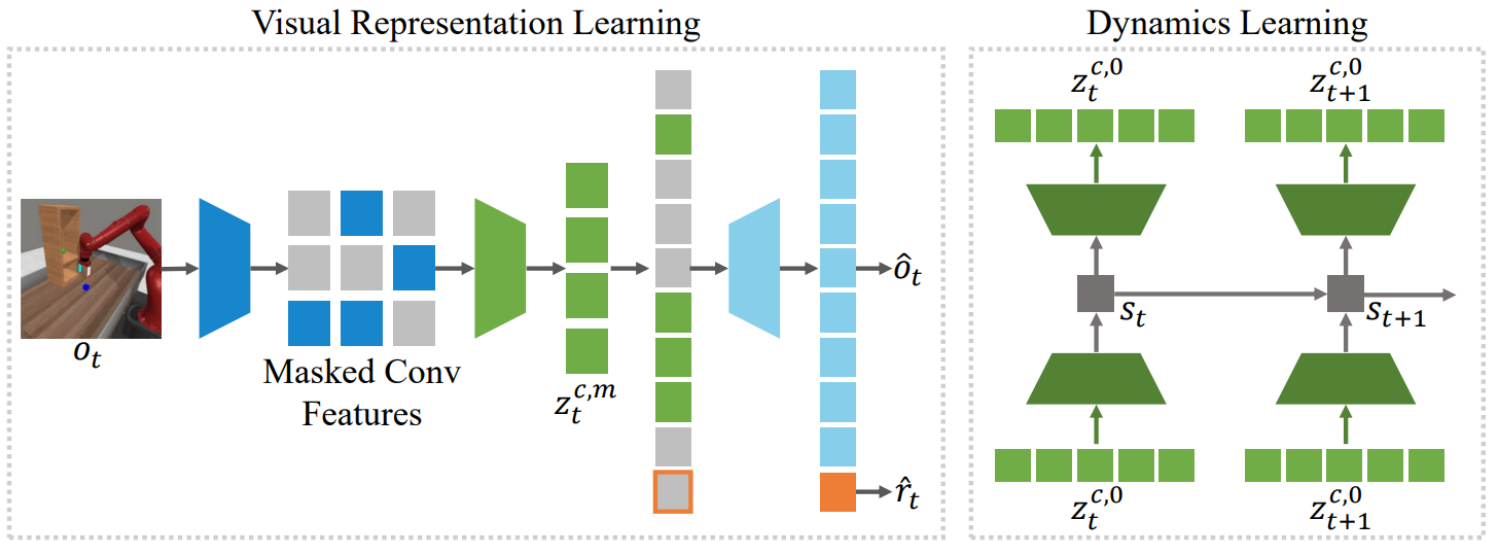
Results

Figure 2: **Visualizations** on the Kinetics-400 [93] validation set (masking ratio 90%). For each video sequence, we sample a clip of 8 frames with a frame gap of 4 and show the original video (top), SiamMAE output (middle), and masked future frames (bottom). Reconstructions are shown with f_1 as the first frame of the video clip and f_2 as the remaining frames, using a SiamMAE pre-trained ViT-S/8 encoder with a masking ratio of 95%.

MAE + Control (MWM)

Learning a world model using MAE's latent space

- DreamerV2, but the representation model(and reward model) is learned via MAE.
- Interestingly, masking is performed CNN outputs. Each token is no longer limited to a patch but contains all information of its receptive field.

