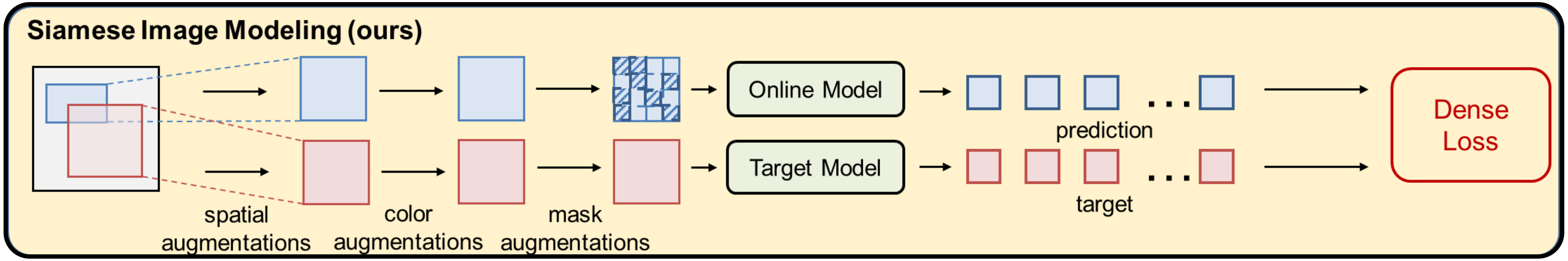
Siamese Image Modeling for Self-Supervised Vision Representation Learning

Chenxin Tao[1*], Xizhou Zhu[2*], Weijie Su[3*],
Gao Huang[1], Bin Li[3], Jie Zhou[1], Yu Qiao[4], Xiaogang Wang[5], Jifeng Dai[1,4]

[1]Tsinghua University, [2]SenseTime Research, [3]University of Science and Technology of China,
[4]Shanghai Artificial Intelligence Laboratory, [5]The Chinese University of Hong Kong

TUE-AM-204

Contribution



- A new form of self-supervised learning that can learn **semantic alignment** and **spatial sensitivity** with a single dense loss.

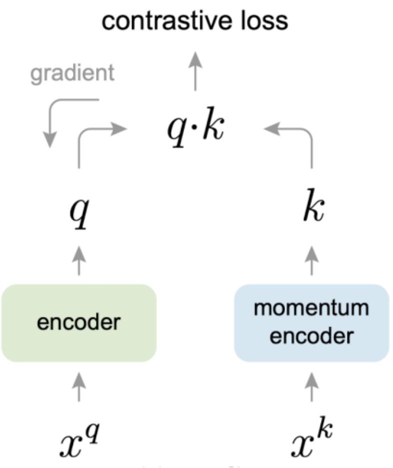
- Compared with MIM methods, reconstructing another view helps to obtain good semantic alignment.

- Compared with ID methods, matching the dense correspondence between two views can help to learn spatial sensitivity.

- SiameseIM is able to surpass both MIM and ID methods over a wide range of tasks. SiameseIM obtains more improvements in few-shot, long-tail and robustness-concerned scenarios.

