# SCAPE: A Simple and Strong Category-Agnostic Pose Estimator Supplemental Materials

Yujia Liang, Zixuan Ye, Wenze Liu, and Hao Lu

National Key Laboratory of Multispectral Information Intelligent Processing Technology; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

## 1 Appendix

This supplementary material includes the following parts:

- Additional results of SCAPE and on all splits of MP-100;
- More details about the proposed framework;
- Additional ablation experiments;
- Further Exploration for CAPE;
- The applicability of SCAPE on more scenarios.
- The code for SCAPE and and the testing for Painter [9] on MP100.
- More qualitative results;
- Visual comparison of attention maps between our state-of-the-art methods and ours.

### A Additional results of SCAPE on MP-100

**Additional metrics reported.** As mentioned in Section 4.1, PCK has some limitation on evaluating the performance. Here we report two additional metrics including Area Under the Curve (AUC) [3] and Normalized Mean Error (NME) [2] for CapeFormer and ours on the MP-100 dataset as shown in Table 1.

**Complete results for Table 3** We supplement the complete results of Table 3 in main content. across five splits on both 1-shot and 5-shot setting on the MP-100 dataset as shown in Table 2.

### B More details about the framework

**Downsampling the feature output by transformer backbones.** Due to the inconsistent resolution output by ResNet ($8 \times 8$) and transformer backbones ($16 \times 16$), we add an average pooling layer at the end of the used transformer backbones for workload reduction and fair comparison.

**The structural differences between SCAPE (without GKP and KAR) and SCAPE (with GKP and KAR).** SCAPE (w/o GKP and KAR) consists of six layers of feature interaction. To enhance the quality of the attention map, we extends SCAPE by incorporating two designs, Global Keypoint Feature Perceptor (GKP) and Keypoint Attention Refiner (KAR). In practice, the first two self-attention blocks are replaced by Light GKP. The GKP is a cross-attention layer which refines the initial support keypoint tokens, with a total of $M = 2$ layers in SCAPE. And the KAR is inserted into each feature interaction (self-attention) stage to refine the attention maps among keypoints. In a nut shall, to maintain the efficiency, we decrease the number of feature interaction layers $N = 6$ in SCAPE (w/o GKP and KAR) to $N = 4$ in SCAPESCAPE (w GKP and KAR). Moreover, the Lite-SCAPE model has $M = 1$ GKP module and $N = 2$ interactive modules.

**The results of Lite-SCAPE on the MP-100 dataset.** Table 3 provides the full metrics as supplement for Table 4 in main content, where one can see that Lite-SCAPE performs well on all five splits of the MP-100 dataset.

**Table 1:** comparison with the results of AUC(↑) and NME(↓) on 5 splits under the 1-shot setting of the MP-100 dataset. Best performance is in **boldface**.

| method | metric | split1 | split2 | split3 | split4 | split5 | mean |
|---|---|---|---|---|---|---|---|
| CapeFormer | NME↓ | 0.088 | 0.110 | 0.111 | 0.116 | 0.108 | 0.106 |
| | AUC↑ | 88.64 | 86.39 | 86.18 | 85.81 | 86.51 | 86.70 |
| SCAPE | NME↓ | **0.078** | **0.101** | **0.097** | **0.105** | **0.101** | **0.096** |
| | AUC↑ | **89.59** | **87.12** | **87.34** | **86.28** | **86.96** | **87.45** |

**Table 2:** Performance across different Transformer-based backbones, we report all 5 splits on the MP-100 dataset, considering both 1-shot and 5-shot settings. Best performance is in **boldface**.

