Accidental Turntables: Learning 3D Pose by Watching Objects Turn

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

Abstract

*We propose a technique for learning single-view 3D object pose estimation models by utilizing a new source of data — in-the-wild videos where objects turn. Such videos are prevalent in practice (*e.g. *cars in roundabouts, airplanes near runways) and easy to collect. We show that classical structure-from-motion algorithms, coupled with the recent advances in instance detection and feature matching, provide surprisingly accurate relative 3D pose estimation on such videos. We propose a multi-stage training scheme that first learns a canonical pose across a collection of videos and then supervises a model for single-view pose estimation. The proposed technique achieves competitive performance with respect to the existing state-of-the-art on standard benchmarks for 3D pose estimation without requiring any pose labels during training. We also contribute an Accidental Turntables Dataset, containing a challenging set of 41,212 images of cars in cluttered backgrounds, motion blur, and illumination changes that serve as a benchmark for 3D pose estimation.*

1. Introduction

Understanding object pose and its structure is a central computer vision problem. Many images have been manually annotated with pose information in multiple datasets containing various types of objects. Still, this manual annotation process is labor-intensive and prone to unavoidable human annotation errors. On the other hand, mechanical devices that precisely change an object's pose are widely utilized when performing high-precision 3D scanning. They allow a particular object to have its pose modified in a controlled manner while capturing its appearance through a variety of image sensors. One of the simplest devices of this kind is a *turntable* – a rotating platform that slowly changes the pose of an object through an electric motor (Fig. 1a). Unfortunately, despite its simplicity, *turntables* are not very practical. They need to be as large as the object at hand, *e.g.*, setting up turntables for cars or airplanes would require a lot of work.

(a) Classic turntable (b) Accidental turntable

Figure 1. **Classic turntable vs. accidental turntable. (a)** Classic turntables rotate and scan objects in a controlled environment to estimate their 3D pose and shape. **(b)** A turning object in a video leads to an accidental turntable. Structure-from-motion, coupled with object detection [15] and feature matching [28], provides surprisingly accurate relative 3D pose estimation (top) and 3D reconstruction (bottom) — the red pyramids indicate the estimated relative poses of video frames. We utilize a collection of such videos to train and evaluate models for single-frame 3D pose estimation in realistic settings. See more accidental turntables here: https://youtu.be/8rFNRri8-TI

Fortunately, we don't need to place those objects in actual turntables. Many are already performing similar motions on their own (Fig. 1b) — cars moving along roundabouts, airplanes landing and parking, ships maneuvering across canals, and so on. In the real world, video recordings of objects performing these types of motions depict them in uncontrolled environments; *i.e.*, cluttered background, occluders, changes in illuminations, motion blur, unpredictable pose changes, and many other nuisance factors. Thanks to many recent advances in computer vision, we show that we are able to bypass many of those nuisance factors and apply Structure from Motion (SfM) to reliably and precisely recover relative pose estimation from videos of real objects (Fig. 1b). We call these types of videos **Accidental Turntables** – objects presenting motion patterns that allow us to observe them from (almost) all possible angles. We demonstrate that these videos, after suitable automatic pre-processing (Sec. 3), are an excellent source of supervision for pose estimation models and, perhaps more importantly, can be mined from the internet, enabling the creation

Website: https://zezhoucheng.github.io/acci-turn

of bigger and more diverse datasets.

However, using the supervision from SfM does not allow us to directly perform pose estimation with respect to a canonical object frame. To this end, we propose to learn a *relative* pose estimation model and show that its training leads to the emergence of a canonical object pose (§ 4.1). In the second stage, we propose a calibration and training procedure (§ 4.2) that allows pose estimation in a canonical frame (§ 4.3). We show that models trained in this fashion *only* using our newly collected dataset from *real videos* significantly outperform other models trained on SfM and perform on par with existing unsupervised approaches on standard benchmarks, *e.g.*, the Freiburg and ImageNet cars datasets.

We summarize our contributions as follows. 1) a procedure for automatically processing accidental turntable videos and annotating its frames with relative pose transformations; 2) a multi-stage training scheme that allows training accurate pose estimation models with respect to arbitrary canonical frames; and 3) a new dataset with 41,212 real images of cars from turntable videos with their corresponding pose annotation.

2. Related Work

Datasets for pose estimation. A number of datasets provide 3D pose annotations for objects in the wild [1, 32, 34, 40, 39, 11] or in controlled environments [10, 37, 38, 41, 16]. These datasets have been widely used for training supervised pose estimation models [36, 21, 18, 13]. However, manually annotating 3D poses is very tedious and thus not scalable. Unsupervised pose estimation models [31, 25, 24, 22] learn to predict 3D pose without any human annotations. Videos [31, 26] that capture multiple views of objects have been the main source of training data in prior works [31, 25, 22]. However, to acquire such videos, a person needs to hold a camera and slowly move around a *static* object. This is a time-consuming procedure, especially for large-size objects (*e.g.* cars, and airplanes), and has limited the size of existing video datasets. For example, the Freiburg Cars dataset [31] consists of 52 car videos, and EPFL car dataset [26] only provides 20 cars. Such limited data may further constrain the performance of prior methods. In this work, we identify a new source of data for unsupervised pose estimation – videos where objects turn. The turning of objects (*e.g.*, vehicles) is such a natural phenomenon in daily life that these videos are quite easy to collect. We build a new dataset consisting of 313 car videos with a total of 141,784 frames. Our 3D pose annotations are generated by SfM [29, 30] enhanced by recent progress in object detection [15] and feature matching [28].

