Supplementary Material: Displets: Resolving Stereo Ambiguities using Object Knowledge

Fatma Güney
MPI Tübingen
fatma.guney@tue.mpg.de

Andreas Geiger
MPI Tübingen
andreas.geiger@tue.mpg.de

Abstract

In this **supplementary document**, *we first show how disparities relate to the 3D planes used in our representation. Next, we provide an illustration of the sampled displets for a single test image and visualize the influence of the displets with respect to the associated superpixels. For reproducibility, we also list the learned parameter settings which we used throughout all our experiments and show the change in performance with respect to these parameters. Finally, we show additional qualitative results on KITTI validation set. On our project website*[1] *we also provide a* **supplementary video** *demonstrating the semi-convex hull optimization process which we use in order to simplify the CAD models.*

1. Plane Representation

In our formulation, each superpixel is represented as a random variable $\mathbf{n}_i \in \mathbb{R}^3$ describing a plane in 3D ($\mathbf{n}_i^T\mathbf{x} = 1$ for $\mathbf{x} \in \mathbb{R}^3$ on the plane). We denote the disparity of plane $\mathbf{n}_i$ at pixel $\mathbf{p} = (u, v)^T$ by $\omega(\mathbf{n}_i, \mathbf{p})$. In the following, we show how $\omega(\mathbf{n}_i, \mathbf{p})$ can be derived for a (rectified) pinhole stereo camera with known intrinsics and extrinsics:

$$\begin{aligned} u &= \frac{fx}{z} + c_u, \quad x = (u - c_u)\frac{z}{f} \\ v &= \frac{fy}{z} + c_v, \quad y = (v - c_v)\frac{z}{f} \\ d &= \frac{fL}{z} \end{aligned}$$

$$\begin{aligned} 1 &= x\mathbf{n}_x + y\mathbf{n}_y + z\mathbf{n}_z \\ 1 &= (u - c_u)\frac{z}{f}\mathbf{n}_x + (v - c_v)\frac{z}{f}\mathbf{n}_y + z\mathbf{n}_z \\ \frac{1}{z} &= (u - c_u)\frac{\mathbf{n}_x}{f} + (v - c_v)\frac{\mathbf{n}_y}{f} + \mathbf{n}_z \end{aligned}$$

$$\begin{aligned} d &= (u - c_u)L\mathbf{n}_x + (v - c_v)L\mathbf{n}_y + fL\mathbf{n}_z \\ d &= L\mathbf{n}_x u + L\mathbf{n}_y v + fL\mathbf{n}_z - c_u L\mathbf{n}_x - c_v L\mathbf{n}_y \quad (1) \\ d &= au + bv + c \quad (2) \end{aligned}$$

Here, f, c_u, c_v denote the intrinsic camera calibration parameters and L is the baseline.

2. Illustration of Displets

Fig. 1 visualizes a subset of the sampled displets for a single test image which serve as input to our CRF. Wrong displet proposals are eliminated during inference if they don't agree with the observations or overlap with over displets.

[1] http://www.cvlibs.net/projects/displets/

Figure 1: **Illustration of Displets.** This figure shows a subset of sampled displets for an image by overlaying the corresponding disparity maps on top of the image regions they have been proposed for.

3. Displet Parameters

The influence of a displet on its associated superpixels is determined by a penalty function λ_{ki}. We take λ_{ki} as a weighted sigmoid function of the distance transform

$$\lambda_{ki} = \lambda_\theta \frac{1}{1 + \exp(\lambda_a - \lambda_b \, \mathrm{DT}_{ki})}$$

where DT_{ki} denotes the Euclidean distance transform of superpixel i with respect to the boundary of displet k. The model parameters λ_θ, λ_a and λ_b are learned from training data as described in the next section. In Fig. 2, we visualize λ_{ki} for a couple of displets and the associated superpixels. Less transparency indicates a higher penalty, e.g., we allow more deviation at the displet boundaries. In contrast, λ_{ki} takes a very large value (λ_θ) in the center of the displet (see Section 4), effectively approximating a hard constraint ($\lambda_{ki} = \infty$). Furthermore, we take the fitness score κ_k as the displet log likelihood ($-E_{\hat{\Omega}}$) in order to increase the plausibility of high quality displets. In practice, we subtract the largest fitness score of all displets originating from the same object proposal region $\mathcal{O}$ in order to calibrate for the scores of different regions.

4. Parameter Settings

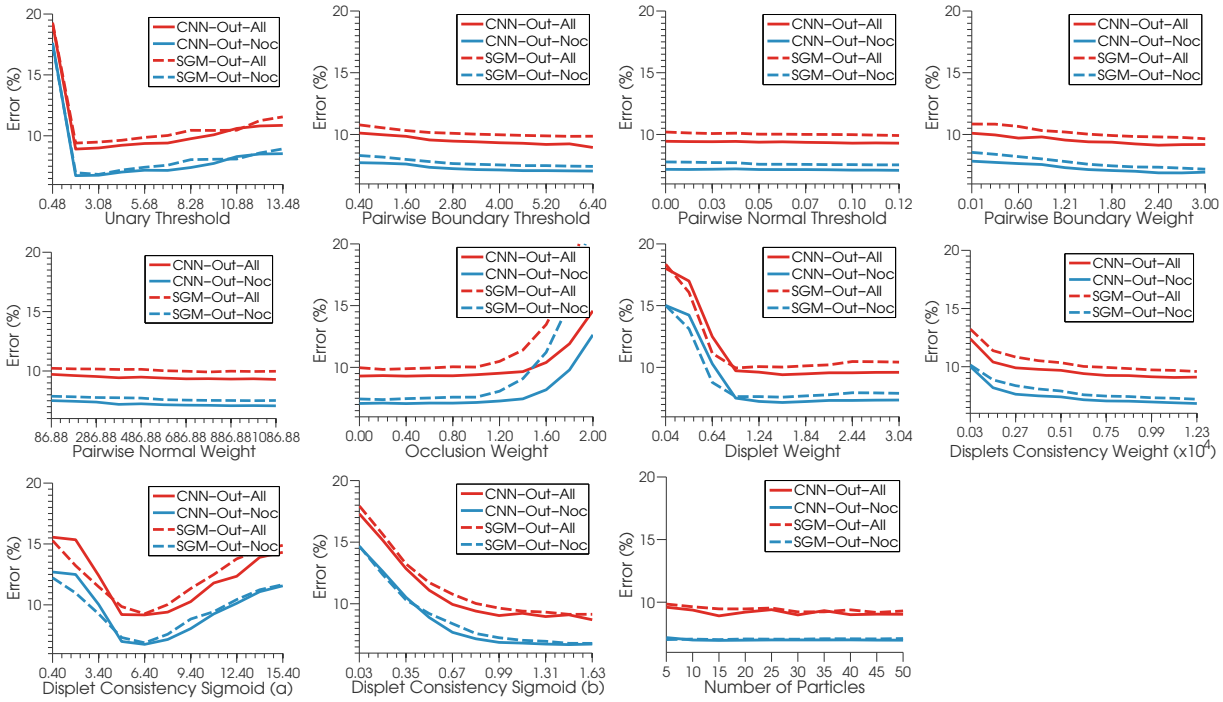
Table 1 lists the parameter values we used throughout our experiments. The parameters in the left table are obtained using block coordinate descent on 50 randomly selected images from the KITTI training set as described in the paper submission. The values of the parameters in the right table have been determined empirically. Furthermore, Fig. 3 shows the sensitivity when varying the main parameters in our model for both CNN and SGM features. For this experiment, we vary one parameter at a time while setting all other parameters to their optimal values listed in Table 1. We observe that our method is not very sensitive with respect to small parameter changes for most of the parameters. The performance in reflective regions is highly dependent on the displet parameters as the induced sigmoid function defines the contribution of the inferred displets to the individual superpixels.

