Geometry-aware Deep Network for Single-Image Novel View Synthesis —Supplementary Material—

Miaomiao Liu
CECS, ANU
Canberra, Australia
Miaomiao.Liu@anu.edu.au

Xuming He
ShanghaiTech University
Shanghai, China
hexm@shanghaitech.edu.cn

Mathieu Salzmann
CVLAB, EPFL
Lausanne, Switzerland
mathieu.salzmann@epfl.ch

1. Sherman-Morrison formula

We first provide more detail for the Sherman-Morrison formula, which allows us to explicitly compute the inverse of homographies. The Sherman-Morrison formula can be stated as follows:

Theorem 1 *Assume* $\mathbf{A}$ *is invertible, and* $\mathbf{u}$ *and* $\mathbf{v}$ *are column vectors. Furthermore, assume* $1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{u} \neq 0$. *Given*

$$\mathbf{B} = \mathbf{A} + \mathbf{u}\mathbf{v}^{\mathrm{T}}, \tag{1}$$

the inverse $\mathbf{B}^{-1}$ *can be obtained as*

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{u}}. \tag{2}$$

In our context, the homography is defined as

$$\mathbf{H} = \mathbf{K}(\mathbf{R} - \mathbf{t}\tilde{\mathbf{n}}^{\mathrm{T}})\mathbf{K}^{-1}.$$

Let us first ignore $\mathbf{K}$ and concentrate on the central part

$$\tilde{\mathbf{H}} = \mathbf{R} - \mathbf{t}\tilde{\mathbf{n}}^{\mathrm{T}}\,, \tag{3}$$

where $\mathbf{R}$ is a rotation matrix and is thus invertible, i.e., $\mathbf{R}^{-1} = \mathbf{R}^{\mathrm{T}}$. Therefore, Eq. 3 satisfies the conditions of $\mathbf{B}$ in Eq. 1, and the inverse $\tilde{\mathbf{H}}^{-1}$ can be written as

$$\tilde{\mathbf{H}}^{-1} = \mathbf{R}^{\mathrm{T}} + \frac{\mathbf{R}^{\mathrm{T}}\mathbf{t}\tilde{\mathbf{n}}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}}{1 - \tilde{\mathbf{n}}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{t}}\,.$$

Re-introducing $\mathbf{K}$, and following the standard rule for matrix product inversion, lets us write the inverse $\mathbf{H}^{-1}$ as

$$\mathbf{H}^{-1} = \mathbf{K}\tilde{\mathbf{H}}^{-1}\mathbf{K}^{-1}.$$

2. Experiments

In this section, we provide additional results on the two datasets. We further illustrate the $m = 16$ synthesized images obtained from the homographies generated by our method, and show additional examples of the selection maps our network predicts. We then provide the visualisation of our estimated depth and normal maps for KITTI dataset and discuss failure cases of our approach.

2.1. Additional Results

We provide additional qualitative results on the KITTI dataset in Figs. 1, 2 and 3, and on the ScanNet dataset in Figs. 4 and 5. As those in the main paper, they clearly illustrate the benefits of our approach over the state-of-the-art appearance flow baseline [1]; specifically, accounting for geometry lets us produce much more realistic novel views. Note also that our complete approach (Ours-Full), with the refinement network, typically yields sharper results than our basic framework without refinement (Ours-Geo). This can be seen, e.g., in the third to seventh rows of Fig. 1.

In Table 1, we analyze the influence of the quality of the depth and normal estimates and of learning the selection maps on ScanNet. Note that, compared to Table 2 in the main paper which shows a similar analysis for KITTI, we eliminated the factor 'gtNor' because it is computed from 'gtDep'. In essence, the behavior is the same as for KITTI. The best results are obtained with the ground-truth depth maps, which leaves room for our method to improve as progress in depth estimation is made. More importantly, our learnt selection maps give a significant boost to our results, whether using ground-truth depth or estimated one.

