InfoSAM: Fine-Tuning the Segment Anything Model from An Information-Theoretic Perspective

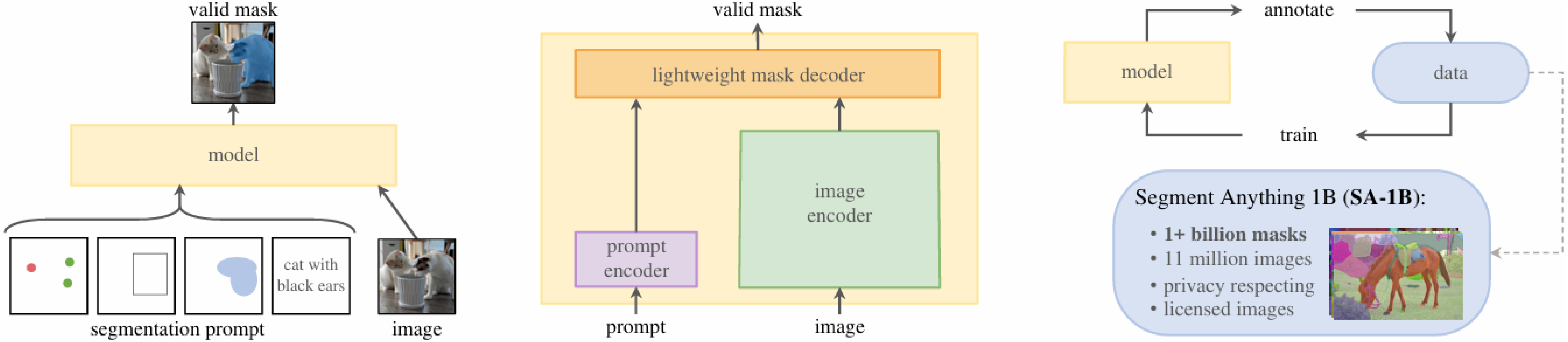
Yuanhong Zhang*[1], Muyao Yuan*[1], Weizhan Zhang[1], Tieliang Gong[1], Wen Wen[1], Jiangyong Ying[2], Weijie Shi[1]

[1] Xi'an Jiaotong University
[2] China Telecom E-surfing Vision Technology
* Equal contribution

Project Page

Background & Motivation

- **Segment Anything**[1]**:** shows impressive zero-shot performance on generic object segmentation



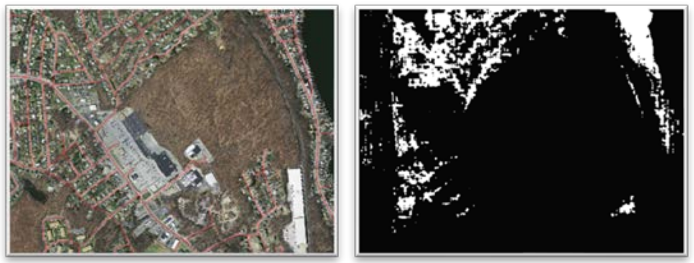
(a) **Task**: promptable segmentation | (b) **Model**: Segment Anything Model (**SAM**) | (c) **Data**: data engine (top) & dataset (bottom)

"interactive" | "Encoder-Decoder" | "over 1 billion masks"

- **Segment Anything:** still struggles with domain-specific real-world segmentation tasks

