Learning Single-View 3D Reconstruction with Limited Pose Supervision

Guandao Yang[1] Yin Cui[1,2] Serge Belongie[1,2] Bharath Hariharan[1]
[1] Department of Computer Science, Cornell University [2] Cornell Tech

Abstract

It is expensive to label images with 3D structure or precise camera pose. Yet, this is precisely the kind of annotation required to train single-view 3D reconstruction models. In contrast, unlabeled images or images with just category labels are easy to acquire, but few current models can use this weak supervision. We present a unified framework that can combine both types of supervision: a small amount of camera pose annotations are used to enforce pose-invariance and view-point consistency, and unlabeled images combined with an adversarial loss are used to enforce the realism of rendered, generated models. We use this unified framework to measure the impact of each form of supervision in three paradigms: semi-supervised, multi-task, and transfer learning. We show that with a combination of these ideas, we can train single-view reconstruction models that improve up to 7 points in performance (AP) when using only 1% pose annotated training data.

1. Introduction

The ability to understand 3D structure from single images is a hallmark of the human visual system and a crucial step in visual reasoning and interaction. Of course, a single image by itself does not have enough information to allow 3D reconstruction, and a machine vision system must rely on some prior over shape: all cars have wheels, for example. The crucial question is how a machine vision system can acquire such priors.

One possibility is to leverage datasets of 3D shapes [4], but obtaining such a dataset for a wide variety of categories requires either 3D modeling expertise or 3D scanning tools and is therefore expensive. Another option, extensively explored recently [27, 21], is to show the machine many different views of a multitude of objects from calibrated cameras. The machine can then use photometric consistency between rendered views of hypothesized shape and the corresponding view of the real object as a learning signal. Although more tractable than collecting 3D models, this approach is still very expensive in practice: one needs to either physically acquire *thousands* of objects and place them on a turntable, or ask human annotators to annotate images in the wild with both the camera parameters and the precise *instance* that the image depicts. The assumption that *multiple*, *calibrated* views of *thousands* of objects are available is also biologically implausible: a human infant must physically interact with objects to acquire such training data, but most humans can understand airplane shape very easily despite having played with very few airplanes.

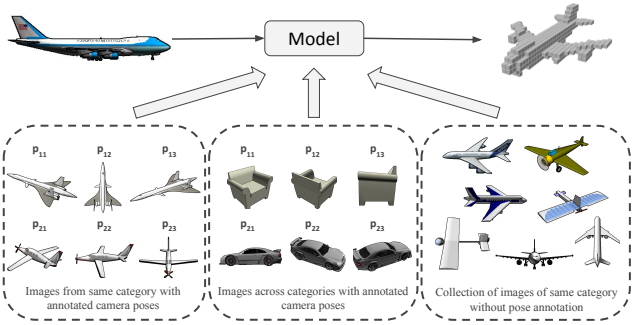

Figure 1. We propose a unified framework for single-view 3D reconstruction. Our model can be trained with different types of data, including pose-annotated images from the same object category or across multiple categories, and unlabeled images.

Our goal in this paper is to learn effective single-view 3D reconstruction models when calibrated multi-view images are available for very few objects. To do so we look at two additional sources of information. First, what if we had a large collection of images of a category but without any annotation of the precise instance or pose? Such a dataset is easy to acquire by simply downloading images of this category from the web (Fig. 1, lower right). While it might be hard to extract 3D information from such images, they can capture the distribution of the visual appearance of objects from this category. Second, we look at annotations from other semantic classes (Fig. 1, lower middle). These other classes might not tell us about the nuances of a particular class, but they can still help delineate what *shapes in general* look like. For example, most shapes are compact, smooth, tend to be convex, etc.

