Supplementary Material for SIGNet

Additional ablation studies on loss augmentations: As mentioned in our paper, the heuristic loss functions are not effective even after careful hyper-parameter tuning. This motivated us to design a learnable loss function (transfer network), which does improve upon the baseline as shown in Table 4 of our paper.

Method	Error metrics				Accuracy ($\delta <$)		
	AbsRel	SqRel	RSME	RSME_{log}	1.25^1	1.25^2	1.25^3
Yin *et al.*[51]	0.155	1.296	5.857	0.233	**0.793**	0.931	0.973
Warp Loss	0.169	1.246	6.233	0.254	0.750	0.917	0.968
Mask Loss	0.165	1.204	**5.593**	0.232	0.769	0.926	0.974
Edge Loss	0.163	1.230	5.961	0.243	0.774	0.924	0.970
Transfer	**0.150**	**1.141**	5.709	**0.231**	0.792	**0.934**	**0.974**

Table 1: Depth predictions for different loss augmentations (without using scale normalization). Here Warp Loss, Mask Loss and Edge Loss are on par or not as good as the baseline, whereas Transfer Network surpasses the baseline in almost all the metrics.

Why does ExtraNet only work for PoseNet? In the ablation studies in Section 5.3, we tested the contribution of semantic information in each module. The result suggests that vanilla PoseNet benefits from semantics only marginally, which might due to its simple structure. By adding Extra Network (ExtraNet) to PoseNet, our model gained further improvement. ExtraNet does not benefit DepthNet because the latter has already had a complicated structure as shown in Figure 1.

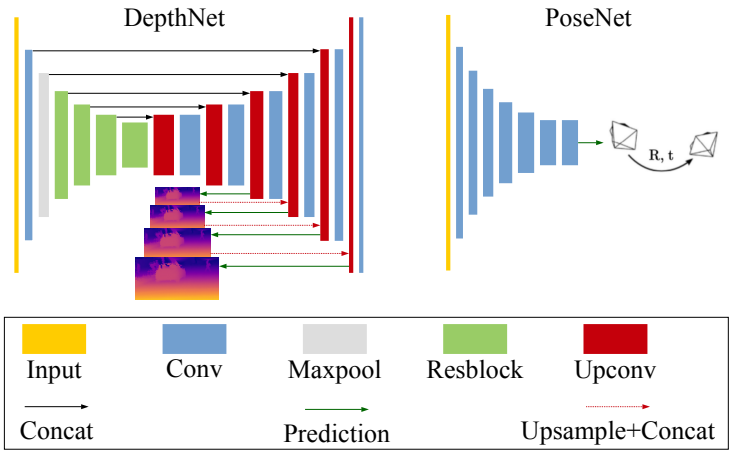

Figure 1: Network structures for DepthNet and PoseNet

More visualization results for depth estimation: In the rest of the supplementary material, we will present extra visualization results to help readers understand where our semantic-aided model improved the most. We compared the prediction result from our best model in Table 1 with Yin *et al.* [51] and ground truth. We followed [13] to plot the prediction result using disparity heatmaps. The following results show that our model can gain improvement from regions belonging to cars and other dynamic classes.