Camouflaged Scenes

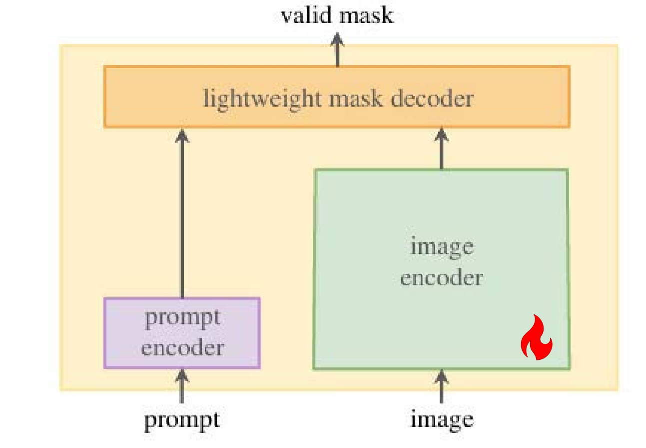
Remote Sensing



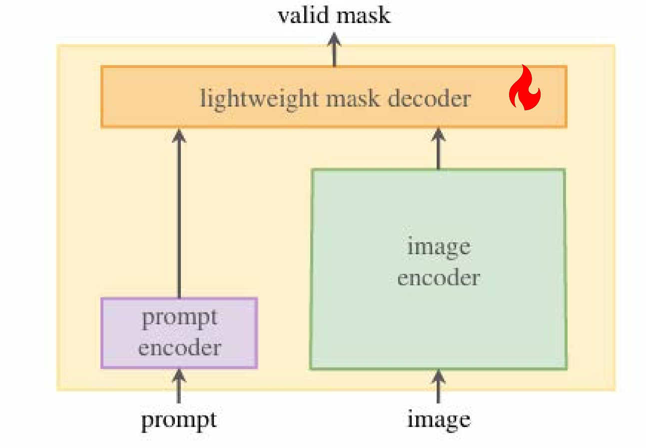
Agriculture

Background & Motivation

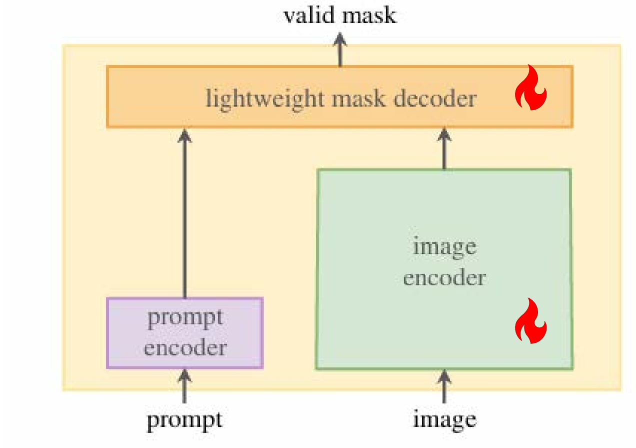
- **Parameter-Efficient Fine-Tuning (PEFT):** a promising approach to unleash the potential of SAM in novel scenarios.



SAM-COBOT[2]



HQSAM[3]



CAT-SAM[4]

Tunable

"Ignore the beneficial information encoded in the pre-trained SAM ! "

- We argue that:
  - There exists **domain-invariant information** that emerges from extensive pre-training.
  - This information is embedded in the feature distributions **between the encoder and decoder**, yet it can be easily **overridden or suppressed** during fine-tuning.

Challenges & Our solutions

Challenges:

- How to **extract** a good domain-invariant information? → To compress
- How to effectively **transfer it** to the fine-tuned models? → To distill

Formulation:

- Let z_i^T / z_i^S and z_m^T / z_m^S denote the image encoder features and mask decoder tokens from the teacher (pre-trained SAM) and student (fine-tuned SAM), respectively. We formulate the extraction and transfer process as the information flow. We use **matrix-based Rényi's theory** to quantify such information.
- **Objective 1**: to prioritize domain-invariant relations, we **constrain the information flow** via an upper bound I_c:

$$I_\alpha\left(z_i^T, z_m^T; r^T\right) \leq I_c$$

- **Objective 2**: To **maximize the extracted information** between pre-trained SAM and fine-tuned SAM:

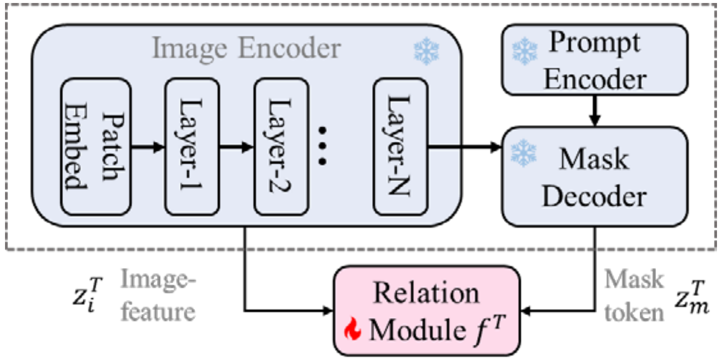
$$\max_{\omega} I_\alpha\left(r^T; r^S\right)$$

- The Lagrangian formulation explicitly implements this trade-off:

$$\max_{\omega} I_\alpha\left(r^T; r^S\right) - \beta I_\alpha\left(z_i^T, z_m^T; r^T\right)$$

