Using Self-Contradiction to Learn Confidence Measures in Stereo Vision Supplementary Material

Christian Mostegel Markus Rumpler Friedrich Fraundorfer Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology*
{surname}@icg.tugraz.at

Figures 1, 2 and 3 are extensions to Fig. 5 in the original paper. These figures each show the sparsification curves for three frames of the KITTI2012 dataset [1]. To give a fair overview of good and bad examples, we ranked the frames depending on the benefit of our training data. As we ran multiple combinations of query algorithm (SGM [4] and SPS [6]) and confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]), we decided to base the ranking on the combination with the lowest total Area Under the Sparsification Curve (AUSC) which turned out to be "SGM-Park" (see Tab. 1 in the original paper). Thus, we used the ratio $AUSC_{SGM-Park-Ours}/AUSC_{SGM-Park-Laser}$ for ranking the frames of the KITTI training dataset in ascending order. Based on this ranking we selected the three best frames (Fig. 1), the three frames around the median index (Fig. 2) and the three worst frames (Fig. 3).

For each of these nine frames we show two additional figures (one for SGM [4] (Fig. 4-12) and one for SPS [6] (Fig. 13-21)). Each of these figures shows the RGB input image, the depth image produced by the query algorithm (SGM or SPS), the label images generated with the laser ground truth [1] and with our approach, as well as the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser [1] and Ours).

For the label image generation, note that the amount of generated positive training data is strongly influenced by the camera motion. We require at least one reference measurement which is more accurate than the query measurement and has a minimum relative observation angle of at least $10°$ (compare Fig. 4 and Fig. 9). This leads to a low coverage in the middle of the road, while on the side of the image the labels are more dense.

For the confidence prediction, note that the prediction output trained on our label images (Ours) is less noisy and impact of the distance to border features becomes smoother (e.g. Fig. 4 left margin and bottom right corner of the image). It seems that the random forests over-fit if we only use the limited amount of label images generated with the laser ground truth. In contrast, our more diverse training data reduces the chance of over-fitting, which in turn leads to an improved overall prediction performance.

As a matter of completeness, Figure 22 shows a different version of Fig. 6 in the original paper. While Fig. 6 in the original paper shows the mean, minimum and maximum values across the three confidence prediction approaches (Ensemble [2], GCP [5], Park [3]), Fig. 22 shows each approach individually. The sequences in Fig. 22 were sorted by the optimal AUSC values.

*The research leading to these results has received funding from the EC FP7 project 3D-PITOTI (ICT-2011-600545) and from the Austrian Research Promotion Agency (FFG) as Vision+ project 836630 and together with OMICRON electronics GmbH as Bridge1 project 843450.

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[2] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 305–312, June 2013. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

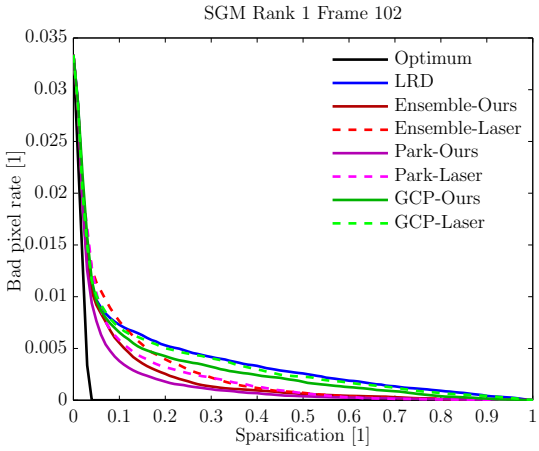
[3] M.-G. Park and K.-J. Yoon. Leveraging stereo matching with learning-based confidence measures. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 101–109, June 2015. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

[4] M. Rothermel, K. Wenzel, D. Fritsch, and N. Haala. SURE: Photogrammetric Surface Reconstruction from Imagery. In *Proceedings LC3D Workshop*, 2012. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 24

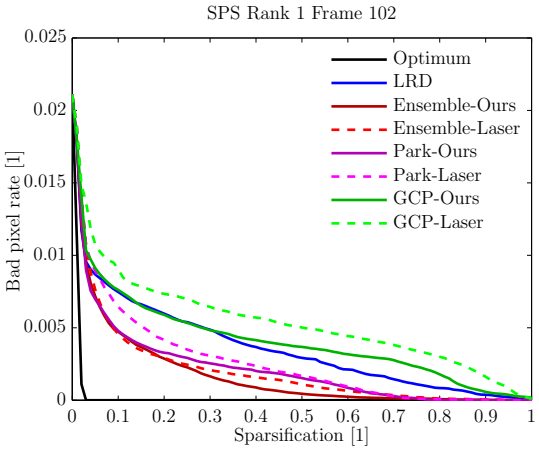
[5] A. Spyropoulos, N. Komodakis, and P. Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1621–1628, June 2014. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

[6] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation.

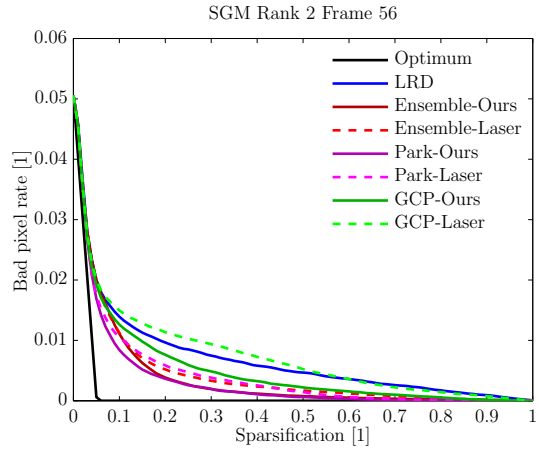
In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 756–771. Springer International Publishing, 2014. 1, 3, 4, 5, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24



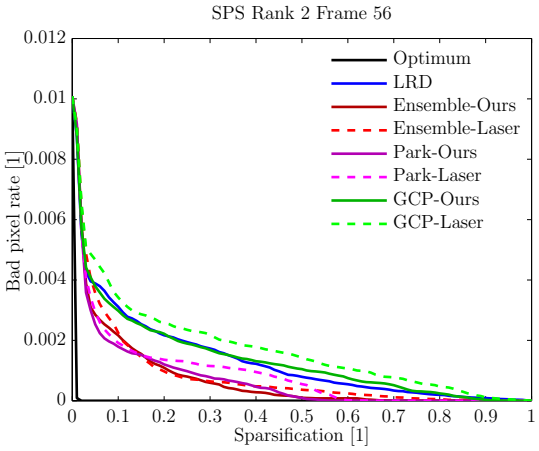
(a) SGM Frame 102 (details in Fig. 10).



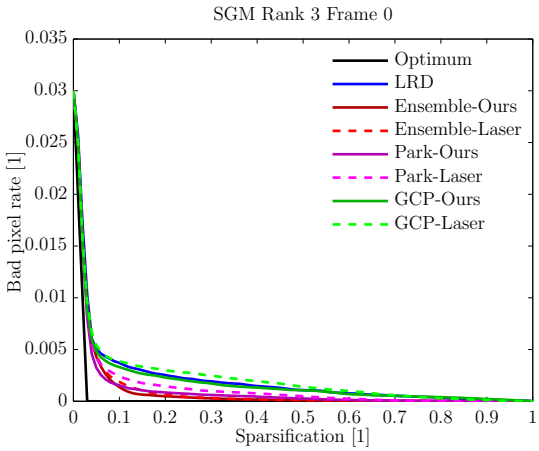
(b) SPS Frame 102 (details in Fig. 19).



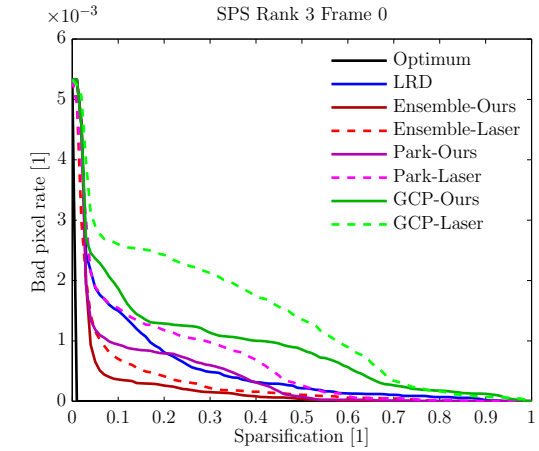
(c) SGM Frame 56 (details in Fig. 7).



(d) SPS Frame 56 (details in Fig. 16).

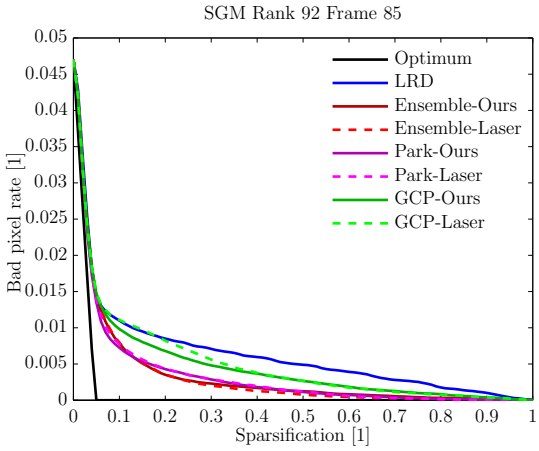


(e) SGM Frame 0 (details in Fig. 4).

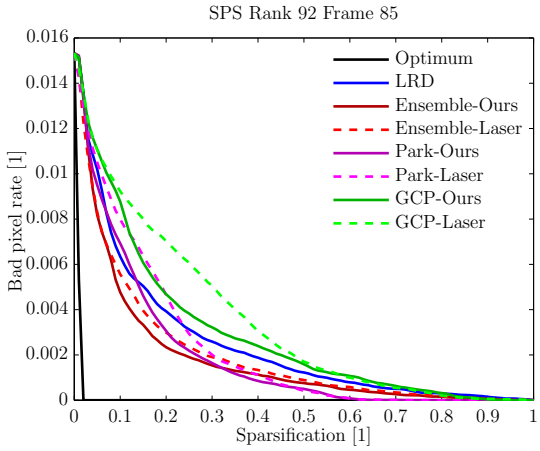


(f) SPS Frame 0 (details in Fig. 13).

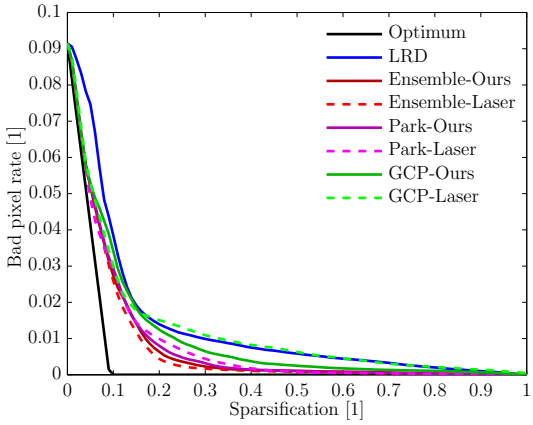
Figure 1. Sparsification curves for the three **best frames** with (102,56, and 0) of the KITTI training dataset. The ranking was obtained by the ratio $AUSC_{SGM-Park-Ours}/AUSC_{SGM-Park-Laser}$. We display all combinations of query algorithm (SGM [4] and SPS [6]), confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). As a baseline method we also show the Left-Right disparity Difference (LRD).