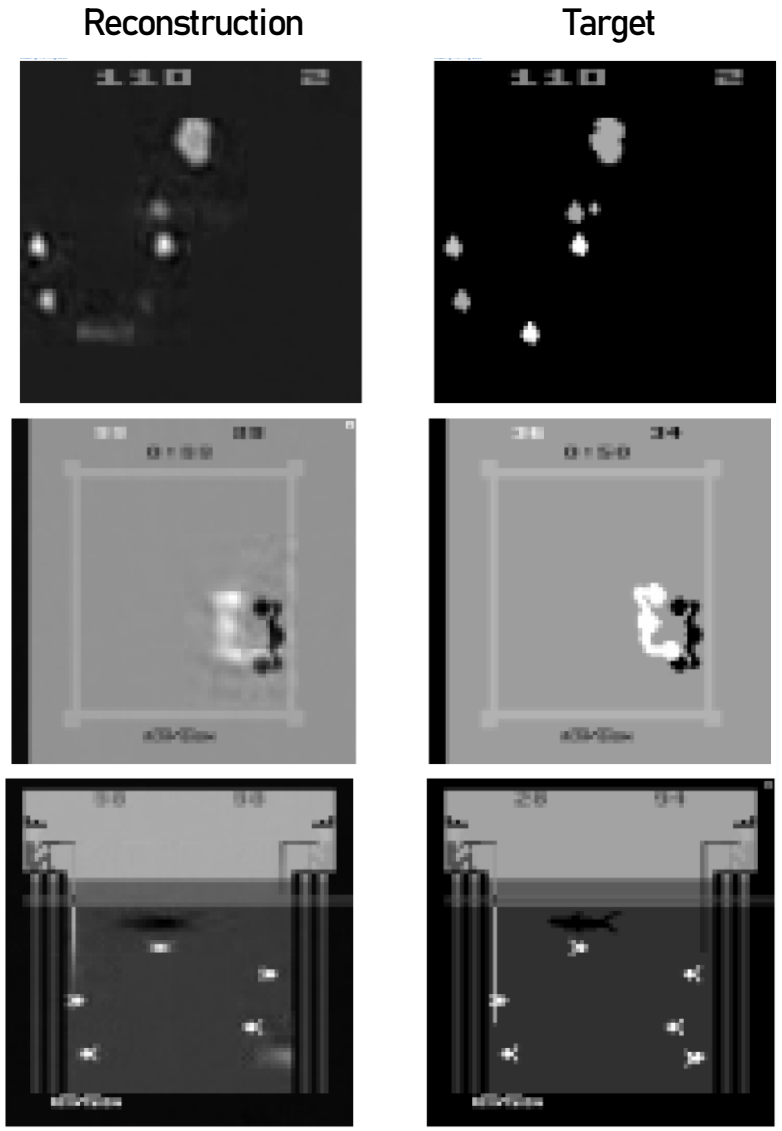


[5] Masked World Models for Visual Control, Seo et. al

Our Implementation

MWM on Atari

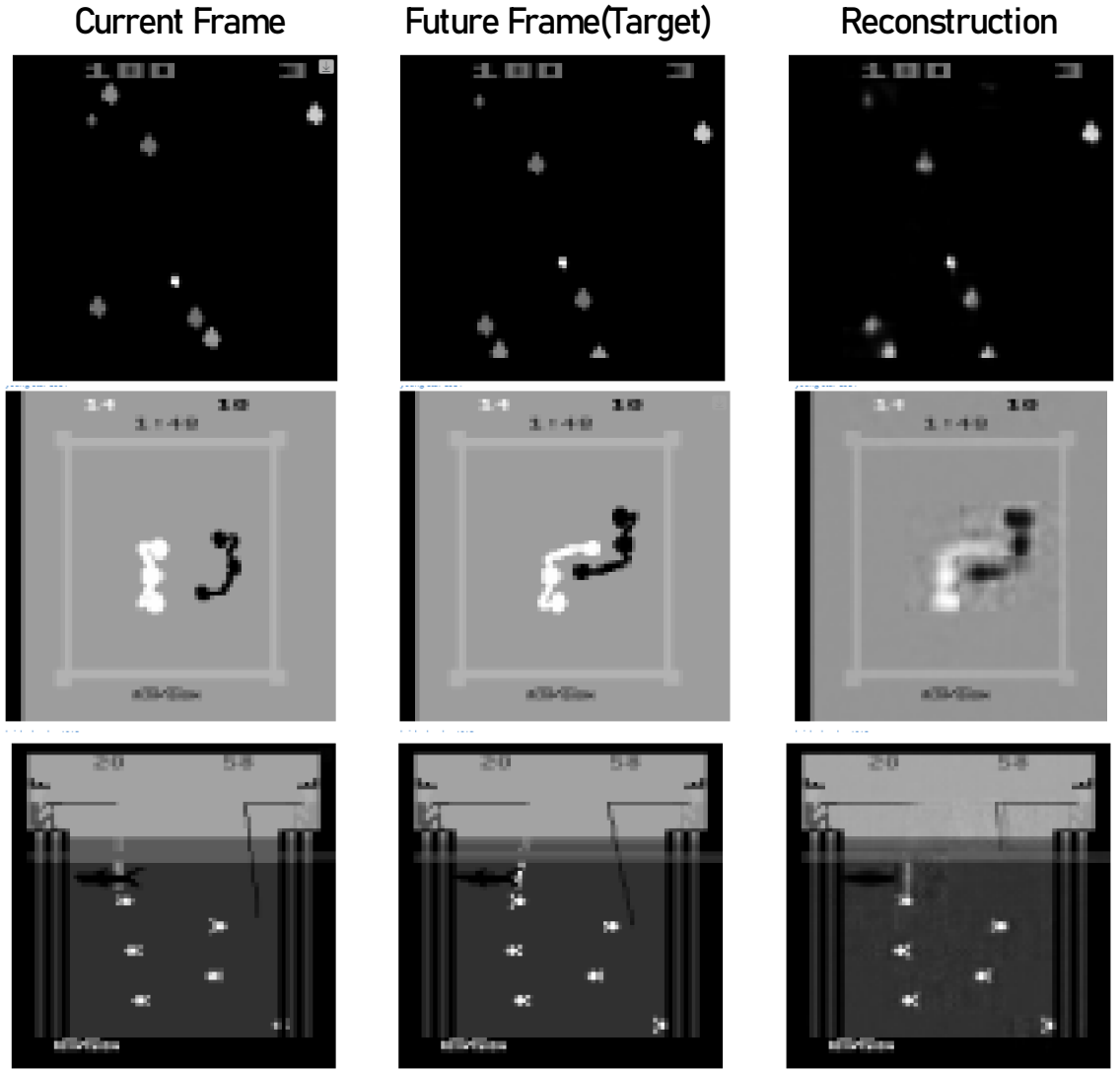
- Learned to reconstruct only (no world model).
- Fails on small, random objects.
- Intuitively, it's almost impossible to reconstruct a tiny object that's completely masked out.
- Empirically, MWM performs better than CURL on games with large objects, but deteriorates when objects get smaller.