| method | backbone | shot | split1 | split2 | split3 | split4 | split5 | mean |
|---|---|---|---|---|---|---|---|---|
| SCAPE | ViT-B | 1 | 91.74 | 87.57 | 87.70 | 86.49 | 87.46 | 88.19 |
| | | 5 | 94.83 | 90.65 | 90.94 | 90.98 | 90.19 | 91.52 |
| SCAPE | Swin-S | 1 | 91.66 | 87.01 | 86.98 | 85.97 | 87.91 | 87.91 |
| | | 5 | 95.18 | 91.25 | 91.78 | 90.74 | 91.10 | 92.01 |
| SCAPE | ViT-S (DINOv2) | 1 | 94.47 | 89.55 | 89.81 | 89.04 | 90.85 | 90.74 |
| | | 5 | 96.29 | 92.11 | 90.48 | 92.27 | 92.11 | 92.65 |
| CapeFormer | ViT-B (DINOv2) | 1 | 93.43 | 89.03 | 87.50 | 86.32 | 89.31 | 89.11 |
| | | 5 | 95.34 | 92.10 | 90.84 | 90.60 | 90.71 | 91.92 |
| SCAPE | ViT-B (DINOv2) | 1 | **95.01** | **90.65** | **90.65** | 90.50 | **92.97** | **91.95** |
| | | 5 | **97.10** | **93.28** | **92.02** | **92.83** | **94.67** | **93.98** |

**Table 3:** Lite-SCAPE on the MP-100 dateset on 5 splits under the 1-shot setting

| split1 | split2 | split3 | split4 | split5 | mean(PCK) |
|---|---|---|---|---|---|
| 90.01 | 85.17 | 85.45 | 84.91 | 85.14 | 86.13 |

**Table 4:** Design of Global Keypoint Feature Perceptor .

| | Support image | Query image | PCK |
|---|---|---|---|
| D1 | ✓ | | 90.0 |
| D2 | | ✓ | 89.3 |
| D3 | ✓ | ✓ | ✓ 90.3 |

**Table 5:** The Global Keypoint Feature Perceptor must be a separate module and cannot be integrated into the feature fusion layer.

| KGP | Interactor | Support image | PCK |
|---|---|---|---|
| 0 | 6 | | 89.8 |
| 0 | 6 | ✓ | 89.3 |
| 2 | 4 | | 90.3 |

## C Additional Ablation Studies

**The design of Global Keypoint Feature Perceptor.** In Section 4.4, we focus on the design of GKP, as depicted in Table 4, GKP enriches the semantic content of the initial keypoint tokens by interacting again with the support image. In implementation of attention, the term "key" and "value" encompass not only support images but also include $F_q$ (query images). By comparing D1, D2 and D3, interacting the support image is essential with GKP and allowing the support keypoints to pre-examine the $F_q$ intended for fusion can further improve performance. Then, as the core of our GKP lies in interacting with the support image. We explore the possibility of incorporating this interaction into the following feature interactor. That is to say, in addition to concatenating $F_s$ and $F_q$ as before, we also concatenate the support image and update all three in feature interaction. According to Table 5, performance of the second line drops when incorporating the support image. The expected match for the feature interaction module is only the query image, and including support images disrupts matching. Therefore, the support image cannot be introduced during the feature interaction stage.

**Global Keypoint Feature Perceptor can expedite convergence.** Learning curves in Fig.1 S1 illustrate that GKP accelerates convergence and enhances final performance. With the assistance of the GKP, SCAPE achieves optimal performance of 90.3 within 160 epochs, whereas previous methods were trained for 210 epochs.

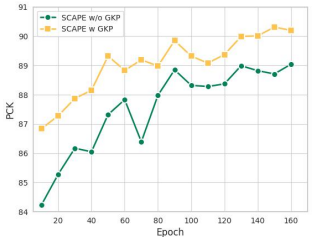


**Fig. 1: The result of SCAPE w/ and w/o the GKP.** We train the model for a total of 160 epochs under the 1-shot setting on split-1 of MP-100, and evaluate the PCK metric every 10 epochs. The x-axis and y-axis represents the number of epochs and the PCK respectively.