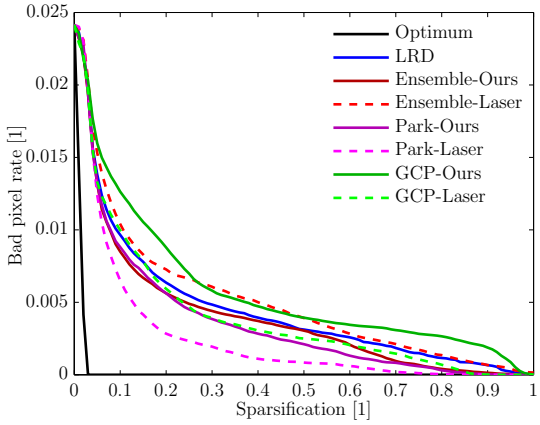
(a) SGM Frame 85 (details in Fig. 8).



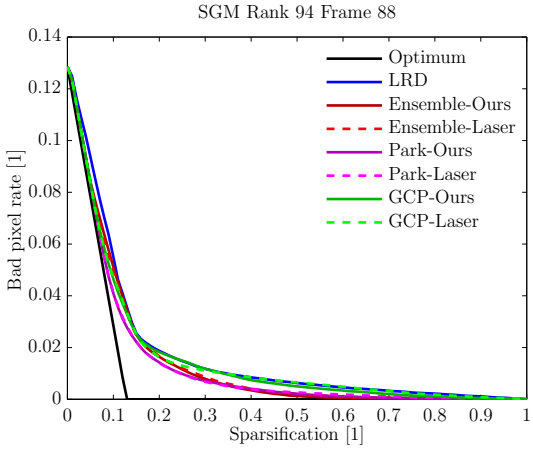
(b) SPS Frame 85 (details in Fig. 17).



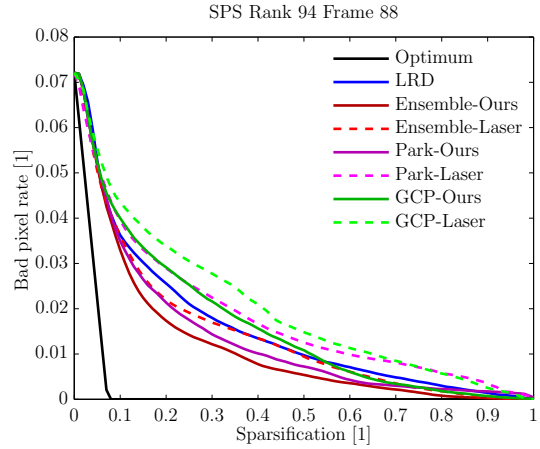
(c) SGM Frame 126 (details in Fig. 11).



(d) SPS Frame 126 (details in Fig. 20).

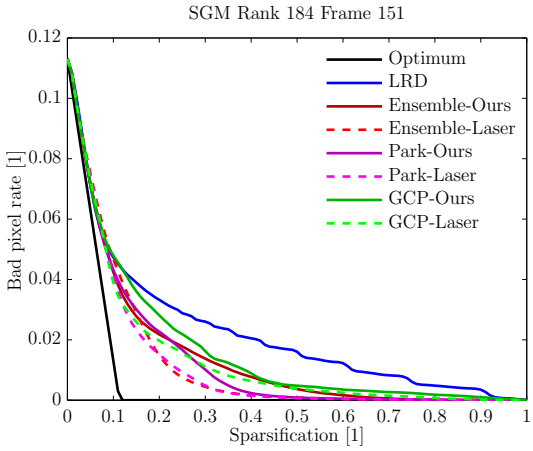


(e) SGM Frame 88 (details in Fig. 9).

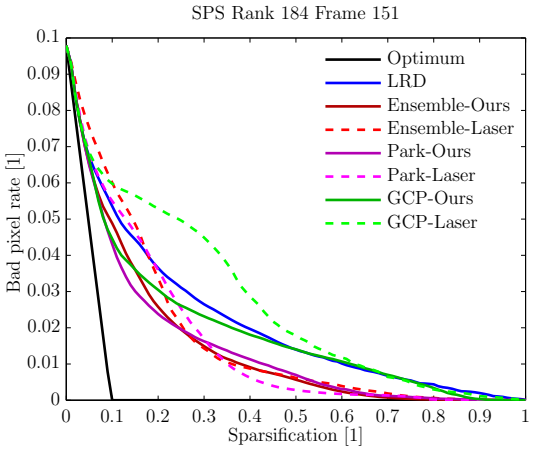


(f) SPS Frame 88 (details in Fig. 18).

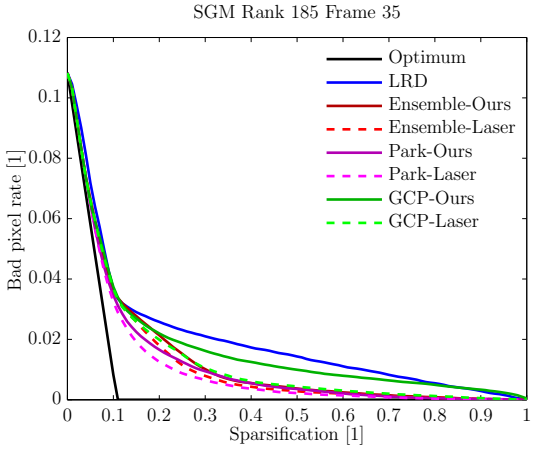
Figure 2. Sparsification curves for the three **median frames** (85,126,88) of the KITTI training dataset. The ranking was obtained by the ratio $AUSC_{SGM-Park-Ours}/AUSC_{SGM-Park-Laser}$. We display all combinations of query algorithm (SGM [4] and SPS [6]), confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). As a baseline method we also show the Left-Right disparity Difference (LRD).



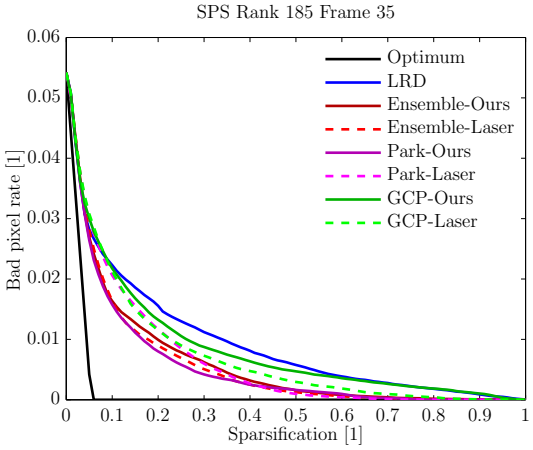
(a) SGM Frame 151 (details in Fig. 12).



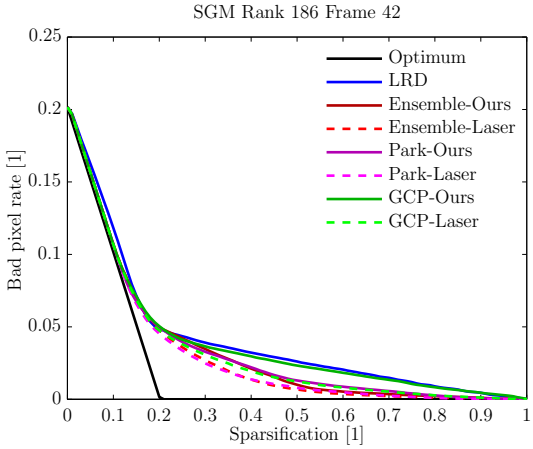
(b) SPS Frame 151 (details in Fig. 21).



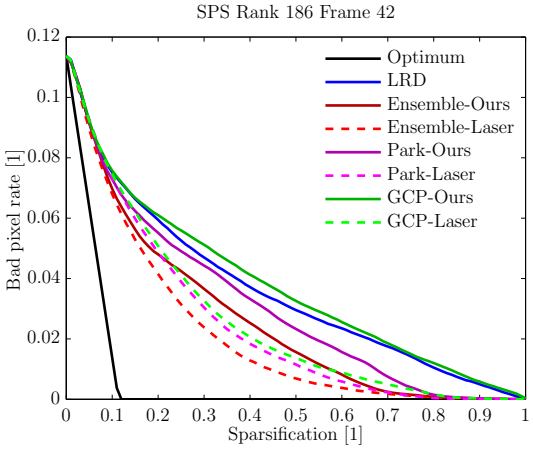
(c) SGM Frame 35 (details in Fig. 5).



(d) SPS Frame 35 (details in Fig. 14).



(e) SGM Frame 42 (details in Fig. 6).



(f) SPS Frame 42 (details in Fig. 15).

Figure 3. Sparsification curves for the three **worst frames** (151,35,42) of the KITTI training dataset. The ranking was obtained by the ratio $AUSC_{SGM-Park-Ours}/AUSC_{SGM-Park-Laser}$. We display all combinations of query algorithm (SGM [4] and SPS [6]), confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). As a baseline method we also show the Left-Right disparity Difference (LRD).

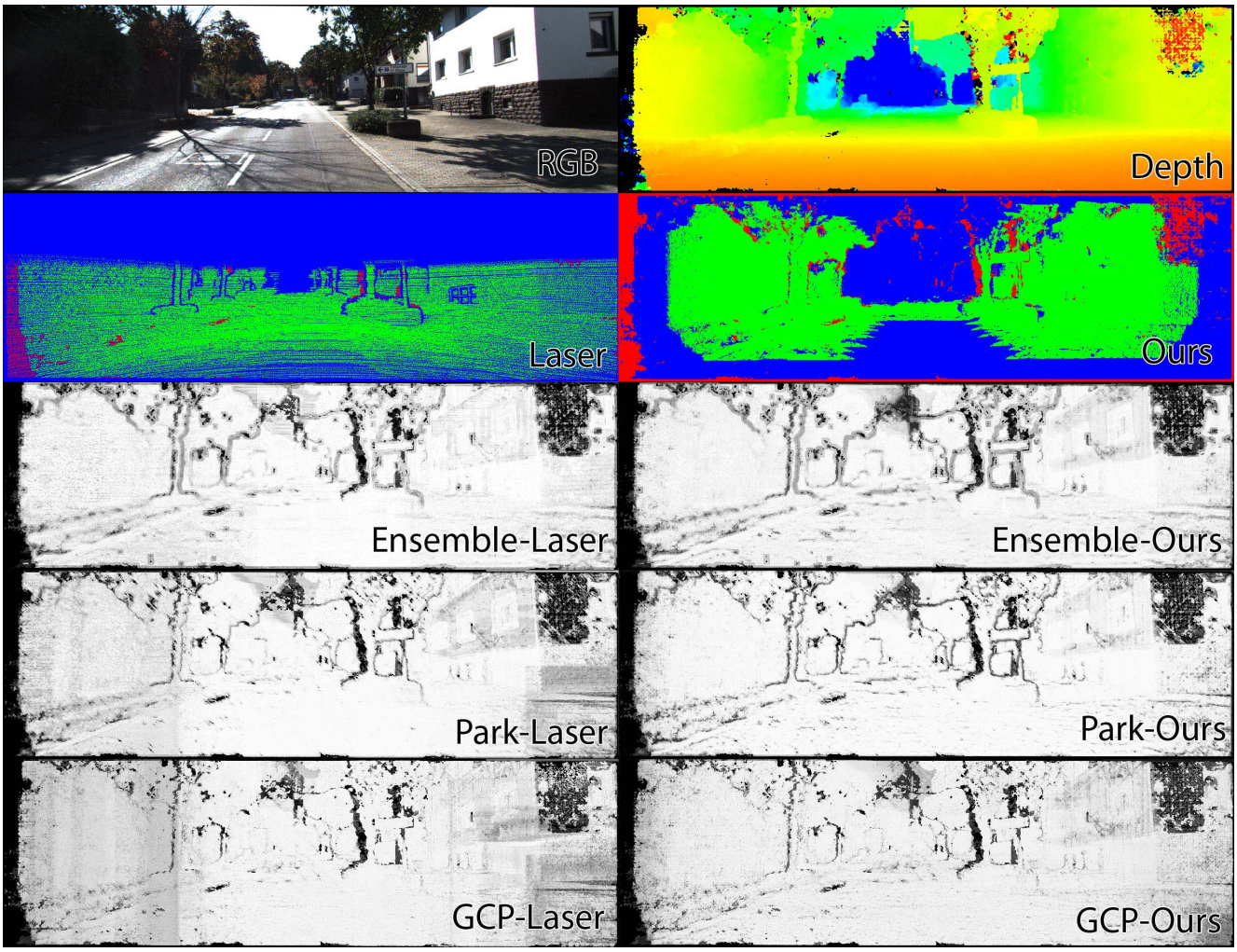


Figure 4. Visual comparison of outputs for frame 0 and SGM [4] as query algorithm (rank 3 - best). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

