JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

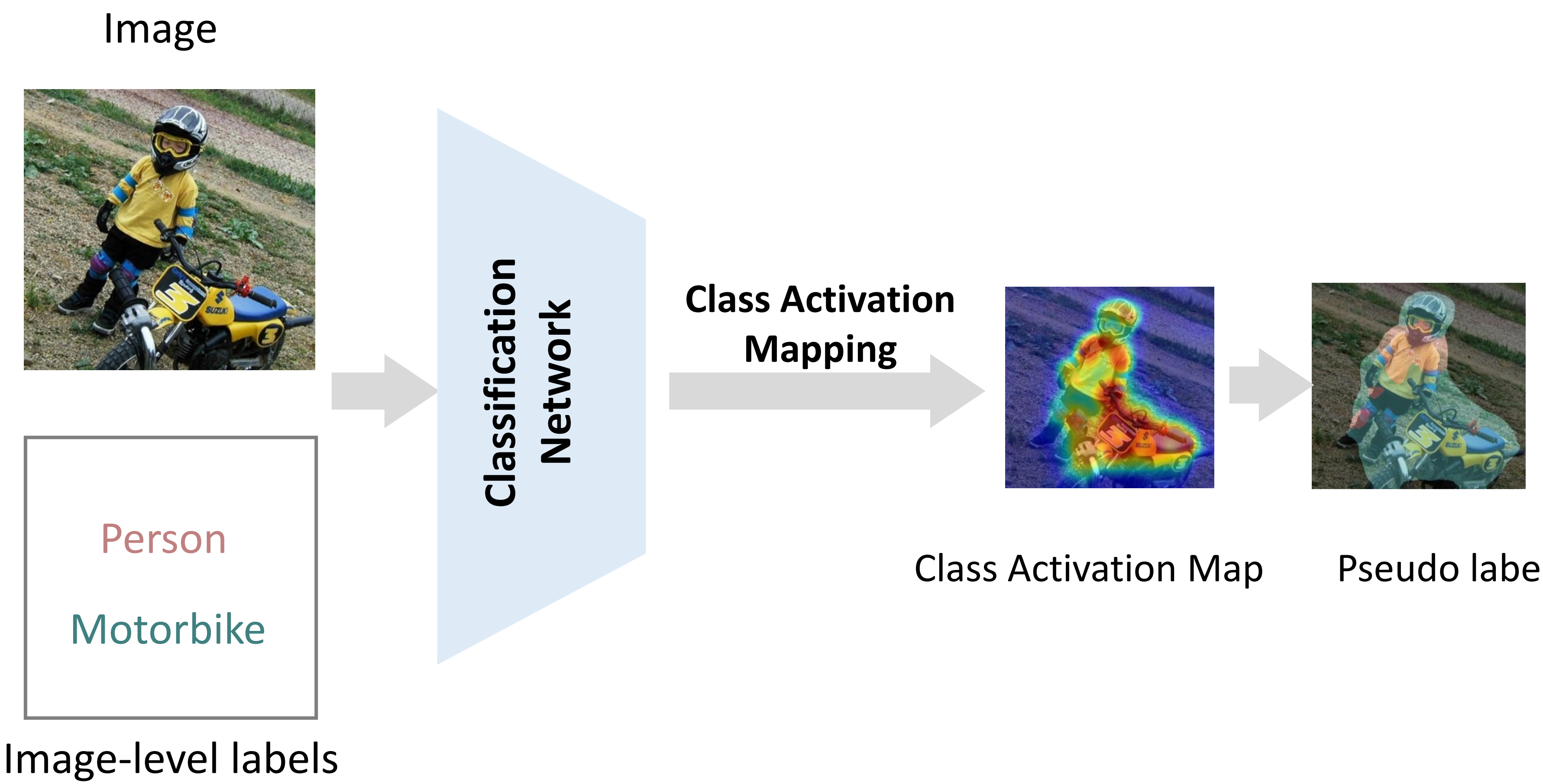
Token Contrast for Weakly-Supervised Semantic Segmentation