2.2. Synthesized Candidate Images

In Fig. 6, we show the synthesized images obtained from our $m = 16$ predicted homographies for one input image. When compared with the ground-truth novel view, we can see that different homographies account for the motion of different regions in the image. For instance, the homography corresponding to the top-left image accounts for the motion of the road. By contrast, the homography corresponding to the bottom-right image models the motion of the buildings. Correctly combining these images then allows us to obtain a realistic novel view, as shown in the top row of Fig. 6.

gtDep	estDep	estNor	Seed	SelMap	ℓ_1
✓	✗	✓	✓	✗	0.174
✓	✗	✓	✗	✓	0.159
✗	✓	✓	✓	✗	0.184
✗	✓	✓	✗	✓	0.167

Table 1. **Influence of the quality of the depth and normal estimates and of learning the selection maps on ScanNet.** From left to right: gtDep denotes the ground-truth depth; estDep and estNor denote the estimated depth and normal, respectively; Seed and SelMap denote the hard-segmentations corresponding to the seed regions and the selection map obtained with our selection network, respectively.

2.3. Selection Maps

In Fig. 7, we provide additional results from our selection network. While our seed regions typically cover only parts of the road, trees, sky, and buildings, our predicted selection maps can extend them to larger planar and semantically-coherent regions.

2.4. Depth and Normal prediction

In Fig. 8, we provide the visualisation of the estimated depth and normal map from our network for sampled images from KITTI test set. It shows that our estimation can well capture the scene structure compared with the ground truth.

2.5. Failure Cases

In Fig. 9, we show typical failure cases of our approach. The failure cases are mainly due to i) moving objects, whose locations cannot be explained by camera motion (see the first row); 2) the need to hallucinate large portions of the image (e.g., because of backward motion), in which case our method tends to generate background and miss foreground objects (see the last two examples).

References

[1] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301. Springer, 2016. 1, 3, 4, 5

