A Refined 3D Pose Dataset for Fine-Grained Object Categories

Yaming Wang[1] Xiao Tan[2] Yi Yang[2] Ziyu Li[2] Xiao Liu[2] Feng Zhou[2] Larry S. Davis[1]
[1]University of Maryland [2]Baidu Research

Abstract

Most existing 3D pose datasets of object categories are limited to generic object types and lack of fine-grained information. In this work, we introduce a new large-scale dataset that consists of 409 fine-grained categories and 31,881 images with refined 3D pose annotation. Specifically, we augment three existing fine-grained object recognition datasets (StanfordCars, CompCars and FGVC-Aircraft) by finding a specific 3D model for each sub-category from ShapeNet and manually annotating each 2D image with a full set of 7 continuous perspective parameters. Since the 2D projection of fine-grained 3D shapes can be an exact fit of object segmentation, we further improve the annotation quality by initializing from the human annotation and conducting a local search of the pose parameters to maximize the IoUs between the projected mask and the segmentation reference predicted from state-of-the-art segmentation networks. We provide full statistics of the annotations with qualitative and quantitative comparisons suggesting that our dataset can be a complementary source for studying 3D pose estimation. The dataset can be downloaded at `http://users.umiacs.umd.edu/~wym/3dpose.html`.

1. Introduction

Estimating 3D object pose from a single 2D image is an inevitable step in various industrial applications, such as vehicle damage detection [10], novel view synthesis [36, 23], grasp planning [28] and autonomous driving [5]. To address this task, collecting suitable data is of vital importance. However, due to the expensive annotation cost, most existing large scale 3D pose datasets such as Pascal3D+ [34] and ObjectNet3D [33], are collected for generic object types and may lack of accurate pose information, since different objects in one hyper class (*i.e.*, cars) are only matched with a few generic 3D shapes, leading to a high projection error that affects human annotators to find the accurate pose, as demonstrated in Figure 1.

In this work, we introduce a new benchmark pose estimation dataset for fine-grained object categories. Specifically, we augment three existing fine-grained recognition datasets, StanfordCars [14], CompCars [35] and FGVC-Aircraft [20], with two types of useful 3D information: (1) for each object in the image, we manually annotate the full perspective projection represented by 7 continuous pose parameters; (2) we provide an accurate match of the computer aided design (CAD) model for each fine-grained object category. The resulting augmented dataset consists of more than 30,000 images for over 400 fine-grained object categories. Table 1 shows the general statistics of our dataset.

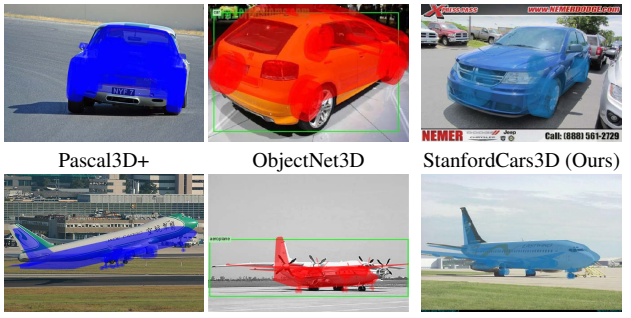

Figure 1. While both Pascal3D+ and ObjectNet3D contain more complicated scenarios with more generic categories for 3D pose estimation, we provide more accurate pose annotations on a large set of fine-grained object classes as a complementary source for studying 3D pose estimation.

To the best of our knowledge, our dataset is the very first one that employs fine-grained category aware 3D models in pose annotation. To fully utilize the valuable fine-grained information, we further develop an automatic pose refinement mechanism to improve over the human annotations. Thanks to the fine-grained shapes, an accurate pose parameter also leads to the optimal segmentation overlap between the projected 2D mask from the 3D model and the target object ground truth segmentation. We hence conduct a local greedy search over the 7 full perspective pose parameters, initialized from the human annotation, to maximize the segmentation overlap objective. To avoid human effort on annotating ground truth segmentation, we utilize state-of-the-art image segmentation models including both Mask R-CNN [9] and DeepLab v3+ [3] to obtain the as-accurate-as-possible segmentation references. Figure 2 illustrates this

Dataset	# class	# image	annotation	fine-grained
3D Object [25]	10	6,675	discretized view	✗
EPFL Car [22]	1	2,299	continuous view	✗
IKEA [16]	11	759	2d-3d alignment	✗
Pascal3D+ [34]	12	30,899	2d-3d alignment	✗
ObjectNet3D [33]	100	90,127	2d-3d alignment	✗
StanfordCars3D	196	16,185	2d-3d alignment	✓
CompCars3D	113	5,696	2d-3d alignment	✓
FGVC-Aircraft3D	100	10,000	2d-3d alignment	✓
Total (Ours)	409	31,881	2d-3d alignment	✓

Table 1. Comparison between our 3D pose estimation dataset (StanfordCars3D + CompCars3D + FGVC-Aircraft3D) and other benchmark datasets. Our dataset can be viewed as a complementary source to the existing large scale 3D pose dataset (Pascal3D+ and ObjectNet3D) with a different focus on more intra-class categories and fine-grained details.

