Supplementary Material: Self-Supervised Visibility Learning for Novel View Synthesis

Yujiao Shi [1], Hongdong Li [1], Xin Yu [2]
[1]Australian National University and ACRV [2]University of Technology Sydney
yujiao.shi@anu.edu.au, hongdong.li@anu.edu.au, xin.yu@uts.edu.au

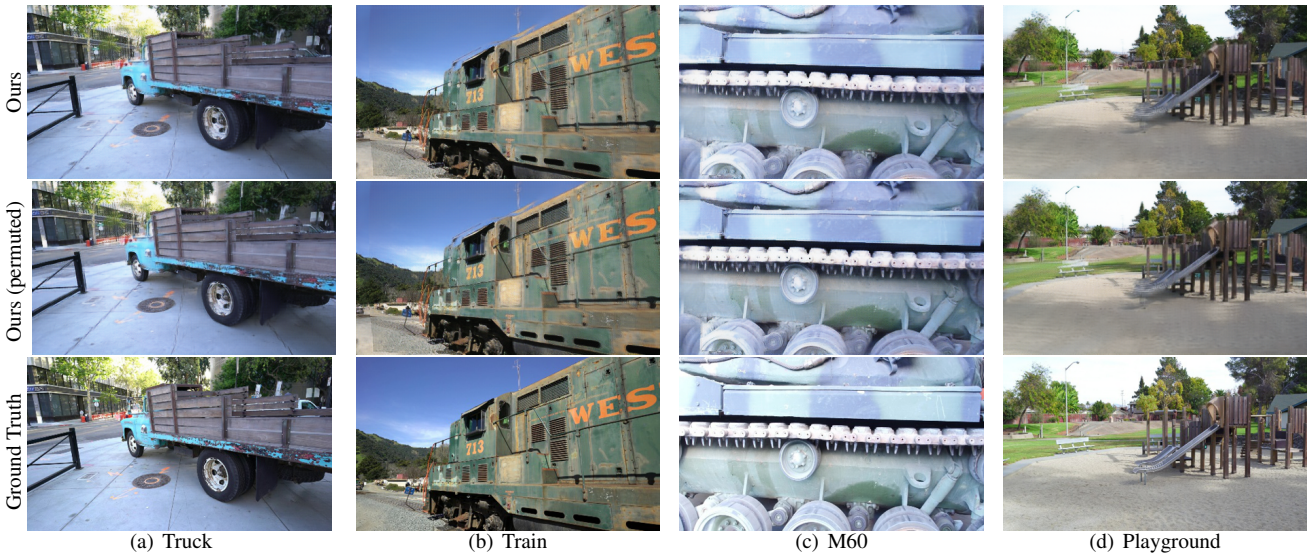
1. Robustness to Source-View Permutations

In real world scenarios, the input source views are usually unordered. Thus, we take into account source-view permutation invariance when designing our framework. To demonstrate this, we randomly permute the input source views and feed them to the same trained model, denoted as "Ours (permuted)". Quantitative results are presented in Tab. 1 and qualitative visualizations are provided in Fig. 1.

From the results, it can be seen that there is negligible difference on the synthesized images with different input view orders. This demonstrates the robustness of our method to source-view permutations.

Table 1: Robustness of our method to different input view orders.

	Tanks and Temples												DTU		
	Truck			Train			M60			Playground					
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
Ours	0.233	0.708	21.33	0.386	0.542	18.81	0.250	0.732	19.20	0.245	0.710	22.12	0.177	0.721	19.20
Ours (permuted)	0.233	0.708	21.34	0.385	0.542	18.81	0.250	0.732	19.13	0.246	0.710	22.13	0.177	0.721	19.20