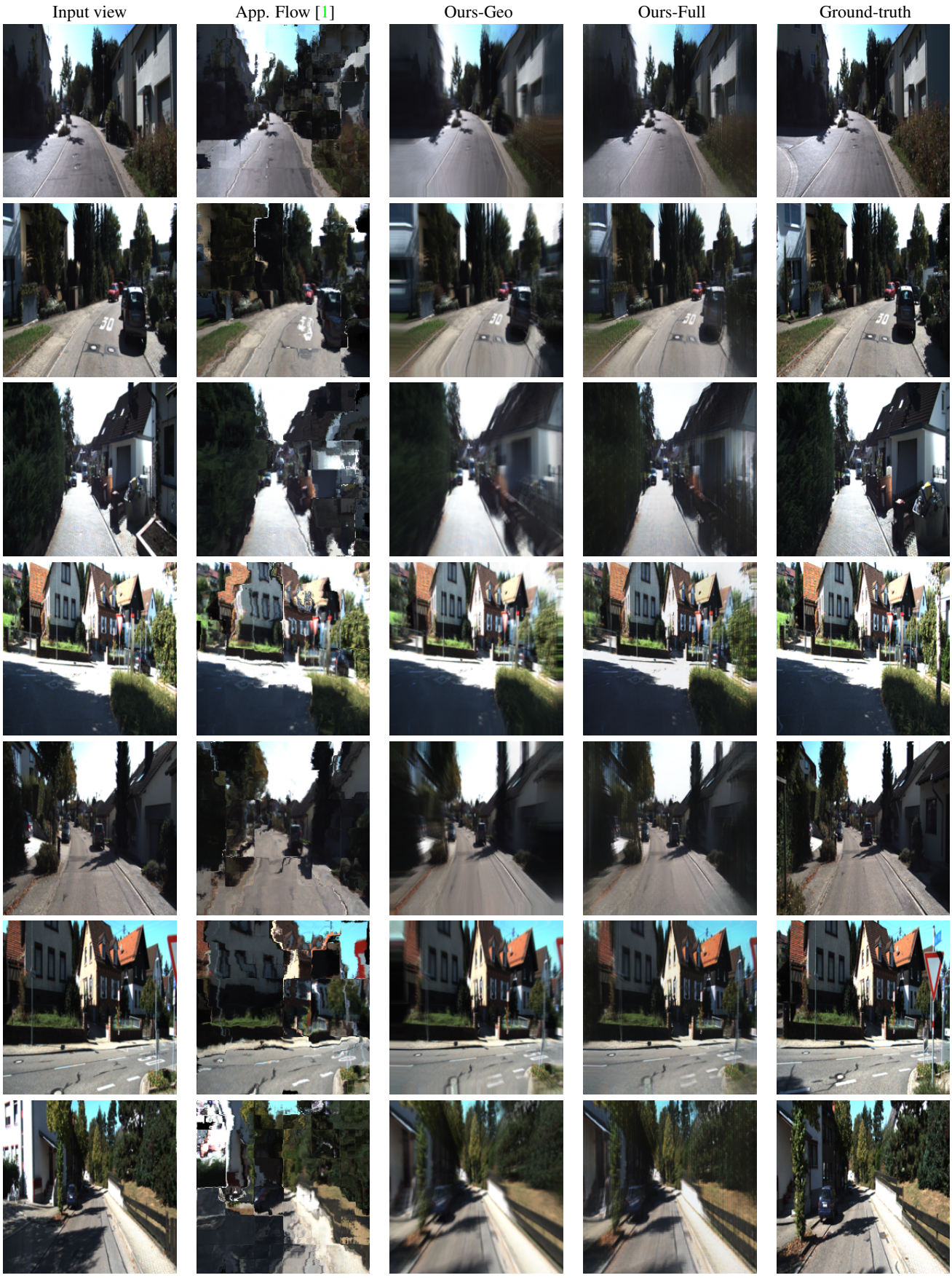

Figure 1. **Qualitative comparison of our approach with the appearance flow method of [1] on KITTI.** While appearance flow yields artifacts, our approach, which reasons about 3D geometry, yields more realistic results.

