Learning to Relate Depth and Semantics for Unsupervised Domain Adaptation Supplementary Materials

Suman Saha* ETH Zurich
Anton Obukhov* ETH Zurich
Danda Pani Paudel ETH Zurich
Menelaos Kanakis ETH Zurich
Yuhua Chen ETH Zurich
Stamatios Georgoulis ETH Zurich
Luc Van Gool ETH Zurich, KU Leuven

In this document, we provide supplementary materials for our main paper submission. The main paper reported our experimental results using three standard UDA evaluation protocols (EPs) where the SYNTHIA dataset [7] is used as the synthetic domain. To demonstrate our proposed method's effectiveness on an entirely new UDA setting, in Sec. S1, we report semantic segmentation results of our method on a new EP: Virtual KITTI → KITTI. In this setup, we use synthetic Virtual KITTI [3] as the source domain and real KITTI [4] as the target domain. We show that our proposed method consistently outperforms the SOTA DADA method [9] when evaluated on this new EP with different synthetic and real domains. In Sec. S2, we present a t-SNE [8] plot comparing our method with [9]. We also share additional qualitative results on SYNTHIA → Cityscapes (16 classes). Sec. S3 details our network design. To demonstrate that the proposed CTRL is not sensitive to a particular network design (in our case, the residual auxiliary block [5]), we train a standard multi-task learning network architecture (i.e., a shared encoder followed by multiple task-specific decoders without any residual auxiliary block) with CTRL and notice a similar improvement trend over the baselines. The set of experiments and the results are discussed in Sec. S4.

S1. Virtual KITTI → KITTI

Following [2], we train and evaluate our model on 10 common classes of Virtual KITTI and KITTI. In KITTI, the ground-truth label is only available for the training set; thus, we use the official unlabelled test images for domain alignment. We report the results on the official training set following [2]. The model is trained on the annotated training samples of VKITTI and unannotated samples of KITTI. For this experiment, we train our model without (w/o) ISL. Table S1 reports the semantic segmentation performance (mIoU%) of our approach. Our model outperforms DADA [9], with significant gains coming from the following classes: "sign" (+8.1%), "pole" (+5.7%), "building" (+2.7%), and "light" (+1.9%). Notably, these classes are practically highly relevant to an autonomous driving scenario. In Figure S2, we present some qualitative results of DADA and our models trained following the new Virtual KITTI → KITTI UDA protocol.

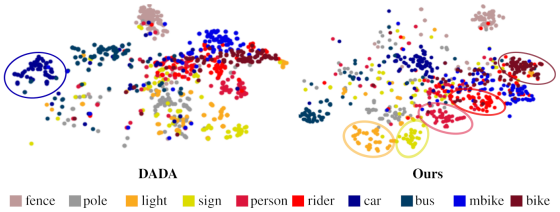

Figure S1: t-SNE comparison of features learned by DADA [9] and CTRL. It leads to more structured feature space and better class separation in the target domain. Circled classes have a better separation than the other method.

S2. SYNTHIA → Cityscapes

This section presents a t-SNE [8] plot of the feature embeddings learned by the proposed model guided by CTRL, and [9]. Fig. S1 shows 10 top-scoring classes of each method; distinct classes are circled. As can be seen from the figure, CTRL leads to more structured feature space, which concurs with our analysis of the main paper. Both models are trained and evaluated following the UDA protocol SYNTHIA → Cityscapes (16 classes). Furthermore, we present additional qualitative results of our model for semantic segmentation and monocular depth estimation. Figures S3, S4 show the results of the qualitative comparison of our method with [9]. Note that our proposed method has higher spatial acuity in delineating small objects like "human", "bicycle", and "person" compared to [9]. Figure S5 shows some qualitative monocular depth estimation results.

Corresponding author: Suman Saha (suman.saha@vision.ee.ethz.ch)
* Equal contribution.

Table S1: Semantic segmentation performance (IoU and mIoU, higher is better, %) comparison to the prior art. All models are trained and evaluated using the UDA evaluation protocol Virtual KITTI → KITTI.