Figure 2: **Illustration of Displet Influence λ_{ki}.** We penalize superpixels which do not agree ($\mathbf{n}_i \neq \hat{\mathbf{n}}_{k,z_i}$) with active displets ($d_k = 1$). Less transparency indicates a higher penalty, e.g., we allow more deviation at the displet boundaries.

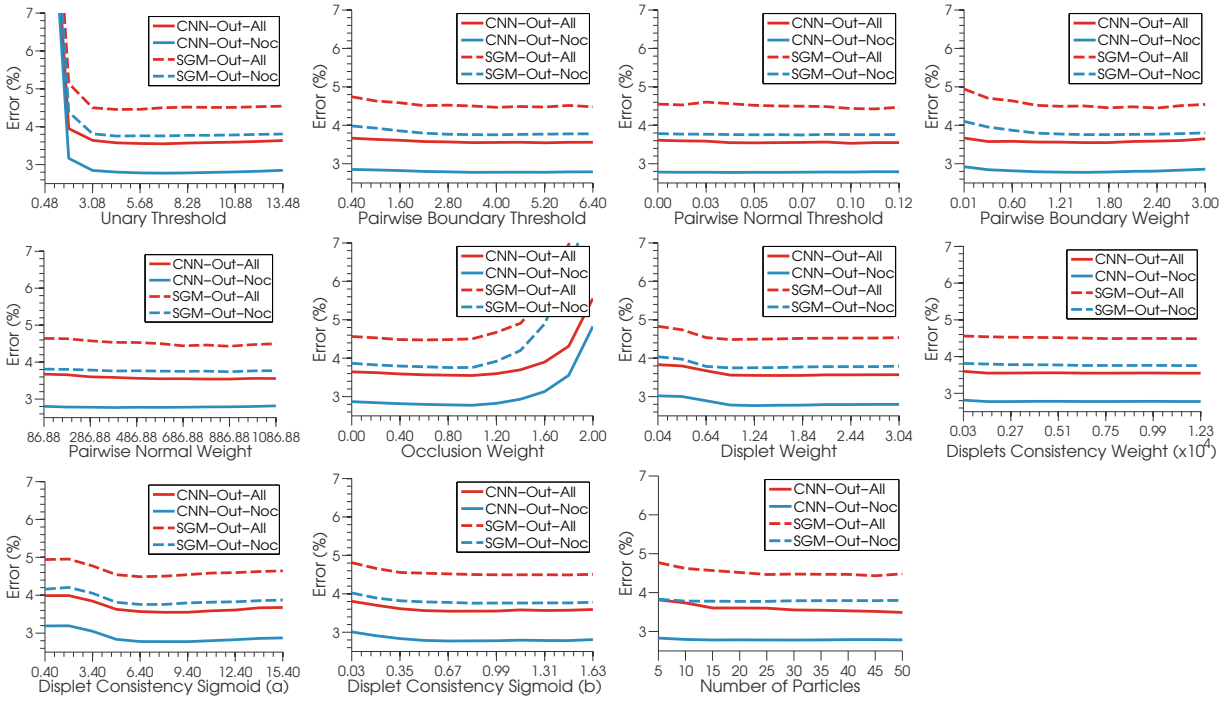
Parameter	Value
Unary Threshold (τ_1)	6.98
Pairwise Boundary Threshold (τ_2)	3.40
Pairwise Normal Threshold (τ_3)	0.06
Pairwise Boundary Weight (θ_1)	1.50
Pairwise Normal Weight (θ_2)	586.87
Displet Weight (θ_3)	1.53
Occlusion Weight	1.00
Displet Consistency Weight (λ_θ)	6×10^4
Displet Consistency Sigmoid (λ_a, λ_b)	[7.90 0.82]

Parameter	Value
Number of Particles	30
Number of Superpixels	1000
Number of Iterations	40
TRWS - Number of Iterations	200
Particle Standard Deviations	[0.5 0.5 5]

Table 1: **Model Parameters.** Here, we show the values of the learned parameters in our model.



(a) **Reflective Regions.** See text for details.



(b) **All Regions.** See text for details.

Figure 3: **Sensitivity to Parameters.** This figure shows the change in performance while varying one parameter at a time and keeping all others fixed at their optimal values. Errors are with respect to an error threshold of 3 pixels.

5. Results with Degraded Input Disparity Maps

As our sampling procedure is stochastic, it can handle degraded input disparity maps to some extend. Fig. 4 illustrates this fact on very challenging scenes for which the initial disparity maps contain large number of outliers.