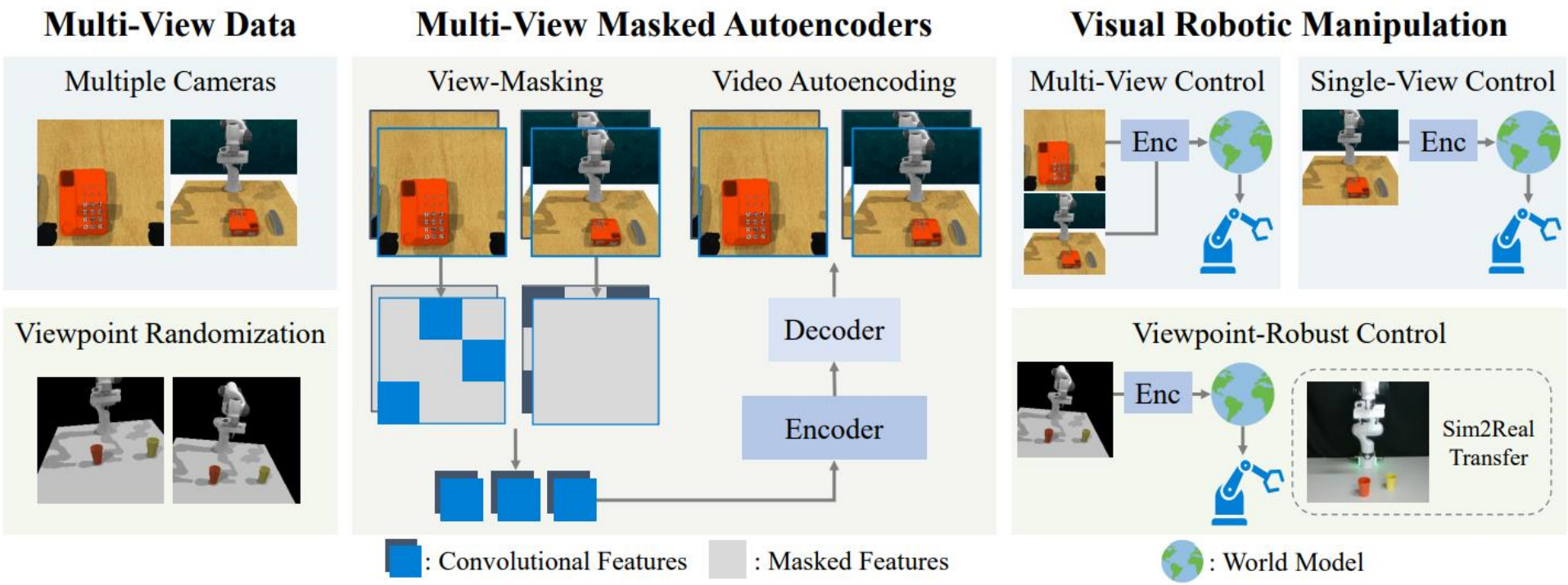
Our Implementation

SiamMAE on Atari

- Implemented in 'MWM style' (i.e., CNN output masking)
- Somewhat cherry-picked :)