To distill   To compress

Our solutions

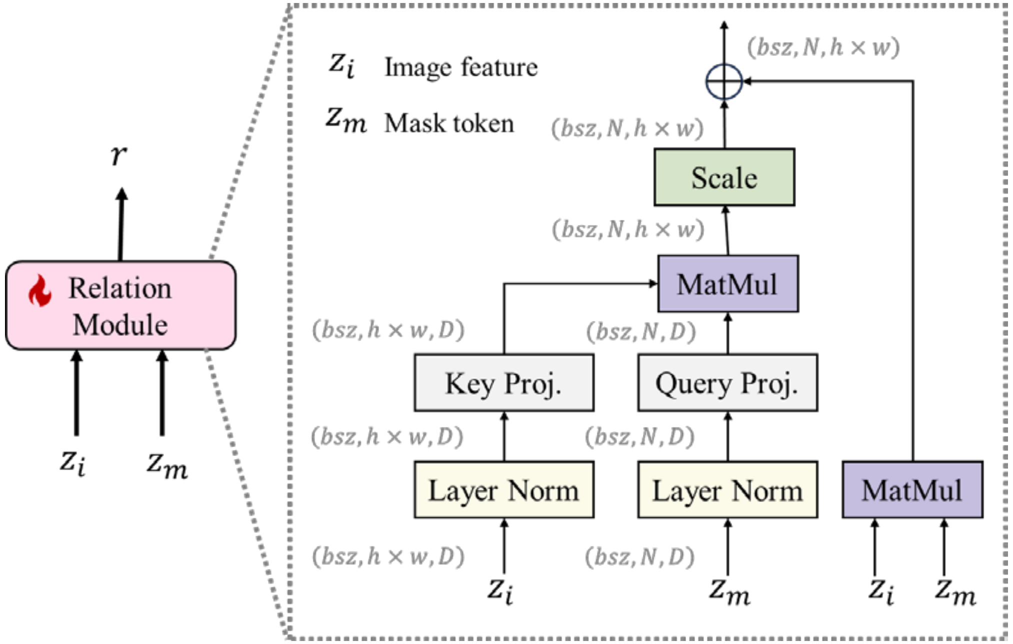


Extraction & Compression

☐ An Information View of SAM Distillation:

For Challenge 1: Compressing Intra-SAM Relations

✓ Attention-based module f^T designed for extraction



✓ Guide relations toward domain-invariant cues

$$I_\alpha(z_i^T, z_m^T; r^T) \le I_c$$

$$\Downarrow$$

$$\mathcal{L}_r = I_\alpha(z_i^T, z_m^T; r^T)$$
$$= \cancel{S_\alpha(G_i^T, G_m^T)} + S_\alpha(G_r^T) - S_\alpha(G_i^T, G_m^T, G_r^T)$$

$$\Downarrow$$

Set $\alpha = 2$

$$\boxed{\mathcal{L}_r = -log_2\|G_r^T\|_F^2 + log_2\left\|G_{imr}^T\right\|_F^2}$$

Our solutions

☐ An Information View of SAM Distillation:

For Challenge 2: Maximizing Inter-SAM Relations

✓ Transfer the relationships by minimizing their distance.

$$\max_{\omega} I_{\alpha}\left(r^{T}; r^{S}\right)$$

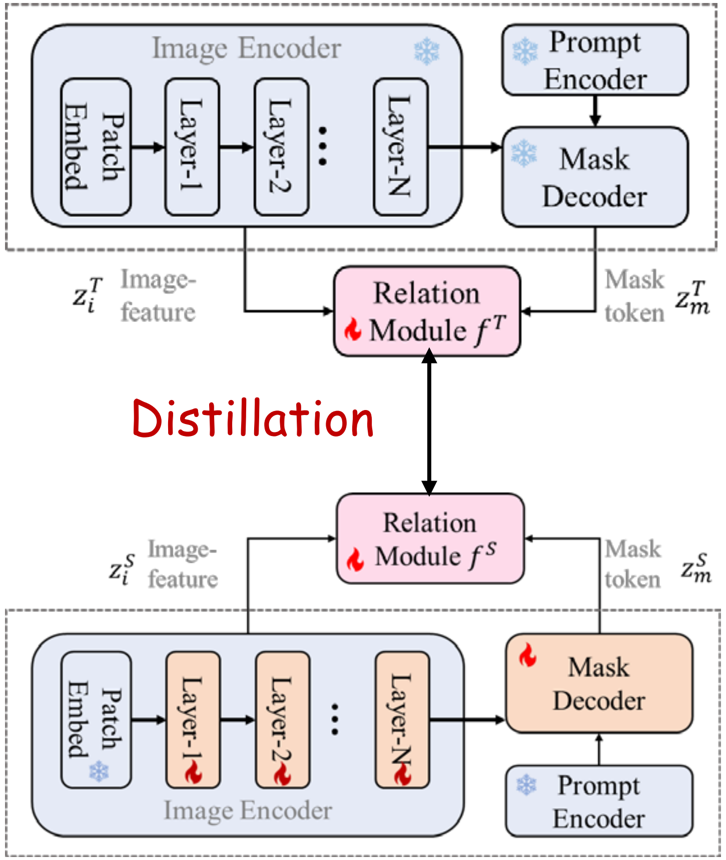
$$\Downarrow$$

$$\begin{aligned}\mathcal{L}_d &= -I_{\alpha}(r^{T}; r^{S}) \\ &= -S_{\alpha}(G_r^T) - S_{\alpha}(G_r^S) + S_{\alpha}(G_r^T, G_r^S)\end{aligned}$$

