CapeNext: Rethinking and Refining Dynamic Support Information for Category-Agnostic Pose Estimation

Yu Zhu, Dan Zeng*, Shuiwang Li, Qijun Zhao, Qiaomu Shen, Bo Tang

Codes and supplementary materials are at https://github.com/yzrs/CapeNext.

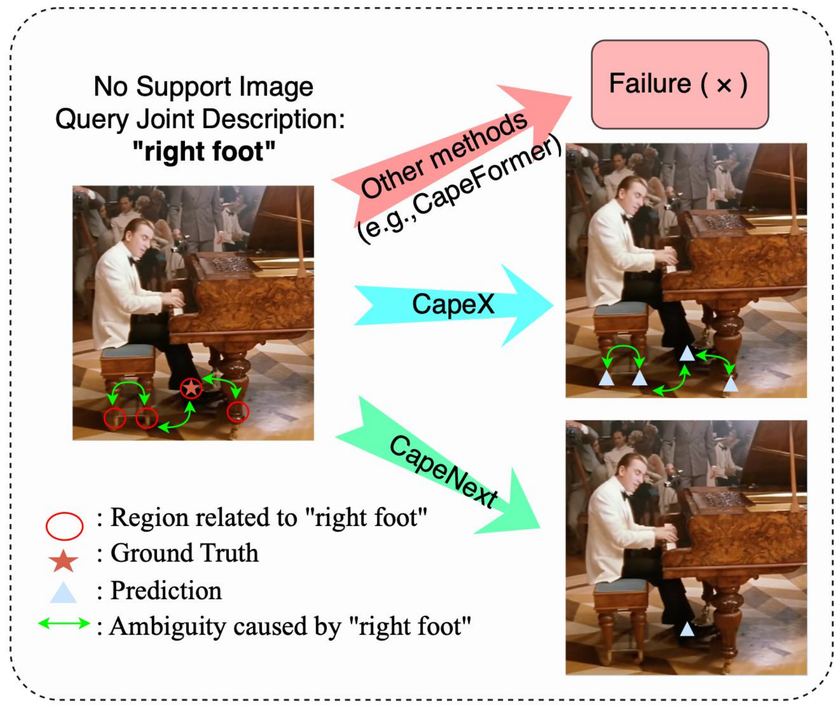
Introduction

CAPE enables keypoint prediction for arbitrary categories, transforming pose estimation into a keypoint matching problem between keypoint feature from support samples and image feature from query image.

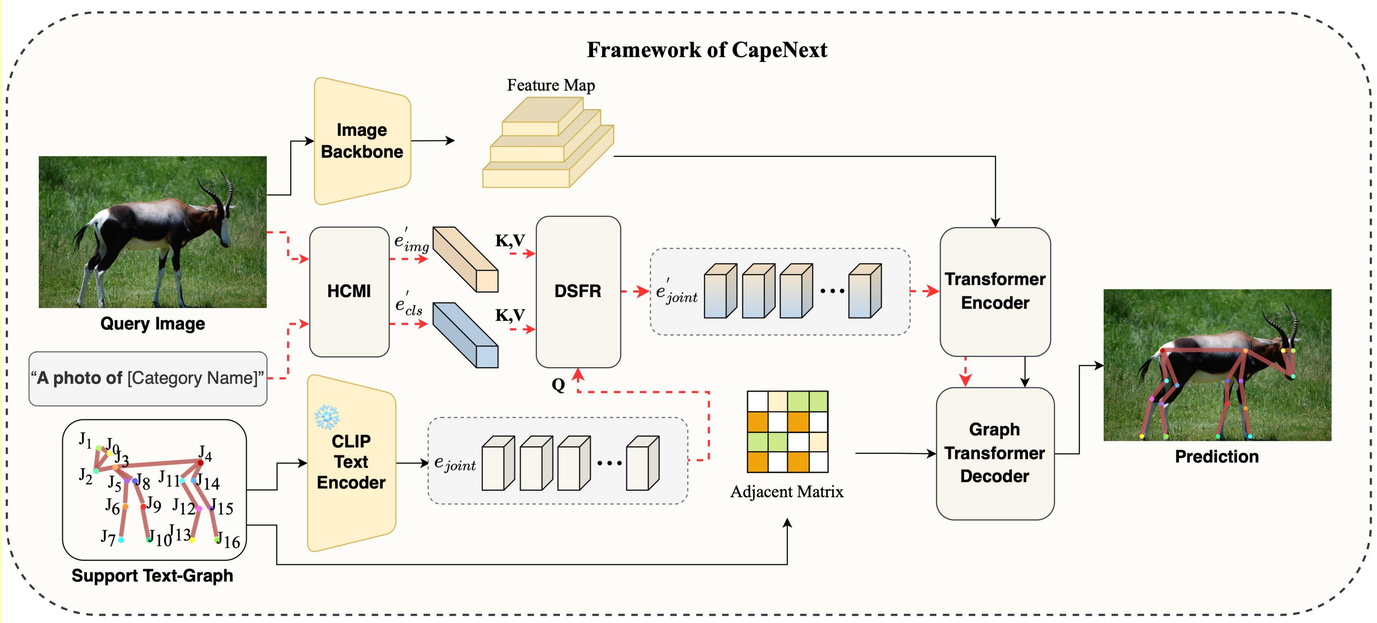
1. **Traditional CAPE methods** extract keypoint feature from annotated support images.
- **Rely heavily on support image quality**: If the instance in the support image is occluded, resulting in invisible keypoints, these methods struggle to extract valid support keypoint features and fail to predict corresponding keypoints in the query image.

