Supplementary Material: Intra-Batch Supervision for Panoptic Segmentation on High-Resolution Images

Daan de Geus Gijs Dubbelman
Eindhoven University of Technology
d.c.d.geus@tue.nl, g.dubbelman@tue.nl

In this document, we provide the following supplementary material in addition to the main manuscript:

- More extensive implementation details (Section 1).
- Additional qualitative results, showing the *confusion* problem and the effect of our IBS to solve it (Section 2).

All code is available through `https://ddegeus.github.io/intra-batch-supervision/`.

1. Implementation details

We provide more extensive implementation details for the neural networks we use for the experiments in our main manuscript. For both Panoptic FCN [5] and Mask2Former [1], we use and extend the official code repositories, which are both based on PyTorch [9] and Detectron2 [10]. All networks use a ResNet-50 backbone [4], which is initialized with weights pre-trained on ImageNet [3]. In general, we use the original implementation details of both Panoptic FCN and Mask2Former. Whenever we use different settings, we indicate this explicitly.

For both networks, we add the focal loss [6] for IBS to the original losses of the network, and calculate the total loss as

$$L_{total} = L_{orig} + \lambda_{IBS} L_{IBS}, \tag{1}$$

where L_{orig} are the original losses, L_{IBS} is the focal loss for IBS, λ_{IBS} is the loss weight for the IBS loss, and L_{total} is the resulting total loss.

1.1. Panoptic FCN

Panoptic FCN is optimized with stochastic gradient descent, and uses a polynomial learning rate schedule with initial learning rate lr_0 and decay 0.9. The weight decay is 10^{-4}. We use $\lambda_{IBS} = 1$, and we empirically find that the network is robust to relatively small variations to this value ([0.5; 5.0]). The embedding dimension is set to $C = 256$, instead of 64 as in the original work. We find that this improves the performance both with and without IBS. In all experiments, we apply random horizontal flipping to the images and ground-truth before feeding them to the network.

Cityscapes. For crop-based training on Cityscapes [2], $lr_0 = 0.02$. Following the original settings, we train for 65k steps with batches of 32 crops of 512×1024 pixels, after they are randomly resized with a factor between 0.5 and 2.0. For full-image training, we use $lr_0 = 0.005$. We train for 100k steps with batches of 4 images, after randomly resizing them with a factor between 0.5 and 2.0.

Mapillary Vistas. For crop-based training on Mapillary Vistas [8], $lr_0 = 0.02$. We train for 150k steps with batches of 32 crops of 1024×1024 pixels, after the images are randomly resized such that the shortest side is between 1024 and 2048 pixels. For full-image training, $lr_0 = 0.005$. We train for 150k steps with batches of 8 images, after the images are randomly resized such that the shortest side is between 1024 and 2048 pixels.

1.2. Mask2Former

Mask2Former is optimized using AdamW [7], and uses an initial learning rate $lr_0 = 0.0001$. The weight decay is 0.05. IBS is applied to the predictions made at each transformer decoder layer, except the first one. To apply IBS, we identify and extract the per-segment *embeddings* belonging to thing segments based on the ground-truth segment they are matched to with the bipartite matching algorithm that is also used for the other losses. We use $\lambda_{IBS} = 100$ to balance the losses in such a way that ratio between the magnitudes of L_{IBS} and L_{orig} is similar to the one for Panoptic FCN. For Mask2Former, we only do experiments with crop-based training. In all experiments, we apply random horizontal flipping to the images and ground-truth before feeding them to the network.

Cityscapes. Following the original settings, we train for 90k steps using batches of 16 crops of 512×1024 pixels, which are taken from the original images after first resizing them with a random factor between 0.5 and 2.0.

Mapillary Vistas. We train for 300k steps using batches of 16 crops of 1024×1024 pixels, which are taken from the original images after first randomly resizing them such that the shortest side is between 1024 and 2048 pixels. The training settings for instance segmentation are equal to those for panoptic segmentation, except that no stuff predictions are made for instance segmentation.

