3D Object Structure Recovery via Semi-supervised Learning on Videos

Qian He
heqian@shanghaitech.edu.cn
Desen Zhou
zhouds@shanghaitech.edu.cn
Xuming He
hexm@shanghaitech.edu.cn

School of Information Science and Technology
ShanghaiTech University
Shanghai, China

Abstract

This paper addresses the problem of joint 3D object structure and camera pose estimation from a single RGB image. Existing approaches typically rely on both images with 2D keypoint annotations and 3D synthetic data to learn a deep network model due to difficulty in obtaining 3D annotations. However, the domain gap between the synthetic and image data usually leads to a 3D object interpretation model sensitive to the viewing angle, occlusion and background clutter in real images. In this work, we propose a semi-supervised learning strategy to build a robust 3D object interpreter, which exploits rich object videos for better generalization under large pose variations and noisy 2D keypoint estimation. The core design of our learning algorithm is a new loss function that enforces the temporal consistency constraint in the 3D predictions on videos. The experiment evaluation on the IKEA, PASCAL3D+ and our object video dataset shows that our approach achieves the state-of-the-art performance in structure and pose estimation.

1 Introduction

Estimating 3D object structure and/or pose from 2D images is a fundamental problem in computer vision and of great importance in a broad range of applications, such as autonomous driving [2] and object manipulation with robots [34]. Despite the success of 3D object reconstruction through multi-view geometry [1] or jointly with pose estimation in the SLAM pipelines [9], it remains challenging to recover object structure and/or camera pose from a single RGB image [30, 33]. Without relying on having correspondence between multi-view images, single-image 3D object structure recovery is usually less restrictive and can be applied in broader scenarios than the geometry-based approaches.

Inspired by recent breakthrough in many computer vision problems (e.g., [13, 20]), deep networks have been explored to infer 3D properties of objects from a single input image, including dense depth map [10] or 3D mesh representation [7], 3D landmark-based structure [15] or camera pose [5], and joint estimation of structure and pose [30]. One main challenge in single-image 3D structure estimation is the lack of image data with ground-truth 3D annotations as collecting such annotations is expensive and time-consuming.

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

To address this problem, Wu *et al.* proposed a two-stage deep network that first estimates 2D keypoints of an object and then lift them into a 3D representation [30]. The network is trained using real images with 2D annotations for its keypoint estimation module and 3D synthetic data for the 3D interpretation module. To bridge the domain gap between the real and synthetic data, the entire network is fine-tuned to minimize the re-projection errors on the real images. Such synthetic-data augmentation approaches, however, suffer from two limitations: First, with limited 2D annotated data, it is challenging to cover the entire pose space of an object category, and hence the fine-tuned network has difficulty in dealing with large pose variations. In addition, the 3D network module is sensitive to the quality of 2D keypoint estimation, and its performance deteriorates significantly with adverse viewing condition such as occlusion.

In this paper, we tackle the problem of joint 3D object structure and camera pose estimation, aiming to improve the robustness and generalization of deep network-based interpretation models. To this end, we propose to utilize rich video data of objects to enhance the learning of a 3D object interpretation network. Using video sequences provides two advantages in network learning: First, object poses and viewing angles have larger variations in many video data, which can enrich the image-based training set. Second, we are able to exploit 3D and temporal consistency constraint to learn from videos with minimal labeling effort. An example of our 3D object interpretation pipeline is shown in Figure 1.

Specifically, we build a robust 3D object interpreter sharing the same overall structure as the 3D-INN [30]. Unlike the 3D-INN, we adopt the stacked hourglass network [22] as the 2D keypoint estimator for its better localization performance. To train the network, we develop a semi-supervised learning strategy that utilizes both the original training data as in [30] and a weakly labeled video dataset, in which only the object category of each video is annotated. Our training procedure starts from a pre-trained 2D keypoint estimator and 3D interpretation module (based on the original data) and refines the entire model on both 2D annotated images and weakly labeled videos. To achieve this, we design a hybrid learning loss function consisting of a re-projection loss term defined on the 2D image set and a 3D structure loss term defined on video sequences. Our 3D structure loss encourages the smooth transitions between neighboring frames and consistent 3D model estimation. Moreover, we explore a simple curriculum learning method that starts from short video sequences and gradually increases the video length, which further improves the model performance.