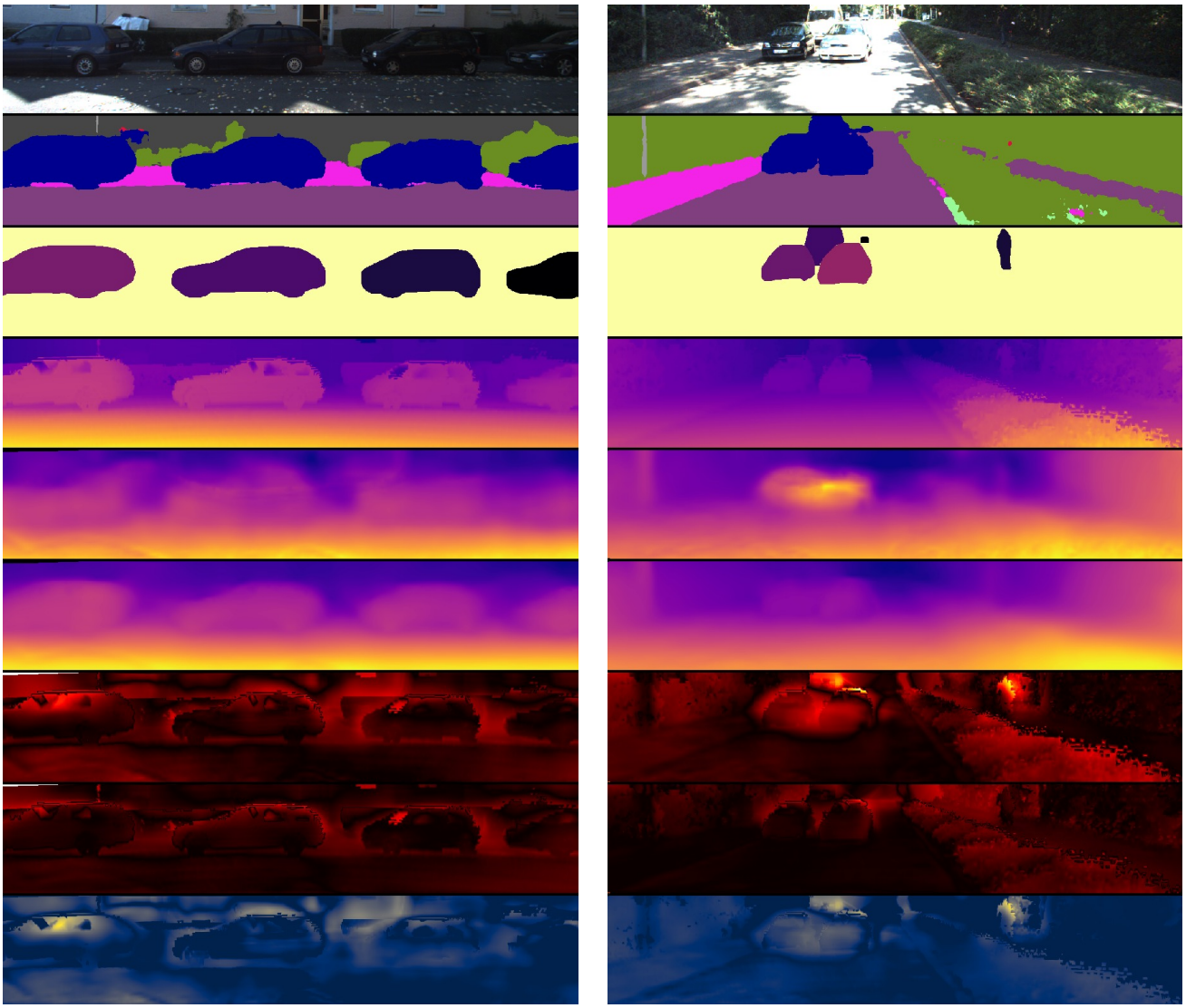

Figure 2: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

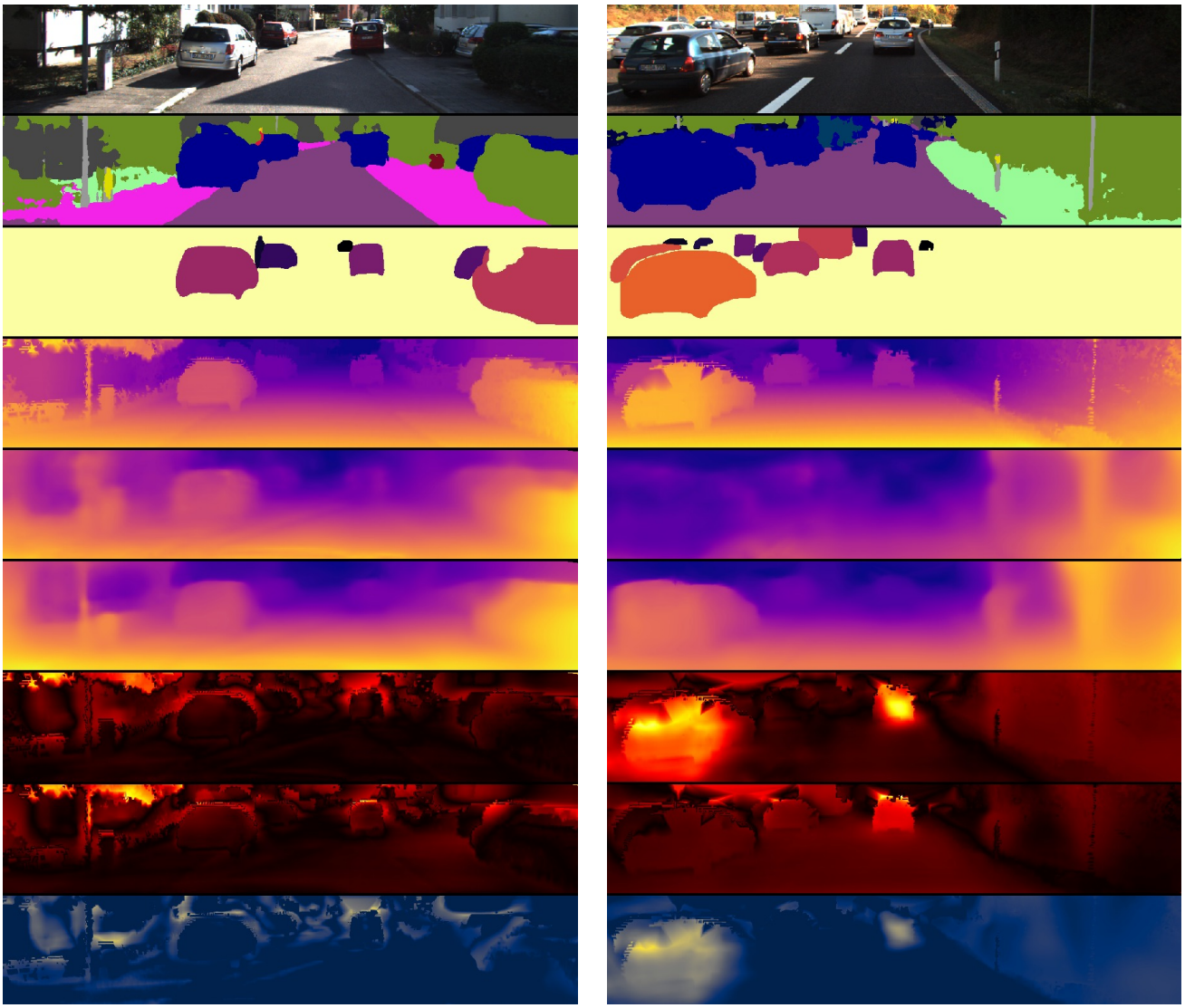

Figure 3: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

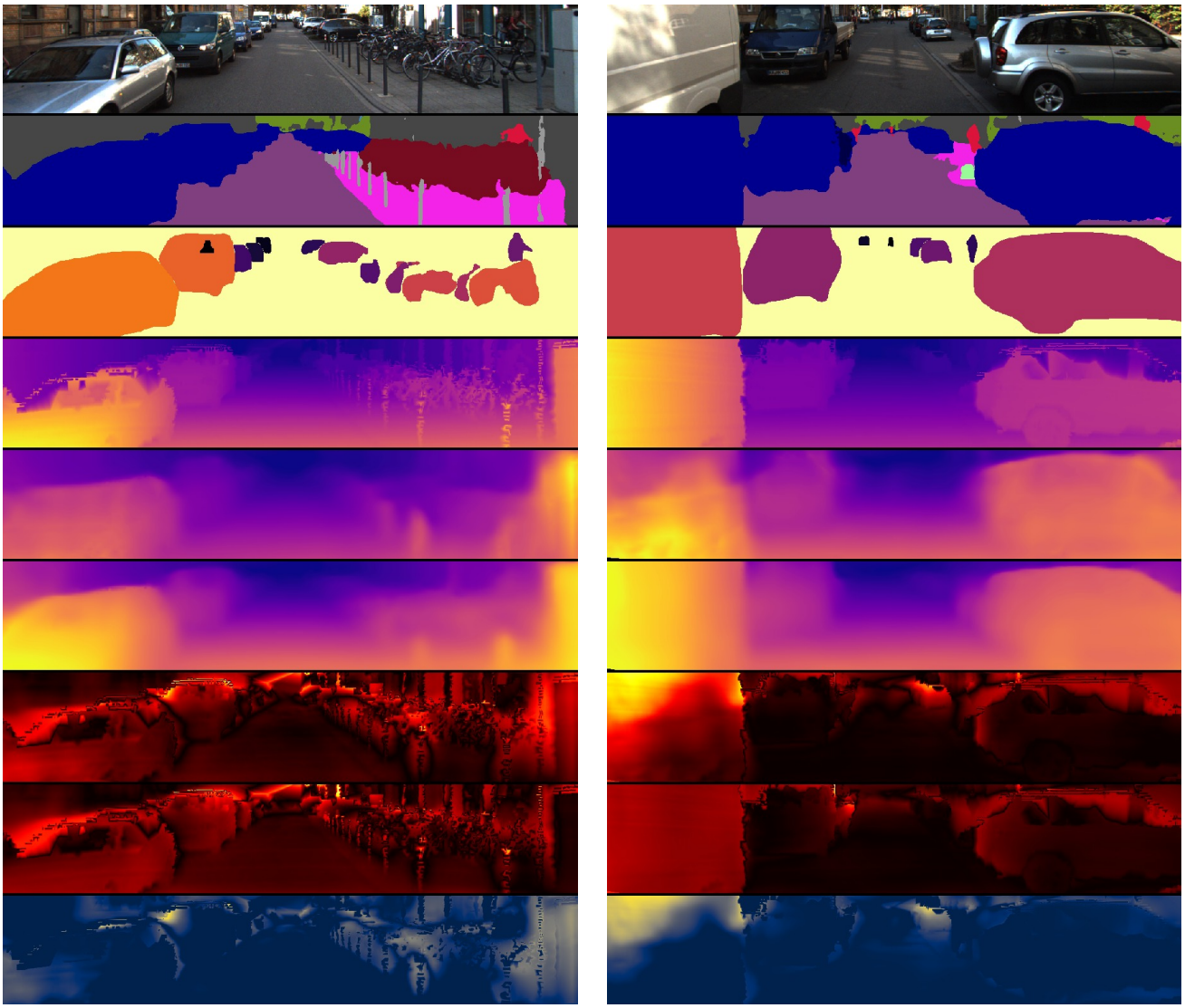