2. Qualitative results

We provide additional qualitative results to illustrate both the *confusion* problem with crop-based training, and the effectiveness of IBS to solve this problem.

To demonstrate the specific confusion problem, we show several examples of individual thing predictions by Panoptic FCN and Mask2Former in Figures 1, 3 and 5. In the predictions made by the networks *without IBS*, it is clearly visible that the masks overlap multiple ground-truth thing instances – which is what we call confusion – and that this confusion mostly occurs between instances of the same class. In these figures, we also demonstrate that IBS solves this confusion problem to a great extent, resulting in much more accurate thing predictions. In Figures 2, 4 and 6, we also provide the full panoptic and instance segmentation predictions for the same images and networks. This way, the impact of confusion on the overall result is visualized. Note that each instance should receive a unique color and text label in the visualized panoptic prediction, so a prediction in which two or more instances share a color and text label is a case of confusion. From these figures, it is also clear the networks *with IBS* make considerably more accurate panoptic predictions, especially for large thing segments. Although there are still some small imperfections in the predicted masks, these predictions are considerably more suitable for downstream processing, as there are fewer grouped or missed objects.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 2022.

[2] 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.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[5] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully Convolutional Networks for Panoptic Segmentation. In *CVPR*, 2021.

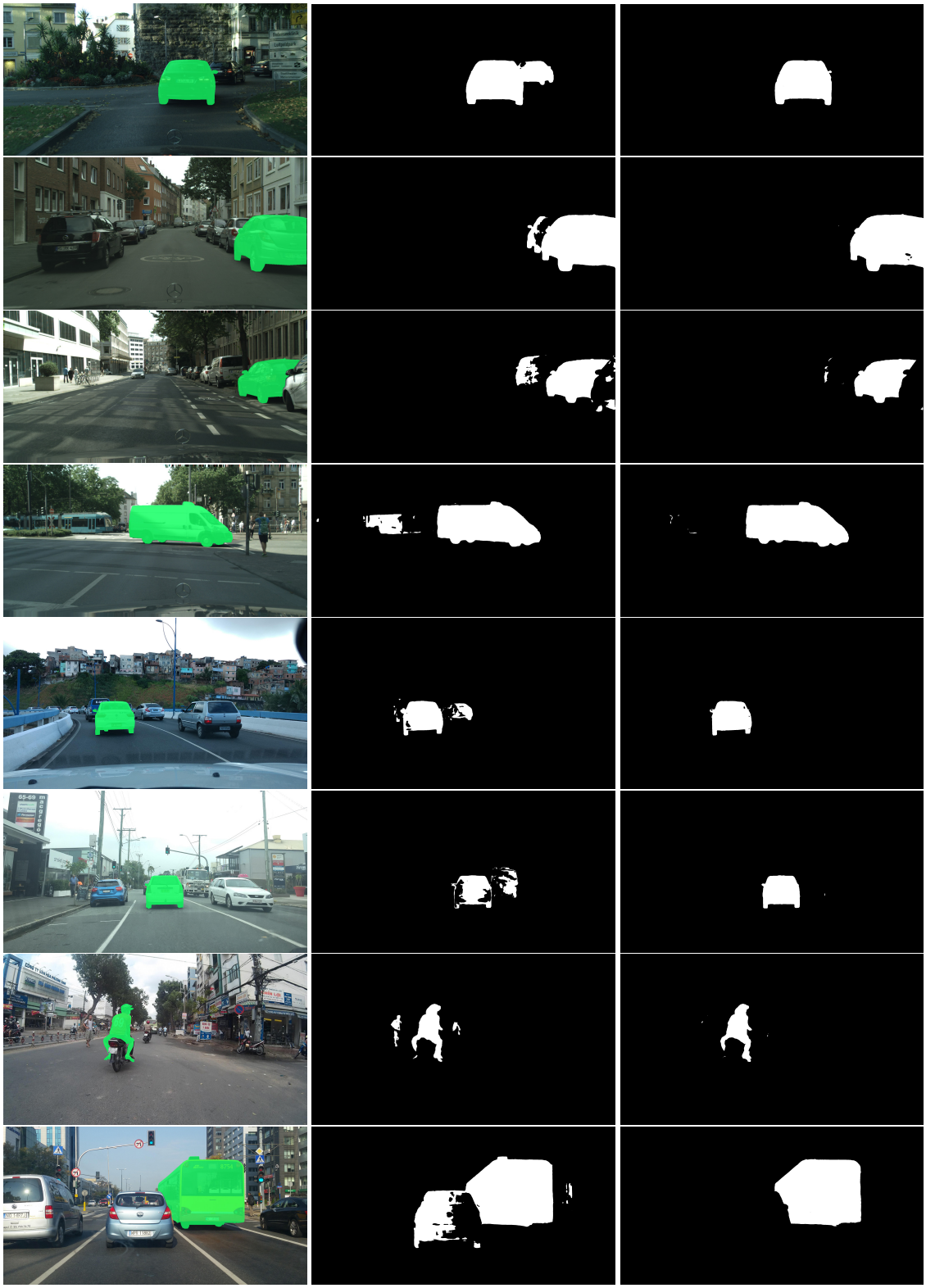
[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *ICCV*, 2017.