We evaluate our method on the challenging IKEA, PASCAL3D+ dataset and our object video sequence dataset. Our approach achieves superior performance over the state of the art [30] while our ablative study also demonstrates the effectiveness of each module in our method design. The main contributions are summarized as follows:

- We propose a semi-supervised learning method that utilizes video information to boost 3D structure and camera pose estimation. We design a novel loss function on video so that the method is more robust towards different camera poses and structures.
- We develop a new object video dataset from [8] for the semi-supervised learning setting. For each sequence, we provide the object class annotation which is easy to label.
- We train a robust 3D object interpreter that achieves the state-of-the-art performance for three object classes in IKEA dataset and on the PASCAL3D+ dataset.

Our dataset and code are available at https://github.com/hailieqh/3dobj_structure.

2 Related Work

Monocular 3D Reconstruction Reconstructing 3D object from a single image is a long-standing challenge in computer vision. Unlike the multi-view reconstruction using structure-from-motion (SFM) [1] or space carving [24], single-image approaches do not rely on feature point matching or foreground segmentation and thus can be applied to broader scenarios, such as textureless objects and cluttered scenes. As it is an ill-defined problem, early works extensively make use of 3D CAD models as priors, and align them with input images to obtain 3D reconstruction [2, 3, 14, 18]. However, such methods have difficulty in handling object variations beyond the available CAD library.

Recently, inspired by the success in semantic understanding, deep networks have been widely used to solve the problem of single-image 3D reconstruction [4, 11, 29, 31]. The majority of these approaches, however, focus on recovering dense 3D volumetric shapes and do not build an abstract structure model in terms of object parts or landmarks. By contrast, our work aims to reconstruct a skeleton model of the target object class.

There have been several attempts to recover the 3D keypoints of objects from a single image [15, 27, 36], which usually first detect 2D landmarks and then fit a 3D parametric skeleton model to minimize the re-projection errors. One main challenge is that, unlike the volumetric representation, it is expensive and tedious to obtain 3D annotations of keypoints in order to learn a skeleton object model. To address this, [30] proposed 3D Interpreter Network(3D-INN), which uses synthetic data to learn a 2D-to-3D projection and fine-tune the joint model of 2D keypoint estimation and 3D reconstruction in an end-to-end manner. Our work is built upon the 3D-INN and recent 2D landmark estimator [22] but exploits videos to reduce the domain gap due to training using synthetic data.

Camera Pose Estimation There is a large body of work on camera pose estimation from a single image, which can be roughly categorized into two groups: one is to directly regress camera pose via single image [6, 21], and the other is first estimate keypoints of a object then estimate pose from 2D keypoints [30, 37]. Several researchers [30, 33] also propose deep networks to jointly optimize 3D object structure and camera pose.

Weakly Supervised Learning with Videos In many vision tasks in which the annotations are scarce or expensive to obtain, videos have been used to provide additional weak supervision to improve model learning, such as action recognition [17], semantic segmentation [26], and pose estimation [35]. In this work, we exploit temporal consistency of 3D model and camera motion in weakly labeled videos to learn a robust 3D object model. Note that videos are widely used in multi-view 3D object reconstruction with SFM [16] while here we focus on 3D model estimation from a single image.

3 Approach

Our goal is to jointly estimates the 3D object structure and camera pose from a single RGB image in a robust manner. To this end, we adopt a two-module deep network as in [30], which first predicts the 2D keypoints in image plane and then maps the keypoints into its 3D shape representation. In order to build a robust model, we develop a semi-supervised learning strategy to exploit a rich set of video sequences. Below we first describe our model design in Sec. 3.1, followed by the semi-supervised learning strategy in Sec. 3.2.