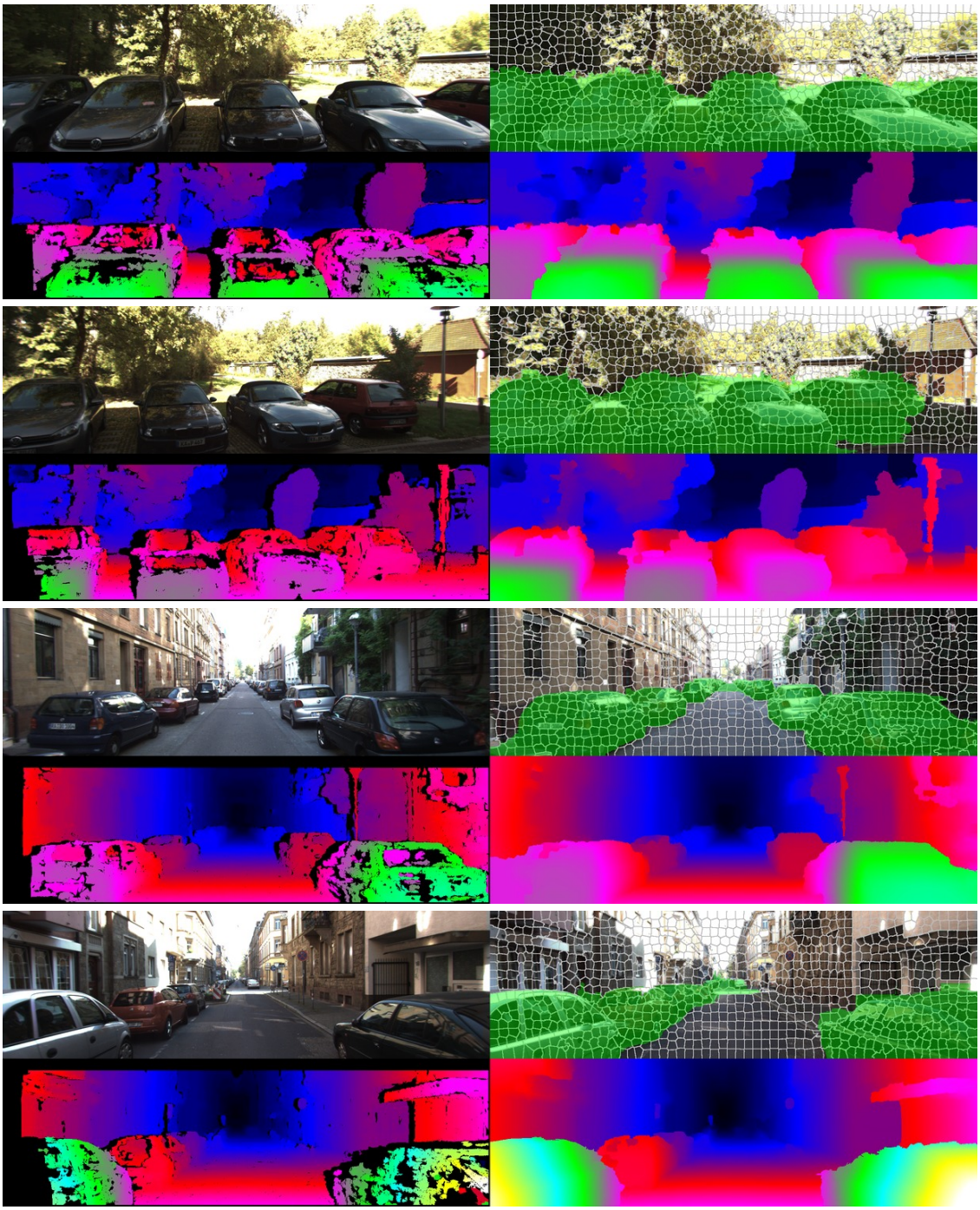


Figure 4: **Effect of Input Maps on Results.** Each subfigure shows: The input image, the superpixels and the semantic segments (first row), the input disparity maps, and our result by using these as input (second row).

6. Additional Qualitative Results

In this section, we show additional qualitative results on *randomly* sampled images from KITTI validation set. For each scene, we show the input image, the superpixels and semantic segments which our method takes as input as well as our inference results without and with displets and the corresponding error images, encoded from black (≤ 1 pixel error) to white (≥ 5 pixels error).