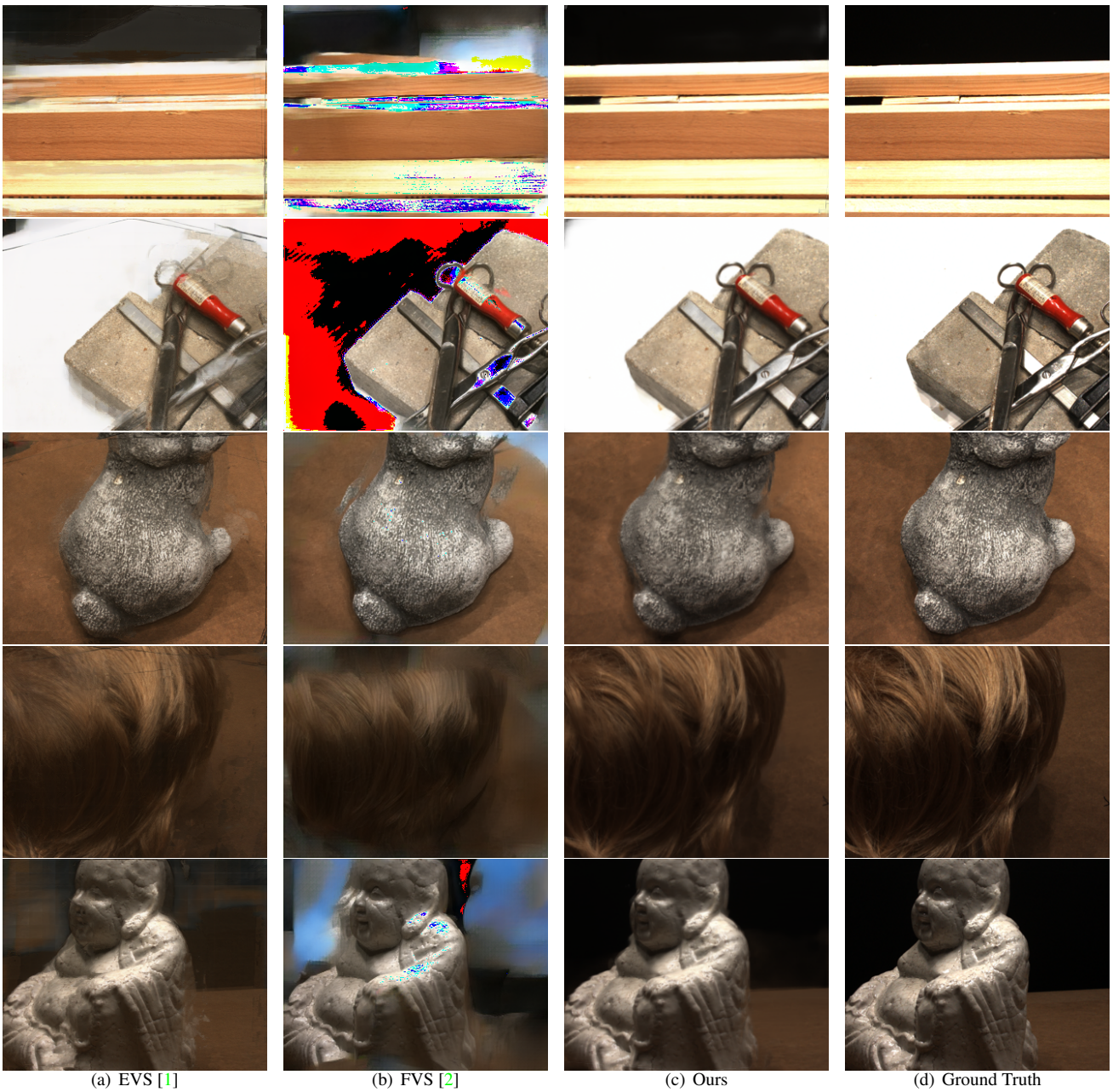
(a) Truck (b) Train (c) M60 (d) Playground

Figure 1: Qualitative comparison of generated images by our method with different input view orders. The four examples are the same as those in Fig. 1 in the main paper.

2. Additional Visualization on the DTU Dataset

In this section, we present more visualization examples on the DTU dataset, as shown in Fig. 2. Scenes in the DTU dataset are simpler than those in the Tanks and Temples dataset, *i.e.*, there is always a single salient object in each image. However, there are large textureless regions and some of the object surfaces (*e.g.*, the second example in Fig. 2) are reflective.

For those regions, COLMAP almost fails to predict the depths, and thus the images synthesized by FVS [2] suffer severe artifacts. EVS estimates target-view depths and source-view visibility from the source view depths. When the source view depths are inaccurate and/or disagree with each other, the error will be accumulated to the final synthesized image. In contrast, our method does not rely on the accuracy of source view depths and is able to handle textureless and reflective regions. Thus our synthesized images are much closer to ground-truth images.



(a) EVS [1] (b) FVS [2] (c) Ours (d) Ground Truth

Figure 2: Additional qualitative visualization of generated results on the DTU dataset with six views as input.

3. Additional Visualization on the Source-View Visibility Estimation (SVE)

In this section, we provide more visualization results on our visibility-aware aggregation mechanism, which is indicated by Eq. (2) in the main paper. We choose the four examples presented in Fig. 1 in the main paper for this visualization. The results are illustrated in Fig. 3, Fig. 4, Fig. 5 and Fig. 6, respectively. The camera movements between source and target views are provided under each input source view image in the Figures.

