Deep Supervision with Shape Concepts for Occlusion-Aware 3D Object Parsing

–Supplementary Material–

1 Introduction

In this supplementary material, Section 2 details the 3D annotation for CAD models and real images as well as our approach to compute the car yaw angle given a 3D skeleton. Next, in Section 3, we present the instance segmentation algorithm in detail for our experiments on PASCAL3D+. Finally, we provide more quantitative results on KITTI-3D in Section 4 and demonstrate more qualitative results on KITTI-3D, PASCAL VOC and IKEA datasets in Section 5.

2 Details of 3D Annotation

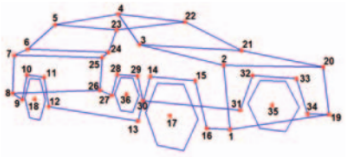
In the Section 2.1, we first provide more details about the 3D keypoint annotation for CAD models of car and furniture. Subsequently, we introduce the method to compute yaw angle of a car given the corresponding 3D skeleton. Finally, in Section 2.3, we detail our approach of 3D skeleton annotation for real images in KITTI-3D.

2.1 CAD Model Annotation

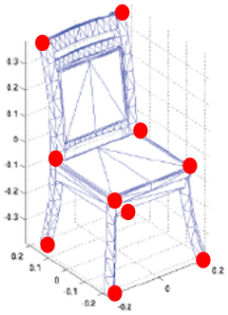
We follow Zia et al. [4] to annotate 36 keypoints for each car CAD model. The 3D skeleton of a car is shown in Figure 1a. Each keypoint represents a particular semantic part of the car such as the wheel and corners of the front/back windshield. For IKEA dataset, we annotate 14 keypoints for both chair and sofa CAD models as shown in Figure 1b and Figure 1c, respectively. Note that the definition of keypoints on the sofa seating area (shown in Figure 1c) are inconsistent with 3D-INN[3]. Additionally, the keypoints on armrests of a chair are merged to the keypoints on the seating area if armrests do not exist.

2.2 Yaw Computation

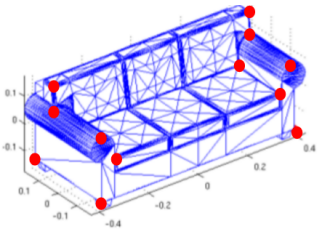
In the Table 1 of the main paper, we report the mean error of estimated yaw angles over all fully visible cars in KITTI-3D for different methods. Yaw angle is defined as the angle of car head direction with respect to X axis on the XZ plane or ground plane. Given



(a) Car Keypoints



(b) Chair Keypoints



(c) Sofa Keypoints

Figure 1: Visualization of keypoint definitions on car, chair and sofa classes. Invisible keypoints are not shown in Figure 1c.

an estimated 3D skeleton, we first project all keypoints onto the XZ plane and compute the average direction over all lines perpendicular to the left-to-right correspondence lines. Then, we can obtain the yaw angle as the angle of this average direction with respect to the X axis. For example in Figure 1a, lines (1,19),(17,35) and (5,23) are correspondence lines. Finally, we compute the absolute error between the estimated yaw direction and the ground truth yaw and average the error over all test images.

2.3 KITTI-3D

KITTI [1] dataset provides road scene images captured from mounted cameras on moving cars in real driving scenario. Object locations and viewpoints are annotated on a subset of the images in KITTI. Zia et al. [4] further select 2040 car images and label each of them with visible 2D keypoints. In the following, we introduce our method to reconstruct 3D skeleton based on ground truth visible 2D keypoints and viewpoint.

To obtain 3D ground truth, we fit a PCA model trained on the 3D keypoint annotations on CAD data, by minimizing the 2D projection error for the known 2D keypoints. First, we compute the mean shape M and 5 principal components, $P_1, \cdots, P_5$ from 472 3D skeletons of our annotated CAD models. M and P_i $(1 \leq i \leq 5)$ are 3×36 matrices where each column contains 3D coordinates of a keypoint. Next, the ground truth 3D structure X is represented as $X = M + \sum_{i=1}^{5} \alpha_i P_i$, where α_i is the weight for P_i. To avoid distorted shapes caused by large α_i, we constrain α_i to lie within $-2.7\sigma_i \leq \alpha_i \leq 2.7\sigma_i$ where σ_i is the standard deviation along the ith principal component direction. Next, we densely sample object poses $T_p = \{T_i\}$ (3x3 rotation matrices) in the neighborhood of the labeled pose. Finally, we compute the optimal 3D structure coefficients $\alpha = \{\alpha_i\}$ by minimizing its 2D