This paper presents a framework that can effectively use all these sources of information. First, we design a unified model architecture and loss functions that combine pose su-

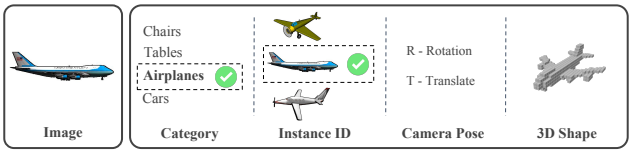

Figure 2. Different forms of training annotations for single-view 3D reconstruction. Note that some annotations (e.g. category) are cheaper to obtain than others (e.g. 3D shapes); and conversely some offer a better training signal than others.

pervision with weaker supervision from unlabeled images. Then, we use our model and training framework to evaluate and compare many training paradigms and forms of supervision to come up with the best way of using a small number of pose annotations effectively. In particular, we show that:

1. Images without instance or pose annotations are indeed useful and can provide significant gains in performance (up to 5 points in AP). At the same time a little bit of pose supervision (< 50 objects) gives a large gain (> 20 points AP) when compared to not using pose information at all.

2. Category-agnostic priors obtained by pooling training data across classes work just as well as, but not better than, category-specific priors trained on each class individually.

3. *Fine-tuning* category-agnostic models for a novel semantic class using a small amount (i.e. only 1%) of pose supervision significantly improves performance (up to 7 points in AP).

4. When faced with a novel category with nothing but a tiny set of pose-annotated images, a category-agnostic model trained on pooled data and fine-tuned on the category of interest outperforms a baseline trained on only the novel category by an enormous margin.

In summary, our results convincingly show large accuracy gains to be accrued from combining multiple sources of data (unlabeled or labeled from different classes) with a single unified model.

2. Training Paradigms

For single-view 3D reconstruction, we consider four types of annotations for an image as illustrated in Fig 2. Our goal is to minimize the need for the more expensive annotations (instance ID, camera pose and 3D shape). Towards this end, we look at three different training paradigms.

2.1. Semi-supervised single-category

In this setting, we assume all images are from a single category. Noting the fact that camera pose and model-instance annotations are difficult to collect in the wild, we

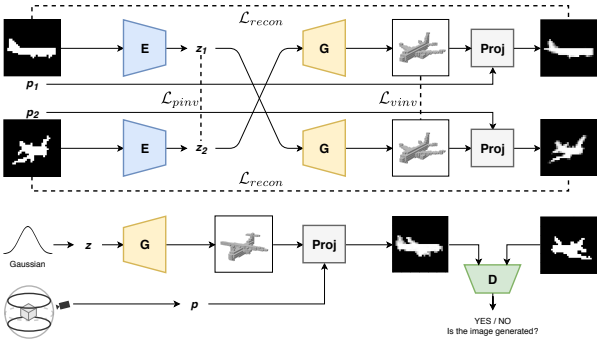

Figure 3. Overview of the proposed model architecture.

restrict to a semi-supervised setting where only some of the images are labeled with camera pose and most of them are unlabeled. Formally, we are given a dataset of images annotated with both camera pose and the instance ID: $\mathcal{X}_l = \{(\mathbf{x}_{ij}, \mathbf{p}_{ij}, i)\}_{i,j}$, where $\mathbf{x}_{ij}$ represents the j-th image of the i-th instance when projected with camera pose $\mathbf{p}_{ij}$. We also have a dataset without any annotation: $\mathcal{X}_u = \{\mathbf{x}_i\}_i$. The goal is to use $\mathcal{X}_l$ and $\mathcal{X}_u$ to learn a category-specific model for single image 3D reconstruction.

2.2. Semi-supervised multi-category

An alternative to building a separate model for each category is to build a category-agnostic model. This allows one to combine training data across multiple categories, and even use training images that do not have any category labels. Thus, instead of a separate labeled training set $\mathcal{X}_l^c$ for each category c, here we only assume a combined dataset $\mathcal{X}_l^{multi} = \mathcal{X}_l^{c_1} \cup \mathcal{X}_l^{c_2} \cup \cdots \cup \mathcal{X}_l^{c_n}$. Similarly, we assume access to an unlabeled set of images $\mathcal{X}_u^{multi}$ (now without category labels). Note that this multi-category setting is harder than the single-category since it introduces cross-category confusion, but it also allows the model to learn category-agnostic shape information across different categories.