Figure 4: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

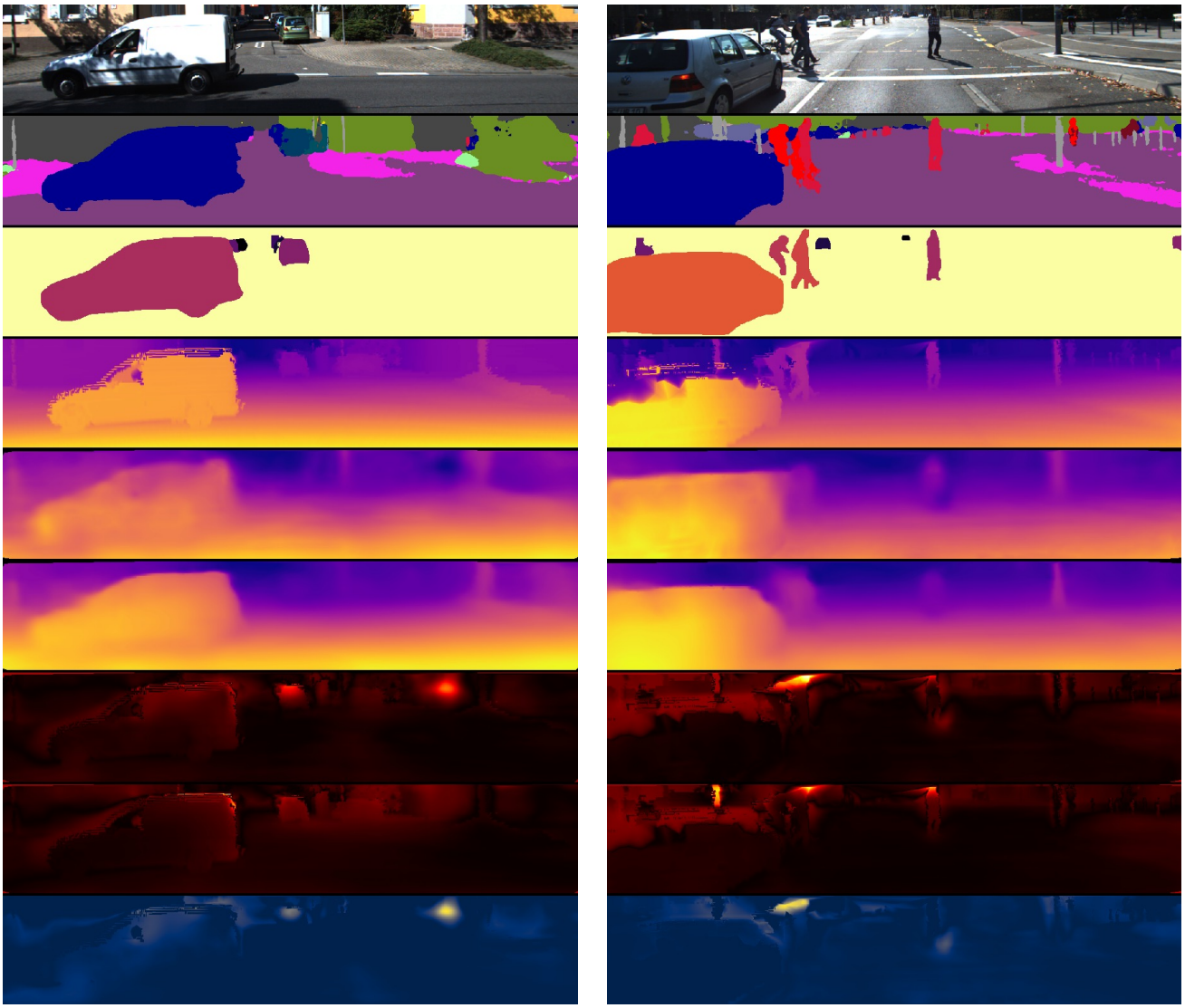

Figure 5: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

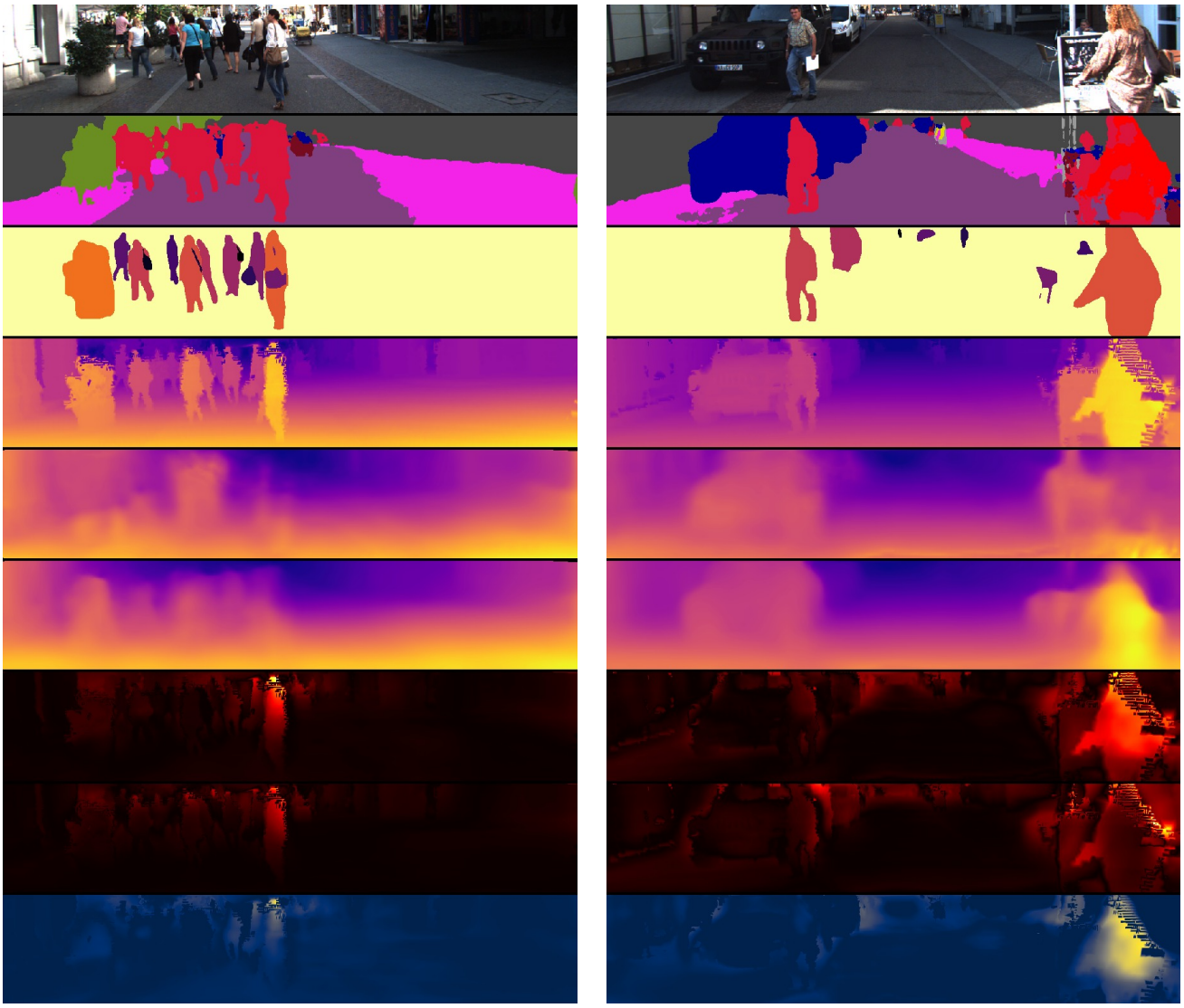