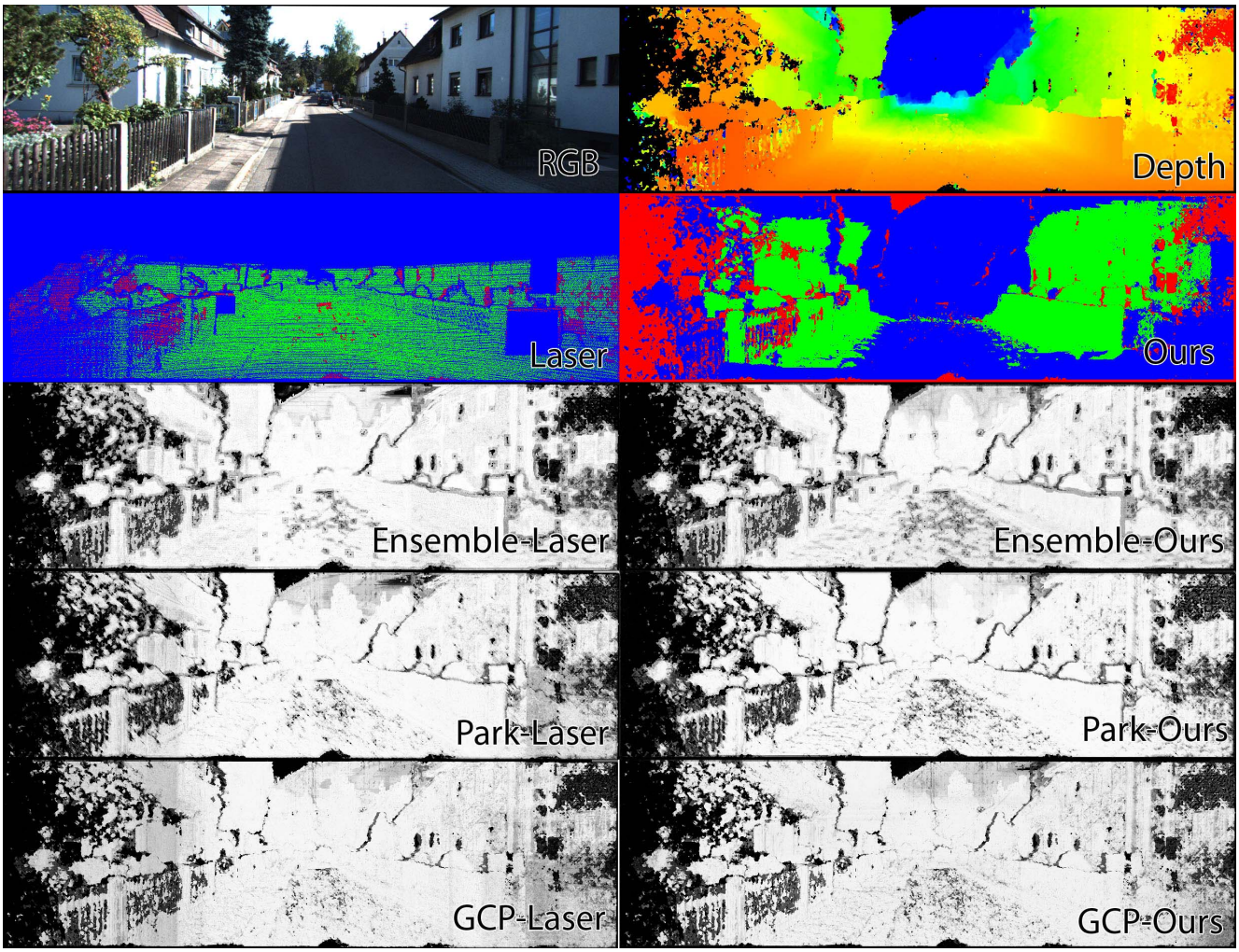


Figure 5. Visual comparison of outputs for frame 35 and SGM [4] as query algorithm (rank 185 - worst). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

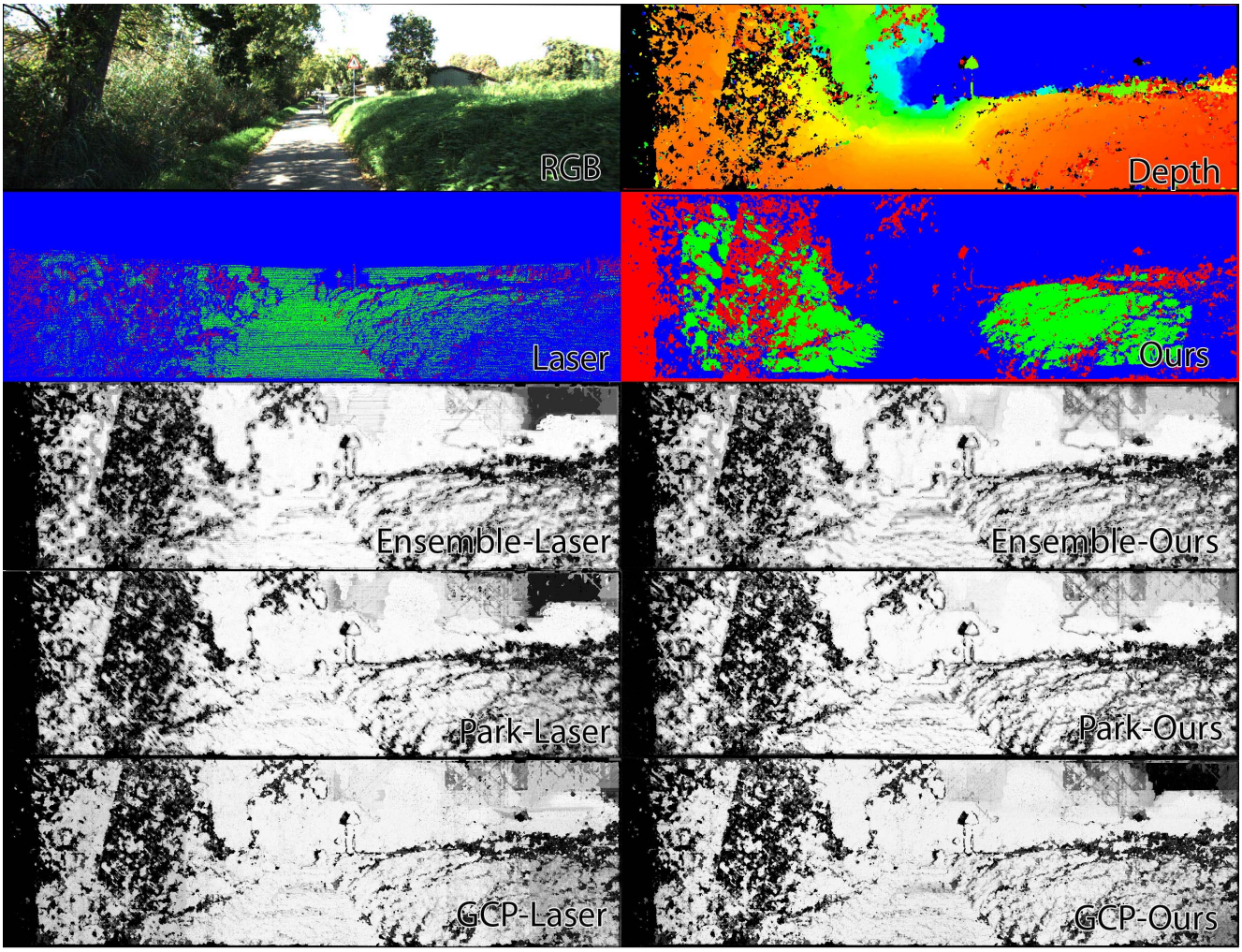


Figure 6. Visual comparison of outputs for frame 42 and SGM [4] as query algorithm (rank 186 - worst). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

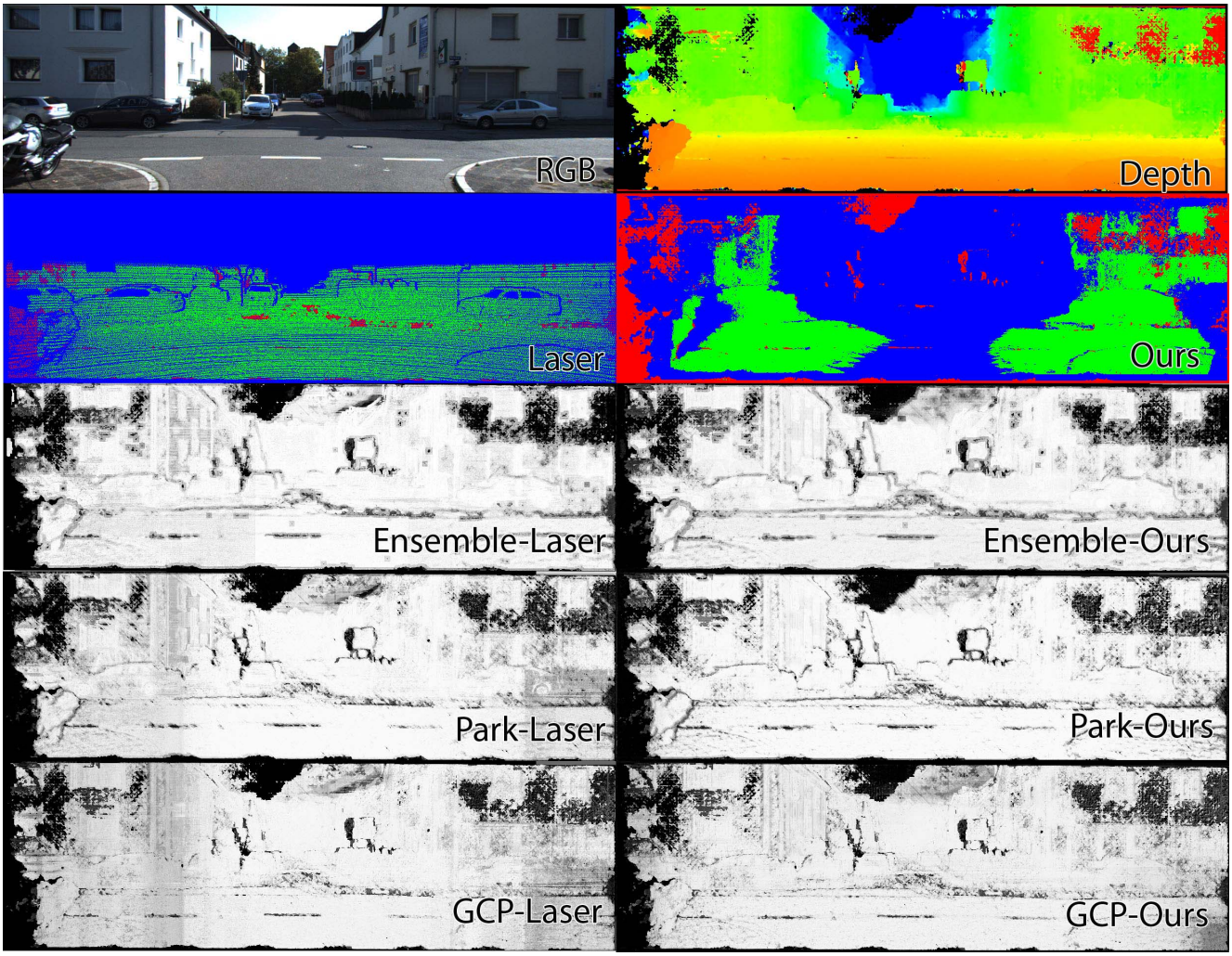


Figure 7. Visual comparison of outputs for frame 56 and SGM [4] as query algorithm (rank 2 - best). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