Supervised pose estimation. With groundtruth 3D pose annotations, supervised pose estimation works have been focusing on developing novel representations of 3D pose [45, 18, 23], learning objectives [36, 18, 43, 42], or network architectures [7, 8]. The difficulty in annotating 3D poses results in the scarcity of pose annotations. This issue is partially relieved by augmenting the existing datasets with synthetic data [33]. The integration of pose estimation and object detection has been explored in the task of 3D object detection [11, 6]. Our models are built upon the prior supervised learning methods [45, 43, 42], but we use pose annotations automatically generated by SfM, instead of human annotations.

Unsupervised pose estimation. Unsupervised pose estimation models learn 3D object pose without any human annotations. Prior works are either based on analysis-by-synthesis [24, 22] or SfM [31, 25]. The analysis-by-synthesis frameworks train a pose estimation model by reconstructing the input images in a pose-aware manner. The SfM-based methods start by estimating the pose labels with SfM on videos that capture 360° views of static objects. However, SfM only provides relative pose estimations among video frames. The absolute pose estimations from SfM are not consistent across videos (*i.e.*, objects in the same pose from two videos may have quite different absolute pose representations). To tackle this issue, Sedaghat *et al*. [31] calibrate the SfM pose estimations via aligning 3D reconstructions of objects; Novotny *et al*. [25] train a model to estimate the relative pose and observe that canonical poses emerge in the models trained in this manner. Similar to Novotny *et al*. [25], we train a model to estimate the relative pose from SfM (Sec. 4.1). Differently, we find that such a training strategy is not sufficient to learn a high-quality pose predictor. We instead use the model trained in this way as a tool to calibrate the SfM estimations across videos (Sec. 4.2), followed by training a novel pose estimator on the calibrated pose annotations (Sec. 4.3).

Accidental data in computer vision. Researchers have discovered interesting phenomena that occur accidentally but turn out to be useful for computer vision tasks in the literature. Torralba *et al*. [35] demonstrate that outdoor scenes can be recovered from accidental pinhole images. Li *et al*. [17] train a depth estimator on a collection of Internet videos of people imitating mannequins, *i.e.*, freezing in diverse poses. The depth information is obtained from SfM and multi-view stereo algorithms. Similar to Li *et al*. [17], we train a pose estimation model on a collection of Internet videos and the 3D pose annotations are automatically generated from SfM. Unlike people imitating mannequins, we only require the object turns in the video, which is a quite natural behavior in practice (*e.g.*, a car moving along roundabouts).

Figure 2. **Samples from accidental turntable dataset.** Accident turntables are prevalent in practice. For instance, a car donuts (1st row), a car moves along a roundabout (2nd and 3rd row), or a car does not turn but passes by a camera (4th row). All car instances exhibit at least 180° azimuth changes relative to the camera.

3. Accidental Turntable Dataset

In this section, we provide the details of our data collection and the generation of 3D pose annotations with SfM algorithms on our dataset. We name the collected video dataset as **Accidental Turntable Dataset**, highlighting its connections to classic turntables (Fig. 1).

Data source. The main criterion of our data collection is that the object turns in the video. Such videos are abundant on the Internet and quite easy to acquire. In this work, we focus on the car category which is one of the most common moving objects in the wild (at least in America). We leave the extension to other categories (*e.g.*, airplanes and boats) in our future work, but include some examples of the reconstructions at the end of the paper. We collect 313 car video clips from YouTube containing a total of 141,784 frames. Each video consists of a single moving car instance that exhibits multiple views in motion. Fig. 2 provides video samples from our dataset. We provide a full list of YouTube links to the collected videos in the Supplementary Material.

Challenges. Even though our dataset consists of a large number of car videos serving as a new source of training data for machine learning models, in-the-wild videos pose technical challenges for automatic extraction of 3D pose using SfM. For example, to exploit the classical SfM algorithms to estimate the object pose, object segmentation is required to remove the background; Motion blur and texture-free object surfaces necessitate robust interest points detection; Discriminative feature description and robust feature matching are needed to avoid the ambiguity of pose estimation on symmetric objects (*e.g.*, cars).