2.3. Few-shot transfer learning

Collecting a large dataset that can cover all categories is infeasible. Therefore, we also need a way to adapt a pre-trained model to a new category. This strategy can also be used for adapting a *category-agnostic* model to a specific category. We assume that for this adaptation, a dataset $\mathcal{X}_l^{(new)}$ containing a very small number of images with pose and instance annotations (< 100) are available for the category of interest. We also assume that the semi-supervised multi-category dataset described above is available as a pre-training dataset: $\mathcal{X}_l^{pre} = \mathcal{X}_l^{multi}$ and $\mathcal{X}_u^{pre} = \mathcal{X}_u^{multi}$.

3. A Unified Framework

We need a model and a training framework that can utilize both images with pose and instance annotations, and

images without any labels. To utilize these data, we propose a unified model architecture with an encoder E, a generator G, and a discriminator D as illustrated in Figure 3. In addition, we make use of a perspective "projector" module P that takes a voxel and a viewpoint as input, and it renders the voxel from the inputted viewpoint. The training process alternates between an iteration on images labeled with pose and instance, and an iteration on unlabeled images.

3.1. Training on pose-annotated images

The encoder is provided with pairs of images x_{i1}, x_{i2} of the same 3D object i taken from different camera poses $\mathbf{p}_1$ and $\mathbf{p}_2$, and embeds each image into latent vectors $\mathbf{z}_1$, $\mathbf{z}_2$. The generator (decoder) G need to predict the 3D voxel grid from $\mathbf{z}_1$ and $\mathbf{z}_2$. The 3D voxel grid produced by the generator should be: 1) a good reconstruction of the object and 2) invariant to the pose of the input image [27]. This requires that the latent shape representation also be invariant to the camera pose of the input image. With these intuitions in mind, we explore the following three losses.

Reconstruction loss: Let $(\mathbf{x}_1, \mathbf{p}_1)$ and $(\mathbf{x}_2, \mathbf{p}_2)$ be two pairs of image-pose pair sampled from a 3D-model, then the voxel reconstructed from $E(\mathbf{x}_1)$ should produce the same image as $\mathbf{x}_2$ if projected from camera pose $\mathbf{p}_2$. Same for the other view. Let $P(\mathbf{v}, \mathbf{p})$ represent the image generated by projecting voxel $\mathbf{v}$ using camera pose $\mathbf{p}$. We define the reconstruction loss to address this requirement as:

$$\begin{aligned}\mathcal{L}_{recon} =& \|P(G(E(\mathbf{x}_2)), \mathbf{p}_1) - \mathbf{x}_1\|_{1+2} + \\ & \|P(G(E(\mathbf{x}_1)), \mathbf{p}_2) - \mathbf{x}_2\|_{1+2}\end{aligned} \tag{1}$$

where $\|\cdot\|_{1+2} = \|\cdot\|_1 + \|\cdot\|_2$ is the summation of ℓ_1 and ℓ_2 reconstruction losses.

Pose-invariance loss on representations: Given two randomly sampled views of an object, the encoder E should be able to embed their latent representations close by, irrespective of pose. Therefore, we define a *pose-invariance* loss on the latent representations:

$$\mathcal{L}_{pinv} = \|E(\mathbf{x}_1) - E(\mathbf{x}_2)\|_2 \tag{2}$$

Pose-invariance loss on voxels: Similarly, the 3D voxel output reconstructed by the generator G from two different views of the same object should be the same. Thus, we introduce a *voxel-based* pose invariance loss:

$$\mathcal{L}_{vinv} = \|G(E(\mathbf{x}_1)) - G(E(\mathbf{x}_2))\|_1 \tag{3}$$

Losses are illustrated by the dashed lines in Fig. 3. Each training step on the images with pose annotations tries to minimize the combined supervised loss.

3.2. Training on unlabeled images

In order to learn from unlabeled images, we use an adversarial loss, as illustrated in the bottom of Fig. 3. When projected from a random viewpoint, the 3D voxel grid generated from G should be able to produce an image that is indistinguishable from a real image. Specifically, we first sample a vector $\mathbf{z} \sim \mathcal{N}(0, I)$ and a viewpoint $\mathbf{p}$ uniformly sampled from the range of camera poses observed in the training set. Then the generator G will take the latent vector $\mathbf{z}$ and reconstruct a 3D shape. This 3D shape will be projected to an image using the random pose $\mathbf{p}$. No matter which camera pose we project, the projected image should look like an image sampled from the dataset.

