Poster ID: WED-PM-337

Towards All-in-one Pre-training via Maximizing Multi-modal Mutual Information

Weijie Su*, Xizhou Zhu*, Chenxin Tao*, Lewei Lu, Bin Li, Gao Huang,
Yu Qiao, Xiaogang Wang, Jie Zhou, Jifeng Dai

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Presented By Weijie Su

Large Vision Model Pre-training

Method & Model	Stage 1	Stage 2	Stage 3
SwinV2-G (3B) [a]	Masked Image Modeling$_{\text{pixel}}$	Image Classification	-
BEiT-3 (2B) [b]	CLIP	Dense Distillation	Masked Data Modeling
FD-SwinV2-G (3B) [c]	Masked Image Modeling$_{\text{pixel}}$	Image Classification	Dense Distillation

[a] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution." CVPR 2022.
[b] Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv 2022.
[c] Wei, Yixuan, et al. "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation." arXiv 2022.

All existing large vision model pre-training methods are multi-stage

Problems of Multi-stage Pre-training

- **Difficult to Locate the Problematic Pre-training Stage** when the final performance is poor

- **Catastrophic Forgetting**
  - the linear classification accuracy of FD-DINO [1] in stage 2 (76.1) is worse than that of stage 1 (78.2)

Method	Backbone	res.	F. D.	IN-1K		ADE20K
				f.t.	linear	
BEiT [1]	ViT-B	224^2		83.2	37.6	47.1
MAE [17]	ViT-B	224^2		83.6	68.0	48.1
SimMIM [45]	ViT-B	224^2		83.8	56.7	47.6
SimMIM [45]	Swin-B	224^2		84.8	24.8	48.3
WiSE-FT CLIP [40]	ViT-L	336^2		87.1	-	-
DINO [3]	ViT-B	224^2		82.8	78.2	46.2
FD-DINO	ViT-B	224^2	✓	**83.8** (+1.0)	76.1	**47.7** (+1.5)

Problems of Multi-stage Pre-training

- **Difficult to Locate the Problematic Pre-training Stage** when the final performance is poor

- **Catastrophic Forgetting**
  - the linear classification accuracy of FD-DINO [1] in stage 2 (76.1) is worse than that of stage 1 (78.2)

Method	Backbone	res.	F. D.	IN-1K		ADE20K
				f.t.	linear	
BEiT [1]	ViT-B	224^2		83.2	37.6	47.1
MAE [17]	ViT-B	224^2		83.6	68.0	48.1
SimMIM [45]	ViT-B	224^2		83.8	56.7	47.6
SimMIM [45]	Swin-B	224^2		84.8	24.8	48.3
WiSE-FT CLIP [40]	ViT-L	336^2		87.1	-	-
DINO [3]	ViT-B	224^2		82.8	78.2	46.2
FD-DINO	ViT-B	224^2	✓	**83.8** (+1.0)	76.1	**47.7** (+1.5)

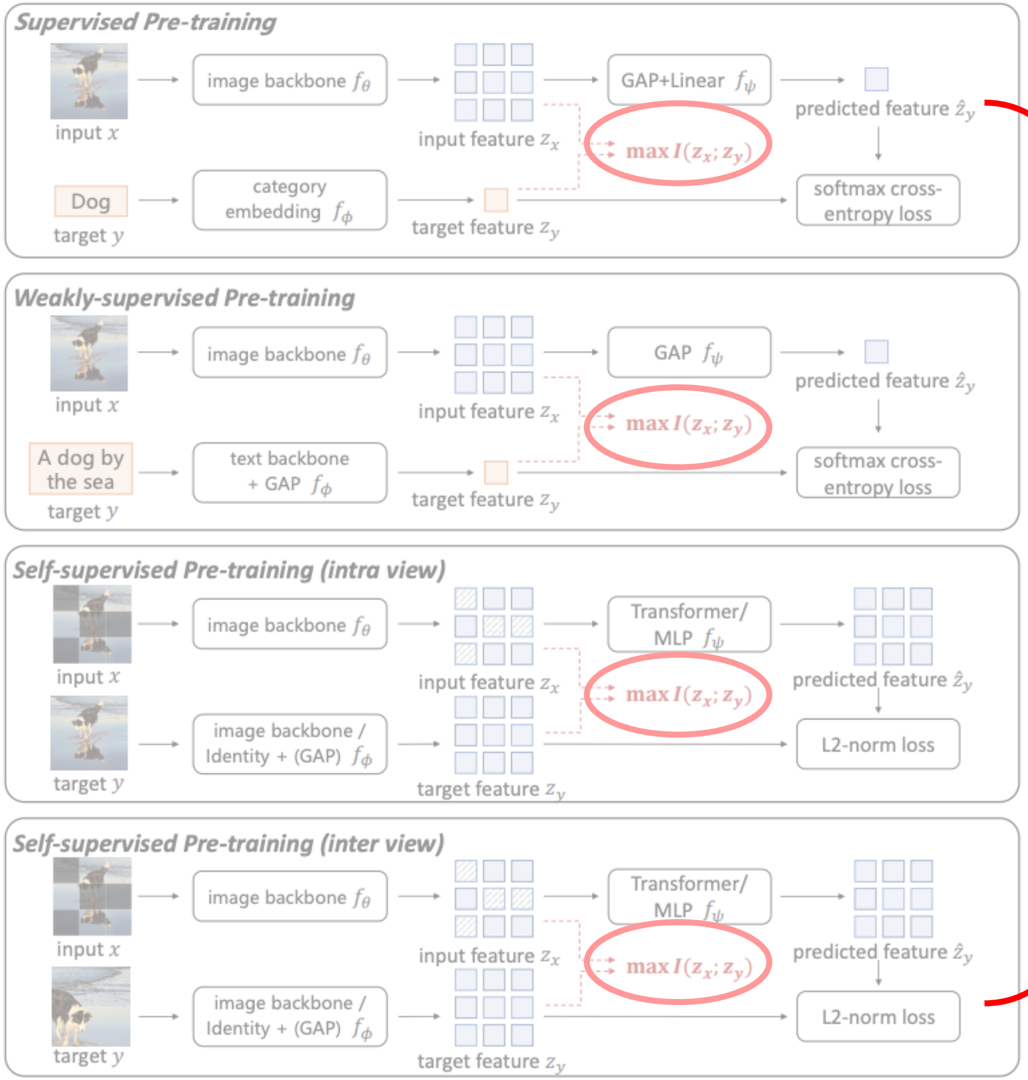
Problems of Multi-stage Pre-training

- **Difficult to Locate the Problematic Pre-training Stage** when the final performance is poor
- **Catastrophic Forgetting**
  - the linear classification accuracy of FD-DINO [1] in stage 2 (76.1) is worse than that of stage 1 (78.2)