Pose estimation with SfM. To tackle the above-mentioned challenges, we use the MaskRCNN [15] pre-trained on MS-COCO dataset [19] to remove the background clutter. We find that the MaskRCNN provides highly accurate object detection and segmentation on in-the-wild car videos. We use SfM algorithms implemented by COLMAP [30, 29] with SuperPoint [5] as the feature extractor and SuperGLUE [28] as the feature matcher to estimate the object pose on cropped object images. We sequentially match the next 10 frames per video frame, instead of exhaustively matching every pair of frames in a video. Sequential matching reduces the ambiguity in matching repeated patterns (*e.g.* left and right wheels of a car). SfM, coupled with MaskRCNN, SuperPoint, and SuperGLUE, provides surprisingly accurate pose estimation, in comparison with classical SIFT [20] and nearest neighbor matching. We provide a detailed study on the effect of feature extraction and matching on SfM in Sec. 5.2.

Statistics. Our dataset consists of 313 car videos with 141,784 frames in total. SfM automatically samples frames with sufficient large relative pose change and reliable feature matching. Adjacent frames in a video usually have tiny differences in the pose. Thus, most of the frames are filtered out by SfM. We end up collecting 41,212 frames with SfM pose estimations. Our dataset covers cars with diverse shapes, colors, textures, and poses (see examples in Fig. 2).

4. Methodology

This section introduces our framework for learning 3D object pose from the proposed accidental turntable dataset. Fig. 3 illustrates an overview of the proposed framework. SfM estimates the relative pose of objects with respect to the object in the first frame per video, followed by optimizing the pose parameters with the bundle adjustment. However, the object pose in the first frame may vary dramatically across videos. It is thus meaningless to train a model directly on the absolute pose labels from SfM. Instead, we start by training a model to estimate the *relative* pose of frame pairs (Fig. 3 left). We observe that a canonical pose emerges in our pose estimation model train in this way (see Sec. 5.2). This provides us a tool to calibrate the pose estimation from SfM to a canonical frame (Fig. 3 middle). In the second stage, we train a pose estimation model directly on the calibrated *absolute* pose annotations similar to standard supervised learning methods [36, 33, 42] (Fig. 3 right). We denote our model trained in the first stage as $f(x)$ and the model in the second stage as $g(x)$, where x is the input image. Our accidental turntable dataset is denoted by $\{(x_i, R_i)\}$, where $R \in \mathrm{SO}(3)$ is the SfM pose estimation.

4.1. Relative pose estimation

In this stage, we train a single-view pose estimation network $f(x)$ to predict the relative pose between pairs of

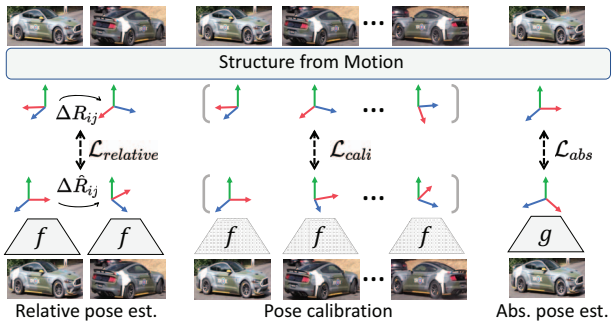

Figure 3. **Approach overview. Left:** a pose estimation model $f(x)$ is trained to predict the *relative* pose of image pairs (denoted by ΔR_{ij}). **Middle:** the emergence of the canonical pose in $f(x)$ enables us to calibrate the pose estimations from SfM to a uniform frame. The model $f(x)$ is frozen in the pose calibration step. **Right:** after the pose calibration, a pose estimation model $g(x)$ is trained on the *absolute* pose annotations.

video frames. The loss function is defined as

$$\mathcal{L}_{\text{relative}} = \sum_{(i,j)}^{N} \text{dist}(R_i R_j^T, \hat{R}_i \hat{R}_j^T) \quad \text{with} \quad \hat{R}_i = f(x_i) \tag{1}$$

where $\text{dist}(\cdot,\cdot)$ is a distance function between two rotation matrices (*e.g.* L_2 or geodesic distance). $\hat{R}_i$ is a 3×3 rotation matrix predicted from the model $f(x_i)$ on the input x_i. The frame pair x_i and x_j are sampled from the same video. N is the total number of frame pairs sampled from our video dataset. $\Delta R_{ij} = R_i R_j^T$ is the relative rotation matrix that transforms the pose of the frame x_j to x_i. We use the 6D continuous rotation representation [45] as the intermediate output of our model $f(x)$, from which the 3×3 rotation matrices $\hat{R}$ are recovered by the Gram-Schmidt orthogonalization [45]. Our first training stage is similar to the learning strategy proposed by Novotny *et al*. [25]. Differently, we only use the model $f(x)$ trained in this stage as a tool to calibrate the SfM pose annotations (Sec. 4.2). Moreover, we demonstrate that the model $g(x)$ trained in our second stage significantly outperforms the stage-one model $f(x)$ as well as Novotny *et al*. [25]. We provide detailed comparisons between $f(x)$ and $g(x)$ in Sec. 5.2.