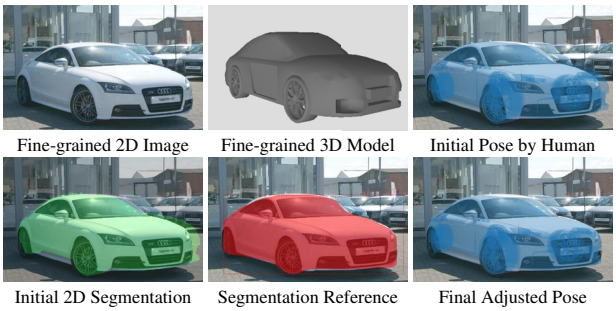

Figure 2. For an image with a fine-grained category (Top left), we first find its corresponding fine-grained 3D model (Top middle) and manually annotate its rough pose (Top right). Since the problem is to estimate the object pose such that the projection of the 3D model aligns with the image as well as possible, we further optimize the segmentation overlap between the projected 2D mask (Bottom left) and the "groundtruth" mask (Bottom middle) estimated from state-of-the-art CNN models to obtain the final 3D pose (Bottom right).

process.

To verify the effect of the segmentation based refinement, we conduct quantitative and qualitative comparisons. Qualitatively, the human evaluation shows that around 50% of annotations are improved significantly from its original labels. Quantitatively, our annotation provides a significantly tighter segmentation overlap on car and airplane categories compared to Pascal3D+ [34] and ObjectNet3D [33].

In summary, we collect a new large-scale 3D pose dataset for fine-grained objects with more accurate annotations. The dataset contains a full perspective model parameters including the camera focal length, which can be a more challenging benchmark for developing algorithms beyond only estimating viewpoint angles (azimuth) [8] or recovering the rotation matrices [19]. We further propose a simple but effective way to automatically refine the pose annotation based on the segmentation cues. With the correct 3D fine-grained model, this method can automatically refine object pose while significantly alleviating the human label effort.

2. Related Work

3D Pose Estimation Dataset. Annotating 3D pose in 2D images requires expensive effort. Due to the 3D ambiguity from 2D images, earlier 3D pose datasets are limited not only in scale but also precision [25]. For example, EPFL Car dataset [22] consists of 2,299 images of 20 car instances captured at different azimuth angles with other parameters including elevation and distance kept almost the same for all the instances. Pascal3D+ [34] and ObjectNet3D [33] are probably the only two large-scale 3D pose datasets for generic object categories. However, both of them assume a camera model with 6 parameters to annotate. Moreover, they use only a few CAD models to match all different objects in one hyper class (*i.e.*, cars, airplanes). The projection error could be large due to the lack of accurate CAD models and may affect the pose accuracy in human annotation. Being aware of these problems, we use a full perspective model and project fine-grained CAD models to match objects in the 2D images to provide a more accurate pose annotation.

Fine-Grained Recognition Dataset. Fine-grained recognition is a new challenge for automatically discriminating categories with only small subtle visual differences [31, 14, 27]. Due to its importance in real world applications, a number of fine-grained datasets has been released, ranging from plants and animals [30, 1, 13, 11, 24, 18, 21, 32] to human-made objects [29, 20, 35, 7, 14, 17]. Almost all existing fine-grained datasets are lack of 3D pose or 3D shape labels [14], and pose estimation for fine-grained object categories are not well-studied. Our work fills the gap by annotating poses and matching CAD models on three existing popular fine-grained recognition datasets including StanfordCars [14], CompCars [35] and FGVC-Aircraft [20].