$$\Downarrow$$

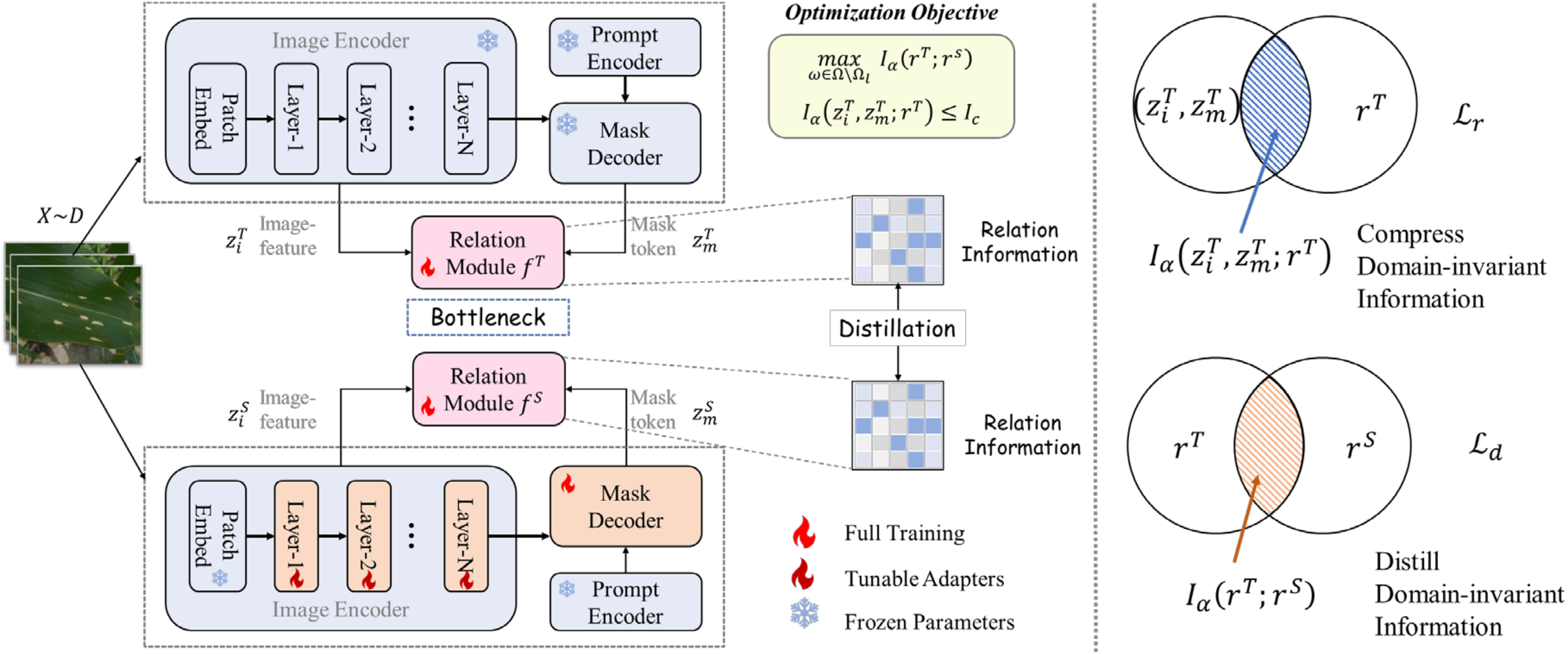
Set $\alpha = 2$

$$\boxed{\mathcal{L}_d = log_2\|G_r^T\|_F^2 + log_2\|G_r^S\|_F^2 - log_2\|G_r^{TS}\|_F^2}$$



Our solutions

☐ Overview of InfoSAM:



$$\boxed{\mathcal{L}_{info} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_d}$$