4.2. Pose calibration

The pose predictor $f(x)$ trained in the first stage provides us a tool to calibrate the pose annotations from SfM into a uniform pose frame, thanks to the emergence of canonical pose (see Sec. 5.2 for more details). If the pretrained $f(x)$ provides perfectly accurate pose estimation per input x, there exists a global rotation ΔR for each video that aligns our pose annotations $\{R_i\}$ to the pose predictions $\{\hat{R}_i\}$:

$$\hat{R}_i = \Delta R R_i \quad \forall i \in 1, \ldots, K \tag{2}$$

Where K is the number of frames in the target video, however, the pose predictions $\{\hat{R}_i\}$ are inaccurate in practice due to the limited performance of the pretrained pose predictor $f(x)$. We thus target at a rotation matrix ΔR^* that aligns $\{R_i\}$ and $\{\hat{R}_i\}$ with minimal calibration error. We define the calibration error as,

$$\mathcal{L}^*_{cali} = \frac{1}{K} \sum_{i}^{K} \text{dist}(\hat{R}_i, \Delta R^* R_i) \tag{3}$$

where $\text{dist}(\cdot,\cdot)$ is a distance function between two rotation matrices. We adopt the geodesic distance $\| \log R^T \hat{R} \|_{\mathcal{F}} / \sqrt{2}$ in our implementation. The pose calibration is then formulated as an optimization problem:

$$\min_{\Delta R} \quad \mathcal{L}_{\text{cali}}(\hat{R}, \Delta R R) \tag{4}$$

$$\text{s.t.} \quad \Delta R \in SO(3) \tag{5}$$

This problem can be solved by the classical Procrustes analysis [12]. In practice, we find that a simple search-based optimization method works reliably. Concretely, the optimal global rotation ΔR^* is searched from the set $\{\Delta R_j : \Delta R_j = \hat{R}_j R_j^T\}$. Moreover, the calibration error L^*_{cali} is closely related to the noise level of the calibrated pose annotations. Large calibration error typically means the failure of calibration and a higher level of noise in the calibrated pose annotations (see Sec. 5.2 for our empirical studies). Therefore, the calibration error L^*_{cali} may serve as a heuristic to filter out noisy pose labels.

4.3. Absolute pose estimation

We now could apply any supervised learning methods for pose estimation on our calibrated dataset $\{(x_i, R_i^{\text{cali}})\}$. In this work, we adopt the framework proposed by Xiao *et al*. [43, 42] to train our pose estimator. Concretely, we use three Euler angles as our pose representation, including azimuth $\alpha \in [-\pi, \pi]$, elevation $\beta \in [-\pi/2, \pi/2]$, and roll $\gamma \in [-\pi, \pi]$. The Euler angles are decomposed from the rotation matrices R^{cali} and divided into Z_θ disjoint angular bins with bin size $B_\theta = \pi/12$. The model is trained to predict the bin indices $y_\theta \in \{1, \ldots, Z_\theta\}$ via a classification loss and within-bin offsets δ_θ via a regression loss:

$$\mathcal{L}_{\text{abs}} = \sum_{\theta \in \alpha,\beta,\gamma} \mathcal{L}_{cls}(y_\theta, p_\theta) + \lambda \mathcal{L}_{\text{reg}}(\delta_\theta, \hat{\delta}_\theta) \tag{6}$$

where p_θ is the probability of the object pose in the bin y_θ; $\hat{\delta}_\theta \in [0, 1]$ is the predicted offsets within the bin y_θ; $(p_\theta, \hat{\delta}_\theta) = g(x)$ are both outputs of our pose estimation

model $g(x)$. We use the cross-entropy loss as the classification loss $\mathcal{L}_{cls}$ and the smooth-L1 loss as the regression loss $\mathcal{L}_{\text{reg}}$; λ is the weight on the regression loss ($\lambda = 1$ by default).

At the inference time, the pose prediction $\hat{\theta}$ on the input x is obtained by combining the prediction of the bin classifier and the offsets within the predicted angular bin:

$$\hat{\theta} = (j + \hat{\delta}_{\theta,j})B_\theta \quad \text{with} \quad j = \arg\max_i p_{\theta,i} \tag{7}$$

where $p_{\theta,i}$ is the probability of object pose in the i-th bin, and $\hat{\delta}_{\theta,j}$ is the predicted offsets within the i-th bin.

5. Experiments

Implementation details. We use a standard ResNet50 network with three fully-connected layers as our pose estimation model. We initialize our model with ImageNet pretrained weights and fine-tune it during training. In the first training stage, we do not apply any data augmentation. In the second training stage, we use standard data augmentations including in-plane rotation and flipping. We conduct hyperparameter search and checkpoint selection on a validation set separate from our training and test set. The validation set consists of 338 non-truncated and non-occluded car images from PASCAL3D+ [40]. Similar to prior work [24, 22, 43, 42], we use a tightly cropped object image as the input to our pose estimation model. The input image is resized and padded to 224×224. We use the Adam optimizer with a learning rate of 1E-4 and weight decay of 5E-4. In the second training stage, we train our model on videos with a calibration error $\mathcal{L}^*_{\text{cali}}$ (Eqn. 3) lower than 7°.