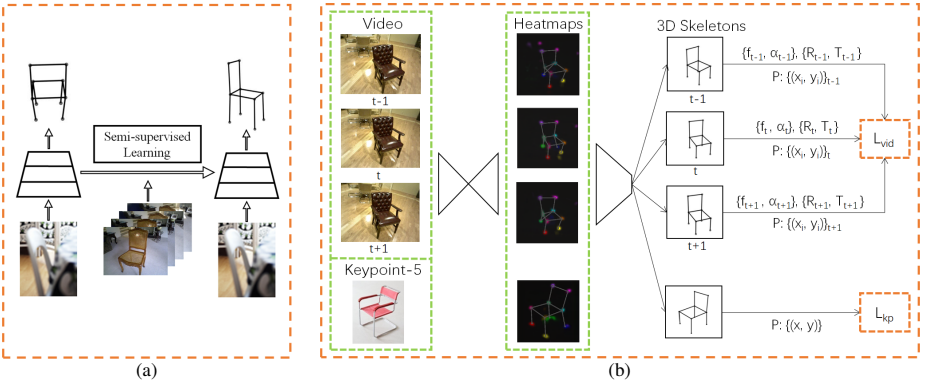

Figure 1: (a) **An overview of our approach to 3D object structure recovery.** After training with weakly annotated video data, our method generates better 3D interpretation. (b) **Overview of our semi-supervised learning for the 3D object interpreter network.** We combine Keypoint-5 data and video data within a batch to be the input of our network. The figure shows how we calculate the loss.

3.1 Model Architecture

Given an input image I of an object category, we aim to predict its 3D skeleton defined by a set of keypoints and the relative camera pose w.r.t. a canonical coordinate system of the 3D object. Formally, we denote the keypoints as $Y = \{y_1, \cdots, y_N\} \in \mathbb{R}^{3\times N}$, and parameterize the camera pose by its rotation $R \in \mathbb{R}^{3\times 3}$, translation $\mathbf{t} \in \mathbb{R}^3$, and focal length f. In order to encode shape prior, we adopt the skeleton representation using a linear combination of predefined base shapes $S_k \in \mathbb{R}^{3\times N}$ so that $Y = \sum_{k=1}^{K} \alpha_k S_k$, where $\{\alpha_k\}$ are the weight parameters and K is the number of base shapes. Let $\alpha = \{\alpha_k\}_{k=1}^{K}$, the 3D object interpretation problem can then be formulated as estimating $S = \{\alpha, R, \mathbf{t}, f\}$ from the image I.

We employ a similar model design as the 3D-INN [30], and build a deep neural network with two modules. The first module predicts 2D keypoint heatmaps from the image and then the second lifts the 2D prediction to 3D by estimating the parametric representation of the 3D object shape and camera pose. For 2D keypoint estimation, we adopt an one-stack hourglass network [22] because of its compact structure and high-quality heatmaps. The hourglass module takes as the input an image of size 256×256, and uses a set of residual modules with pooling that first gradually downsample the convolution feature maps and then upsample the feature maps to a size of 64×64, on which it predicts a set of 2D keypoint heatmaps. We refer the reader to [22] for a detailed description. The second module of our network consists of four fully connected layers with widths of 2048, 512, 128, and $|S|$, respectively, where the output is a concatenation of all parameters in S. Here we represent the rotation matrix R by the *sin* and *cos* values of three Euler angles, and the network predicts those 6 values in its output, denoted by $\mathbf{r} \in \mathbb{R}^6$.

During training stage, we also add a deterministic projection layer as in [30], which allows us to finetune our network using input images with only 2D keypoint ground truth. Under center projection assumption, the 2D coordinates of the keypoints X can be written as

$$X = P(RY + T) = P(R\sum_{k=1}^{K} \alpha_k S_k + T) \tag{1}$$

where $T = \mathbf{t} \cdot \mathbf{1}^T$ ($\mathbf{1} \in \mathbb{R}^N$), and P is the projection matrix which only depends on f.

