Novel View Synthesis with View-Dependent Effects from a Single Image

Supplementary Material

6. Introduction

We present additional details, experimental results, and visualizations that are essential to prove our NVSVDE-Net can learn to model and render novel views with view-dependent effects (VDEs).

In Section 7, we provide additional details on the epipolar projection equations that are used for the projection operations in our relaxed approximation of volumetric rendering. For reproducibility, in Section 8, we provide the details of our network backbones in our NVSVDE-net. In Section 9, we provide the details of our improved camera pose estimation network.

In Section 10 we provide additional visualizations of VDE modeled from *single image inputs*. In Section 11 we provide additional results on NVS from a single image and compare them against Single-view MPIs [40], PixelNeRF [46], MINE [26], BehindScenes [42], and SceneRF [2]. In Section 12, we provide additional visualizations of intermediate network outputs, such as the coarse NVS output, the mean sample distance estimated by the sampler MLP head, and the changes in geometry and VDE weights induced by different camera motions.

In addition to the qualitative results presented in this supplemental, we also attach videos that better show our VDEs and Novel Views against the previous methods (these are also available in `https://shorturl.at/ABIJ3`). Furthermore, we complement the results in our main paper by providing videos for our qualitative results. We refer to these videos by FigX-video in the attached supplemental materials. Higher resolution/length videos can also be found at `https://shorturl.at/ltJT7`.

Finally, we push the boundaries of our model by rendering views with an impressive 40-frame disparity from the input view in Section 13 and discuss our method's limitations and failure cases in Section 14.

7. Epipolar Projections

Our base relaxed volumetric rendering requires projected epipolar colors and probability logits. Such colors are obtained by projecting lines from the target ray sampling points at t_i (or t_i^*) depths to the camera centers of the reference view. The color is then sampled from the intersections with the camera plane. Assuming a pinhole camera model, the coordinates of the epipolar color projection for a pixel in image I are given by

$$g(\boldsymbol{p}, t_i, R_c|\boldsymbol{t}_c, K) = \boldsymbol{p'} \tag{15}$$

where $\boldsymbol{p'}$ is the image pixel coordinate to sample from reference image I via bilinear sampling. $\boldsymbol{p'}$ is computed by:

(i) Obtaining 3D points in world coordinates $\boldsymbol{p}_w$ by lifting $\boldsymbol{p}$ into the depth t_i as described by

$$\boldsymbol{p}_w = t_i K^{-1} \boldsymbol{p}; \tag{16}$$

(ii) Getting reference camera coordinates $\boldsymbol{p}_{c_j}$ by rotating and translating the world coordinates by the reference camera extrinsics (rotation and translation). Given camera-to-world extrinsics, $\boldsymbol{p}_{c_j}$ is given by

$$\boldsymbol{p}_c = R'_c \boldsymbol{p}_{w_j} - R'_c \boldsymbol{t}_c, \tag{17}$$

where R'_c and $\boldsymbol{t}_c$ are the inverse camera rotation matrix and translation vector of the reference view, and

(iii) Projecting camera coordinates into reference view image coordinates, as given by

$$\boldsymbol{p'} = K \frac{\boldsymbol{p}_c}{z_c}, \tag{18}$$

where K are the camera intrinsics and z_c is the Z-axis component of $\boldsymbol{p}_c$.

8. Additional Network Architecture Details

The network backbone F_W extracts geometry and pixel-aligned features W_D and W_V, consisting of an encoder-decoder network architecture with skip connections. An ImageNet [7] pre-trained ResNet-34 [19] is chosen for F_W for most of our experiments and design explorations, as it makes for a fast yet effective feature extractor encoder backbone. For the decoder side, we upscale the deep features via the nearest interpolation to the resolution of its skip connection, followed by a CONV-ELU-SKIP-CONV-ELU block, similar to the well-known Monodepth2 [15]. However, at the decoder stage, we concatenate into the skip connection deep-encoded pixel positional information given by the relative pixel locations (U, V). Pixel positional information is encoded by a 1×1Conv-ELU-1×1Conv (or MLP) head with 16 hidden units that encode the horizontal and vertical pixel positional information into an 8-element vector. We repeat this process until we reach the input resolution. All Conv layers in our decoder are 3×3 convolutions. To better assess the effects of our design choices (volumetric rendering approximations, rendering of VDE, etc.) we do not use any advanced block such as attention or dropout. For a fair comparison, all methods in the Experiments and Results Section share the same network backbone F_W.

For our NVSVDE-Net, we set two output branches at the last decoder stage, one for W_D and one for W_V, with N and

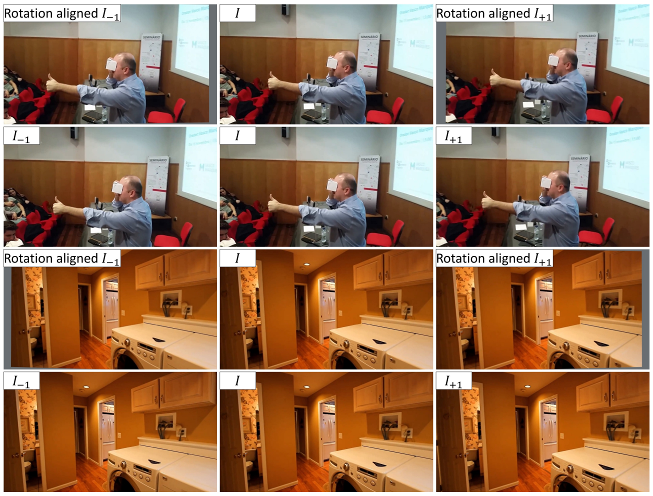

Figure 11. Our improved PoseNet incorporates rotation-aligned views for improved camera pose estimation. Rotation-aligned views allow our PoseNet to extract more relevant visual features for finer pose estimation. See attached *RotAligned.gif* for an animated visualization.