Figure 6: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

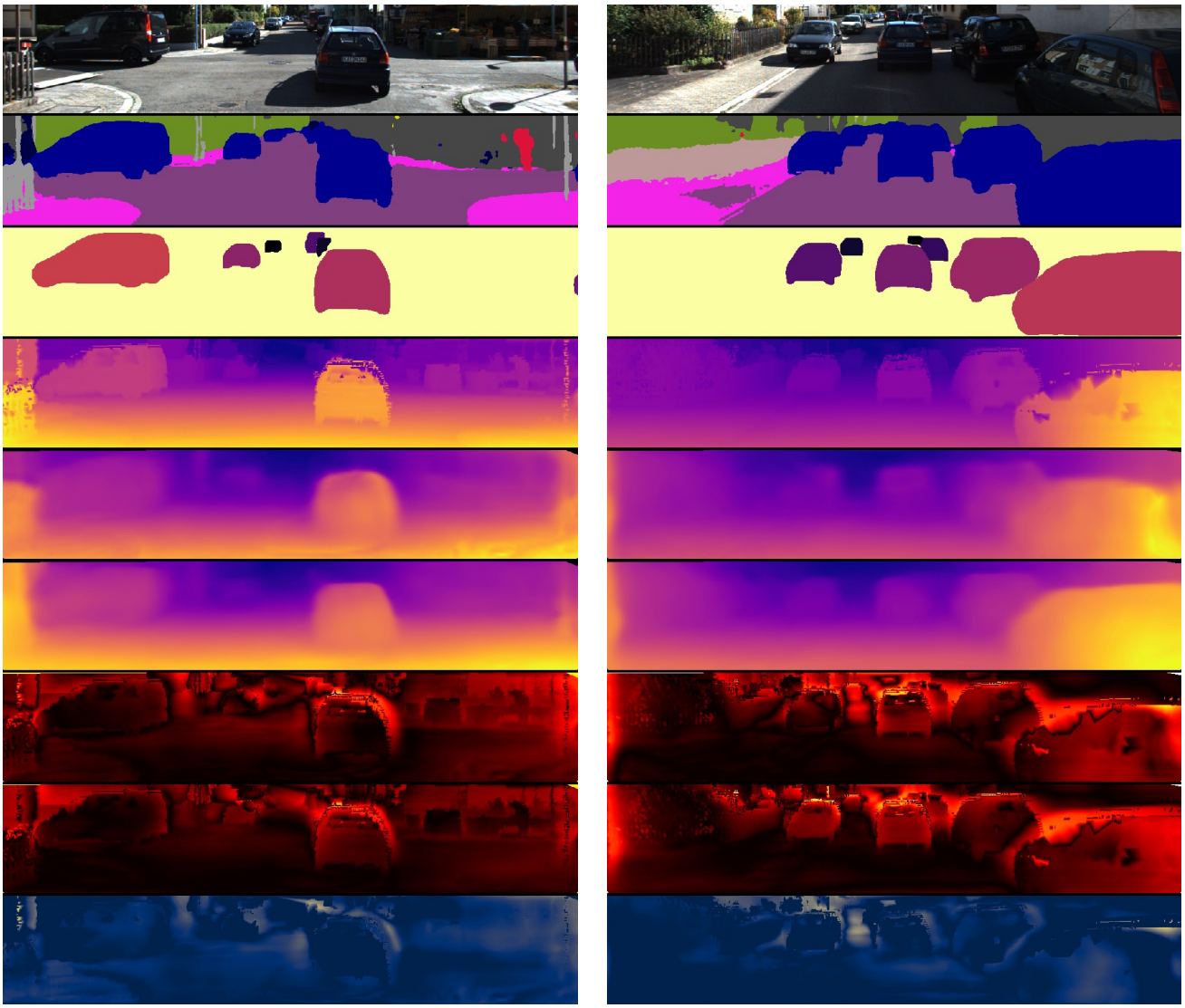

Figure 7: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

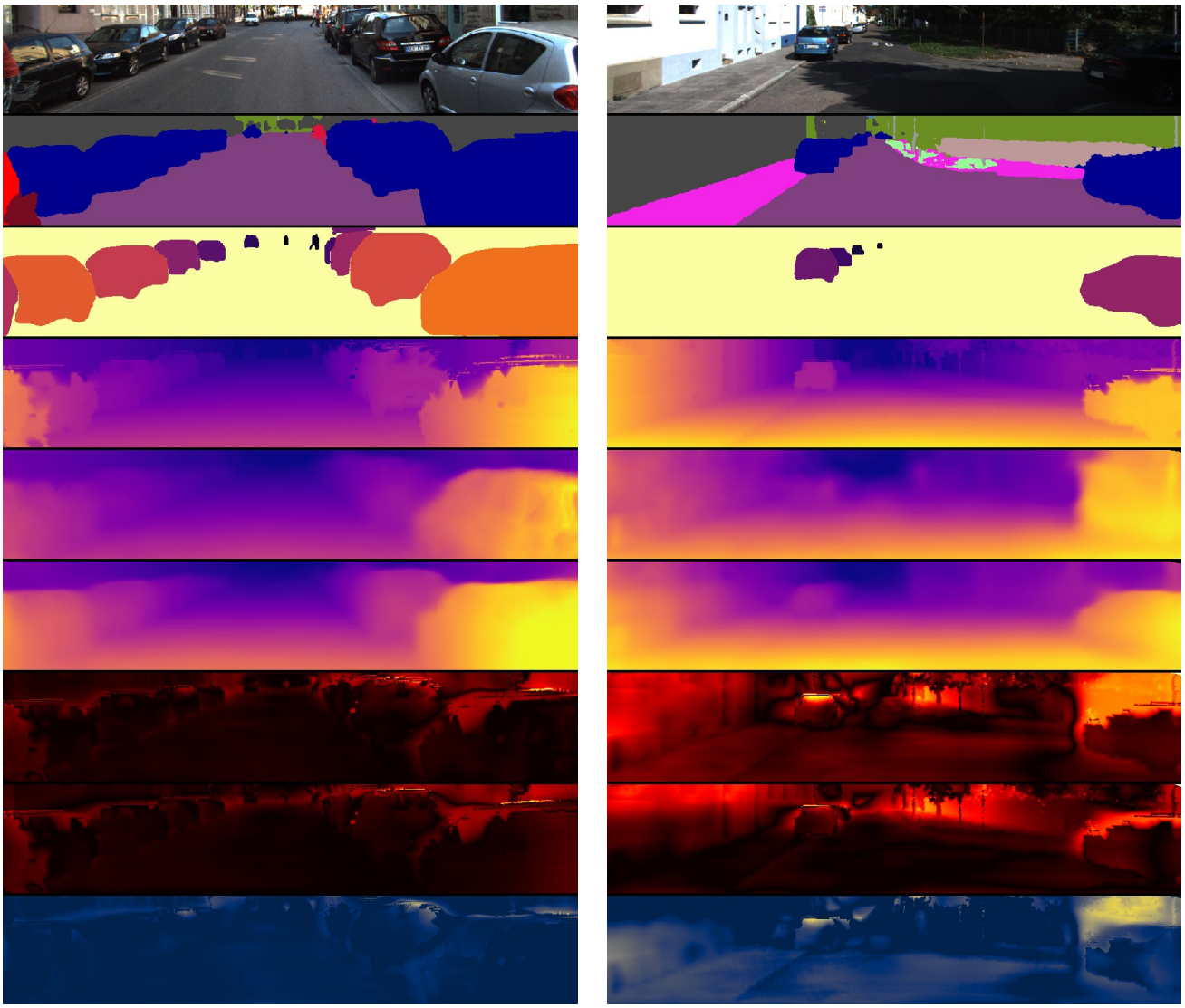

