Transformation-Grounded Image Generation Network for Novel 3D View Synthesis – Supplementary Material

Eunbyung Park[1] Jimei Yang[2] Ersin Yumer[2] Duygu Ceylan[2] Alexander C. Berg[1]
[1]University of North Carolina at Chapel Hill [2]Adobe Research
eunbyung@cs.unc.edu {jimyang,yumer,ceylan}@adobe.com aberg@cs.unc.edu

1. Detailed Network Architectures

We provide the detailed network architecture of our approach in Figure 1.

2. More examples

We provide more visual examples for *car* and *chair* categories in Figures 2 and 3 respectively. In addition to novel views synthesized by our method, we also provide the intermediate output (visibility map and output of DOAFN) as well as views synthesized by other approaches.

3. Test results on random backgrounds

Figure 4 presents test results on synthesized images with random backgrounds. Intermediate stages, such as visibility map, background mask, and outputs of DOAFN are also shown. We compare against L_1 and AFN baselines. Note that L_1 and AFN could perform better on background area if we applied similar approaches used in TVSN, which we considered backgrounds separately.

4. Arbitrary transformations with linear interpolations of one-hot vectors

We show an experiment on the generalization capability for arbitrary transformations. Although we have trained the network with 17 discrete transformations in the range [20,340] with 20-degree increments, our trained network can synthesize arbitrary view points with linear interpolations of one-hot vectors. For example, if [0,1,0,0,...0] and [0,0,1,0,...0] represent 40 and 60-degree transformations respectively, [0,0,0.5,0.5,0,...0] represents 50 degree. More formally, let $\mathbf{t} \in [0,1]^{17}$ be encoding vector for the transformation parameter $\theta \in [20, 340]$ and s be step size ($s = 20$). For a transformation parameter $i \times s \leq \theta < (i+1) \times s$, i and $i+1$ elements of the encoding vector $\mathbf{t}$ is

$$\mathbf{t}^i = 1 - \frac{\theta - (i \times s)}{s}, \quad \mathbf{t}^{i+1} = 1 - \mathbf{t}^i \tag{1}$$

Figure 5 shows some of examples. From the third to the sixth columns, we used linearly interpolated one-hot vectors to synthesize views between two consecutive discrete views that were in the original transformation set (the second and the last columns).

5. More categories

We picked cars and chairs, since both span a range of interesting challenges. The car category has rich variety of reflectance and textures, various shapes, and a large number of instances. The chair category was chosen since it is a good testbed for challenging 'thin shapes', e.g. legs of chairs, and unlike cars is far from convex in shape. We also wanted to compare to previous works, which were tested mostly on cars or chairs. In order to show our approach is well generalizable to other categories, we also performed experiments for *motorcycle* and *flowerpot* categories. We followed the same experimental setup. We used the entire *motocycle*(337 models) and *flowerpot*(602 models) categories. For each category, 80% of 3D models are used for training, which leaves around 0.1 million training pairs for the *motorcycle* and 0.2 million for the *flowerpot* category. For testing, we randomly sampled instances, input viewpoints, and desired transformations from the rest 20% of 3D models. Figure 6 shows some of qualitative results.

References

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv:1512.03012*, 2015. 4, 5

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 4, 5

[3] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4, 5

[4] A. Lamb, V. Dumoulin, and A. Courville. Discriminative regularization for generative models. *arXiv:1602.03220*, 2016. 4, 5

[5] A. B. L. Larsen, S. K. Snderby, H. Larochelle, and OleWinther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 4, 5

[6] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016. 4, 5

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. chen. Improved techniques for training gans. In *NIPS*, 2016. 4, 5