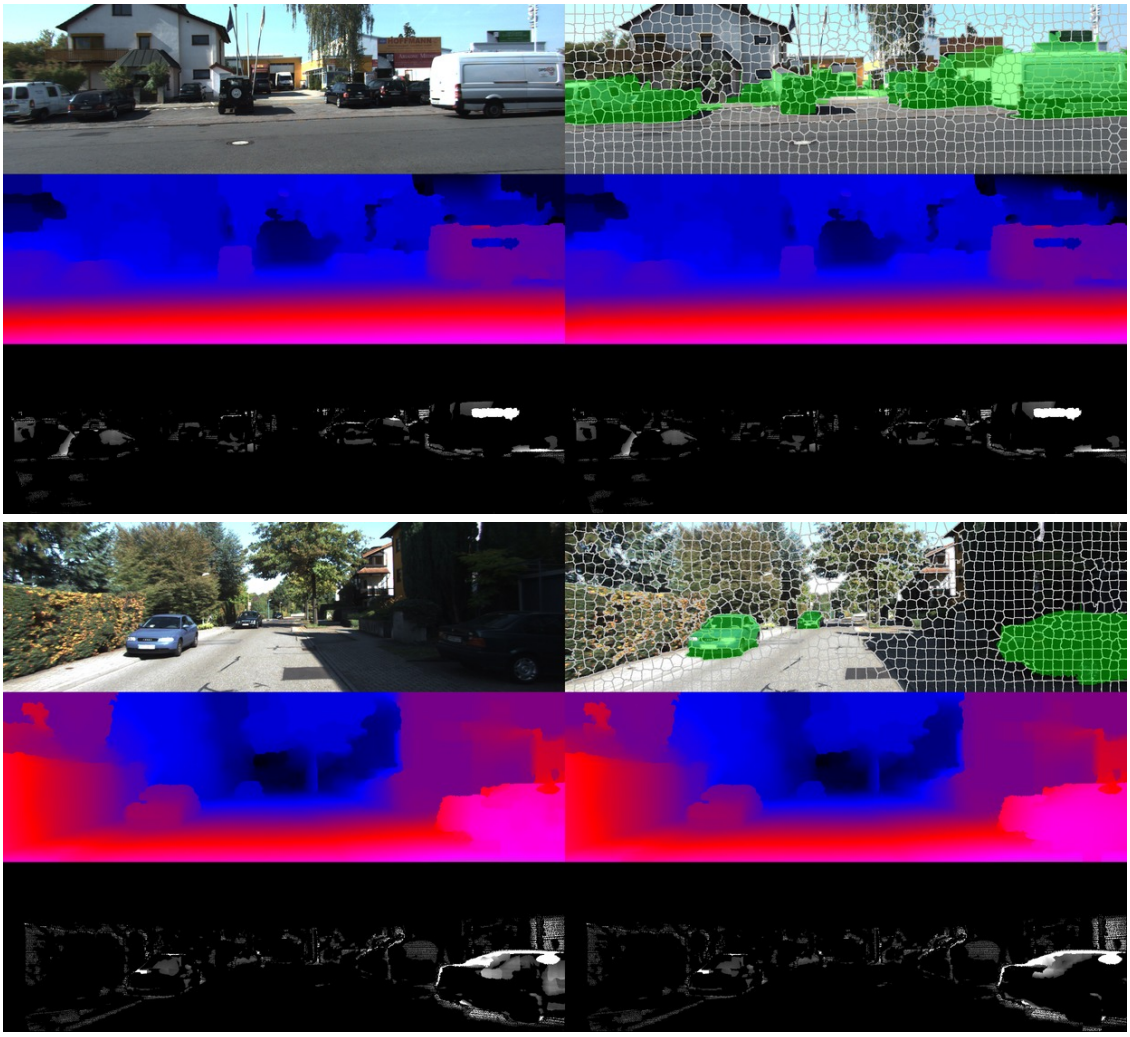


Figure 5: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

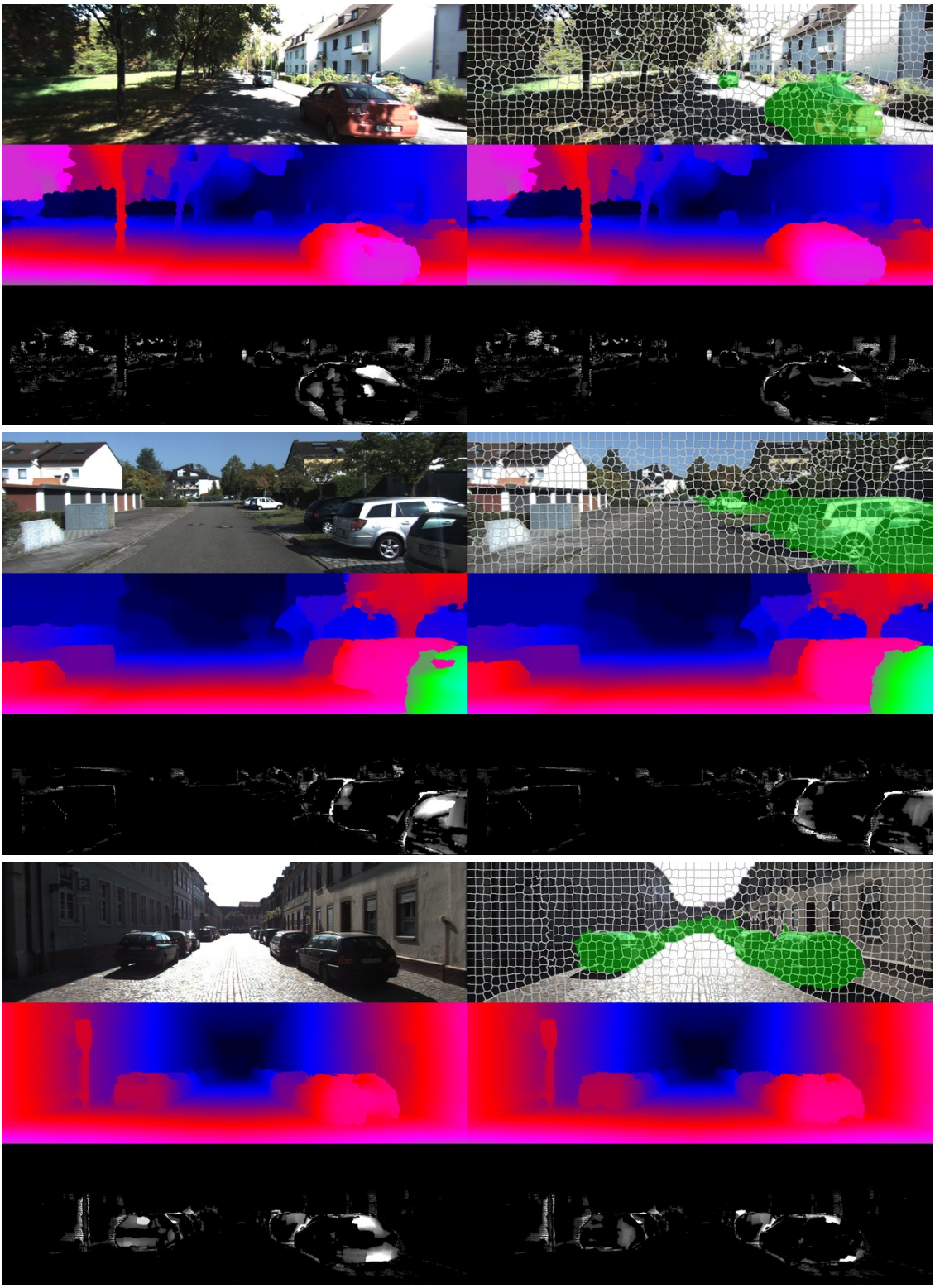


Figure 6: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