[7] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

[8] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *ICCV*, 2017.

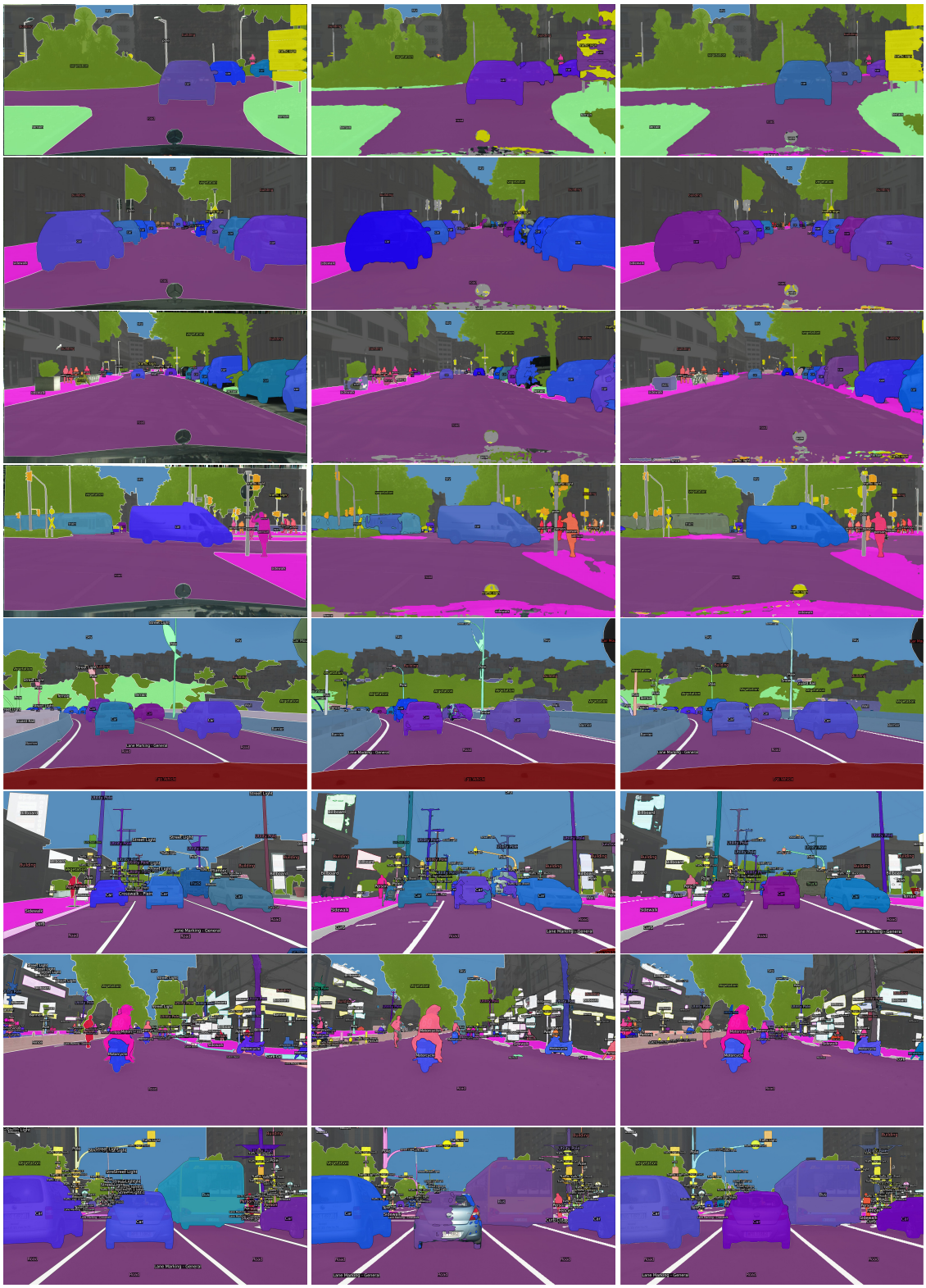
[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.