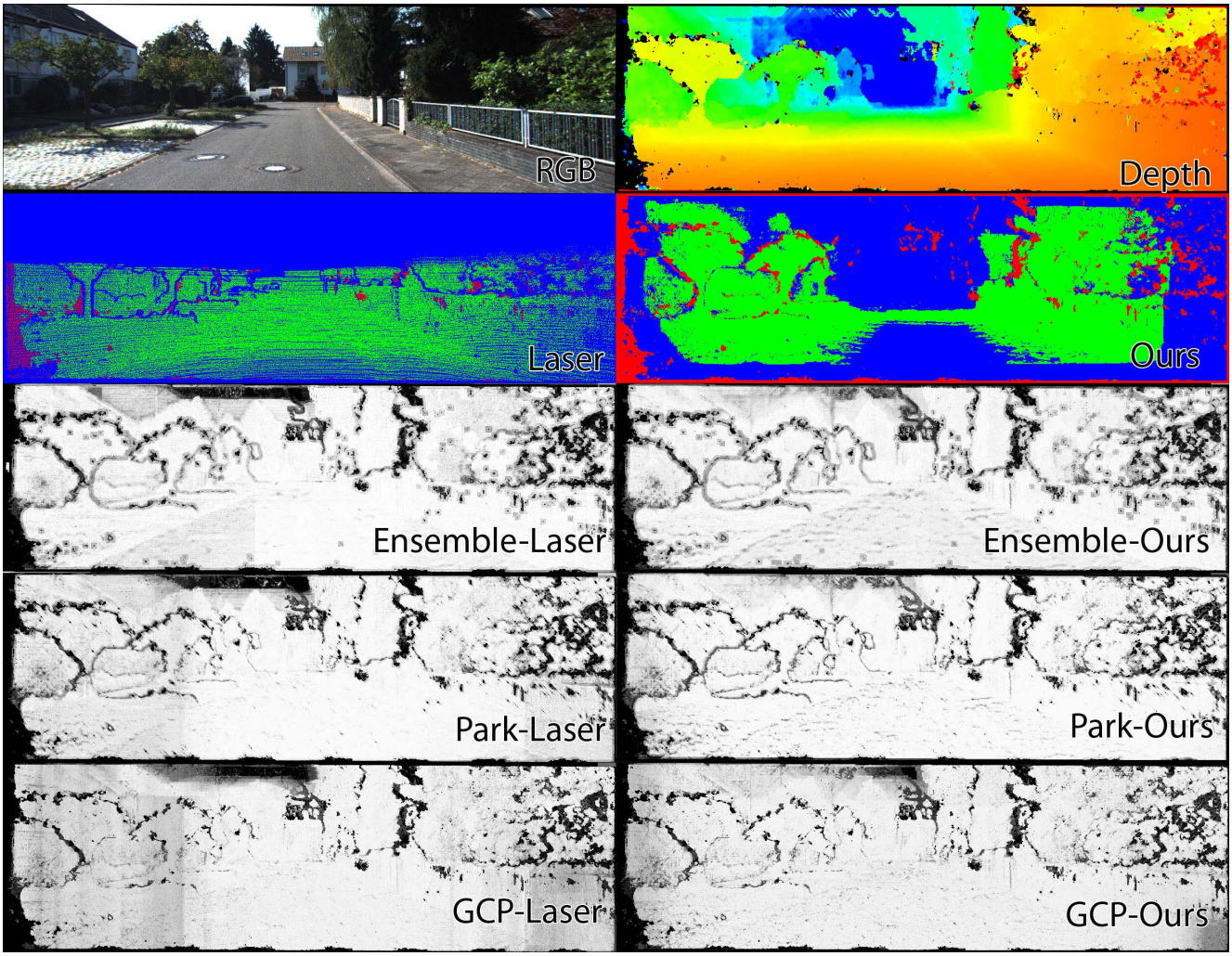


Figure 8. Visual comparison of outputs for frame 85 and SGM [4] as query algorithm (rank 92 - median). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

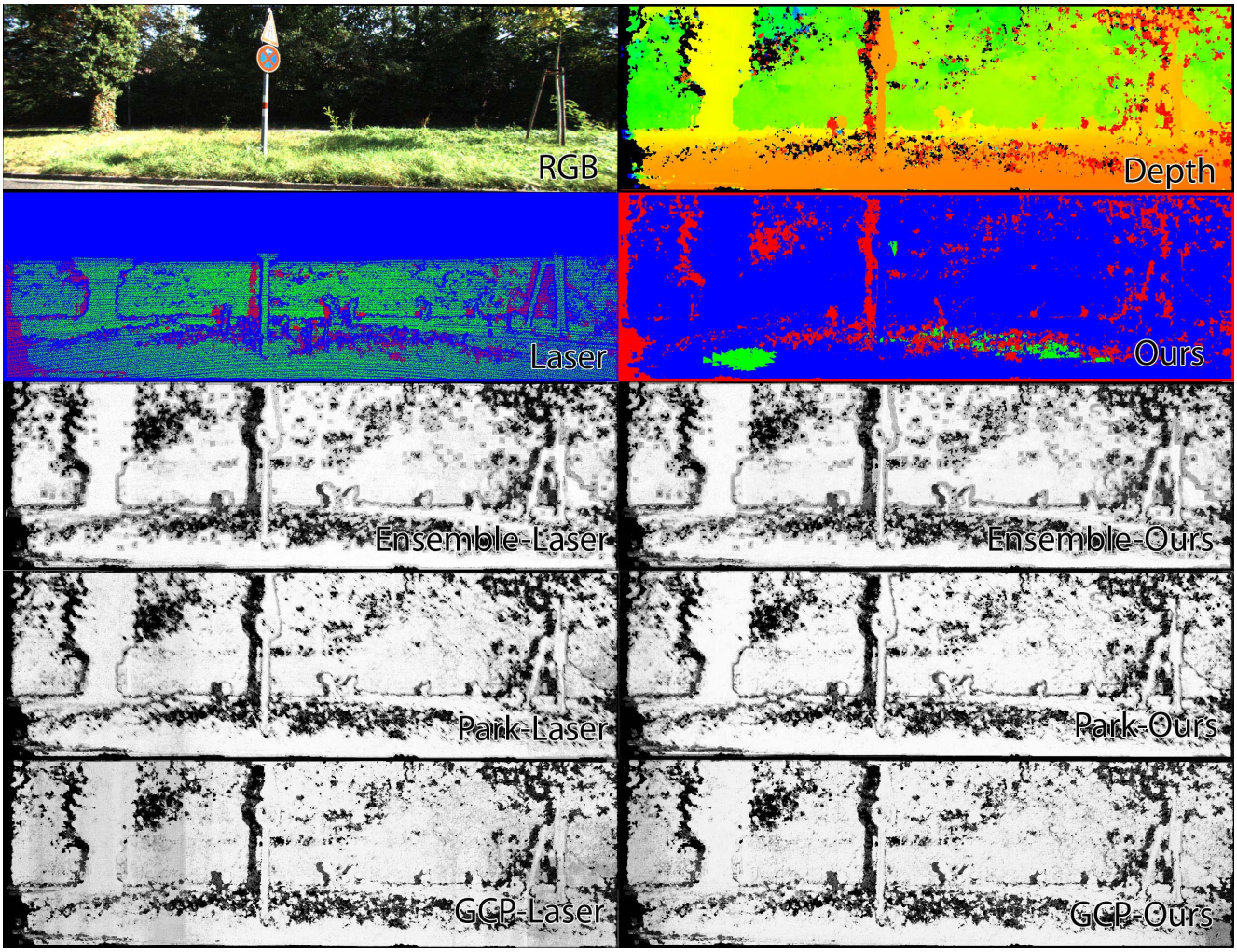


Figure 9. Visual comparison of outputs for frame 88 and SGM [4] as query algorithm (rank 94 - median). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

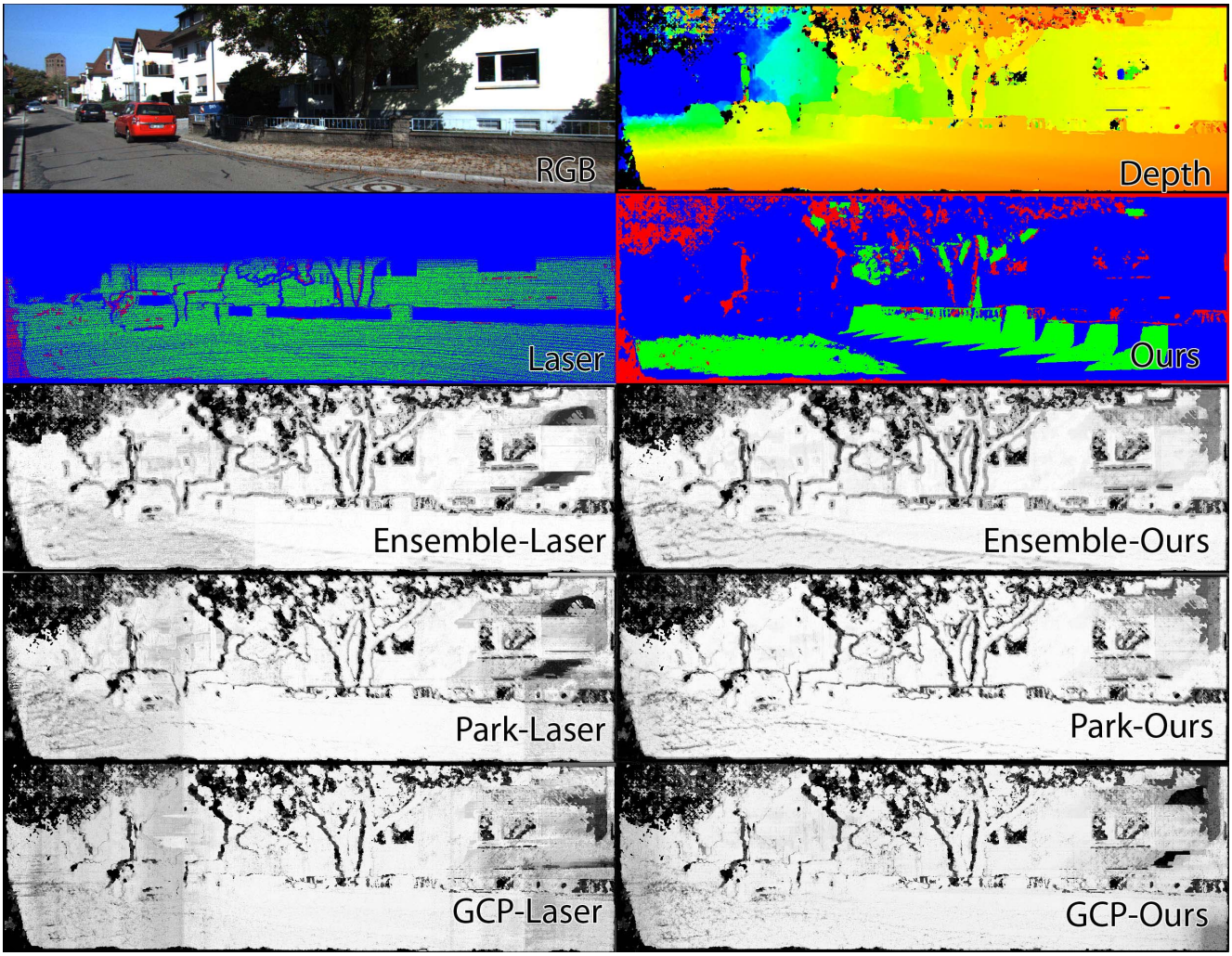


Figure 10. Visual comparison of outputs for frame 102 and SGM [4] as query algorithm (rank 1 - best). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