Network Pre-training We pre-train the two modules of our network separately as in [30]. For the 2D hourglass network, we train it on a dataset of real images with 2D keypoint annotations. The 2D-to-3D module is trained on keypoint maps generated by 40,000 synthetic objects with 3D ground-truth shapes (including both structure and view). The synthetic objects are generated by randomly sampling the structural parameters α and viewpoint parameters P, R and T. To obtain a baseline network, we finetune the entire network on the real image dataset to minimize the reprojected 2D keypoint errors. Empirically, we found the baseline network has already achieved better performance than the original 3D-INN (See Sec. 4.3).

3.2 Semi-supervised Learning of 3D Interpretation

While the synthetic data augmentation enables model training without 3D ground truth, the performance of the 3D object recovery is limited by the domain gap between real and synthetic data as well as availability of a large image set with 2D annotations for fine-tuning. To address these issues, we develop a semi-supervised learning method that utilizes weakly labeled object video sequences, which increase the diversity of training data with almost no additional labeling cost.

Formally, we augment the original 2D annotated image dataset D_{2d} with a set of short object video sequences D_{vid}. Each video sequence v in D_{vid} has J frames, i.e., $v = \{I_v^1, \cdots, I_v^J\}$. The core of our approach is a new loss function that integrates both 2D re-projection errors on D_{2d}, as well as temporal and multi-view consistency of 3D predictions on D_{vid}. We now introduce these two loss terms below.

3.2.1 2D Re-projection Loss

Given the 2D image dataset D_{2d}, we denote the 2D annotations as X^{gd}. The re-projection loss is defined by the squared L^2 distance between the 2D projection of the estimated 3D object keypoints and the 2D annotations,

$$L_{kp}(\Omega) = \sum_{i=1}^{|D_{2d}|} \|P_i(R_i \sum_{k=1}^{K} \alpha_{i,k} S_k + T_i) - X_i^{gd}\|_2^2 \tag{2}$$

where Ω denotes the parameters of our network model (omitted in the right-hand side for clarity), i is the index of the training images in D_{2d}, and $\|\cdot\|_2$ is the L^2 norm.

3.2.2 Video Consistency Loss

For each video sequence, we assume all frames share the same underlying object structure α^c and camera focal length f^c, which is reasonable for rigid objects and short videos. Our video consistency loss enforces a temporally coherent 3D interpretation for each sequence, and consists of three terms, including a structure consistency loss, a motion smoothness loss and a 2D projection loss.

The structure consistency loss encodes the constraint that the 3D object structure is stable within each video due to its rigid property. In each video, we treat the underlying object structure and focal length as latent variables, denoted as $\mathbf{h} = \{\alpha^c, f^c\}$. We then minimize the differences between the structure estimation at each frame and the latent variables over

the entire video,

$$L_{st}(\Omega, \mathbf{h}) = w_\alpha \sum_{j=1}^{J} \|\alpha^j - \alpha^c\|_1 + w_f \sum_{j=1}^{J} \|f^j - f^c\|_1 \tag{3}$$

where w_α and w_f are the weights for object structure and camera focal length, respectively, and $\|\cdot\|_1$ is the L^1 norm.

The motion smoothness loss assumes the camera pose changes slowly within each video sequence, and we minimizes the changes between neighboring frames,

$$L_{ms}(\Omega) = w_R \sum_{j=1}^{J-1} \|\mathbf{r}^j - \mathbf{r}^{j+1}\|_1 + w_T \sum_{j=1}^{J-1} \|\mathbf{t}^j - \mathbf{t}^{j+1}\|_1 \tag{4}$$

where w_R and w_T are the weights for rotation and translation respectively.