[8] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 4, 5

[9] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016. 4, 5

[10] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *ECCV*, 2016. 4, 5

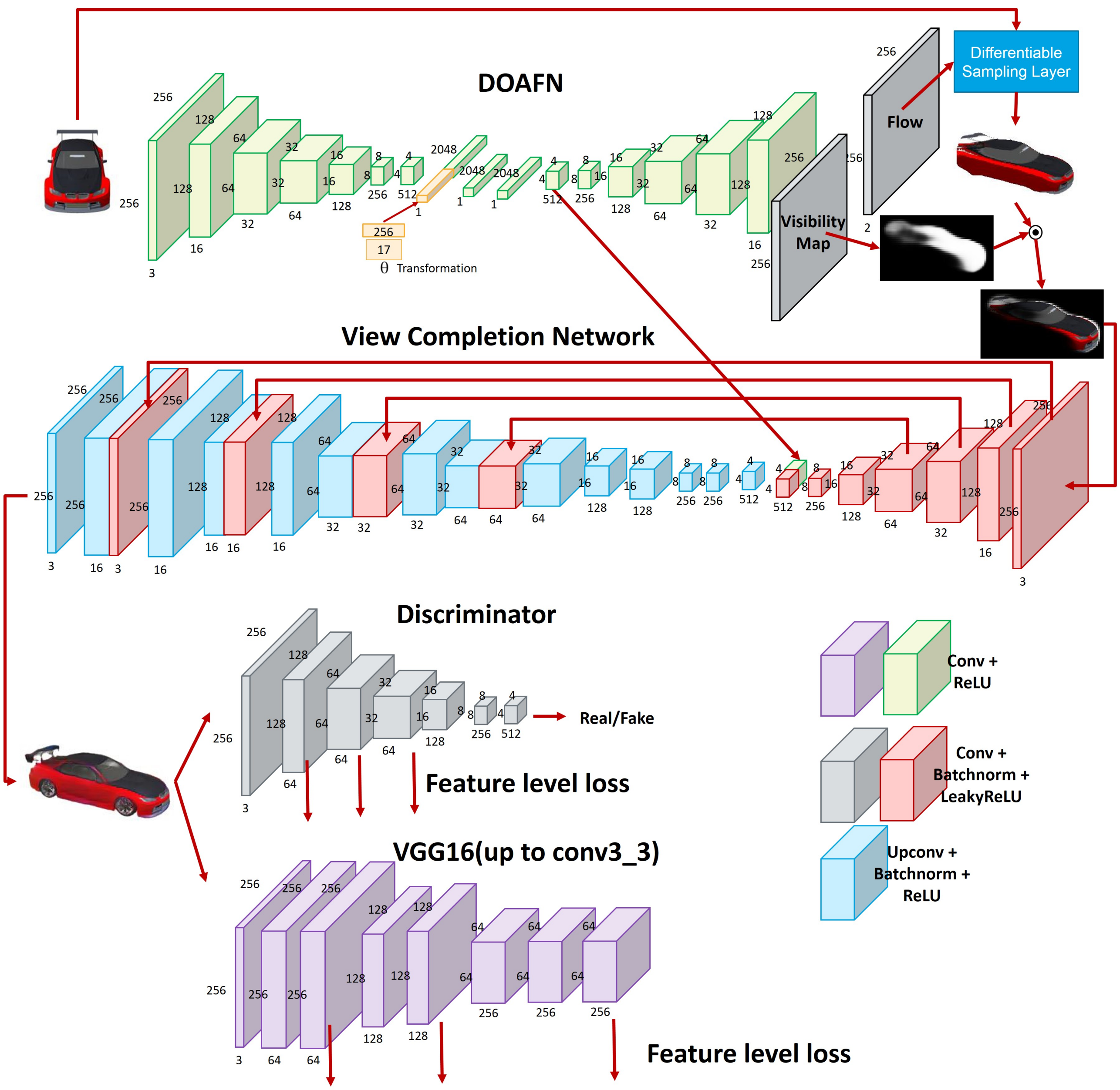

Figure 1. Transformation-grounded view synthesis network architecture

Figure 2. Results on test images from the car category [1]. 1st-input, 2nd-ground truth. From 3rd to 6th are deep encoder-decoder networks with different losses. (3rd-L_1 norm [8], 4th-feature reconstruction loss with pretrained VGG16 network [3, 5, 9, 4], 5th-adversarial loss with feature matching [2, 6, 7], 6th-the combined loss). 7th-appearance flow network (AFN) [10]. **8th-ours(TVSN)**.