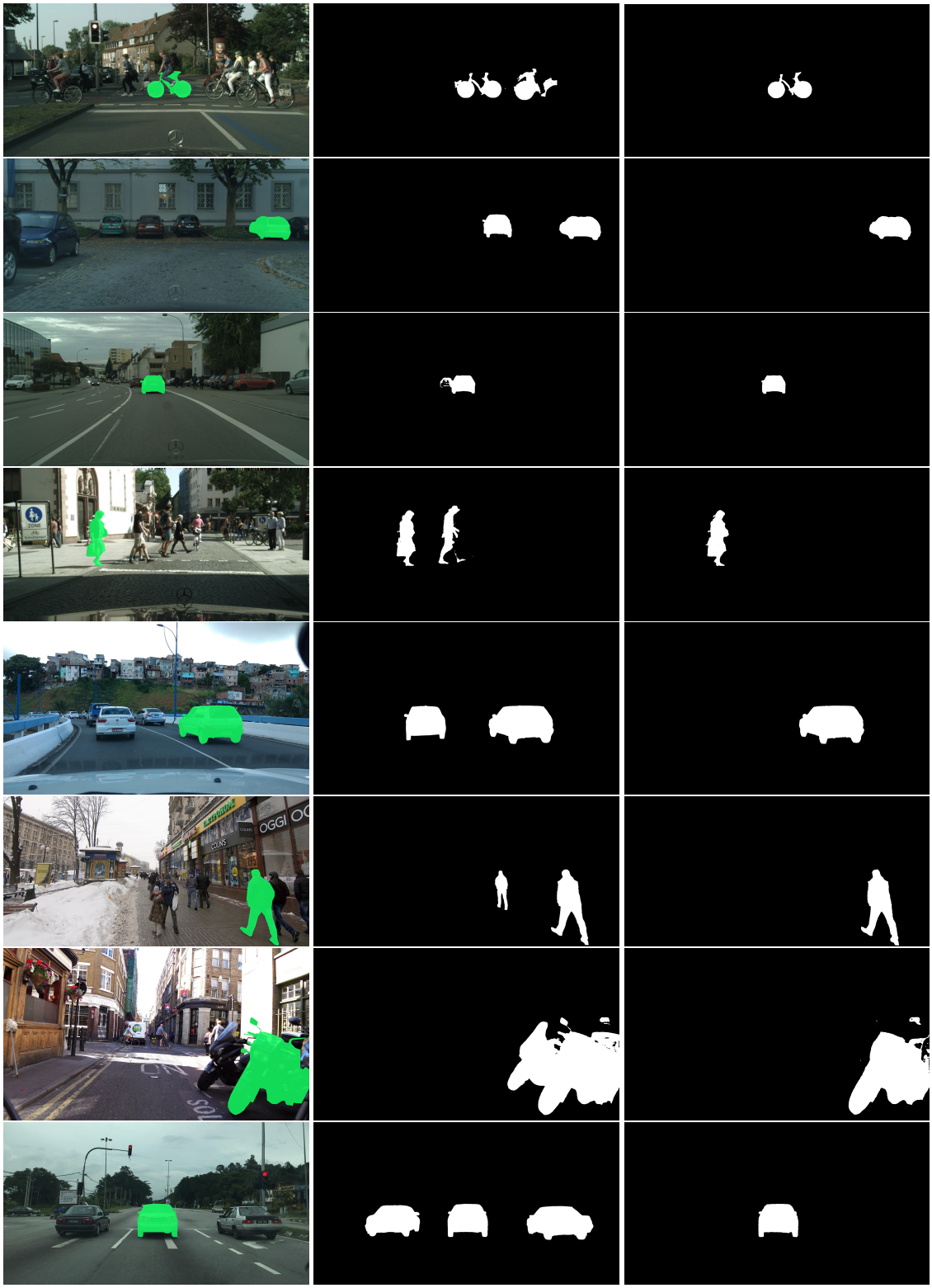
(a) Image with ground-truth segment (b) Predicted segment **without IBS** (c) Predicted segment **with IBS** (ours)