3D Model Dataset. Similar to [33], we adopt the 2d-3d alignment method to annotate object poses, Annotating in such a way requires a source for accessing accurate 3D models of objects. Luckily, there has been substantial growth in the number of 3D models available online over the last decade [4, 6, 12, 15] with well-known repositories like the Princeton Shape Benchmark [26] which contains around 1,800 3D models grouped into 90 categories. In this work, we use ShapeNet [2], the so-far largest 3D CAD model database which has indexed more than 3,000,000 models, with 220,000 models out of which are classified into 3,135 categories including various object types such as cars, airplanes, bicycles, etc. The large amount of 3D shapes allow us to find an exact model to many of the objects in the natural images. For example, ShapeNet provides

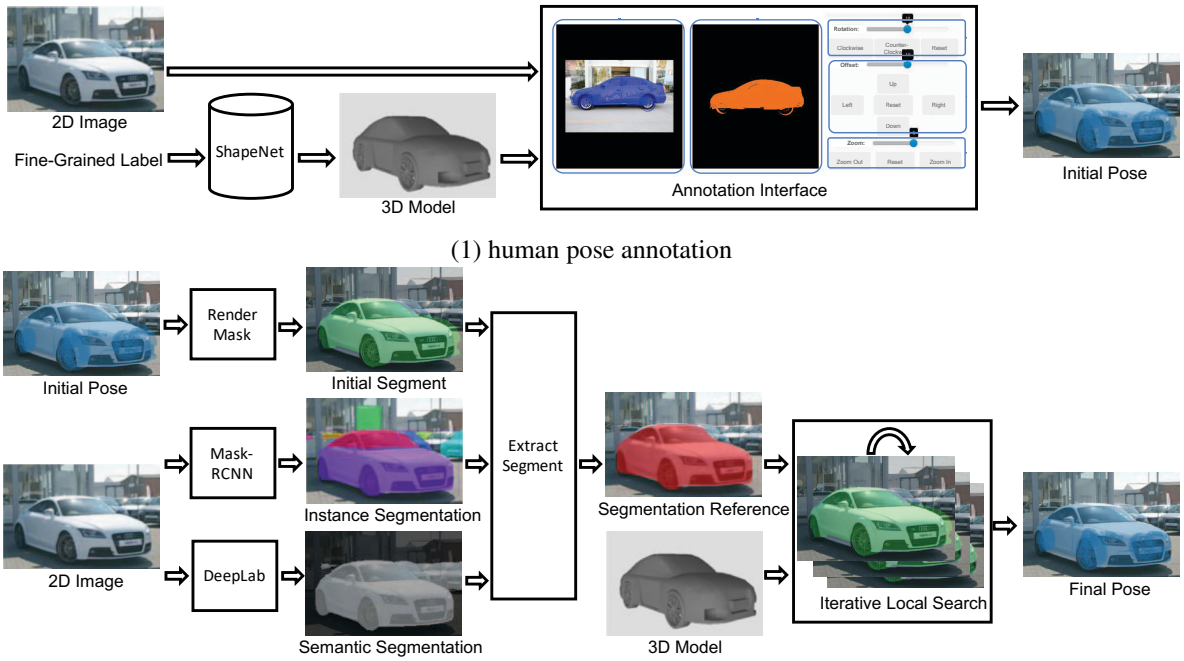

Figure 3. An overview of our whole annotation framework which includes two parts: (1) human initial pose annotation, and (2) segmentation based pose refinement. The human annotation provides a strong initialization for the second-stage pose refinement, hence we only need to conduct local search to adjust the pose.

183,533 models for the car category and 114,045 models for the airplane category. Although we only annotate three fine-grained datasets, our annotation framework can be used to apply to build more 3D pose dataset, thanks to larger-scale datasets like ShapeNet [2] and iNaturalist [27].

3. Dataset Construction

We build three fine-grained 3D pose datasets. Each dataset consists of three parts: 2D images, 3D models and 3D poses. The 2D images are collected from StanfordCars [14], CompCars [35] and FGVC-Aircraft [20] respectively. Annotating the 3D model and pose involves two main steps: (1) human pose annotation, (2) segmentation based pose refinement. Figure 3 illustrates the whole process.

Our human pose annotation process is similar to ObjectNet3D [33] but requires more effort on selecting finer 3D models. We first select the most appropriate 3D model from ShapeNet [2] for each object in the fine-grained image dataset. We then obtain the 7 pose parameters by asking the annotators to align the projection of the 3D model to the corresponding image using our designed interface.

Although a human can initiate the pose annotation with reasonably high efficiency and accuracy, we find it hard for them to adjust the fine detailed poses given a limited amount of time. Our second-stage segmentation based pose refinement further adjusts the pose parameters by performing a local greedy search initialized from the human annotation. We discuss the details of each process in the next subsections.

3.1. 3D Models

To better annotate the 3D pose, we adopt a distinct model for each category. Thanks to ShapeNet [2], we can find the corresponding 3D models with its fine-grained object category. If there is no exact match between a category and the 3D model, we manually select a visually similar one for that category. For StanfordCars [14], we annotate images for all 196 categories, where 148 categories have exact matched 3D models. For CompCars [35], we include 113 categories with matched 3D models. For FGVC-Aircraft [20], we annotate images for all 100 categories with more than 70 matched models. To the best of our knowledge, our dataset is the very first one that employs fine-grained 3D models in 3D pose estimation.