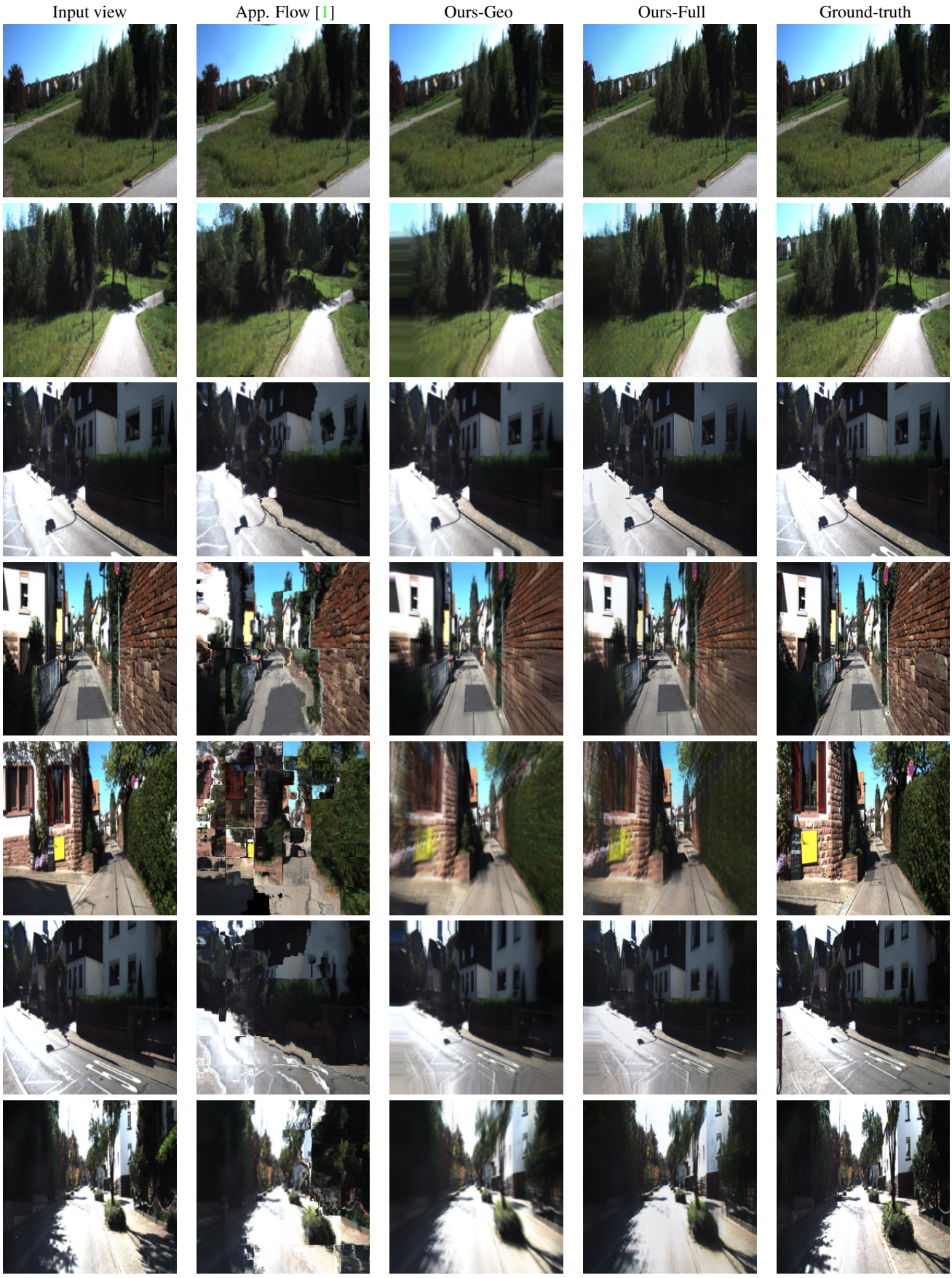

Figure 2. **Qualitative comparison of our approach with the appearance flow method of [1] on KITTI.** While appearance flow yields artifacts, our approach, which reasons about 3D geometry, yields more realistic results.

