# Learning multiplane images from single views with self-supervision

Supplementary Material

## Neural network architecture

The architecture of our neural network is illustrated in Fig. 1. We use intermediate depth supervision in a similar way as in [2], but regressing the depth with the AdaBins [1] strategy. From AdaBins, we use only the main idea of splitting the depth into a set of bins, where the final depth map is regressed with Equation (3) from [1], considering all bins as a *uniform grid*. Note that the intermediate depth supervision is used with the only purpose of helping the network to learn to split the scene into $D$ layers, which represent the depth bins in our intermediate supervision. The intermediate depth predictions are not used during inference. In our experiments, we use $D = 32$, in a similar way to [4].

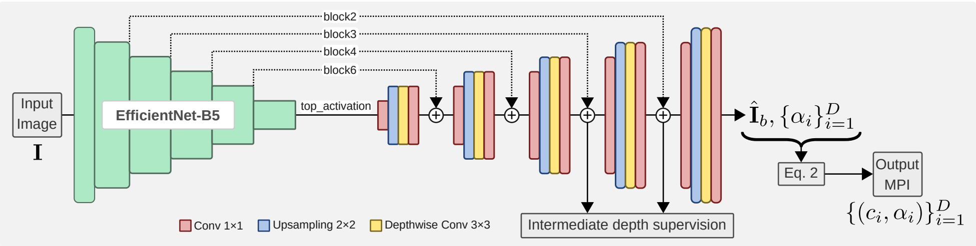


Figure 1: Network architecture used to implement the function $f_\theta$ in our method.

## Training details

In this part, we show some additional training details that could help in replicating our method. During our self-supervised training approach, we generated target viewpoints randomly. For this, we assume a random camera movement, considering pan and tilt with random values in the interval of $[-5, 5]$ degrees. We also generated camera translations considering random values in normalized coordinates in the interval of $[-0.4, 0.4]$ for $(x, y)$ coordinates (w.r.t. the image plane) and $[-0.1, 0.1]$ for $(z, )$ coordinate (movement perpendicular to the image plane). To illustrate this process, we included some samples from the training set of Places II dataset in Fig. 2.


Figure 2: Training samples from Places II dataset with our randomly generated target views.

# Additional ablation results

We show in Table 1 an extended version of Table 3 from the main paper. In this case, we present the loss coefficient values used in each experiment, with some additional training strategies. For instance, we trained our model only with depth supervision, with and without intermediate depth supervision (first two rows). We can see without intermediate depth supervision, our model has very high LPIPS metric, which means that the overall quality of the generated views are poor. Note that for Table 1 we trained our models for 500k iterations due to our limited computational resources.

From Table 2, we can also observe that with higher coefficient values in $\beta$ and $\gamma$, the SSIM and PSNR metrics decrease, but the LPIPS metric is improved. In a practical point of view, the general visual quality of the results with higher VGG and Style losses improves, but the more classical metrics (SSIM and PSNR) get worse. In this experiment, we trained our models longer, for about 5M iterations.

# Additional qualitative results

We provide a qualitative comparison between our method, Single-View Synthesis and 3D-Photography on Fig. 3. All the images shown are from the RealEstate10K test set considering source and target frames are 10 frames apart. It is important to stress that our model was trained on Places II using self-supervision from a single image, while Single-View Synthesis was trained with image pairs from RealEstate10K and 3D-Photograph uses multiple views during inference to estimate depth.

| Training strategy | | | | | | Validation on RE10K | | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{depth}$ | $\mathcal{L}_{pix}$ | $\mathcal{L}_{vgg}(\beta)$ | $\mathcal{L}_{style}(\gamma)$ | Inverse proj. | Cyclic | SSIM ↑ | PSNR ↑ | LPIPS ↓ |
| 1.0$^\star$ | | | | | | 0.750 | 17.153 | 0.357 |
| 1.0 | | | | | | 0.734 | 17.699 | 0.237 |
| | 1.0 | | | | | 0.758 | 19.349 | 0.280 |
| 1.0 | 1.0 | | | | | 0.760 | 19.473 | 0.215 |
| 1.0 | 10.0 | | | | | 0.802 | 20.341 | 0.265 |
| 1.0 | 1.0 | 0.01 | | | | 0.752 | 19.332 | 0.195 |
| 1.0 | 1.0 | 0.01 | 0.0001 | | | 0.735 | 18.748 | 0.183 |
| 1.0 | 1.0 | 0.01 | 0.0001 | ✓ | | 0.761 | 19.556 | 0.182 |
| 1.0 | 1.0 | 0.01 | 0.0001 | | ✓ | 0.765 | 19.773 | 0.182 |