Method	Backbone	res.	F. D.	IN-1K		ADE20K
				f.t.	linear	
BEiT [1]	ViT-B	224^2		83.2	37.6	47.1
MAE [illegible]	ViT-B	224^2		83.6	68.0	48.1
SimMIM [45]	ViT-B	224^2		83.8	56.7	47.6
SimMIM [45]	Swin-B	224^2		84.8	24.8	48.3
WiSE-FT CLIP [40]	ViT-L	336^2		87.1	-	-
DINO [3]	ViT-B	224^2		82.8	78.2	46.2
FD-DINO	ViT-B	224^2	✓	**83.8** (+1.0)	76.1	**47.7** (+1.5)

Our Solution: all-in-one single-stage pre-training under an unified perspective

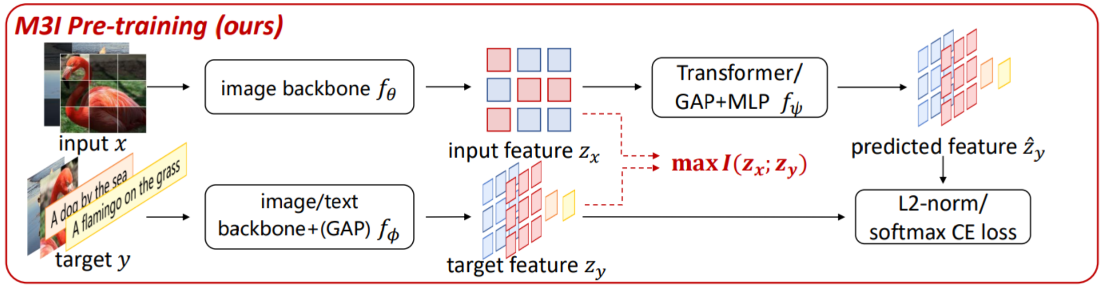
All-in-One: M3I Pre-training



training sample
input transform
target transform

$$(s, t_x, t_y) \sim D_{\text{train}} \quad \text{(sampled training sample)},$$
$$x = t_x(s)\ ,\ y = t_y(s) \quad \text{(extracted training data)},$$
$$z_x \sim p(z_x|x)\ ,\ z_y \sim p(z_y|y) \quad \text{(encoded training representation)},$$

$$\boxed{I(z_x; z_y \mid t_x, t_y)} = \underbrace{\mathbb{E}_{p(t_y)}\Big[H\big(p(z_y|t_y)\big)\Big]}_{\text{regularization term to avoid collapse}}$$
$$\boxed{-\underbrace{\mathbb{E}_{p(s,t_x,t_y,z_x)}\Big[H\big(p(z_y|y)\ ,\ p(z_y|z_x,t_x,t_y)\big)\Big]}_{\text{(cross-entropy) prediction term for target representation}},} \quad (1)$$



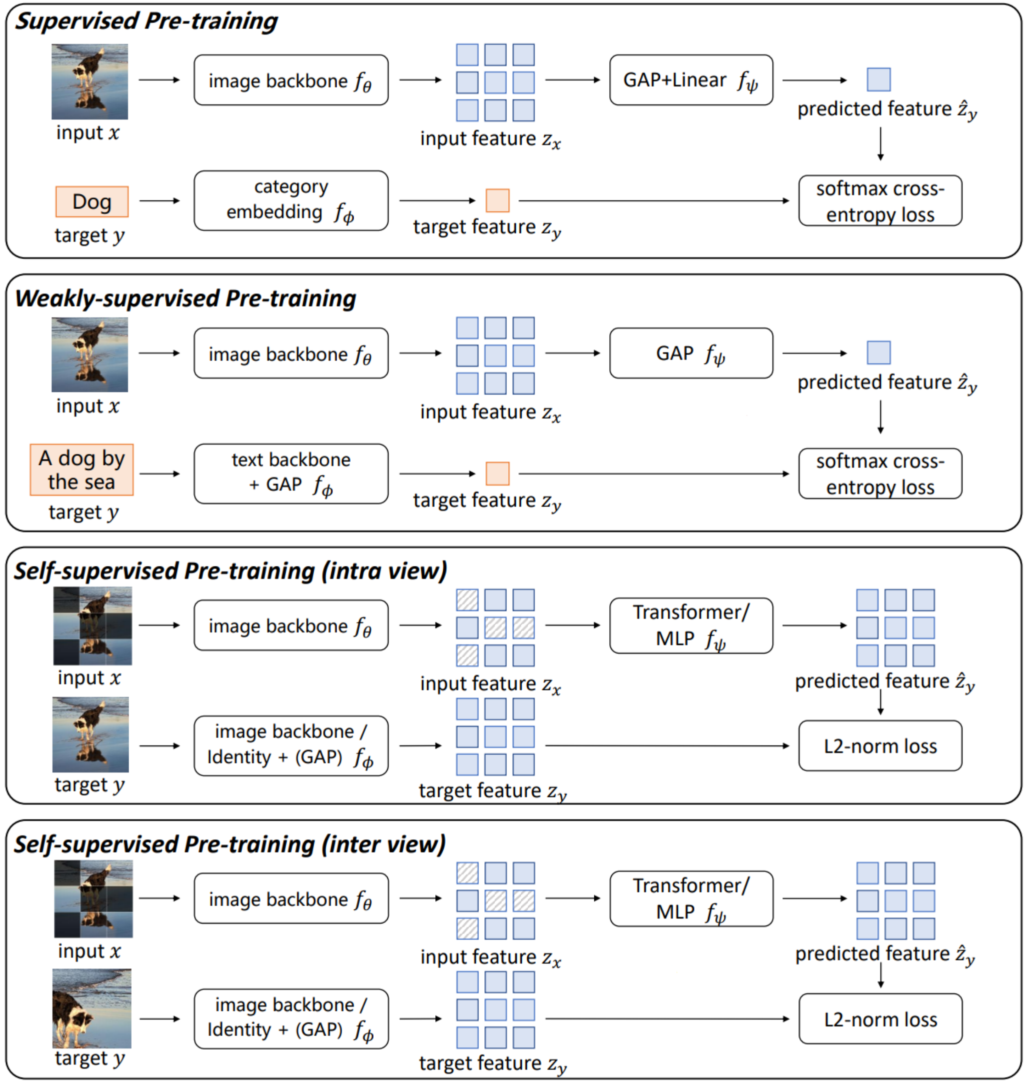
M3I Pre-training – Result on 1B model

Pre-training Approach	Model	Pipeline	Public Data	Private Data	ImageNet val	COCO test-dev	LVIS minival	ADE20k val
M3I Pre-training	InternImage-H [78] (1B)	Single Stage: M3I Pre-training	427M image-text 15M image-category	-	89.2	**65.4**	**62.5**	**62.9**
[47]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification	15M image-category	55M image-category	89.2	63.1	-	59.9
[77]	BEiT-3 (2B)	Stage 1: CLIP Stage 2: Dense Distillation Stage 3: Masked Data Modeling	21M image-text 15M image-category	400M image-text	**89.6**	63.7	-	62.8
[80]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification Stage 3: Dense Distillation	15M image-category	55M image-category	89.4	64.2	-	61.4
†previous best					89.1[a]	64.5[b]	59.8[c]	60.8[d]