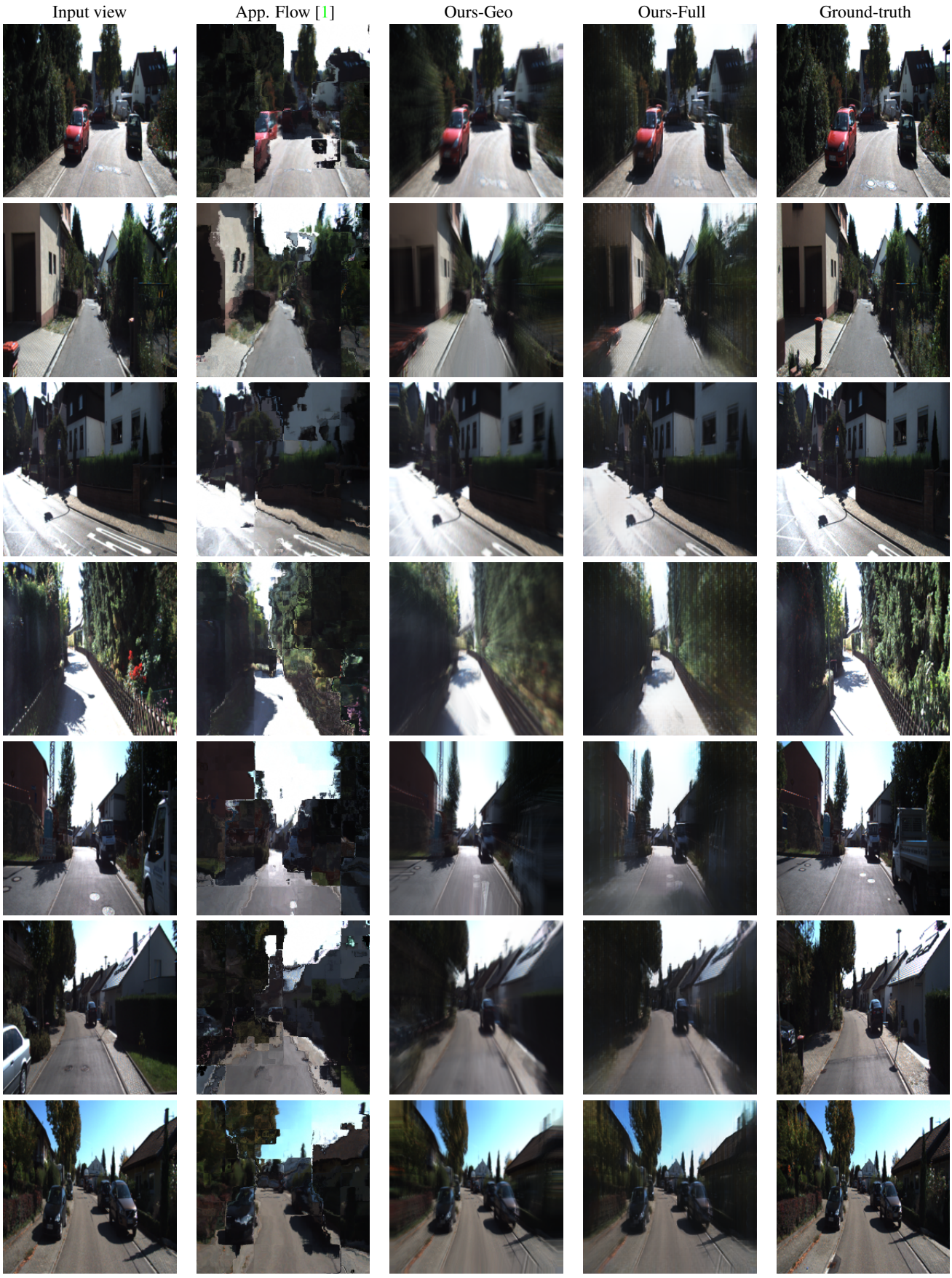

Figure 3. **Qualitative comparison of our approach with the appearance flow method of [1] on KITTI.** While appearance flow yields artifacts, our approach, which reasons about 3D geometry, yields more realistic results.

