Boosting Low-Data Instance Segmentation by Unsupervised Pretraining with Saliency Prompt

Hao Li, Dingwen Zhang*, Nian Liu, Lechao Cheng, Yalun Dai, Chao Zhang, Xinggang Wang, Junwei Han

西北工業大學 NORTHWESTERN POLYTECHNICAL UNIVERSITY
之江实验室 ZHEJIANG LAB
華中科技大學 Huazhong University of Science and Technology
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE
中国科学院大学 University of Chinese Academy of Sciences

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Problem Definition and Contribution

Motivation

Query-based end-to-end instance segmentation (QEIS) methods would lose efficacy when only a small amount of training data is available since it's hard for the crucial queries/kernels to learn localization and shape priors.

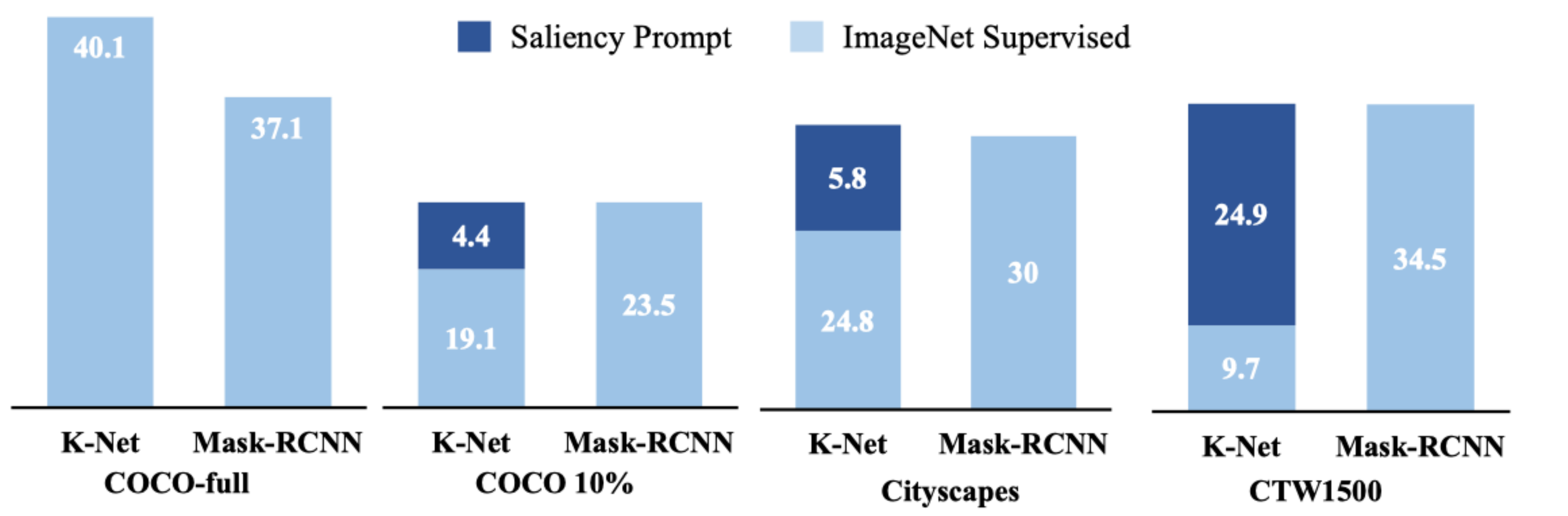


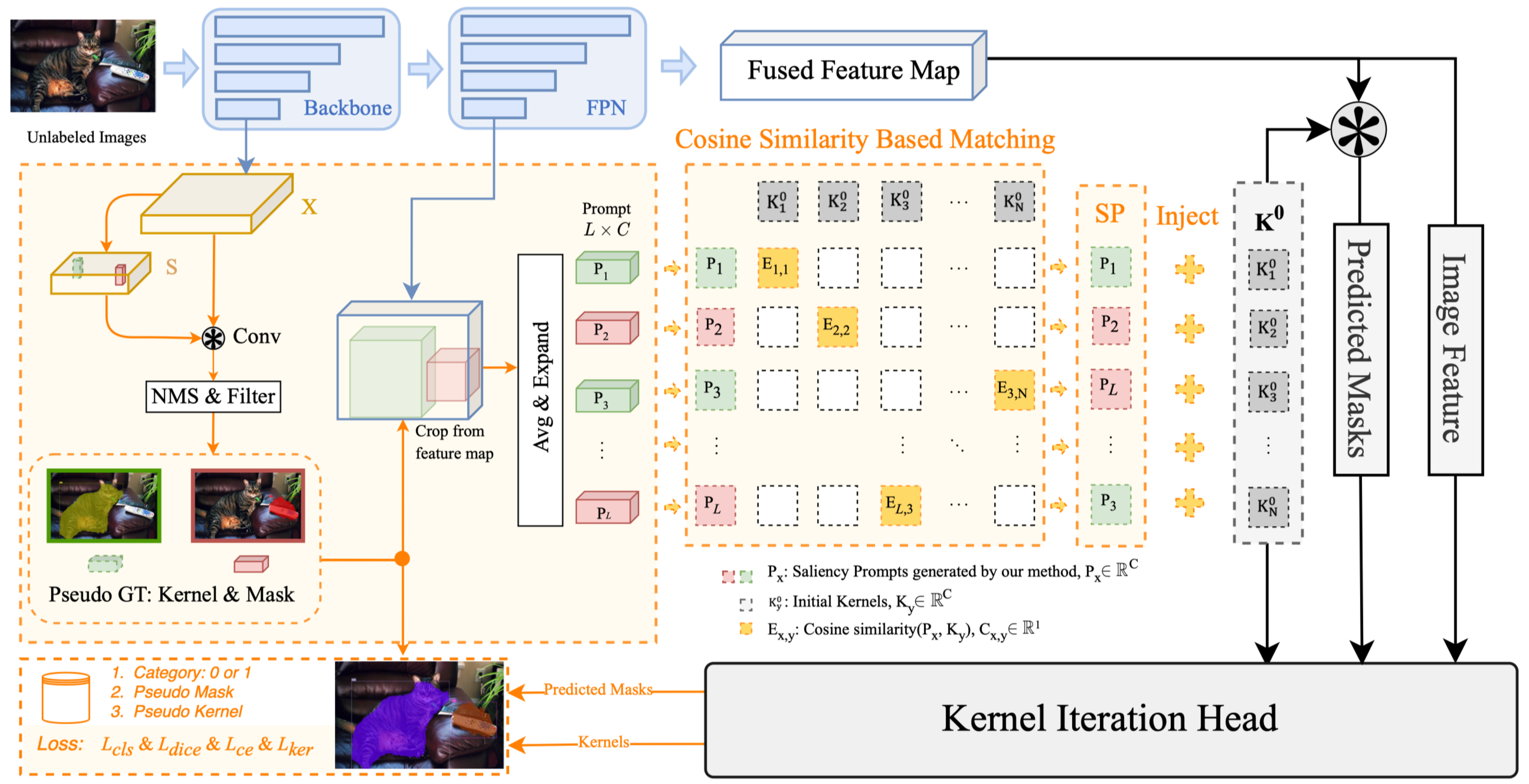
Figure 1: K-Net can outperform Mask-RCNN on large-scale datasets (COCO-full). However, on small datasets (the right three), it can not perform as well as Mask-RCNN since it's hard to learn localization and shape priors. Our proposed unsupervised pre-training method based on saliency prompt not only boosts the vanilla K-Net significantly, but also helps to achieve comparable performance compared with Mask-RCNN.