Benchmarks. We evaluate the performance of our model on the PASCAL3D+ dataset [40] which is a standard benchmark for 3D pose estimation. The test split in the PASCAL3D+ dataset consists of 308 non-occluded and non-truncated car images collected from the PASCAL VOC dataset [9]. More recently, Mariotti *et al*. [22] reports their results on the ImageNet validation set included in PASCAL3D+ which consists of 2712 test images of cars. To make a comparison with Mariotti *et al*., we provide results on both test splits. Following prior works, we measure the prediction error using the standard geodesic distance $\Delta R = \| \log R_{gt}^T R_{pred} \|_{\mathcal{F}} / \sqrt{2}$ between the estimated rotation matrix R_{pred} and the groundtruth R_{gt}. We report the median geodesic error (Med.) and the percentage of predictions with error less than $\pi/6$ (Acc.) relative to the groundtruth.

Pose calibration for evaluation. The pose predictions from our model align with human annotations up to a global rotation, due to the difference between the coordinate frame of our model and that of pose annotation tools adopted by the benchmarks. To evaluate our model on the benchmarks, similar to prior unsupervised learning methods [22, 24], we need to calibrate our pose estimations to the groundtruth annotations. Such pose calibration for evaluation is exactly the same as our pose calibration step described in Sec 4.2. Specifically, we estimate a global calibration matrix ΔR such that $\Delta R R_{pred}$ equals the human annotations R_{gt}. We formulate the pose calibration as an optimization problem and solve it via a simple search-based method (see more details in Sec 4.2). The calibration matrix ΔR is obtained by solving the optimization problem on 100 car images randomly sampled from the training set of PASCAL3D+.

5.1. Pose estimation

Quantitative results. Tab. 1 provides quantitative comparisons with prior unsupervised pose estimation works on PASCAL3D+ test set. Our method significantly outperforms the existing SfM-based methods [31, 25]. Similar to ours, these models are trained on video data with pose annotations from SfM. However, they rely on SfM with SIFT [20] and nearest neighbor (NN) matching, which fails to provide high-quality pose estimations (see more details in Sec 5.2). For this reason, prior SfM-based models collect videos by slowly moving a camera around *static* cars to avoid large motion blur. This tedious procedure limits the size of existing car video datasets. For example, the FreiburgCars dataset [31] consists of 52 car videos; the EPFL car dataset [26] provides only 20 car videos. In comparison, our video dataset (consisting of 313 videos) is easy to collect and prevalent on the Internet. SfM, coupled with the recent progress in object detection [15] and feature matching [28], provides robust and accurate pose estimations on our in-the-wild videos, which is the key to the success of our framework. Our model trained on the accidental turntable dataset achieves higher pose prediction accuracy than when trained on the FreiburgCars dataset.

In comparison with analysis-by-synthesis frameworks [24, 22], our prediction accuracy is significantly higher than that of SSV model [24] which is trained on the CompCars dataset [44] (consisting of 137,000 real car images). ViewNet [22] achieves the highest performance on PASCAL3D+ among existing unsupervised learning methods. However, this method relies on 3D models from ShapeNet [2] to generate a highly curated dataset with controlled variations in viewpoint, translation, lighting, background, etc. In contrast, ViewNet has a harder time learning from real videos (*e.g*. FreiburgCars [31]) where its performance drops remarkably. We're unable to train the ViewNet on our dataset as the source code was not publicly available at the time of this study.

Table 1. **Pose estimation on PASCAL3D+ test sets.** We make comparisons with supervised learning methods trained with human annotations (dubbed Anno.) and unsupervised pose estimation models based on Structure-from-Motion (dubbed SfM) or Analysis-by-Synthesis (dubbed AbS). *ViewNet ignores the in-plane rotation in the evaluation and reports the results on the ImageNet validation set.

	Methods	Supervision	Trainset	Testset	Acc.(%) ↑	Med.(°) ↓
Super.	Tulsiani *et al*. [36]	Anno.	PASCAL3D+	VOC	89	9.1
	Mahendran *et al*. [21]	Anno.	PASCAL3D+	VOC	–	8.1
	Liao *et al*. [18]	Anno.	PASCAL3D+	VOC	93	5.2
	Grabner *et al*. [13]	Anno.	PASCAL3D+	VOC	94	5.1
Unsupervised	VPNet [31]	SfM	FreiburgCars	VOC	–	49.6
	VpDRNet [25]	SfM	FreiburgCars	VOC	–	29.6
	SSV [24]	AbS	CompCars	VOC	67	10.1
	Ours	SfM	FreiburgCars	VOC	72	15.7
	Ours	SfM	Acci.Turn.	VOC	75	15.8
	ViewNet* [22]	AbS	ShapeNet	ImageNet	88	5.6
	ViewNet* [22]	AbS	FreiburgCars	ImageNet	61	16.1
	Ours	SfM	FreiburgCars	ImageNet	84	15.0
	Ours	SfM	Acci.Turn.	ImageNet	86	14.8