3.2. Camera Model

We define the world coordinate system in accordance with the 3D model coordinate system. A point $\mathbf{X}$ on a 3D model is projected onto a point $\mathbf{x}$ in a 2D image:

$$\mathbf{x} = \mathcal{P}\mathbf{X}, \tag{1}$$

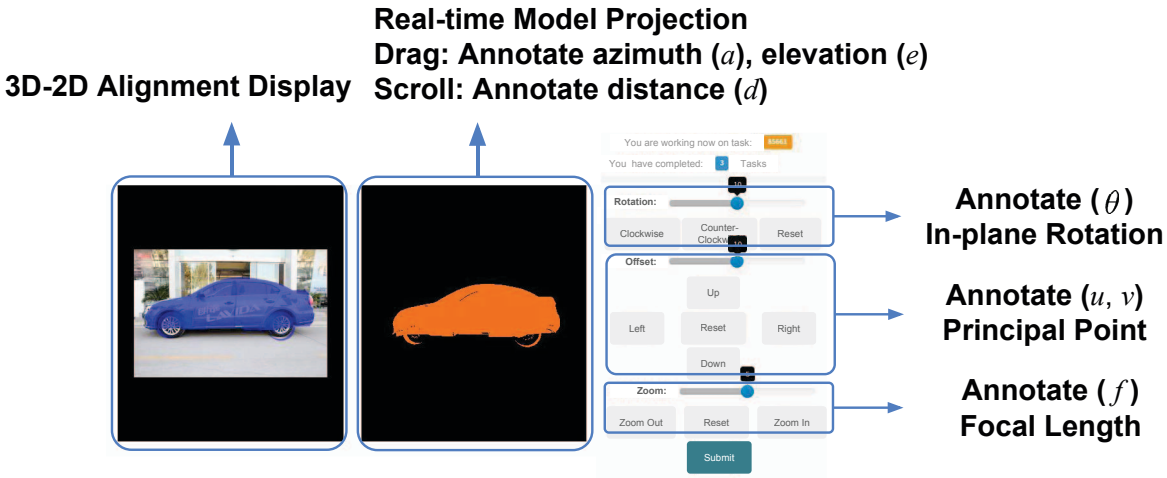

Figure 4. An overview of our annotation interface. Our annotation tool renders the projected 2D mask onto the image in real time to facilitate the annotators to better adjust pose parameters.

via a perspective projection matrix:

$$\mathcal{P} = K\left[R|T\right], \tag{2}$$

where K denotes the intrinsic parameter matrix:

$$K = \begin{bmatrix} f & 0 & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

R encodes a 3×3 rotation matrix between the world and camera coordinate systems, parameterized by three angles, *i.e.*, elevation e, azimuth a and in-plane rotation θ. We assume that the camera is always facing towards the origin of the 3D model. Hence the translation $T = [0, 0, d]^{\mathrm{T}}$ is only defined up to the model depth d, the distance between the origins of the two coordinate systems, and the principal point (u, v) is the projection of the origin of world coordinate system on the image. As a result, our model has 7 continuous parameters: camera focal length f, principal point location (u, v), azimuth a, elevation e, in-plane rotation θ and depth d. Note that since images are collected online, the annotated intrinsic parameters (u, v and f) are approximations. Compared to previous datasets [34, 33] with 6 parameters (f fixed), our camera model considers both the camera focal length f and object depth d in a full perspective projection for finer 2D-3D alignment, which allows for a more flexible pose adjustment and a better shape matching.

3.3. 2D-3D Alignment

We annotate 3D pose information for all 2D images through crowd-sourcing. To facilitate the annotation process, we develop an annotation tool illustrated in Figure 4. For each image during annotation, we choose the 3D model according to the fine-grained label given beforehand. We then ask the annotators to adjust the 7 parameters so that the projected 3D model is aligned with the target object in the 2D image. This process can be roughly summarized as follows: (1) shift the 3D model such that the center of the model (the origin of the world coordinate system) is roughly aligned with the center of the target object in the 2D image; (2) rotate the model to the same orientation as the target object in the 2D image; (3) adjust the model depth d and camera focal length f to match the size of the target object in the 2D image. Some finer adjustment might be applied after the three main steps. In this way we annotate all 7 parameters across the whole dataset. On average, each image takes approximately 1 minute to annotate by an experienced annotator. To ensure the quality, after one round of annotation across the whole dataset, we perform quality check and let the annotators do a second round revision for the unqualified examples.