		VKITTI → KITTI (10 classes)										
Models	Depth	road	building	pole	light	sign	veg	terrain	sky	car	truck	mIoU ↑
Chen *et al*. [2]	✓	81.4	71.2	11.3	26.6	23.6	**82.8**	**56.5**	**88.4**	80.1	12.7	53.5
DADA [9]	✓	**90.9**	76.2	12.4	30.3	30.8	73.5	24.1	**88.4**	**86.8**	**17.2**	53.0
Ours (w/o ISL)	✓	**90.9**	**78.9**	**18.1**	**32.2**	**38.9**	73.7	22.0	88.2	86.2	16.7	**54.6**

Table S2: Semantic segmentation performance (mIoU) of two variants of the proposed model. Both models outperform DADA [9] attesting the robustness of features learned by the proposed CTRL.

UDA Protocol	DADA	Ours∗	Ours
S → C 16 cls	42.6	43.7±0.2	**45.0±0.3**
S → C (LR) 7 cls	63.4	63.8±0.5	**64.7±0.5**
S → M (LR) 7 cls	55.8	61.5±0.6	**62.1±0.4**
S → C (FR) 7 cls	69.2	**71.3±0.5**	70.8±0.4
S → M (FR) 7 cls	67.6	**70.1±0.5**	69.0±0.1

S3. Network Architecture Design

The shared part of the semantic and depth prediction network $\mathcal{F}_e$ consists of a ResNet-101 backbone and a decoder. The decoder consists of four convolutional layers, each followed by a Rectified Linear Unit (ReLU). The decoder outputs a feature map that is shared among both semantics and depth heads. This shared feature map is fed forward to the respective semantic segmentation, monocular depth estimation, and semantics refinement heads. For the task-specific and task-refinement heads, we use Atrous Spatial Pyramid Pooling (ASPP) with sampling rates $[6, 12, 18, 24]$ and the Deeplab-V2 [1] architecture. Our DC-GAN [6] based domain discriminator takes as input a feature map with channel dimension $2\times C+K$, where C is the number of semantic classes, and K is the number of depth levels.

S4. Robustness to Different Network Design

Our proposed model adopts the residual auxiliary block [5] (as in [9]), which was originally proposed to tackle a particular MTL setup where the objective was to improve one primary task by leveraging several other auxiliary tasks. However, unlike [9] which doesn't have any decoder for depth, we introduce a DeepLabV2 decoder for depth estimation to improve both task performances. Our qualitative and quantitative experimental results show an improvement of depth estimation performance over [9]. Furthermore, we are interested to see the proposed model's performance when used with a standard MTL architecture (a common encoder followed by multiple task-specific decoders without any residual auxiliary blocks). To this end, we make necessary changes to our existing network design to have a standard MTL network design. We then train it following UDA protocols. The details of our experimental analysis are given below.

For the standard MTL model (denoted as "Ours*" in Table S2), the depth head is placed after the shared feature extractor $\mathcal{F}_e$. The shared feature extractor consists of a ResNet backbone and decoder network (see Fig. 2). For the second model with residual auxiliary block (denoted as "Ours"), we positioned the depth head after the decoder's third convolutional layer. The semantic segmentation performance of these two variants of the proposed model is shown in Table S2. Both models are evaluated on the five different UDA protocols and outperform state-of-the-art DADA [9] results. The results show that our proposed CTRL is not sensitive to architectural changes and can be used with standard encoder-decoder MTL frameworks. Our findings may be found beneficial for the domain-adaptive MTL community, e.g., in answering a question whether *learning additional complementary tasks (surface normals, instance segmentation) performs domain alignment.*

References

[1] *Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.* IEEE Trans. Pattern Anal. Mach. Intell., *40(4):834–848, 2017. S2*

[2] *Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In* IEEE Conf. Comput. Vis. Pattern Recog., *pages 1841–1850, 2019. S1, S2*