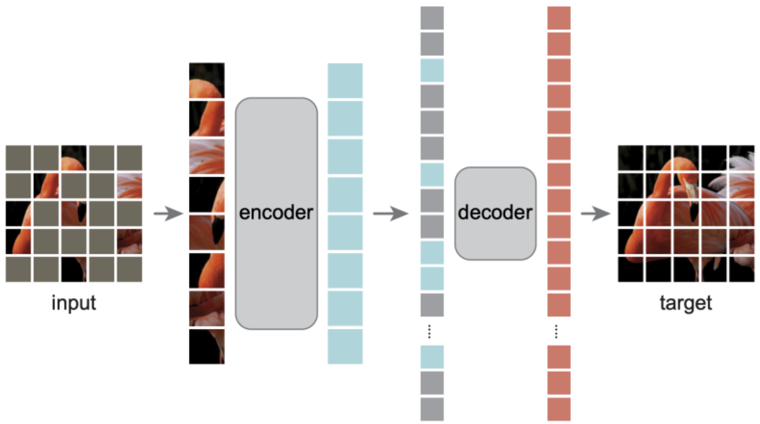
Motivation – Two Mainstream SSL Methods

Instance Discrimination



- lack **spatial sensitivity**:
  modeling the local structure within an image

Masked Image Modeling

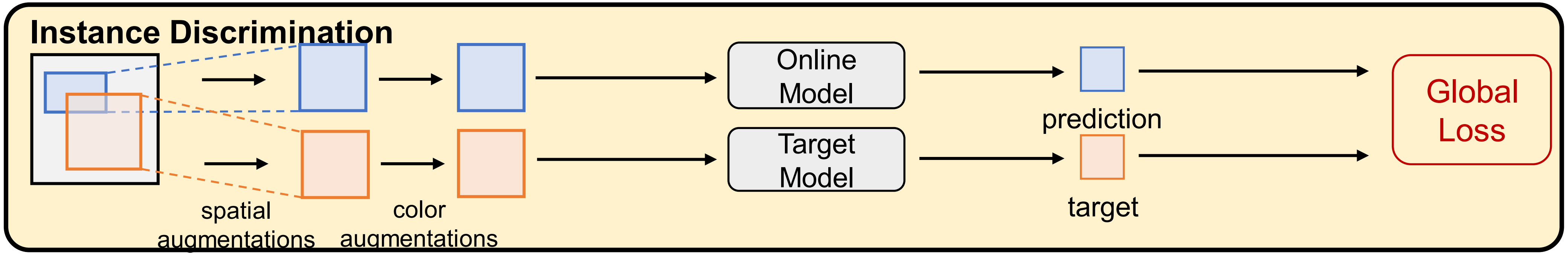


- lack **semantic alignment**:
  projecting semantically similar views into nearby representations