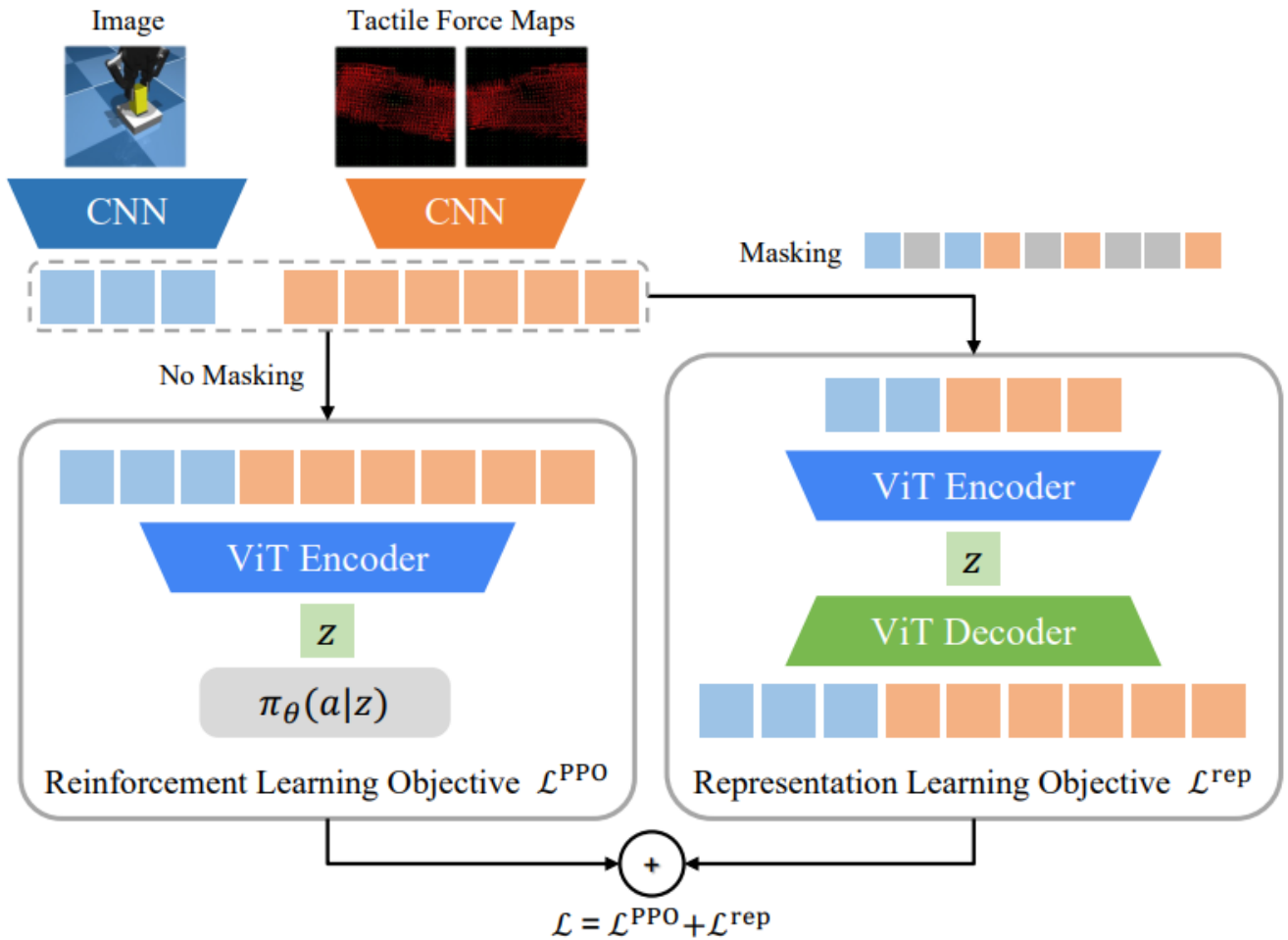


MWM + Multi-View



[6] Multi-View Masked World Models for Visual Robotic Manipulation, Seo et. al

MWM + Multi-Modality



[7] The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning, Sferrazza, Seo et. al

Empirical Takeaways

Why should I consider using MAE?

- Effective, efficient, requires minimal augmentation.
- Transformer architecture allows easy extension to multi-modality.

Which masking strategy should I use?

- Random.
- If temporality is involved, consider tube masking.

What masking ratio should I use?

- Start with 0.75 for images, 0.9 if temporality is involved.

How big should encoder/decoder be?

- Encoder: Larger the better (tradeoff).
- Decoder: Reasonably large, but not too large (4 layer, 512 width transformer).

Why Does MAE Work?

Theoretical works on MAE have just begun

- Some try to understand MAE via hierarchical latent variable models.
- Others try to connect MAE with contrastive learning.
- One of which is by Zhang et. al, which states that there's an implicit contrastive loss that lower-bounds MAE loss.
- Take this with a grain of salt, though.

[9] Understanding Masked Autoencoders via Hierarchical Latent Variable Models, Kong et. al
[10] Understanding Masked Autoencoders From a Local Contrastive Perspective, Yue et. al
[11] How Mask Matters: Towards Theoretical Understandings of Masked Autoencoders, Zhang et. al

Why Does MAE Work?

"MAE implicitly creates an augmentation graph on masked images"

- Start by creating a bipartite graph between masked & unmasked images
- Step 1. MAE loss is lower-bounded by the masked-unmasked alignment loss.
- Gross simplification: L2 loss is lower-bounded by dot product loss.
- Interpretation: MAE is aligning network output with its unmasked target.

Theorem 3.2. *Under Assumption 3.1, the MAE loss can be lower bounded by*

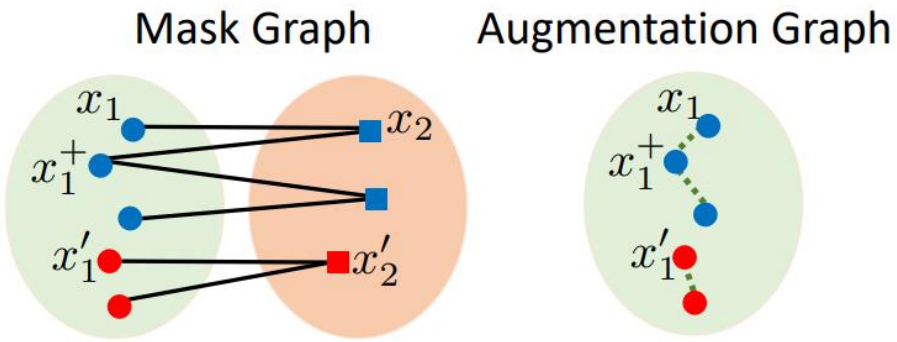
$$\mathcal{L}_{MAE}(h) \geq \mathcal{L}_{asym}(h) - \varepsilon + const, \tag{3}$$