4. Qualitative Illustration on the Proposed Soft Ray-Casting (SRC)

The proposed soft ray-casting (SRC) mechanism is one of the key components in our framework. It converts a multi-modal surface existence probability along a viewing ray to a single-modal depth probability.

We demonstrate its necessity by removing it from our whole pipeline, denoted as "Ours w/o ray-casting" in the main paper. Fig. 7 presents the qualitative comparison. For the depth visualization, we apply a softargmax to the surface probability distribution (in "Ours w/o ray-casting") or the depth probability distribution (in "Our whole pipeline"). From Fig. 7, it can be observed that the generated depth and image by "Ours w/o ray casting" are more blurred than "Our whole pipeline". This is mainly because the surface probability distribution is always multi-modal and it does not reflect pure depth information, demonstrating the effectiveness of the proposed soft ray-casting mechanism.

5. Residual Visualization Between Images Before and After Refinement

We visualize the aggregated images by Eq.(2) in the paper and the corresponding output images after the refinement network in Fig. 8. To show the differences, we also present the residual images between the aggregated images and final output images.

References

[1] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7781–7790, 2019. 2

[2] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020. 2

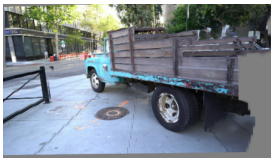

Rot = $[0.90°, 1.73°, -0.71°]$
Trans = $[-25.56\text{cm}, -2.50\text{cm}, 3.32\text{cm}]$


Rot = $[-0.95°, -4.10°, 1.67°]$
Trans = $[24.88\text{cm}, 0.04\text{cm}, 22.57\text{cm}]$


Rot = $[-6.46°, -8.52°, -2.26°]$
Trans = $[-51.16\text{cm}, 22.23\text{cm}, 100.32\text{cm}]$


Rot = $[-1.34°, -11.69°, 1.28°]$
Trans = $[-23.63\text{cm}, -3.57\text{cm}, 57.15\text{cm}]$


Rot = $[-6.45°, -11.41°, -2.27°]$
Trans = $[-68.72\text{cm}, 23.02\text{cm}, 93.47\text{cm}]$

Rot = $[-0.41°, -1.08°, 0.94°]$
Trans = $[24.33\text{cm}, 30.47\text{cm}, -4.96\text{cm}]$

(a) Input Source-View Images (b) Warped Source-View Images

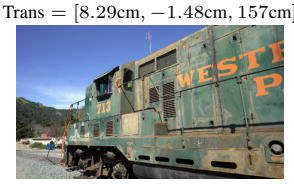

(c) Target-View Ground Truth (d) Aggregated Image

Figure 3: Qualitative illustration on our visibility-aware aggregation (*Truck*).


Rot = $[11.73°, 5.42°, 9.12°]$
Trans = $[7.70\text{cm}, 1.27\text{cm}, 175.36\text{cm}]$


Rot = $[16.40°, 1.85°, 14.23°]$
Trans = $[42.91\text{cm}, -8.60\text{cm}, 121.41\text{cm}]$


Rot = $[15.83°, 2.58°, 13.22°]$
Trans = $[8.29\text{cm}, -1.48\text{cm}, 157\text{cm}]$




Rot = $[0.10°, -2.71°, 0.55°]$
Trans = $[-21.16\text{cm}, -0.34\text{cm}, 17.46\text{cm}]$