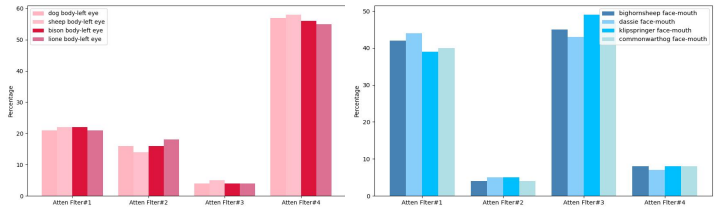


**Fig. 2: Similar support keypoints exhibit analogous weight distributions (for the weight assigner).** Even though they belong to different categories, their keypoints are the same, and the inter-node relationships is similar, resulting in closely aligned weight distributions.

**Weight visualization of Weight Assigner in Keypoint Attention Refiner.** With multiple node relationships in different aspects generated by multiple Attention Filters, the Weight Assigner in KAR is used to selectively keep and discard the relationships. Since the weights are obtained by the corresponding support keypoint token, as mentioned in Section 3.4, therefore support keypoint tokens with similar meaning (left eye of different animal) should yield more similar weights. We visualize the weights in Fig. 2, one can see that the similar support keypoint tokens generate similar weights for the 4 Attention Filters. Likewise, as shown in Fig. 3, dissimilar support keypoints bring distinct weight distributions.

**Ablation study on unshared q and k for support keypoint tokens and query images.** The linear projection before self-attention can be formulated as:

$$\begin{aligned} K_s = W_{K1}F_s\,, K_q = W_{K2}F_q\,, K = \text{concat}(K_s, K_q)\,, \\ Q_s = W_{Q1}F_s\,, Q_q = W_{Q2}F_q\,, Q = \text{concat}(Q_s, Q_q)\,, \end{aligned} \tag{1}$$

where $W_{K1}$, $W_{K2}$, $W_{Q1}$ and $W_{Q2}$ represent the linear projection matrix to generate key $K_s$, $K_q$ and query $Q_s$, $Q_q$ for support feature $F_s$ and query feature $F_q$. To eliminate the potential impact of increased parameters in Eq. (1) compared to the vanilla self-attention, we conduct an experiment where the linear layers

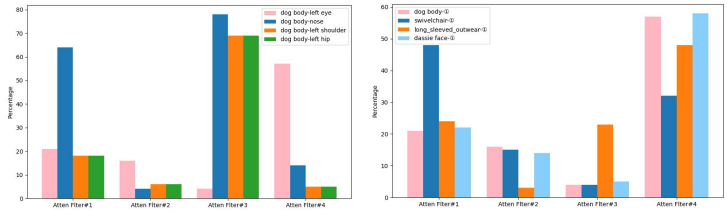


**Fig. 3: The weight distributions of dissimilar support keypoints differ.** On the left side, nodes within the same category exhibit varying distributions due to different ways of modeling inter-node relationships. The right side illustrates the distinct weight distributions for the first keypoints in different categories.

are shared, *i.e.*, $W_{K1} = W_{K2}$ and $W_{Q1} = W_{Q2}$ to maintain the same number of parameters with the vanilla self-attention. Table 6 shows that the our method performs comparably either with or without additional parameters, which implies that it is the mechanism rather than the more parameters that boosts the performance.

**Table 6:** Ablation study on parameter increase resulting from Non-Sharing of q and k in Feature Interactor

| share | layer | PCK |
|---|---|---|
| ✓ | 1 | 89.1 |
| ✓ | 2 | 89.3 |
| | 1 | 89.8 |

## D Further Exploration for CAPE

**The generality of GKP and KAR.** i) We apply our GKP to the state-of-the-art CapeFormer on split 1, enhancing PCK by 0.6. ii) By incorporating KAR into the first three layers of the Encoder for interaction, PCK is further improved by 0.8. iii) The combined effect of these two modules results in an overall enhancement of 1.1 PCK.