Figure 4. **Pose prediction on Pascal3D+ test set. Left**: our model achieves high accuracy of pose estimation on cars in diverse appearances, poses, and shapes. **Right**: the performance drops on large, occluded objects (1st row), low-resolution images (2nd row) or out-of-domain data (last two rows). The solid arrows indicate the pose predictions from our model and the dashed arrows are the groundtruth annotations. The blue arrow directs towards the frontal side of cars and the red points toward the right side. The angular distances between the predictions and the groundtruth are less than 7° for examples on the left while higher than 90° on the failure cases.

Qualitative results. Fig. 4 visualizes our pose predictions on the Pascal3D+ test set. Our model provides accurate pose estimation on diverse cars in terms of appearance, poses, and shapes. The performance of our model drops in several cases: the object is highly occluded; the image is in low resolution; the domain gap between the input and our dataset is large (*e.g*. cartoon cars, snow-covered cars). These issues can be potentially relieved by collecting more videos to further enrich the diversity of cars in our dataset.

5.2. Analysis

The emergence of canonical pose. The key to the success of the proposed model is the emergence of the canonical pose in our first training stage. Fig. 5 provides images from our dataset with similar pose annotations after the calibration step (Sec. 4.2). On the one hand, Fig. 5 clearly demonstrates that the calibrated pose annotations well align in a uniform frame. On the other hand, the calibration fails on several videos due to the limited performance of our stage-one model (Fig. 5 bottom). A typical failure case is that the pose predictor misidentifies the frontal view of a car as the rearview. Such failure cases of pose calibration introduce noisy pose annotations into our dataset. Fortunately, we find that the noise level of the annotations is closely correlated with the calibration error $\mathcal{L}^*_{\text{cali}}$ (Eqn. 3). We thus use the calibration error $\mathcal{L}^*_{\text{cali}}$ as a heuristic to filter out noisy annotations in our second training stage. We provide a detailed analysis below.

The effect of the noise level in the annotations. We use the calibration error $\mathcal{L}^*_{\text{cali}}$ (Eqn. 3) as an indicator of the noise level of the pose annotations. A higher threshold on the calibration error corresponds to a larger number of training images yet more noisy annotations, and vice versa. Fig. 6 presents the performance of our model under different noise levels of the annotations. It demonstrates that neither clean-yet-small data nor large-yet-noisy data lead to higher performance than mid-size data with mid-level noise.

The effect of two-stage training. As demonstrated in Fig 5, the model trained in the first stage provides a tool to calibrate the pose annotations of our dataset. However, the performance of the stage-one model lags behind the state-of-the-art analysis-by-synthesis frameworks (*e.g*. SSV [24] and ViewNet [22]). We hypothesize that training to predict the relative pose is a suboptimal learning strategy for

Figure 5. **Canonical pose emerges in our first training stage (Sec. 4.1).** For each reference image (top), we present four matches (including one failure case) of which the pose annotations have less than 5° angular distance to that of the reference frame. The calibration error $\mathcal{L}^*_{\text{cali}}$ (Eqn. 3) is higher than 25° on these failure cases while lower than 10° on the well-calibrated video instances. This provides us with a heuristic to filter out noisy annotations.

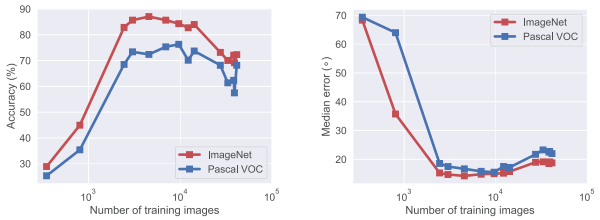

Figure 6. **The effect of annotation noise level on 3D pose prediction.** We report the performance of our pose estimation model under different noise levels of pose annotations. A higher level of annotation noise corresponds to a larger number of training images. We report both prediction accuracy (left panel) and median error (right panel) on two test splits included in PASCAL3D+.

the task of absolute pose estimation. As shown in Tab. 2, the model trained in our second training stage significantly outperforms the one trained in the first stage. This suggests that learning with absolute pose annotations is a more effective training method. However, our stage-two training is not possible without the pose calibration and stage-one model. Therefore, the proposed two training stages are complementary and both play an important role in our framework.

The effect of network initialization. The recent self-supervised learning (SSL) [14, 3] has significantly improves the unsupervised pose estimation [4] and part discovery [27]. We initialize our pose estimation network with ImageNet-pretrained models by default. However, ImageNet classification labels require extensive human labor. A natural question is how the recent SSL methods help us further reduce the requirement of human annotations. Tab. 3 provides a comparison of different initialization strategies. Supervised ImageNet pretraining and unsupervised contrastive pretraining [14, 3] have similar performance in the task of pose estimation, while both outperform the random initialization in a large margin.