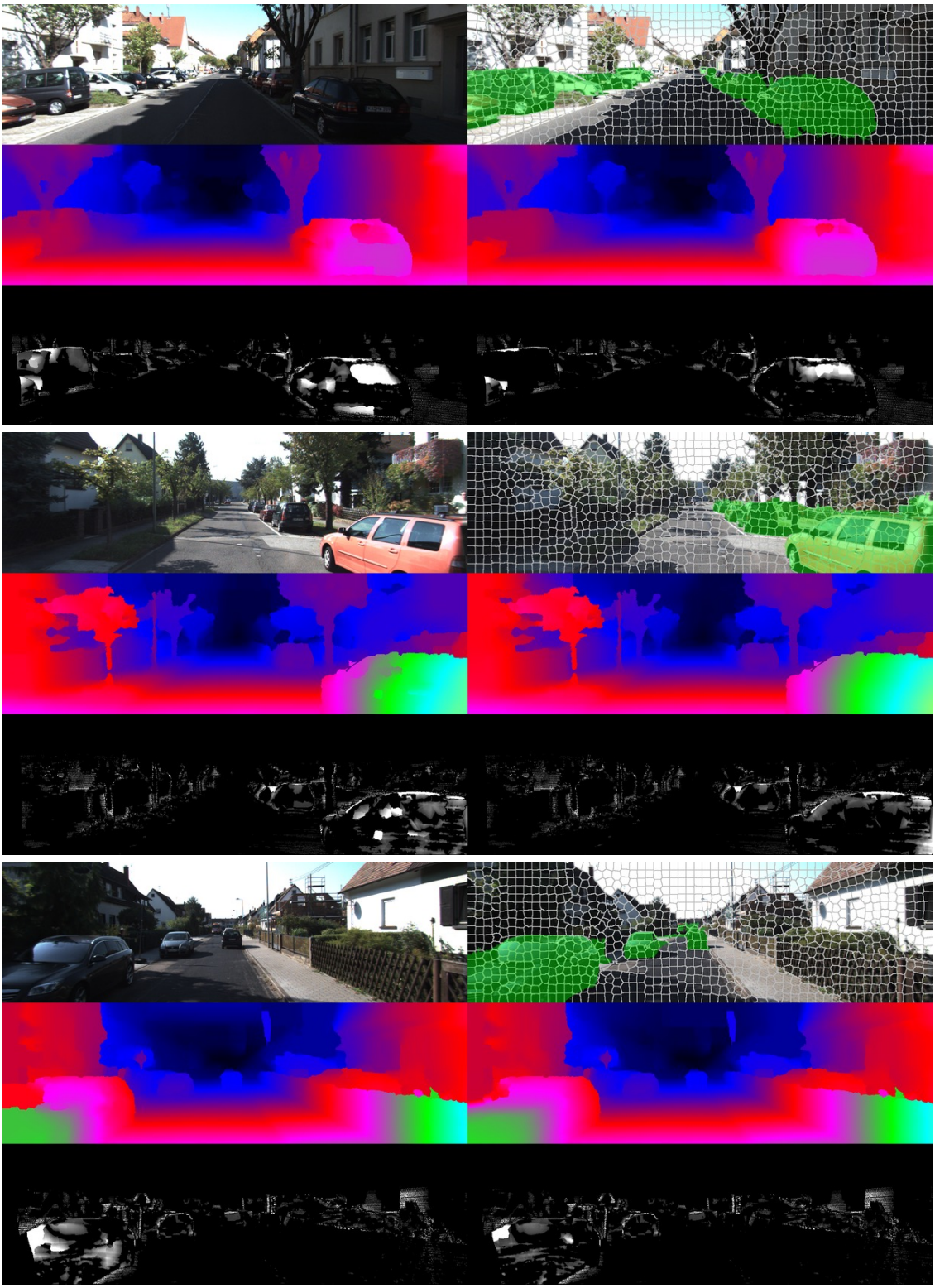


Figure 7: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

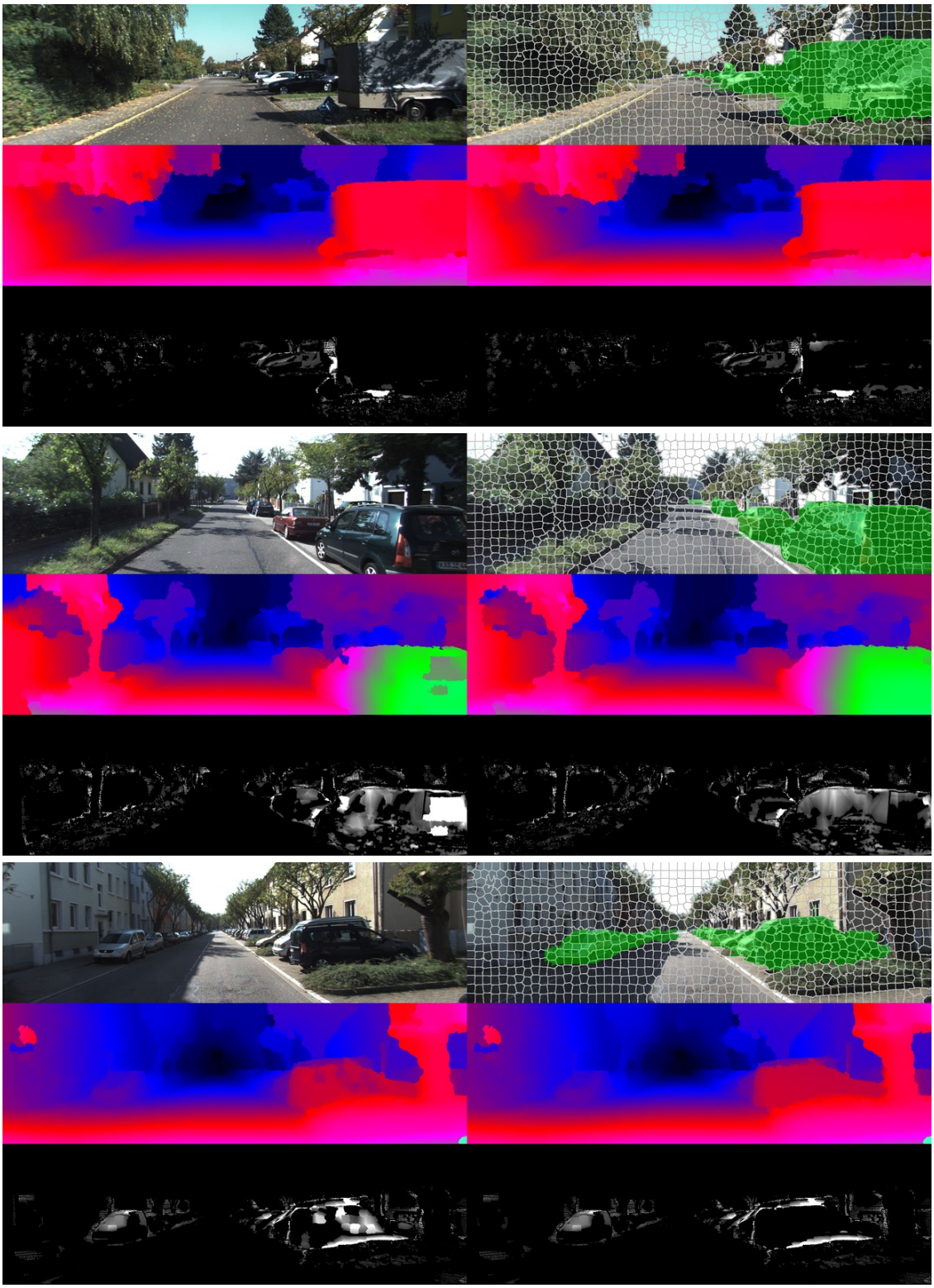


Figure 8: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

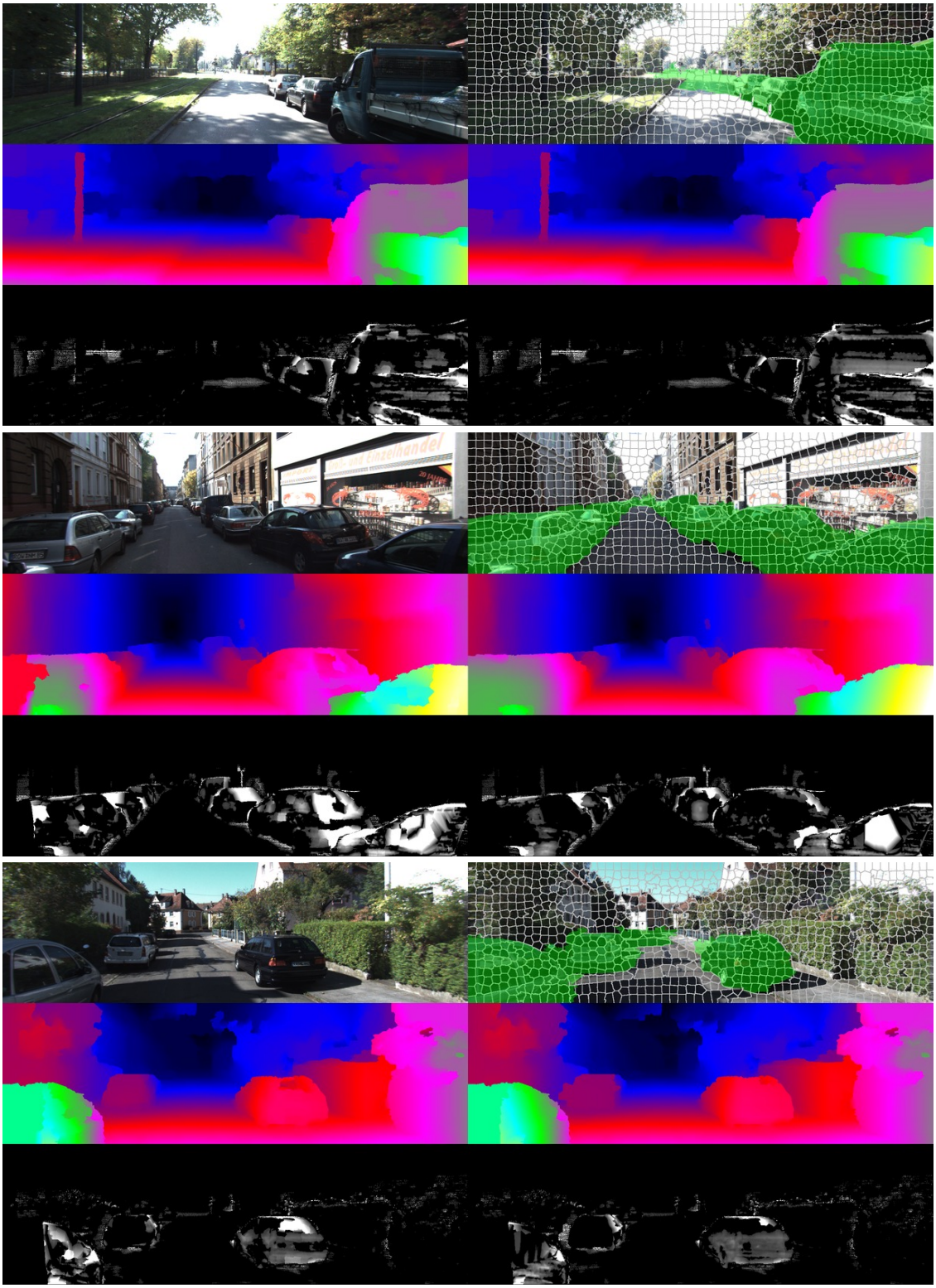