3.4. Segmentation Based Pose Refinement

Although human annotators already provide reasonably accurate annotation in the first stage, we notice that there are still potential to further improve the annotation quality. This is because humans are good at providing a strong initial pose estimate but finetuning the detailed pose parameters is very annoying. Realizing that ultimate problem is to estimate the object pose such that the projection of the 3D model aligns with the image, we design a simple but effective iterative greedy search algorithm to automatically refine pose parameters by maximizing

$$\max_{\boldsymbol{p}} IoU(S(\boldsymbol{p}, \mathcal{M}), s^*), \tag{4}$$

where s^* is the 2D object segmentation reference and $S(\boldsymbol{p}, \mathcal{M})$ maps a 3D model $\mathcal{M}$ to a 2D mask according to

Algorithm 1 Iterative local pose search algorithm:

Input: 3D model $\mathcal{M}$, Human pose annotation $\boldsymbol{p}_0$, segmentation reference s^*, 2D mask generator $S(\boldsymbol{p}, \mathcal{M})$, segmentation evaluation function $IoU(s_1, s_2)$, pose parameter update unit $\boldsymbol{\epsilon}$, update step size α.

Output: Optimized pose parameter $\boldsymbol{p}^*$.

for each image with segmentation reference s^* **do**
  Initialize pose parameters $\boldsymbol{p} = \boldsymbol{p}_0$.
  Initialize 2D mask $s = S(\boldsymbol{p}, \mathcal{M})$
  Initialize $iou = IoU(s, s^*)$
  repeat
    Update $iou_{last} = iou$.
    for each dimension i in $\boldsymbol{p}$ **do**
      Update $\boldsymbol{p}_i^+ = \boldsymbol{p}_i + \alpha\boldsymbol{\epsilon}_i$
      Update $\boldsymbol{p}_i^- = \boldsymbol{p}_i - \alpha\boldsymbol{\epsilon}_i$
      Render new 2D mask $s^+ = S(\boldsymbol{p}^+, \mathcal{M})$
      Render new 2D mask $s^- = S(\boldsymbol{p}^-, \mathcal{M})$
      Update $iou^+ = IoU(s^+, s^*)$
      Update $iou^- = IoU(s^-, s^*)$
      Update $iou = \max(iou, iou^+, iou^-)$
      Update $\boldsymbol{p} = \arg\max(iou, iou^+, iou^-)$
    end for
    if $iou == iou_{last}$ **then**
      Update $\alpha = \alpha/2$
      if $\alpha <= threshold$ **then**
        Set as convergence.
      end if
    else
      Continue.
    end if
  until converge
end for

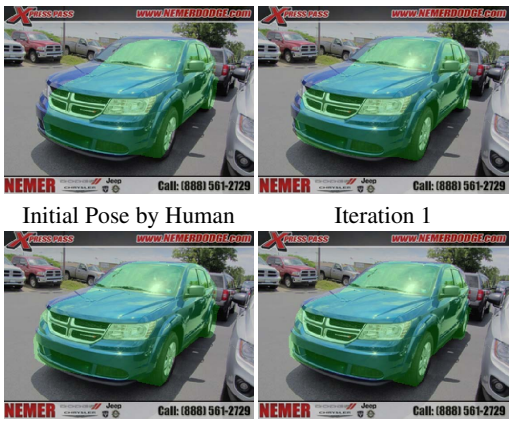

Figure 5. Iterative local greedy search for the fine detailed pose, initialized from human annotation. The green highlights are the 2D masks projected by the 3D model during pose optimization.

the pose parameter $\boldsymbol{p} = (a, e, \theta, d, f, u, v)$.

The algorithm aims to finetune the 7 pose parameters to maximize the segmentation overlap between the projected 2D mask from the 3D model and the segmentation reference. We use the traditional *intersection over union* as the segmentation overlapping criterion. The algorithm greedily updates pose parameters, it is hence a local search algorithm with guarantee to converge to a local optimum. During the local search process, we observe it converges in 3-10 iterations with 1 minute per image on average. Algorithm 1 shows the overall process. Figure 5 illustrates the local search algorithm.

3.5. Segmentation Reference

To conduct the local greedy search, we ideally need the ground truth target object segmentation. Although annotating segmentation through crowd-sourcing is possible, we find using existing state-of-the-art image segmentation models such as Mask R-CNN [9] and DeepLab v3+ [3] can already provide satisfying segmentation results. For example, on the Pascal VOC2012 segmentation benchmark, DeepLab v3+ can reach average IoUs of 93.2 on the "car" class and 97.0 on the "aeroplane" class respectively. Mask R-CNN, although not accurate as DeepLab, can obtain instance-level segmentation, which is particularly useful for images with more than 1 instance from the same class. In the end, we use a combination of both models to find the most appropriate segmentation reference. Figure 6 illustrates the process.

