Deep Single-View 3D Objet Reconstruction with Visual Hull Embedding (*Supplementary Material*)

Abstract

In this supplementary material, we provide: **1)** more discussions pertaining to the difference between our method and previous methods using pose and silhouette, **2)** our detailed network structure, **3)** more results on the PASCAL 3D+ dataset, and **4)** details of the experiment analyzing the result sensitivity with respect to focal length and visual hull distortion.

Difference from prior art using pose and silhouette

As mentioned in the main paper, our method differs from previous methods that use object pose and silhouette to supervise 3D shape prediction (Yan et al. 2016; Tulsiani et al. 2017; Gwak et al. 2017; Zhu et al. 2017; Wu et al. 2017; Tulsiani, Efros, and Malik 2018). As illustrated in Fig. 1 left, these methods project the predicted shape onto the image plane using ground-truth or estimated (Zhu et al. 2017; Tulsiani, Efros, and Malik 2018) object pose, and use the discrepancy between the projected and ground-truth (or estimated (Wu et al. 2017)) silhouettes (or other observations such as depth map (Tulsiani et al. 2017; Wu et al. 2017)) *as a loss to train or finetune* the network. In contrast, we explicitly embed *inside of the network* a single-view visual hull and use convolutions to process it for shape refinement. Additionally, our visual hull construction is actually an *inverse process* of the shape-to-image projection scheme used by these methods. But note that these two approaches are orthogonal and we can also use the projection loss to train or finetune our network.

Detailed Network Structure

Figure 2 presents a detailed structure of our 3D reconstruction network which is omitted in the main paper due to space limitation. The four sub-networks V-Net, P-Net, S-Net and R-Net are primarily based on 2D and 3D convolutional layers, where the V-Net is adapted from (Choy et al. 2016). The SPVH layer connects these sub-networks.

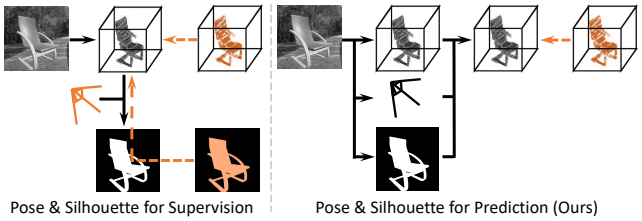

Figure 1: Difference between methods using object pose and silhouette to supervise shape learning (left) and our method (right). Orange color indicates ground truth. Solid lines represent forward computation during inference and dashed lines indicate supervision signals in the training stage.

More Results on PASCAL 3D+ Dataset

More implementation details

To train (finetune) our method on this dataset, we simply set the focal length to be 2000 for all images since no focal length is provided. With this fixed focal length, we recomputed the object distances using the image keypoint annotations and the CAD models through reprojection error minimization. We only recomputed the distances labels; other pose labels including azimuth, elevation and in-plane rotations as well as the object centers on the image were kept the same as the original dataset. We use the silhouettes obtained by projecting the pseudo ground-truth shapes to train the S-Net.

Discussion of the dataset

The PASCAL 3D+ dataset (Xiang, Mottaghi, and Savarese 2014) was originally proposed for 3D object detection. In this dataset, up to 10 CAD models are used to annotate 300 to 2,000 images in each category and serve as the pseudo ground truth shape. As a result, some images are associated with inaccurate 3D structures, as shown in Fig. 3. The silhouette and pose labels are consequently also noisy, making it very challenging for our method.