**Whether CAPE-specific models are needed**

i) Models like VIT, trained in the DINOv2 [4] manner, demonstrate fine-grained matching effects across various tasks [5, 7]. Stable Diffusion [6] has also exhibited similar capabilities in DIFT [8]. We evaluated these approachs on the MP100 dataset following the procedure outlined in DIFT [8]. Provided support image and heatmap ground truth for target points, we utilized the visual extractor from Stable Diffusion to extract support features. These features were then multiplied by the heatmap ground truth to obtain support keypoints. Calculating the cosine similarity between keypoints and the query image produced

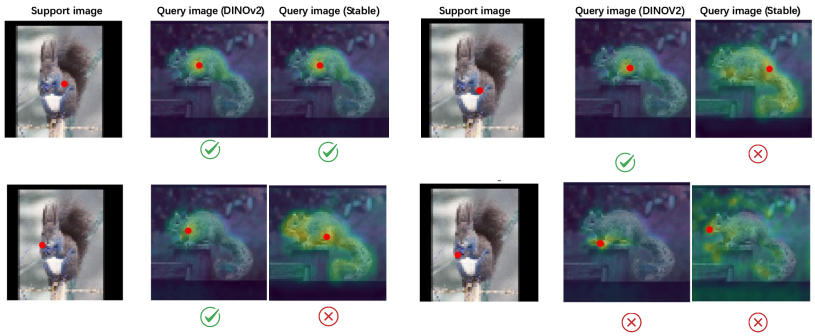


**Fig. 4: Visualization of similarity maps.** The second column corresponds to DINOv2, and the third column represents Stable Diffusion. Red points indicate the final predicted points, with a checkmark denoting correct predictions and a cross indicating incorrect predictions.

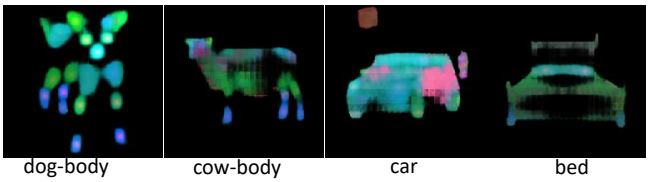


**Fig. 5: Visualization of the rendered style maps for keypoint by Painter.**

a similarity map, $S$. We identified the target point P as the `argmax`(S), yielding the final target point. We applied a similar evaluation to DINOv2. We visualized some of these similarity maps, as shown in Fig. 4. While they effectively handle simple cases, they face challenges when encountering occlusion or significant appearance differences between support and query images.

ii) Several recent papers on generalist models learning assert the ability to solve diverse tasks with just a few task examples, as seen in Painter [9] or general approach for visual prompting [1]. It would be interesting to evaluate whether CAPE-specific models are even needed. Here we show the result of generalist model ('Images Speak in Images') [9] on MP-100 with identical metrics and settings. As shown in the table below, the performance is significantly lower than the CAPE methods. Per visualizations in Fig. 5, the model can only output the shape of furniture and vehicles, but fails to precisely localize the points. Hence, we believe that generalist model cannot replace CAPE models at this moment. The evaluation code is available in the supplementary material folder.

The results of the three approaches are reported in Table 7 using consistent metrics. Through a combination of quantitative metrics and visualizations, it becomes evident that the proprietary model of CAPE is needed. Moreover, these general models are notably heavier compared to SCAPE.

**Table 7: Results for various general models on MP100 split1.** Specifically, Stable Diffusion performs better when provided with textual cues (right eye) compared to cases without prompts.

| Method | spilt1 |
|---|---|
| DINOv2-L | 78.1 |
| DIFT | 60.0 |
| DIFT (propmt) | 66.8 |
| Painter | 23.9 |
| SCAPE (ours) | 91.6 |

## E The applicability of SCAPE on more scenarios

We tested the applicability of our method in cross-category multiple-instance scenarios as well as cross-style scenarios. Indeed, in an effort to seek the most similar keypoint of support keypoint, the **current CAPE** only finds one-to-one correspondence and cannot establish one-to-many correspondence. As shown below, we have tested the applicability of our approach on multiple-instance scenarios across categories (Row 1-3) and even across domains (Row 4). We think the **future CAPE** could certainly benefit from datasets and frameworks of multi-instance scenarios (including one-to-many correspondence of the same categories).