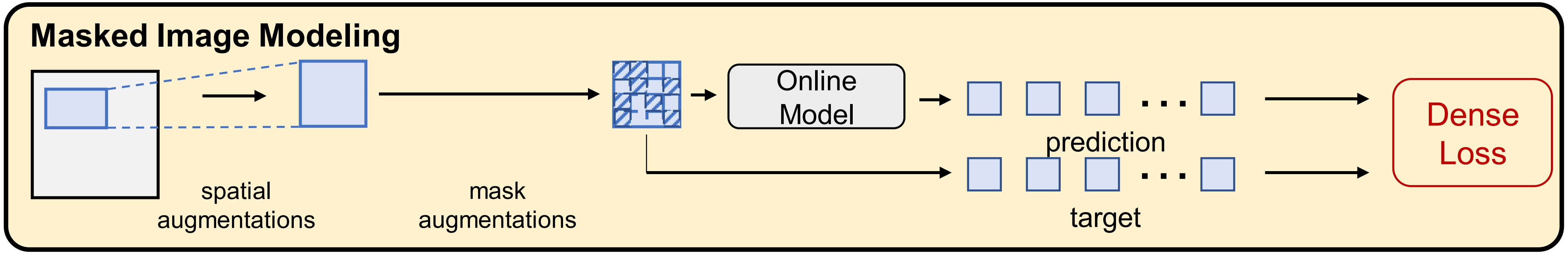
Typical Work

Characteristics

Motivation – Two Mainstream SSL Methods



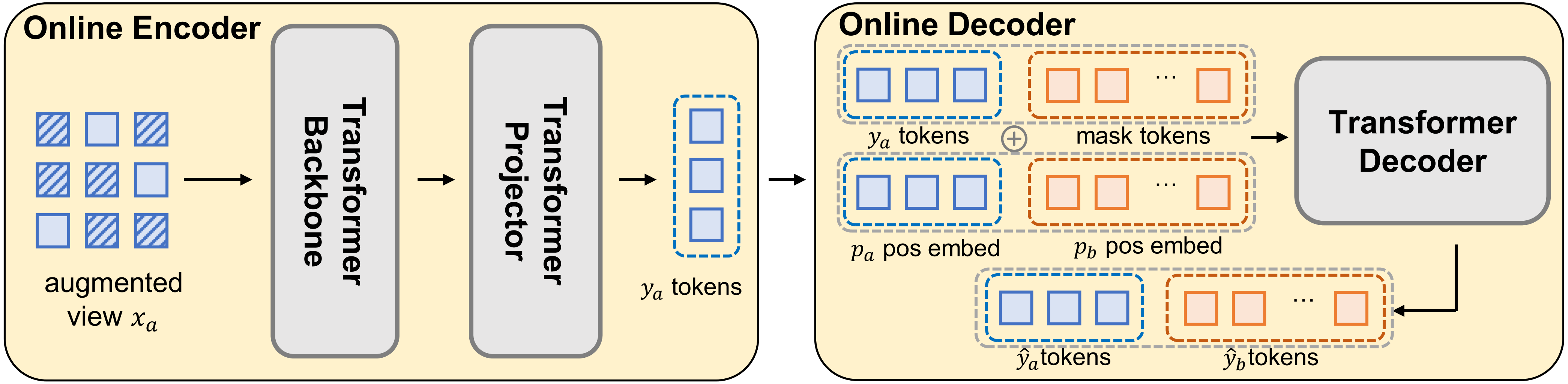
semantic alignment can be achieved by matching different augmented view from the same image



spatial sensitivity can be achieved by predicting dense representations from masked images

