Dynamic Support Information Mining for Category-Agnostic Pose Estimation (Supplementary Material)

Pengfei Ren, Yuanyuan Gao, Haifeng Sun, Qi Qi, Jingyu Wang*, Jianxin Liao*
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{rpf, gaoyuanyuan, hfsun, qiqi8266, wangjingyu, liaojx}@bupt.edu.cn

In the supplemental material, we provide:

- more details of network structure in Sec. 1,
- cross super-category experiments in Sec. 2,
- more qualitative results in Sec. 3,

Note that all the notation and abbreviations here are consistent with the main manuscript.

1. Details of Network Structure

We use the lowest resolution feature map among the multi-scale features output by HRNet-32 [1] as the low-resolution feature map $\mathbf{F}_{lr}$. We concatenate the multi-scale features and obtain high-resolution feature map $\mathbf{F}_{hr}$ through a 1×1 convolutional layer. The channel dimension of the visual feature map and keypoint features is 256. The GCN used for keypoint reconstruction and query keypoint information interaction both contain 4 residual graph convolution modules as [8]. For the deformable attention module, each keypoint predicts four offset vectors for image feature sampling. The number of heads for the multi-head self-attention is set to 8. Similar to CapeFormer [4], we adopt 3 iteration refinements. Inspired by CapeFormer [4], Instance Order Encoding and Keypoint Identifier Encoding adopt fixed sinusoidal embedding [6] to avoid overfitting to the trained categories.

2. Cross Super-Category Experiments

In order to further evaluate the generalization ability of SPDNet, we adopt the 'Leave-One-Out' strategy to perform cross-super-category experiments. Specifically, we use one super-category data as the test set and the remaining data as the train set. Similar to previous work [4, 7], the super-categories to be evaluated include the human body, human face, vehicle, and furniture. As shown in Table 1, using the same super-category splits, SPDNet outperforms CapeFormer [4] in four super-categories, especially in the Furniture category. Cross super-categories pose estimation is more challenging, demonstrating our method's superiority in robustness and generalization capabilities.

Table 1. Cross super-category evaluation (PCK). Experiments are conducted under 1-shot setting.

Method	Human Body	Human Face	Vehicle	Furniture
ProtoNet [5]	37.61	57.80	28.35	42.64
MAML [2]	51.93	25.72	17.68	20.09
Fine-tune [3]	52.11	25.53	17.46	20.76
POMNet [7]	73.82	79.63	34.92	47.27
CapeFormer [4]	83.44	80.96	45.40	52.49
SPDNet	**83.84**	**81.24**	**45.53**	**53.08**

3. More Qualitative Results

In this section, we present more qualitative results. As shown in the first row of Fig. 1, due to the similar appearance of antlers and eyes, CapeFormer mistakenly regarded antlers as eyes, while our method accurately predicted the positions of the left and right eyes. At the same time, as shown in the second and third rows of Fig. 1, when an eye cannot be observed due to self-occlusion, CapeFormer's prediction makes a severe error, while our method accurately predicts the positions of the occluded eye based on the symmetry structures of the eyes. Additionally, our method can focus on fine-grained visual information. As shown in the last row, our method can pay attention to some visual regions that are easily overlooked due to low-light environment.

Benefitting from structured relationship modeling, the keypoints predicted by SPDNet have a more reasonable overall structure. For example, as shown in the first row of Fig. 2, our method can well maintain the cuboid structure of the bed, while CapeFormer's prediction has obvious structural deformation. Existing methods cannot understand the target category from a global perspective, so the support structure of the chair predicted by CapeFormer is