$$\text{and } \mathcal{L}_{asym}(h) = -\mathbb{E}_{x_1,x_2} h(x_1)^\top h_g(x_2) = -\operatorname{tr}(H_g^\top \bar{A}_M H), \tag{4}$$

Why Does MAE Work?

"MAE implicitly creates an augmentation graph on masked images"

- Two masked images that have the same unmasked target should also get closer. This should occur a lot more as masking ratio increases.
- We can think this pair as the two views of some (unknown) augmentation function, on which we apply contrastive alignment loss.
- Also note that any augmentation function induces a graph among input images, of which two nodes are connected when they can be two views of that augmentation function,



Why Does MAE Work?

"MAE implicitly creates an augmentation graph on masked images"

| Step 2. The masked-unmasked alignment loss is lower-bounded by the alignment loss on the implicit augmentation graph.

Theorem 3.3. *The asymmetric alignment loss on the mask graph (Eq. (4)) can be lower bounded by the symmetric alignment loss on the augmentation graph (Eq. (5)):*

$$\mathcal{L}_{asym}(h) \geq \frac{1}{2}\mathcal{L}_{align}(h) + const. \tag{6}$$

| ~~Step 3. Have a debate with the author on GitHub.~~

Why Does MAE Work?

Can two inputs share the same target in the first place?

- Two masked images sharing the *exact* same target?
- Extreme cases: No masking & All masking
- My interpretation (verified by silence):
  If two masked images are from the same class, their targets are likely to be semantically similar, and that (hopefully) still creates implicit connections. This makes their argument weaker, which leads to next concern.

Is the lower-bound meaningfully tight?

- Unanswered (according to area chair)

Why Does MAE Work?

Theoretically explaining why mask ratio matters

- Based on this, authors provide a theoretical guarantee on downstream classification.
- This can be summarized into: "Mask ratio should be high enough to create enough intra-class connections, but not too much to create inter-class connections."

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:*

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

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

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

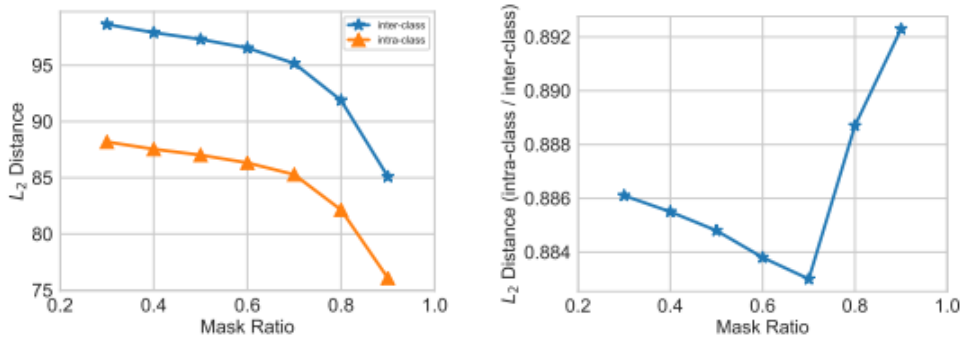
$$\forall\, h \in \mathcal{H}, \quad \mathcal{L}_{\text{U-MAE}}(h) \geq \frac{1}{4L}\sum_{i=k+1}^{N_1} \lambda_i^2 - \varepsilon + const, \tag{15}$$

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

Why Does MAE Work?

The sweet spot of mask ratio

- We can roughly measure inter/intra class connectivity via the relative L2 distance.
- Intuitively, we want low inter-class distance and high intra-class distance.
- On ImageNet, relative L2 distance hits minimum around 0.75!



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

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

Fin