Figure 1: **Confusion problem for crop-based training of Panoptic FCN.** Predictions for individual thing instances with and without IBS. Top four images: Cityscapes *val*; bottom four: Mapillary Vistas *validation*. (b) The predictions by Panoptic FCN without IBS suffer from confusion, and (c) IBS largely solves this problem, leading to more accurate predictions. Full panoptic results for these images are shown in Figure 2.



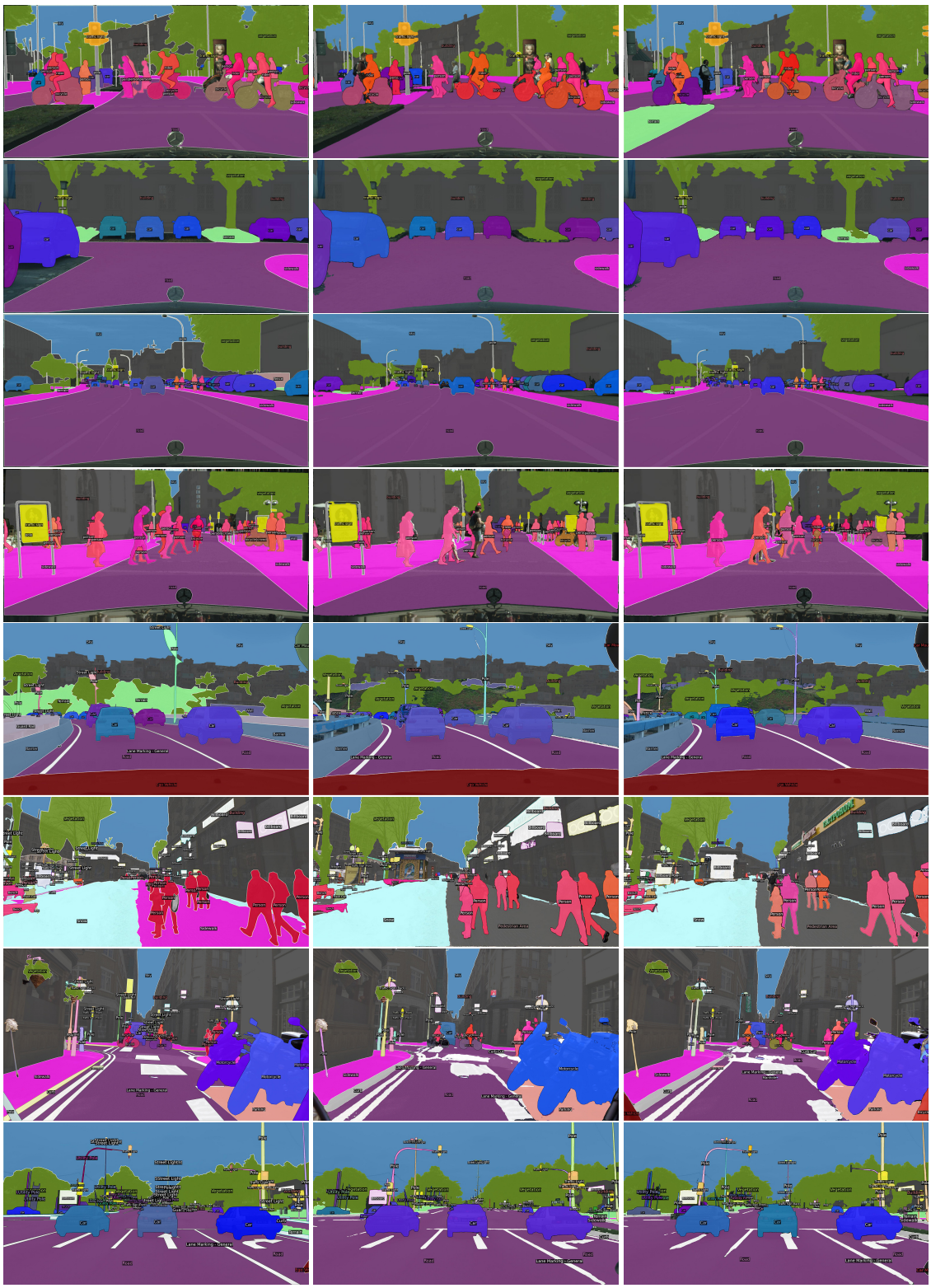
(a) Ground truth (b) Without IBS (c) With IBS (ours)