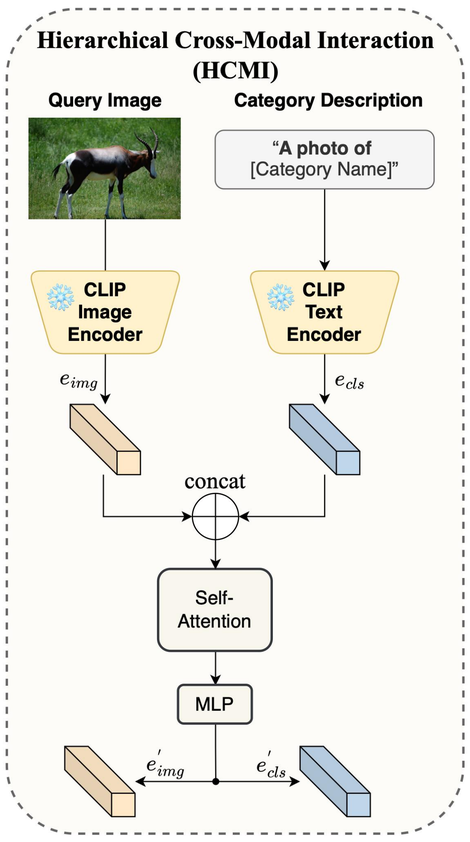
2. **Text-based CAPE methods** extract keypoint feature from static textual keypoint descriptions.
- **Cross-category ambiguity**: The same text description may refer to entirely different keypoints for different categories.
- **Insufficient intra-category discriminability**: Even within the same category, variations in appearance and pose across instances render static text ineffective.