Table 2. **The effect of two-stage training on 3D pose prediction.** The second stage trains the model to regress to absolute pose after using the first stage model to calibrate the relative pose annotations. This procedure leads to a significant improvement in pose estimation accuracy (%) and median error (°), in spite of the training datasets.

Trainset	Stage	PASCAL VOC		ImageNet	
		Acc. ↑	Med. ↓	Acc. ↑	Med. ↓
Acci. Turn.	1	42	38.8	46	32.9
	2	**75**	**15.8**	**86**	**14.8**
FreiburgCars	1	36	44	47	31.9
	2	**72**	**15.7**	**84**	**15.0**

Table 3. **The effect of network initialization on 3D pose prediction.** ImageNet pretrained models provide a significant improvement over random initialized ones but self-supervised counterparts are competitive alternatives without having to resort to extra human annotations.

Initialization	PASCAL VOC		ImageNet	
	Acc. ↑	Med. ↓	Acc. ↑	Med. ↓
Random	58	25	70	20.2
Contrastive [14]	74	15.7	85	**14.3**
ImageNet	**75**	**15.8**	**86**	14.8

Pose distribution. Figure 7 compares the pose distribution of the Acci- dental Turntables dataset and PASCAL3D+. The distribution of azimuth is more balanced in our dataset, where PASCAL3D+ has more cars with large elevations.

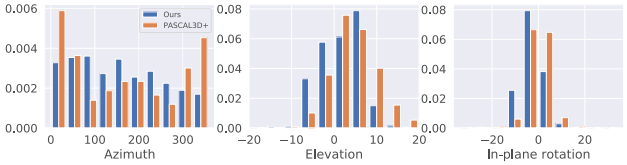

Figure 7. Distribution of the poses in the proposed accidental turntable dataset and the PASCAL3D+.

Feature extraction and matching for SfM. Feature extraction and matching is the core of SfM algorithms. The classical SIFT [20] and simple nearest neighbor matching (NN) remain the default components in popular SfM packages (*e.g.* COLMAP [29, 30]), despite of the recent success of learning-based methods [5, 28]. We observe that SfM with SIFT and NN does not work reliably on our in-the-wild video dataset. Fig. 8 compares the 3D reconstruction and pose estimation from COLMAP under different feature extraction and matching algorithms on two videos from

our dataset. SfM with SIFT and NN only provides partial 3D reconstruction and pose estimation on a small subset of frames. Its performance drops significantly on texture-free objects (Fig. 8 bottom). Simply replacing SIFT with Superpoint [5] leads to more complete 3D reconstruction and pose estimations. SfM with Superpoint and SuperGlue [28] provides the highest quality of shape reconstruction and pose estimations. Our experimental results can be explained by the following observations: SIFT detects few interest points on most cars due to the texture-free surface; SIFT extracts features in a small local region, which results in large ambiguity in matching duplicated patterns (*e.g.* frontal and rear wheels of a car); large motion blur further destabilizes the feature-matching process; In comparison, Superpoint provides rich interest points even in texture-free regions; Lastly, SuperGLUE aggregates long-range contextual information via an attention mechanism, which we find significantly reduces the ambiguity in matching repeated patterns. Fig. 9 provides more examples from our accidental turntable dataset. The performance of SfM may drop on highly-occluded objects (*e.g.* the car is occluded by smoke in Fig. 9 bottom).

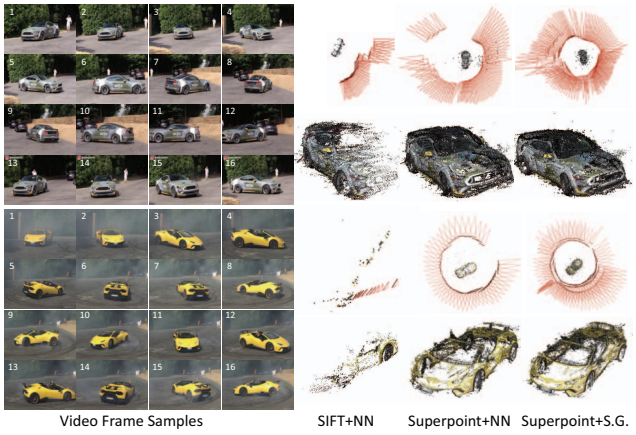

Figure 8. **Feature extraction and matching for structure-from-motion. Left**: video samples from the proposed accidental turntable dataset. **Right**: pose estimations (top) and dense 3D reconstruction (bottom) under different feature extraction (SIFT [20] or Superpoint [5]) and matching (nearest neighbor (NN) or SuperGlue (S.G.) [28]) algorithms. The red square pyramids indicate the location of the estimated camera pose. Each video consists of more than 200 frames and the car turns around 720°.

