Supplemantary Material: EpipolarNVS

Gaëtan Landreau[1,2]
gaetan.landreau@meero.com
Mohamed Tamaazousti [2]
mohamed.tamaazousti@cea.fr

[1] Meero
Paris, France

[2] Université Paris-Saclay,
CEA-LIST,
F-91120 Palaiseau, France

This document provides additional information regarding our EpipolarNVS contribution for single-image novel view synthesis.

A Extended encoding strategy

As explained in the core paper, we extended our original relative pose encoding strategy to better apprehend the specificity of KITTI [3] and Synthia [5] datasets.

As shown on the Figure 1, most of the car trajectories are made based on a straight path and only a few turns exist in the sequence **00**. In a similar fashion way, the sequence **01** is almost an end-to-end complete pure translation.

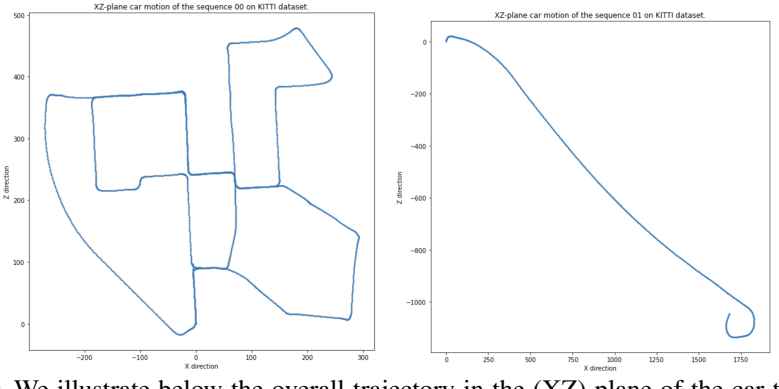


Figure 1: We illustrate below the overall trajectory in the (XZ) plane of the car that drove across German's street to acquire the 00 (left) and 01 (right) sequence of the KITTI[3] dataset.

B Spectral loss

Theoretical insights are presented in this section regarding the spectral loss function that was used for training in addition to the usual MAE.

As already mentioned in the main paper, such loss directly takes inspiration from [2], one of the latest state-of-the-art paper related to the super-resolution issue. Authors emphasise in their work on the fundamental role high frequencies have in the image generation process.

The Gaussian filter w_{gauss} we used is straightforwardly defined by a mean $\mu = \frac{k_s-1}{2}$ and a variance $\sigma = (\frac{k_s}{k_s+1})^2$, with k_s=5.

Such formulation allows to end up with the Spectral loss function that was introduced in the paper:

$$\mathcal{L}_{spectral} = ||I_t^{HF} - \hat{I}_t^{HF}||_2^2 \tag{1}$$

We therefore extensively focus through this loss on the highest frequencies of the target image, to enforce during training the network to retrieve as much as much as possible fine and complex structures.

C Experiments

C.1 Dataset characteristics

We first provide here some additional information regarding the different datasets we used for our experiments.

Regarding the ShapeNet dataset, we decided to not use the same dataset as [4, 6] but rather worked with the rendered ShapeNet images from DISN [7]. It offers at least three main improvements over the ones used in [4]:

- Intrinsic camera parameters are available.
- Each object within a class has 36 different views (against 18 for the dataset provided by [4]).
- Rendered images have a non null elevation angle and the azimuth one is sampled on a regular 10° basis. A random noise term is added on each rendered view to slightly jitter the camera pose.

Considering real-world Synthia [5] and KITTI [3] datasets, original images used in [4] also only contain extrinsic matrices, leaving apart the intrinsic information our architecture requires. We, therefore, build up our own train/test sets, with the same scenes as the ones used in [4]. Images were resized to 256×256 for speed-up and convenience purposes and ground-truth intrinsic matrices were adjusted accordingly. Images we work with on these real scene scenarios are more challenging than the ones used in [4][6] since dealing with center-cropped images discards fast-moving elements (on image borders) from the scenes.