In addition to the 3D loss terms, we also include a 2D re-projection loss that enforces the projection of 3D keypoints to be consistent with the 2D heatmaps predicted by the hourglass module. This 2D loss term is based on empirical observation that the estimated 2D heatmaps are usually more reliable and can be used to regularize the 3D estimation. Specifically, we first pass the predicted heatmaps through a softmax layer to get a confidence map at each pixel location. For the frame I^j, we denote the confidence values in the location of our 2D keypoint prediction as $\{c_{j,n}\}_{n=1}^{N}$, and define the 2D loss as

$$L_{rp}(\Omega) = w_{2D} \sum_{j=1}^{J} \sum_{n=1}^{N} c_{j,n} \|\mathbf{x}_{j,n} - \mathbf{x}_{j,h_n}\|_2^2 \tag{5}$$

where w_{2D} is the weight parameter, $\mathbf{x}_{j,n}$ is the re-projected 2D locations of the n-th keypoint (based on Eqn 1) and $\mathbf{x}_{j,h_n}$ is the corresponding 2D heatmap peak location.

Our overall loss function for the semi-supervised learning is defined by weighted average of video consistency loss defined on the video set and the 2D Re-projection loss defined on the annotated 2D image set. Let $H = \{\mathbf{h}^v | v \in D_{vid}\}$ be the latent variables of all sequences,

$$L_{full}(\Omega) = w_{kp} * L_{kp}(\Omega) + \min_H L_{vid}(\Omega, H) \tag{6}$$

$$L_{vid}(\Omega, H) = \sum_{v \in D_{vid}} \left(L_{st}^v(\Omega, \mathbf{h}^v) + L_{ms}^v(\Omega) + L_{rp}^v(\Omega) \right) \tag{7}$$

where w_{kp} is the weight to balance the 2D and video datasets.

3.2.3 Model Learning

During the training process, we optimize the overall loss by iteratively updating Ω and H in the loss function. Specifically, we first fix an initial Ω pretrained with L_{kp} only and update H. In our case, we solve the L^1 loss optimization problem in Eqn 3, which amounts to updating α^c and f^c by taking median of α^j and f^j at each dimension, respectively [1]. Then we update Ω with respect to fixed $\mathbf{h}$ by computing the gradients.

Instead of using a fixed length of video during training, we explore a curriculum learning strategy in our semi-supervised learning with videos. Specifically, we generate a series of training sets using video sequences with gradually increasing lengths, and train the model in multi-steps starting from short sequences. Concretely, we build training sets of videos with length 3, 7, 11 and 15 respectively. We first train on the dataset with 3 video frames and

[1] Empirically, we found that we can speed up the learning by relaxing Eqn 3 to an L^2 loss optimization problem and updating H by taking average of network outputs across video frames in the initial stage of training.

Figure 2: **Examples of Keypoint-5 and IKEA.** Note that these two datasets have very different background and viewing angles.

save the best model during training. We then retrain this model on data batches with 7 video frames, and gradually step forward to training on the datasets with 15 video frames.

4 Experiment

We evaluate our method quantitatively on two challenging 2D image dataset, IKEA [19] and PASCAL3D+ [32]. Additionally, we conduct a set of ablation study to analyze the components of our model and learning strategy.

We use three datasets for training our 3D object interpretation network: in addition to the 2D Keypoint-5 and synthetic datasets from [30], we build a new weakly labeled object video dataset from the ObjectScan [8] database.

4.1 Datasets and Metrics

Keypoint-5: We pre-train and validate our hourglass module on Keypoint-5 for 2D keypoint estimation. Keypoint-5 is a relatively small dataset which contains 5 categories, including *bed*, *chair*, *sofa*, *swivel chair*, and *table*. Each category has 1,000 to 2,000 images, of which 80% are for training and 20% are for validation. We use the median of annotations as the ground truth. In the semi-supervised learning, we also include the Keypoint-5 dataset for computing the 2D re-projection loss in Eqn 2.