- We propose Siamese Image Modeling, which reconstructs the dense representations of an augmented view, based on another augmented view from the same image.

Siamese Image Modeling



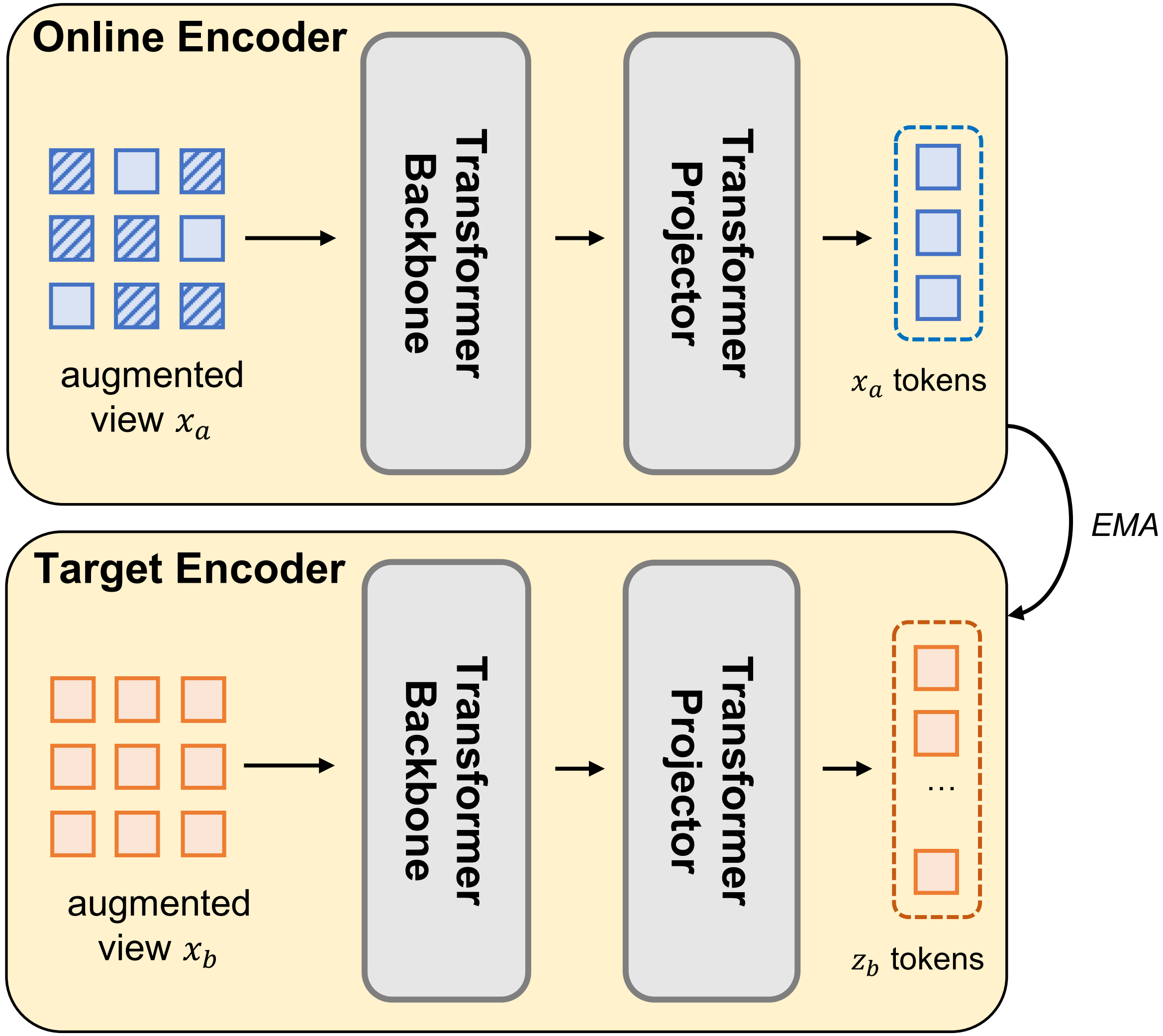
- Augmentation: spatial + color augmentations for both views, mask augmentation for online view