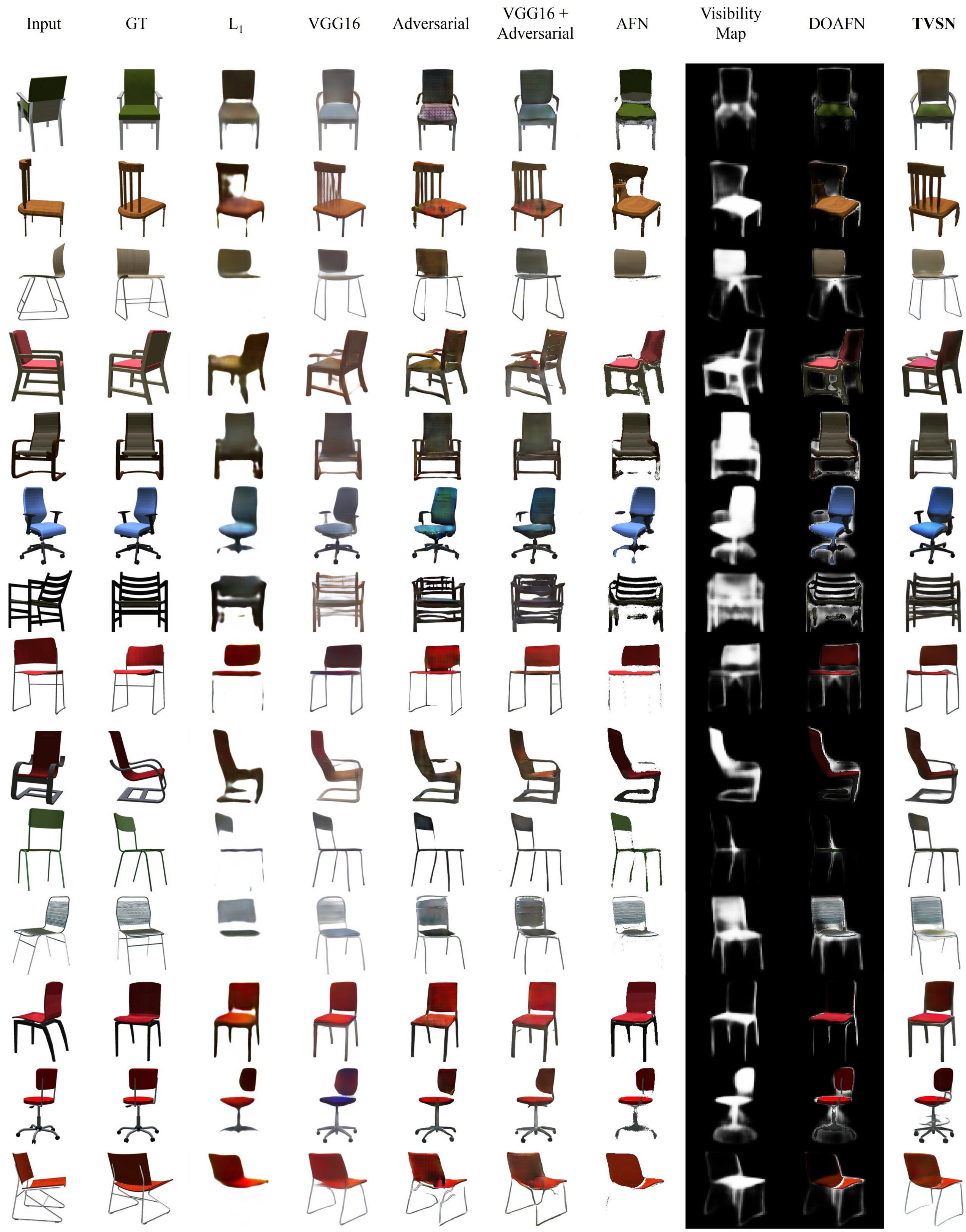

Figure 3. Results on test images from the car category [1]. 1st-input, 2nd-ground truth. From 3rd to 6th are deep encoder-decoder networks with different losses. (3rd-L_1 norm [8], 4th-feature reconstruction loss with pretrained VGG16 network [3, 5, 9, 4], 5th-adversarial loss with feature matching [2, 6, 7], 6th-the combined loss). 7th-appearance flow network (AFN) [10]. **8th-ours(TVSN)**.