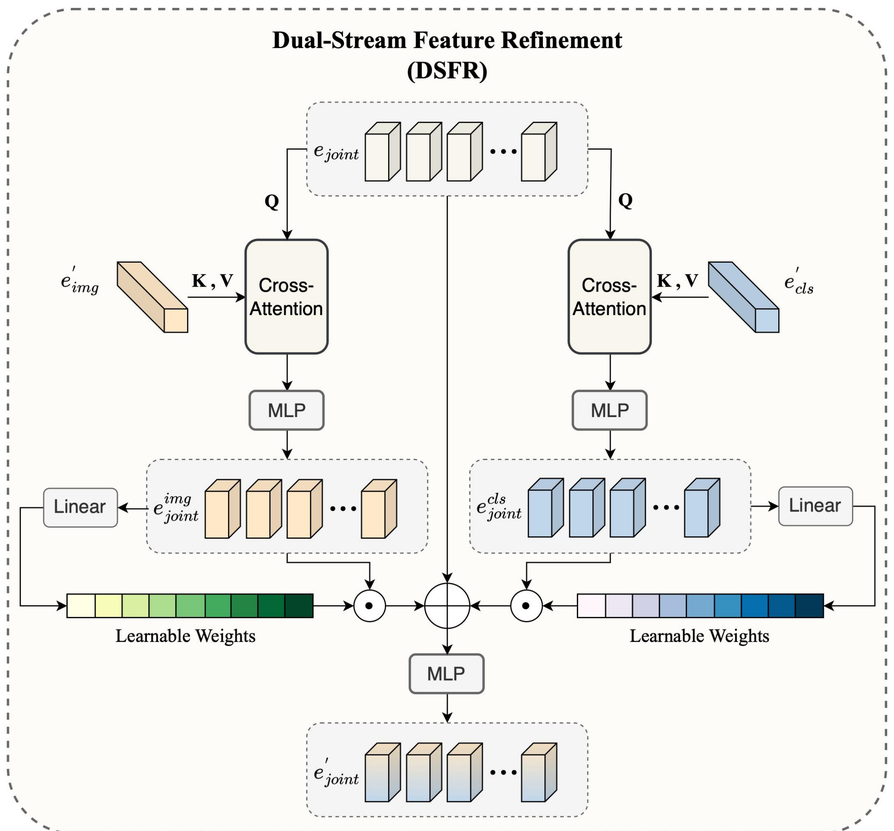
CapeNext aims to solve the issues above by multimodal features.



Method







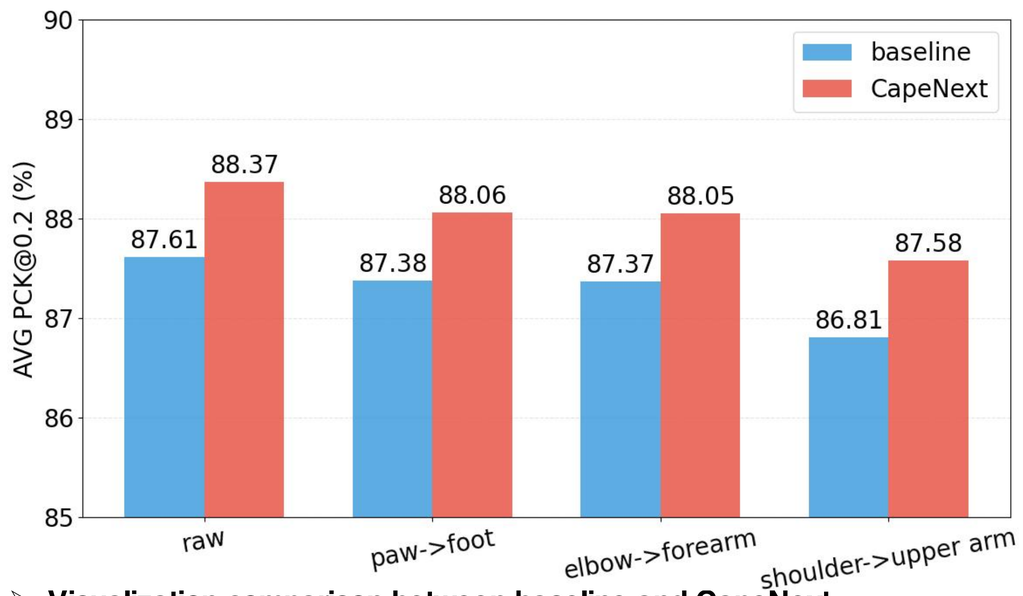
- **HCMI** bridges the hierarchy and modality gap between these visual and semantic embeddings.
- **DSFR** dynamically adapts the CLIP-generated joint embedding to enhance cross-modal alignment between the query image's visual feature, the category feature and the support joint feature.