Table 1: Ablation study considering different training strategies in our method. In $^\star$, intermediate depth supervision was not used during training.

| Training strategy | | | | | | Validation on RE10K | | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{depth}$ | $\mathcal{L}_{pix}$ | $\mathcal{L}_{vgg}(\beta)$ | $\mathcal{L}_{style}(\gamma)$ | Inverse proj. | Cyclic | SSIM ↑ | PSNR ↑ | LPIPS ↓ |
| 1.0 | 1.0 | 0.01 | 0.0001 | ✓ | | 0.786 | 19.960 | 0.176 |
| 1.0 | 1.0 | 0.01 | 0.0001 | | ✓ | 0.788 | 20.032 | 0.179 |
| 1.0 | 1.0 | 0.1 | 0.01 | | ✓ | 0.778 | 19.623 | 0.164 |

Table 2: Comparison of inverse projection and cyclic training, also considering different values for $\beta$ and $\gamma$.

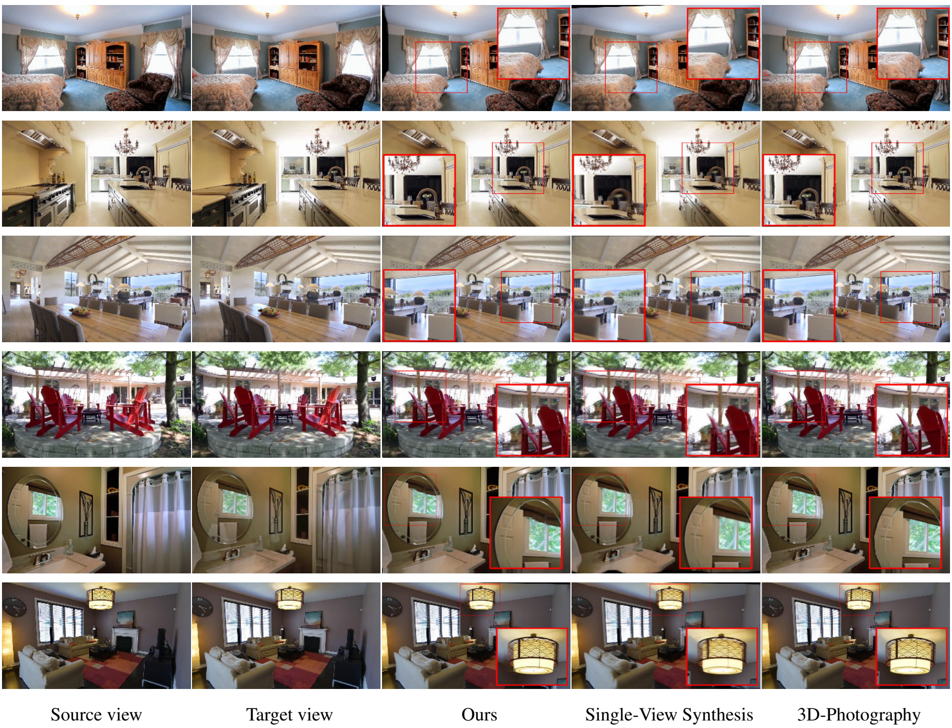


Figure 3: Qualitative results for our method compared with Single-View Synthesis [4] and 3D-Photography [3] on images from RealEstate10K.

In Fig. 4, we included one example of MPI generated by our method from Places II. Note that the produced MPI has learned transitions between layers, which helps our method to produce smooth transitions between different views, as can be also noticed from our demonstration videos.

# Misalignment problem on Mannequin Challenge

We show in Fig. 5 some examples of the the alignment problem between the target frame and predictions made by our method and Single-View Synthesis [4]. As one may notice, even though the predicts have good visual quality they do not align with the target frame provided by the Mannequin Challenge dataset.

# References


[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *arXiv preprint arXiv:2011.14141*, 2020.

[2] Diogo C Luvizon, Gustavo Sutter P Carvalho, Andreza A dos Santos, Jhonatas S Conceicao, Jose L Flores-Campana, Luis GL Decker, Marcos R Souza, Helio Pedrini, Antonio Joia, and Otavio AB Penatti. Adaptive multiplane image generation from a single internet picture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2556–2565, 2021.