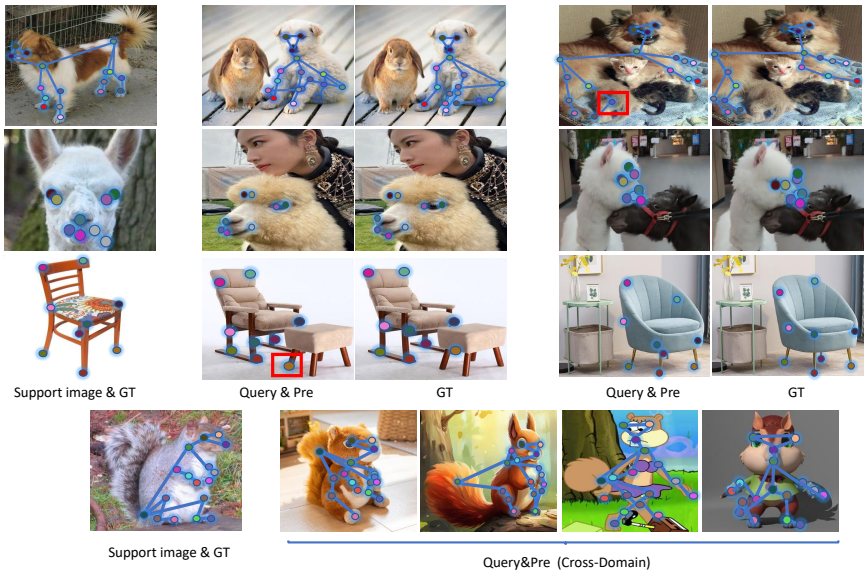


**Fig. 6: The applicability of SCAPE on multiple-instance scenarios across categories**

### F The code for SCAPE and testing for Painter

The code of SCAPE and testing for Painter can be found in the attachment.

### G More Qualitative Results

More qualitative results for visual comparison between the previous best method CapeFormer and ours are shown in Fig. 7 and Fig. 8.

### H Visualization of Attention Maps

Fig. 9, Fig. 10 depict the attention maps between support keypoints and the query image of the last three layers. The attention maps of both CapeFormer and ours are shown here for better understanding.

## References

1. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting **35**, 25005–25017 (2022)
2. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: CVPR. pp. 360–368 (2018)
3. Khan, M.H., McDonagh, J., Tzimiropoulos, G.: Synergy between face alignment and tracking via discriminative global consensus optimization. In: ICCV. pp. 3811–3819. IEEE (2017)
4. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
5. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervised learning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging **41**(7), 1837–1848 (2022)
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
7. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models **36** (2024)
8. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion **36**, 1363–1389 (2023)
9. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: CVPR. pp. 6830–6839 (2023)