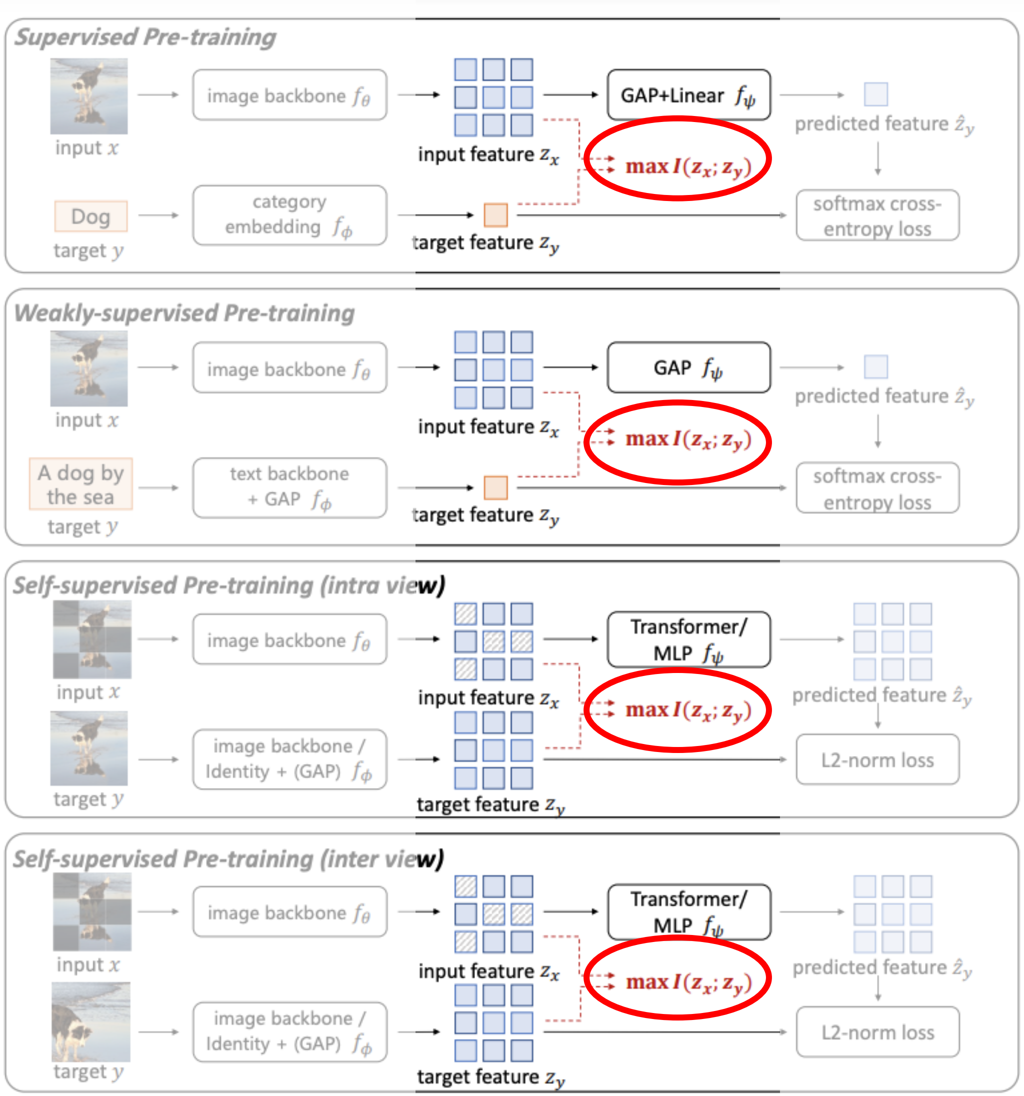
† previous best results on these tasks with only public training data. Results reference: a. MOAT, b. Group DETR v2, c. GLIPv2, d. Mask DINO