N_v numbers of channels, respectively. Our NVSVDE-Net incorporates additional Linear-ELU-Linear MLP heads F_D and F_V with 32 hidden units. Finally, our Sampler Linear-ELU-Linear-ELU-Linear MLP head with $4N$ input units and $2N^*$ output units utilizes 64 hidden units.

9. Additional Details on our Improved PoseNet

We provide detailed descriptions of each layer in our improved PoseNet in Table 3. In addition, it is worth noting that the camera pose network is not trained with random resized and cropped patches (as the NVS networks) but with full images at 1/2 resolution. The same resolution, 1/2, is used during testing.

A core feature of our improved PoseNet, is the processing of rotation-aligned images, which allows for extracting more relevant visual features. Fig. 11 shows that it is much "easier" to understand the diagonal translation motion in the top scene and the Z-axis motion on the bottom scene when the images are rotation aligned. See attached *RotAligned.gif* for an animated visualization.

10. Additional Visualization of VDEs

Due to their high sparsity, and low-frequency nature, VDEs such as glossy reflections are hard to visualize in still images. For this reason, we include several examples of our single-view-based VDE in the attached *VDE-video.mp4* video. We kindly suggest viewing it.

11. Additional Results on NVS from a Single Image

For the sake of completeness, we also include the comparison with the previous methods of Single-view MPI [40] and MINE [26] in this supplemental. Table 5 shows the extended version of Table 1 with results for Single-view MPI and MINE which, for a fair comparison, were also trained with the same ResNet34 network backbone and under the same conditions as those methods in Section 4. Again, our NVSVDE-Net outperforms in terms of all metrics by a considerable margin. Interestingly, Single-view MPI [40] yields 0.4 dB better PSNR than SceneRF (but still 0.4 dB lower than ours) for the MC dataset, but with worse LPIPS, which is reflected in its qualitative results.

We also provide additional qualitative results on RE10 and MC in Figs. 13 and 14. Single-view MPI [40] and MINE [26] methods attempt to solve a very ill-posed problem of estimating densities and colors for all pixels in all image planes in the MPIs but yield very blurred results, as shown in Figs. 13 and 14. In contrast, our NVSVDE-Net with a "relaxed" volumetric rendering approximation yields the sharpest novel views with baked-in view-dependent effects. To complement Figs. 13 and 14 we also attach *RE10k-results.mp4* and *MC-results.mp4* videos in this supplementary materials. We kindly suggest to view them. As can be observed in the third sample of *RE10k-results.mp4*, previous methods either provide blurred results for the highly reflective regions (such as Single-view MPI and MINE) or completely warps the reflective surface (such as SceneRF) instead of modeling VDEs from a single image. It is worth noting that the warping in methods such as SceneRF is due to the method modeling the reflections to be at the symmetrical "mirror" depth, which causes reflections to be considered as located farther than the reflective surface. In contrast, our NVSVDE-Net infuses VDEs into the input images (last column of *RE10k-results.mp4*) and then synthesizes novel views based on our relaxed volumetric rendering approximation.

As shown in Fig. 14 and Video *MC-results.mp4*, our NVSVDE-Net generates the highest quality of novel views, even for complex camera motions such as those in the second and third rows.

11.1. Additional Comparisons Against Trilinear Density Interpolation

Instead of proposing a novel relaxed volumetric rendering as in our NVSVDE-Net, the previous work of [48] proposed to perform a trilinear sampling of density in a predicted density volume for novel view synthesis. In contrast, in our relaxed VR, F_D and F_V perform a more complex non-linear mapping from source to target densities (instead of trilinear sampling), allowing for a higher level of details in the novel