Figure 9: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

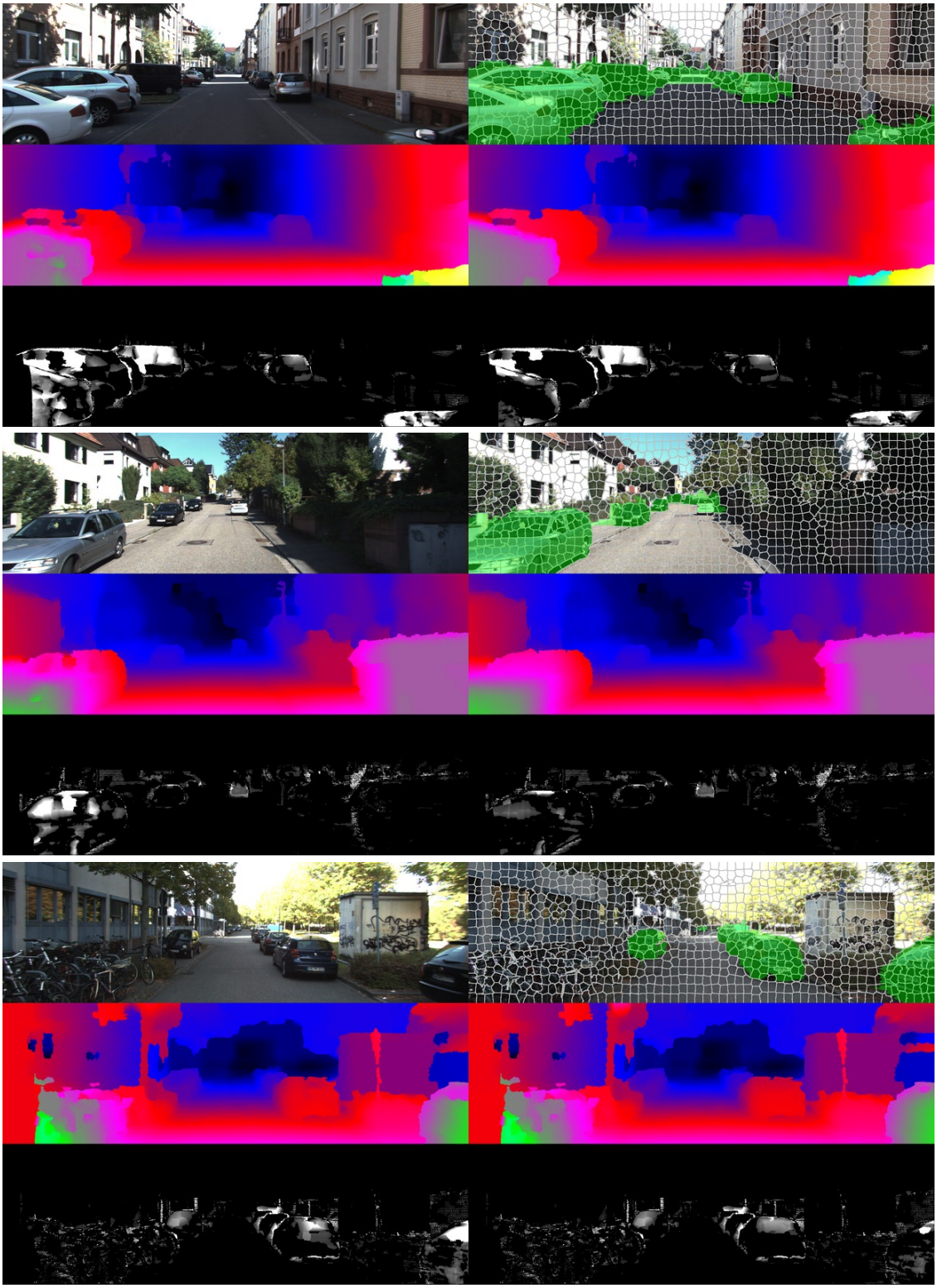


Figure 10: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

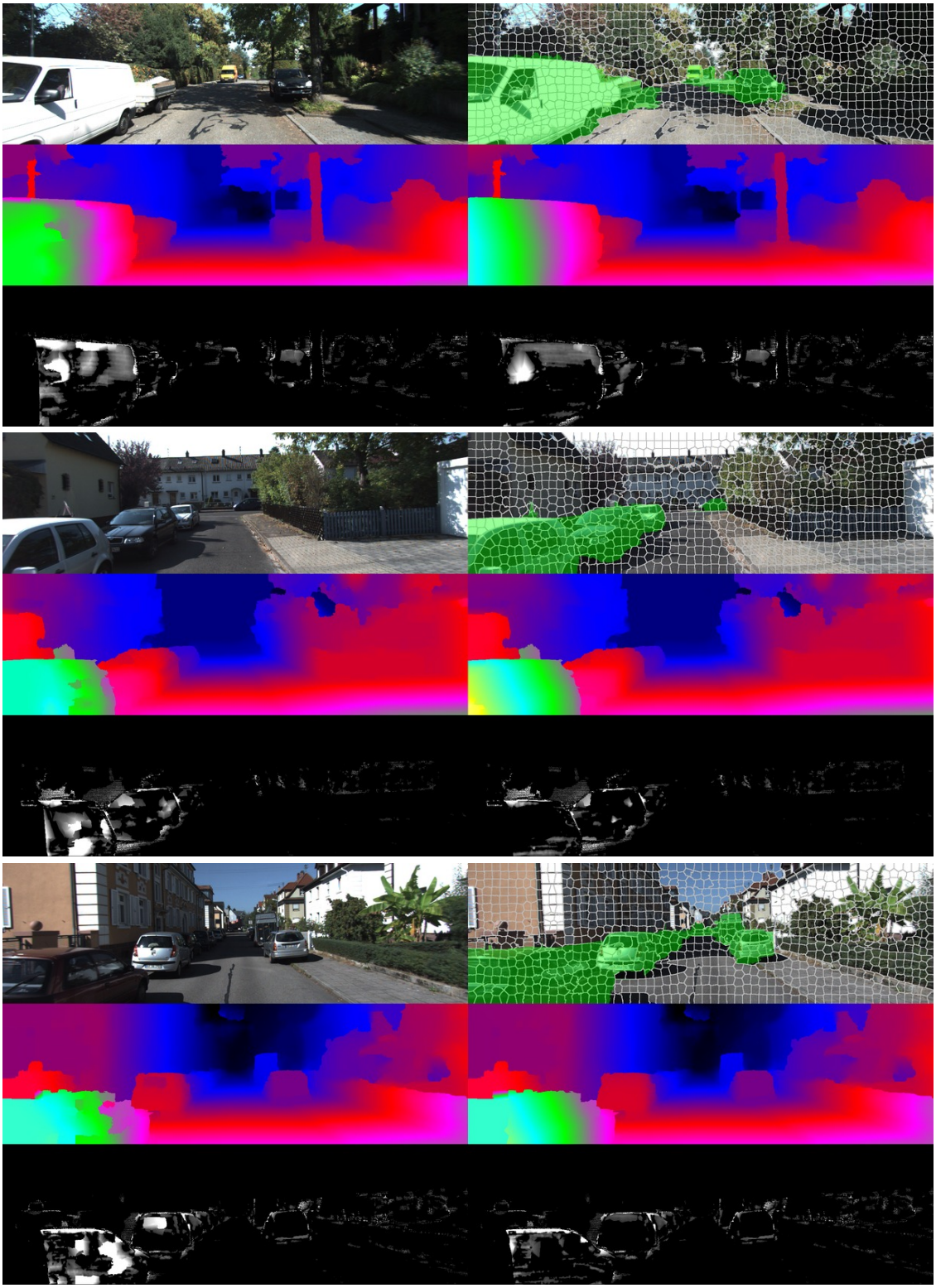