Finally, we get consideration for the same setting used in [4] for the maximal latitude between the source and target view: full $\pm 180°$ azimuthal range is permitted for the ShapeNet classes while a maximum of ± 10 frames are considered for the real world datasets Synthia[5] and KITTI [3].

C.2 Ablation studies

We conduct ablation studies to highlight and understand how some meaningful properties of our encoding strategy behave.

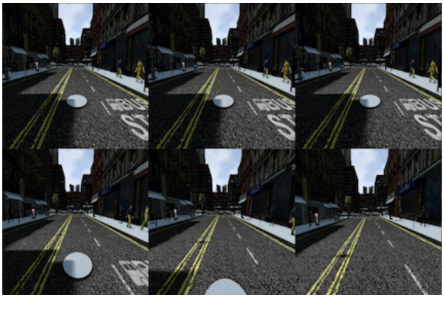
Benefit of the extended encoded pose strategy

Dealing with real-world scene datasets and the maximum 10 frames difference that could occurred between the source and the target view is one of the most tricky scenarios for single-image novel view synthesis. We thus conduct a first ablation study to validate the intuition behind this additional channel that encodes the relative largest motion in the (XZ) plane. Please note that the Spectral loss has not been added in this ablation study, leading to slightly different results from the original ones reported in the main paper.

Method	Synthia		KITTI	
	MAE (↓)	SSIM (↑)	*MAE* (↓)	SSIM (↑)
Encoded pose	0.077	0.602	0.109	0.576
Extended encoded pose	**0.066**	**0.622**	**0.086**	**0.605**

Table 1: Benefit on Synthia [5] and KITTI [3] datasets of our extended encoding strategy. Adding such fourth channel helps the network to better perform on real-world datasets. The grid $\mathbf{G}_{15}$ has been used in both cases.

As shown in Table 1, and considering the neural network architecture as fixed, the fourth channel we have added to our representation $E_{s\to t}$ clearly helps the network to perform better in the task it has been trained for. The SSIM gained more than 2 points on average while the MAE significantly decreased (by almost 20% on average for the two datasets) to reach competitive results with [4].



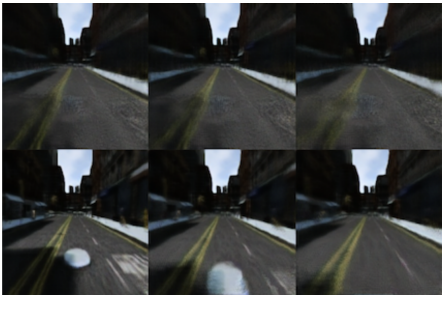


Figure 2: Source fixed view I_s (top row, Left) and three consecutive target views (bottom row, Left) from Synthia [5] test set. - Predictions made by our model with the encoded pose (top row,Right) and the extended encoded pose strategy (bottom row,Right). Adding an additional channel in our extended pose encoding allows the network to better apprehend the motion that occurred in Synthia [5] and KITTI[3]. The grid $\mathbf{G}_{15}$ has been used in both cases.

We highlight on Figure 2 the positive influence this last channel has on the our model. While the manhole cover (disappearing on the target views sequence) is entirely discarded by the network trained with the 3 channel pose encoding representation, the extended version

we proposed managed to grasp the car's motion.

Spectral loss influence

A second ablation study has been conducted to highlight to which extent the spectral loss function positively impacts the training of our model architecture.

Datasets	Metrics	$\mathcal{L}_1$ only	$\mathcal{L}_1 + \mathcal{L}_{spectral}$
ShapeNet - Car	L1 (↓)	0.019	**0.016**
	SSIM (↑)	0.912	**0.928**
	PSNR (↑)	22.61	**24.23**
ShapeNet - Chair	L1 (↓)	0.037	**0.032**
	SSIM (↑)	0.892	**0.901**
	PSNR (↑)	19.19	**19.55**
Synthia	L1 (↓)	0.066	**0.065**
	SSIM (↑)	0.622	**0.631**
	PSNR (↑)	19.24	**19.44**
KITTI	L1 (↓)	0.086	**0.082**
	SSIM (↑)	0.605	**0.609**
	PSNR (↑)	16.99	**17.11**

Table 2: Impact of the Spectral loss function.