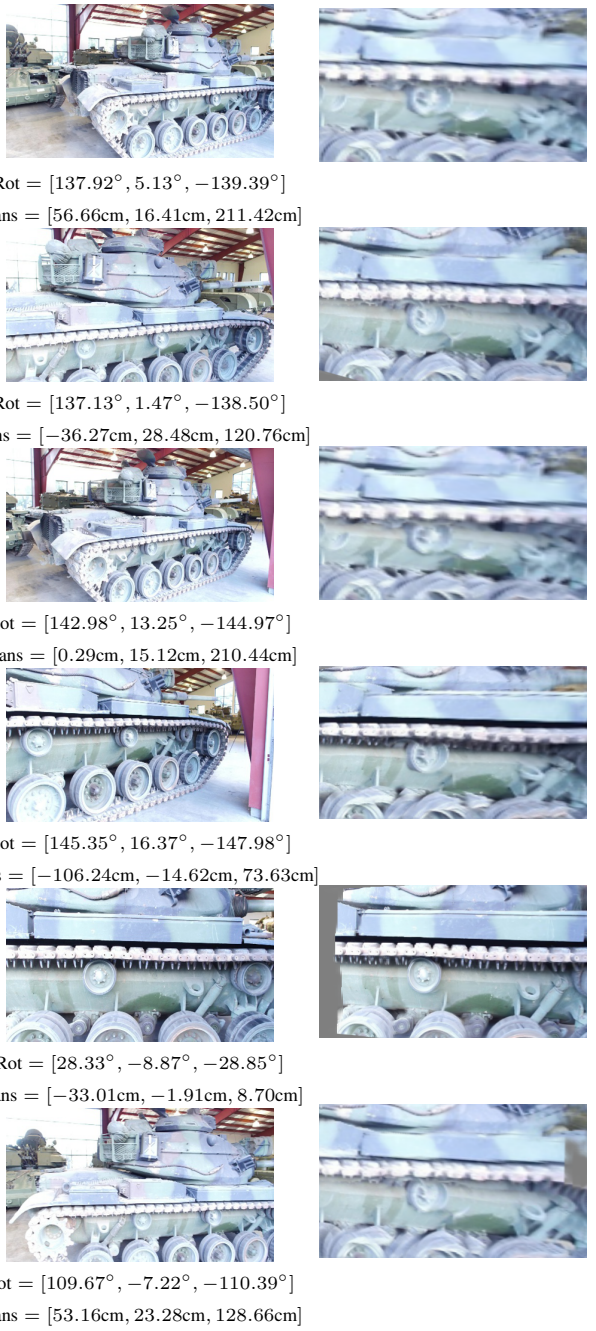

Rot = $[20.18°, 4.12°, 18.33°]$
Trans = $[-52.76\text{cm}, 5.25\text{cm}, 260.19\text{cm}]$


Rot = $[18.20°, 1.54°, 15.74°]$
Trans = $[-24.79\text{cm}, 2.41\text{cm}, 198.15\text{cm}]$

(a) Input Source-View Images (b) Warped Source-View Images


(c) Target-View Ground Truth (d) Aggregated Image

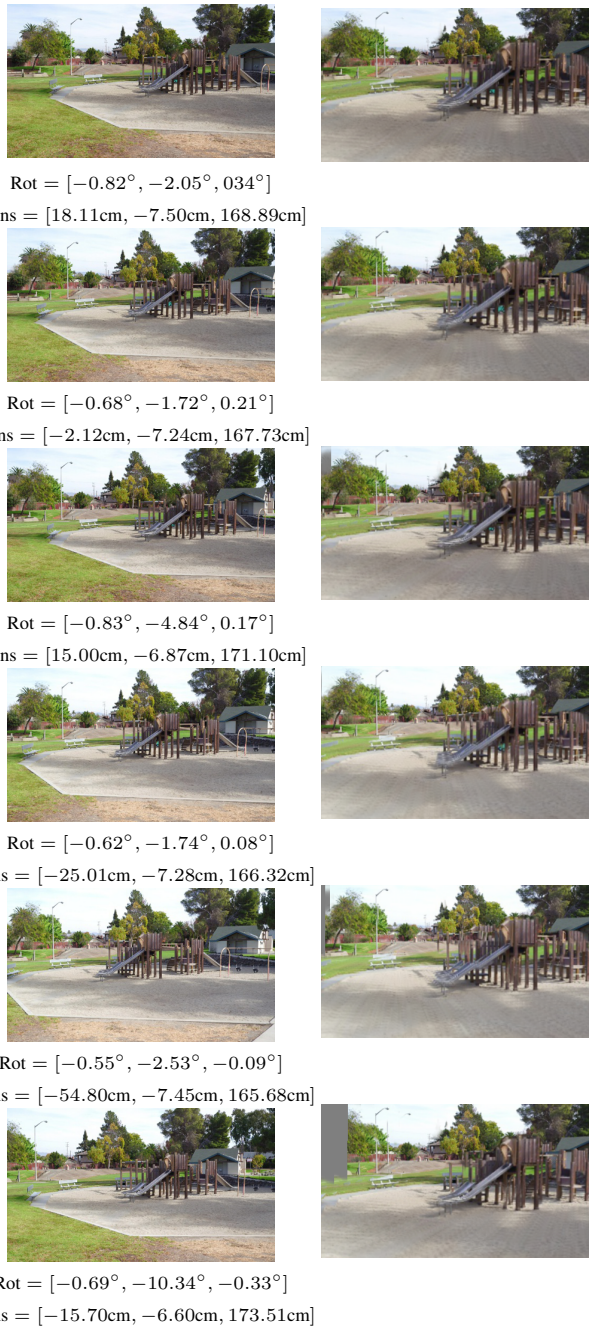
Figure 4: Qualitative illustration on our visibility-aware aggregation (*Train*).