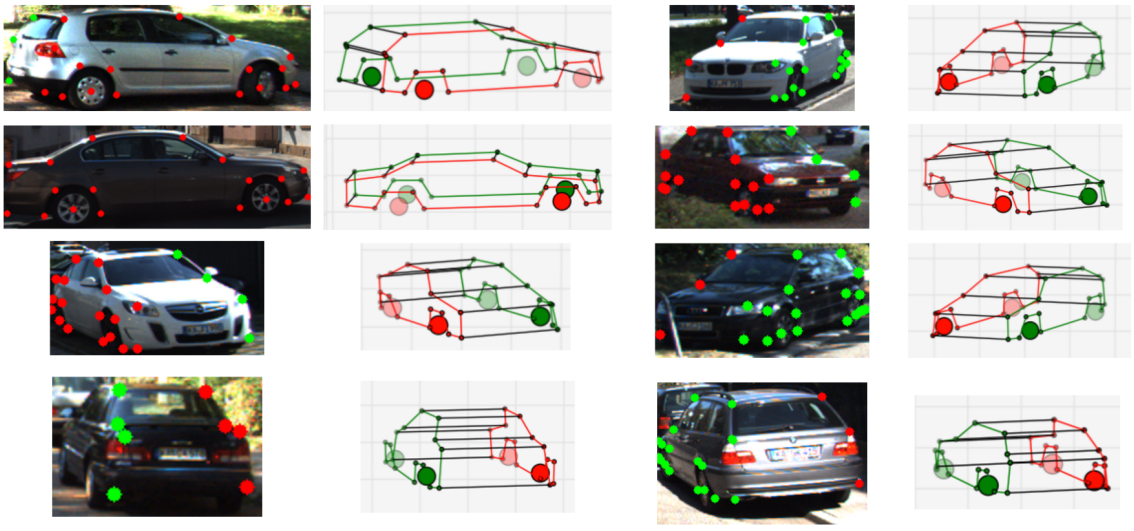


Figure 2: Examples of 2D and 3D annotations in KITTI-3D. Visible 2D keypoints annotated by Zia et al.[4] are shown on the car images. The corresponding 3D skeleton is shown to the right of each image.

projection error with respect to 2D ground truth Y:

$$\begin{aligned}\alpha^* &= \arg_\alpha \min_{T_i \in T_p, K} \|\text{Proj}(KT_i(M + \sum_{i=1}^{N} \alpha_i P_i) - Y\|_2^2 \qquad s.t. \ -2.7\sigma_i \leq \alpha_i \leq 2.7\sigma_i \\ &= \arg_\alpha \min_{T_i \in T_p, s, \beta} \|s\text{Proj}(T_i(M + \sum_{i=1}^{N} \alpha_i P_i)) + \beta - Y\|_2^2 \quad s.t. \ -2.7\sigma_i \leq \alpha_i \leq 2.7\sigma_i\end{aligned} \tag{1}$$

where K is the camera intrinsic $K = [s_x, 0, \beta_x; 0, s_y, \beta_y; 0, 0, 1]$ with the scaling $s = [s_x; s_y]$ and shifting $\beta = [\beta_x; \beta_y]$. Proj$(x)$ computes the 2D image coordinates from 2D homogeneous coordinates x.

We solve (1) by optimizing $\{\alpha_i\}, \beta, s$ given a fixed T_i and then searching for the lowest error among all sampled T_i. As we can see, more annotated points in Y are better in regularizing (1) for the better estimations of $\{\alpha_i\}, \beta, s$ given a fixed T_i. Thus, we only provide 3D keypoint labels for fully visible cars because the most of occluded or truncated cars do not contain enough visible 2D keypoints for minimizing (1). We show some examples of 3D annotation in KITTI-3D.

3 Instance Segmentation

Based on the 3D skeleton of the car category as shown in Figure 1a, we can define a rough 3D mesh using the 3D keypoints. For example, the rectangular area bounded by keypoint 1,2,20,19 forms the surface of the head of a car. Recall that our network DISCO localizes the complete set of 2D keypoints regardless of the visibility for an object. Therefore, we can compute the projected area of each predefined mesh surface and the union of all projected surfaces is the instance segmentation mask. Note that the surface of a car wheel is a hexagon where the center is the wheel keypoint and the corners are defined based on the surrounding keypoints. We exploit this segmentation mask in Section 4.3 in our main paper.