Key Contribution

1. This paper first points out that the QEIS models lack spatial distribution and shape awareness and perform poorly in low-data regimes.
2. Introduce a new pre-training method that boosts QEIS models by giving Saliency Prompt for queries/kernels.
3. From a practical perspective, our pre-training method helps QEIS models achieve a similar convergence speed and comparable performance with CNN-based models in low-data regimes.

Our Method

Overview.
Orange colors denotes our pre-training method with the corresponding supervision. Blue and gray modules denote a vanilla QEIS model, here we use K-Net for example.



- **Saliency Masks Proposal**

Responsible for generating pseudo masks from unlabeled images based on the saliency mechanism.

$$\mathbf{Y}_{i,j} = \mathrm{Conv}\left(\mathbf{S}_{i,j}, \mathbf{X}\right) \in \mathbb{R}^{H \times W}$$

- **Prompt-Kernel Matching**

Transfers pseudo masks into prompts and injects the corresponding localization and shape priors to the best-matched kernels.

$$\mathbf{E}_{n,l} = \frac{\mathbf{K}_n^0}{\|\mathbf{K}_n^0\|_2} \cdot \frac{\mathbf{P}_l}{\|\mathbf{P}_l\|_2}. \quad \delta(n) = \underset{l \in [1,\dots,L]}{\arg\max}\, \mathbf{E}_{n,l}.$$

- **Kernel Supervision**

Applied to supply supervision at the kernel level for robust learning.

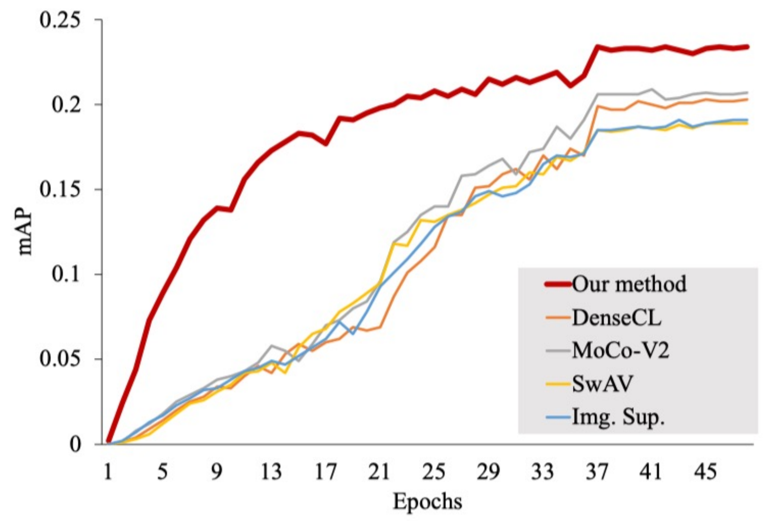
$$\mathcal{L}_{ker} = \sum_{l=0}^{L} \sum_{i} (1 - \mathrm{Cos}(\mathrm{Linear}(\mathbf{S}_l), \mathbf{K}_{n_l}^i)),$$

Experimental result

- **Compared with SOTA unsupervised methods**

Table 1. Instance segmentation fine-tune results on COCO with 5% and 10% annotated images based on K-Net.

	Pre-train	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
5% images	Img. Sup.	14.8	29.1	13.7	4.3	15.5	24.4
	DenseCL	16.7	31.2	15.9	5.1	17.5	27.7
	SwAV	15.7	30.3	14.7	4.6	25.9	16.6
	MoCo-v2	17	32	16.2	5.3	18.3	27.1
	SP(ours)	**19.9**	**35.7**	**19.9**	**6.0**	**21.0**	**32.6**
10% images	Img. Sup.	19.1	35.7	18.2	6.7	20	31.6
	DenseCL	20.3	36.4	20.3	6.6	21.8	33.6
	SwAV	18.9	34.8	18.3	6.8	20.8	30.6
	MoCo-v2	20.7	37.7	20.4	6.4	22.1	34.2
	SP(ours)	**23.5**	**41.4**	**23.7**	**7.9**	**24.8**	**38.6**