3.6. Dataset Statistics

We plot the distributions of the 7 parameters in Figure 7 for StanfordCars3D, CompCars3D and FGVC-Aircraft3D respectively. Unsurprisingly, all the parameters are not uniformly distributed due to the nature of the original dataset. The most challenging parameter across the three datasets is azimuth (a), which varies across the $360°$, while elevation (e) and in-plane rotation (θ) are relatively concentrated in a small range around $0°$ since the images of cars and airplanes are often taken from the ground view. The distribution of focal length (f) and model depth (d) are also not widespread enough because objects in these fine-grained images are normalized and cropped to a standard size. Although the parameter distribution issue may raise concerns about learning trivial solutions, we believe that our effort still provides a reasonable diversity on pose annotation. For example, the distribution of azimuth (a) are quite different across the three datasets and complementary to each other. This could encourage building a more generalized pose estimation model.

3.7. Dataset Split

We split the three datasets in this way. For StanfordCars3D, we follow the standard train/test split provided by

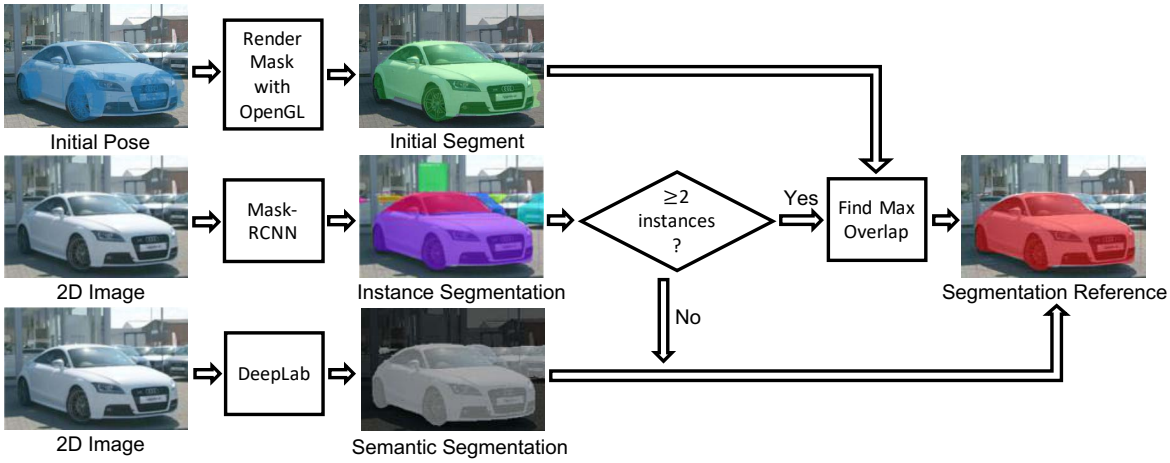

Figure 6. An illustration of our reference segmentation extraction process. Ideally, we can ask human annotators to annotate the ground truth segment for the target object in a 2D image. However, we find CNNs such as Mask-RCNN and DeepLab can already provide accurate enough segmentation predictions for the pose refinement.