Lixiang Ru   Heliang Zheng   Yibing Zhan   Bo Du

rulixiang@outlook.com

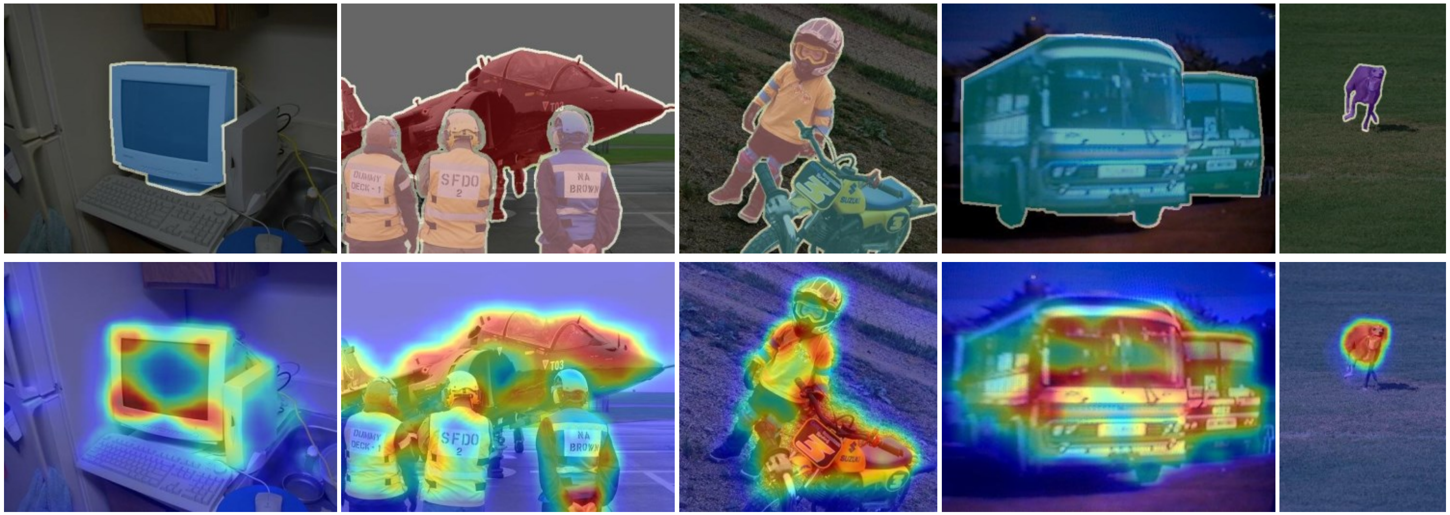
Wuhan university   JD Explore Academy

Weakly-supervised Semantic segmentation

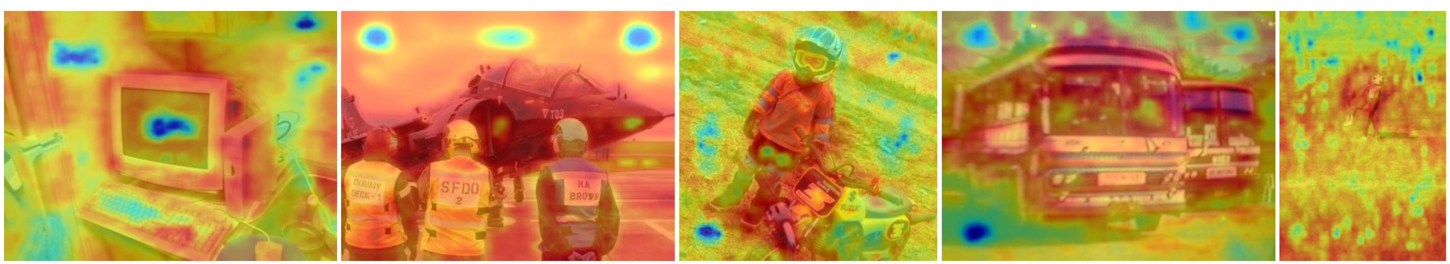


Zhou, Bolei, et al. "Learning deep features for discriminative localization." CVPR 2016.

CAMs generated with deep ViTs cannot recognize different semantics.