- Online branch makes the prediction:

$$y_b = g\left(Concat\left(y_a + p_a, \left\{m + p_b^{(u,v)}\right\}_{u=1,v=1}^{N_h,N_w}\right)\right)$$

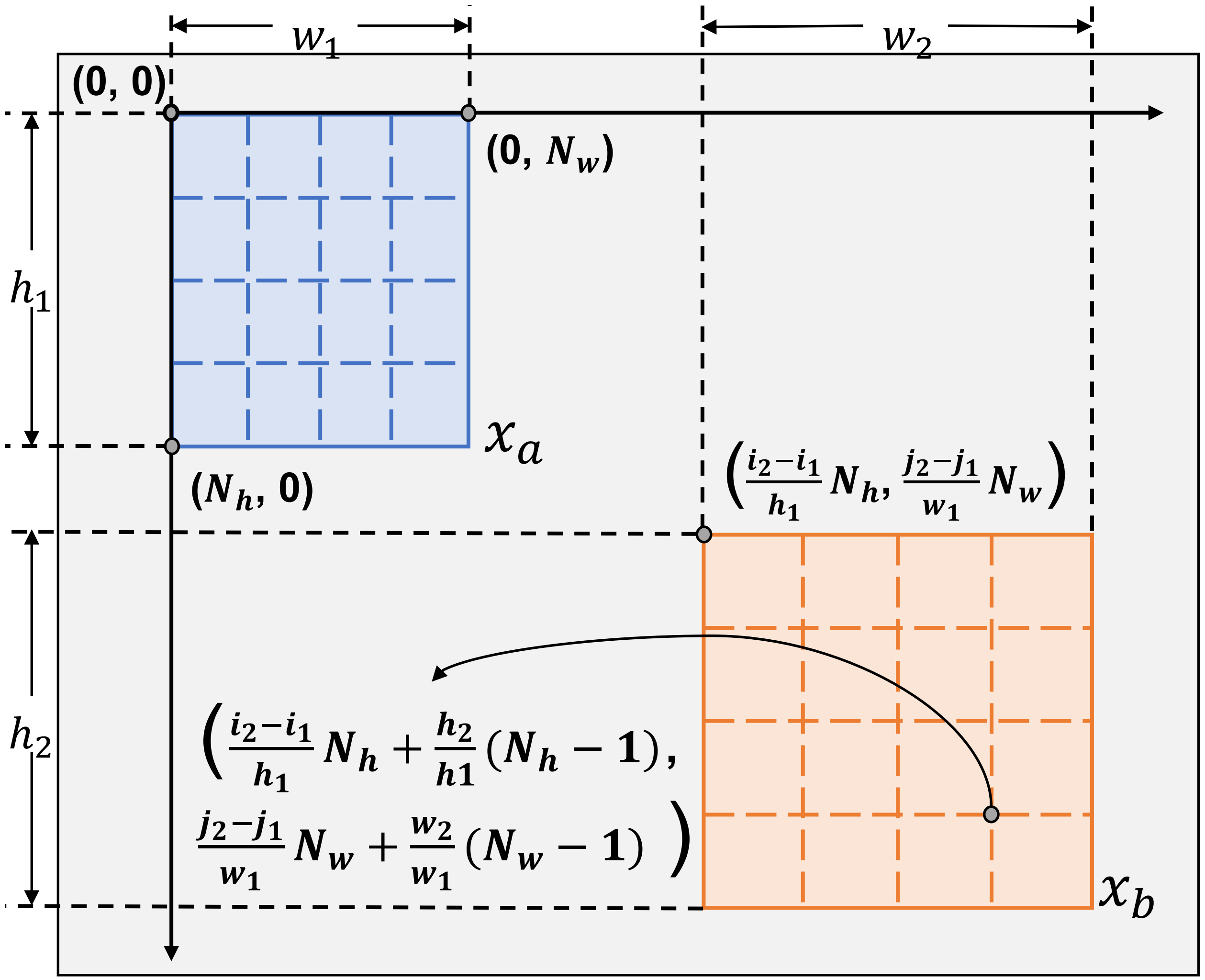
- m indicates mask token, p_a, p_b are the positional embedding of view x_a and x_b