Experiments

- Compare with PEFT baselines across various downstream segmentation tasks

Method	Natural Images			Medical				Agriculture		Remote Sensing	
	CAMO			ISIC 2017		Kvasir		Leaf		Road	
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F^\omega_\beta \uparrow$	Jac ↑	Dice ↑	$S_\alpha \uparrow$	$E_\phi \uparrow$	IoU ↑	Dice ↑	IoU ↑	Dice ↑
SAM	$79.7_{\pm 0.02}$	$88.8_{\pm 0.09}$	$79.6_{\pm 0.01}$	$61.0_{\pm 0.12}$	$71.7_{\pm 0.14}$	$71.4_{\pm 0.16}$	$77.9_{\pm 0.17}$	$37.6_{\pm 0.11}$	$47.0_{\pm 0.16}$	$7.2_{\pm 0.24}$	$12.9_{\pm 0.29}$
decoder-only	$84.9_{\pm 0.38}$	$92.7_{\pm 0.34}$	$81.8_{\pm 0.33}$	$85.9_{\pm 0.34}$	$92.2_{\pm 0.20}$	$90.9_{\pm 0.05}$	$95.2_{\pm 0.18}$	$55.6_{\pm 1.12}$	$68.8_{\pm 1.17}$	$47.6_{\pm 0.47}$	$64.1_{\pm 0.47}$
BitFit	$87.5_{\pm 0.13}$	$94.5_{\pm 0.08}$	$85.3_{\pm 0.48}$	$87.7_{\pm 0.14}$	$93.2_{\pm 0.08}$	$92.5_{\pm 0.12}$	$96.3_{\pm 0.20}$	$69.2_{\pm 0.67}$	$80.3_{\pm 0.68}$	$58.1_{\pm 0.06}$	$73.1_{\pm 0.06}$
AdaptFormer	$87.9_{\pm 0.10}$	$94.8_{\pm 0.21}$	$86.2_{\pm 0.19}$	$87.6_{\pm 0.24}$	$93.2_{\pm 0.15}$	$93.3_{\pm 0.68}$	$97.0_{\pm 0.81}$	$75.0_{\pm 0.11}$	$84.8_{\pm 0.08}$	$61.1_{\pm 0.15}$	$75.5_{\pm 0.12}$
LoRA	$87.7_{\pm 0.59}$	$94.6_{\pm 0.50}$	$85.1_{\pm 0.64}$	$87.8_{\pm 0.24}$	$93.3_{\pm 0.13}$	$93.0_{\pm 0.14}$	$96.6_{\pm 0.11}$	$71.4_{\pm 0.54}$	$82.1_{\pm 0.62}$	$59.0_{\pm 0.19}$	$74.0_{\pm 0.17}$
Adapter	$88.2_{\pm 0.44}$	$94.8_{\pm 0.34}$	$86.7_{\pm 0.92}$	$87.7_{\pm 0.23}$	$93.2_{\pm 0.16}$	$93.4_{\pm 0.12}$	$97.1_{\pm 0.15}$	$74.4_{\pm 0.16}$	$84.3_{\pm 0.28}$	$60.5_{\pm 0.10}$	$75.1_{\pm 0.08}$
HQ-SAM	$85.1_{\pm 0.10}$	$92.6_{\pm 0.10}$	$81.0_{\pm 0.61}$	$86.3_{\pm 0.32}$	$92.4_{\pm 0.19}$	$91.1_{\pm 0.50}$	$95.5_{\pm 0.57}$	$66.2_{\pm 0.44}$	$77.8_{\pm 0.43}$	$54.9_{\pm 0.16}$	$70.6_{\pm 0.13}$
SU-SAM	$88.3_{\pm 0.21}$	$95.0_{\pm 0.22}$	$86.2_{\pm 0.59}$	$87.8_{\pm 0.18}$	$93.2_{\pm 0.09}$	$93.8_{\pm 0.02}$	$97.5_{\pm 0.06}$	$74.7_{\pm 0.53}$	$84.5_{\pm 0.56}$	$60.2_{\pm 0.26}$	$74.8_{\pm 0.22}$
ConvLoRA-SAM	$87.5_{\pm 0.39}$	$94.5_{\pm 0.17}$	$85.4_{\pm 0.41}$	$87.7_{\pm 0.22}$	$93.2_{\pm 0.11}$	$92.9_{\pm 0.13}$	$96.6_{\pm 0.28}$	$71.4_{\pm 0.44}$	$82.2_{\pm 0.37}$	$59.6_{\pm 0.22}$	$74.4_{\pm 0.20}$
LoRA+Ours	$\mathbf{88.3}_{\pm 0.05}$	$\mathbf{95.2}_{\pm 0.00}$	$\mathbf{85.8}_{\pm 0.59}$	$\mathbf{88.1}_{\pm 0.08}$	$\mathbf{93.5}_{\pm 0.05}$	$\mathbf{93.4}_{\pm 0.11}$	$\mathbf{96.8}_{\pm 0.09}$	$\mathbf{72.2}_{\pm 0.06}$	$\mathbf{82.8}_{\pm 0.04}$	$\mathbf{59.9}_{\pm 0.20}$	$\mathbf{74.6}_{\pm 0.17}$
Adapter+Ours	$\mathbf{88.6}_{\pm 0.09}$	$\mathbf{95.1}_{\pm 0.05}$	$\mathbf{87.1}_{\pm 0.37}$	$\mathbf{88.0}_{\pm 0.05}$	$\mathbf{93.4}_{\pm 0.00}$	$\mathbf{94.4}_{\pm 0.12}$	$\mathbf{97.9}_{\pm 0.09}$	$\mathbf{75.6}_{\pm 0.27}$	$\mathbf{85.2}_{\pm 0.23}$	$\mathbf{61.4}_{\pm 0.30}$	$\mathbf{75.8}_{\pm 0.27}$

InfoSAM outperforms other PEFT techniques across various datasets from different domains.