As seen in Table 2, constraining the training on the high frequencies helps the network to generate more realistic novel views. Such quantitative improvement is visually confirmed in Figure 3 where the same object instance is generated through both configurations with the ShapeNet *Car* class.

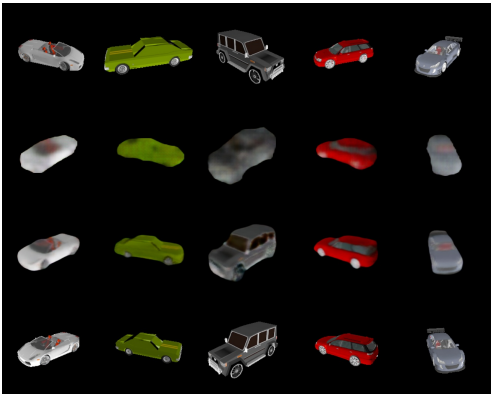


Figure 3: Inference results from the ShapeNet [1] *Car* test set. From the top row to the bottom one: Source images I_s, Ours prediction with $\mathcal{L}_1$ only at training, Ours prediction with $\mathcal{L}_{spectral} + \mathcal{L}_1$ at training, Ground truth - Target images I_t. From a general perspective, tires and windows are better retrieved at inference time when high frequencies have been constrained during training.

Discrete grid $\mathbf{G}_r$ granularity and sampling strategy

We present a third ablation study to get some knowledge regarding the granularity our grid sampling needs. Beyond the three grids we tested out, we also consider a random sampling strategy, that consists of sampling a fixed number (corresponding to 1% of pixels for a 256×256 image) of locations. Table 3 summarises the different results of this experiment on the real-world datasets.

Method	Synthia		KITTI	
	MAE (↓)	SSIM (↑)	*MAE* (↓)	SSIM (↑)
Random Sampling (655 pix. sampled)	**0.0816**	**0.593**	0.1290	0.549
$\mathbf{G}_{15}$ grid (225 pix. sampled)	0.0823	0.589	0.1222	0.562
$\mathbf{G}_{20}$ grid (400 pix. sampled)	0.0857	0.576	0.1241	0.560
$\mathbf{G}_{25}$ grid (625 pix. sampled)	0.0908	0.575	**0.1217**	**0.563**

Table 3: Sampling strategy influence over the real-world datasets Synthia [5] and KITTI [3].

Overall, there are no significant differences between the strategies that were tested in this ablation study. However, the random and the grid sampling strategy differ in an important aspect: the latter performs significantly faster and roughly takes 4 times less time to form a batch of triplets $(I_s, I_t, E_{s\to t})$ than the random sampling strategy. Using a regular grid $\mathbf{G}_r$ always use the same pixel locations from I_s to build $E_{s\to t}$ while the random sampling strategy imposes to picked up new samples all the time.

We performed the same ablation study on the synthetic ShapeNet dataset [1] and drew identical observations.

C.3 Inference results

We finally present in this last part additional qualitative results from our model, on all the four datasets we have get consideration for in this study. We present for each dataset 8 different scenes or object instances.