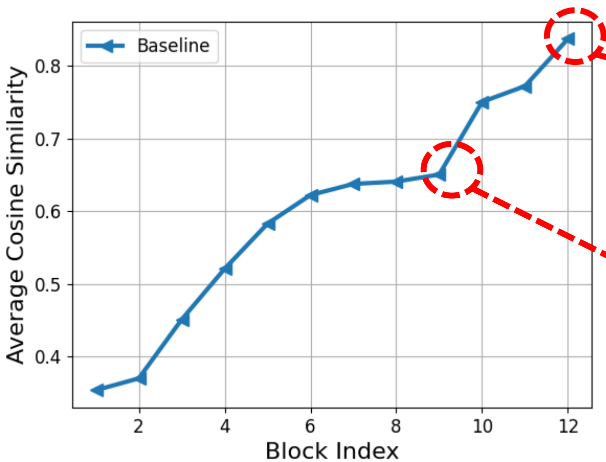


(a) CAM using ViT-S，9 attention blocks

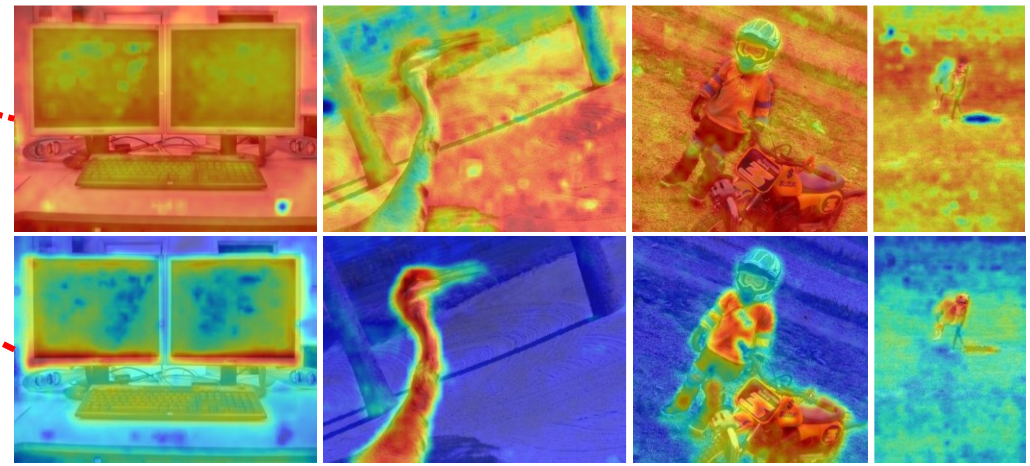


(b) CAM using ViT-B，12 attention blocks

Observation



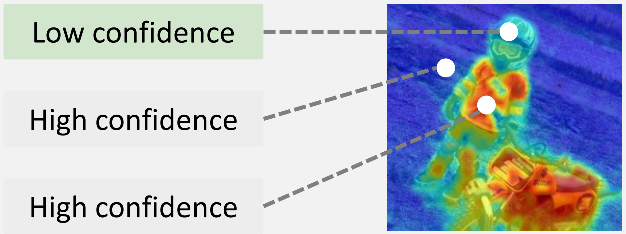
(a) Cosine similarity of patch tokens



(b) CAMs generated from final and intermediate layer

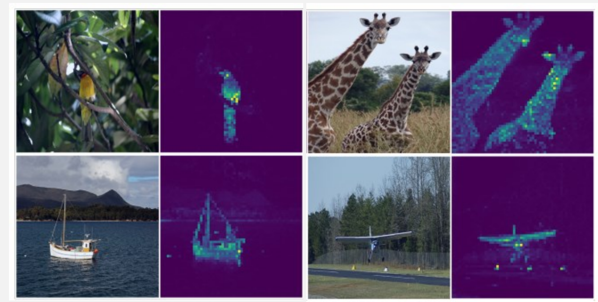
- The intermediate layers can still preserve the semantic diversity.

Observation

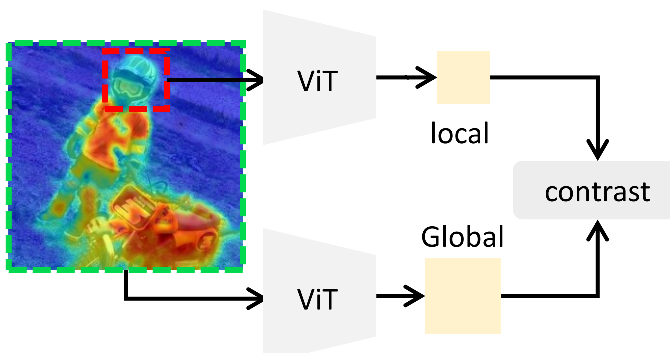


Uncertain regions in CAMs

DINO [Caron etal, 2021]