Achieves SoTA performance on various benchmarks in public-data only setting

4 Types of Visual Pre-training Methods



Unified Framework: Maximizing Mutual Information



Unified Framework: Maximize Mutual Information

training sample

input transform

target transform

$$(s, t_x, t_y) \sim D_{\text{train}} \qquad \text{(sampled training sample)},$$

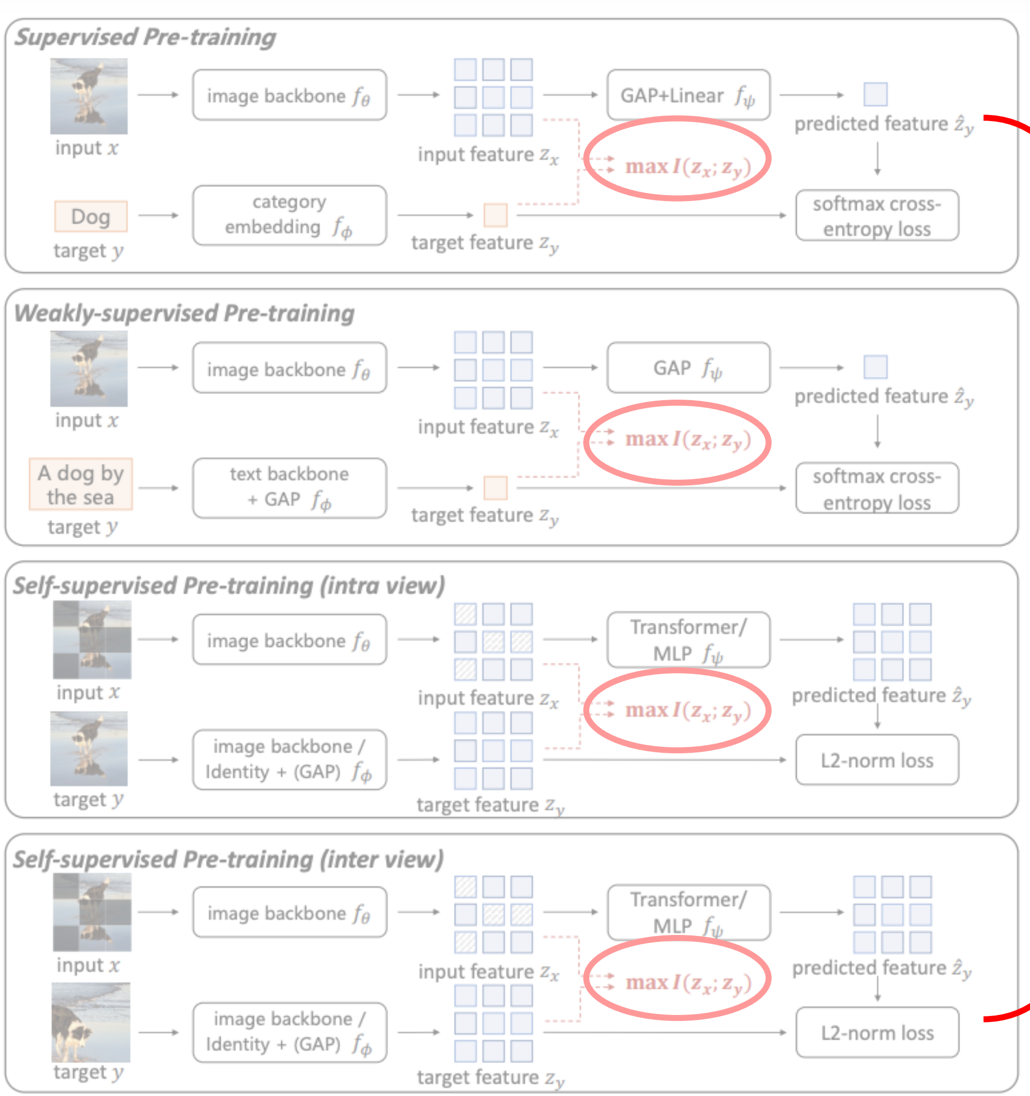
$$x = t_x(s)\ ,\ y = t_y(s) \qquad \text{(extracted training data)},$$

$$z_x \sim p(z_x|x)\ ,\ z_y \sim p(z_y|y) \qquad \text{(encoded training representation)},$$

$$\boxed{I(z_x; z_y \mid t_x, t_y)} = \underbrace{\mathbb{E}_{p(t_y)}\Big[H\big(p(z_y|t_y)\big)\Big]}_{\text{regularization term to avoid collapse}}$$

$$\boxed{-\underbrace{\mathbb{E}_{p(s,t_x,t_y,z_x)}\Big[H\big(p(z_y|y)\ ,\ p(z_y|z_x,t_x,t_y)\big)\Big]}_{\text{(cross-entropy) prediction term for target representation}},} \qquad (1)$$

All-in-One: M3I Pre-training



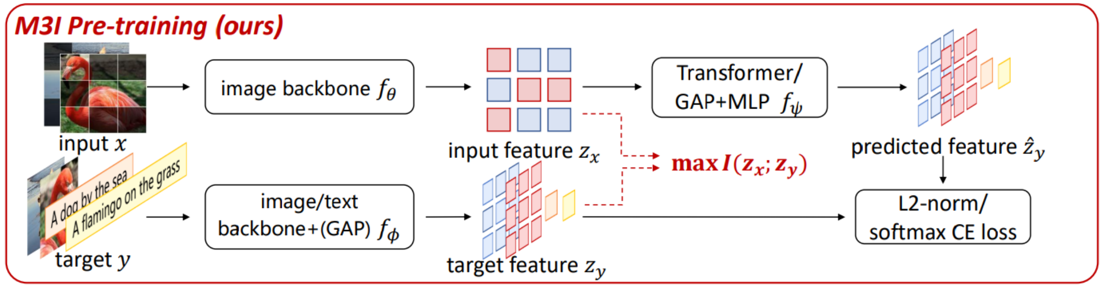
training sample

input transform

target transform

$$
\begin{aligned}
&(s, t_x, t_y) \sim D_{\text{train}} && \text{(sampled training sample)},\\
&x = t_x(s)\ ,\ y = t_y(s) && \text{(extracted training data)},\\
&z_x \sim p(z_x|x)\ ,\ z_y \sim p(z_y|y) && \text{(encoded training representation)},
\end{aligned}
$$

$$
\boxed{I(z_x; z_y \mid t_x, t_y)} = \underbrace{\mathbb{E}_{p(t_y)}\Big[H\big(p(z_y|t_y)\big)\Big]}_{\text{regularization term to avoid collapse}} \boxed{- \underbrace{\mathbb{E}_{p(s,t_x,t_y,z_x)}\Big[H\big(p(z_y|y)\ ,\ p(z_y|z_x, t_x, t_y)\big)\Big]}_{\text{(cross-entropy) prediction term for target representation}},} \quad (1)
$$



Unified Framework: All Instantiations

Pre-training Method	Typical Work	Input Data x	Target Data y	Input Representation z_x	Target Representation z_y	Regularization $H(p(z_y \| t_y))$	Distribution Form $\hat{P}$
Supervised Pre-training :							
Image Classification	ViT [24]	view1	category	dense feature	category embedding	negative categories	Boltzmann
Weakly-supervised Pre-training :							
Contrastive Language-Image Pre-training	CLIP [55]	view1	text	dense feature	text embedding	negative texts	Boltzmann
Self-supervised Pre-training (intra-view) :							
Auto-Encoder	-	view1	view1	dense feature	dense pixels	-	Gaussian
[1]Dense Distillation	FD [80],BEiT v2 tokenizer [54]	view1	view1	dense feature	dense feature	stop gradient	Gaussian
Global Distillation	-	view1	view1	dense feature	global feature	stop gradient	Boltzmann
Masked Image Modeling$_{\text{pixel}}$	MAE [30]	masked view1	view1	dense feature	dense pixels	-	Gaussian
[2]Masked Image Modeling$_{\text{feature}}$	data2vec [4],MILAN [35], BEiT [5],BEiT v2 [54]	masked view1	view1	dense feature	dense feature	stop gradient	Gaussian
Masked Image Modeling$_{\text{global}}$	-	masked view1	view1	dense feature	global feature	stop gradient	Gaussian
Self-supervised Pre-training (inter-view) :							
Novel View Synthesis	-	view2	view1	dense feature	dense pixels	-	Gaussian
Dense Instance Discrimination	DenseCL [79]	view2	view1	dense feature	dense feature	negative samples	Boltzmann
[3]Instance Discrimination	MoCo [31],BYOL [27], Barlow Twins [89]	view 2	view1	dense feature	global feature	negative samples / stop gradient / decorrelation	Boltzmann / Gaussian
Siamese Image Modeling$_{\text{pixel}}$	-	masked view2	view1	dense feature	dense pixels	-	Gaussian
Siamese Image Modeling$_{\text{feature}}$	SiameseIM [67]	masked view2	view1	dense feature	dense feature	stop gradient	Gaussian
Siamese Image Modeling$_{\text{global}}$	MSN [3]	masked view2	view1	dense feature	global feature	negative samples	Boltzmann