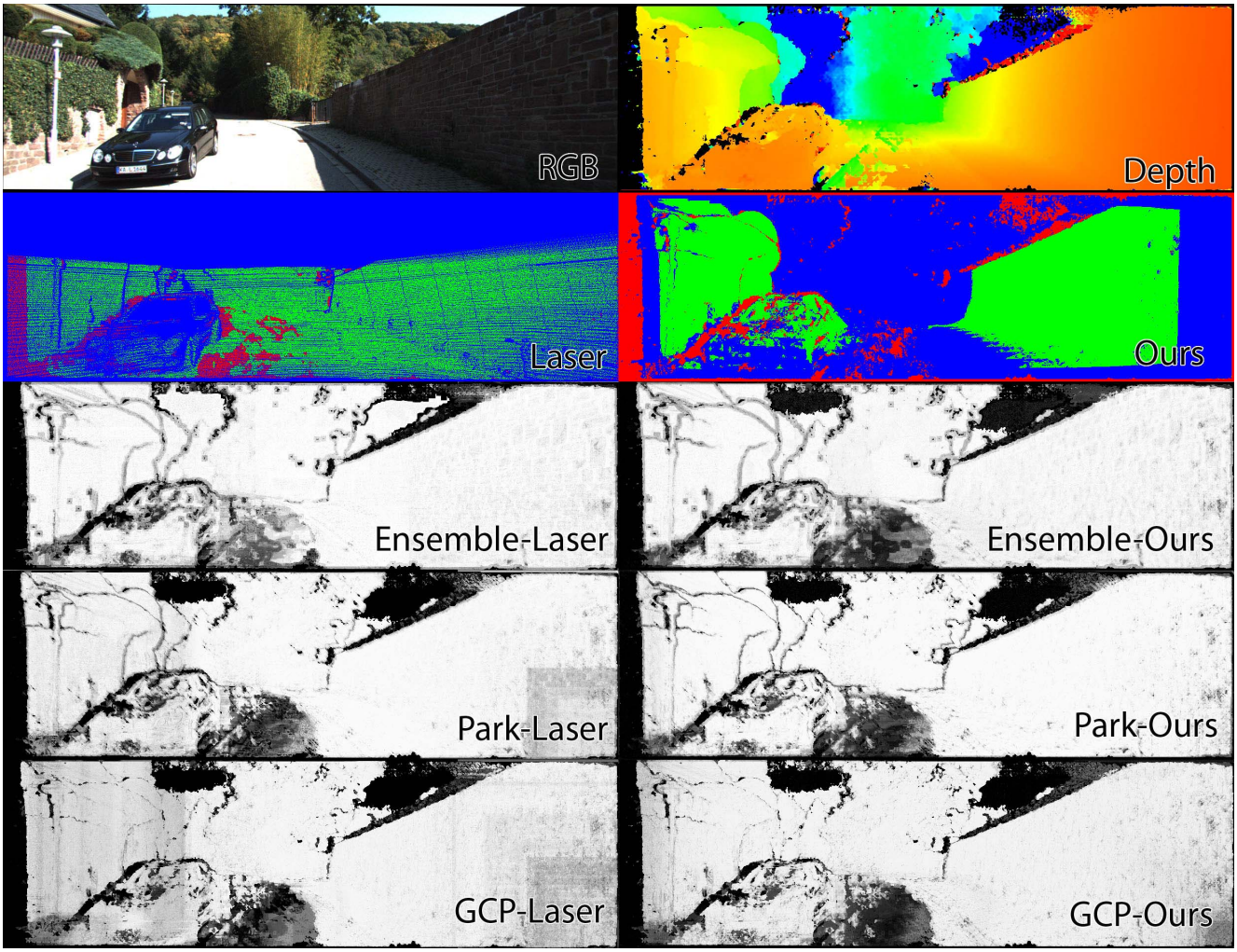


Figure 11. Visual comparison of outputs for frame 126 and SGM [4] as query algorithm (rank 93 - median). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

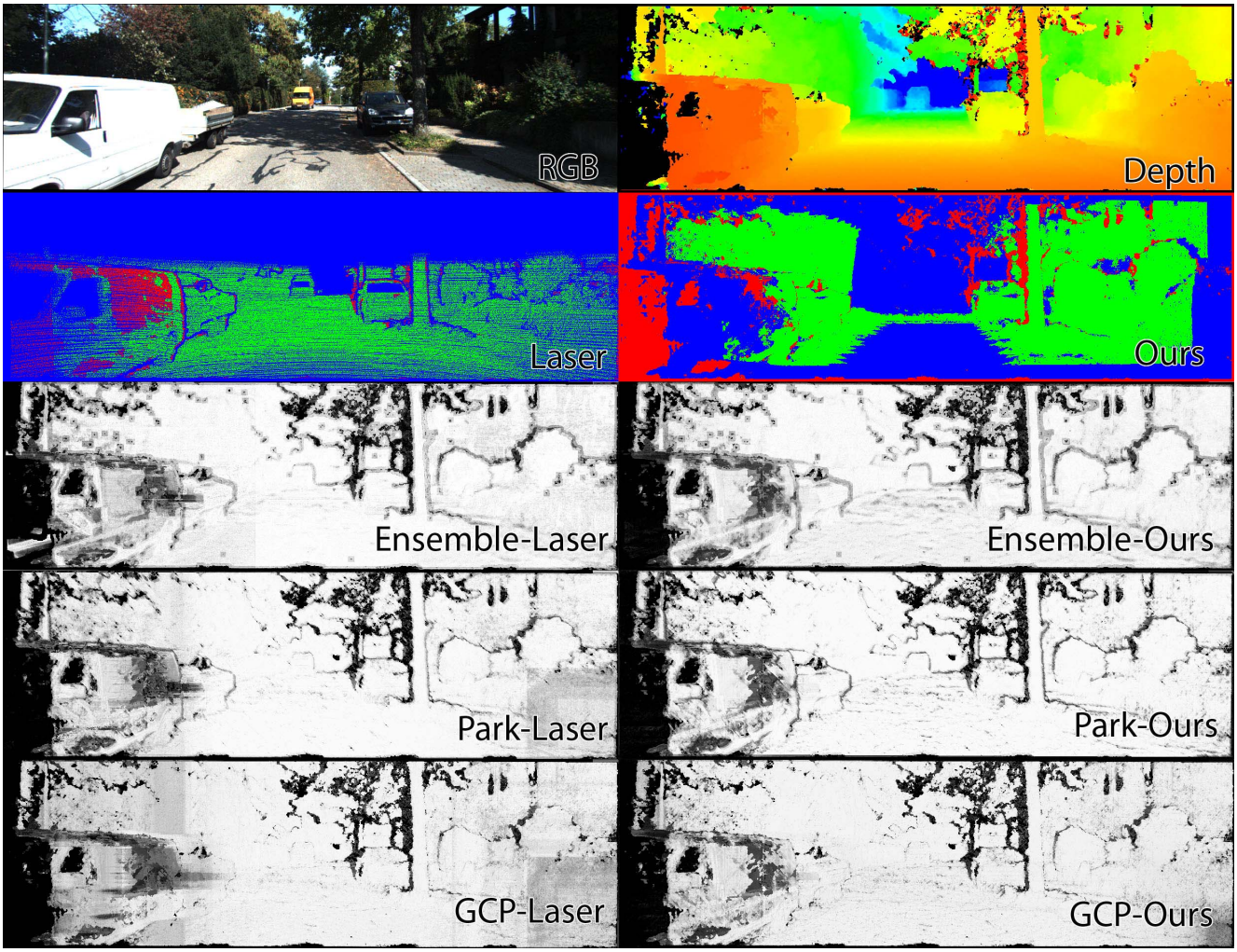


Figure 12. Visual comparison of outputs for frame 151 and SGM [4] as query algorithm (rank 184 - worst). Top row: RGB input image and SGM [4] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

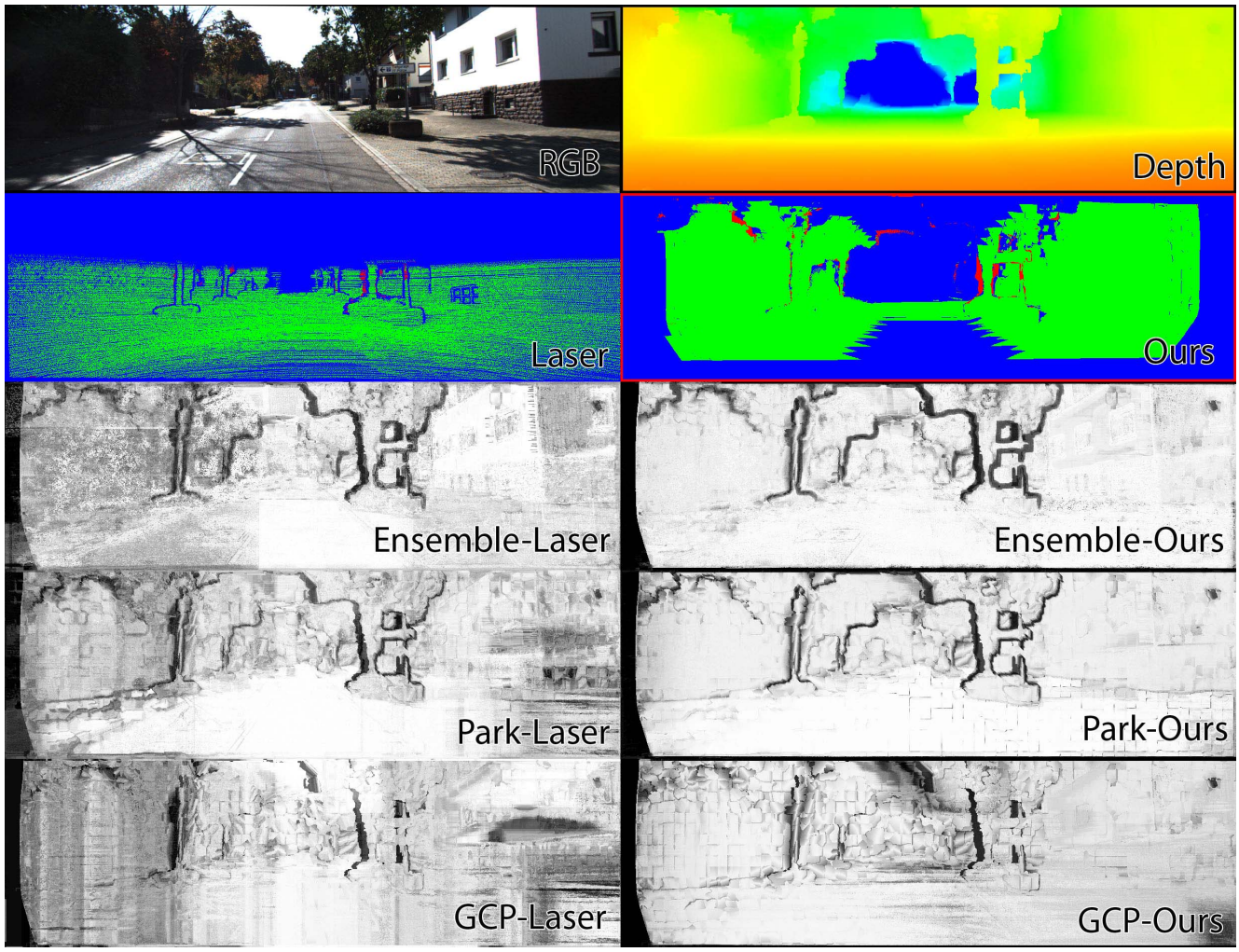


Figure 13. Visual comparison of outputs for frame 0 and SPS [6] as query algorithm (rank 3 - best). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