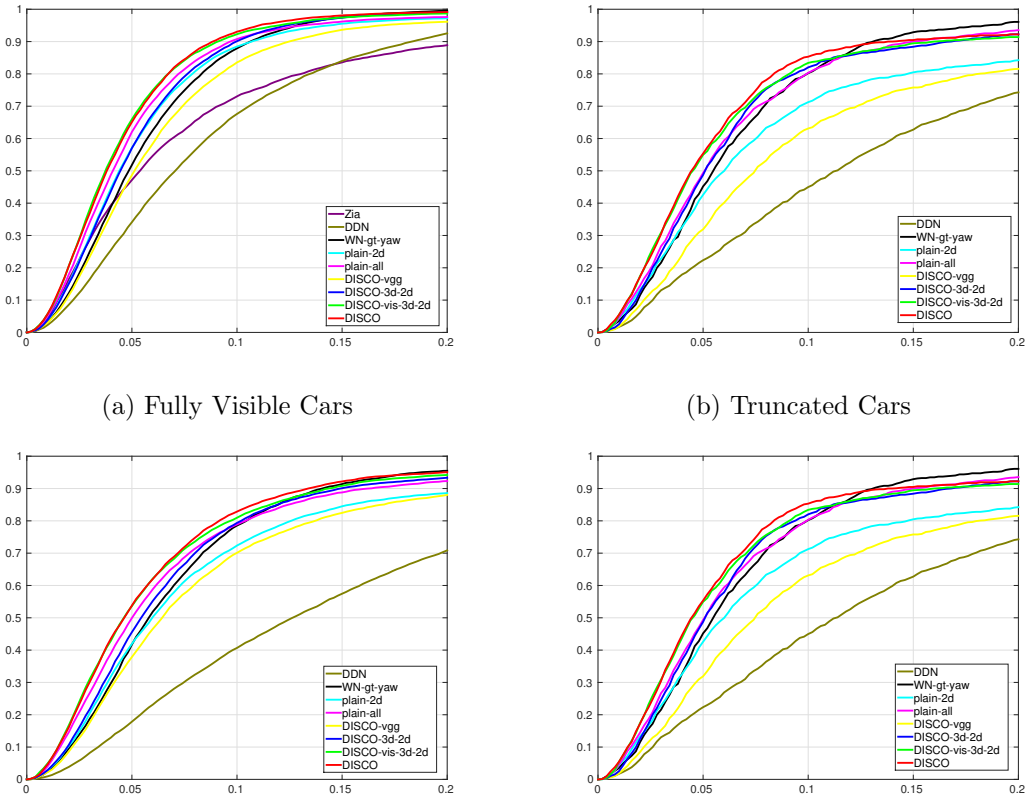
4 Quantitative Results

In this section, we demonstrate more quantitative results for KITTI-3D and IKEA dataset. Figure 3 shows the PCK curves of different methods for Table 1 in the main paper. We only plot 2D PCK curves at $\alpha \in [0, 0.2]$ because most of methods converge to 100% accuracy after $\alpha = 0.2$. We can see that DISCO outperforms other baseline methods as well as its own variants on all occlusion categories. Further, Figure 4 presents 3D PCK curves of different methods on KITTI-3D as well as 3D-INN and DISCO on the IKEA benchmark. DISCO is significantly superior to 3D-INN on IKEA chair. On IKEA sofa, DISCO is better than 3D-INN at $\alpha \in [0, 0.1]$ while being inferior at $\alpha \in [0.1, 0.25]$. Note that our keypoint annotation for the sofa class is different from [3], which directly degrades the performance of DISCO. Moreover, as mentioned in the main paper, 3D-INN uses shape bases to constrain the prediction, which enables the wrong prediction to maintain coarse object shapes.

Next, Table 1 reports the PCK at $\alpha = 0.1$ and yaw angle accuracies of DISCO and its variants on ground truth and detection bounding boxes. 3D-yaw in Table 1 measures the mean difference in yaw angle, in degrees, of the estimated 3D skeleton with respect to the ground truth. We use RCNN [2] to compute the bounding box locations for the car class. All detection bounding boxes have at least 0.7 IoU scores on the ground truth locations. Surprisingly, DISCO performs better on the detected bounding boxes over the ground truth bounding boxes. This result shows that DISCO is robust to the location noises in detection algorithms.