Synthetic data: For training the 2D-to-3D module, we generate 40,000 objects for training and 1,000 for validation. Given a specified range for each parameter of our 3D model and camera pose, we first randomly sample a model parameter from a Gaussian distribution and then project the 3D locations to get 2D keypoints. The pairs of 2d keypoints and 3D model parameters are used to pre-train the 2D-to-3D module.

Object videos: We build our weakly labeled object video dataset from a large database of object scans [8], which contains more than ten thousand 3D scans of real objects acquired by PrimeSense camera. We select a subset of RGB sequences for three categories, including *chair*, *sofa*, *table*, and divide them into training and test set. The original frame rate is 30Hz and we subsample them to 5Hz. For each training sequence, we select 36 continuous frames. In our experiments, we use 91 sequences of *chair* class, 52 of *sofa* class, and 30 of *table* class for training and 15 sequences of each class for qualitative evaluation in testing.

IKEA dataset: IKEA dataset contains four object classes: *chair*, *sofa*, *table* and *bed*, which have 195, 164, 202 and 61 images respectively. Here we focus on the class of *chair*, *sofa* and *table* due to lacking of video data for *bed* class. The annotations include structure and rotation of the object in each image. Fig 2 shows a few examples of Keypoint-5 and IKEA data. Unlike Keypoint-5, the objects in IKEA dataset can be heavily occluded.

PASCAL3D+: PASCAL3D+ dataset [32] contains 12 categories of the PASCAL VOC 2012

Method	Average recall %					
	IKEA Chair		IKEA Sofa		IKEA Table	
	Structure	Pose	Structure	Pose	Structure	Pose
Zhou-perp [36]	60.76	-	58.02	-	-	-
Su [25]	-	37.69	-	35.65	-	-
3D-INN [30]	87.84	63.46	88.03	64.65	85.71*	55.02
Li [18]	**89.9**	-	83.4	-	-	-
Ours	89.68	**71.25**	**92.66**	**71.40**	**88.52**	**56.64**

Table 1: 3D structure and pose estimation results on IKEA dataset. * is generated from the code released by the authors.

dataset and part of images from ImageNet, which are augmented with 3D annotations. In our experiments, we use the dataset to evaluate the performance for pose estimation. We report our results on the validation sets of *chair* and *sofa* of PASCAL VOC 2012, which include 642 and 336 images respectively.

Evaluation metric: For evaluating 3D structure estimation, we follow the protocol in [30], which first computes the root-mean-square error (RMSE) between predictions and ground truth, and then calculates the average recall of the keypoints using a range of minimum deviation thresholds. For pose estimation, we use the absolute error of azimuth angle predictions and also calculate the average recall by the same procedure.

Implementation Details: We first train the one-stack hourglass module on Keypoint-5 to generate 2D keypoint predictions as heatmaps, and then train our 2D-to-3D module on synthetic data to predict 3D parameters from the heatmaps. We combine these two modules and finetune the full network on Keypoint-5 to get our baseline model. For our semi-supervised learning setting, we freeze the 2D prediction module and train the rest of our model on videos and Keypoint-5 data. We use the RMSprop for optimization, a maximum of 300 epochs in training, and $8e-6$ as the learning rate. No dropout layers are applied to the 2D-to-3D module.

To determine the video length M, we validate on a set of lengths in $\{3,5,7,9,11,13,15\}$ frames. For curriculum learning, we incrementally train our model with video length of $\{3,7,11,15\}$, each of which takes 50 epochs. Additionally, we validate our model hyperparameters on Keypoint-5 by a grid search. The final weights for our loss $\{w_{kp}, w_\alpha, w_f, w_R, w_T\}$ are set to $\{1, 10000, 10000, 50, 50, 150\}$ throughout the experiments.

4.2 Results on IKEA and Video Dataset

Given the trained 3D object interpretation networks, we first conduct our evaluation on three categories, *chair*, *sofa*, and *table* in the IKEA dataset. The quantitative results are summarized in Table 1, in which we also include the reported performances of four prior methods. It is evident that our approach achieves significantly higher average recall than other approaches and consistently across all three categories.