Extension to other categories. There are a fair number of turntable videos for other categories on Youtube. For example, airplanes turn along the runway[1]; landing or takeoff of airplanes usually induces more than 90-degree pose changes relative to the camera[2]; cruises turn[3]. Fig. 10 shows SfM with Superpoint, and SuperGlue provides reasonable pose estimation and 3D reconstruction on these categories. Even though we focus on cars in this work, our dataset is much larger, easier to collect, and more useful to train a pose estimator than existing car datasets (*e.g.*, FreiburgCars).

[1] https://youtu.be/khesztRJKUw
[2] https://youtu.be/Z7CutgNEMfA?t=30
[3] https://youtu.be/CgaJgRdI3FQ

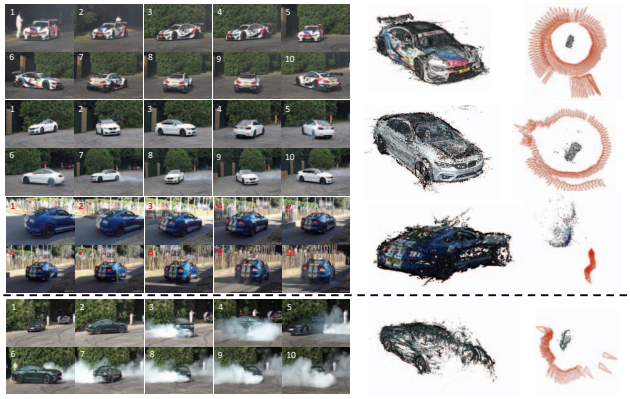

Figure 9. **More examples from the Accidental Turntables dataset.** SfM provides accurate 3D reconstructions and pose estimations on either texture-rich (**1st row**) or texture-free (**2nd row**) objects, as well as objects moving along a straight line without any turns (**3rd row**). The performance drops on highly-occluded objects (**bottom**).

Figure 10. Accidental turntables for airplanes and cruise. **Left:** video frame samples. **Right:** pose estimation and 3D reconstruction from structure-from-motion.

6. Discussion and Conclusion

We propose to learn 3D pose estimation models from a new source of data: videos where objects turn. We demonstrate that classical structure-from-motion algorithms, coupled with the recent advances in feature matching and object detection, provide surprisingly accurate pose estimations and 3D reconstructions on in-the-wild car videos. We also provide a novel learning framework that successfully trains a high-quality 3D pose predictor on the collected video datasets. We plan to release our **Accidental Turntable dataset** along with the pose estimations and 3D reconstructions from the enhanced SfM for the research community.

Acknowledgements. The research is funded in part by National Science Foundation (USA) under grants #1749833 and #1908669 to Subhransu Maji. Our experiments were performed on the University of Massachusetts GPU cluster funded by the Mass. Technology Collaborative.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7822–7831, 2021. 2

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 7

[4] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9897–9906, 2021. 7

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 3, 7, 8

[6] Gilad Divon and Ayellet Tal. Viewpoint estimation—insights & model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–268, 2018. 2

[7] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018. 2

[8] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. *Advances in Neural Information Processing Systems*, 33:8614–8625, 2020. 2

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[10] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020. 2

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[12] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 4

[13] Alexander Grabner, Peter M Roth, and Vincent Lepetit. 3d pose estimation and 3d model retrieval for objects in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3022–3031, 2018. 2, 6

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 7

[15] 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, 2, 3, 5

[16] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 2

[17] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019. 2

[18] Shuai Liao, Efstratios Gavves, and Cees GM Snoek. Spherical regression: Learning viewpoints, surface normals and 3d rotations on n-spheres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9767, 2019. 2, 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3, 5, 7, 8

[21] Siddharth Mahendran, Haider Ali, and René Vidal. 3d pose regression using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2174–2182, 2017. 2, 6

[22] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Viewnet: Unsupervised viewpoint estimation from conditional generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10418–10428, 2021. 2, 5, 6

[23] Kieran Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*, 2021. 2

[24] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3981, 2020. 2, 5, 6

[25] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5218–5227, 2017. 2, 4, 5, 6

[26] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785. IEEE, 2009. 2, 5

[27] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Ganorcon: Are generative models useful for few-shot segmentation? *arXiv preprint arXiv:2112.00854*, 2021. 7

[28] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 3, 5, 7, 8

[29] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 3, 7

[30] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2, 3, 7

[31] Nima Sedaghat and Thomas Brox. Unsupervised generation of a viewpoint annotated car dataset from videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1322, 2015. 2, 5, 6

[32] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang 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. 2

[33] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pages 2686–2694, 2015. 2, 3

[34] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 2

[35] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381. IEEE, 2012. 2

[36] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 2, 3, 6

[37] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[38] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373. IEEE, 2020. 2

[39] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European conference on computer vision*, pages 160–176. Springer, 2016. 2

[40] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 2, 5

[41] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2

[42] 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. 2, 3, 4, 5

[43] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3d objects. *arXiv preprint arXiv:1906.05105*, 2019. 2, 4, 5

[44] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 5

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 2, 4