Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research/Review)

Supplementary document for STATE: Learning structure and texture representations for novel view synthesis

Xinyi Jing[1]*, Qiao Feng[1]*, Yu-Kun Lai[2], Jinsong Zhang[1], Yuanqiang Yu[1], and Kun Li[1] (✉)

1 Overview

This supplementary document is structured as follows: Section 2 provides more visual results for ablation study, comparisons with state-of-the-art methods and our generated images at different views. More results of the user study are shown in Section 3.

2 More Visual Results

In this section, we provide more visual results for Section 4.2 and Section 4.3 in the main paper, including comparisons with TBN [1], pixelNeRF [2] and four alternative designs. Comparisons with four alternative designs are shown in Figure 1. It can be seen that the w/o Tex. model can generate correct structures, *e.g.*, the legs and arms of chair, but the textures of the source images cannot be well maintained. The w/o Str. model can recover the detailed textures, *e.g.*, the headlights of car, but fails to keep the shape consistency. Compared with the w/o SVA model, it can be seen that the color and shape of the results of full method are closer to the ground-truths, especially on *Chair* dataset. The w/o Cos. model cannot ensure the color consistency, such as the lights of car and the seats of chair. On the contrary, our full model can achieve color, texture and structure consistency. Several visual comparisons on *Car*, *Chair* and *Human* datasets are shown in Figures 2-4. Thanks to the disentanglement of structure representation and texture representation, STATE successfully recovers the invisible regions and detailed textures for any number of input views.

In addition, we demonstrate the robustness and generalization capability of our approach in Figure 5. The first row gives the source images, and the second to the fifth rows show the generated images. Note that the objects in the test images are not included in the training set. It can be seen that our method generates reasonable images with consistent structure and texture from only four input images for various novel views.

3 User Study

In this section, we provide more results of the user study. Figure 6 shows the number of people choosing each method for twelve cases. It can be seen that more people choose C than A or B in terms of texture, structure and overall in twelve cases, which means more users agree that our results are better than those of TBN [1] and pixelNeRF [2].

1 College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. E-mail: Xinyi Jing, jingxinyi@tju.edu.cn; Qiao Feng, exculibar@tju.edu.cn; Jinsong Zhang, jinszhang@tju.edu.cn; Yuanqiang Yu, yuyuanqiang@tju.edu.cn; Kun Li, lik@tju.edu.cn(✉).
2 School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. E-mail: Yu-Kun Lai, LaiY4@cardiff.ac.uk.
* Contributed equally.

Manuscript received: 2014-12-31; accepted: 2015-01-30.

References

[1] K. Olszewski, S. Tulyakov, O. Woodford, H. Li, and L. Luo. Transformable bottleneck networks. In *Proc. International Conference on Computer Vision*, pages 7648–7657, 2019.