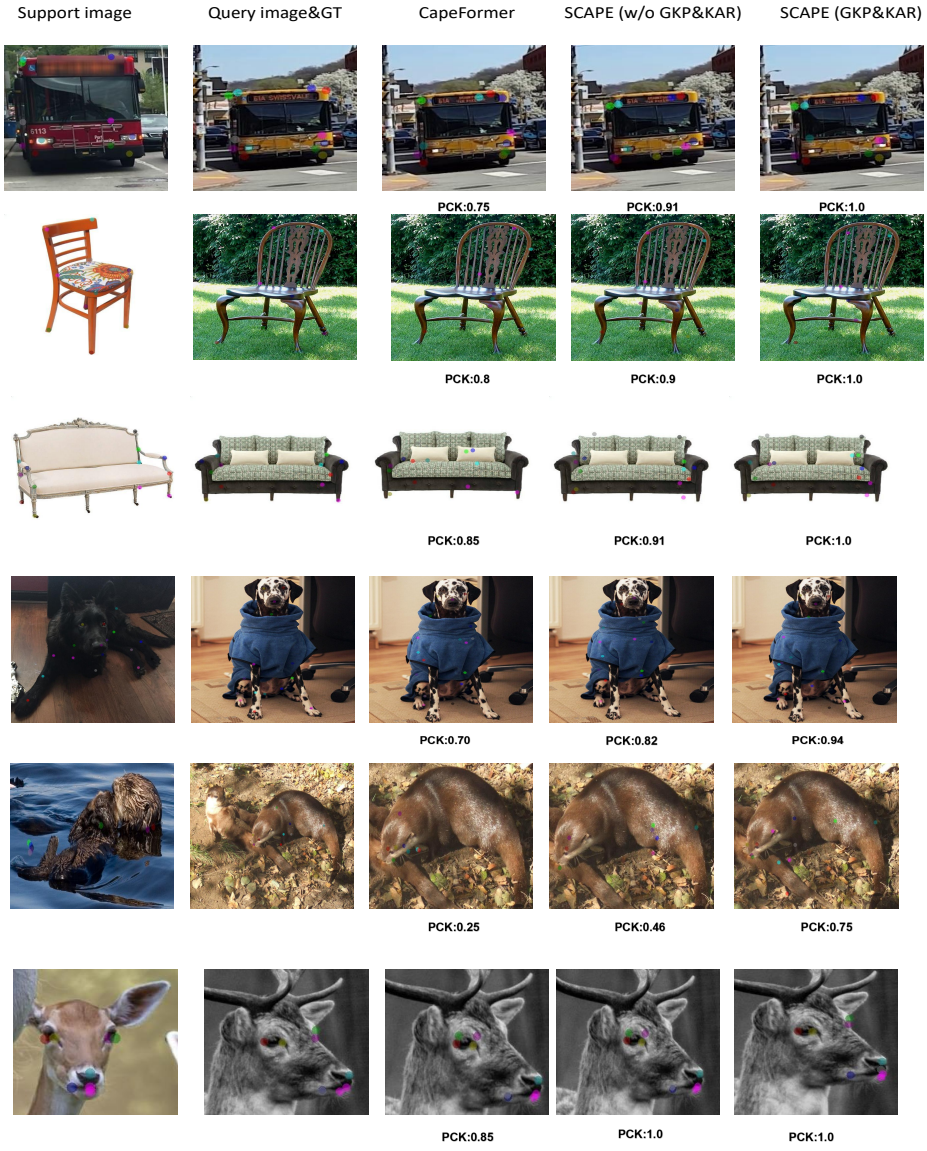


**Fig. 7: Qualitative results. #1**

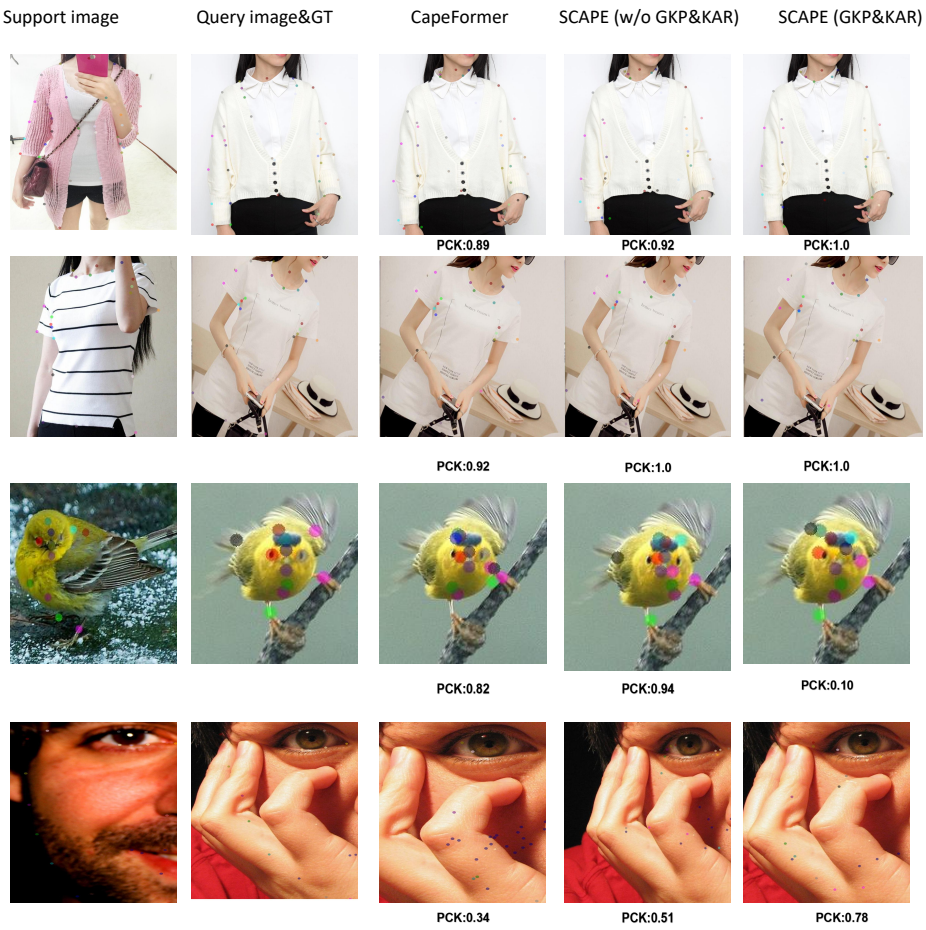


**Fig. 8: Qualitative results. #2**

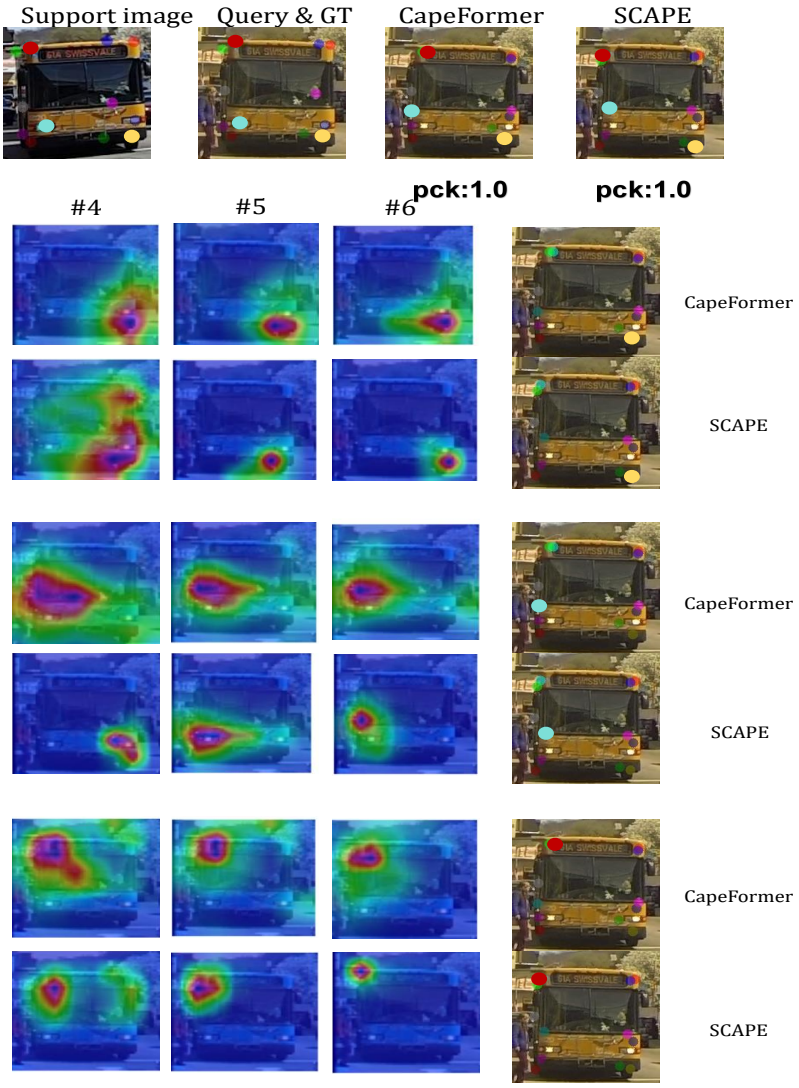