Experiments

Compare with distillation baselines across various domains

Method	Natural Images			Medical				Agriculture		Remote Sensing	
	CAMO			ISIC 2017		Kvasir		Leaf		Road	
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	Jac ↑	Dice ↑	$S_\alpha \uparrow$	$E_\phi \uparrow$	IoU ↑	Dice ↑	IoU ↑	Dice ↑
Teacher	$79.7_{\pm 0.02}$	$88.8_{\pm 0.09}$	$79.6_{\pm 0.01}$	$61.0_{\pm 0.12}$	$71.7_{\pm 0.14}$	$83.0_{\pm 0.10}$	$88.8_{\pm 0.29}$	$37.6_{\pm 0.11}$	$47.0_{\pm 0.16}$	$7.2_{\pm 0.24}$	$12.9_{\pm 0.29}$
Student	$88.2_{\pm 0.44}$	$94.8_{\pm 0.34}$	$86.7_{\pm 0.92}$	$87.7_{\pm 0.23}$	$93.2_{\pm 0.16}$	$93.4_{\pm 0.12}$	$97.1_{\pm 0.15}$	$74.4_{\pm 0.16}$	$84.3_{\pm 0.28}$	$60.5_{\pm 0.10}$	$75.1_{\pm 0.08}$
Logit	$88.4_{\pm 0.08}$	$94.9_{\pm 0.05}$	$87.1_{\pm 0.22}$	$87.2_{\pm 0.43}$	$92.9_{\pm 0.29}$	$93.2_{\pm 0.19}$	$96.5_{\pm 0.19}$	$73.0_{\pm 0.35}$	$83.3_{\pm 0.29}$	$50.9_{\pm 0.08}$	$67.2_{\pm 0.06}$
PKD	$87.0_{\pm 0.43}$	$94.1_{\pm 0.23}$	$84.3_{\pm 0.97}$	$86.5_{\pm 0.26}$	$92.5_{\pm 0.17}$	$92.2_{\pm 0.25}$	$96.0_{\pm 0.17}$	$70.2_{\pm 1.15}$	$81.1_{\pm 1.08}$	$56.9_{\pm 0.61}$	$72.2_{\pm 0.56}$
PKT	$87.8_{\pm 0.40}$	$94.5_{\pm 0.35}$	$86.2_{\pm 0.46}$	$87.4_{\pm 0.12}$	$93.0_{\pm 0.07}$	$93.7_{\pm 0.41}$	$97.3_{\pm 0.53}$	$74.2_{\pm 0.51}$	$84.2_{\pm 0.52}$	$60.7_{\pm 0.20}$	$75.2_{\pm 0.16}$
IBD	$85.2_{\pm 0.47}$	$92.6_{\pm 0.35}$	$82.4_{\pm 0.31}$	$85.1_{\pm 0.74}$	$91.7_{\pm 0.45}$	$91.5_{\pm 0.14}$	$95.3_{\pm 0.05}$	$72.2_{\pm 0.12}$	$82.7_{\pm 0.07}$	$44.9_{\pm 0.18}$	$61.5_{\pm 0.18}$
VID	$87.9_{\pm 0.22}$	$94.8_{\pm 0.34}$	$86.3_{\pm 0.32}$	$87.6_{\pm 0.44}$	$93.1_{\pm 0.29}$	$93.7_{\pm 0.16}$	$97.4_{\pm 0.07}$	$75.1_{\pm 0.08}$	$84.9_{\pm 0.17}$	$60.7_{\pm 0.19}$	$75.4_{\pm 0.19}$
SemCKD	$86.2_{\pm 0.16}$	$93.5_{\pm 0.21}$	$82.8_{\pm 1.54}$	$85.4_{\pm 0.27}$	$91.8_{\pm 0.19}$	$92.4_{\pm 0.07}$	$96.2_{\pm 0.03}$	$72.0_{\pm 0.04}$	$82.8_{\pm 0.10}$	$53.5_{\pm 0.17}$	$69.4_{\pm 0.17}$
ReviewKD	$86.7_{\pm 0.07}$	$94.0_{\pm 0.09}$	$84.6_{\pm 0.63}$	$85.5_{\pm 0.26}$	$91.9_{\pm 0.15}$	$92.4_{\pm 0.33}$	$96.4_{\pm 0.26}$	$72.6_{\pm 0.64}$	$83.1_{\pm 0.47}$	$57.3_{\pm 0.11}$	$72.6_{\pm 0.11}$
TinySAM	$83.7_{\pm 0.39}$	$91.6_{\pm 0.31}$	$81.1_{\pm 0.35}$	$79.4_{\pm 1.12}$	$87.8_{\pm 0.84}$	$88.5_{\pm 0.31}$	$93.5_{\pm 0.24}$	$48.6_{\pm 1.14}$	$61.0_{\pm 0.95}$	$25.7_{\pm 1.19}$	$39.6_{\pm 1.71}$
MobileSAM	$87.1_{\pm 0.36}$	$94.1_{\pm 0.27}$	$85.1_{\pm 0.09}$	$86.7_{\pm 0.13}$	$92.6_{\pm 0.09}$	$92.5_{\pm 0.12}$	$96.3_{\pm 0.14}$	$71.9_{\pm 0.30}$	$82.6_{\pm 0.39}$	$59.2_{\pm 0.09}$	$74.1_{\pm 0.08}$
InfoSAM(Ours)	$\mathbf{88.6_{\pm 0.09}}$	$\mathbf{95.1_{\pm 0.05}}$	$\mathbf{87.1_{\pm 0.37}}$	$\mathbf{88.0_{\pm 0.05}}$	$\mathbf{93.4_{\pm 0.00}}$	$\mathbf{94.4_{\pm 0.12}}$	$\mathbf{97.9_{\pm 0.09}}$	$\mathbf{75.6_{\pm 0.27}}$	$\mathbf{85.2_{\pm 0.23}}$	$\mathbf{61.4_{\pm 0.30}}$	$\mathbf{75.8_{\pm 0.27}}$