In particular, our method outperforms the recent state of the art [30] in both structure and pose metric despite that we add only a small set of additional short videos (fewer than 100 per class) in training. This shows that the video set does effectively improve the diversity of the training data and narrows the domain gap between the original training (based on Keypoint-5+synthetic data) and the challenging test cases in the IKEA dataset.

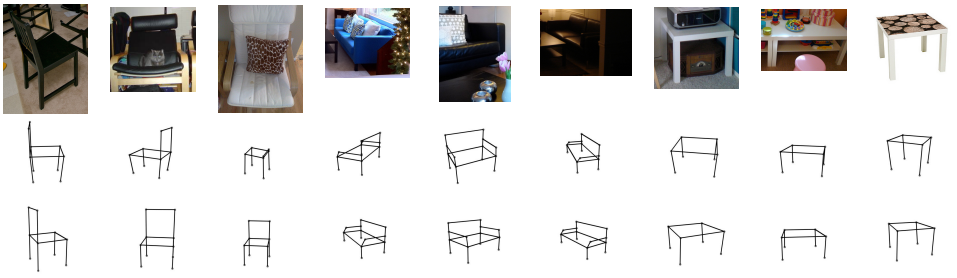

Figure 3: **Qualitative results on IKEA.** Each row: Top: input image; Middle: baseline results; Bottom: our results.

We also visually compare our results with baseline model as shown in Figure 3 for the IKEA dataset and Figure 4 for our video dataset. We can see that our method generates much better results compared to the baseline. In particular, our method perform well even under heavy occlusion (e.g. *chair* and *table*), and it can also deal with multiple instances and cluttered scenes.

We note that each frame of a video sequence has a slightly different 2D projection, self-occlusion and background caused by gradual change of the camera viewpoints. Due to lack of sufficient training data, the 2D keypoint estimation module may fail to produce consistent predictions under these variations. The keypoint errors generated by the 2D module then propagate to the 2D-to-3D module and lead to unstable structure predictions in the baseline model. Our semi-supervised learning strategy allows the 2D-to-3D module to cope with the inaccurate 2D estimations in a more robust manner and thus produces better 3D predictions.

4.3 Ablation Study

To understand the effect of each module in our approach, we conduct an ablation study to investigate their contributions towards the final performance. We consider three different variations of our method: 1) Baseline, which is our model trained with Keypoint-5 and synthetic data only in the same way as the 3D-INN [30]; 2) Video, which is our model trained with all three datasets, including the weakly labeled video sequences (of a fixed length); 3) Video+Curriculum learning, which is the full model with both semi-supervised and curriculum learning.

Table 2 shows the comparisons between three different variants of our method on the *chair* and *sofa* class. Our Baseline model has already outperformed the original 3D-INN due to its better 2D keypoint estimation module based on the hourglass. Our Video model outperforms the Baseline model consistently in both metrics and classes, which indicates that the semi-supervised learning plays an important role in improving the average recalls. Moreover, we found the curriculum learning can slightly improve the model by gradually adding more challenging training examples, which achieves the best overall performances with less overall training time (two thrids of the vanilla Video model learning). We note that our approach achieves the performance gains with a moderate-sized video dataset and expect that overall performances would increase with more training video sequences.

4.4 Results on PASCAL3D+ Dataset

To demonstrate the generality of our method, we also evaluate our trained models on the PASCAL3D+ dataset [32] for object pose estimation. Note that we only train the models

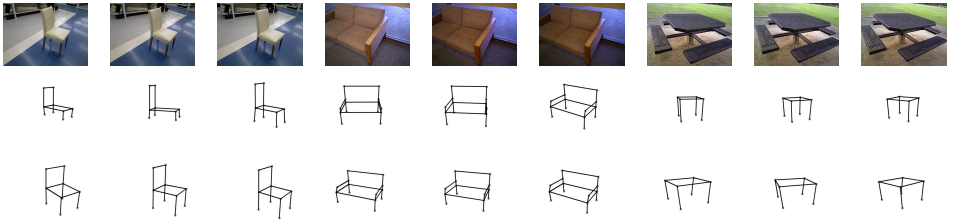

Figure 4: **Qualitative results on our video dataset.** Each row: Top: input video frames; Middle: baseline results; Bottom: our results.