4. Experiments

4.1. Dataset and evaluation metrics

We use voxelized 3D shapes from the ShapeNetCore [4] dataset. We first rotate the voxelized 3D model around its center using a rotation vector $\mathbf{r} = [r_x, r_y, 0]$ uniformly sampled from a fixed range. We then project the rotated 3D voxel into a binary mask as the image where the rotation vector $\mathbf{r}$ is the camera pose. A model is trained with $r\%$ of pose supervision if $r\%$ of model instances are annotated with poses. All training images are used as unlabeled images. We evaluate the moels using Intersection-over-Union (IoU) with threshold and Average Precision (AP).

4.2. Semi-supervised single-category

In this setting, we train a separate model for each category. We experiment with varying amounts of pose supervision from 0% to 100%.

Comparison with prior work: We first compare with prior work that uses full pose/instance supervision. We train our models with 50% of the images annotated with instance and pose. Performance comparisons are shown in Table 1. The performance of our model is comparable with prior work across multiple metrics. However, note that due to differences in the setting across different approaches, the numbers are not exactly commensurate.

Are unlabeled images useful? Is using unlabeled images and an adversarial loss to provide additional supervision and regularization useful? We compare three models on the `chair` category: 1) a model trained with both pose-annotated and unlabeled images; 2) a model trained on just the pose-annotated images; and 3) a model trained on only the unlabeled images. Fig. 4 shows that pose supervision is necessary as 1% of pose supervision significantly increases the performance over the unsupervised model. Second, the model that combines pose annotations with unlabeled images outperforms the one that uses only pose-annotated images. The lesser the pose annotation available, the larger the gain. When there are enough images with pose annotations (e.g. $> 50\%$), leveraging unlabeled data is unnecessary.

Table 1. Comparison between our model and prior work on single-view 3D reconstruction (single category).

Category	MVC [20]	McRecon [10]		PTN [27]	Ours (50% pose annotations)			
	IoU	AP	$IoU_{0.4}$	$IoU_{0.5}$	IoU	AP	$IoU_{0.4}$	$IoU_{0.5}$
airplanes	0.55	0.59	0.37	-	**0.57**	**0.75**	**0.56**	0.57
cars	0.75	0.82	0.56	-	**0.78**	**0.92**	**0.77**	0.77
chairs	0.42	0.48	0.35	**0.49**	**0.44**	**0.60**	**0.43**	0.42

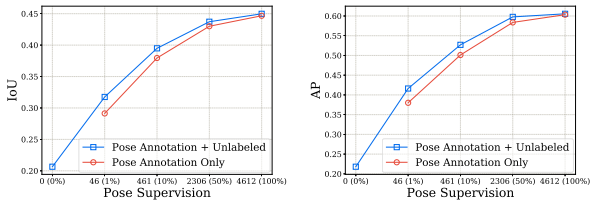

Figure 4. Comparison between three variations of our models.

Table 2. Performance of category-agnostic models.

Test categories	Single		Multi	
	IoU	AP	IoU	AP
airplanes	0.57	0.75	0.57	0.73
cars	0.78	0.92	0.78	0.93
chairs	0.44	0.60	0.44	0.57

4.3. Semi-supervised multi-category

We next experiment with a *category-agnostic* model on combined training data from 3 categories. Results are reported in Table 2. With the same amount of pose supervision (50%) for each category, the category-agnostic model achieves similar performance compared with the category-specific models. This suggests that the model is able to remedy the removal of category information by learning a category-agnostic representation.

4.4. Few-shot transfer learning

To evaluate whether the model can *transfer* the knowledge and adapt it to a new class with very limited annotated training data, we use the category-agnostic model, pre-trained on the dataset described in Sec 4.3, and adapt it to three *unseen* categories: `benches`, `vessels`, and `carbinets`. For each of the novel categories, only 1% of the pose-annotated data is provided (i.e. each novel category contains about 13 3D-shapes).