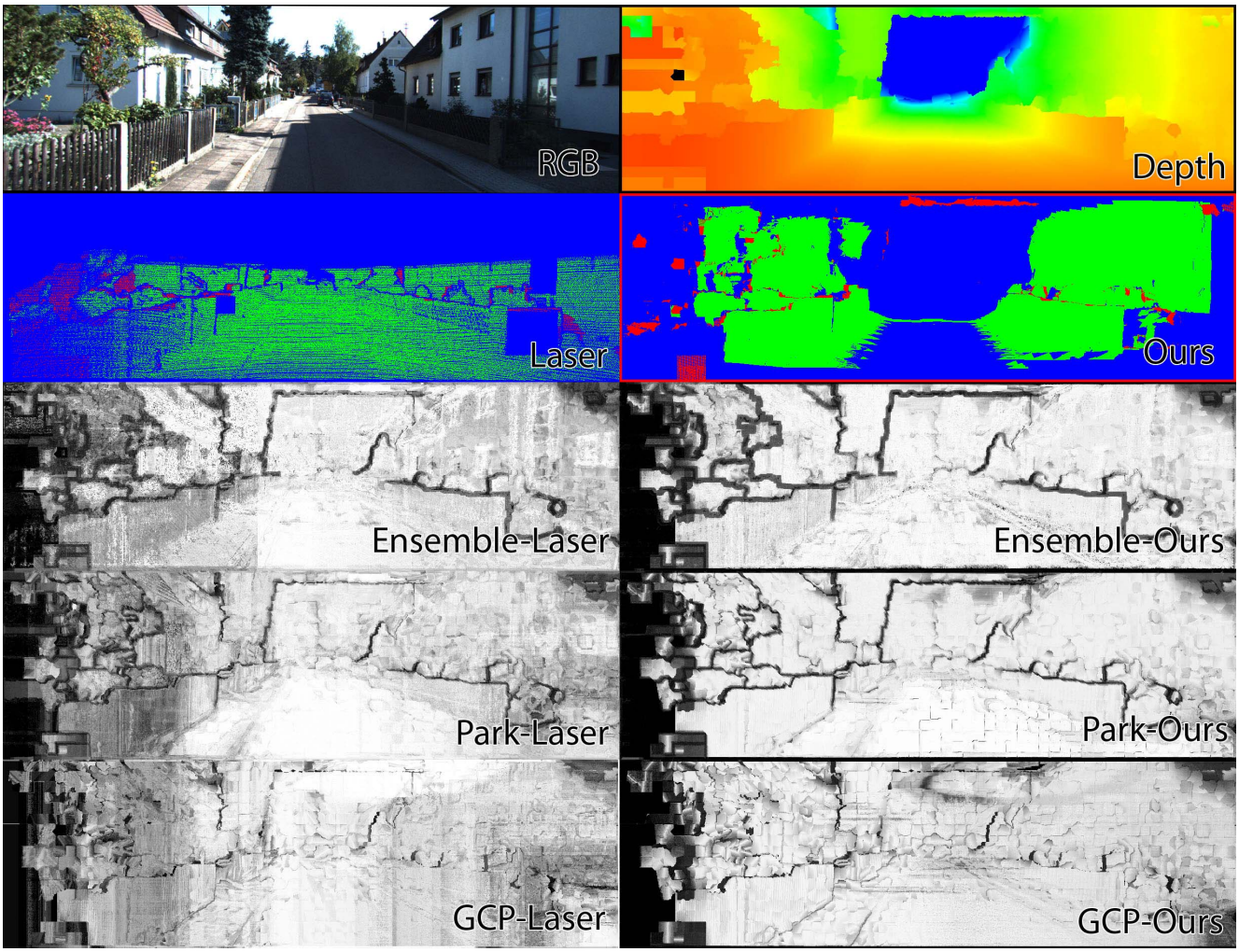


Figure 14. Visual comparison of outputs for frame 35 and SPS [6] as query algorithm (rank 185 - worst). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

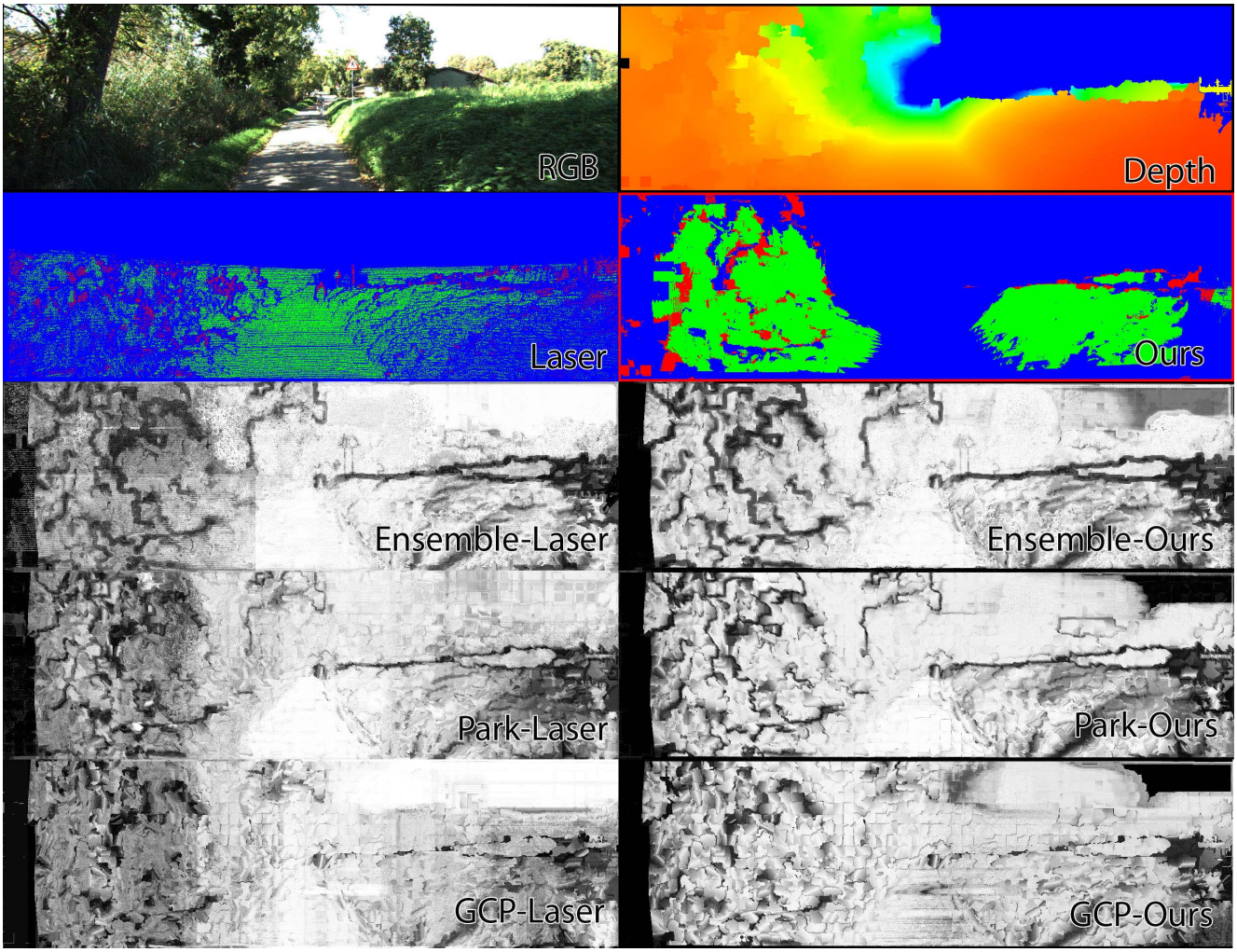


Figure 15. Visual comparison of outputs for frame 42 and SPS [6] as query algorithm (rank 186 - worst). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

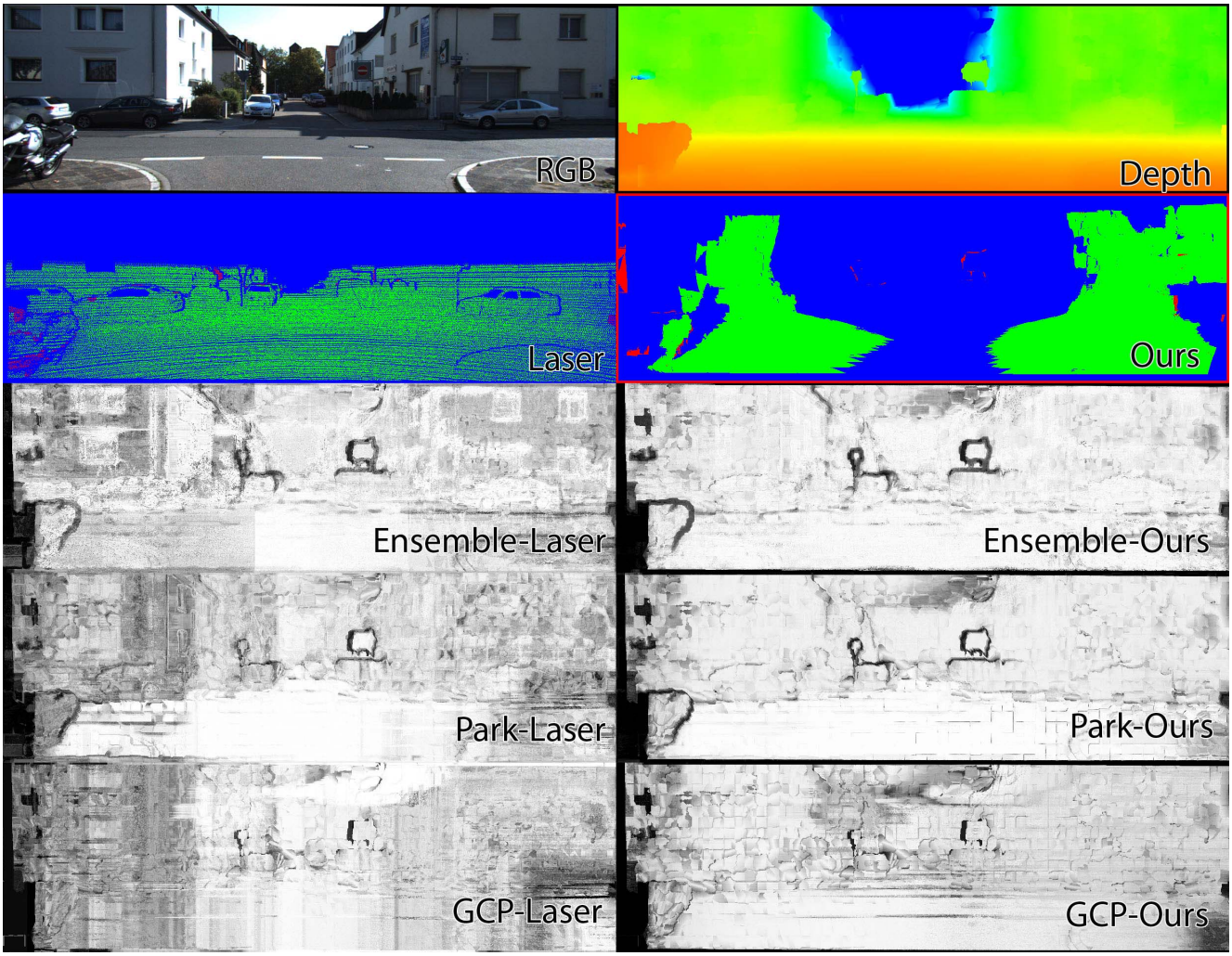


Figure 16. Visual comparison of outputs for frame 56 and SPS [6] as query algorithm (rank 2 - best). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