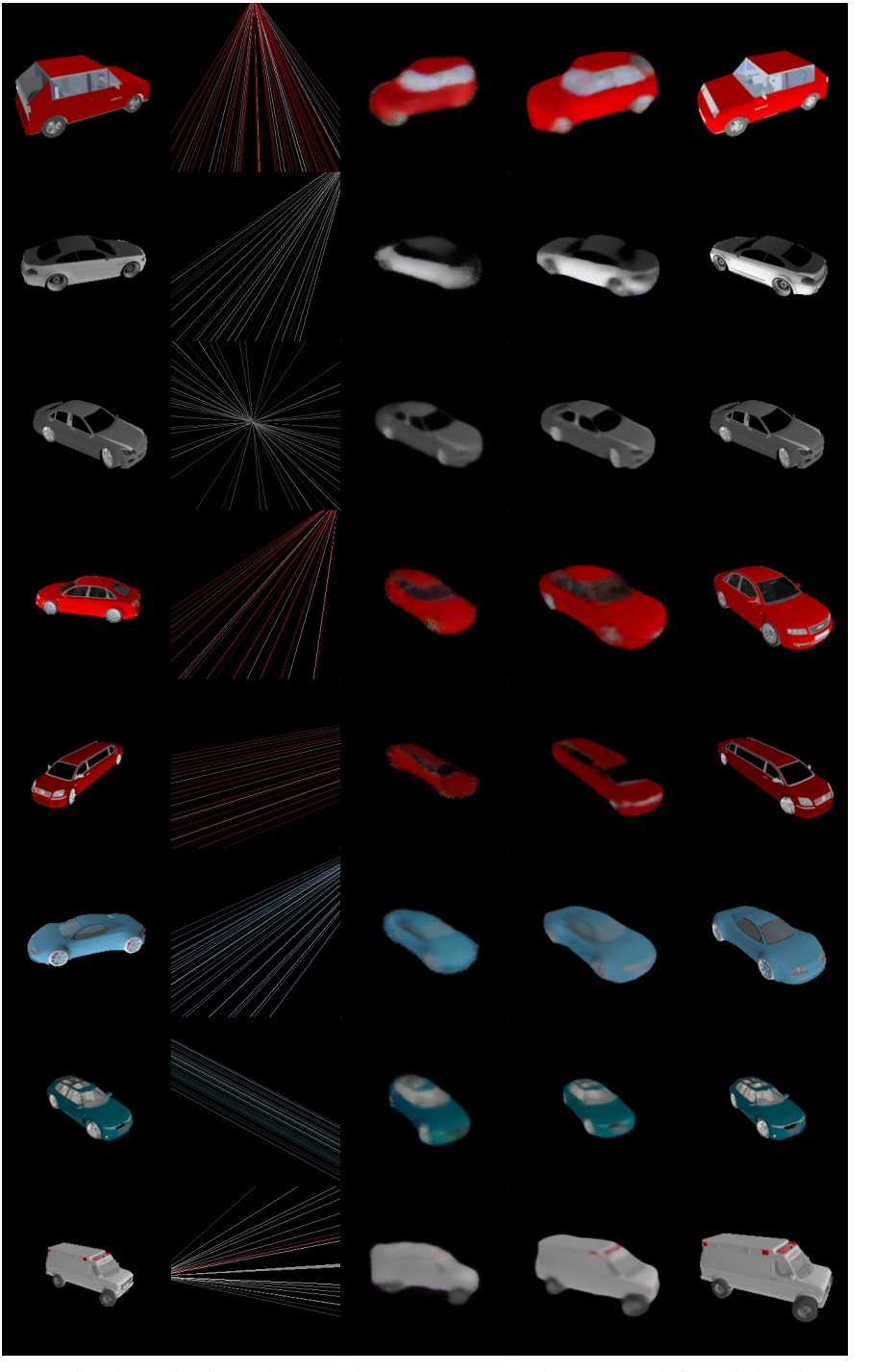


Figure 4: Visual results from the test ShapeNet [1] *Car* class. From left to right: the source image I_s, the Encoded Pose $E_{s \to t}$, the prediction of [2], our prediction and the target image I_t.

Figure 5: Visual results from the test ShapeNet [1] *Chair* class. From left to right: the source image I_s, the Encoded Pose $E_{s\rightarrow t}$, the prediction of [4], our prediction and the target image I_t.