Figure 4. **Qualitative results of our approach on ScanNet.**

Figure 5. **Qualitative results of our approach on ScanNet.**

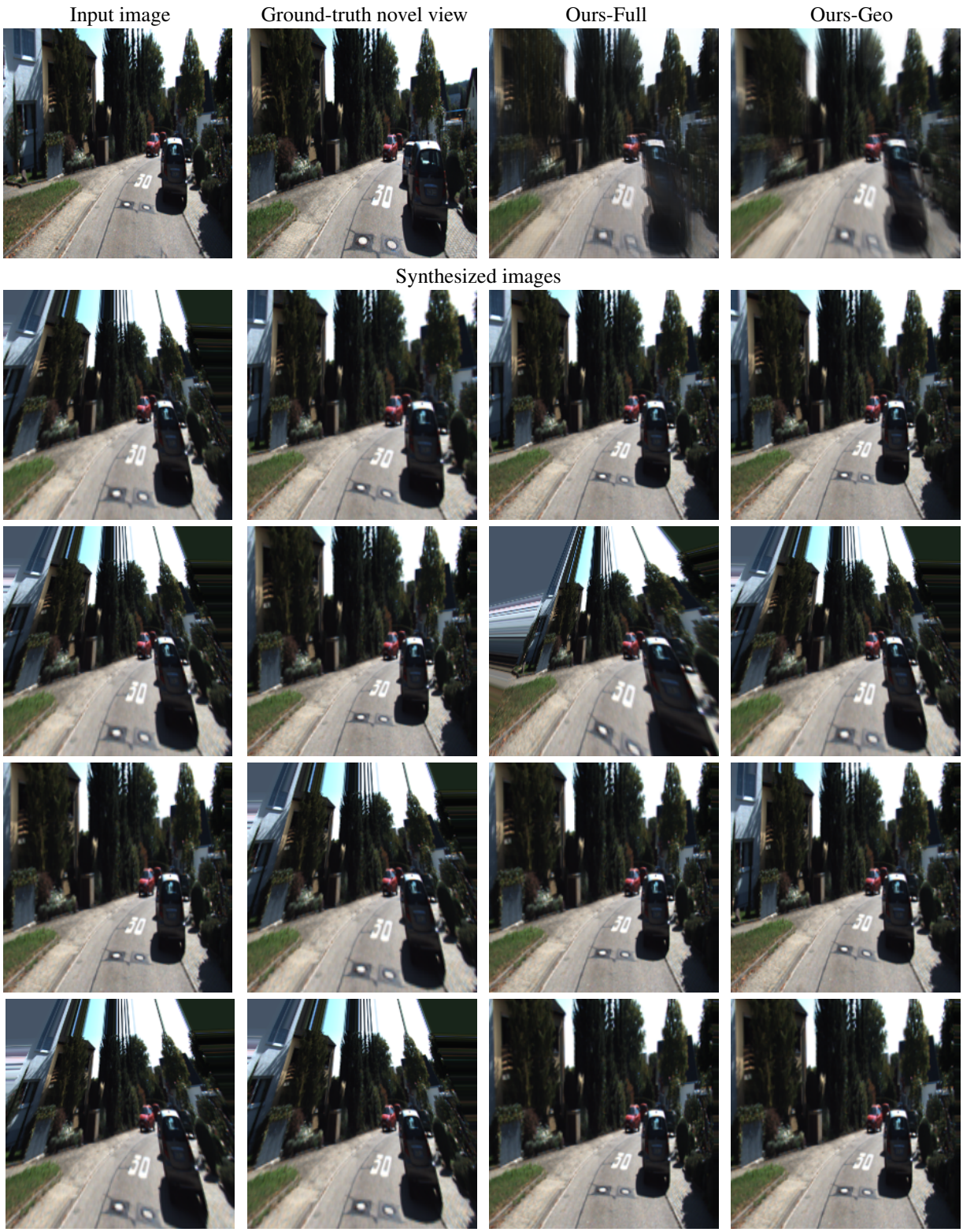

Figure 6. **Synthesized images from the 16 estimated homographies.** In the top row, we show the input image, the ground-truth novel view and the results of our complete model (Ours-Full) and of our model without refinement (Ours-Geo). The remaining images correspond to images synthesized with our estimated homographies. Note that different homographies correctly account for the motion of different regions between the input and novel view. For instance, the top-left image models the motion of the road, while the bottom-right one accounts for the motion of the buildings.