Failure cases

Some failure cases of our method are shown in Fig. 4. Note that "failure case" refers to the examples where the IoU metrics were lowered after the refinement process. Figure 4

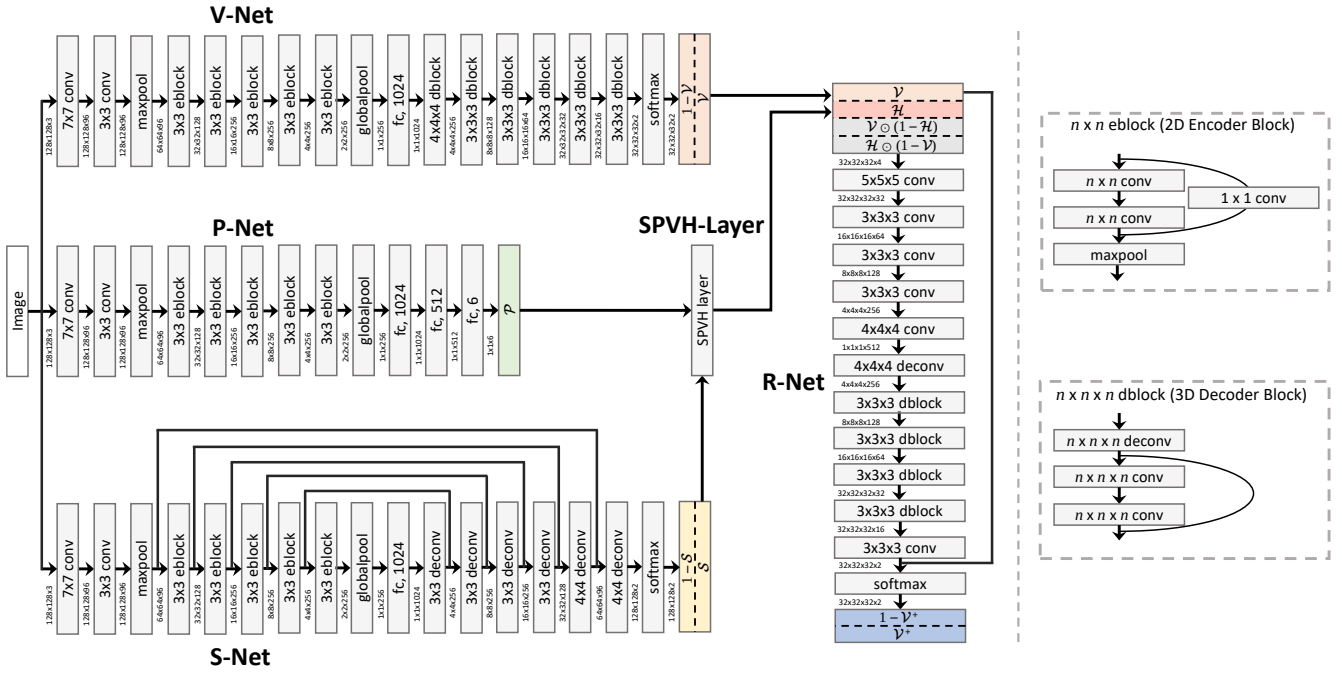

Figure 2: Detailed network structure. The rectangles represent the network layers (conv, deconv/fractionally-strided-conv, fc, *etc*.), layer blocks (right column) as well as inputs and outputs. The numbers between the layers indicate feature map sizes. All max-pooling layers decrease spacial resolution of feature map by factor of 2. Other spatial resolution changes are due to conv and deconv layers with strides larger than 1 (*e.g*., stride 2 for shrinking/enlarging resolution by 2 in conv/deconv layers). The whole network runs at 55 fps on an NVIDIA M40 GPU (∼18 milliseconds per image).

Figure 3: Randomly selected images that are annotated in the PASCAL 3D+ dataset. The images are superposed with the silhouettes obtained by projecting the pseudo ground-truth shapes. In this dataset, each category of the four only contains up to 10 CAD models used to annotate 300 to 2,000 images. As a consequence, some images are apparently associated with inaccurate 3D structures (see the last few columns). Moreover, the pseudo ground-truth silhouettes and poses we used to train our S-Net and P-Net are also noisy for many images, making it very challenging for our method. (**Best with on screen with zoom-in**)