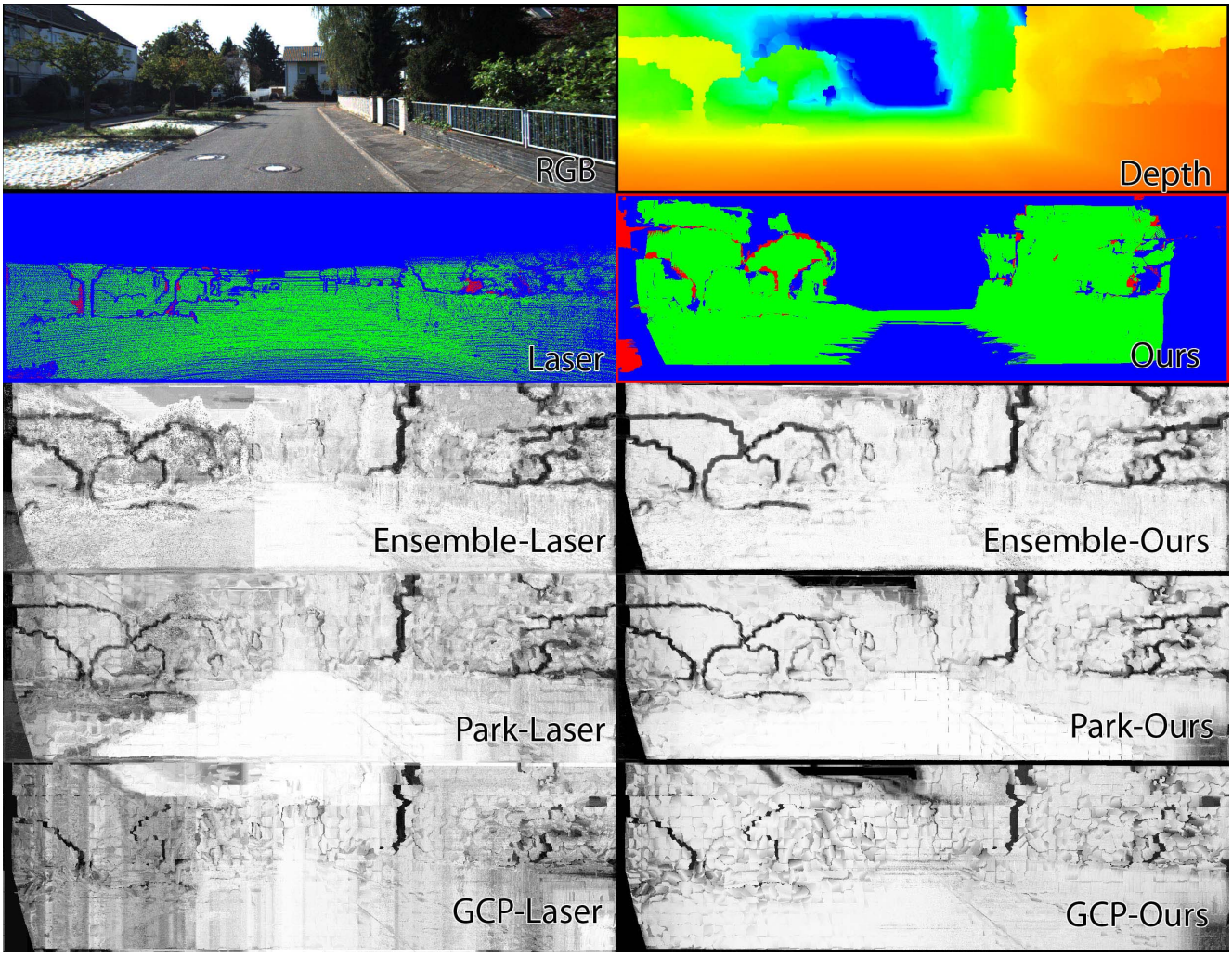


Figure 17. Visual comparison of outputs for frame 85 and SPS [6] as query algorithm (rank 92 - median). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

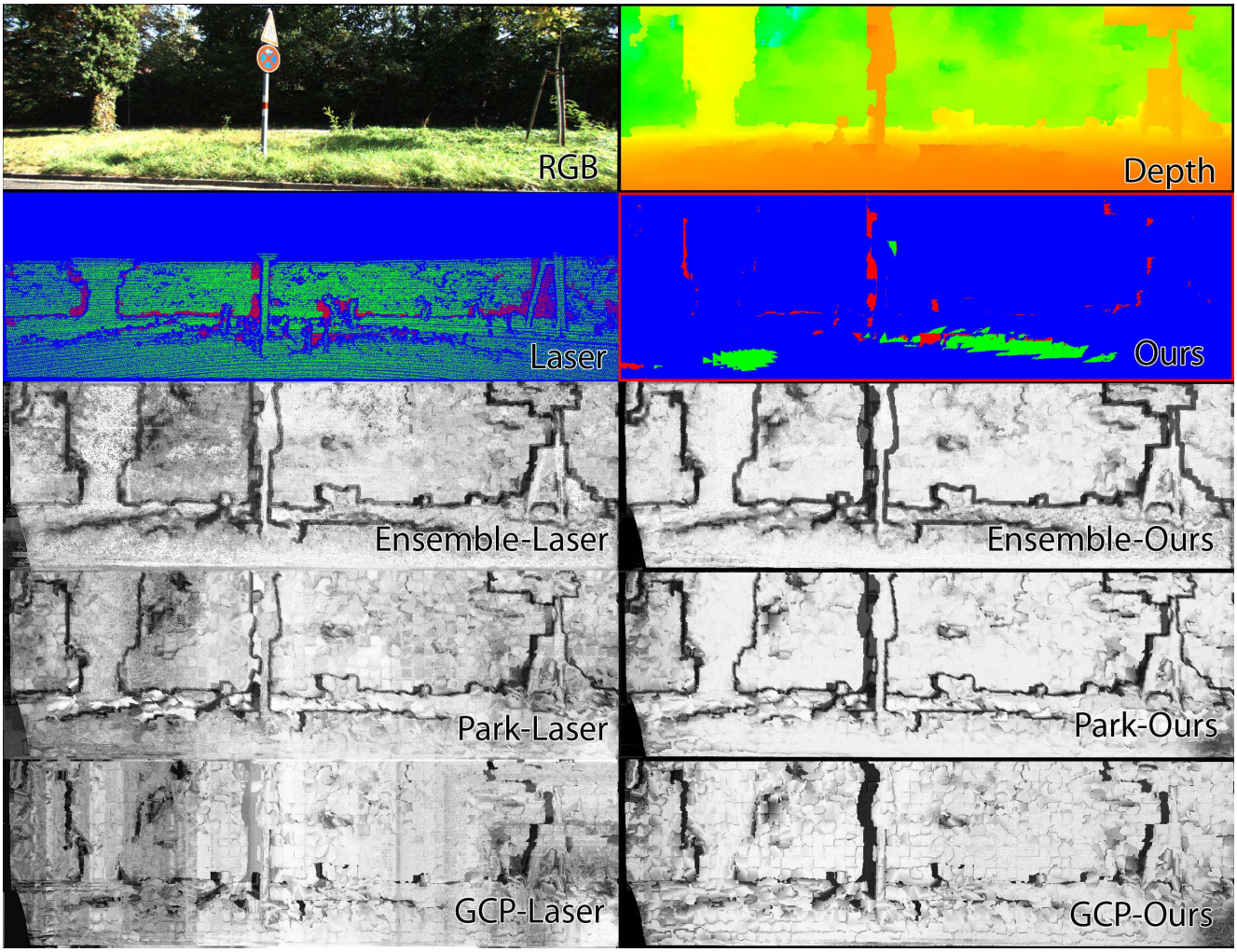


Figure 18. Visual comparison of outputs for frame 88 and SPS [6] as query algorithm (rank 94- median). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

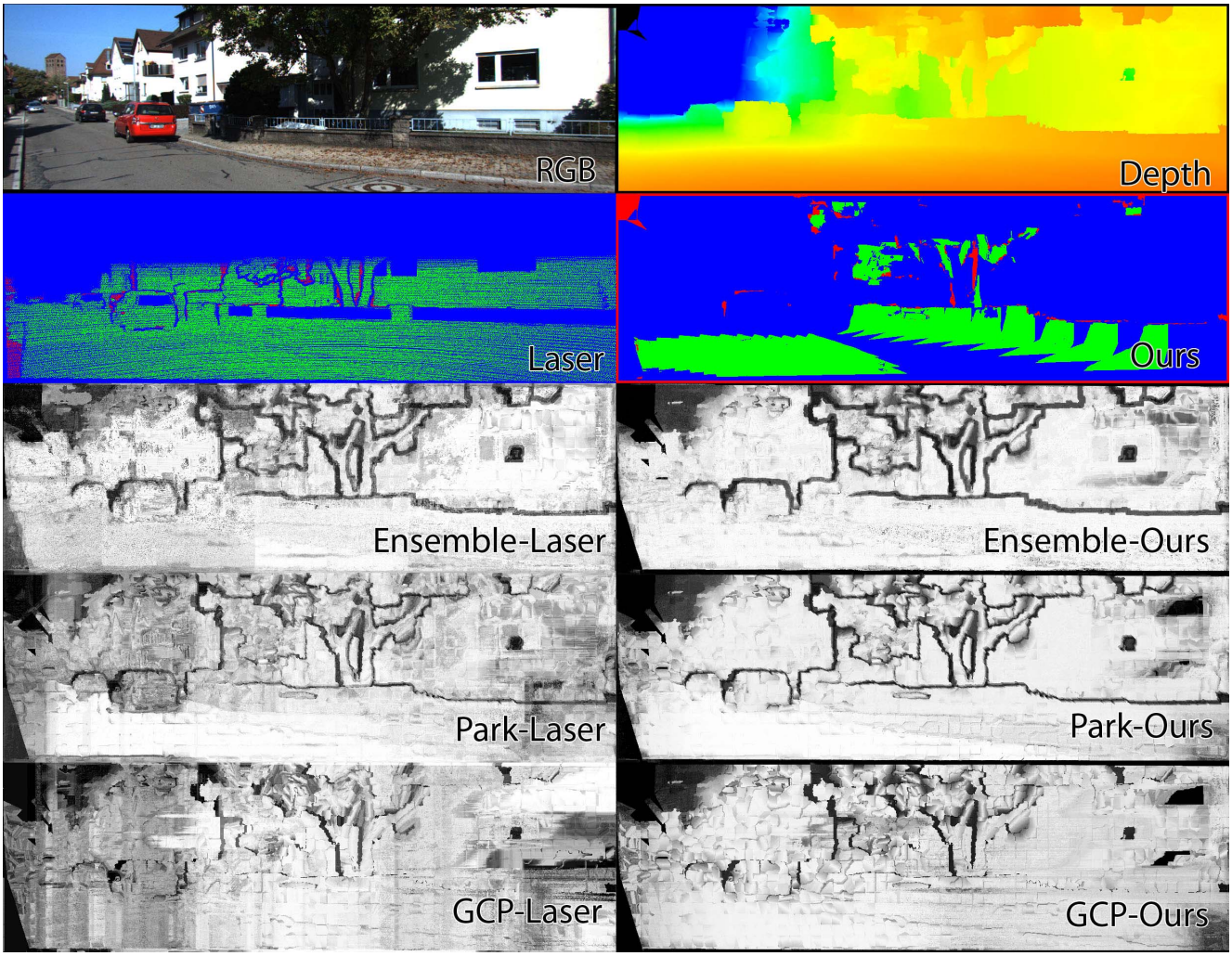


Figure 19. Visual comparison of outputs for frame 102 and SPS [6] as query algorithm (rank 1 - best). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