Unified Framework: All Instantiations

12 SSP Methods, some of which have not been explored as pre-training before

Pre-training Method	Typical Work	Input Data x	Target Data y	Input Representation z_x	Target Representation z_y	Regularization $H(p(z_y\|t_y))$	Distribution Form $\hat{P}$
Supervised Pre-training :							
Image Classification	ViT [24]	view1	category	dense feature	category embedding	negative categories	Boltzmann
Weakly-supervised Pre-training :							
Contrastive Language-Image Pre-training	CLIP [55]	view1	text	dense feature	text embedding	negative texts	Boltzmann
Self-supervised Pre-training (intra-view) :							
Auto-Encoder	-	view1	view1	dense feature	dense pixels	-	Gaussian
[1]Dense Distillation	FD [80],BEiT v2 tokenizer [54]	view1	view1	dense feature	dense feature	stop gradient	Gaussian
Global Distillation	-	view1	view1	dense feature	global feature	stop gradient	Boltzmann
Masked Image Modeling$_{\text{pixel}}$	MAE [30]	masked view1	view1	dense feature	dense pixels	-	Gaussian
[2]Masked Image Modeling$_{\text{feature}}$	data2vec [4],MILAN [35], BEiT [5],BEiT v2 [54]	masked view1	view1	dense feature	dense feature	stop gradient	Gaussian
Masked Image Modeling$_{\text{global}}$	-	masked view1	view1	dense feature	global feature	stop gradient	Gaussian
Self-supervised Pre-training (inter-view) :							
Novel View Synthesis	-	view2	view1	dense feature	dense pixels	-	Gaussian
Dense Instance Discrimination	DenseCL [79]	view2	view1	dense feature	dense feature	negative samples	Boltzmann
[3]Instance Discrimination	MoCo [31],BYOL [27], Barlow Twins [89]	view 2	view1	dense feature	global feature	negative samples / stop gradient / decorrelation	Boltzmann / Gaussian
Siamese Image Modeling$_{\text{pixel}}$	-	masked view2	view1	dense feature	dense pixels	-	Gaussian
Siamese Image Modeling$_{\text{feature}}$	SiameseIM [67]	masked view2	view1	dense feature	dense feature	stop gradient	Gaussian
Siamese Image Modeling$_{\text{global}}$	MSN [3]	masked view2	view1	dense feature	global feature	negative samples	Boltzmann

Compare 12 SSP Methods

Pre-training Method	Input Data x	Target Representation z_y	ImageNet Top1	COCO AP^{box}
Self-supervised Pre-training (intra-view)				
(a) Auto-Encoder	view1	dense pixels	77.5	$0.0^{\dagger}$
(b) Dense Distillation	view1	dense feature	78.8	32.4
(c) Global Distillation	view1	global feature	77.1	27.9
(d) Masked Image Modeling$_{pixel}$	masked view1	dense pixels	83.1	46.8
(e) Masked Image Modeling$_{feat}$	**masked view1**	**dense feature**	**83.3**	**47.4**
(f) Masked Image Modeling$_{gloal}$	masked view1	global feature	83.2	47.5
Self-supervised Pre-training (inter-view)				
(g) Novel View Synthesis	view2	dense pixels	78.8	33.0
(h) Dense Instance Discrimination	view2	dense feature	83.2	50.1
(i) Instance Discrimination	view2	global feature	83.0	46.4
(j) Siamese Image Modeling$_{pixel}$	masked view2	dense pixels	78.9	38.1
(k) Siamese Image Modeling$_{feat}$	**masked view2**	**dense feature**	**83.7**	**49.8**
(l) Instance Discrimination$_{mask}$	masked view2	global feature	82.9	46.2

Multi-Input Multi-target

$$(\boldsymbol{s}, \boldsymbol{t}_x, \boldsymbol{t}_y, X, Y) \sim D_{\text{train}} \qquad \text{(sample inputs and targets)}$$

$$\boldsymbol{z}_x = f_\theta\left(X = \{x_i\}_{i=1}^N\right) \qquad \text{(encode multiple inputs jointly)}$$

$$\boldsymbol{z}_y^k = f_{\phi_k}(Y_k),\ Y_k = \{y_{k_j}\}_{j=1}^{M_k} \qquad \text{(encode multiple targets separately)}$$

$$\hat{\boldsymbol{z}}_y^k = f_{\psi_k}(\boldsymbol{z}_x, \boldsymbol{t}_x, \boldsymbol{t}_y) \qquad \text{(predict multiple targets separately)}$$

$$I\left(\boldsymbol{z}_x;\{\boldsymbol{z}_y^k\}_{k=1}^K|\boldsymbol{t}_x, \boldsymbol{t}_y\right) \geq \sup_{\{f_{\psi_k}\}_{k=1}^K} \underbrace{\mathbb{E}_{p(\boldsymbol{t}_y)}\left[H\left(p(\{\boldsymbol{z}_y^k\}_{k=1}^K|\boldsymbol{t}_y)\right)\right]}_{\text{regularization term to avoid collapse}}$$

$$+ \boxed{\underbrace{\sum_{k=1}^K \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{t}_x,\boldsymbol{t}_y)}\left[\log \hat{P}_k\left(\boldsymbol{z}_y^k \mid \hat{\boldsymbol{z}}_y^k\right)\right]}_{\text{(log-likelihood) prediction term for target representation}}}\ ,$$

$$\Rightarrow L(\boldsymbol{s}, \boldsymbol{t}_x, \boldsymbol{t}_y) = \sum_{k=1}^K -\log \hat{P}_k\left(\boldsymbol{z}_y^k(\phi_k) \mid \hat{\boldsymbol{z}}_y^k(\theta, \psi_k)\right), \qquad (4)$$