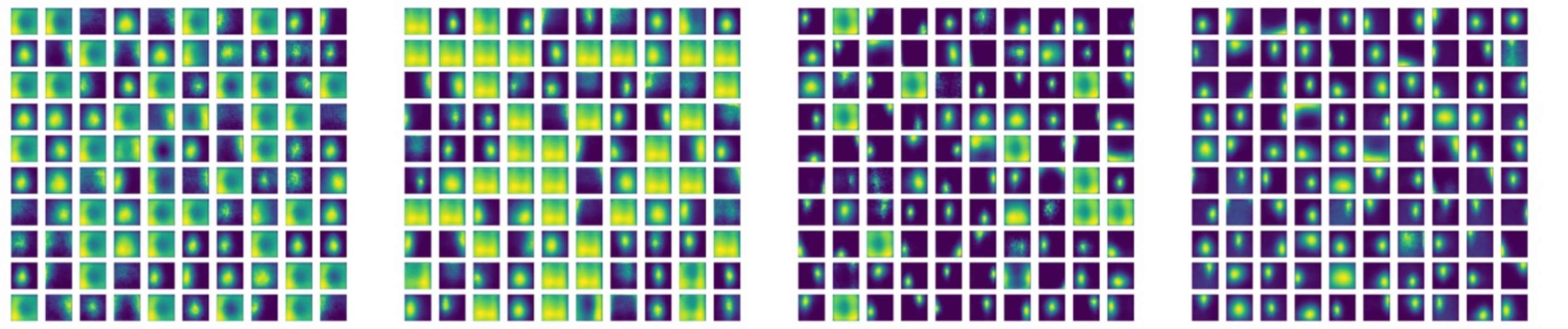


- **Deployed on QueryInst and Mask2Former**

Model	Pre-train	CTW1500 Epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Cityscapes Epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask2Former [7]	Img. Sup.	80	38.8	67.6	41.6	15.6	41.2	57.4	24	29.1	52.4	-	5.6	23.9	55
	DenseCL	80	43.2	71.6	48.5	18.4	47.6	59.9	24	27.5	48.9	-	5.2	23.8	53.7
	SwAV	80	41.2	69.1	46.1	17.6	45.2	58.1	24	30.3	53.3	-	5.4	23.3	59
	MoCo-v2	80	43.3	71.2	49.2	18.9	47.6	59.4	24	30.7	54.3	-	5.4	25.5	56.4
	SP(ours)	**20**	**52.9**	**83.4**	**62.1**	**29.4**	**56.4**	**67.6**	**24**	**31.8**	**55.8**	**-**	5.1	**26.5**	**59.0**
QueryInst [15]	Img. Sup.	80	28.3	53.7	28.6	9.8	29	41.8	24	29.1	53.2	-	6.7	27.4	50.7
	DenseCL	80	31.6	56.7	33.4	10.4	32.5	46.6	24	30.8	54.7	-	8.6	28.9	54.5
	SwAV	80	24.6	50	23.1	8.1	25	36.3	24	30.7	54.4	-	7.9	28.5	53.9
	MoCo-v2	80	31.6	56.8	32.8	12.6	32	45.8	24	31.4	54.4	-	8.1	28.4	56.1
	SP(ours)	**20**	**39.2**	**66.8**	**43.1**	**16.7**	**42.2**	**51.9**	**24**	**32.8**	**57.3**	**-**	**8.8**	**29.2**	**57.0**

These results indicate that our pretraining method can help the kernels/queries of QEIS models to learn localization and shape prior effectively and help gain competitive performance improvement.

- **Visual analysis for each query/kernel**



(a) Train from scratch. (b) Pre-trained by MoCo-v2. (c) Pre-trained by our method. (d) Trained on COCO train2017-full.

These results demonstrate that the kernels pre-trained with Saliency Prompt have learned effective spatial distribution and shape discrimination ability.

GitHub: WeChat: