DiffusionMTL: Learning Multi-Task Denoising Diffusion Model from Partially Annotated Data

Supplemental Material

Hanrong Ye and Dan Xu✉
Department of Computer Science and Engineering, HKUST
Clear Water Bay, Kowloon, Hong Kong
{hyeae, danxu}@cse.ust.hk

1. Supplemental Implementation Details

1.1. Additional Details about DiffusionMTL

Multi-Task Denoising Diffusion Network. This section provides additional details about the implementation of DiffusionMTL on different datasets. For our experiments, we set the default diffusion steps to 2 using a linear variance scheduler with a range from 10^{-3} to 10^{-2}. All self-attention blocks in the Denoiser use a single head.

Loss functions. For semantic segmentation, human parising, saliency detection, and boundary detection, we use cross-entropy loss. For depth and surface normal estimation, we opt for L1 loss. The multi-task loss balance weights are the same as those used in [3].

1.2. Implementation Details on Different Datasets

For all three partial-labeling benchmarks (PASCAL, NYUD, and Cityscapes), we use exactly the same image-task label mappings as those used in [3].

PASCAL On PASCAL-Context [2] (abbreviated as "PASCAL"), in the one-label setting, there are 1000, 999, 1000, 1000, 999 images separately labeled for semantic segmentation, human parsing, surface normal estimation, saliency detection, and boundary detection. In the random-label setting, there are 450, 2553, 2480, 2445, and 2557 images labeled for semantic segmentation, human parsing, surface normal estimation, saliency detection, and boundary detection, respectively, We pad the images to a resolution of 512×512. We use the Adam optimizer and a polynomial learning rate scheduler with a base learning rate of 2×10^{-5}. All models are trained for 100 epochs with a batch size of 6. We adopt the same data augmentations as in [5], which include random scaling, cropping, random horizontal flipping, and color jittering.

NYUD [4] In the one-label setting, 265 images are labeled for semantic segmentation, 265 images are labeled for monocular depth estimation, and 265 images are labeled for surface normal estimation. In the random-label setting, 392, 408, and 385 images are respectively labeled for these tasks. The images are resized to a resolution of 288×384. We use the Adam optimizer and a polynomial learning rate scheduler with a base learning rate of 2×10^{-5}. All models are trained for 200 epochs with a batch size of 4. We adopt the same data augmentations as in [3], which include random cropping and random horizontal flipping.

Method	#Params	FLOPS	Train GPU Mem	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Perf Δ_m ↑
MTL Baseline	157M	608G	6163M	49.71	56.00	74.50	16.85	62.80	-2.85%
XTC [3]	173M	608G	6409M	55.08	56.72	77.06	16.93	63.70	+0.37%
MTINet [5]	281M	589G	11533M	54.32	57.73	77.12	16.41	64.20	+1.21%
InvPT [6]	141M	1182G	9993M	56.96	57.05	77.19	16.80	63.20	+1.27%
DiffusionMTL (Prediction)	133M	628G	5703M	**59.43**	56.79	77.57	16.20	64.00	+3.23%
DiffusionMTL (Feature)	133M	676G	5811M	57.78	**58.98**	**77.82**	**16.11**	**64.50**	**+3.65%**

Table 1. One-label setting on PASCAL with ResNet-18 backbone.

Cityscapes [1] As we only evaluate two tasks on the Cityscapes dataset, the one-label setting is equivalent to the random-label setting. The training split contains 1,487 labeled images for semantic segmentation and 1,488 labeled images for monocular depth estimation. We adopt a learning rate of 10^{-4}. All models are trained for 200 epochs with a batch size of 8. The images are resized to a resolution of 128×256. We adopt the data augmentations in [3], which include random cropping and random horizontal flipping.

2. Additional Quantitative Study

2.1. Comparison with SOTA refinement methods.

We conduct extensive experiments to compare our proposal with previous SOTA MTL refinement methods, including MTI-Net [5] and InvPT [6], based on the ResNet-18 baseline under the one-label setting on PASCAL dataset. The results, presented in Table 1, demonstrate the superior performance of DiffusionMTL across all tasks.

2.2. Comparison under Fully-Annotated Setting

Our method can be applied to fully-annotated benchmarks. We conduct experiments on fully-annotated PASCAL dataset using ResNet-18 and show the results in Table 2. Our method demonstrates stronger performance compared to both the baseline as well as the state-of-the-art (SOTA) method XTC [3] and InvPT [6].

Method	#Params	FLOPS	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Perf Δ_m ↑
STL Baseline	219M	817G	52.56	62.21	82.75	14.12	68.90	-
MTL Baseline	157M	608G	62.91	57.37	81.82	14.49	66.40	+0.90%
XTC [3]	173M	608G	63.29	57.93	82.09	14.48	66.50	+1.34%
InvPT [6]	141M	1182G	64.38	59.49	83.52	14.75	66.80	+2.31%
DiffusionMTL (Prediction)	133M	628G	64.31	58.68	83.07	14.44	67.10	+2.44%
DiffusionMTL (Feature)	133M	676G	**64.62**	**60.14**	**83.99**	**14.17**	**67.80**	**+3.84%**

Table 2. Fully-annotated setting on PASCAL with ResNet-18.

2.3. Computation and Memory Cost Comparison.

We have already shown the parameters and FLOPs comparison with the MTL baseline and XTC in Table 3 of our main paper. We further provide the training GPU memory in Table 1 of this document. Our method shows higher parameter/memory efficiency and comparable computational costs with significantly better performance.

3. Additional Qualitative Study

3.1. Denoising Effectiveness of DiffusionMTL

To assess the denoising performance of our model, we visually examine the noisy multi-task prediction maps generated through the diffusion process, as well as the denoised outputs produced by Prediction Diffusion based on ResNet-18 on Cityscapes dataset under a one-label training setting. The obtained results are showcased in Fig. 1 and Fig. 2. The effectiveness of our proposed DiffusionMTL is demonstrated by its ability to successfully denoise the noisy prediction maps, resulting in significantly improved multi-task predictions that align better with the ground-truth labels. These results serve as additional validation for our motivation behind designing a robust multi-task denoising diffusion framework, addressing the challenges inherent in the multi-task partially supervised learning problem.

3.2. Comparison with SOTA

In order to further demonstrate the performance advantage of DiffusionMTL, we present a set of randomly selected samples generated by our model and the previous state-of-the-art model (*i.e.*, XTC [3]) on Cityscapes in Fig. 3 and Fig. 4. We further compare the results on PASCAL in Fig. 5 and Fig. 6. These models are trained under the same one-label multi-task partially supervised learning setting. The superiority of prediction maps generated by DiffusionMTL in terms of accuracy is evident on both datasets. This compelling evidence serves to further validate the effectiveness of our proposed denoising diffusion model.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 111:98–136, 2010. 1

[3] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[5] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 1

[6] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 1, 2

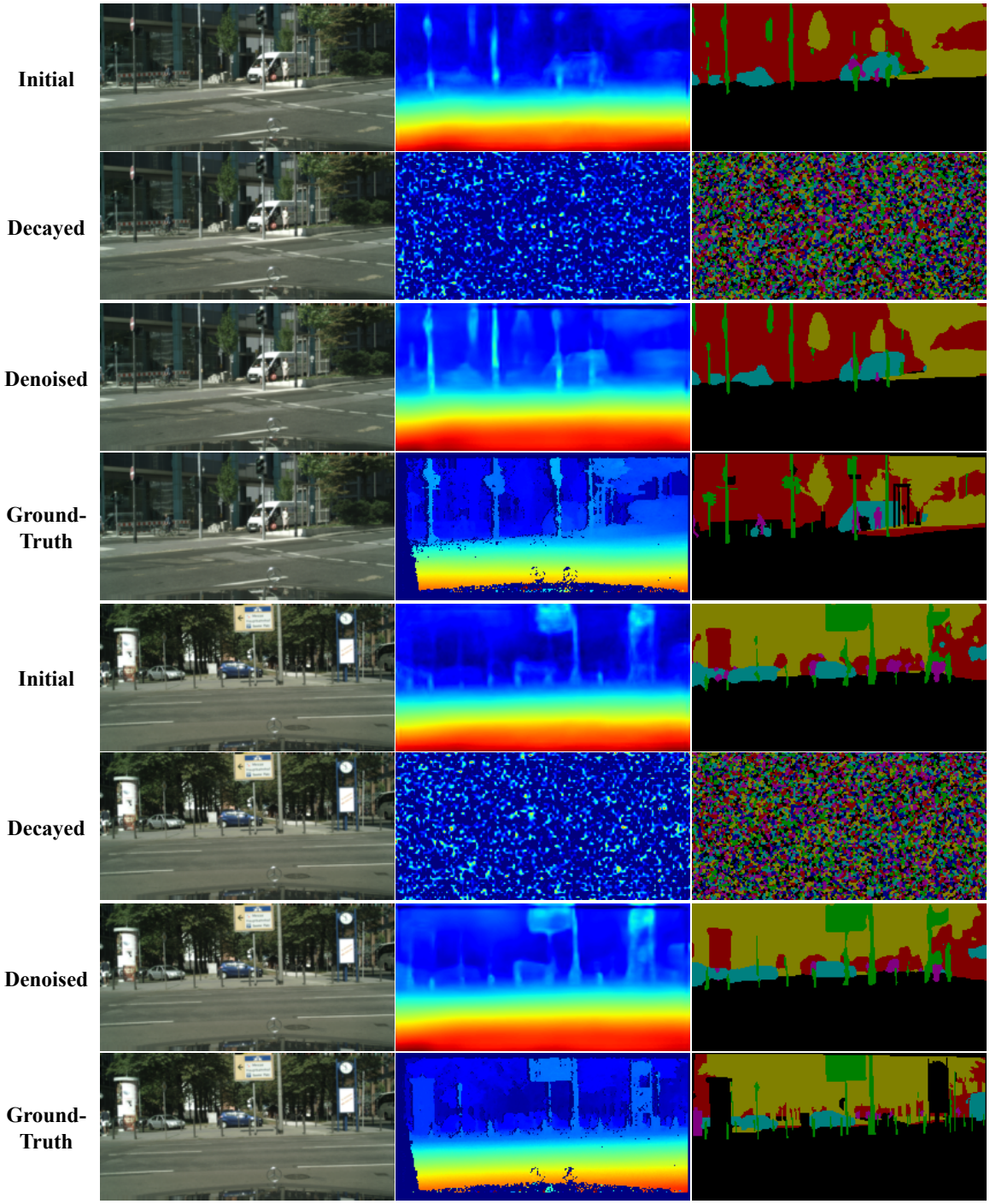

Figure 1. Qualitative comparison of the initial multi-task predictions, decayed predictions, our denoised results, and ground-truth labels on Cityscapes under one-label setting. Our DiffusionMTL is able to rectify noisy input and generate clean prediction maps. The model used in this comparison is trained on the Cityscapes dataset under the one-label MTPSL setting.

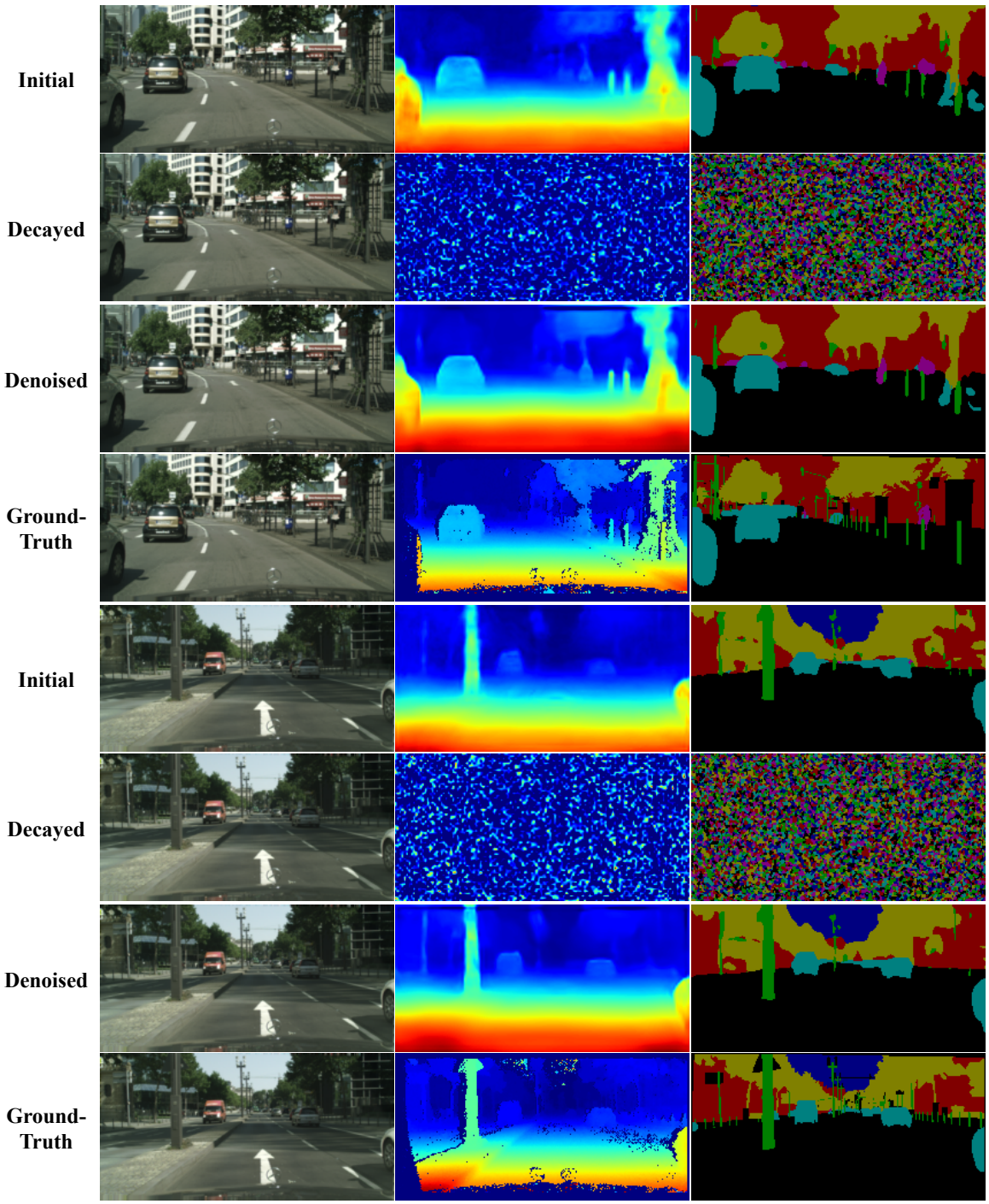

Figure 2. Qualitative comparison of the initial multi-task predictions, decayed predictions, our denoised results, and ground-truth labels on Cityscapes under one-label setting. Our DiffusionMTL is able to rectify noisy input and generate clean prediction maps. The model used in this comparison is trained on the Cityscapes dataset under the one-label MTPSL setting.

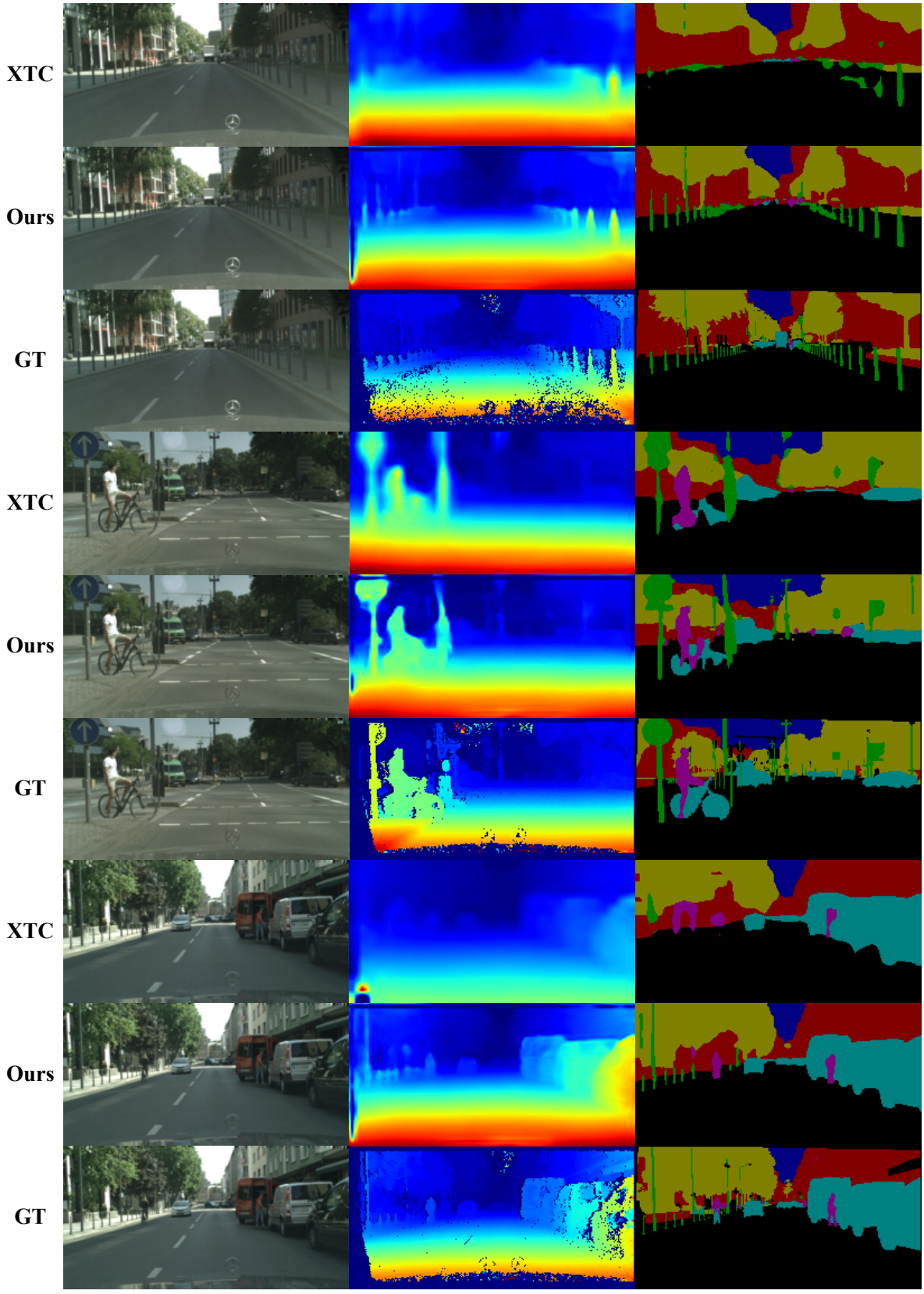

Figure 3. Qualitative comparison between our method and the state-of-the-art method (*i.e.* XTC [3]) for depth estimation and semantic segmentation tasks in Cityscapes dataset, using the same ResNet-18 backbone. Our DiffusionMTL approach outperforms the previous state-of-the-art method in producing superior prediction maps. Notably, each training sample is labeled for only one task.