Siamese Image Modeling



- Target branch computes the target representation z_b

Siamese Image Modeling

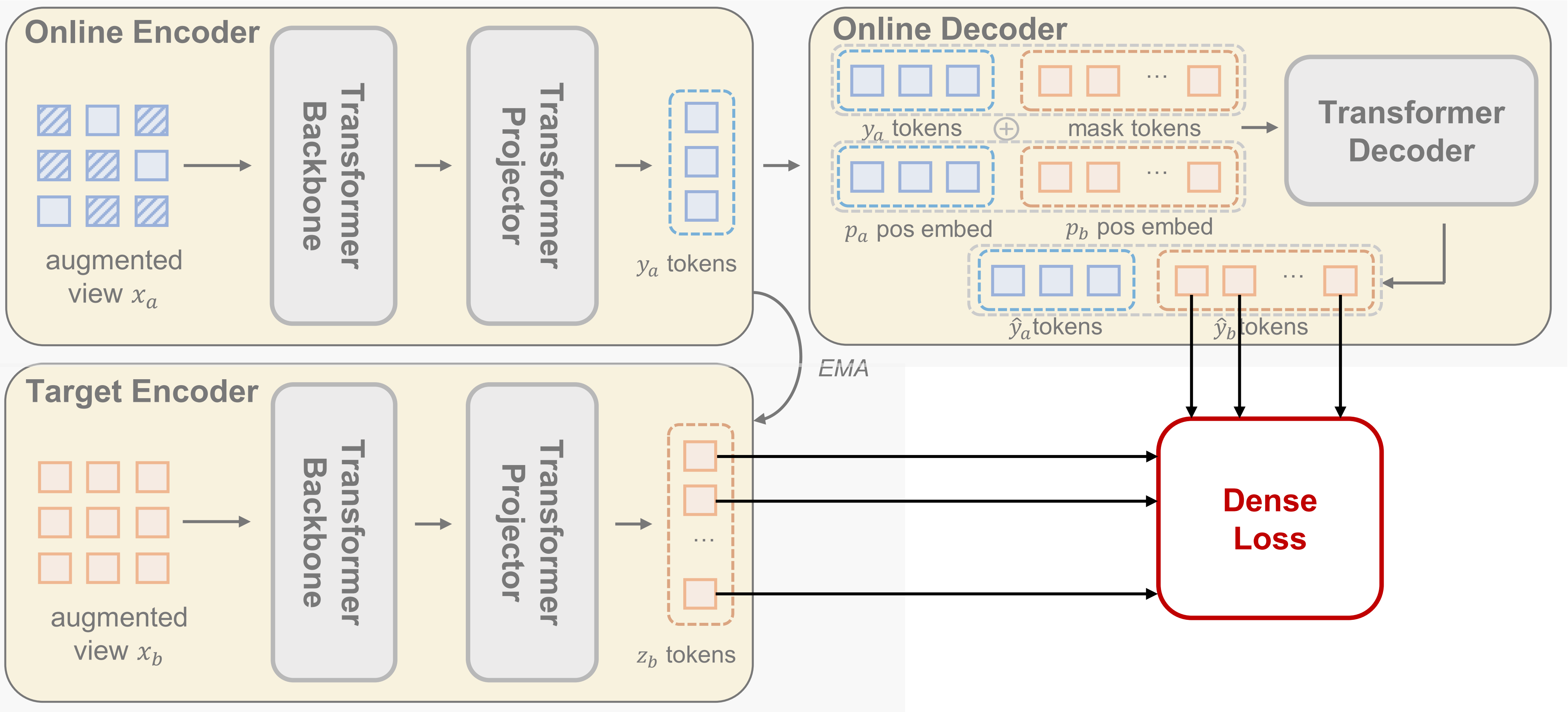


- Positional Embedding

$$\tilde{p}_a^{(u,v)} = (u-1, v-1)$$

$$\tilde{p}_b^{(u,v)} = (\frac{h_2}{h_1}(u-1) + \frac{i_2-i_1}{h_1}N_h, \frac{w_2}{w_1}(v-1) + \frac{j_2-j_1}{w_1}N_w)$$

Siamese Image Modeling



➢ Dense Loss

$$L = \mathbb{E}_{\{y_b^i, z_b^i\}}\left[-||y_b^i - z_b^i||^2 + \lambda \sum_{u \in N} \left(u^T y_b^i\right)^2\right]$$

Experiments – Main Results

Method	Epochs	ImageNet FT	ImageNet LIN	ImageNet $FT_{1\%}$
Supervised	300	81.8	-	-
DINO*	800†	82.8	78.2	-
iBOT*	1600†	84.0	79.5	-
DenseCL‡	400	82.2	69.7	49.9
MoCo-v3	600†	83.0	76.7	63.4
BEiT	800	83.2	-	-
MAE	400	83.1	62.5	-
MAE	1600	83.6	68.0	51.1
SiameseIM	400	83.7	76.8	61.8
SiameseIM	1600	**84.1** (+0.5)	**78.0** (+1.3)	**65.1** (+1.7)

(a) Image classification.

Method	Epochs	COCO AP^b	COCO AP^m	ADE20k mIoU
Supervised	300	47.9	42.9	47.4
DINO*	800†	50.1	43.4	46.8
iBOT*	1600†	51.2	44.2	50.0
DenseCL‡	400	46.6	41.6	44.5
MoCo-v3	600†	47.9	42.7	47.3
BEiT	800	49.8	44.4	47.1
MAE	400	50.6	45.1	45.0
MAE	1600	51.6	45.9	48.1
SiameseIM	400	50.7	44.9	49.6
SiameseIM	1600	**52.1** (+0.5)	**46.2** (+0.3)	**51.1** (+3.0)

(b) Common object detection and semantic segmentation.