Outputs	Layer descriptions	Inputs	Channels	Feature sizes
Image Encoder				
I	Input image	-	3	H×W
Conv1	3×3Conv(s2), ELU, ResBlock	I	32	H/2×W/2
Conv2	3×3Conv(s2), ELU, ResBlock	Conv1	64	H/4×W/4
Conv3	3×3Conv(s2), ELU, ResBlock	Conv2	128	H/8×W/8
Conv4	3×3Conv(s2), ELU, ResBlock	Conv3	128	H/16×W/16
Joint Encoder				
I_1, I_2^R	Concatenated rotation aligned pair	-	6	H×W
Conv1	3×3Conv(s2), ELU, ResBlock	I	32	H/2×W/2
Conv2	3×3Conv(s2), ELU, ResBlock	Conv1	64	H/4×W/4
Conv3	3×3Conv(s2), ELU, ResBlock	Conv2	128	H/8×W/8
Conv4$_{joint}$	3×3Conv(s2), ELU, ResBlock	Conv3	128	H/16×W/16
Pose Estimator				
Conv5	3×3Conv(s2), ELU, ResBlock	Conv4$_1$, Conv4$_2$	128	H/32×W/32
Conv6	3×3Conv(s2), ELU, ResBlock	Conv5	256	H/64×W/64
Conv7	1×1Conv, ELU	Conv6	256	H/64×W/64
$R_0, \boldsymbol{t}_0$	1×1Conv, ELU	GAP(Conv7)	6	1
Δ Pose Estimator				
Conv5	3×3Conv(s2), ELU, ResBlock	Conv4$_{joint}$, Conv4$_1$, Conv4$_2$	128	H/32×W/32
Conv6	3×3Conv(s2), ELU, ResBlock	Conv5	256	H/64×W/64
Conv7	1×1Conv, ELU	Conv6	256	H/64×W/64
$\Delta R, \Delta \boldsymbol{t}$	1×1Conv, ELU	GAP(Conv7), $R_0, \boldsymbol{t}_0$	6	1
I_1	Input image	-	3	H×W
I_2	Input image (target view in NVS)	-	3	H×W
Conv4$_1$	Image Encoder	I_1	128	H/16×W/16
Conv4$_2$	Image Encoder	I_2	128	H/16×W/16
$R_0, \boldsymbol{t}_0$	Pose Estimator	Conv4$_1$, Conv4$_2$	6	1
I_2^R	$I_2^R = I(g(R_0))$	I_2, R_0,	3	H×W
Conv4$_{joint}$	Joint Encoder	I_1, I_2^R	128	H/16×W/16
$\Delta R, \Delta \boldsymbol{t}$	Δ Pose Estimator	Conv4$_{joint}$, Conv4$_1$, Conv4$_2$, $R_0, \boldsymbol{t}_0$	6	1
$R, \boldsymbol{t}$	$R = R_0 + \Delta R, \boldsymbol{t} = \boldsymbol{t}_0 + \Delta \boldsymbol{t}$	$R_0, \Delta R, \boldsymbol{t}_0, \Delta \boldsymbol{t}$	6	1

Table 3. Detailed network architecture of our improved PoseNet. s2: Stride of 2. GAP: Global Average Pooling. ResBlock(x): ELU(3×3Conv(ELU(3×3Conv)) + x). ELU: Exponential Linear Unit [5].

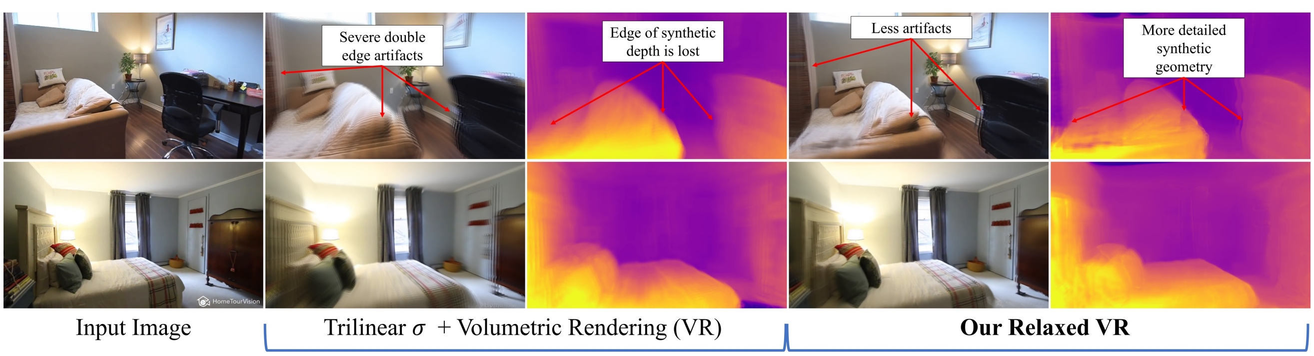

Figure 12. Relaxed-VR VS σ-interpolation-VR.

views and geometries as evidenced by Fig. 12. Fig. 12 depicts results at large viewpoint changes (40 frames apart, approximately 2.5× the training viewpoint changes in the training set), showing that our approach can better handle density discontinuities due to large camera motion thanks to F_D and F_V.

Our relaxed volumetric rendering approach is not only more accurate, as shown in Table 4, but also computationally less expensive than trilinear sampling, which requires sampling N source densities not only in x, y (like

Methods	VDE	MAE↓	PSNR↑	PSNR$_{lf}$↑	SSIM↑	LPIPS↓
Trilinear σ (I''_c)	Yes	0.0343	23.9083	29.7102	0.8267	0.2560
Relaxed VR (I''_c)	Yes	**0.0325**	**24.1020**	**29.9808**	**0.8343**	**0.2365**

Table 4. Relaxed-VR VS σ-interpolation-VR.

Methods	VDE	MAE↓	PSNR↑	PSNR$_{lf}$↑	SSIM↑	LPIPS↓
RealEstate10k (RE10k) Dataset [50]						
PixelNerf [46]	No	0.0417	22.8455	28.0945	0.7818	0.3256
BehindScenes [42]	No	0.0466	22.9949	28.5941	0.8068	0.2762
MINE [26]	No	0.0415	23.1657	27.8785	0.8041	0.2976
Single-view MPI [40]	No	0.0374	23.6260	28.9447	0.8112	0.2925
SceneRF [2]	No	0.0373	23.6087	28.9636	0.8130	0.2709
NVSVDE-Net (Ours)	**Yes**	**0.0319**	**24.3131**	**30.2529**	**0.8397**	**0.2325**
MannequinChallenge (MC) Dataset [27]						
PixelNerf [46]	No	0.0511	21.3047	25.2781	0.7580	0.3455
BHindScenes [42]	No	0.0463	21.4307	25.9280	0.7831	0.3101
SceneRF [2]	No	0.0467	21.5992	25.8119	0.7796	0.3080
MINE [26]	No	0.0487	21.6922	25.6230	0.7803	0.3306
Single-view MPI [40]	No	0.0460	22.0378	26.2500	0.7873	0.3251
NVSVDE-Net (Ours)	**Yes**	**0.0405**	**22.4274**	**27.0263**	**0.8130**	**0.2733**

Table 5. Single view-based NVS results. ↓/↑ denotes the lower/higher, the better.

ours) but also in z. Our sampling operation has a memory complexity of $H\times W\times N$, while trilinear sampling requires $H\times W\times N^2$, at least on the vanilla PyTorch.