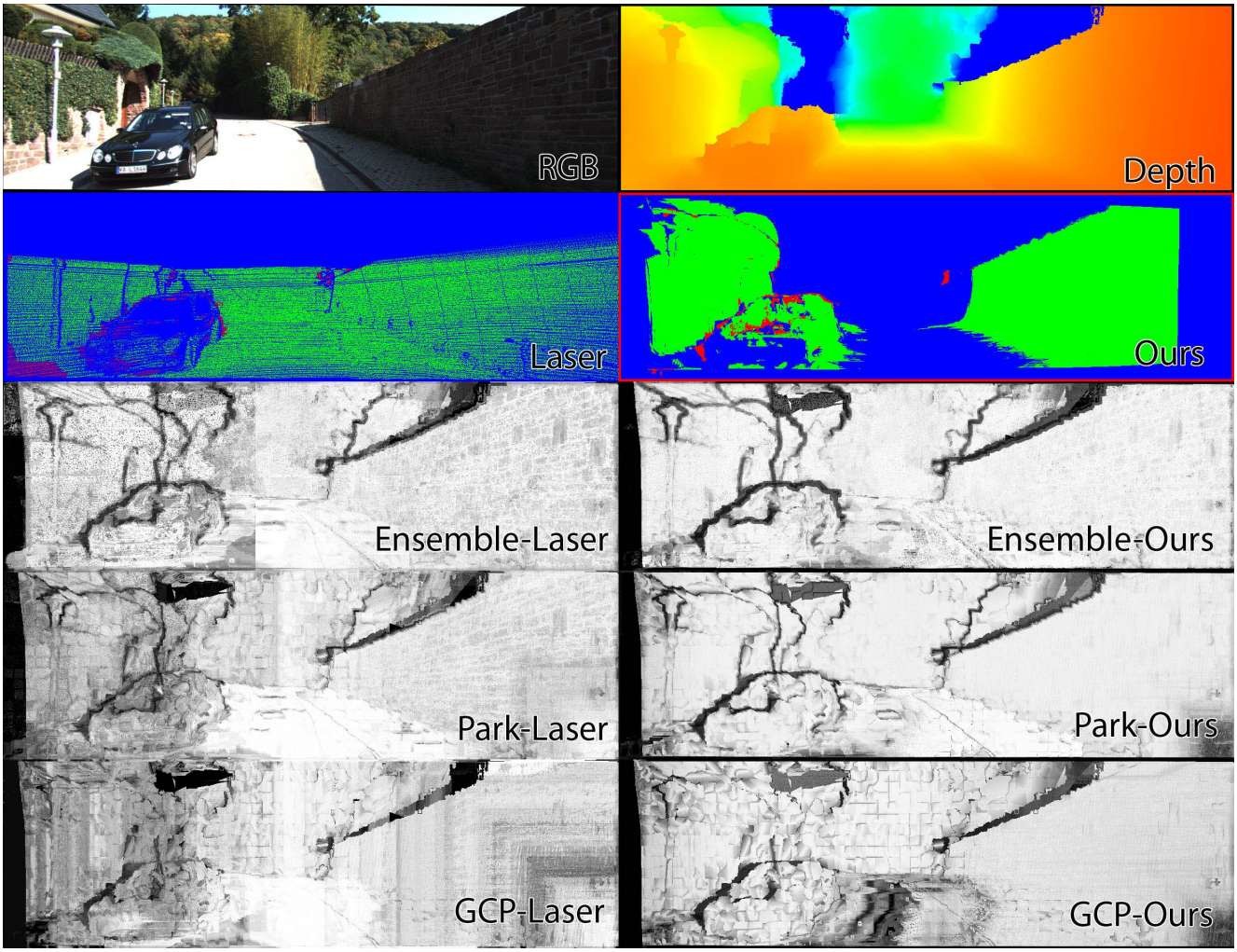


Figure 20. Visual comparison of outputs for frame 126 and SPS [6] as query algorithm (rank 93 - median). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

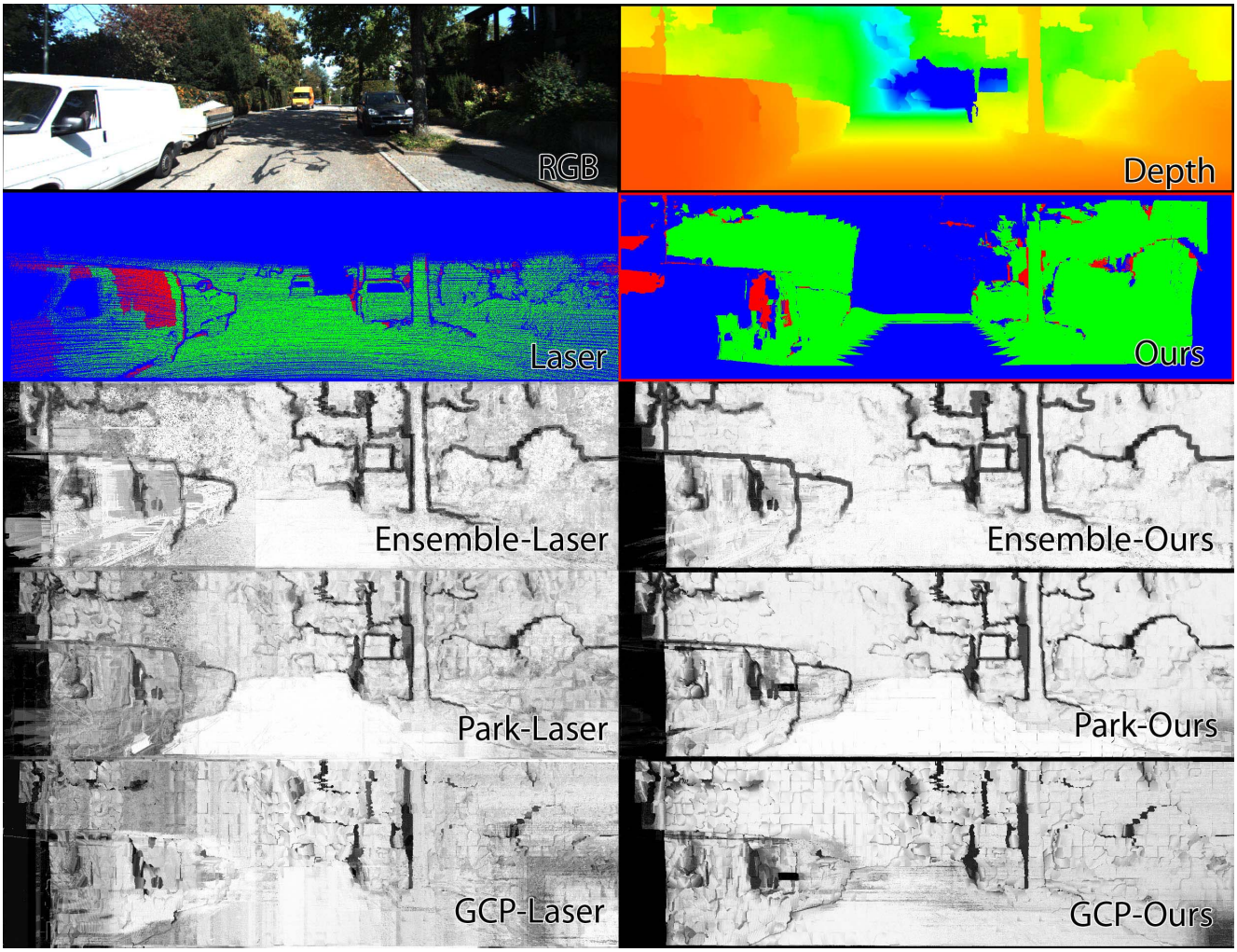


Figure 21. Visual comparison of outputs for frame 151 and SPS [6] as query algorithm (rank 184 - worst). Top row: RGB input image and SPS [6] depthmap. The color in the depth images ranges from blue (far away) to red (very close). Second row: Label images computed with laser ground truth and with our approach. In the label images the color green stands for positive samples, red for negative and blue is ignored during training and evaluation. In the remaining rows, we display the confidence prediction output for all combinations of confidence prediction algorithm (Ensemble [2], GCP [5], Park [3]) and training data (Laser and Ours). The color ranges from black (low confidence) to white (high confidence).

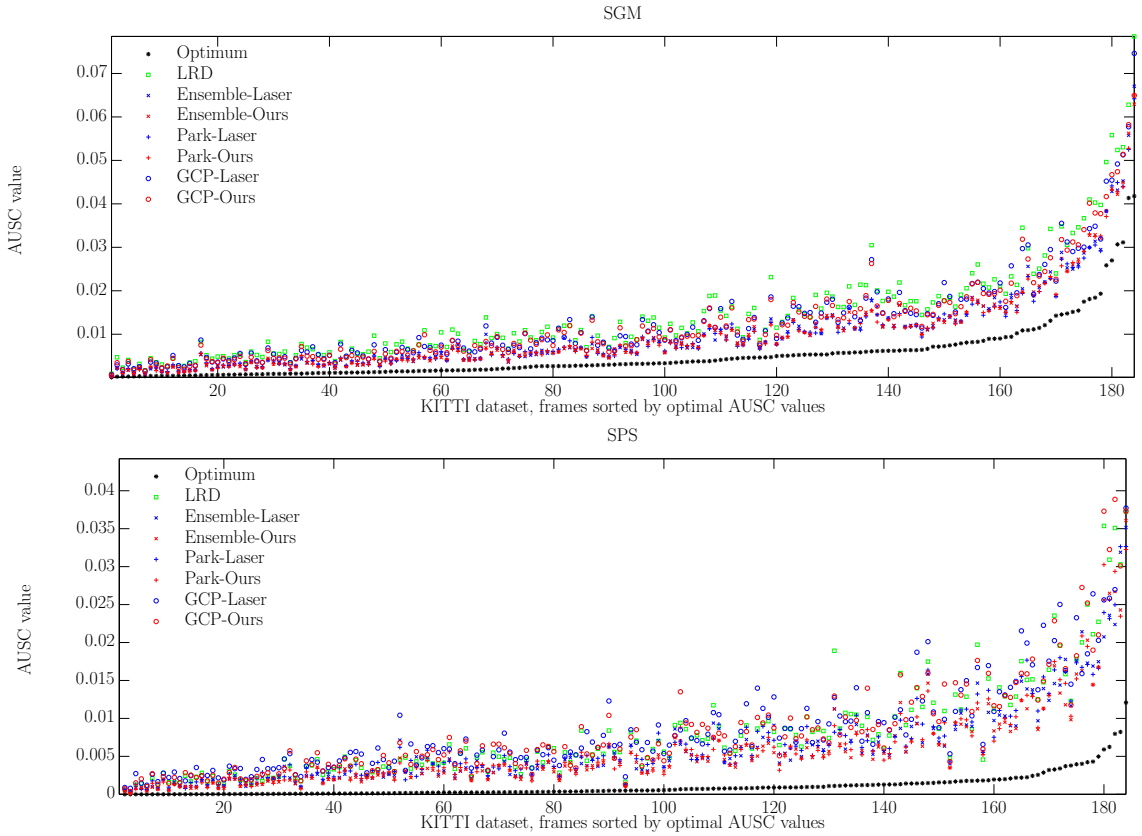


Figure 22. Area under the Sparsification Curve (AUSC) values for all frames of the KITTI training dataset minus the eight frames used for training. The frames were sorted according to the optimal area under the curve value. We display all combinations of query algorithm (SGM [4] and SPS [6]), confidence prediction algorithm (Ensemble [2],GCP [5],Park [3]) and training data(Laser and Ours). As a baseline method we also show the Left-Right disparity Difference (LRD).