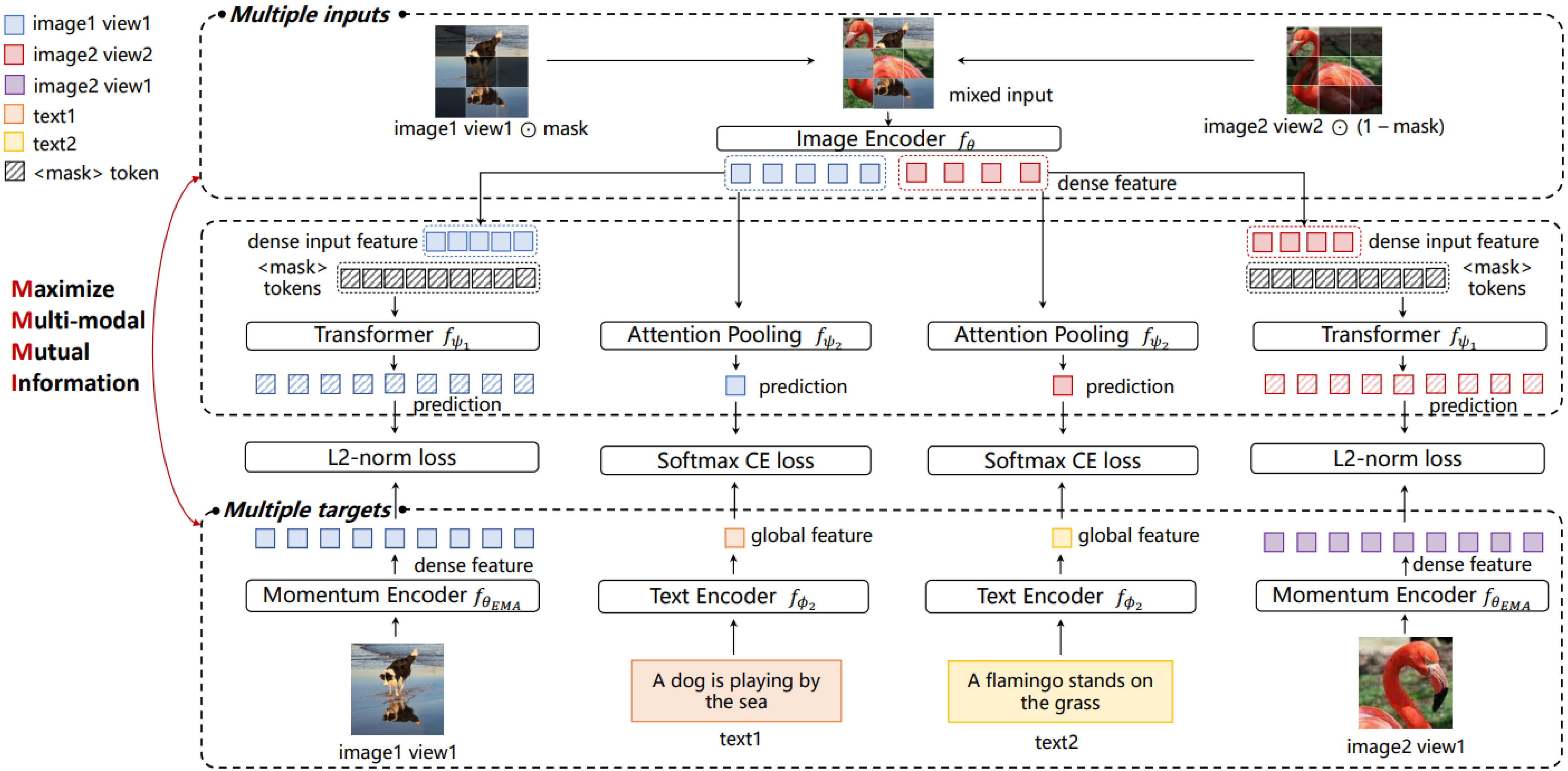
Ablation of Multi-Target

Pre-training Method	ImageNet	COCO	LVIS		ADE20k
	Top1	AP^{box}	AP^{box}	AP^{box}_{rare}	mIoU
Image Classification	81.8	46.6	33.0	25.5	45.1
Best Intra-view SSP	83.3	47.4	31.2	21.9	40.1
Best Inter-view SSP	83.7	49.8	35.2	26.9	47.7
Ours					
M3I Pre-training w/o mix	83.7	50.3	36.6	27.2	48.7
M3I Pre-training	**83.9**	**50.8**	**37.5**	**29.6**	**49.0**