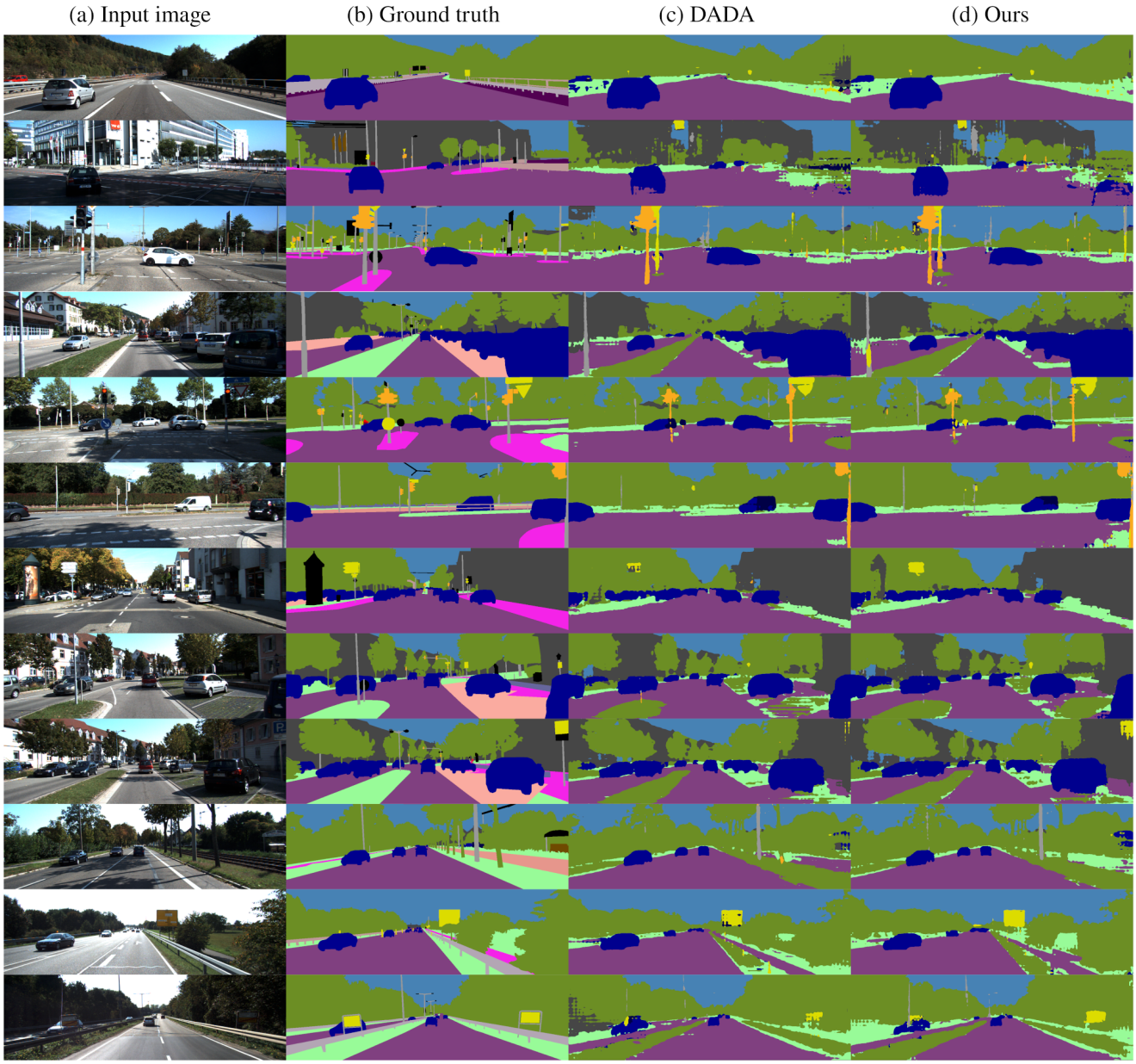

Figure S2: Qualitative semantic segmentation results with VKITTI → KITTI (10 classes) UDA evaluation protocol. (a) Input image from the target domain KITTI; (b) ground truth annotations; (c) DADA [9] predictions; (d) our model predictions. We follow the color encoding scheme of Cityscapes to colorize the label maps.