11.2. Additional Qualitative Results on Ablation Studies

Fig. 15 shows results for our NVSVDE-Net with different network encoder backbones for F_W. Interestingly, even though our NVSVDE-Net (Swin-t) does not yield the highest PSNR in Table 2, it presents the most discriminative VDE activation maps and the most detailed depth maps, suggesting further improvements could be achieved by fine-tuning the network architecture design. In contrast, the NVSVDE-Net (R18), which incorporates a weaker encoder backbone, struggles to predict VDE activation maps focusing on the reflective scene regions.

12. Additional Visualizations of Intermediate Network Outputs

Fig. 17 depicts the different intermediate network outputs in our NVSVDE-Net. We show the infused VDEs and VDE activation map in I^v_{+1} and $\hat{V}$ respectively for a positive Z-axis camera translation in the first row. In the second row, Fig. 17 shows a comparison between our coarse synthetic view I''_{+1} and the final rendered view I'_{+1}. Note that the fine-grained ray sampling in I'_{+1} fixes the double edge artifacts in I''_{+1}.

The third row of Fig. 17 depicts the estimated geometry (inverse depth) of the input view $\hat{D}$, the mean sample distance $\bar{t}$ for the coarse synthetic view (a constant due to uniform sampling), the mean estimated sample distance $\overline{t^*}$, and the novel view geometry $\hat{D}_{+1}$. $\overline{t^*}$ resembles the scene disparity, which shows that ray samples are being taken around the highest-density regions in the scene. The novel view geometry $\hat{D}_{+1}$ can be estimated from the fine-grained ray sampling distances and weights by

$$\hat{D}_c(\boldsymbol{p}) = t^*(\boldsymbol{p}) \cdot w^*(\boldsymbol{w}). \tag{19}$$

In Fig. 17, $+1$ sub-index represents a view 8 frames apart from the input image I.

We also show the effects of the so-called re-calibration blocks F_D and F_V in our NVSVDE-Net. While the changes in $\hat{V}$ by F_V can be better visualized in the attached *Extreme-NVS.mp4* video, the changes in D^L are depicted in Fig. 16. Fig. 16. shows the corresponding channels of D^L for rendering two different novel views, one with negative Z-axis motion (left column) and the other with positive Z-axis motion (right column). From top to bottom, D^L channels represent the weights for close and far-away ray points. As can be noted, F_D changes the values in D_L to account for the novel camera view. For instance, the chair in row 3 of Fig. 16 is assigned less weight for the left column, as it will be further away in the novel view. Similarly, the chair is assigned more weight in row 3 for the right column, where the chair will be closer to the positive Z-axis motion novel view.

13. Rendering Views Beyond the Training Set

We trained our NVSVDE-Net to render views that are at most 16 frames apart from the single-image input. In this experiment, we render views equivalent to 40 frames apart from the input view and show our results in the attached *Extreme-NVS.mp4* video. We cordially invite the reader to observe the video, also available at the following anonymous repository link: https://shorturl.at/ABIJ3. Despite the inherent challenges associated with extreme Novel View Synthesis (NVS), our method consistently produces realistic views, albeit with certain observable artifacts, as anticipated in any single-view NVS framework.

The *Extreme-NVS.mp4* video illustrates two primary failure scenarios. Firstly, when the camera motion exceeds a certain threshold, our relaxed volumetric rendering encounters challenges in modeling large dis-occluded regions, a known issue prevalent in prior methods relying on projections for rendering [2, 26, 40, 42, 46]. Generative models have effectively addressed this concern but very often show stochastic artifacts. Secondly, under substantial camera motions, our negative disparity-based view-dependent effects (VDEs) struggle to model reflections accurately, leaking incorrect colors into the scene. Addressing this challenge may involve incorporating extreme NVS samples during training