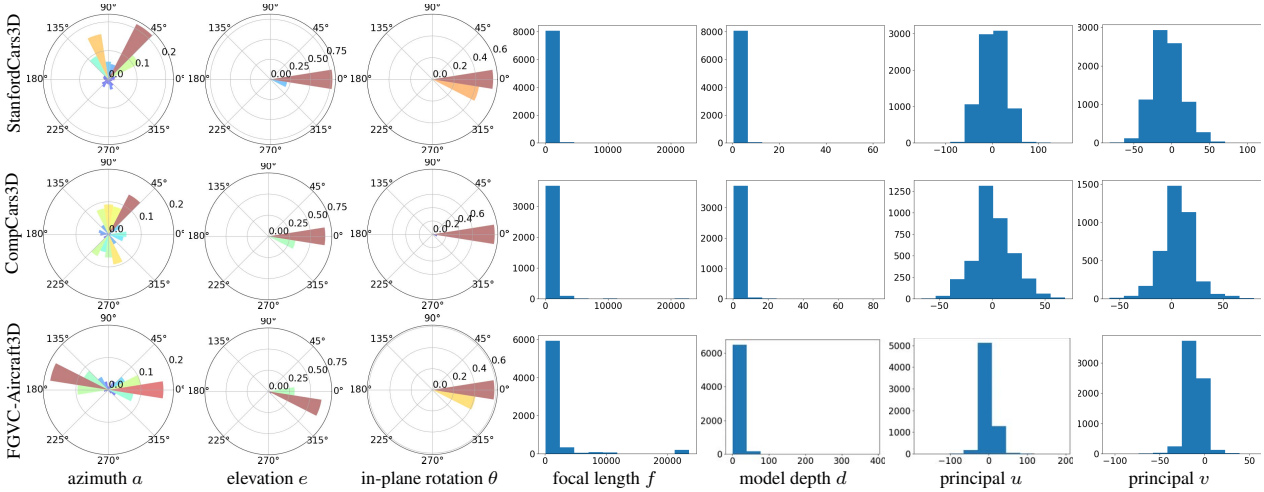

Figure 7. The polar histograms of the three rotation parameters as well as the histograms of the other four parameters in our annotated dataset.

the original dataset [14] with 8144 training examples and 8041 testing examples. For CompCars3D, we randomly sample $2/3$ of our annotated data as training set and the rest $1/3$ as testing set, resulting in 3798 training and 1898 test examples. We provide the train/test split information in the dataset release. For FGVC-Aircraft3D, we follow the standard train/test split provided by the original dataset [20] with 6667 training examples and 3333 testing examples.

4. Dataset Comparison

4.1. Compare to Existing Dataset

We compare our annotation quality with two existing large-scale 3D pose dataset, PASCAL3D+ [34] and ObjectNet3D [33]. It is worth to note that we are not aiming to show the superiority of our dataset, since the two previous datasets consider more general scenarios with multiple objects and challenging occlusion in an image. However, we hope that by comparing to them, we demonstrate our fine-grained pose dataset can become a complementary resource for studying 3D pose estimation in monocular images.

Figure 8 and Figure 9 show the qualitative comparison on the "car" class and the "aeroplane" class respectively. Overall, we find our annotation more satisfying by visually comparing the overlay images which maps the 3D model on the 2D image. To further conduct quantitative comparison, we use segmentation overlap between the projected 2D mask and the ground truth object mask as the evaluation measure. We randomly select 50 "car" images and 50 "aeroplane" images from PASCAL3D+ and ObjectNet3D respectively. We then randomly pick 50 images from StanfordCars3D and FGVC-Aircraft3D. In total, we select 300

Figure 8. Qualitative comparison of ground-truth pose annotation between our StanfordCars3D and two existing large scale 3D pose datatset. We randomly select 5 car images from each dataset. While both Pascal3D+ and ObjectNet3D provide more complicated scenarios with more generic categories for 3D pose estimation, our pose annotations look more accurate thanks to the fine-grained shape matching.

Figure 9. Qualitative comparison of ground-truth pose annotation between our FGVC-Aircraft3D and two existing large scale 3D pose datatset. We randomly select 5 aircraft images from each dataset.

images and annotate them with ground truth segmentation.

Since both PASCAL3D+ and StanfordCars3D consider more complicated scenarios such as multiple objects with cluttered background, we filter out those images containing more than one object with reasonably large size for a fair comparison. Hence the average IoUs can be an optimistic estimate for both baseline datasets. Even with that, our annotation shows a clear segmentation improvement on average IoUs on both "car" and "aeroplane", as demonstrated in Table 2. Particularly, both the mean and the standard deviation of the segmentation IoUs get significantly improved, indicating that our annotations are not only more accurate but more stable as well.

car	PASCAL3D+ [34]	ObjectNet3D [33]	StanfordCars3D
	78.5% $\pm$ 8.6%	84.1% $\pm$ 6.0%	**90.4%** $\pm$ **3.3%**
airplane	PASCAL3D+ [34]	ObjectNet3D [33]	FGVC-Aircraft3D
	62.7% $\pm$ 13.1%	65.1% $\pm$ 11.0%	**78.9%** $\pm$ **9.4%**

Table 2. Comparison on the average IoUs with the standard deviation on the "car" category and "aeroplane" category. Note that in this evaluation, we manually annotate around 50 ground truth segmentation masks for each dataset.

4.2. Compare to Human Annotation

We also analyze how much gain we get by conducting segmentation based pose refinement. To understand this, we utilize the manually annotated ground truth 2D segmentation on the randomly select 100 images from the StanfordCars and FGVC-Aircraft. We then compare the average IoUs between human annotated pose and the refined pose. Table 3 shows the improvement of segmentation overlap on the three datasets. On StanfordCars3D, our second-stage refinement improves average IoUs from 84.1% to 90.4%, which is significant. On FGVC-Aircarft3D, the improvement is even more, from 65.3% to 78.9%. Figure 10 and Figure 11 illustrate the pose improvement qualitatively.

