Accidental Turntables: Learning 3D Pose by Watching Objects Turn — Supplementary Material

Zezhou Cheng[1] Matheus Gadelha[2] Subhransu Maji[1]
[1]UMass Amherst [2]Adobe Research
[1]{zezhoucheng, smaji}@cs.umass.edu, [2]gadelha@adobe.com

1. Accidental Turntables Dataset

Data source. We use 6 Youtube videos as the source of our Accidental Turntables dataset including video1, video2, video3, video4, video5, video6.

More examples. Fig. 1 provides more examples from our Accidental Turntables dataset.

Figure 1. **Samples from the Accidental Turntables dataset.** SfM provides accurate 3D reconstructions (**middle**) and pose estimations (**right**) on either texture-rich (**1st row**) or texture-free (**2nd row**) objects, as well as objects moving along a straight line without any turns (**3rd row**).

2. More Analysis

The effect of annotation noise level on pose estimation In the main text, we use ImageNet-pretrained ResNet50 to initialize our model and analyze the effect of annotation noise level on the performance of pose estimation (Fig. 6 in the main paper). Here we provide additional experimental results under different network initialization including contrastively pretraining and random initialization. Fig. 2 demonstrates that the effect of annotation noise level on the pose estimation performance is consistent across different network initialization, *i.e.*, neither clean-yet-small data nor large-yet-noisy data lead to higher performance than mid-size data with mid-level noise.

3. Implementation

We use the Structure-from-Motion (SfM) and Multiview Stereo (MVS) pipelines implemented in COLMAP [4, 5] [1] and HLOC library [3] [2]. We use the MaskRCNN [2] implemented in Detectron2 [6] to get the object masks. We implement our pose estimation models based on PoseContrast [7] [3].

References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

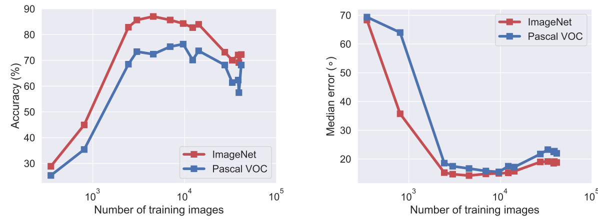
[3] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 1

[4] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

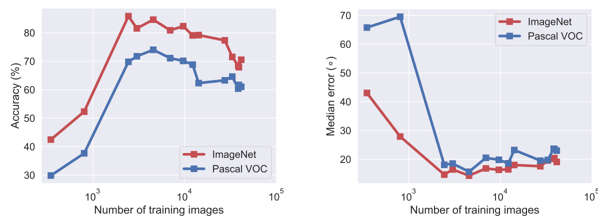
[5] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[6] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

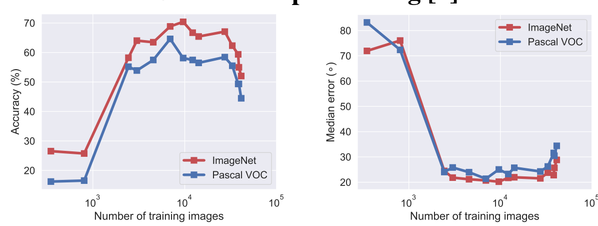
[1] https://colmap.github.io/
[2] https://github.com/cvg/Hierarchical-Localization
[3] https://github.com/YoungXIAO13/PoseContrast



ImageNet pretraining



Contrastive pretraining [1]



Random initialization

Figure 2. **The effect of annotation noise level on 3D pose prediction is consistent across different network initialization.** For each initialization method, we report the performance of the pose predictor under different noise levels of pose annotations. A higher level of annotation noise corresponds to a larger number of training images. We report both prediction accuracy (top row) and median error (bottom row) on two test splits included in PASCAL3D+ (*i.e.*, PASCAL VOC and ImageNet validation set.).

[7] Yang Xiao, Yuming Du, and Renaud Marlet. Posecontrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning. In *2021 International Conference on 3D Vision (3DV)*, pages 74–84. IEEE, 2021. 1