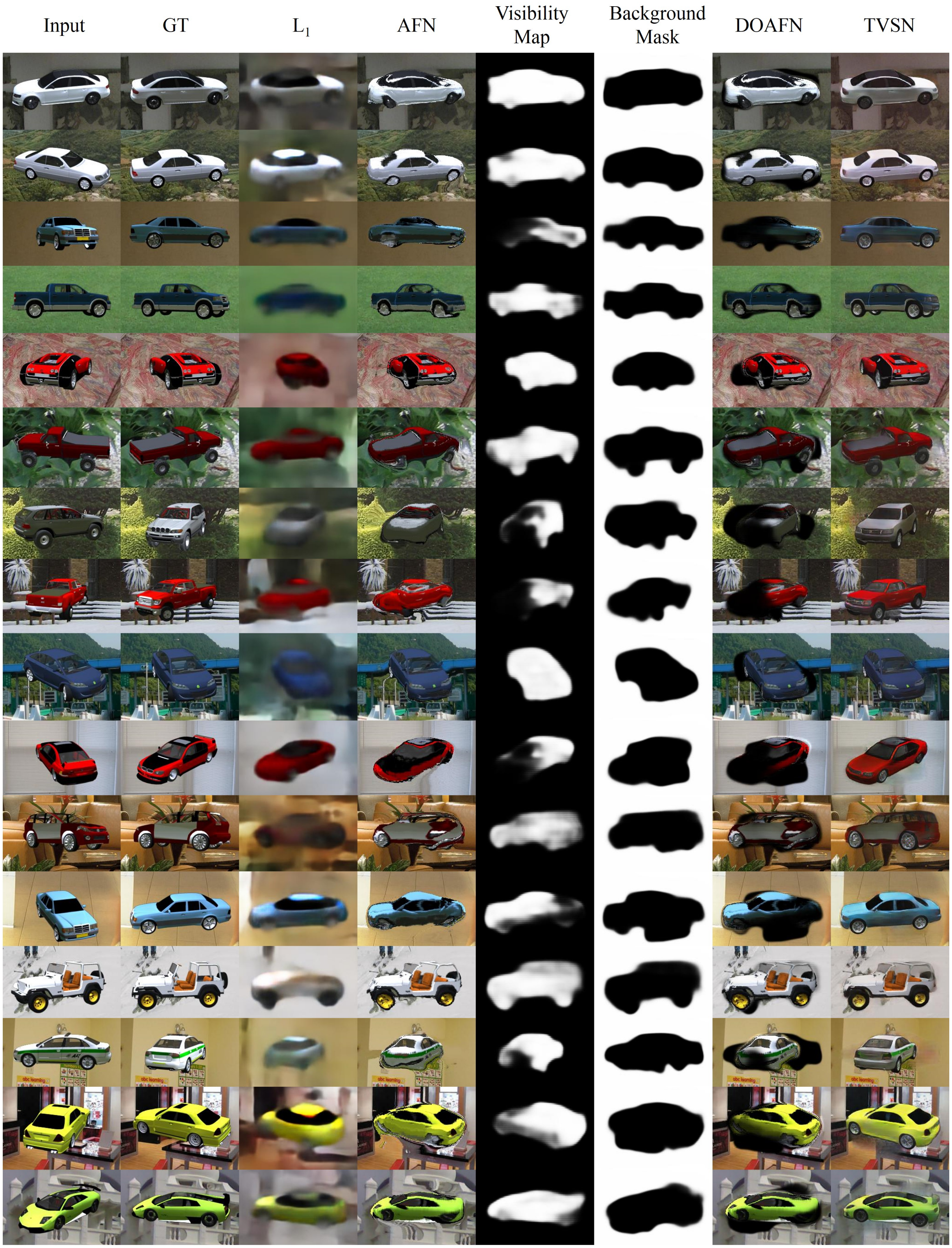

Figure 4. Test results on synthetic backgrounds