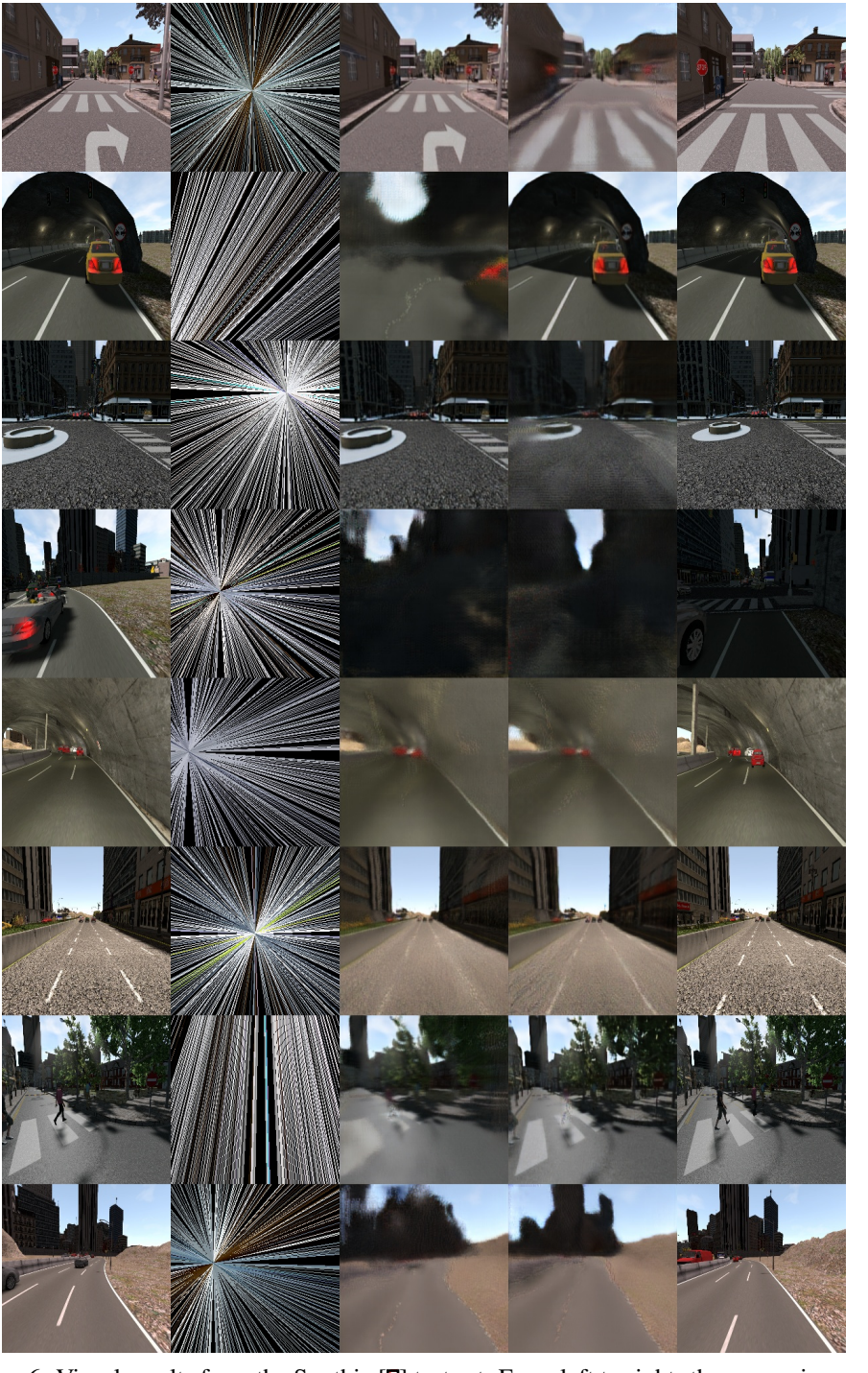


Figure 6: Visual results from the Synthia [5] test set. From left to right: the source image I_s, the Encoded Pose $E_{s\rightarrow t}$, the prediction of [4], our prediction and the target image I_t.