Method	Epochs	AP^b	AP^b_{rare}	AP^m	AP^m_{rare}
Supervised	300	37.2	-	34.9	26.4
iBoT*	1600	36.9	29.1	34.6	28.9
DenseCL‡	400	33.8	25.1	32.1	24.6
MoCo-v3	600†	37.3	25.5	35.3	25.8
MAE	400	38.4	25.4	36.6	25.7
MAE	1600	40.1	29.3	**38.1**	29.1
SiameseIM	400	38.5	28.9	36.1	27.7
SiameseIM	1600	**40.5** (+0.4)	**30.9** (+1.6)	**38.1** (+0.0)	**30.1** (+1.0)

(c) Long-tail object detection on LVIS.

Method	Epochs	IN-A top-1	IN-R top-1	IN-Sketch top-1	IN-C 1-mCE
MSN*	1200†	37.5	50.0	36.3	53.4
iBoT*	1600†	42.4	50.9	36.9	55.5
DenseCL‡	400	30.8	43.8	29.9	48.1
MoCo-v3	600†	32.4	49.8	35.9	55.4
MAE	1600	35.9	48.3	34.5	48.3
SiameseIM	400	38.6	51.6	37.7	55.9
SiameseIM	1600	**43.8** (+7.9)	**52.5** (+2.7)	**38.3** (+2.4)	**57.1** (+1.7)

(d) Robustness evaluation.

- SiameseIM is able to surpass both MIM and ID methods over a wide range of tasks.

Experiments – Main Results

Method	Epochs	ImageNet FT	ImageNet LIN	ImageNet $FT_{1\%}$
Supervised	300	81.8	-	-
DINO*	800†	82.8	78.2	-
iBOT*	1600†	84.0	79.5	-
DenseCL‡	400	82.2	69.7	49.9
MoCo-v3	600†	83.0	76.7	63.4
BEiT	800	83.2	-	-
MAE	400	83.1	62.5	-
MAE	1600	83.6	68.0	51.1
SiameseIM	400	83.7	76.8	61.8
SiameseIM	1600	**84.1** (+0.5)	**78.0** (+1.3)	**65.1** (+1.7)

(a) Image classification.

Method	Epochs	COCO AP^b	COCO AP^m	ADE20k mIoU
Supervised	300	47.9	42.9	47.4
DINO*	800†	50.1	43.4	46.8
iBOT*	1600†	51.2	44.2	50.0
DenseCL‡	400	46.6	41.6	44.5
MoCo-v3	600†	47.9	42.7	47.3
BEiT	800	49.8	44.4	47.1
MAE	400	50.6	45.1	45.0
MAE	1600	51.6	45.9	48.1
SiameseIM	400	50.7	44.9	49.6
SiameseIM	1600	**52.1** (+0.5)	**46.2** (+0.3)	**51.1** (+3.0)

(b) Common object detection and semantic segmentation.

Method	Epochs	AP^b	AP^b_{rare}	AP^m	AP^m_{rare}
Supervised	300	37.2	-	34.9	26.4
iBoT*	1600	36.9	29.1	34.6	28.9
DenseCL‡	400	33.8	25.1	32.1	24.6
MoCo-v3	600†	37.3	25.5	35.3	25.8
MAE	400	38.4	25.4	36.6	25.7
MAE	1600	40.1	29.3	**38.1**	29.1
SiameseIM	400	38.5	28.9	36.1	27.7
SiameseIM	1600	**40.5** (+0.4)	**30.9** (+1.6)	**38.1** (+0.0)	**30.1** (+1.0)

(c) Long-tail object detection on LVIS.

Method	Epochs	IN-A top-1	IN-R top-1	IN-Sketch top-1	IN-C 1-mCE
MSN*	1200†	37.5	50.0	36.3	53.4
iBoT*	1600†	42.4	50.9	36.9	55.5
DenseCL‡	400	30.8	43.8	29.9	48.1
MoCo-v3	600†	32.4	49.8	35.9	55.4
MAE	1600	35.9	48.3	34.5	48.3
SiameseIM	400	38.6	51.6	37.7	55.9
SiameseIM	1600	**43.8** (+7.9)	**52.5** (+2.7)	**38.3** (+2.4)	**57.1** (+1.7)

(d) Robustness evaluation.