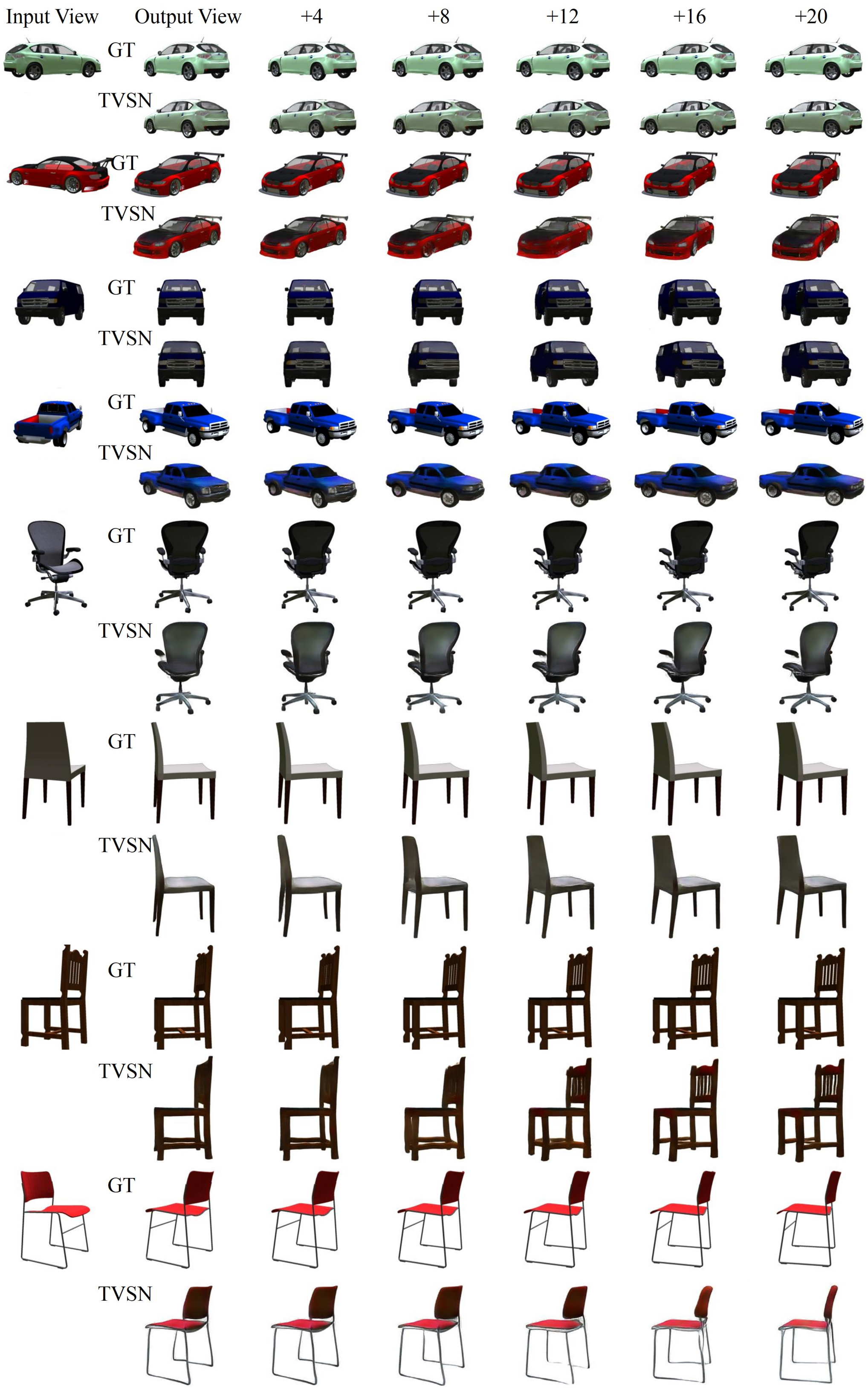

Figure 5. Test results of linear interpolation of one-hot vectors