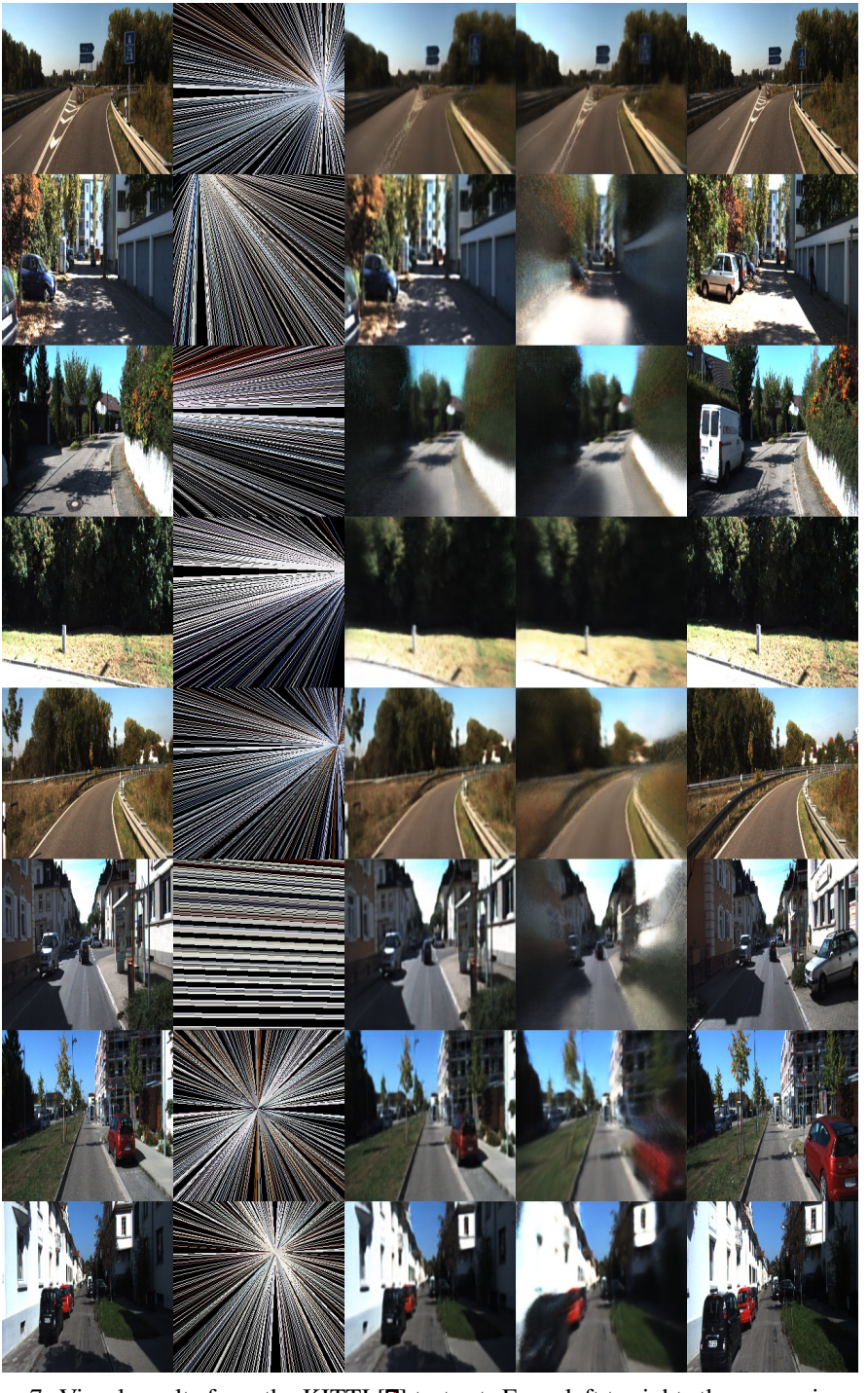


Figure 7: Visual results from the KITTI [3] test set. From left to right: the source image I_s, the Encoded Pose $E_{s \to t}$, the prediction of [4], our prediction and the target image I_t.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, and al. Shapenet: An information-rich 3d model repository. *arXiv*, 2015.

[2] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCV*, 2019.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[4] Kim Juhyeon and Young Min Kim. Novel view synthesis with skip connections. In *ICIP*, 2020.

[5] German Ros, Laura Sellart, Joanna Materzynska, and & al. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

[6] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *ECCV*, 2018.

[7] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.