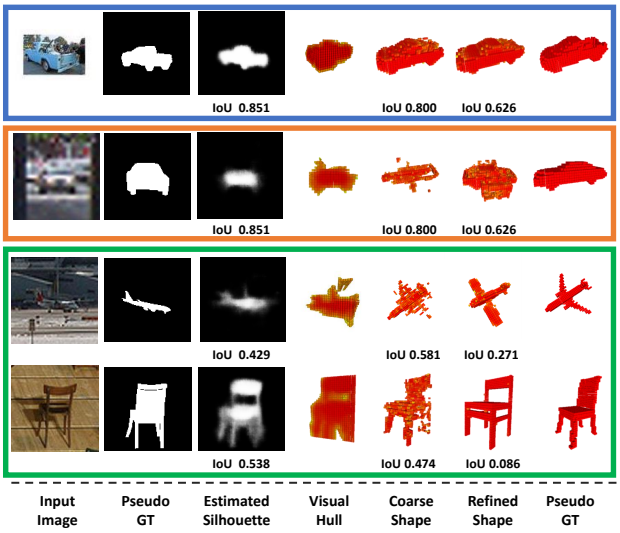

Figure 4: Failure cases on the PASCAL 3D+ dataset where the refinement led to lower IoU. **Blue box:** failure due to erroneous pose estimation (the rotation error is over 70 degrees for the car). **Orange box:** failure due to inaccurate silhouette estimates. **Green box:** due to the inconsistency between the pseudo ground-truth shapes and input images, our refined results, despite appearing much more realistic and consistent with the input images, have larger IoU values. As such, *these examples are not true failure cases for our method*. **(Best view on screen with zoom-in)**

shows that when the pseudo ground-truth shapes are clearly inconsistent with the input images (last two examples), the refined shapes have larger IoU values despite being much more realistic and consistent with the input images compared to the coarse shapes. As such, these examples are not true failure cases for our method, and they demonstrate the limitation of using this dataset for evaluating 3D reconstruction results. Apart from this dataset issue, our visual hull based refinement typically fails due to inaccurate pose and silhouette estimates, as shown in the first two examples of Figure 4.

Sensitivity w.r.t. Focal Length

As mentioned in the main paper, we designed an experiment to test our method under wrong focal lengths with distorted visual hulls. Specifically, we multiply the used focal length as well as the output t_Z (i.e., the depth of object center) from P-Net by a scale factor. This way, the constructed visual hulls are subject to certain degrees of weak-perspective distortion. We directly re-run the network without finetuning. Table 1 shows that for ShapeNet objects the focal length changes only lead to minor IoU drops and the results are still much better than the baseline. For real images, there is even no obvious performance drop with larger focal lengths. These results indicate that *our method still works well with some weak-perspective approximations* and the results are insensitive to the real focal length of the input images especially for distant objects (such as those in Pascal 3D+).

Table 1: Results with different focal lengths

		Mean IoU
ShapeNet	Before Refine.	0.631
	After Refine.	0.680
	0.8× focal & t_Z	0.675
	1.2× focal & t_Z	0.676
	1.5× focal & t_Z	0.672
Pascal 3D+	Before Refine.	0.5518
	After Refine.	0.5872
	0.5× focal & t_Z	0.5866
	1.5× focal & t_Z	0.5872
	2.0× focal & t_Z	0.5873

References

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 628–644.

Gwak, J.; Choy, C. B.; Garg, A.; Chandraker, M.; and Savarese, S. 2017. Weakly supervised generative adversarial networks for 3D reconstruction. *arXiv:1705.10904*.

Tulsiani, S.; Zhou, T.; Efros, A. A.; and Malik, J. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2626–2633.

Tulsiani, S.; Efros, A. A.; and Malik, J. 2018. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, 2897–2905.

Wu, J.; Wang, Y.; Xue, T.; Sun, X.; Freeman, W. T.; and Tenenbaum, J. B. 2017. MarrNet: 3d shape reconstruction via 2.5D sketches. In *Advances In Neural Information Processing Systems (NIPS)*, 540–550.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 75–82.

Yan, X.; Yang, J.; Yumer, E.; Guo, Y.; and Lee, H. 2016. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In *Advances in Neural Information Processing Systems (NIPS)*, 1696–1704.

Zhu, R.; Galoogahi, H. K.; Wang, C.; and Lucey, S. 2017. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *ICCV*, 57–64.