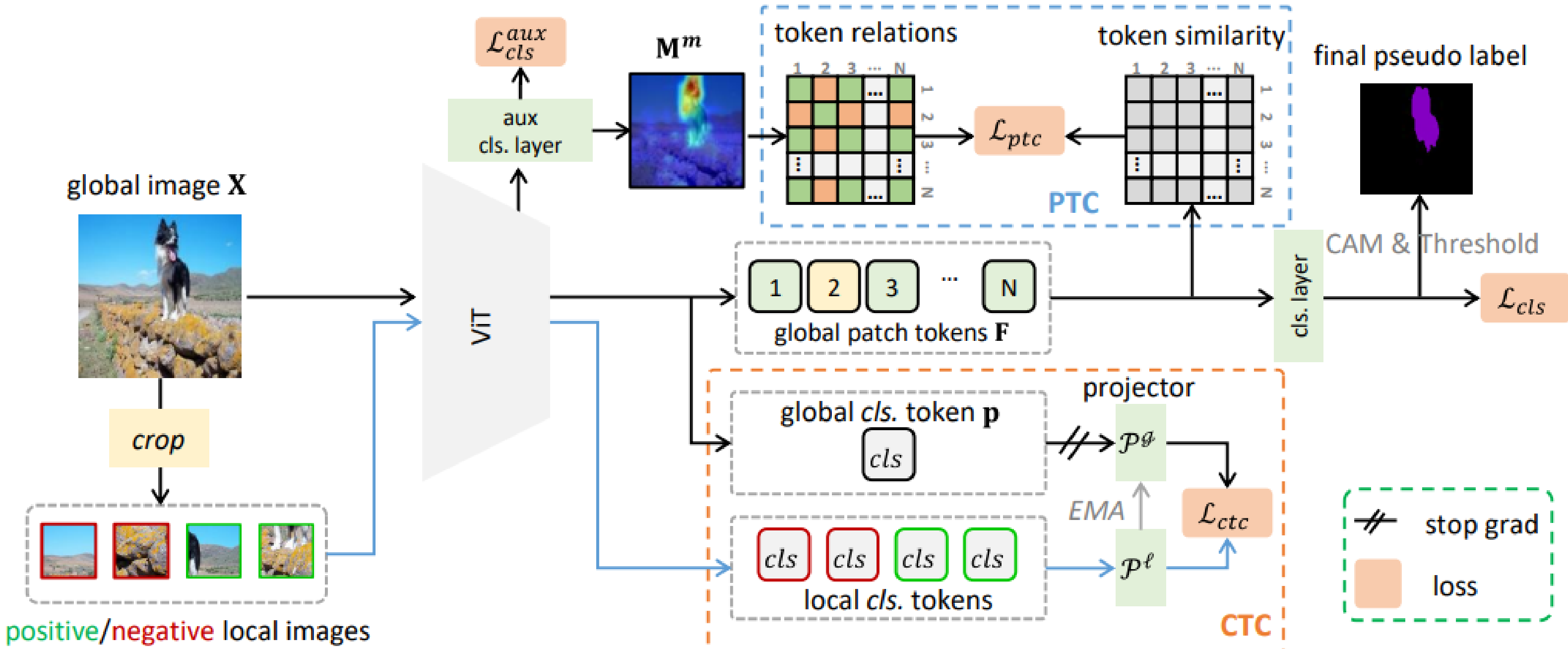
Class token attention map in ViT



- **Differentiate the uncertain regions via aligning the class tokens of local uncertain regions and global regions.**

Caron, Mathilde, et al. "*Emerging properties in self-supervised vision transformers.*" ICCV. 2021

