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.

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 pretrained 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] and HLOC library [3]. We use the MaskRCNN [2] implemented in Detectron2 [6] to get the object masks. We implement our pose estimation models based on PoseContrast [7].

References

[9] https://github.com/YoungXIAO13/PoseContrast
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).