Average recall %				
Method	IKEA Chair		IKEA Sofa	
	Structure	Pose	Structure	Pose
Baseline	87.93	68.14	91.39	68.67
Video (ours)	89.60	70.51	92.34	71.19
Video+Curriculum learning (ours)	**89.68**	**71.25**	**92.66**	**71.40**

Table 2: Ablation study on 3D structure and pose estimation using IKEA *chair* and *sofa* dataset. We compare the results of three variants of our models. See the text for details.

Method	VDPM [62]	DPM-VOC+VP [23]	Su[25]	V&K [28]	3D-INN [60]	Baseline	Ours
Chair	6.8	6.1	15.7	25.1	23.1	25.0	**25.8**
Sofa	5.1	11.8	18.6	43.8	45.8	46.2	**47.3**

Table 3: Pose estimation results on PASCAL3D+. Our method achieves the state-of-the-art performance.

using the Keypoint-5, synthetic and video dataset as described before and directly test them on the PASCAL3D+ dataset. We adopt the standard R-CNN [12] for object detection. As shown in Table 3, our model outperforms all other methods on both *chair* and *sofa* class. We also show the results of our base model for ablative study. Our model outperforms 0.8% on *chair* class compare to our base model, and 1.1% on *sofa* class, which demonstrates the effectiveness of our semi-supervised learning method.

5 Conclusion

We have developed a semi-supervised learning strategy to address the problem of joint 3D object structure and camera pose estimation from a single monocular image. Our method exploits weakly labeled video data, which can be obtained at relatively low cost, to improve the robustness of the resulting 3D object interpretation models. In contrast to the existing methods, our 3D object interpreters are capable of handling large variations in viewing angle, occlusion and background clutter. The experiment evaluation on the IKEA, PASCAL3D+ and our new object video dataset shows that our approach achieves the state-of-the-art performance in both structure and pose estimation.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *ICCV*, 2009.

[2] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[3] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *CVPR*, 2016.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[5] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014.

[6] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *CVPR*, 2016.

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

[8] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv:1602.02481*, 2016.

[9] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014.

[10] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[11] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *CVPR*, 2017.

[15] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015.

[16] Adam G Kirk, James F O'Brien, and David A Forsyth. Skeletal parameter estimation from optical motion capture data. In *CVPR*, 2005.

[17] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[18] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. *CVPR*, 2017.

[19] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing IKEA Objects: Fine Pose Estimation. In *ICCV*, 2013.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[21] Frank Michel, Alexander Kirillov, Erix Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation. *arXiv preprint arXiv:1612.02287*, 2016.

[22] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[23] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[24] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *IJCV*, 1999.

[25] 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 *ICCV*, 2015.

[26] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016.

[27] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR*, 2017.

[28] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*, 2015.

[29] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017.

[30] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *ECCV*, 2016.

[31] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *NIPS*, 2017.

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

[33] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016.

[34] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *ICRA*, 2017.

[35] Feng Zhou and Fernando De la Torre. Spatio-temporal matching for human pose estimation in video. *IEEE transactions on PAMI*, 38(8):1492–1504, 2016.

[36] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3d shape reconstruction from 2d landmarks: A convex formulation. In *CVPR*, 2015.

[37] M Zeeshan Zia, Michael Stark, Bernt Schiele, and Konrad Schindler. Detailed 3d representations for object recognition and modeling. *IEEE transactions on PAMI*, 35 (11):2608–2623, 2013.