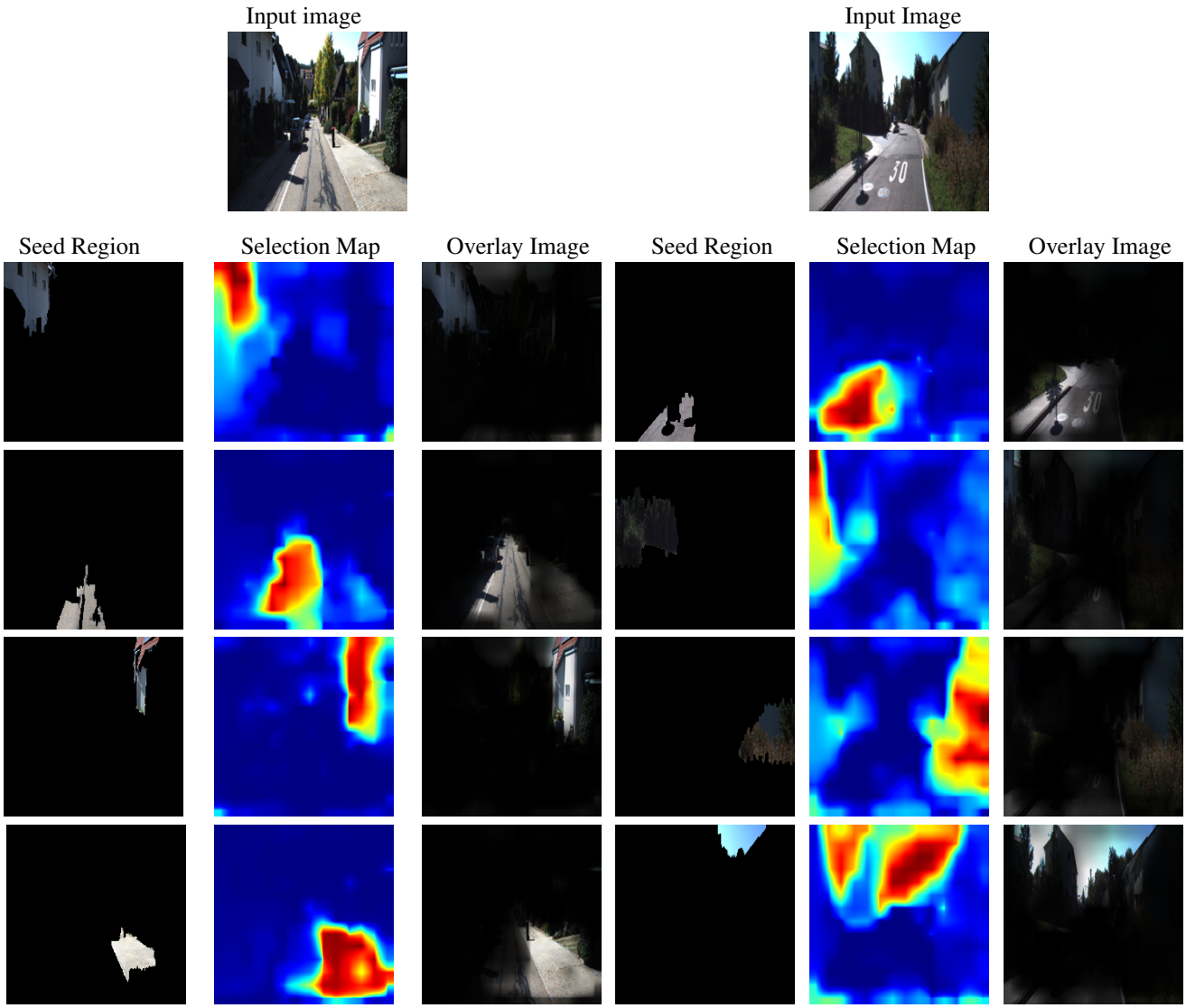

Figure 7. **Sample seed regions and predicted selection maps in the input view.** From left to right: seed region, predicted selection map and predicted selection map overlaid on the input image. Red indicates a high likelihood for a pixel to belong to the plane defined by the seed region and blue a low likelihood.