Figure 2: **Intra-Batch Supervision on Panoptic FCN.** Top four images: Cityscapes *val*; bottom four: Mapillary Vistas *validation*. Each segment is indicated with a unique color and text label, so confusion can be observed when multiple thing instances share a color or text label. Individual thing predictions for these images are shown in Figure 1. Best viewed digitally.



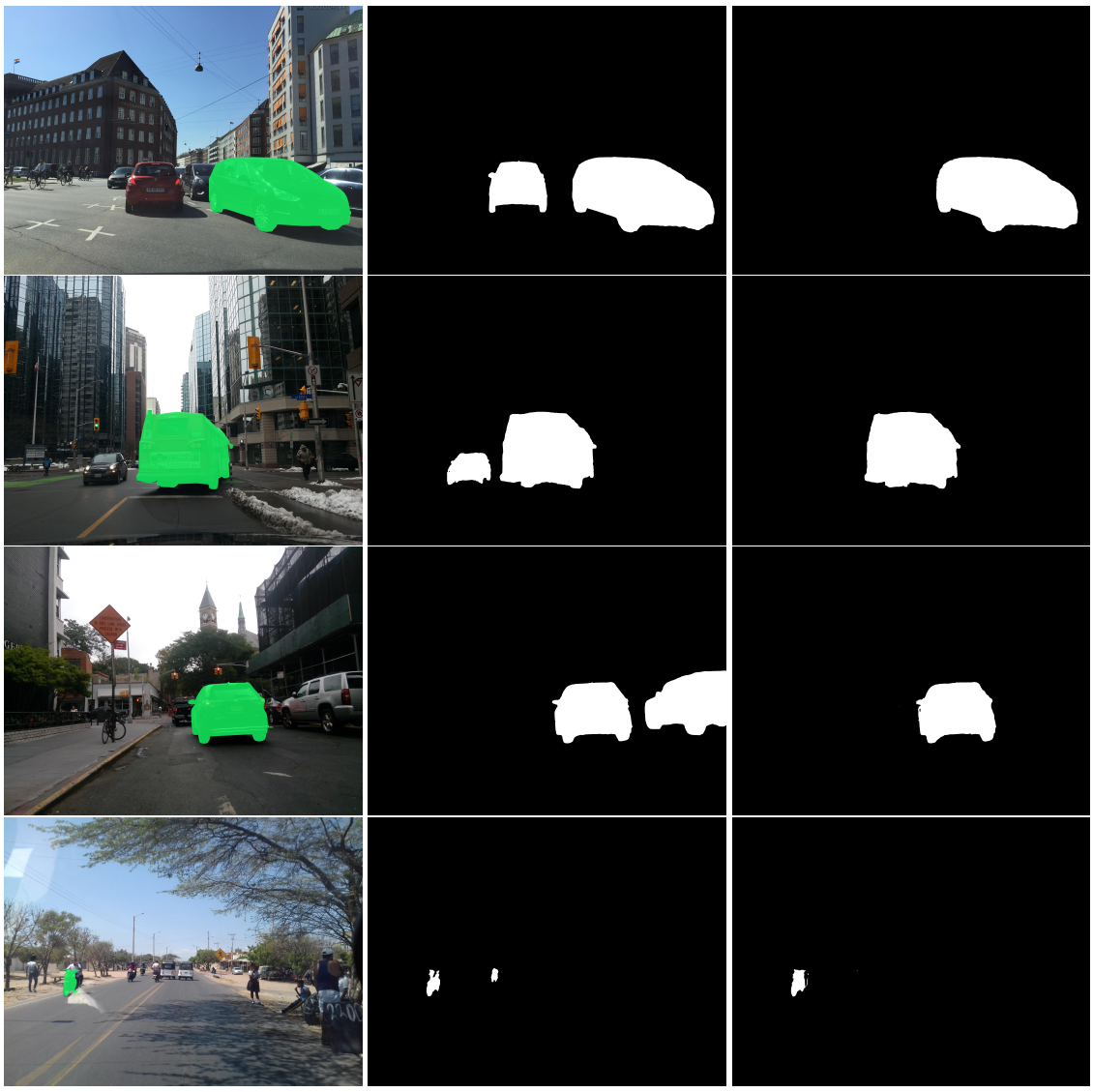
(a) Image with ground-truth segment (b) Predicted segment **without IBS** (c) Predicted segment **with IBS** (ours)

Figure 3: **Confusion problem for crop-based training of Mask2Former.** Predictions for individual thing instances with and without IBS. Top four images: Cityscapes *val*; bottom four: Mapillary Vistas *validation*. (b) The predictions by Mask2Former without IBS suffer from confusion, and (c) IBS largely solves this problem, leading to more accurate predictions. Full panoptic results for these images are shown in Figure 4.