5 Qualitative Results

In this section, we supply more qualitative results of DISCO on KITTI-3D and PASCAL VOC as shown in Figure 5 and Figure 6, respectively. We can see that DISCO effectively localizes 2D and 3D keypoints under different types of occlusions and large variations in object appearance. Additionally, we visually compare 3D-INN and DISCO on IKEA chair



(a) Fully Visible Cars

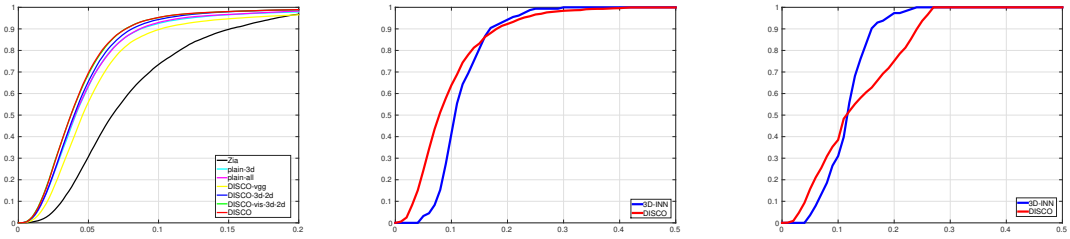
(b) Truncated Cars

(c) Multi-Car Occlusion

(d) Other Occluders

Figure 3: 2D PCK curves of different methods on fully visible cars (Figure 3a), truncated cars (Figure 3b), multi-car occlusion (Figure 3c) and other occluders (Figure 3d) in KITTI-3D. In each figure, X axis stands for α of PCK and Y axis represents the accuracy.

(Figure 7) and IKEA sofa (Figure 8). We observe that DISCO is able to capture the fine-grained 3D geometry details such as the inclined seatback or the armrest but 3D-INN fails to reliably estimate subtle geometric variations.. We attribute this to the fact that DISCO directly exploits rich object appearances for 3D structure prediction while 3D-INN only relies on detected 2D keypoints.



(a) KITTI-3D (b) IKEA Chair (c) IKEA Sofa

Figure 4: 3D PCK (RMSE[3]) curves of different methods on fully visible cars of KITTI-3D (Figure 4a), IKEA chairs (Figure 4b) and IKEA sofas (Figure 4c). In each figure, X axis stands for α of PCK and Y axis represents the accuracy.

Method	2D				3D	3D-yaw
	Full	Truncation	Multi-Car Occ	Other Occ	Full	Full
plain-2D	91.2/88.4	62.4/62.6	76.7/72.4	73.6/71.3	NA	NA
plain-3D	NA				92.2/90.6	6.0/6.5
plain-all	92.2/90.8	72.9/72.6	81.5/78.9	79.5/80.2	92.2/92.9	3.7/3.9
DISCO-3D-2D	92.8/90.1	70.4/71.3	85.0/79.4	82.6/82.0	94.0/94.3	3.2/3.1
DISCO-vis-3D-2D	94.4/92.3	75.9/75.7	85.8/81.0	86.3/83.4	95.3/95.2	2.3/2.3
DISCO-Vgg	84.9/83.5	60.6/59.4	72.0/70.1	62.3/63.1	89.3/89.7	7.0/6.8
DISCO	95.9/93.1	78.9/78.5	87.7/82.9	90.5/85.3	95.5/95.3	2.1/2.2

Table 1: 2D and 3D PCK[$\alpha = 0.1$] accuracies of different methods on KITTI-3D. In each cell, two PCK (%) separated by "/" are the results on detected bounding boxes (left) and ground truth bounding boxes (right). 3D-yaw is the mean error of yaw angle in degree.

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single Image 3D Interpreter Network. In *ECCV*, 2016.

[4] M. Z. Zia, M. Stark, and K. Schindler. Towards Scene Understanding with Detailed 3D Object Representations. *IJCV*, 2015.