[3] *Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In* Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. S1*

[4] *Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset.* The International Journal of Robotics Research*, 32(11):1231–1237, 2013. S1*

[5] *Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In* Adv. Neural Inform. Process. Syst.*, pages 1310–1322, 2018. S1, S2*

[6] *Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks.* arXiv preprint arXiv:1511.06434*, 2015. S2*

[7] *German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large*

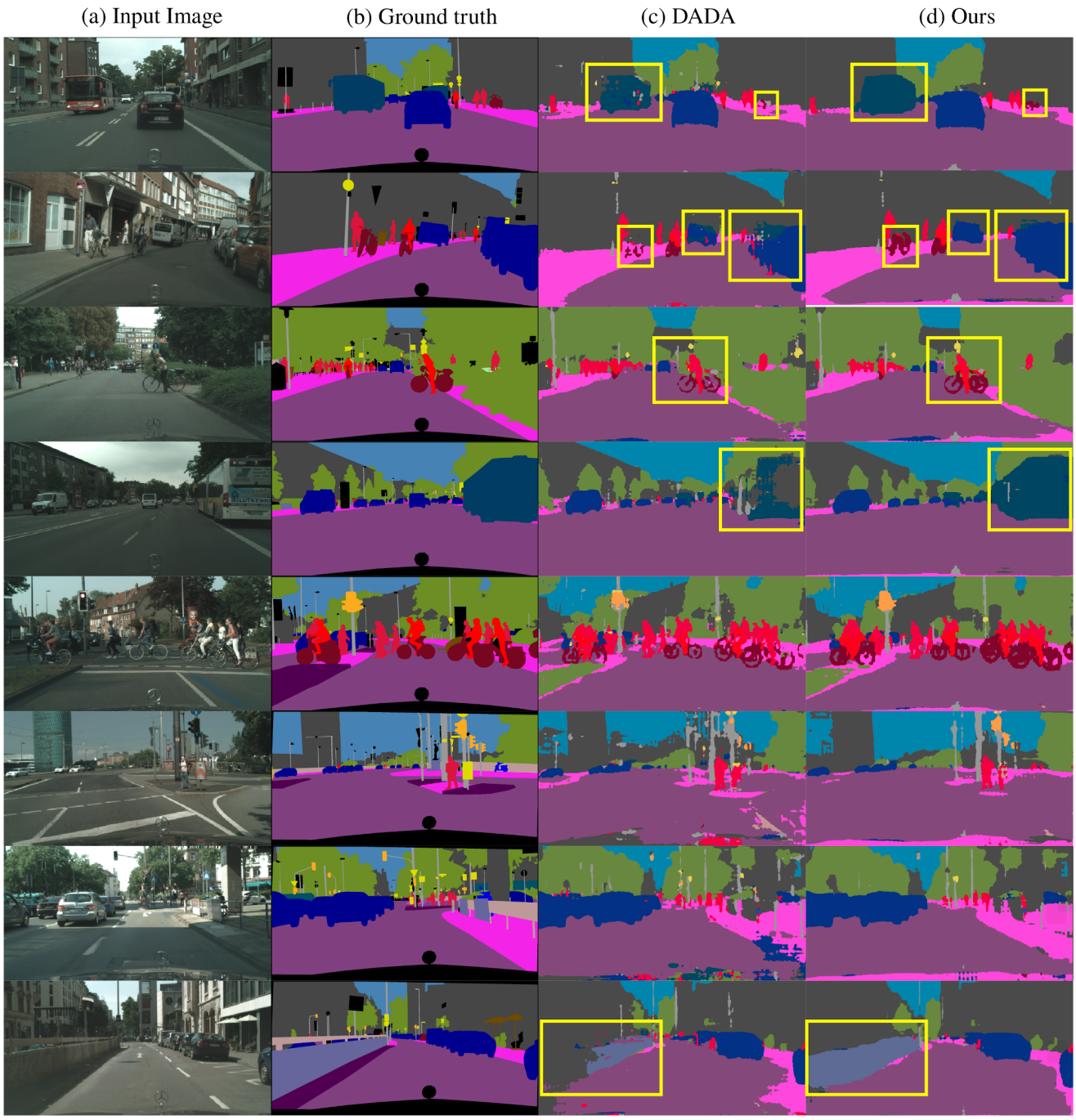

Figure S3: Qualitative semantic segmentation results with EP1: SYNTHIA → Cityscapes (16 classes). (a) Images from Cityscapes validation set; (b) ground truth annotations; (c) DADA [9] predictions; (d) our model predictions. Our method demonstrates notable improvements over [9] on "bus", "person", and "bicycle" classes as highlighted using the yellow boxes.

collection of synthetic images for semantic segmentation of urban scenes. *In* IEEE Conf. Comput. Vis. Pattern Recog., *pages 3234–3243, 2016.* S1