Figure 8: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

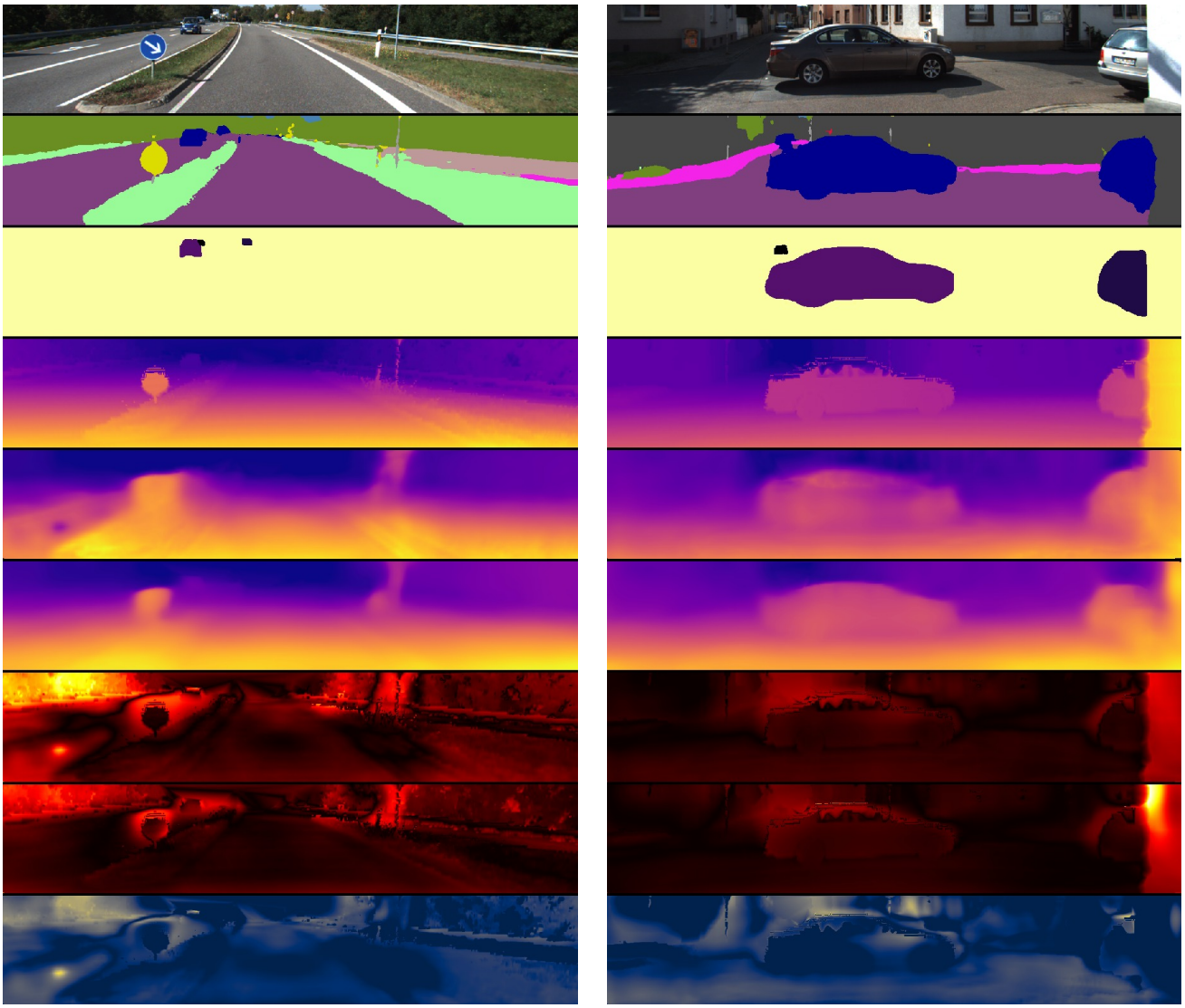

Figure 9: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

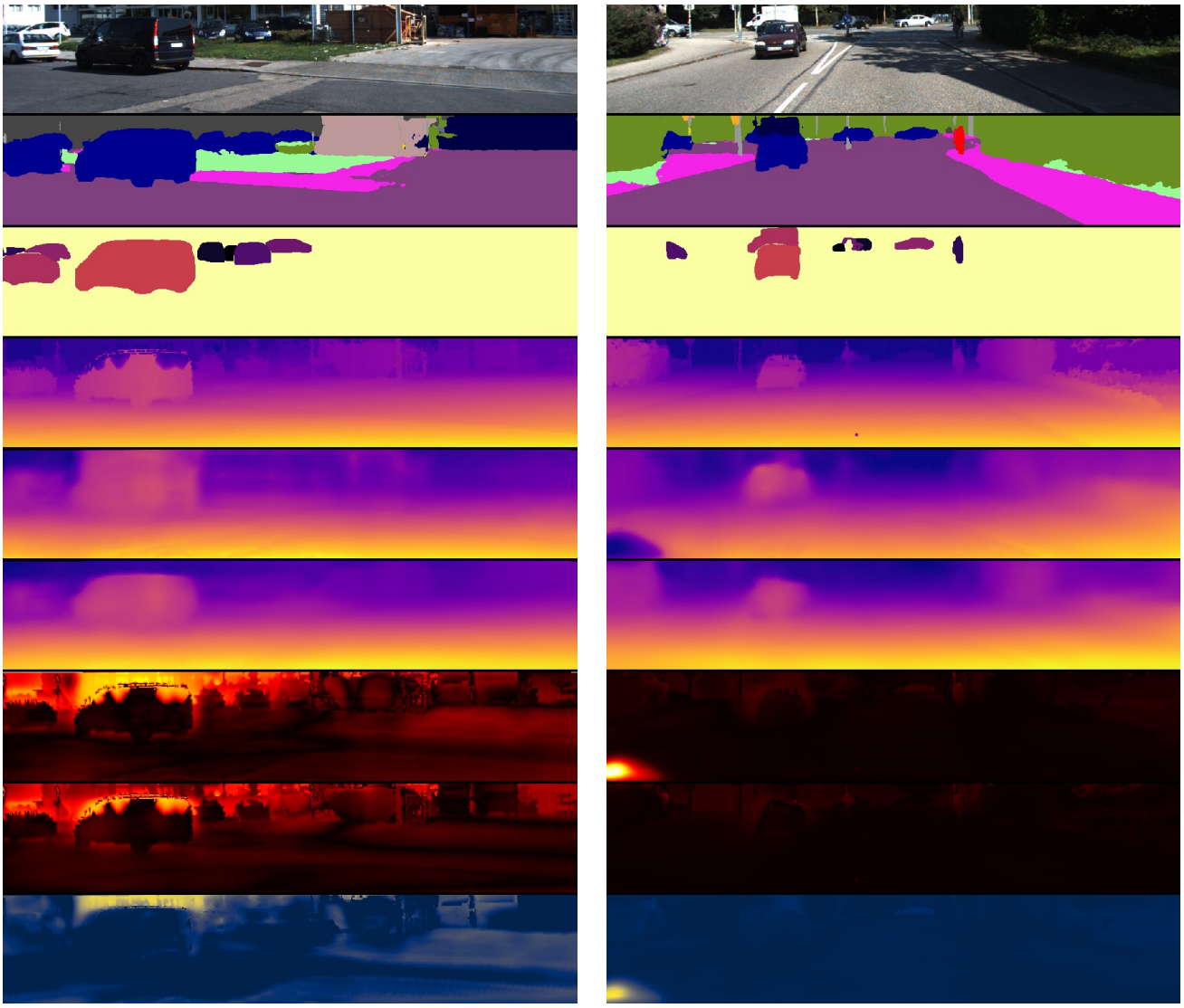

Figure 10: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al.