Overall framework

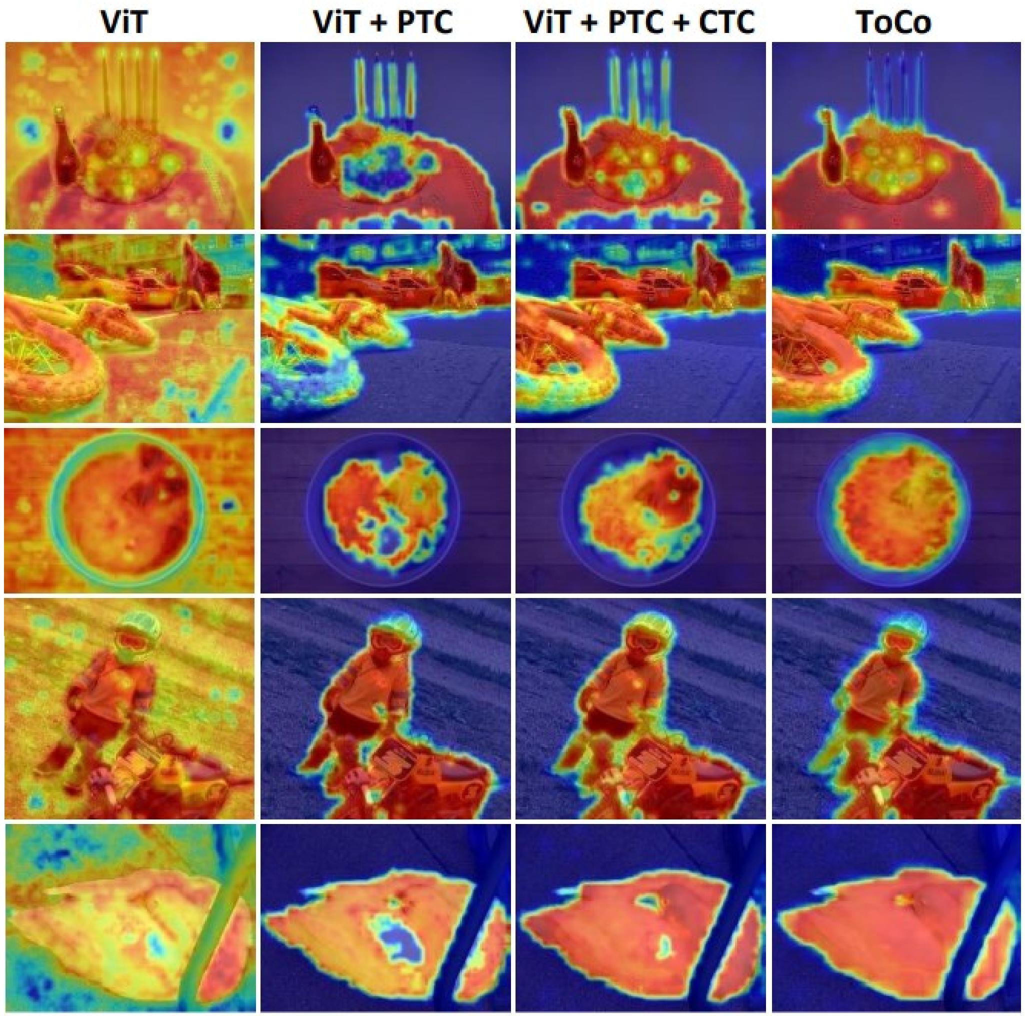


- Patch token contrast loss: $\mathcal{L}_{ptc} = \frac{1}{N^+}\sum_{\mathbf{Y}_i^m=\mathbf{Y}_j^m}\left(1 - CosSim\left(\mathbf{F}_i, \mathbf{F}_j\right)\right) + \frac{1}{N^-}\sum_{\mathbf{Y}_i^m\neq\mathbf{Y}_j^m} CosSim\left(\mathbf{F}_i, \mathbf{F}_j\right)$

- Class token contrast loss: $\mathcal{L}_{ctc} = \frac{1}{N^+}\sum_{\mathbf{q}^+}\log\frac{\exp(\mathbf{p}^\top\mathbf{q}^+/\tau)}{\exp(\mathbf{p}^\top\mathbf{q}^+/\tau)+\sum_{q^-}\exp(\mathbf{p}^\top\mathbf{q}^-/\tau)+\epsilon}$