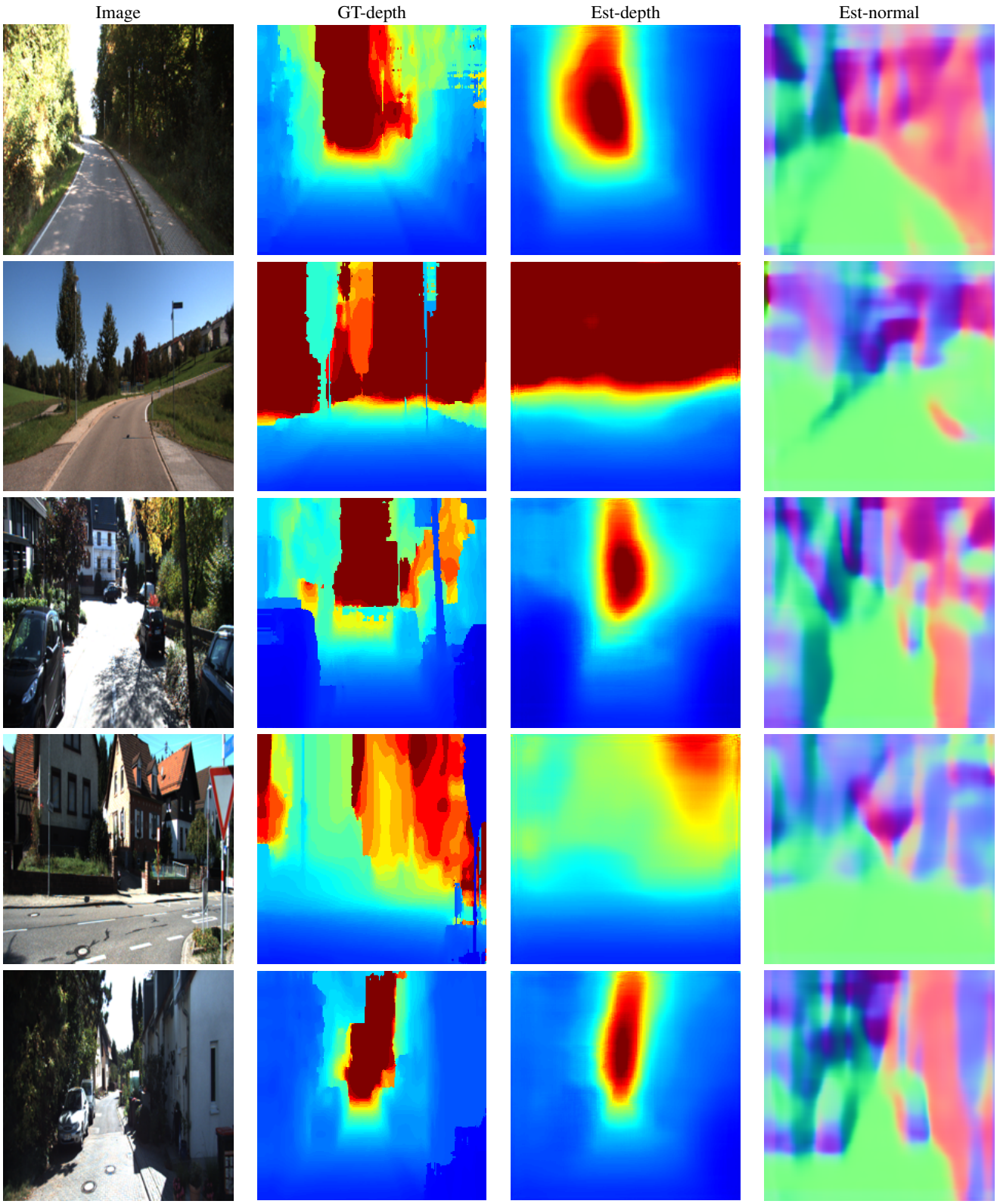

Figure 8. **Visualization of the estimated depth and normal for KITTI.** Color indicates depth (red is far, blure is close).

Figure 9. **Failure cases of our approach on KITTI.** Typical failures correspond to moving objects, or hallucination of large portions of the image (e.g., due to backward motion), in which case our approach tends to generate background instead of foreground objects.