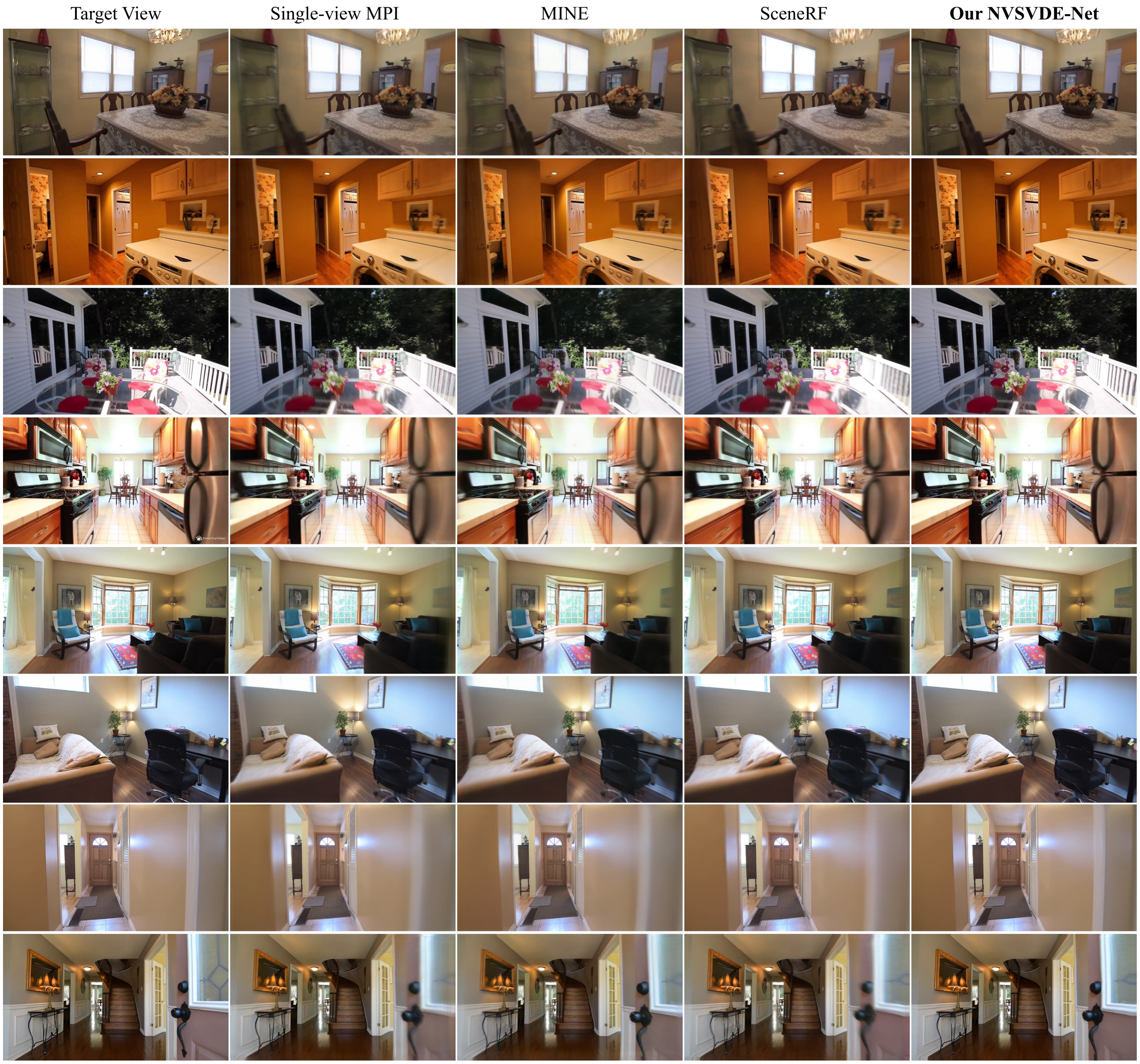

Figure 13. Additional Results on the RE10k [50] dataset. Previous methods struggle to render sharp structures for very close-by objects and reflective regions. Our NVSVDE-Net with explicitly modeled view-dependent effects (VDE) and fine-grained relaxed volumetric rendering yields more detailed synthetic views in all image regions.

and refining the regularization applied to the intermediate output I_c^v of the VDE-infused input image.

14. Limitations and Failure Cases

The architecture of our NVSVDE-Net, as defined in Eq. 5, imposes limitations on rendering high-frequency view-dependent appearances. This, however, proves to be adequate for rendering most glossy reflections in the context of realistic NVS. Future avenues of research will explore the modeling of both low- and high-frequency View-Dependent Embeddings from a single image to mitigate this inherent limitation.

It is also important to note that for the synthesis of view-dependent effects Section 3.2, we exploit two simple yet effective priors for simple glossy/diffuse specular reflections, which are plausible to estimate/render for single image inputs: (i) VDEs follow camera motion w.r.t. their reflective surfaces, (ii) VDE 'motion' cannot be larger than the