Figure 11: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

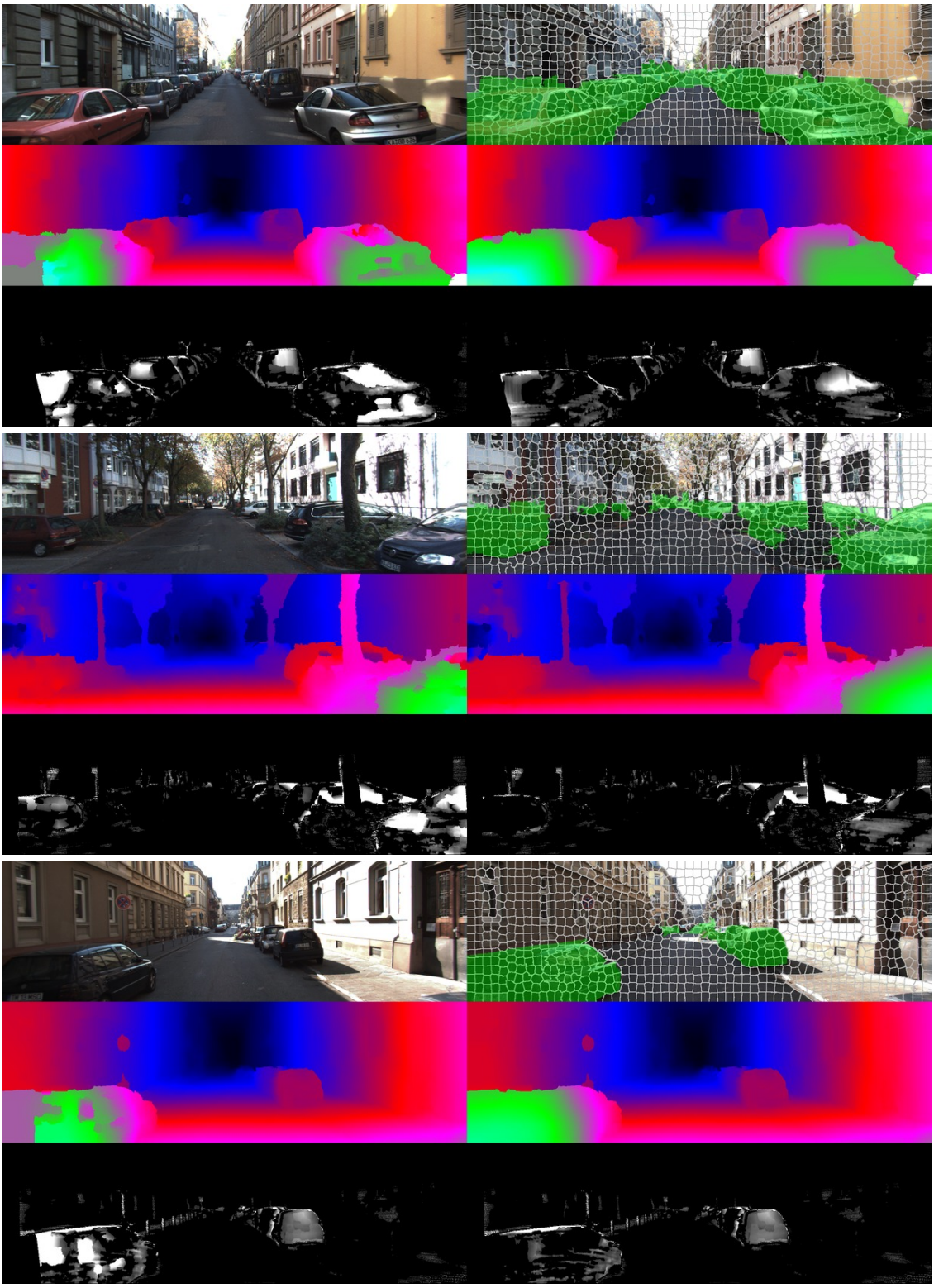


Figure 12: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

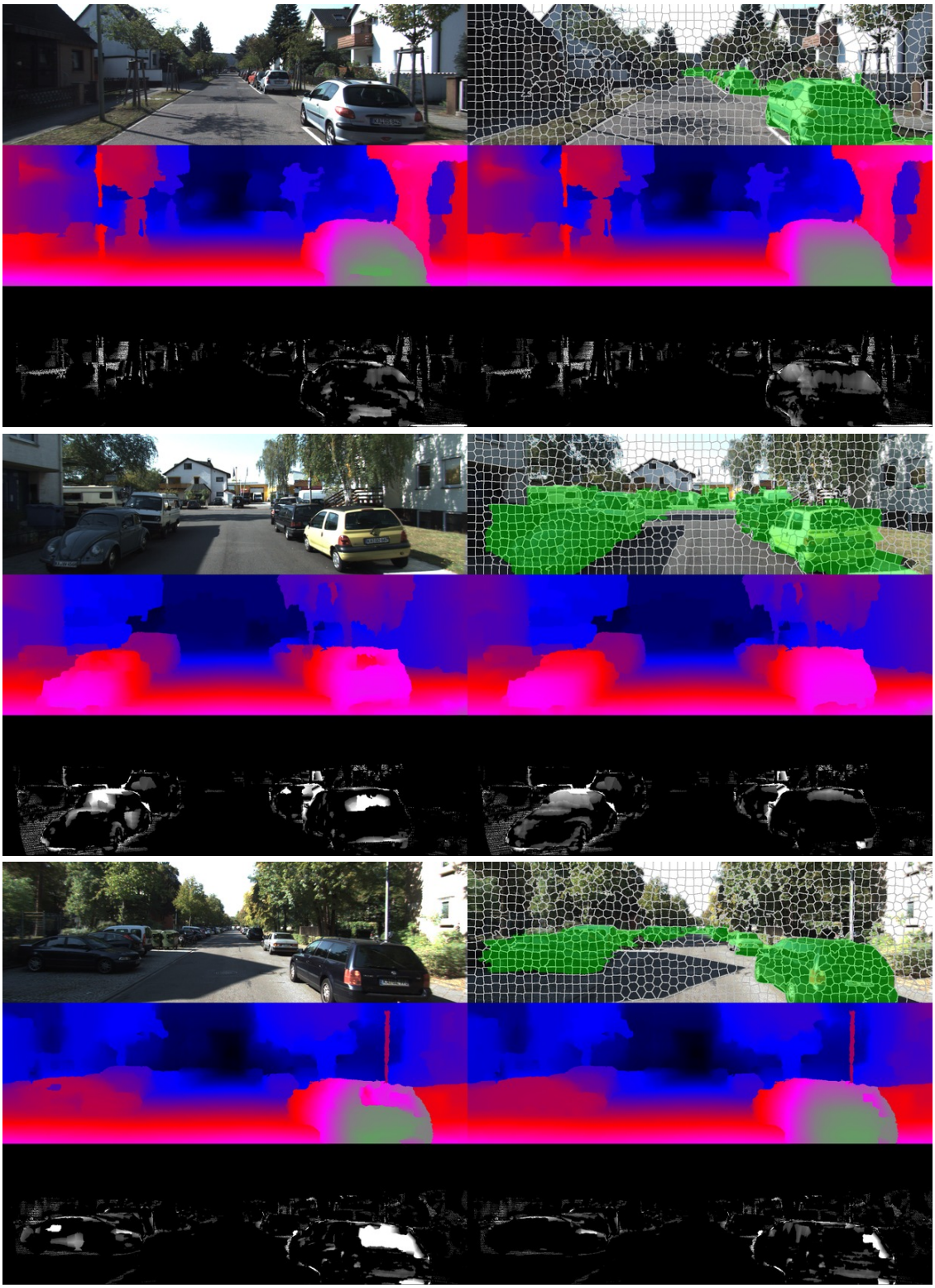