We compare three models in this experiment. **From scratch**: a model trained from scratch on the given novel category without using any pre-training; **Out-of-Category** [27]: the pre-trained category-agnostic model directly applied on the novel classes without any additional training; and **Fine-tuning**: a pre-trained category-agnostic model fine-tuned on the given novel category. The fine-tuning is done by fixing the encoder and training the generator only using pose-annotated images for a few iterations. In this experiment, we varies the amount of pose annotations used for pre-traning. The results are shown in Fig. 5. We observe that fine-tuning a pre-trained model for a novel category performs much better than training from scratch without pre-training. Compared with the out-of-category baseline, fine-tuning improves the performance a lot upon directly using the pre-trained model, especially in the case of limited pose supervision.

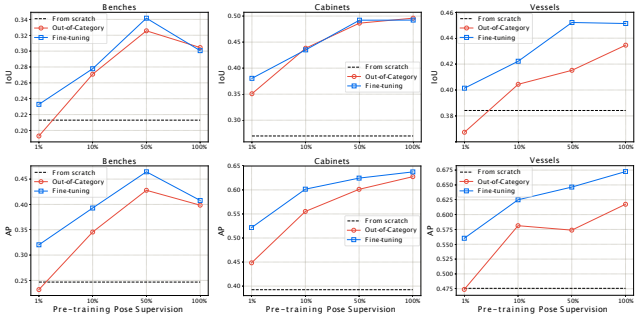

Figure 5. Few-shot transfer learning on novel categories. Each column represents the performance on a novel category (IoU in top row and AP in bottom row). Notice that the horizontal axis shows the amount of pose annotated supervision in *pre-training*.

Table 3. Different training strategies on `chairs` with 1% pose.

Model	IoU	AP
S, P	0.2913	0.3800
S, U	0.2065	0.2180
S, P+U	0.3175	0.4162
M	0.3104	0.3859
FT	**0.3250**	**0.4247**

4.5. How best to use limited annotation?

We now have all the ingredients necessary to answer the question: given a very small number of pose annotations, what is the best way to train a single-view 3D reconstruction model? Table 3 compares multiple training strategies on `chairs`: using just the pose-annotated images of `chairs` (**S, P**), using just unlabeled images of `chairs` (**S, U**), using both pose-annotated and unlabeled images of `chairs` (**S, P+U**), combining multiple categories to train a category-agnostic model (**M**), and fine-tuning a category-agnostic model for `chairs` (**FT**). The fine-tuned model works best.

References

[1] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM (2011)

[2] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[3] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co. (1999)

[4] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015) 1, 3

[5] Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: ECCV (2016)

[6] Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: CVPR. vol. 2, p. 6 (2017)

[7] Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: 3DV (2017)

[8] Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: ECCV (2016)

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)

[10] Gwak, J., Choy, C.B., Garg, A., Chandraker, M., Savarese, S.: Weakly supervised generative adversarial networks for 3d reconstruction. In: 3DV (2017) 4

[11] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)

[12] Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: CVPR (2015)

[13] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

[14] Kodali, N., Hays, J., Abernethy, J., Kira, Z.: On convergence and stability of gans (2018)

[15] Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)

[16] Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)

[17] Rezende, D.J., Eslami, S.A., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3d structure from images. In: NIPS (2016)

[18] Rock, J., Gupta, T., Thorsen, J., Gwak, J., Shin, D., Hoiem, D.: Completing 3d object shape from one depth image. In: CVPR (2015)

[19] Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. PAMI (2009)

[20] Tulsiani, S., Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In: CVPR (2018) 4

[21] Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: CVPR (2017) 1

[22] Ullman, S.: The interpretation of structure from motion. In: Proc. R. Soc. Lond. B. The Royal Society (1979)

[23] Vicente, S., Carreira, J., Agapito, L., Batista, J.: Reconstructing pascal voc. In: CVPR (2014)

[24] Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: ECCV (2016)

[25] Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: NIPS (2016)

[26] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR (2015)

[27] Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: NIPS (2016) 1, 3, 4