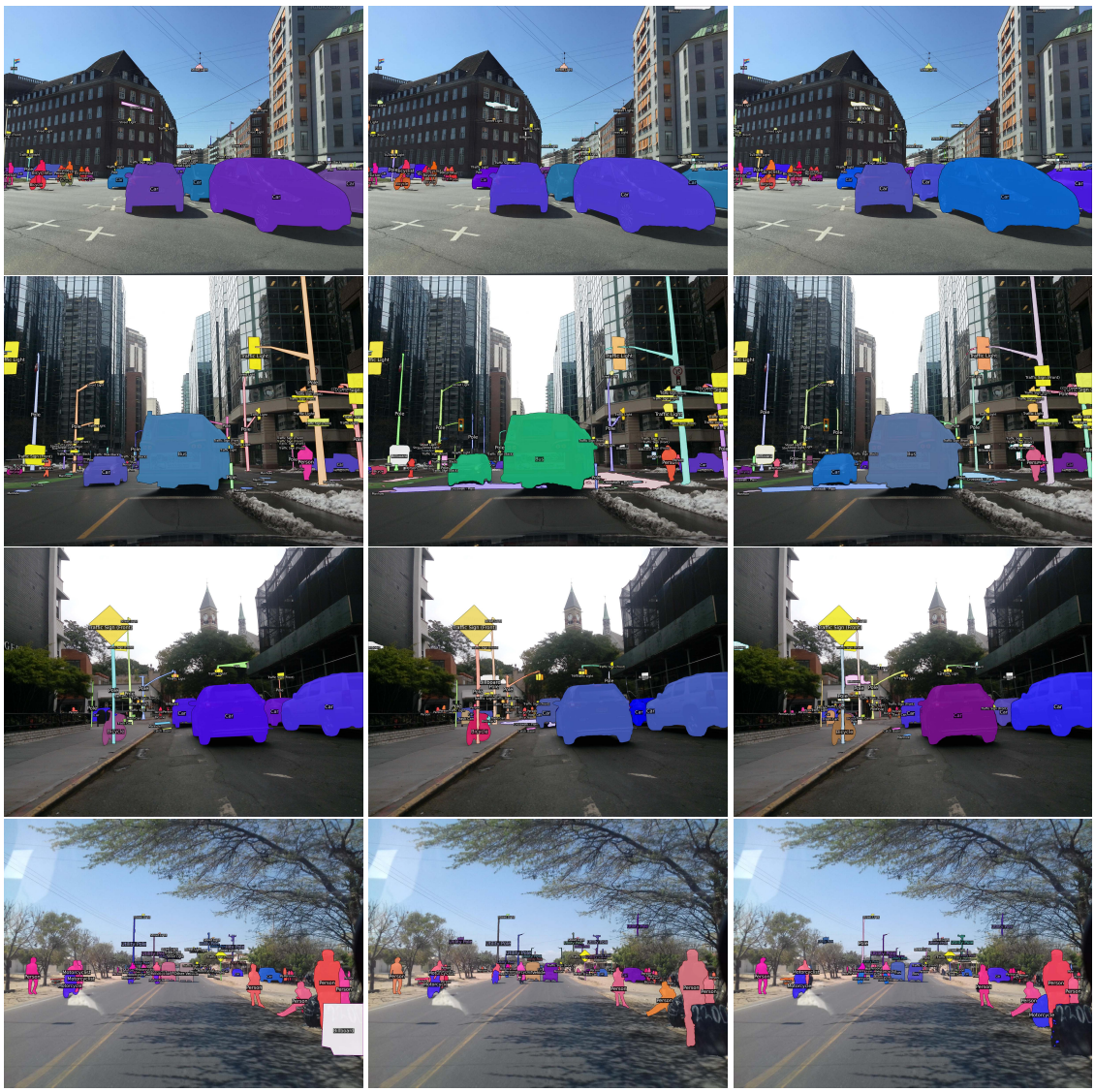
(a) Ground truth (b) Without IBS (c) With IBS (ours)

Figure 4: **Intra-Batch Supervision on Mask2Former.** Top four images: Cityscapes *val*; bottom four: Mapillary Vistas *validation*. Each segment is indicated with a unique color and text label, so confusion can be observed when when multiple thing instances share a color or text label. Individual thing predictions for these images are shown in Figure 3. Best viewed digitally.



(a) Image with ground-truth segment (b) Predicted segment **without IBS** (c) Predicted segment **with IBS** (ours)

Figure 5: **Confusion problem for crop-based training of Mask2Former for instance segmentation.** Predictions for individual thing instances with and without IBS, on the Mapillary Vistas *validation* set. (b) The predictions by Mask2Former without IBS suffer from confusion, and (c) IBS largely solves this problem, leading to more accurate predictions. Full instance segmentation results for these images are shown in Figure 6.



(a) Ground truth (b) Without IBS (c) With IBS (ours)

Figure 6: **Intra-Batch Supervision on Mask2Former for instance segmentation.** Images from the Mapillary Vistas *validation* set. Each instance is indicated with a unique color and text label, so confusion can be observed when when multiple instances share a color or text label. Individual thing predictions for these images are shown in Figure 5. Best viewed digitally.