Figure 13: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).

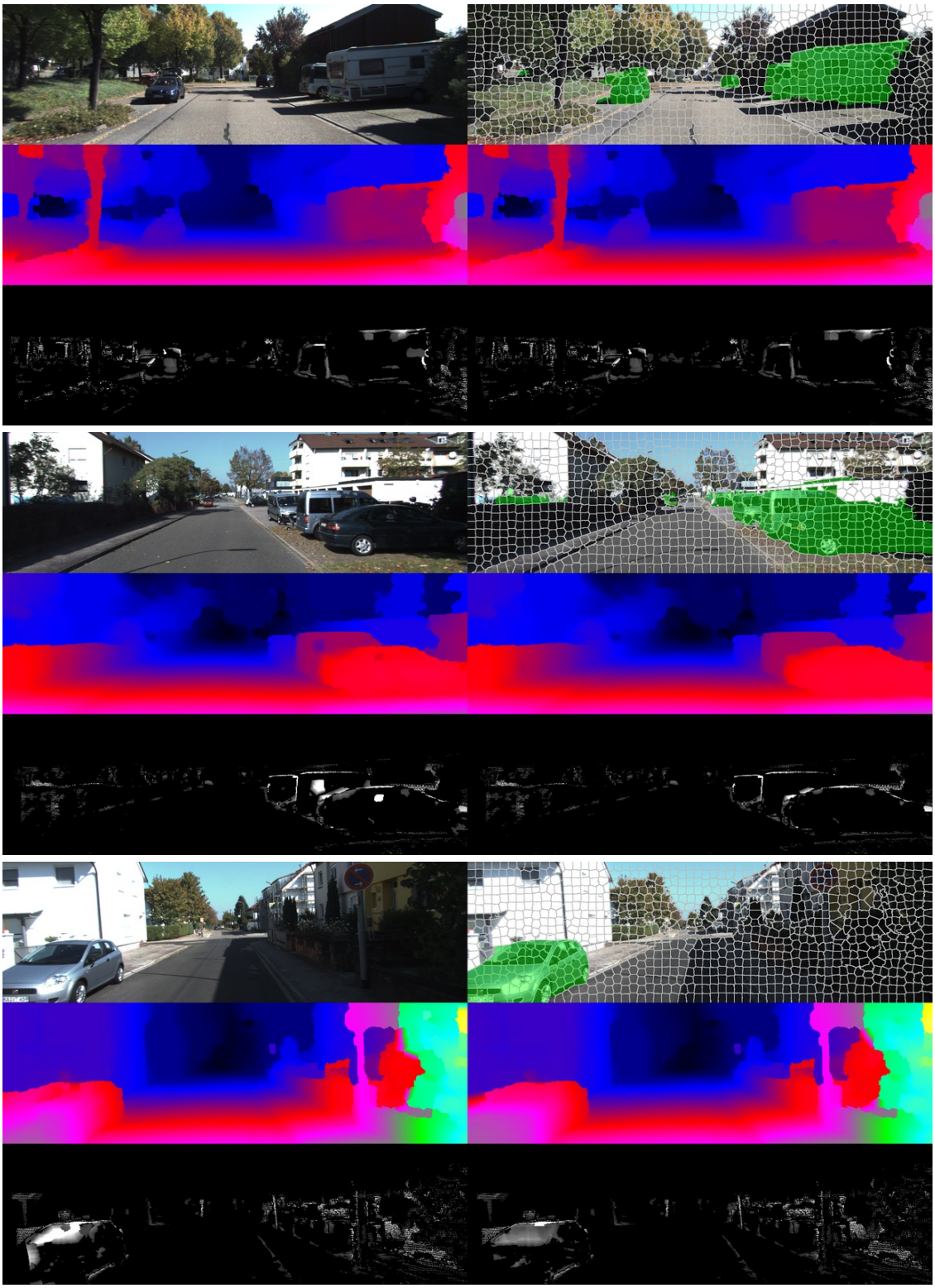


Figure 14: **Qualitative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from ≤ 1 pixel error in black to ≥ 5 pixels error in white (third row).