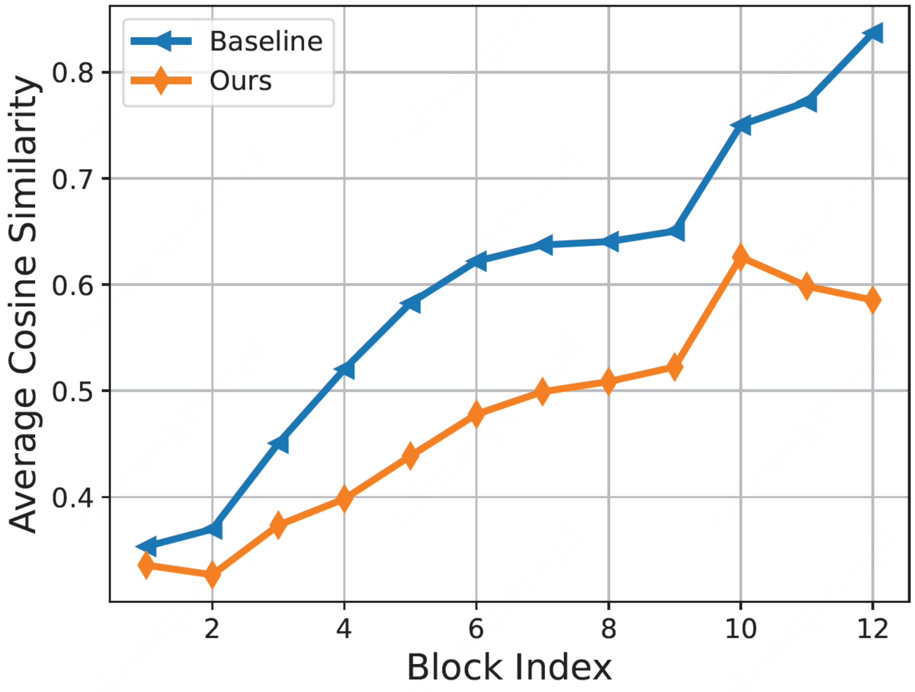
Ablation

Method	Backbone	train	val
RRM [50] AAAI'2020	WR38	–	65.4
1Stage [3] CVPR'2020	WR38	66.9	65.3
AA&LR [52] ACM MM'2021	WR38	68.2	65.8
SLRNet [28] IJCV'2022	WR38	67.1	66.2
AFA [34] CVPR'2022	MiT-B1	68.7	66.5
ViT-PCM [32] ECCV'2022	ViT-B†	67.7	66.0
ViT-PCM + CRF [32] ECCV'2022	ViT-B†	71.4	69.3
ToCo	ViT-B	72.2	70.5
ToCo†	ViT-B†	**73.6**	**72.3**



Method	PTC	CTC	$\mathcal{L}_{seg}$	$\mathcal{L}_{reg}$	M	$\mathbf{M}^m$	*Seg.*
$\mathcal{L}_{cls}$					27.8	–	–
$\mathcal{L}_{cls} + \mathcal{L}^m_{cls}$					27.9	53.8	–
$\mathcal{L}_{cls} + \mathcal{L}^m_{cls}$	✓				62.5	57.8	–
	✓	✓			67.2	60.7	–
	✓	✓	✓		69.9	61.2	66.6
	✓	✓	✓	✓	**70.5**	**62.5**	**68.1**

Analysis of PTC



- PTC module can finely decrease the patch similarity of late layers.