[8] *Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne.* Journal of Machine Learning Research, *9(86):2579–2605, 2008.* S1

[9] *Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation.* *In* Int. Conf. Comput. Vis., *pages 7364–7373, 2019.* S1, S2, S3, S4, S5, S6

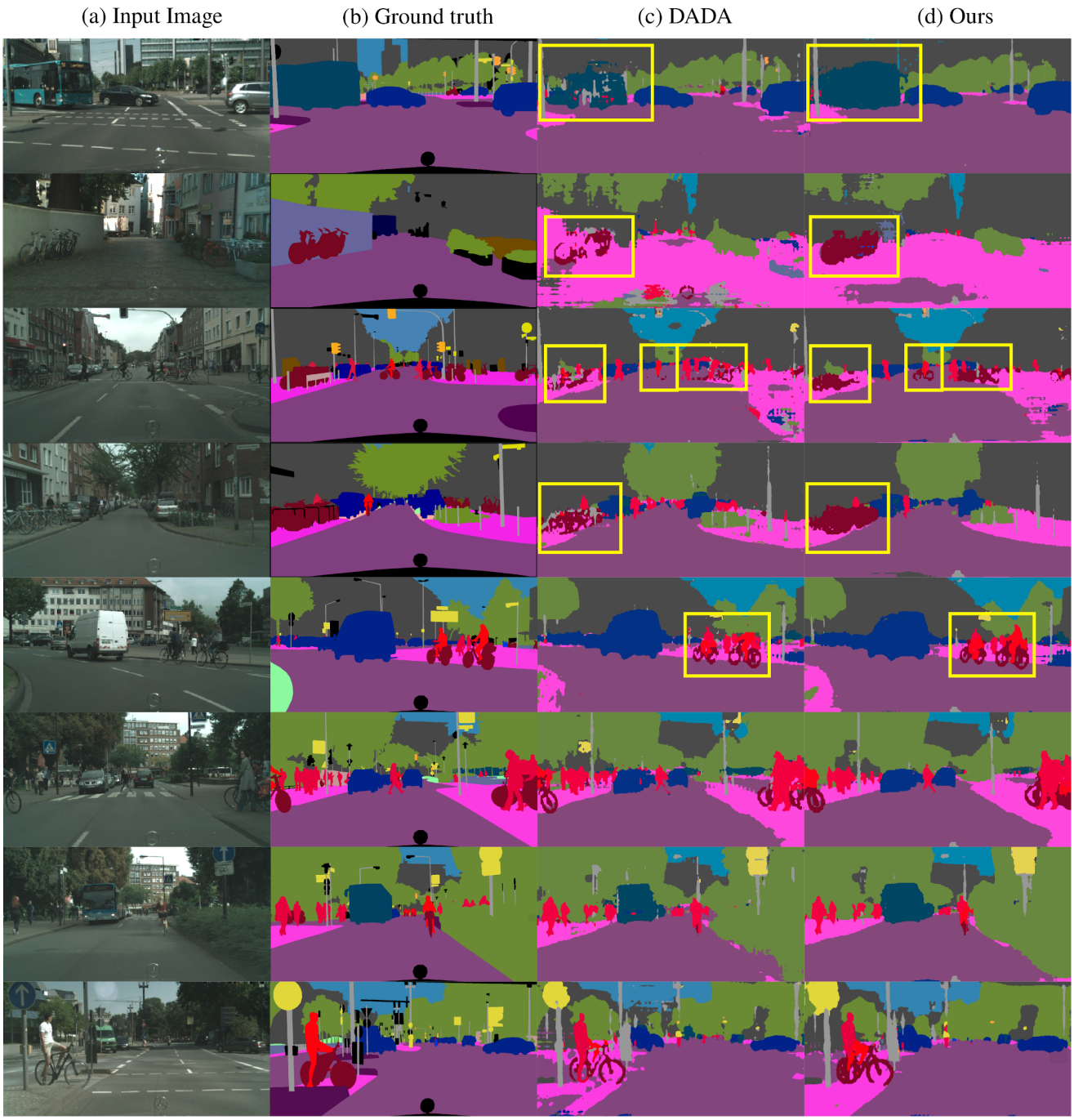

Figure S4: Qualitative semantic segmentation results with EP1: SYNTHIA → Cityscapes (16 classes). (a) Images from Cityscapes validation set; (b) ground truth annotations; (c) DADA [9] predictions; (d) our model predictions. Our method demonstrates notable improvements over [9] on "bus", "person", and "bicycle" classes as highlighted using the yellow boxes.