(a) Input Source-View Images (b) Warped Source-View Images

(c) Target-View Ground Truth (d) Aggregated Image

Figure 5: Qualitative illustration on our visibility-aware aggregation (*M60*).



(a) Input Source-View Images (b) Warped Source-View Images

(c) Target-View Ground Truth (d) Aggregated Image

Figure 6: Qualitative illustration on our visibility-aware aggregation (*Playground*).

Truck Train M60 ground

(a) Generated depth (the first row) and image (the second row) by "Ours w/o ray-casting".

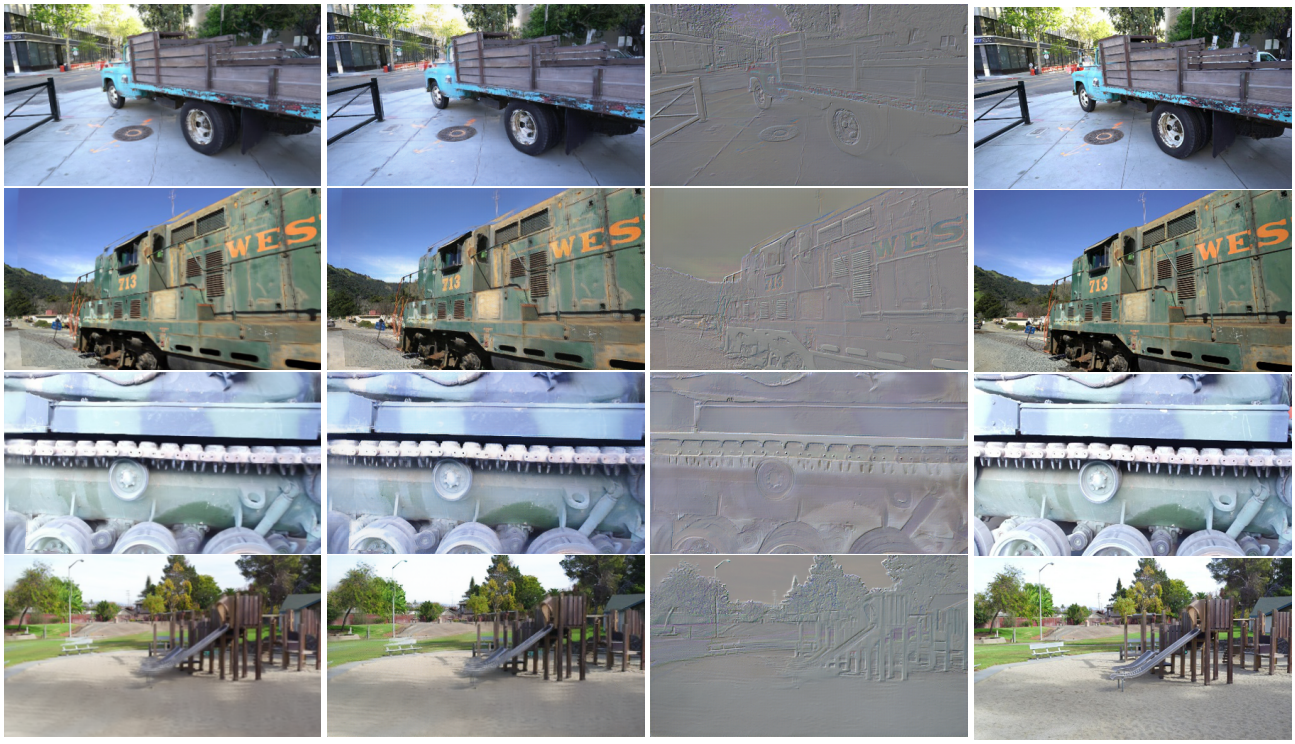
Truck Train M60 ground

(b) Generated depth (the first row) and image (the second row) by "Our whole pipeline".

Truck Train M60 ground

(c) Ground Truth

Figure 7: Qualitative comparison between "Ours w/o ray-casting" and "Our whole pipeline".



Aggregated Images | Output Images | Residual Images | Ground Truths

Figure 8: Visualization of aggregated images, output images after refinement, residual images between the aggregated images and the output images, and the ground truths. The four examples are from scenes "*Truck*", "*Train*", "*M60*" and "*Playground*", respectively. Their corresponding input images are presented in Fig.3, Fig.4, Fig.5, and Fig.6, respectively.