**Fig. 9: Attention map for the last three layers of support keypoints and query image #1.** In cases where predictions are generally accurate, our method exhibits more convergent attention

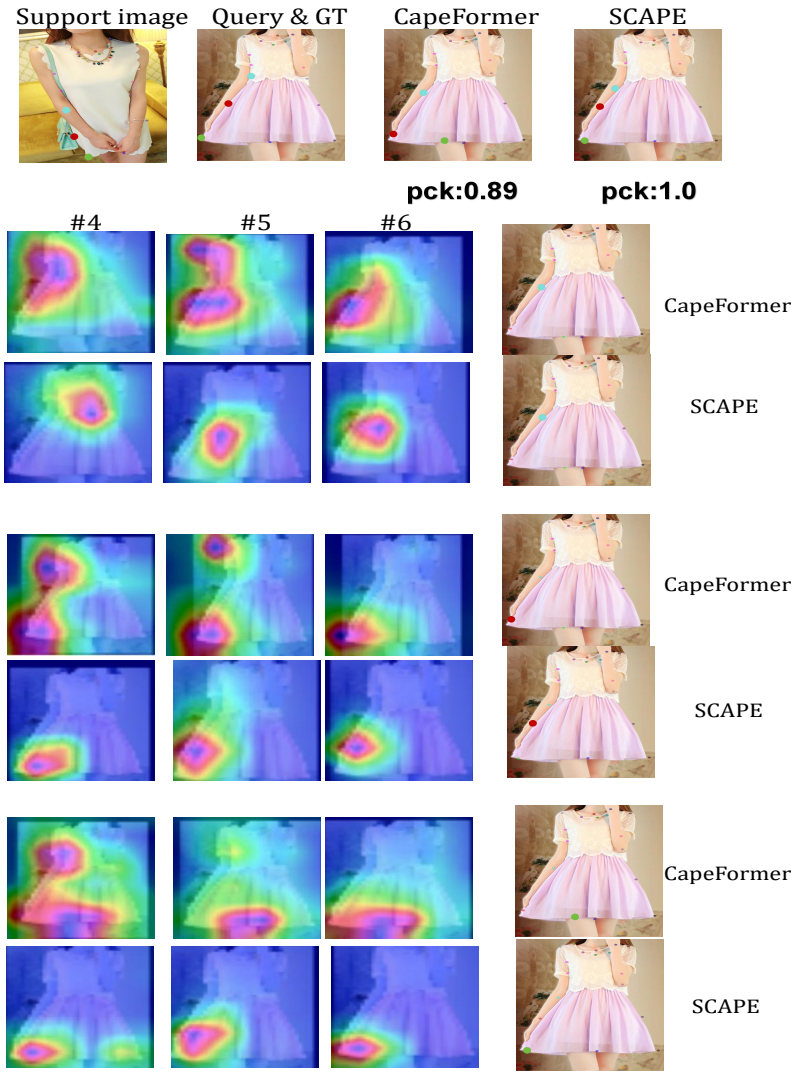


**Fig. 10: Attention map for the last three layers of support keypoints and query image #2**. SCAPE+ can provide more accurate predictions.