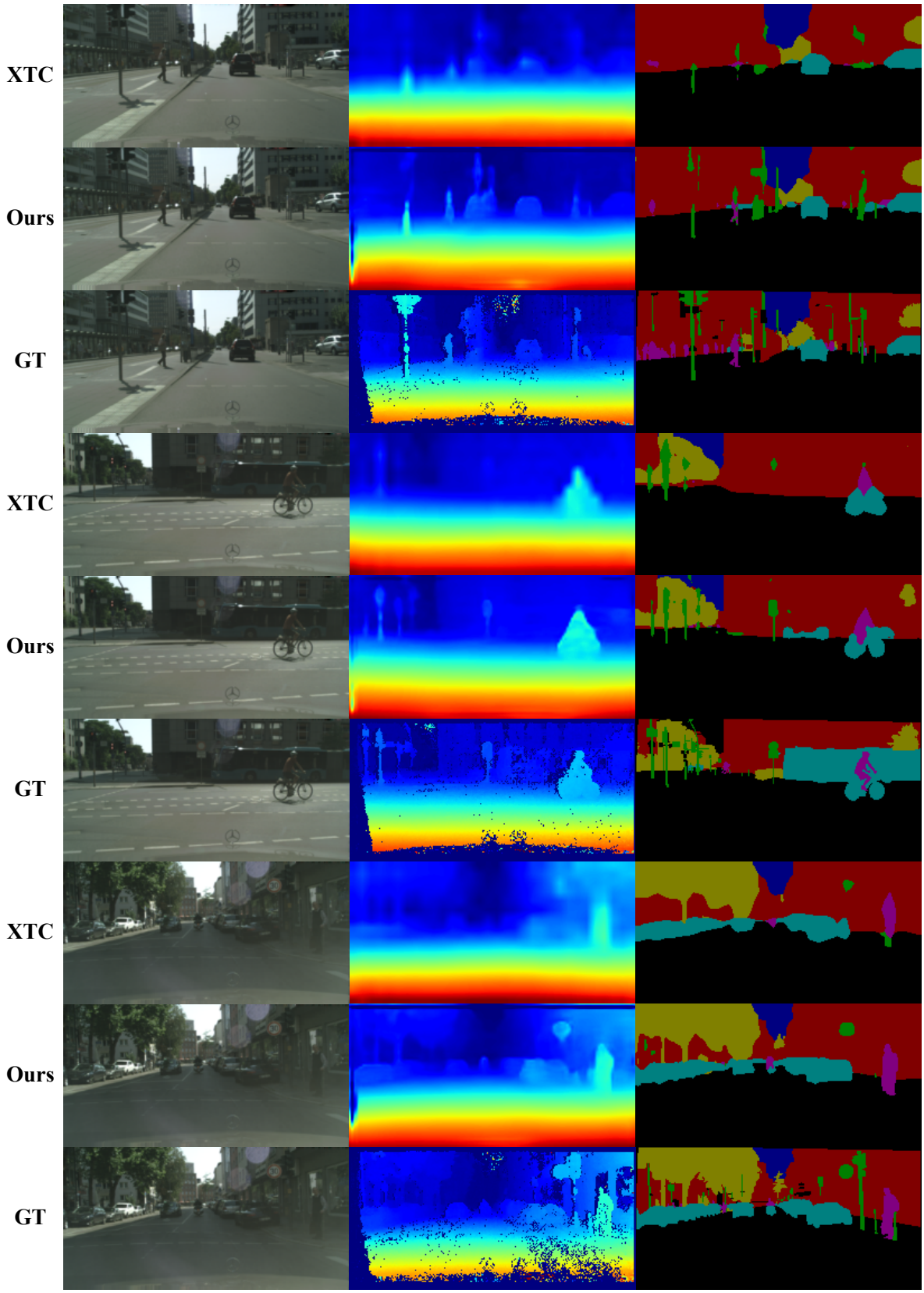

Figure 4. Qualitative comparison between our method and the state-of-the-art method (*i.e.* XTC [3]) for depth estimation and semantic segmentation tasks on the Cityscapes dataset, using the same ResNet-18 backbone. Our DiffusionMTL approach outperforms the previous state-of-the-art method in producing superior prediction maps. Notably, each training sample is labeled for only one task.