Most distillation methods harm PEFT due to the weak teacher, often underperforming vanilla fine-tuning. In contrast, InfoSAM distills only essential knowledge from the teacher.

Experiments

Extended Experiment with SAM2

(a) PEFT Methods Comparison

Method	Medical	Agriculture	Remote Sensing
	S_α (Kvasir)	IoU (Leaf)	IoU (Road)
SAM2	$87.1_{\pm 0.12}$	$42.7_{\pm 0.32}$	$6.9_{\pm 0.13}$
decoder-only	$93.2_{\pm 0.07}$	$71.8_{\pm 0.58}$	$48.5_{\pm 0.47}$
BitFit	$93.8_{\pm 0.09}$	$75.4_{\pm 0.29}$	$59.2_{\pm 0.26}$
AdaptFormer	$93.7_{\pm 0.19}$	$73.6_{\pm 1.10}$	$59.9_{\pm 0.35}$
LoRA	$93.7_{\pm 0.10}$	$75.9_{\pm 0.40}$	$60.8_{\pm 0.32}$
Adapter	$94.4_{\pm 0.06}$	$76.8_{\pm 0.56}$	$60.9_{\pm 0.14}$
LoRA+Ours	$\mathbf{94.0}_{\pm \mathbf{0.09}}$	$\mathbf{76.1}_{\pm \mathbf{0.38}}$	$\mathbf{60.9}_{\pm \mathbf{0.05}}$
Adapter+Ours	$\mathbf{94.5}_{\pm \mathbf{0.17}}$	$\mathbf{77.3}_{\pm \mathbf{0.14}}$	$\mathbf{61.3}_{\pm \mathbf{0.05}}$

(b) Distillation Methods Comparison

Method	Medical	Agriculture	Remote Sensing
	S_α (Kvasir)	IoU (Leaf)	IoU (Road)
Teacher	$87.1_{\pm 0.12}$	$42.7_{\pm 0.32}$	$6.9_{\pm 0.13}$
Student	$94.4_{\pm 0.06}$	$76.8_{\pm 0.56}$	$60.9_{\pm 0.14}$
PKT	$94.0_{\pm 0.25}$	$74.8_{\pm 0.14}$	$57.3_{\pm 0.07}$
VID	$94.1_{\pm 0.47}$	$77.2_{\pm 0.37}$	$61.1_{\pm 0.38}$
ReviewKD	$93.4_{\pm 0.10}$	$72.7_{\pm 0.37}$	$55.9_{\pm 0.50}$
TinySAM	$89.4_{\pm 0.10}$	$45.2_{\pm 0.76}$	$23.9_{\pm 2.61}$
MobileSAM	$93.3_{\pm 0.15}$	$74.1_{\pm 0.35}$	$52.3_{\pm 0.46}$
InfoSAM2(Ours)	$\mathbf{94.5}_{\pm \mathbf{0.17}}$	$\mathbf{77.3}_{\pm \mathbf{0.14}}$	$\mathbf{61.3}_{\pm \mathbf{0.05}}$

InfoSAM consistently performs well with SAM2, thanks to its structure-independent, information-theoretic foundation.