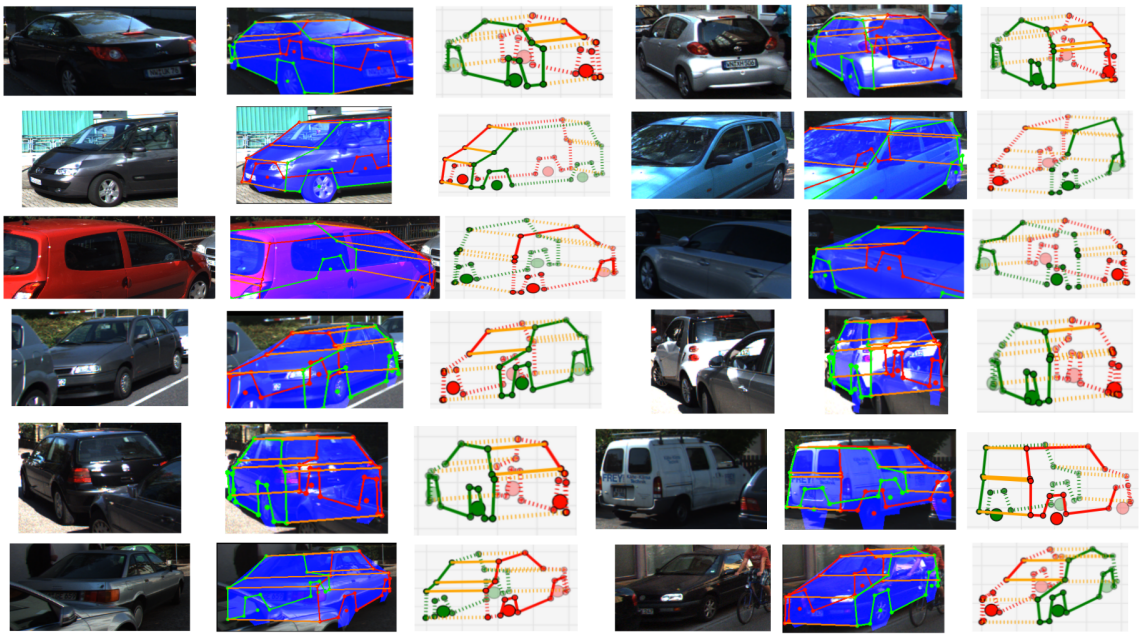


Figure 5: Visualization of correct 2D/3D prediction results by DISCO for fully visible cars (row 1), truncated cars (rows 2-3) and occluded cars (rows 4-6) in KITTI-3D.

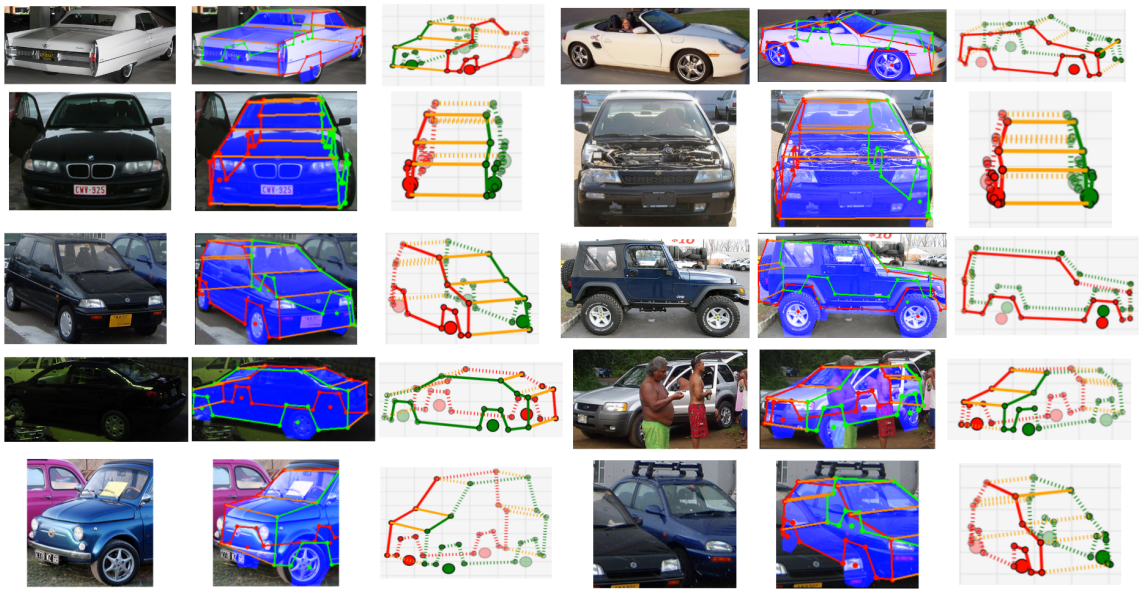


Figure 6: Visualization of correct 2D/3D prediction by DISCO for cars with large appearance variations and various occlusions in PASCAL VOC.