[3] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

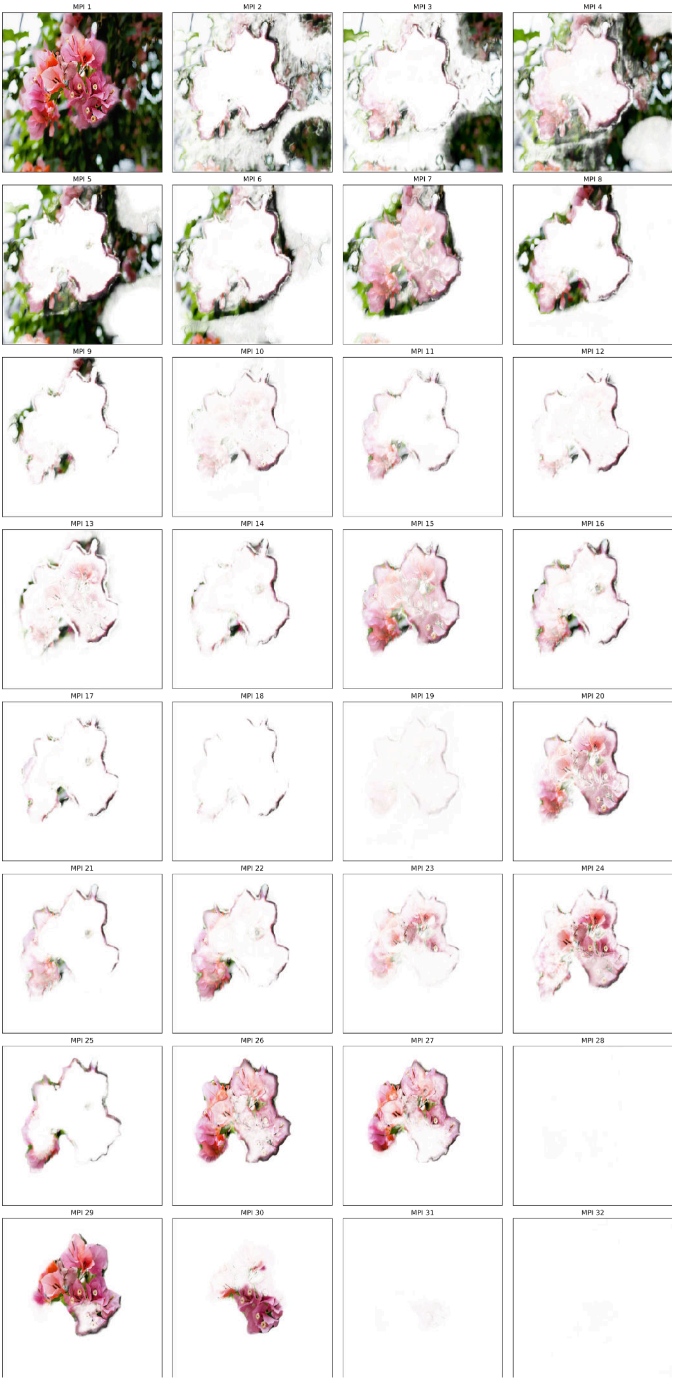


Figure 4: Sample of an MPI produced by our method with $D = 32$.

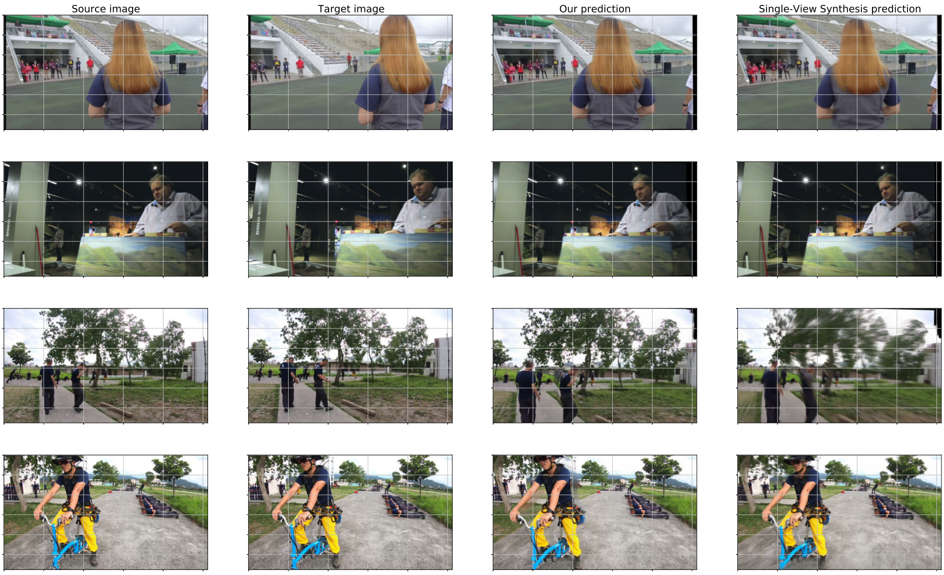


Figure 5: Examples of the misalignment problem on the Mannequin Challenge dataset. Grid lines facilitate to visualize that target and predictions are not correctly aligned.