[2] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

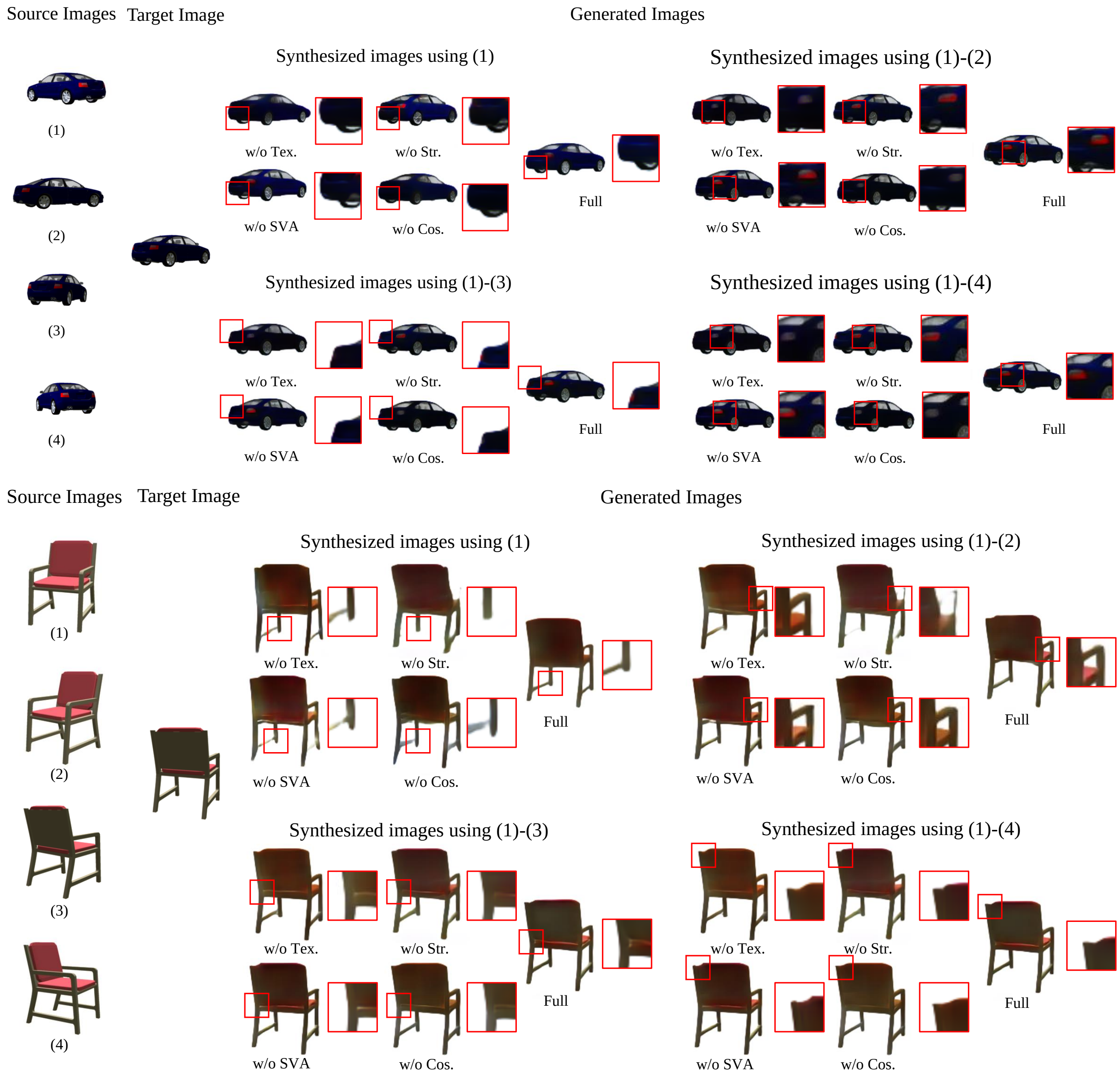
Source Images
Target Image
Generated Images
Synthesized images using (1)
Synthesized images using (1)-(2)
Synthesized images using (1)-(3)
Synthesized images using (1)-(4)
(1)
(2)
(3)
(4)
w/o Tex.
w/o Str.
w/o SVA
w/o Cos.
Full
Source Images
Target Image
Generated Images
Synthesized images using (1)
Synthesized images using (1)-(2)
Synthesized images using (1)-(3)
Synthesized images using (1)-(4)
(1)
(2)
(3)
(4)
w/o Tex.
w/o Str.
w/o SVA
w/o Cos.
Full

Fig. 1 Qualitative comparison with four alternative designs.