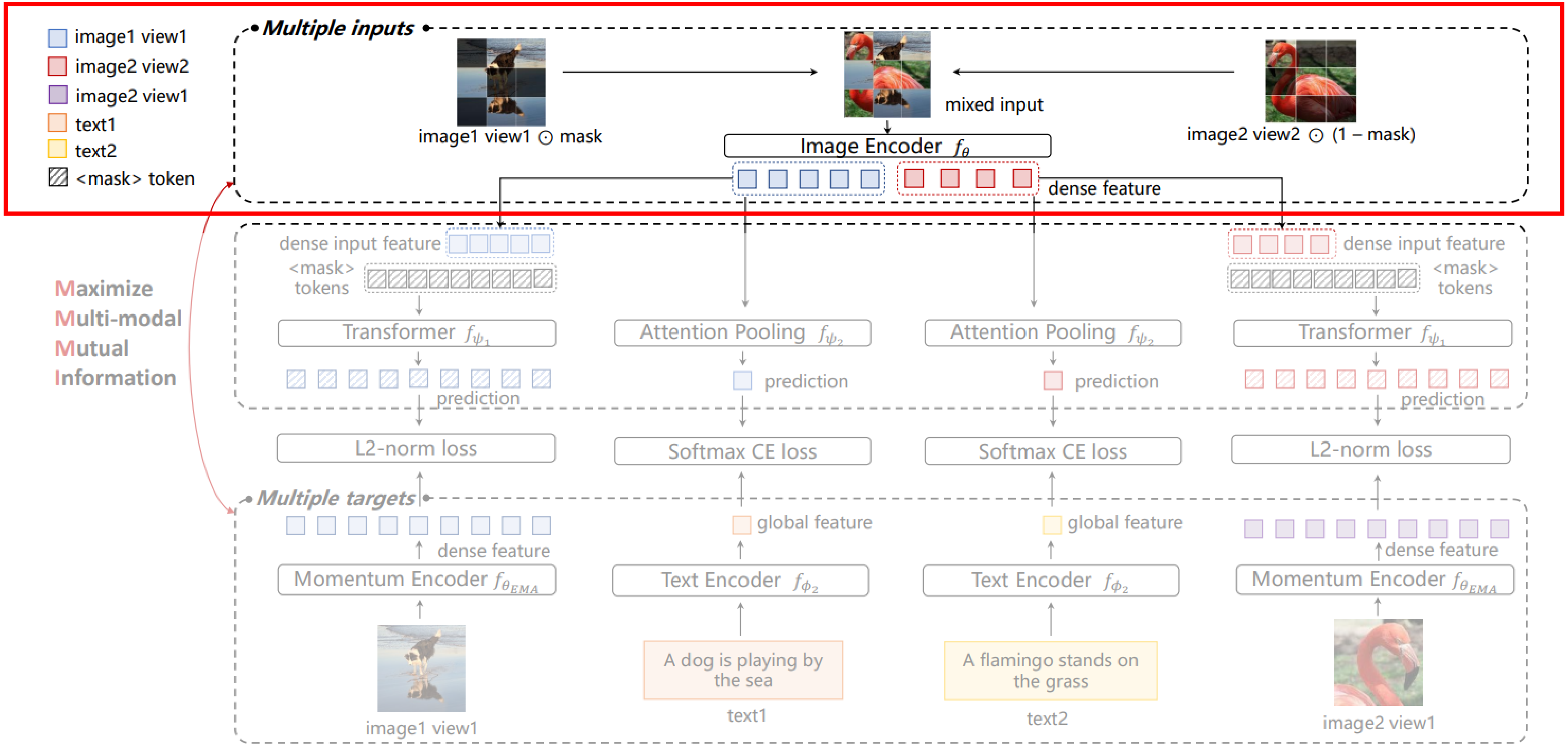
Ablation of Multi-Input

Pre-training Method	ImageNet Top1	COCO AP^{box}	LVIS AP^{box}	LVIS AP^{box}_{rare}	ADE20k mIoU
Image Classification	81.8	46.6	33.0	25.5	45.1
Best Intra-view SSP	83.3	47.4	31.2	21.9	40.1
Best Inter-view SSP	83.7	49.8	35.2	26.9	47.7
Ours					
M3I Pre-training w/o mix	83.7	50.3	36.6	27.2	48.7
M3I Pre-training	**83.9**	**50.8**	**37.5**	**29.6**	**49.0**