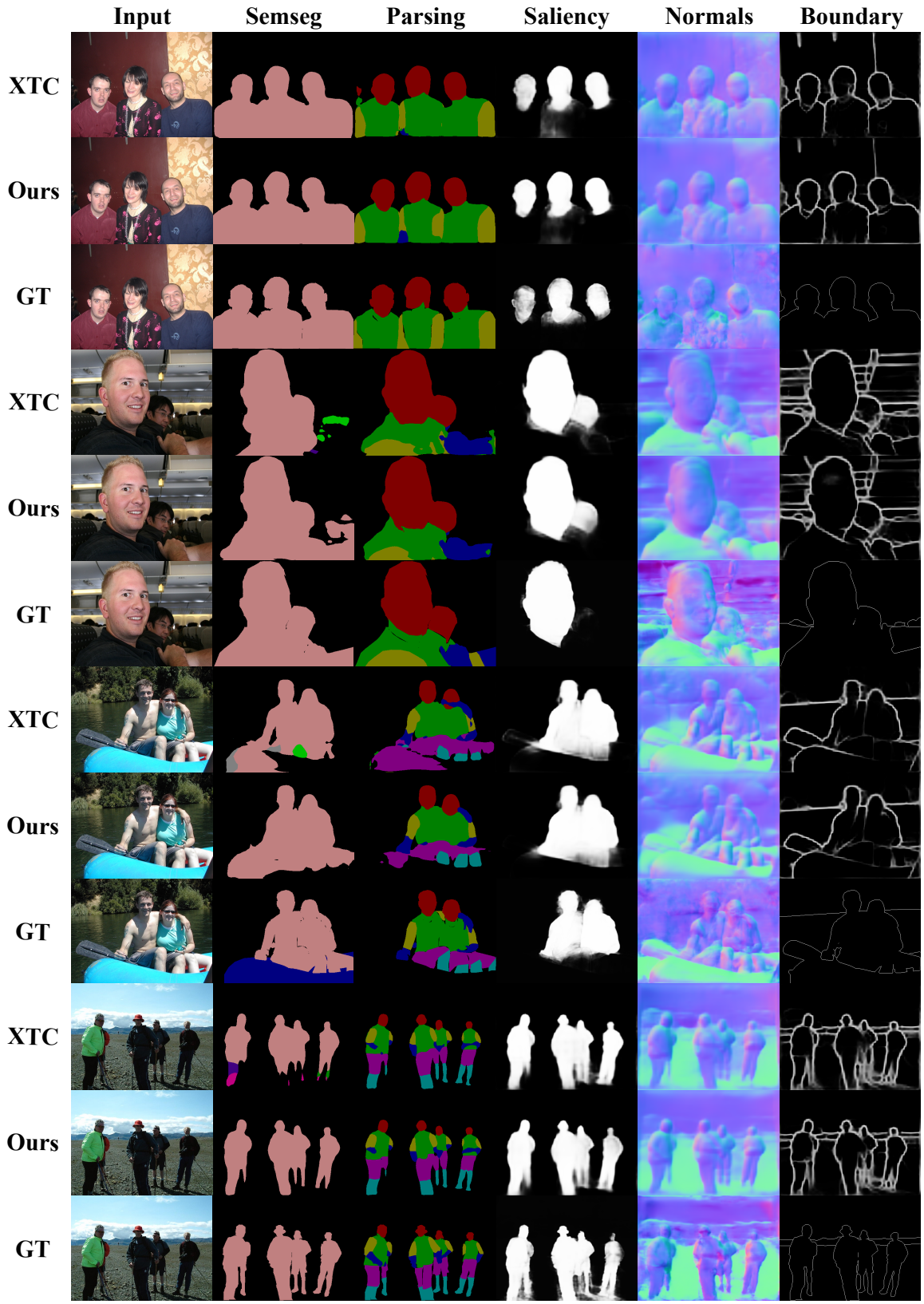

Figure 5. Qualitative comparison between our method and the state-of-the-art method (*i.e.* XTC [3]) in PASCAL dataset, using the same ResNet-18 backbone. Our DiffusionMTL approach outperforms the previous state-of-the-art method in producing superior prediction maps. Notably, each training sample is labeled for only one task.