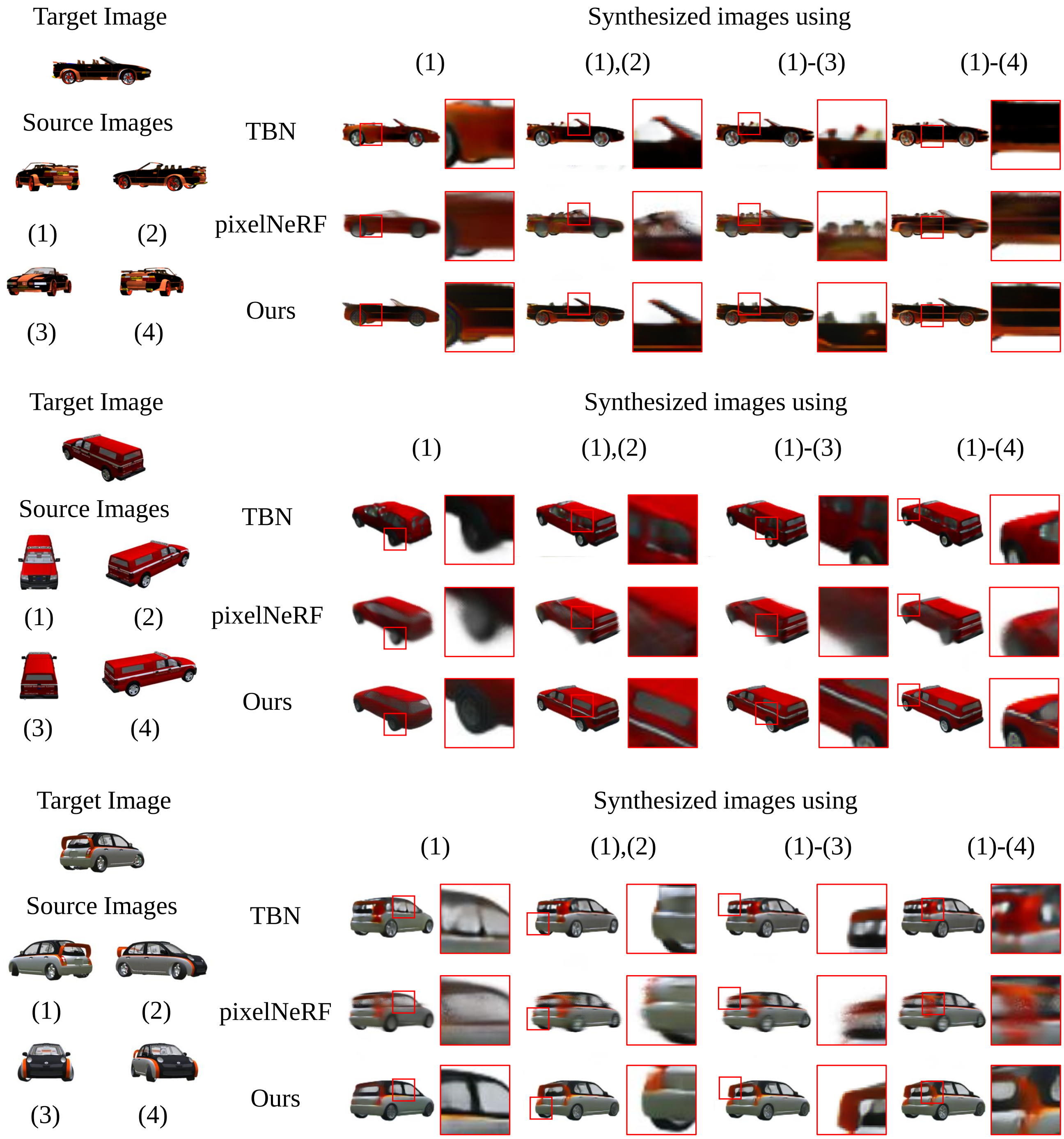

Fig. 2 Qualitative comparison on *Car* dataset.