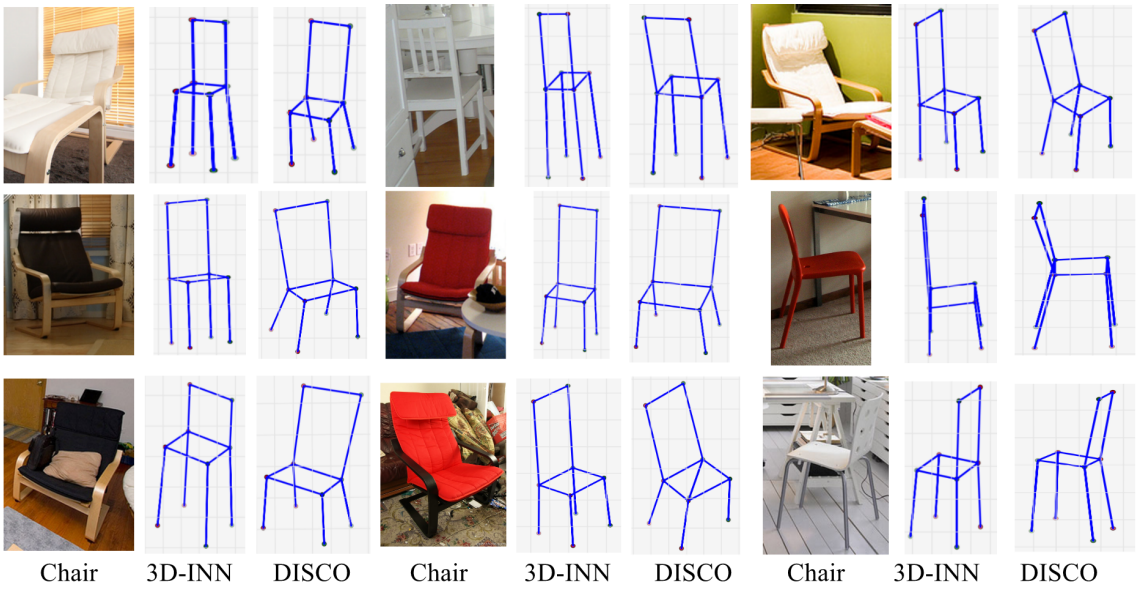


Figure 7: Qualitative comparison between 3D-INN and DISCO for 3D structure prediction on IKEA chair. DISCO is able to delineate more fine-grained 3D geometry such as inclined seatback than 3D-INN.

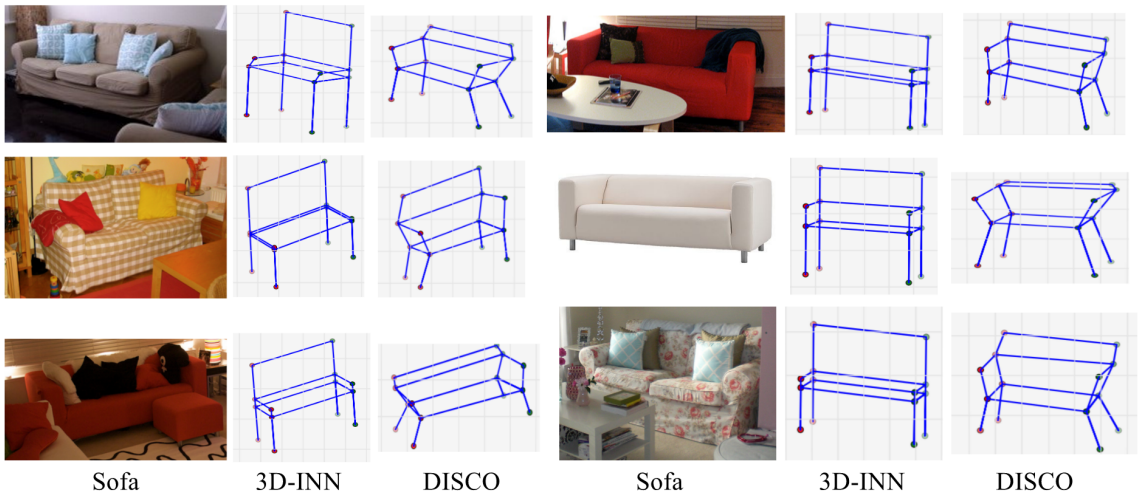


Figure 8: Qualitative comparison between 3D-INN and DISCO for 3D structure prediction on IKEA sofa. Sofa armrests can be reliably predicted by DISCO when 3D-INN fails.