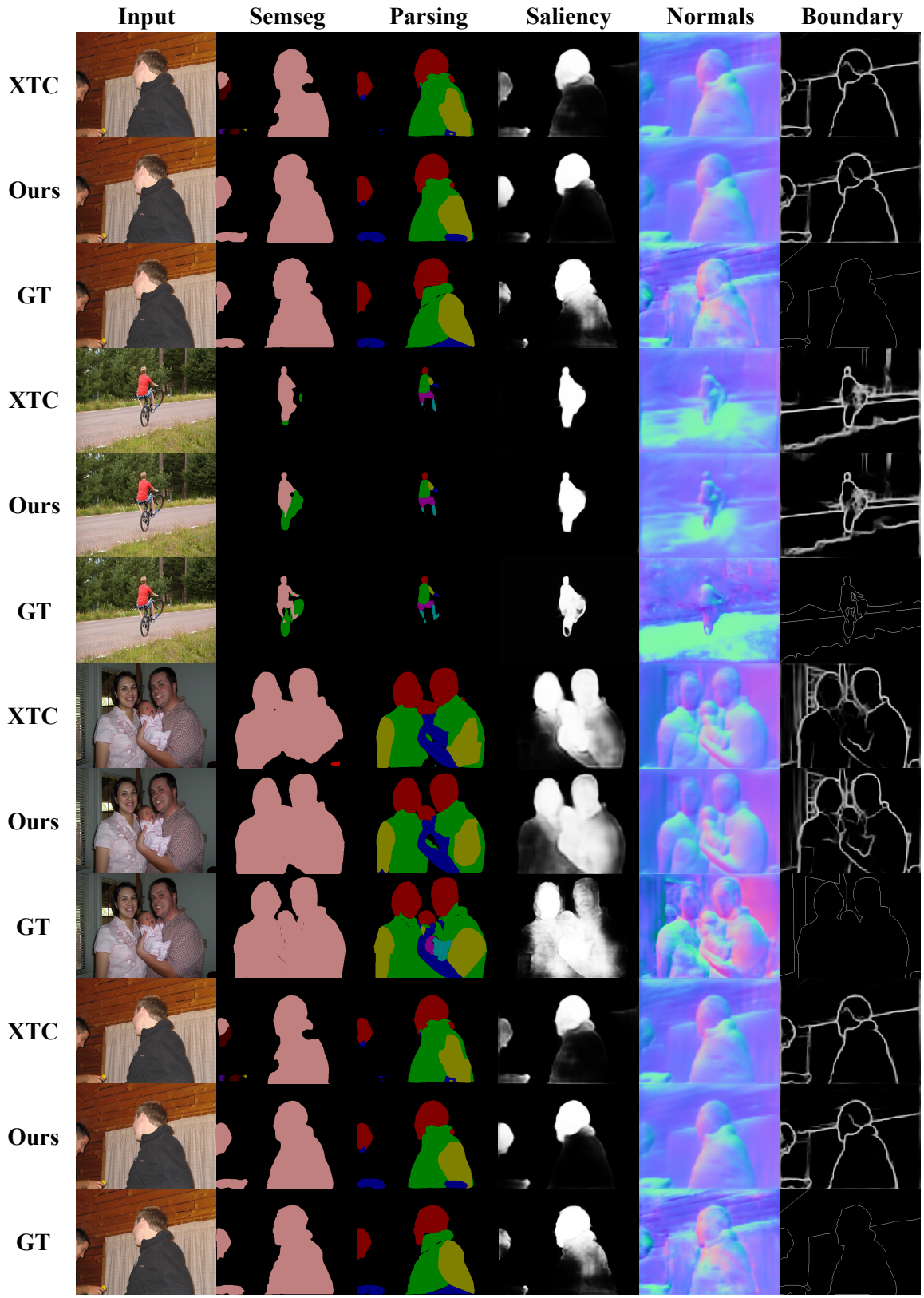

Figure 6. Qualitative comparison between our method and the state-of-the-art method (*i.e.* XTC [3]) in PASCAL dataset, using the same ResNet-18 backbone. Our DiffusionMTL approach outperforms the previous state-of-the-art method in producing superior prediction maps. Notably, each training sample is labeled for only one task.