- SiameseIM obtains more improvements in few-shot, long-tail and robustness-concerned scenarios.

Experiments – Ablation Study

	target type	different views	color aug	mask type	BN/LN in proj & dec	loss norm*	loss type	loss form	FT	LIN	AP^b	AP^m
single view with dense loss:												
MAE	pixel			random	LN	MAE-like	dense	L2	83.1	62.5	46.8	42.0
(a)	pixel			random	LN	MAE-like	dense	L2	82.8	62.3	47.3	42.5
(b)	feature			random	LN	MoCo-like	dense	UniGrad	81.0	48.7	43.5	39.2
(c)	pixel		✓	random	LN	MAE-like	dense	L2	82.0	59.9	46.3	41.8
multiple views with dense loss:												
(d)	pixel	✓		random	LN	MAE-like	dense	L2	78.7	46.2	38.1	34.8
(e)	feature	✓		random	LN	MoCo-like	dense	UniGrad	82.9	69.6	48.5	43.4
(f)	feature	✓	✓	random	LN	MoCo-like	dense	UniGrad	83.0	73.1	47.9	43.2
(g)	feature	✓	✓	random	BN	MoCo-like	dense	UniGrad	83.2	73.6	48.7	43.7
(h)	feature	✓	✓	blockwise	BN	MoCo-like	dense	UniGrad	83.5	74.7	50.0	44.5
(i)	feature	✓	✓	blockwise	BN	MAE-like	dense	UniGrad	83.7	76.8	49.8	44.2
(j)	feature	✓	✓	blockwise	BN	MAE-like	dense	L2	83.3	76.5	49.8	44.2
multiple views with global loss:												
(k)	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.7	72.0	45.9	41.4
MoCo-v3 with mask	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.8	72.2	45.0	40.5

- ➢ Compared with MIM methods, reconstructing another view helps to obtain good semantic alignment.

Experiments – Ablation Study

	target type	different views	color aug	mask type	BN/LN in proj & dec	loss norm*	loss type	loss form	FT	LIN	AP^b	AP^m
single view with dense loss:												
MAE	pixel			random	LN	MAE-like	dense	L2	83.1	62.5	46.8	42.0
(a)	pixel			random	LN	MAE-like	dense	L2	82.8	62.3	47.3	42.5
(b)	feature			random	LN	MoCo-like	dense	UniGrad	81.0	48.7	43.5	39.2
(c)	pixel		✓	random	LN	MAE-like	dense	L2	82.0	59.9	46.3	41.8
multiple views with dense loss:												
(d)	pixel	✓		random	LN	MAE-like	dense	L2	78.7	46.2	38.1	34.8
(e)	feature	✓		random	LN	MoCo-like	dense	UniGrad	82.9	69.6	48.5	43.4
(f)	feature	✓	✓	random	LN	MoCo-like	dense	UniGrad	83.0	73.1	47.9	43.2
(g)	feature	✓	✓	random	BN	MoCo-like	dense	UniGrad	83.2	73.6	48.7	43.7
(h)	feature	✓	✓	blockwise	BN	MoCo-like	dense	UniGrad	83.5	74.7	50.0	44.5
(i)	feature	✓	✓	blockwise	BN	MAE-like	dense	UniGrad	83.7	76.8	49.8	44.2
(j)	feature	✓	✓	blockwise	BN	MAE-like	dense	L2	83.3	76.5	49.8	44.2
multiple views with global loss:												
(k)	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.7	72.0	45.9	41.4
MoCo-v3 with mask	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.8	72.2	45.0	40.5

- Compared with ID methods, dense supervision can improve the spatial sensitivity.

Siamese Image Modeling for Self-Supervised Vision Representation Learning

Thanks for watching!

Contact us:
tcx20@mails.tsinghua.edu.cn