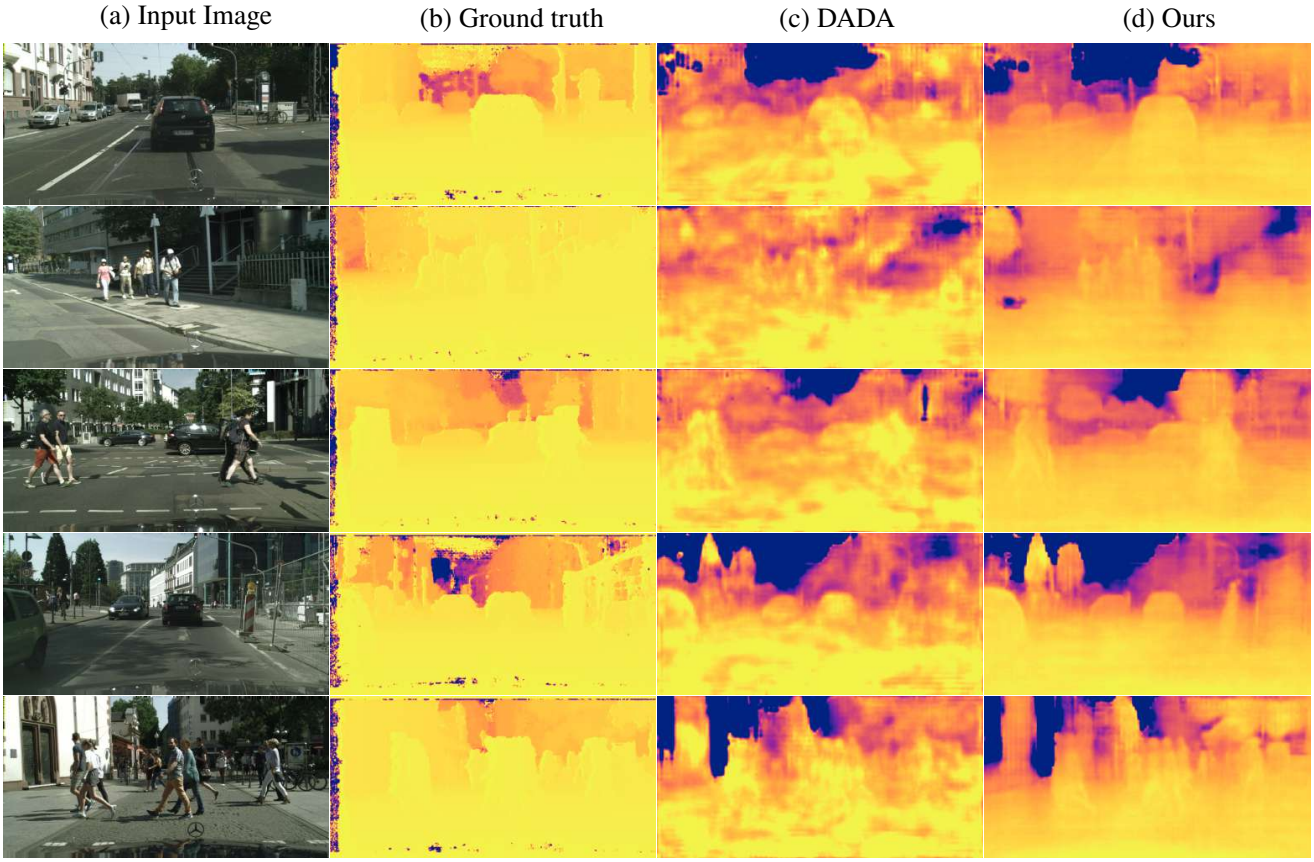

Figure S5: Qualitative monocular depth estimation results with EP1: SYNTHIA $\rightarrow$ Cityscapes (16 classes). (a) Images from Cityscapes validation set; (b) ground truth annotations; (c) DADA [9] predictions; (d) our model predictions.