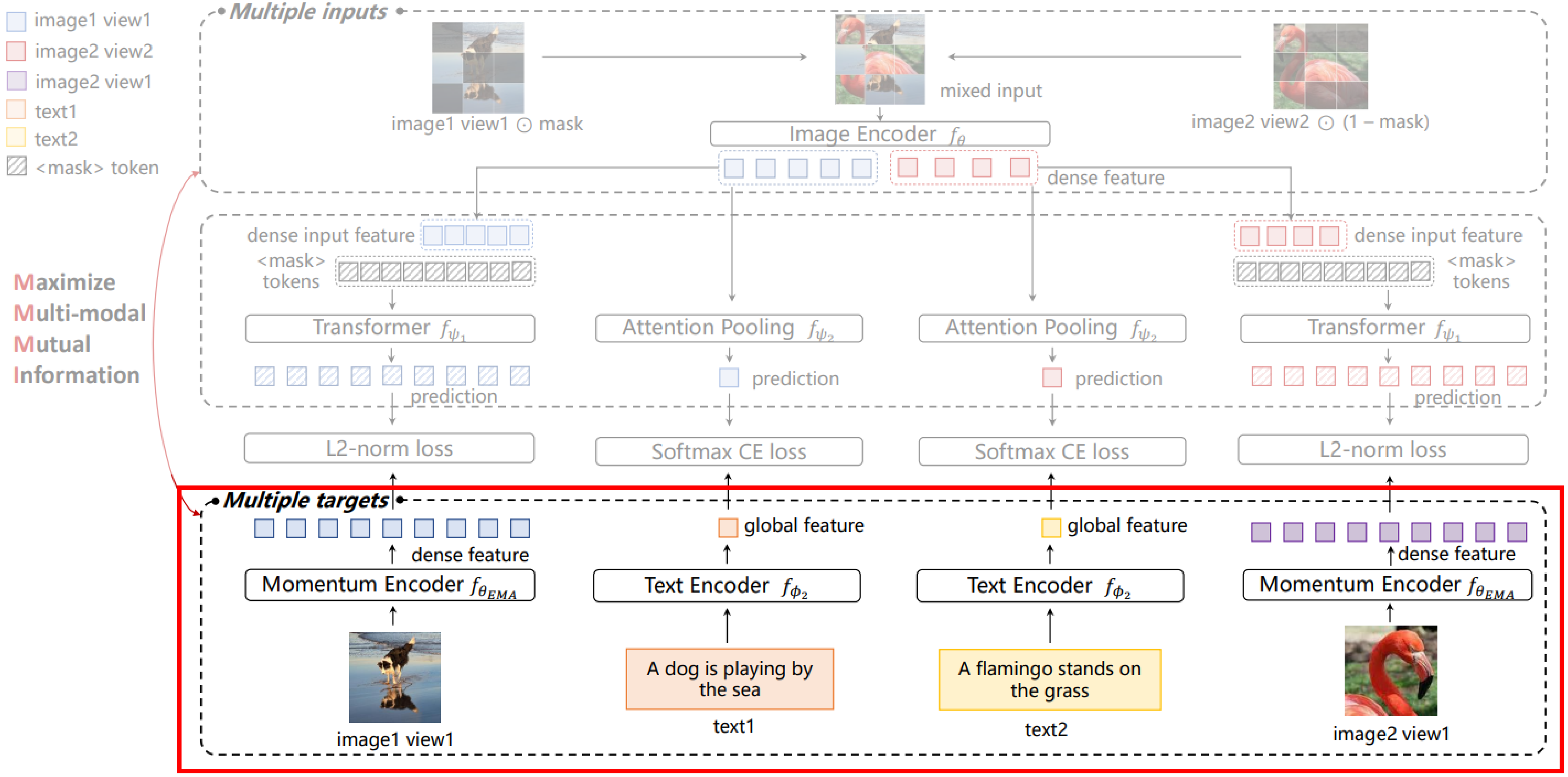
M3I Pre-training



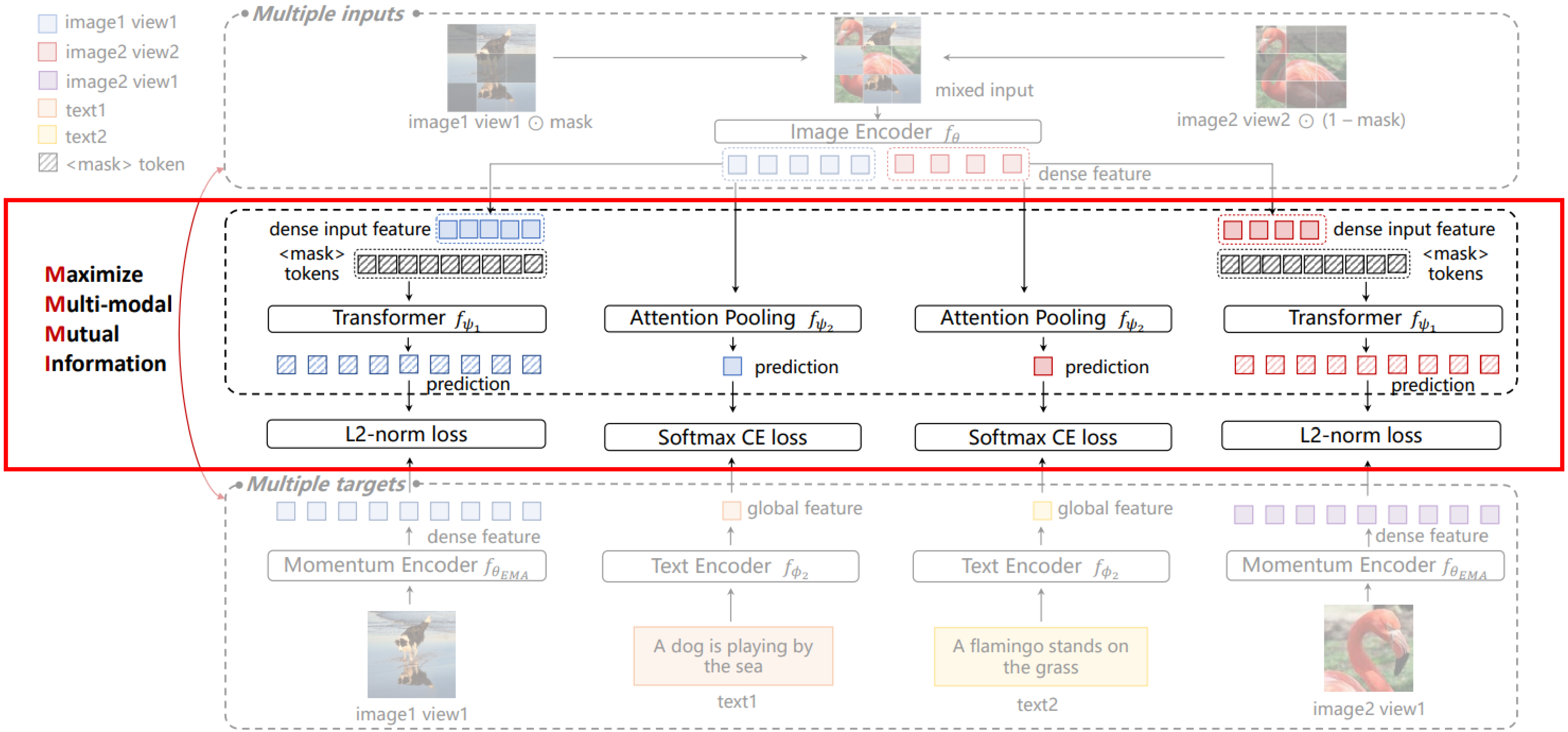
M3I Pre-training



M3I Pre-training



M3I Pre-training



Result on InternImage-H (1B)

Pre-training Approach	Model	Pipeline	Public Data	Private Data	ImageNet val	COCO test-dev	LVIS minival	ADE20k val
M3I Pre-training	InternImage-H [78] (1B)	Single Stage: M3I Pre-training	427M image-text 15M image-category	-	89.2	**65.4**	**62.5**	**62.9**
[47]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification	15M image-category	55M image-category	89.2	63.1	-	59.9
[77]	BEiT-3 (2B)	Stage 1: CLIP Stage 2: Dense Distillation Stage 3: Masked Data Modeling	21M image-text 15M image-category	400M image-text	**89.6**	63.7	-	62.8
[80]	SwinV2-G (3B)	Stage 1: Masked Image Modeling$_{\text{pixel}}$ Stage 2: Image Classification Stage 3: Dense Distillation	15M image-category	55M image-category	89.4	64.2	-	61.4
†previous best					89.1[a]	64.5[b]	59.8[c]	60.8[d]

† previous best results on these tasks with only public training data. Results reference: a. MOAT, b. Group DETR v2, c. GLIPv2, d. Mask DINO

Achieves SoTA performance on various benchmarks in public-data only setting

Result on ViT-B

Task	Metric	ImageNet Pre-train SSP (intra-view)	SSP (inter-view)	SP	M3I (ImageNet)	YFCC Pre-train WSP	M3I (YFCC)
ImageNet w/o Fine-tuning	Top1 acc.	×	×	**83.8** (DeiT-III)	83.3	†37.6 (CLIP)	†39.1
ImageNet Linear Classification	Top1 acc.	79.5 (iBOT)	78.0 (SiameseIM)	**83.8** (DeiT-III)	**83.8**	66.5 (CLIP)	72.3
ImageNet Fine-tuning	Top1 acc.	**84.2** (data2vec)	84.1 (SiameseIM)	83.8 (DeiT-III)	**84.2**	80.5 (CLIP)	83.7
COCO	AP^{box}	51.6 (MAE)	52.1 (SiameseIM)	47.6 (Sup.)	**52.2**	-	51.9
LVIS	AP^{box}	40.1 (MAE)	40.5 (SiameseIM)	37.2 (Sup.)	40.6	-	**40.8**
	AP^{box}_{rare}	38.1 (MAE)	38.1 (SiameseIM)	-	38.2	-	**38.4**
ADE20k	mIoU	50.0 (iBOT)	51.1 (SiameseIM)	49.3 (DeiT-III)	**51.3**	-	**51.3**

M3I Pretraining can maintain all desired properties through a single-stage pre-training

Conclusion

- Multi-stage pre-training methods has several problems
- We proposed a generic pre-training framework that unifies mainstream pre-training approaches
- We proposed an single-stage all-in-one pre-training method, M3I Pre-training
- Our approach surpasses previous pre-training methods in various transfer-learning settings

Poster ID: WED-PM-337

Contact Person: Weijie Su

E-mail: jackroos@mail.ustc.edu.cn

Paper

Code