Figure 10. Selected examples illustrating the second-stage automatic pose refinement improving the initial human pose annotation on StanfordCars3D dataset.

Figure 11. Selected examples illustrating the second-stage automatic pose refinement improving the initial human pose annotation on FGVC-Aircraft3D dataset.

Average IoUs	Human Annotation	Refined Annotation
StanfordCars3D	84.1% ± 6.2%	**90.4% ± 3.3%**
FGVC-Aircraft3D	65.3% ± 19.9%	**78.9% ± 9.4%**

Table 3. Segmentation evaluation of initial human annotation and after iterative pose refinement on the two datasets. Note that in this evaluation, we manually annotate around 50 ground truth segmentation masks for each dataset.

Considering segmentation overlap may not be the only appropriate quantitative measure, we further conduct a human study to compare the pose annotation quality. To do this, we hire 5 professional annotators, show them the 2D-3D alignment of the same image with annotation in the two stages simultaneously and let them rate the relative quality for the 50 selected images in each dataset. The relative comparison consists of "Worse", "Equal" or "Better", indicating the second-stage pose is either significantly worse, roughly equal or significantly better than the first-stage human annotation from the subjective point of view. Table 4 shows the human study result. Surprisingly, the second-stage refined pose is either roughly equal or significantly better than the initial human annotation, suggesting the benefit of utilizing segmentation cues to facilitate the pose search.

5. Conclusions

In this work, we annotate three popular fine-grained recognition datasets with 3D shapes and poses, ending in total 31,881 images with 409 classes. By utilizing image segmentation as an intermediate cue, we further improve the pose annotation quality. It is worth to note that given unlimited time human may ultimately produce high quality annotation, but the segmentation based pose refinement provides a better trade-off between cost and accuracy. In the future, we would like to develop a new annotation system that combines the segmentation based refinement in the loop with the human annotation interface.

	Worse	Equal	Better
StanfordCars3D	13.0%	28.3%	58.7%
FGVC-Aircraft3D	12.8%	40.4%	46.8%

Table 4. Human satisfaction rate by comparing original human annotation with refined pose. "Worse" means refined pose is worse than initial pose. "Better" means refined is better. "Equal" meaning the annotation are roughly the same. From the table, we can see humans are much more satisfied with the refined pose annotation.

References

[1] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014.

[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.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. In *Acm transactions on graphics (tog)*, volume 28, page 73. ACM, 2009.

[5] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3D object detection for autonomous driving. In *CVPR*, 2016.

[6] Xiaobai Chen, Abulhair Saparov, Bill Pang, and Thomas Funkhouser. Schelling points on 3d surface meshes. *ACM Transactions on Graphics (TOG)*, 31(4):29, 2012.

[7] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, pages 1358–1367. IEEE, 2017.

[8] Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. Is 2D information enough for viewpoint estimation? In *BMVC*, 2014.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[10] Srimal Jayawardena et al. *Image based automatic vehicle damage detection*. PhD thesis, The Australian National University, 2013.

[11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, page 1, 2011.

[12] Vladimir G Kim, Wilmot Li, Niloy J Mitra, Siddhartha Chaudhuri, Stephen DiVerdi, and Thomas Funkhouser. Learning part-based templates from large collections of 3d shapes. *ACM Transactions on Graphics (TOG)*, 32(4):70, 2013.

[13] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016.

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshops on 3D Representation and Recognition*, 2013.

[15] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, volume 2014. ., 2014.

[16] Joseph J Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing IKEA objects: Fine pose estimation. In *ICCV*, 2013.

[17] Yen-Liang Lin, Vlad I Morariu, Winston Hsu, and Larry S Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *European Conference on Computer Vision*, pages 466–480. Springer, 2014.

[18] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *European conference on computer vision*, pages 172–185. Springer, 2012.

[19] Siddharth Mahendran, Haider Ali, and René Vidal. 3d pose regression using convolutional neural networks. In *IEEE International Conference on Computer Vision*, volume 1, page 4, 2017.

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[21] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[22] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[23] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3D view synthesis. In *CVPR*, 2017.

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[25] Silvio Savarese and Li Fei-Fei. 3D generic object categorization, localization and pose estimation. In *ICCV*, 2007.

[26] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Shape modeling applications, 2004. Proceedings*, pages 167–178. IEEE, 2004.

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist challenge 2017 dataset. *arXiv preprint arXiv:1707.06642*, 2017.

[28] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *IROS*, 2017.

[29] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3622–3629, 2014.

[30] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[32] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016.

[33] 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 *ECCV*, 2016.

[34] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, 2014.

[35] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015.

[36] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016.