* [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects

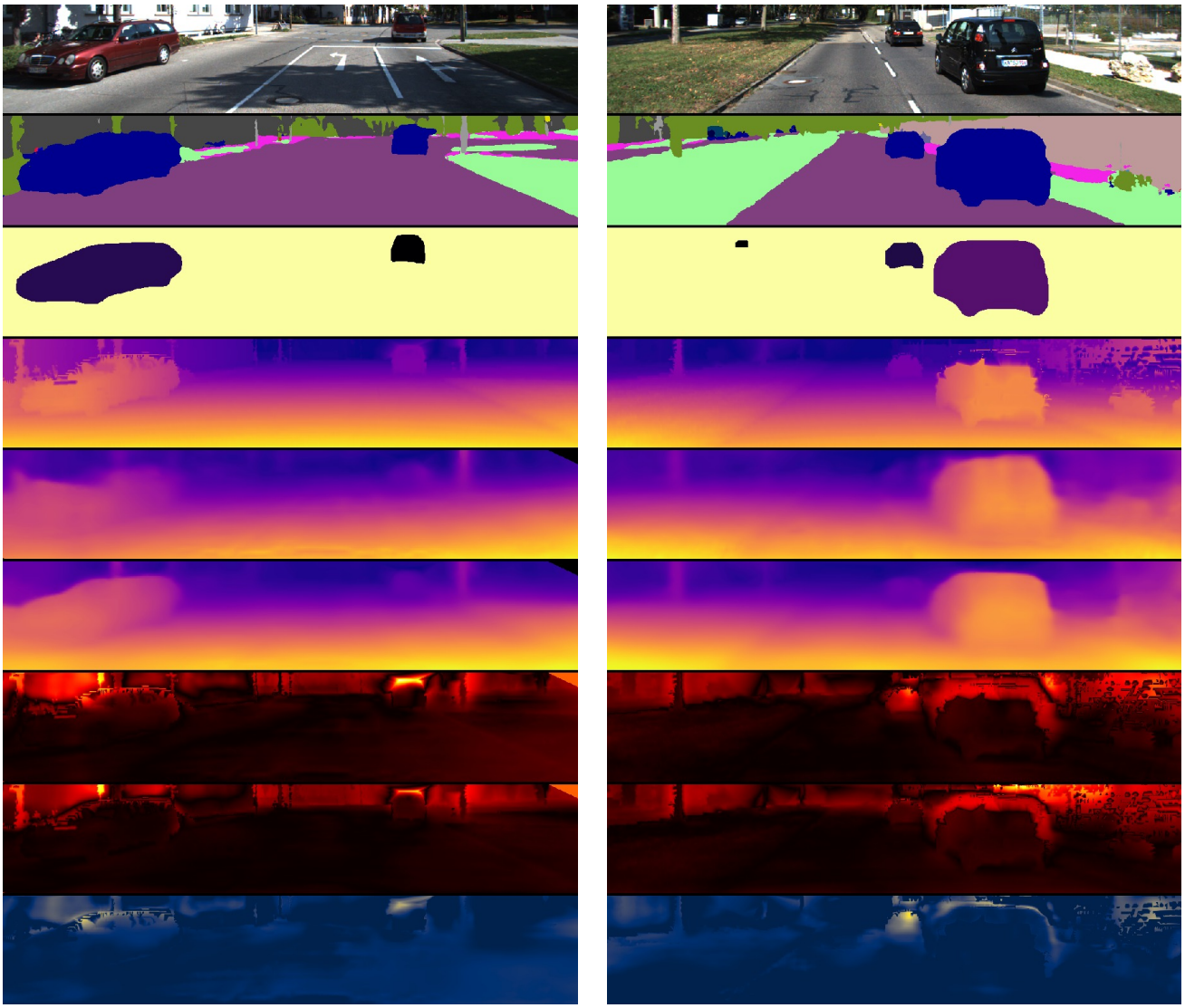

Figure 11: Top to bottom: input image, semantic segmentation, instance segmentation, ground truth disparity map, disparity prediction from baseline(Yin *et al*. [51]) , disparity prediction from ours, AbsRel error map of baseline models, AbsRel error map of ours and the improvement region compared to baseline. For the purpose of visualization, disparity maps are interpolated and cropped[13]. For all heatmaps, darker means smaller value (disparity, error or improvement). Typical image regions where we do better include cars, pedestrians and other common dynamic objects