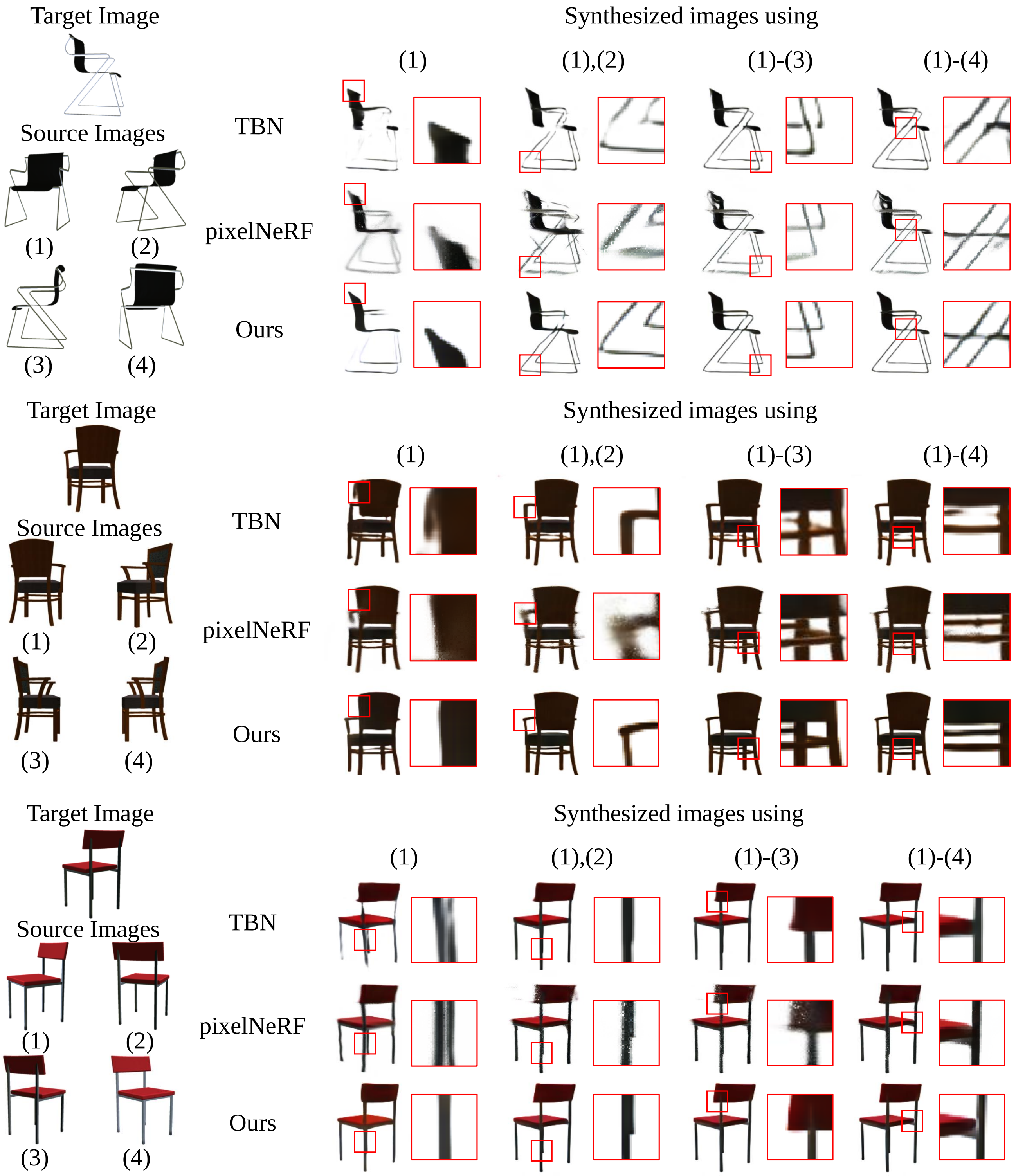

Fig. 3 Qualitative comparison on *Chair* dataset.