Results

➢ **PCK@0.2 performance comparison on MP-100 dataset**

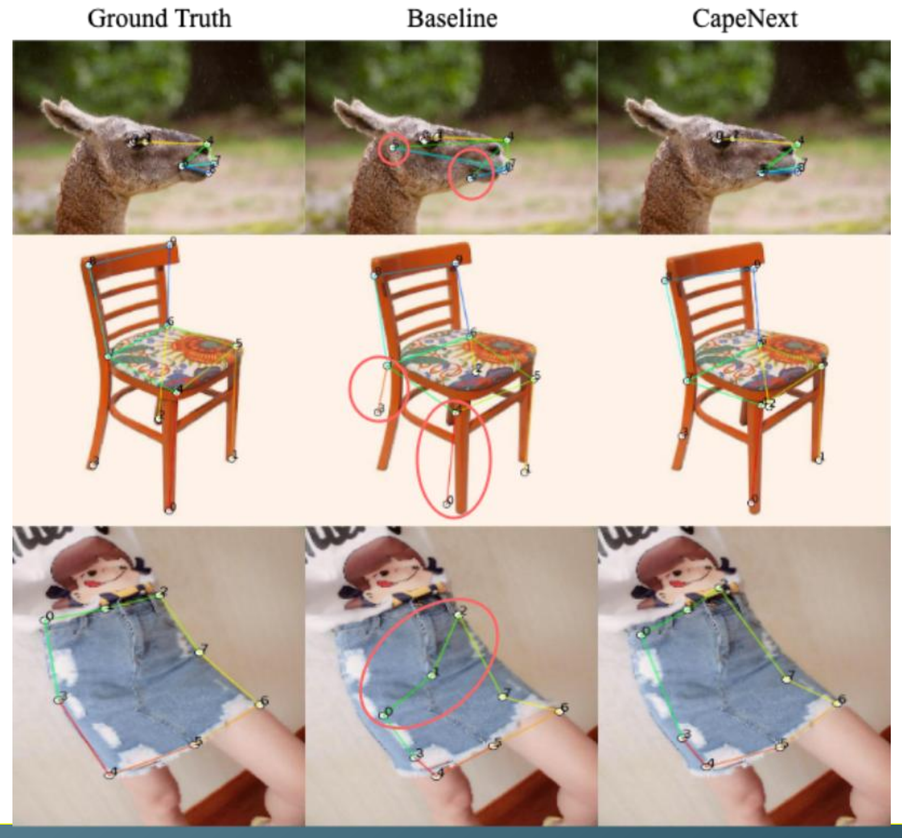
Methods	Img Backbone	Split1	Split2	Split3	Split4	Split5	Avg
POMNet	ResNet-50	84.23	78.25	78.17	78.68	79.17	79.70
CapeFormer	ResNet-50	89.45	84.88	83.59	83.53	85.09	85.31
ESCAPE	ResNet-50	86.89	82.55	81.25	81.72	81.32	82.74
MetaPoint+	ResNet-50	90.43	85.59	84.52	84.34	85.96	86.17
X-Pose	ResNet-50	89.07	85.05	85.26	85.52	85.79	86.14
SDPNet	HRNet-32	91.54	86.72	85.49	85.77	87.26	87.36
GraphCape	Swinv2-T	91.19	**87.81**	**85.68**	85.87	85.61	87.23
CapeX	Swinv2-T	91.9	86.97	84.41	86.13	88.64	87.61
CapeNext	Swinv2-T	**92.44**	87.31	85.44	**86.47**	**90.17**	**88.37**

➢ **PCK@0.2 performance comparison with noisy prompts**



➢ **Visualization comparison between baseline and CapeNext**

CapeNext performs better in fine-grained prediction, which is consistent with the result that its improvement in PCK@0.05 is greater than that of the baseline.



➢ **Ablation study for the multimodal inputs and corresponding modules**

Settings	Split1	Split2	Split3	Split4	Split5	Avg
Baseline	91.9	86.97	84.41	86.13	88.64	87.61
+img emb	**92.77**	**87.36**	84.55	86.31	89.76	88.15
CapeNext(img & cls emb)	92.44	87.31	**85.44**	**86.47**	**90.17**	**88.37**

Settings	Split1	Split2	Split3	Split4	Split5	Avg
w/o HCMI	91.43	83.4	84.97	87.27	86.98	86.81
w/o DSFR	89.04	82.72	82.94	82.97	83.58	84.25
w/o LW	92.59	86.8	85.03	86.52	90.5	88.28

- Left table: CapeNext utilizes both visual features from query image and texual feature from class description.
- Right table: To verify the effectiveness of each designed module, we remove them based on CapeNext.

Conclusion

- CapeNext's results highlight that the rational utilization of query image feature and class information is conducive to CAPE task.
- CapeNext fully leverages multi-modal advantages, achieving significant performance gains and remarkable robustness under noisy inputs.

For More

paper

code

WeChat