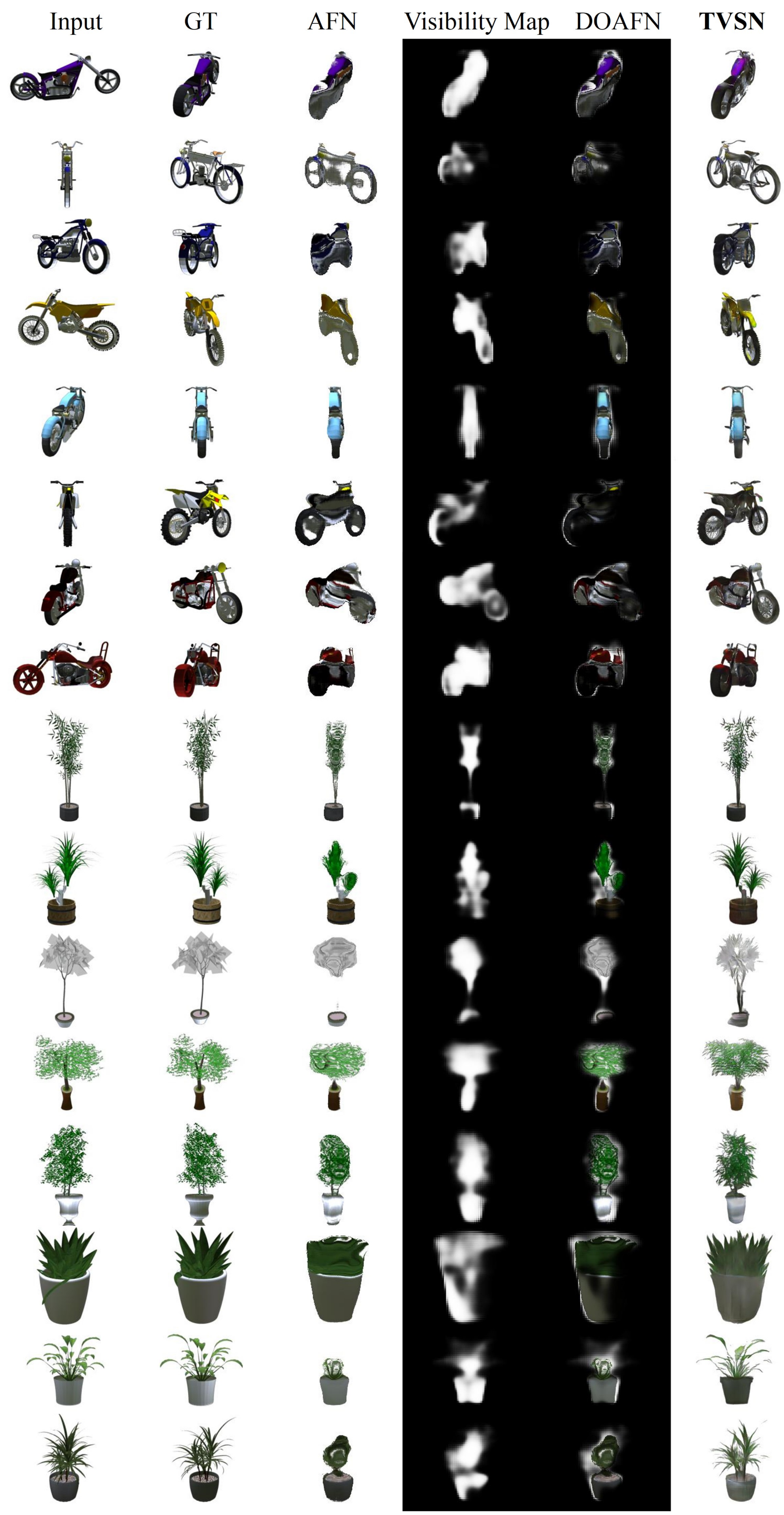

Figure 6. Test results of motorcycle and flowerpot categories