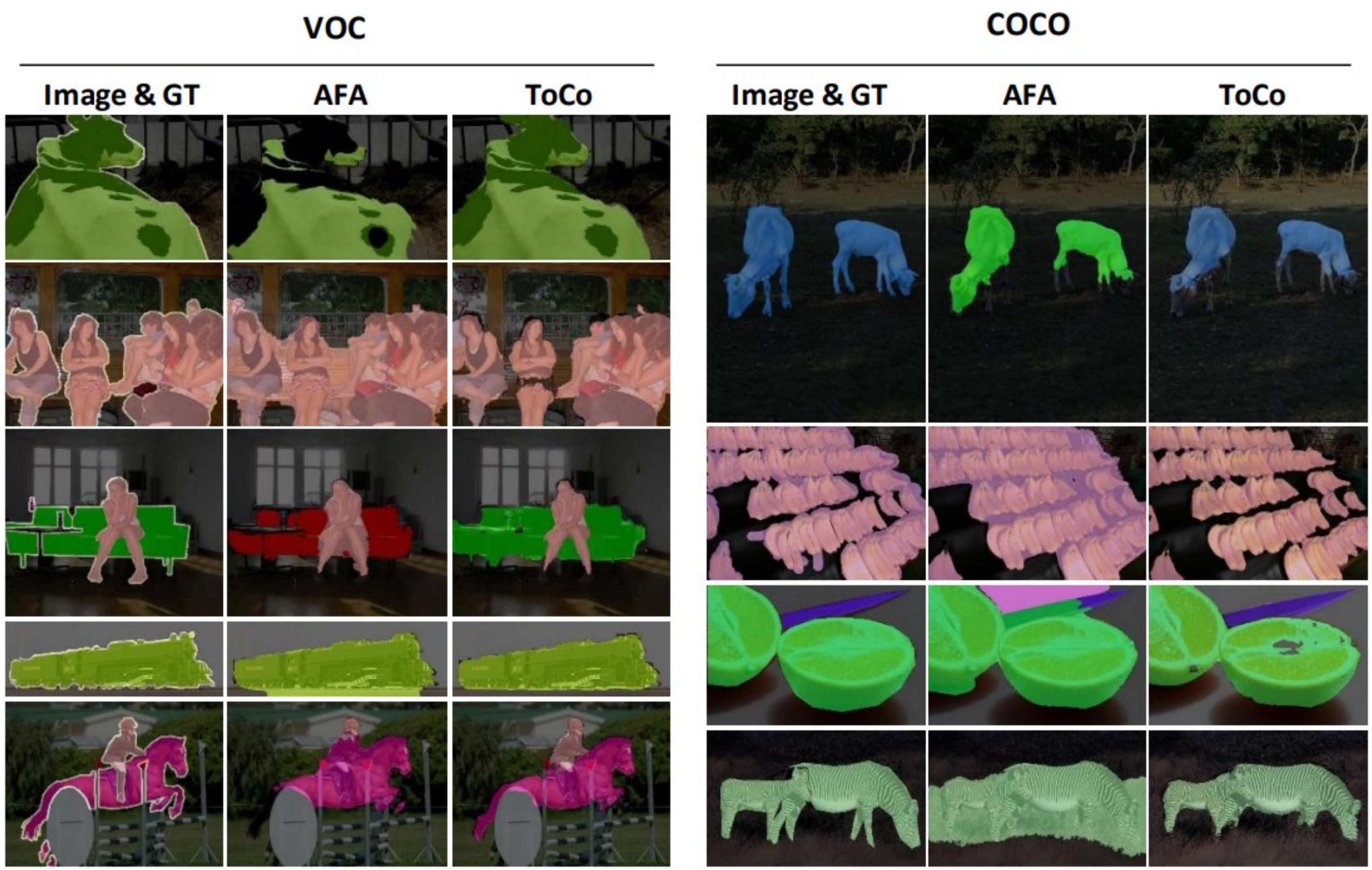
Analysis of CTC

- Attention map in local images can capture non-salient object regions, which are often ignored in global images.

Semantic segmentation results

	Sup.	*Net.*	VOC		COCO
			val	test	val
Multi-stage WSSS methods.					
RIB [21] NeurIPS'2021	$\mathcal{I}+\mathcal{S}$	DL-V2	70.2	70.0	–
EPS [24] CVPR'2021	$\mathcal{I}+\mathcal{S}$	DL-V2	71.0	71.8	–
L2G [19] CVPR'2022	$\mathcal{I}+\mathcal{S}$	DL-V2	72.1	71.7	44.2
RCA [54] CVPR'2022	$\mathcal{I}+\mathcal{S}$	DL-V2	72.2	72.8	36.8
Du *et al.* [13] CVPR'2022	$\mathcal{I}+\mathcal{S}$	DL-V2	72.6	73.6	–
RIB [21] NeurIPS'2021	$\mathcal{I}$	DL-V2	68.3	68.6	43.8
ReCAM [11] CVPR'2022	$\mathcal{I}$	DL-V2	68.4	68.2	45.0
VWL [33] IJCV'2022	$\mathcal{I}$	DL-V2	69.2	69.2	36.2
W-OoD [22] CVPR'2022	$\mathcal{I}$	WR38	70.7	70.1	–
MCTformer [47] CVPR'2022	$\mathcal{I}$	WR38	71.9	71.6	42.0
ESOL [25] NeurIPS'2022	$\mathcal{I}$	DL-V2	69.9	69.3	42.6
Single-stage WSSS methods.					
RRM [50] AAAI'2020	$\mathcal{I}$	WR38	62.6	62.9	–
1Stage [3] CVPR'2020	$\mathcal{I}$	WR38	62.7	64.3	–
AFA [33] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
SLRNet [28] IJCV'2022	$\mathcal{I}$	WR38	67.2	67.6	35.0
ToCo	$\mathcal{I}$	ViT-B	69.8	70.5[1]	41.3
ToCo†	$\mathcal{I}$	ViT-B†	**71.1**	**72.2**[2]	**42.3**

Semantic segmentation results



https://github.com/rulixiang/ToCo