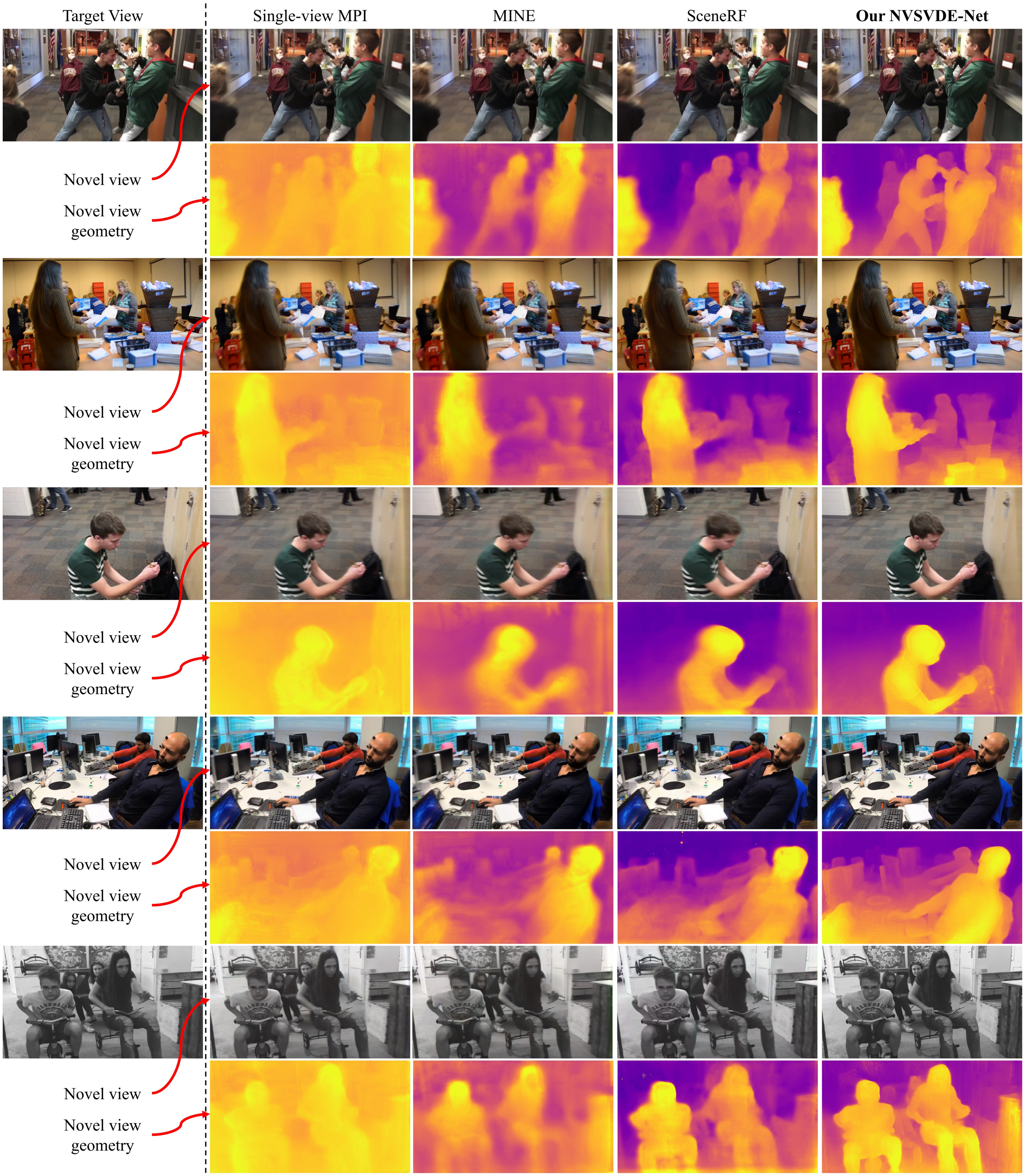

Figure 14. Additional Results on the MC [27] dataset.

rigid flow of the reflective surface itself. These assumptions can fail with complex reflections (e.g., on very concave or convex reflective surfaces). However, (i) and (ii) hold for the simpler glossy reflections/highlights we aim to model,