Experiments

Ablation Study

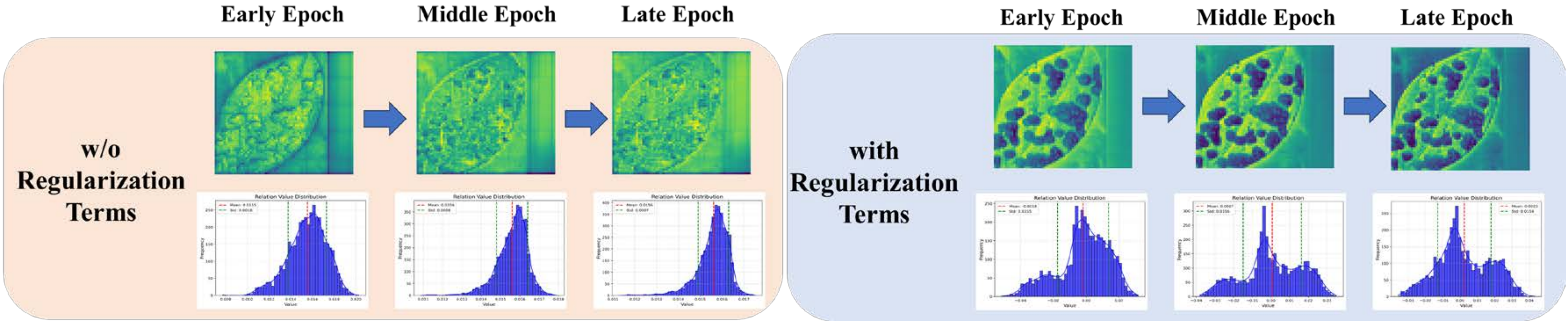
✓ Ablation study results of two losses

L_r	L_d	Medical S_α (Kvasir)	Agriculture IoU (Leaf)	Remote Sensing IoU (Road)
		93.4	74.4	60.5
	✓	93.6 (+0.2)	75.2 (+0.8)	61.0 (+0.5)
✓	✓	94.4 (+1.0)	75.6 (+1.2)	61.4 (+0.9)

✓ Effects of the Relation Module

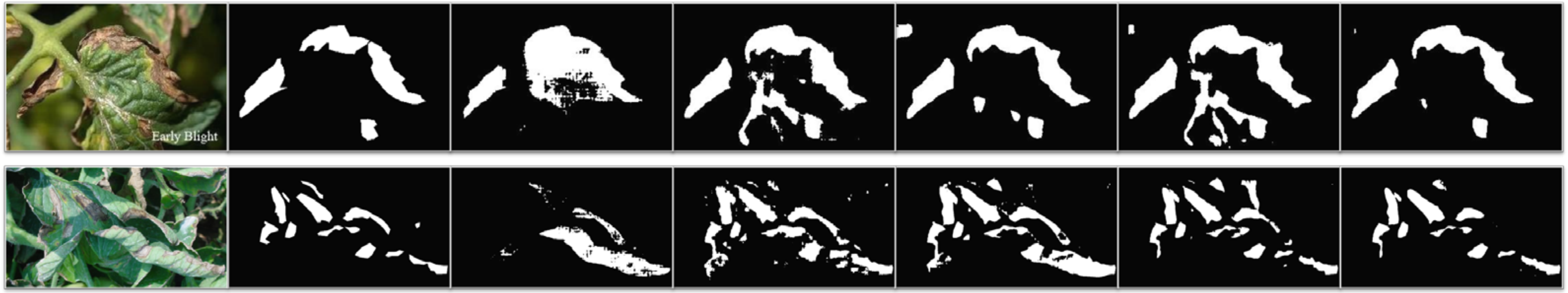
Model	Method	Agriculture IoU (Leaf)	Remote Sensing IoU (Road)
TinySAM	w/o RM	$48.6_{\pm 1.14}$	$28.7_{\pm 1.69}$
	w RM	$\mathbf{50.3}_{\pm \mathbf{0.76}}$	$\mathbf{33.9}_{\pm \mathbf{0.32}}$
MobileSAM	w/o RM	$71.9_{\pm 0.30}$	$59.2_{\pm 0.09}$
	w RM	$\mathbf{73.8}_{\pm \mathbf{0.22}}$	$\mathbf{61.3}_{\pm \mathbf{0.35}}$

✓ Evolution of relation maps and their statistical distributions over epochs, without and with the regularization term.



Experiments

☐ Visualization results



RGB Image | GT | SAM | HQSAM | SU-SAM | ConvLoRA-SAM | InfoSAM

RGB Image | GT | SAM | HQSAM | SU-SAM | ConvLoRA-SAM | InfoSAM

Reference

[1] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4015-4026

[2] Peng Z, Xu Z, Zeng Z, et al. Parameter efficient fine-tuning via cross block orchestration for segment anything model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 3743-3752.

[3] Ke L, Ye M, Danelljan M, et al. Segment anything in high quality[J]. Advances in Neural Information Processing Systems, 2023, 36: 29914-29934.

[4] Xiao A, Xuan W, Qi H, et al. CAT-SAM: conditional tuning network for few-shot adaptation of segmentation anything model[J]. arXiv e-prints, 2024: arXiv: 2402.03631.