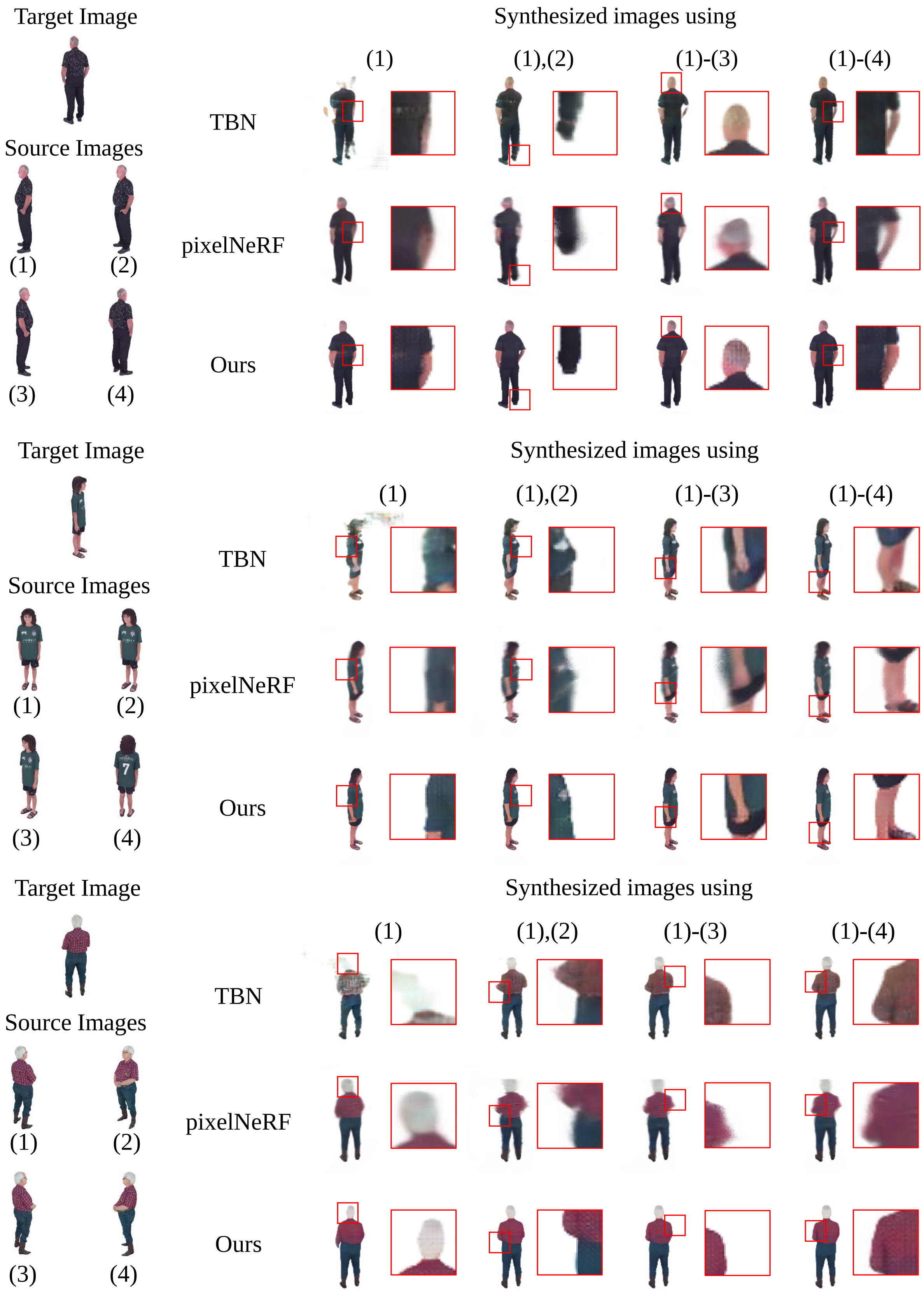

Fig. 4 Qualitative comparison on *Human* dataset.

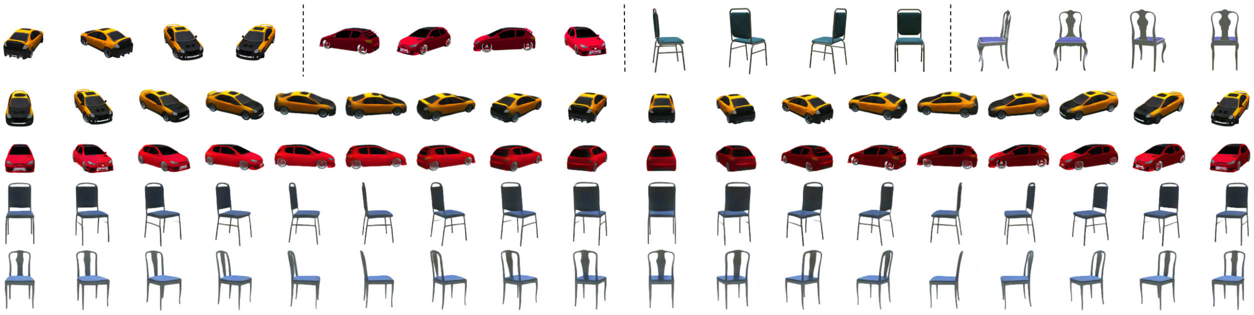

Fig. 5 Our generated images for various views from only four input images.

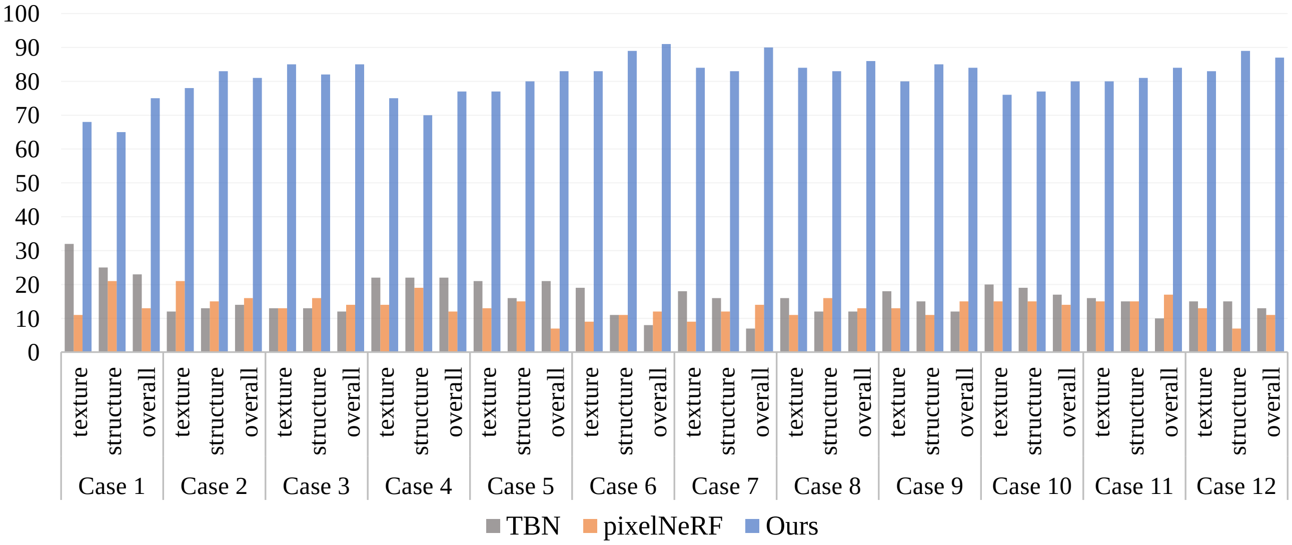

Fig. 6 User study result.