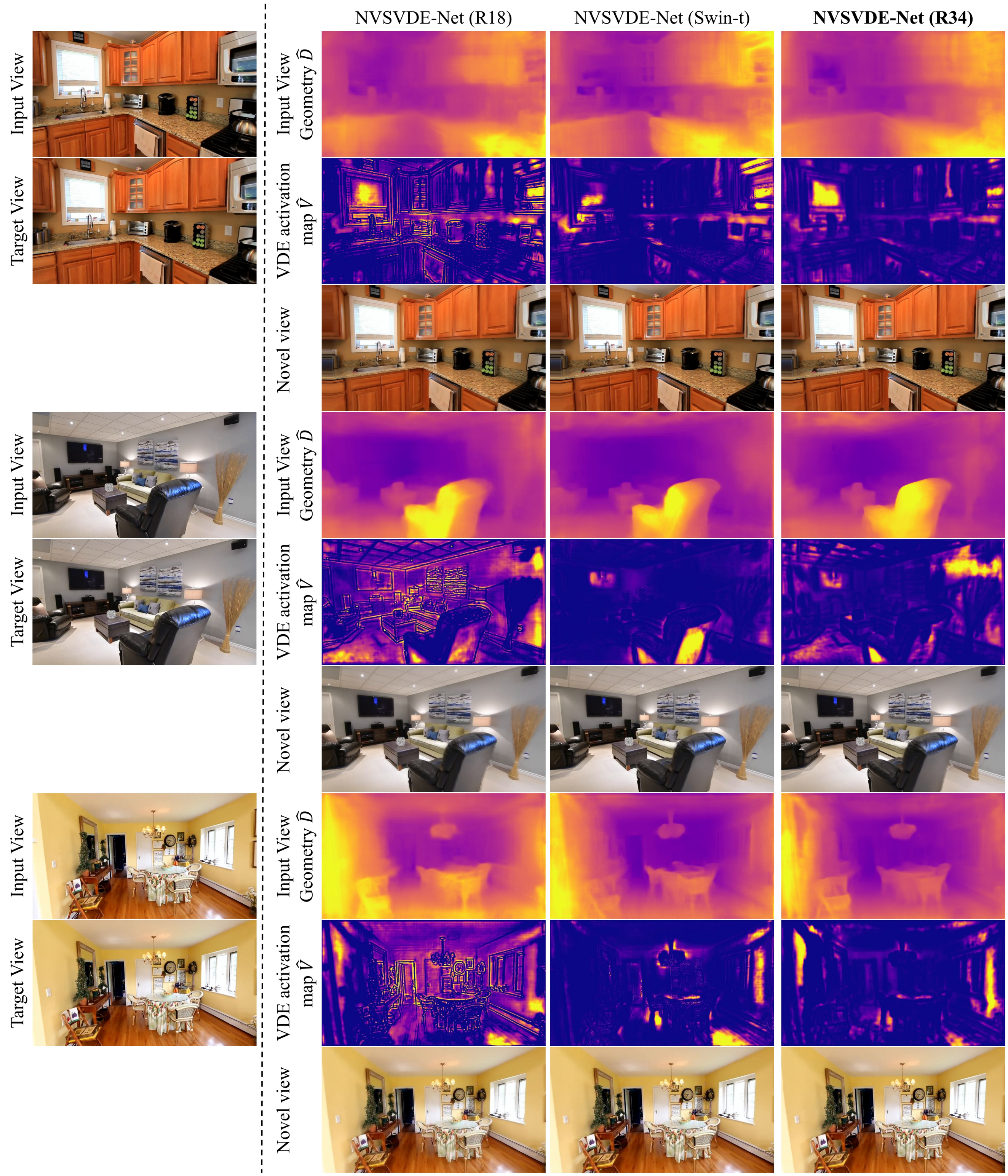

Figure 15. Qualitative comparison among different network backbones for our NVSVDE-Net. The Swin-t [30] backbone yields more discriminative VDE activation maps $\hat{V}$ that better focus on the most reflective surfaces in the input image and relatively more detailed input view geometries $\hat{D}$.

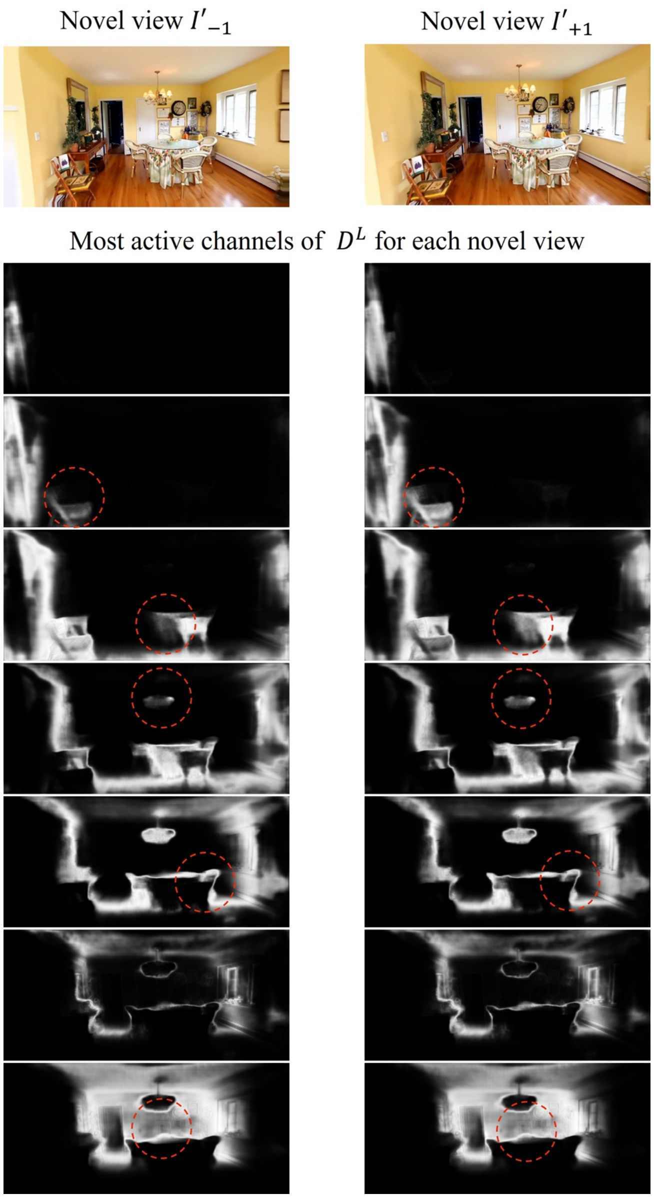

Figure 16. Visualization of the effects of the adjustment MLP block F_D. Differences in channel activations of D^L show the recalibration carried by F_D for rendering novel view geometry at the target camera views.

which are plausible to estimate/render from single image inputs.

Another limitation of our approach is apparent when dealing with extreme NVS scenarios, where novel views with baselines significantly surpassing those encountered during training are required. As demonstrated in Section 13, our network encounters challenges in generating artifact-free novel views when tasked with rendering views 40 frames apart from the input view, a significant departure from the training set's 16-frame disparity. The impediments arise from the sheer size of occlusions, rendering them impractical to inpaint through projected colors, and the limited context available to the sampler MLP head for predicting sampling distances from the few valid projected D^L and I_c^v values. Future research directions include incorporating generative model properties into our NVSVDE-Net for realistic wide-baseline occlusion inpainting.

