Category-Agnostic Pose Estimation (CAPE)

Keywords: Category-Agnostic Pose Estimation (CAPE), 2D Skeleton/Keypoints

Previous Work: Category-Agnostic Pose Estimation
Conventional 2D pose estimation [1, 4] is typically restricted to specific object categories on which the model is trained, such as humans [3], animals [8], and vehicles [6]. These methods cannot generalize to novel object types without retraining. To overcome this limitation, category-agnostic pose estimation (CAPE) [7] has emerged as a paradigm that can estimate poses across diverse object categories using only a few annotated support images, enabling better generalization to arbitrary and previously unseen object categories (see Figure 1). Unlike existing CAPE methods that depends on annotated support images, CapeX [5] only requires a pose-graph abstraction presented as text description to perform pose estimation on input query image.

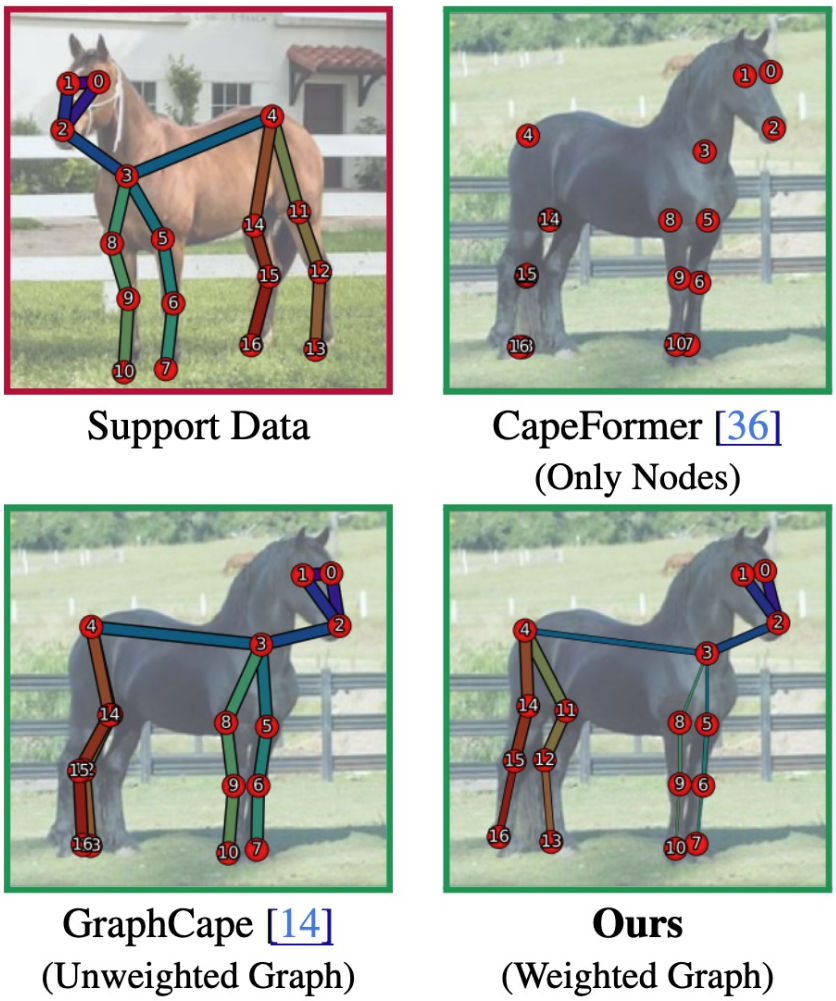

Figure 1: Illustration of Category-Agnostic Pose Estimation. Given support data on the left (i.e. support image, keypoint annotation), the model takes it as an input to estimate keypoints on query image (black horse). Photo is taken from EdgeCape [2].

This Project: Sequential Keypoint Generation for Category-Agnostic Pose Estimation

Goal. Building upon our autoregressive framework for 2D floorplan reconstruction (see fig. 2), the objective is to extend this approach to address the CAPE problem on the MP-100 dataset [7]. Inspired by CapeX [5] (see fig. 3), students need to adapt the Raster2Seq framework (codebase access will be provided) to perform pose estimation using only a query image and a pose graph represented as a keypoint sequence, eliminating the need for annotated support images. The key distinction from CapeX lies in our pose-graph representation: while CapeX employs textual descriptions of keypoints, our method directly utilizes 2D coordinate sequences as the support data for pose estimation.