*Corresponding author

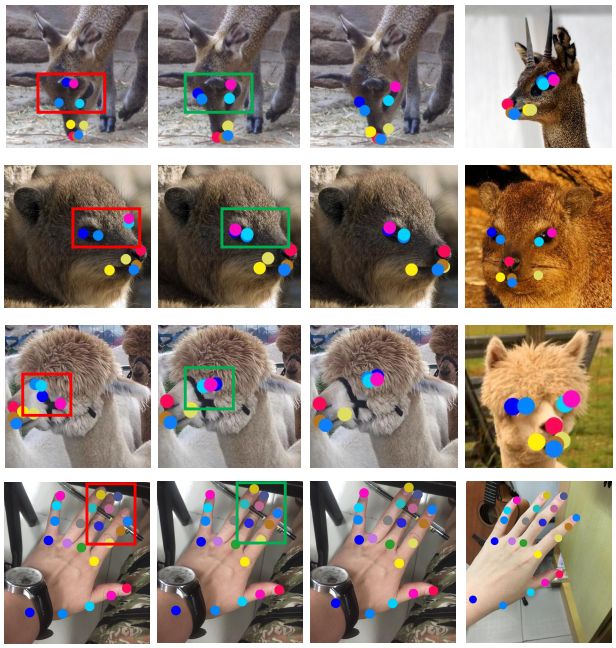

Figure 1. Qualitative results. We visualize the keypoint predictions under the 5-shot setting. We use red and green boxes to highlight keypoints that are difficult to estimate with CapeFormer.

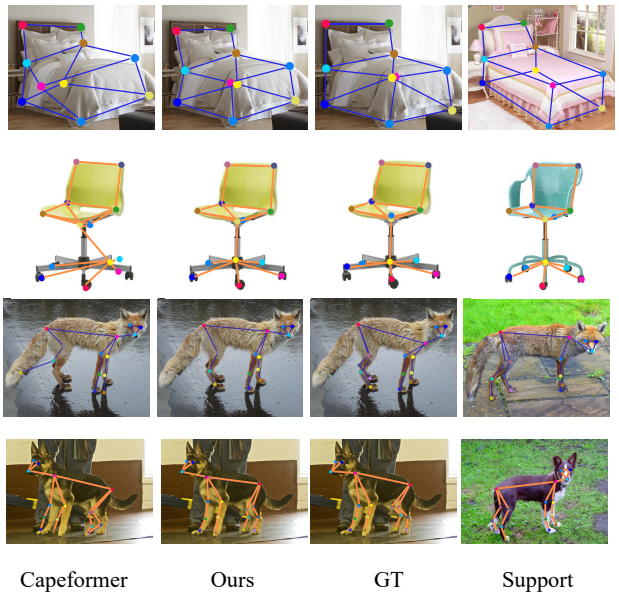

Figure 2. Qualitative results. We visualize the keypoint predictions under 5-shot setting. The bones are not predicted by our network, but provided by the dataset.

unreasonable. In contrast, our method can accurately predict it. At the same time, as shown in the third and fourth rows of Fig. 2, our method can predict the overall pose of the animal more robustly. However, as shown in Fig. 3, our method is prone to estimation errors when faced with multiple coupled complex environments such as local similarity, occlusion and low brightness.

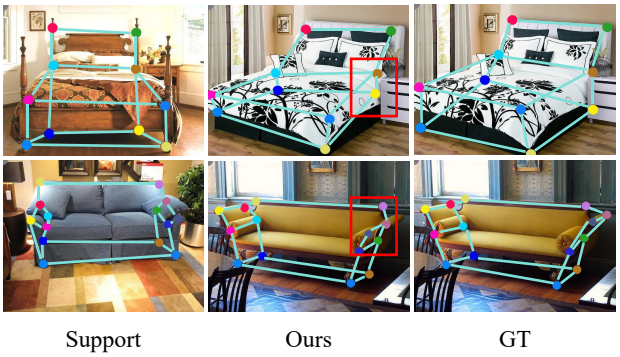

Figure 3. Failure examples. Red boxes highlight keypoints with obvious errors.

In conclusion, our method can perceive fine-grained visual information and global structure information. Furthermore, our method is robust to common occlusion and self-similarity appearance.

References

[1] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. 2017.

[3] Akihiro Nakamura and Tatsuya Harada. Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216*, 2019.

[4] Min Shi, Zihao Huang, Xianzheng Ma, Xiaowei Hu, and Zhiguo Cao. Matching is not enough: A two-stage framework for category-agnostic pose estimation. In *CVPR*, 2023.

[5] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NeurIPS*, 2017.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[7] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *ECCV*. Springer, 2022.

[8] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019.