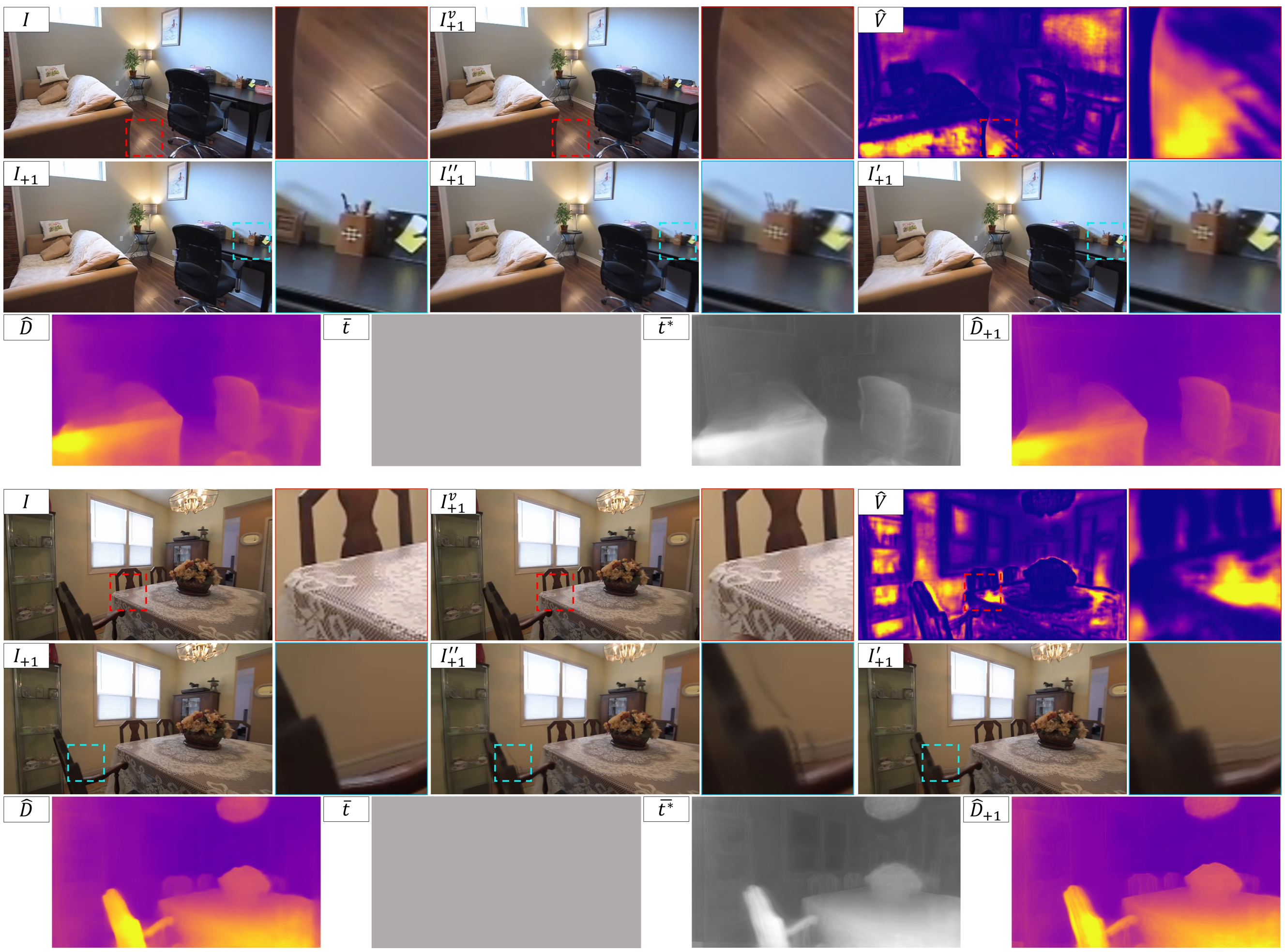

Figure 17. Intermediate outputs of our NVSVDE-Net

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021. 3

[2] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023. 1, 2, 3, 5, 6, 7, 8, 4

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[4] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 6, 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. *ACM Siggraph Computer Graphics*, 22(4): 65–74, 1988. 2, 3

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5491–5500. IEEE, 2022. 2, 3

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 4

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 6

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 2

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019. 2, 5, 6, 1

[16] Juan Luis Gonzalez and Munchurl Kim. Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6851–6860, 2021. 5

[17] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. In *Advances in Neural Information Processing Systems*, pages 12626–12637. Curran Associates, Inc., 2020. 4, 6

[18] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 1

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[21] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pages 577–584, New York, NY, USA, 2005. ACM. 2

[22] Youichi Horry, Ken Anjyo, and Kiyoshi Arai. Tour into the picture: Using spidery mesh interface to make animation from a single image". pages 225–232, 1997. 2

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[25] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4): 76–1, 2020. 2

[26] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. MINE: towards continuous depth MPI with nerf for novel view synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12558–12568. IEEE, 2021. 1, 2, 3, 5, 6, 7, 4

[27] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7, 8, 4

[28] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2

[29] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 806–815, 2023. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7, 8

[31] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4):29:1–29:14, 2019. 6

[32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 405–421. Springer, 2020. 1, 2, 3, 5, 7

[33] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019. 2

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[35] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022. 3

[36] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. 3

[37] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. VQ3D: Learning a 3D-aware generative model on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[39] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 2, 3

[40] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4, 5, 6, 7

[41] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 7

[42] F. Wimbauer, N. Yang, C. Rupprecht, and D. Cremers. Behind the scenes: Density fields for single view reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9076–9086, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 2, 3, 5, 6, 7, 4

[43] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8534–8543. Computer Vision Foundation / IEEE, 2021. 2, 3

[44] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016. 2

[45] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5732–5741. IEEE, 2021. 2

[46] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4578–4587. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 6, 7, 4

[47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[48] Kaichen Zhou, Lanqing Hong, Changhao Chen, Hang Xu, Chaoqiang Ye, Qingyong Hu, and Zhenguo Li. Devnet: Self-supervised monocular depth learning via density volume construction. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, page 125–142, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[49] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 5, 6, 7, 8

[50] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 2, 4, 6, 7, 5