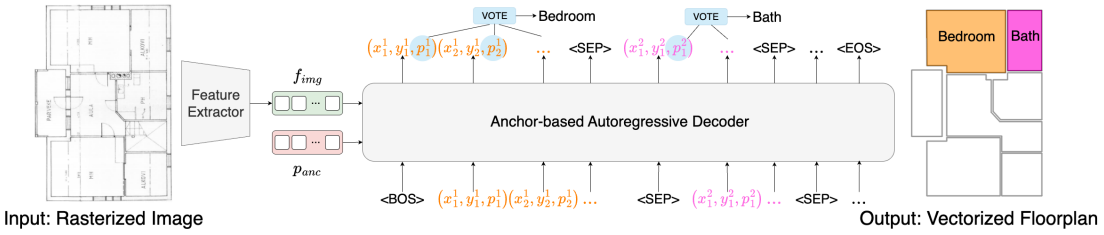

Figure 2: Raster2Seq framework. Given a rasterized floorplan image (left), Raster2Seq converts it into vectorized format, represented as a labeled polygon sequence, separated using special `<SEP>` tokens. The main architectural component of the framework is an anchor-based autoregressive decoder, which predicts the next token given image features (f_{img}), learnable anchors (v_{anc}) and the previously generated tokens. To highlight the order of prediction, the first two predicted polygons are colored in orange and pink, respectively.

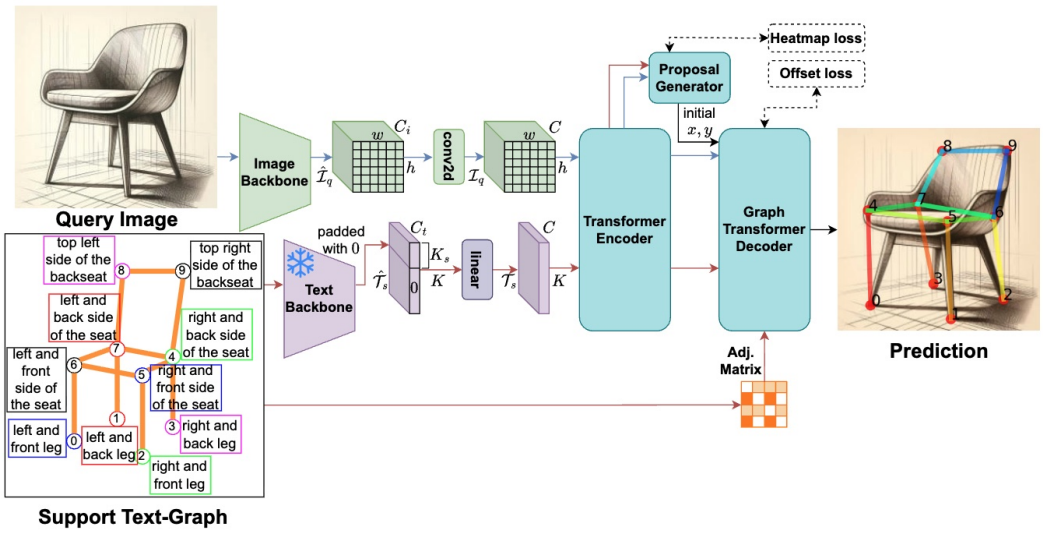

Figure 3: CapeX framework.

Expectation. Comprehensive quantitative and qualitative comparisons against at least three state-of-the-art CAPE baselines is needed to demonstrate its effectiveness.

Related Works

- Raster2Seq framework (please contact Hao for the details and the codebase)
- CapeX [5]

If you are interested, contact Hao Phung (htp26@cornell.edu)

References

[1] M. Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[2] O. Hirschorn and S. Avidan. Edge weight prediction for category-agnostic pose estimation. *arXiv preprint arXiv:2411.16665*, 2024.

[3] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

[4] D. Maji, S. Nagori, M. Mathew, and D. Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2637–2646, 2022.

[5] M. Rusanovsky, O. Hirschorn, and S. Avidan. Capex: Category-agnostic pose estimation from textual point explanation. *arXiv preprint arXiv:2406.00384*, 2024.

[6] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, and R. Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5452–5462, 2019.

[7] L. Xu, S. Jin, W. Zeng, W. Liu, C. Qian, W. Ouyang, P. Luo, and X. Wang. Pose for everything: Towards category-agnostic pose estimation. In *European conference on computer vision*, pages 398–416. Springer, 2022.

[8] H. Yu, Y. Xu, J. Zhang, W. Zhao, Z. Guan, and D. Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021.