MGNiceNet: Unified Monocular Geometric Scene Understanding
-
Supplementary Material

Markus Schön, Michael Buchholz, and Klaus Dietmayer

Institute of Measurement, Control, and Microtechnology, Ulm University, Germany
{markus.schoen,michael.buchholz,klaus.dietmayer}@uni-ulm.de

A Comparison to SOTA Panoptic Segmentation Methods

We compare our method to non-real-time state-of-the-art panoptic segmentation methods on the Cityscapes [2] validation dataset in Tab. 1. MGNiceNet focuses on low latency for real-time applications such as autonomous driving perception systems. Thus, heavier architectures that do not consider this constraint naturally outperform our approach in terms of panoptic quality. However, Tab. 1 shows that our approach can significantly close the gap to heavier approaches compared to the current real-time state-of-the-art method RT-K-Net [15]. Compared to Axial-DeepLab-XL [16], which is the best-performing approach that neither uses Mapillary Vistas [11] nor ImageNet-22K [3] pre-training, our approach can close the gap from 4.2% PQ to 2.7% PQ, a significant improvement of approximately 35%. Compared to the current state-of-the-art approach OneFormer [8] with ConvNeXt-L [10] backbone, which uses both Mapillary Vistas and ImageNet-22K pre-training, our approach is outperformed by 6.1% PQ. We argue that this large gap is due to ImageNet-22K pre-training and the much heavier architecture used in [8]. Since non-real-time state-of-the-art methods, such as OneFormer, do not report runtimes, we compare the number of floating point operations for input images of 1024×2048 pixels. Using this metric, our method outperforms all other methods by a large margin, underlining that current state-of-the-art methods primarily focus on accuracy rather than inference speed. For example, OneFormer requires 497G floating point operations compared to 155G required for our method.

B Ablation Study on Kernel Linking

We perform an additional ablation study to investigate the effect of the Kernel Linking (KL) module. As shown in Tab. 2, KL improves both the performance of panoptic segmentation and depth estimation. This is in-line with results conducted in [20, 21], showing that a unified approach with explicit linking between both tasks can boost single-task performance.

Table 1: Comparison to non-real-time state-of-the-art panoptic segmentation methods on the Cityscapes dataset. Methods marked with † use Mapillary Vistas [11] pre-training, backbones marked with * use ImageNet-22K [3] pre-training.

Method	Backbone	PQ ↑	PQ_{th} ↑	PQ_{st} ↑	#FLOPs ↓
RT-K-Net [15]	RTFormer [17]	60.2	51.5	66.5	-
Mask2Former [1]	ResNet-50 [7]	62.1	-	-	-
PanopticDepth [5]†	ResNet-50 [7]	64.1	58.8	68.1	-
kMaX-DeepLab [19]	ResNet-50 [7]	64.3	57.7	69.1	434G
Axial-DeepLab-XL [16]	Axial-ResNet-XL [16]	64.4	-	-	2447G
Mask2Former [1]	Swin-L [9]*	66.6	-	-	514G
OneFormer [8]	Swin-L [9]*	67.2	-	-	543G
Axial-DeepLab-XL [16]†	Axial-ResNet-XL [16]	67.8	-	-	2447G
kMaX-DeepLab [19]	ConvNeXt-L [10]*	68.4	62.9	72.4	1673G
OneFormer [8]†	ConvNeXt-L [10]*	**70.1**	**64.6**	**74.1**	497G
Ours	RTFormer [17]	61.7	54.6	66.8	**155G**
Ours†	RTFormer [17]	64.0	56.3	69.5	**155G**

Table 2: Ablation study on Kernel Linking (KL) for MGNiceNet. The ablation study is conducted on the Cityscapes [2] validation dataset.

Method	Kernel Linking (KL)	PQ ↑	RMSE ↓
Ours	-	63.4	7.3
	✓	64.0	7.1

C Qualitative Results

Figure 1 shows qualitative results of our method on images of the Cityscapes test dataset and the KITTI [6] Eigen test split [4]. The first three rows per dataset show examples where our method performs well. Our method produces high-quality panoptic segmentation and depth estimation predictions, showing its effectiveness. The last row per dataset shows an example where the prediction is inaccurate. For Cityscapes, our model wrongly predicts a bicycle on the back of the fire truck, leading to an incorrect depth prediction. We argue that this corner case was not seen during training and is one opportunity for future improvement. For KITTI, our model cannot accurately predict the segmentation mask and depth of the bridge across the street. We argue that this is due to the usage of panoptic pseudo labels. Since KITTI does not provide panoptic ground truth, our model is only trained using pseudo labels. Thus, the panoptic prediction contains more incorrect segmentation masks than Cityscapes which also influences depth performance. Figure 2 shows a qualitative comparison of our MGNiceNet with our previous approach in monocular geometric scene understanding MGNet [14]. Compared to MGNet, the unified approach improves both panoptic segmentation and depth estimation performance. For panoptic segmentation, small objects,

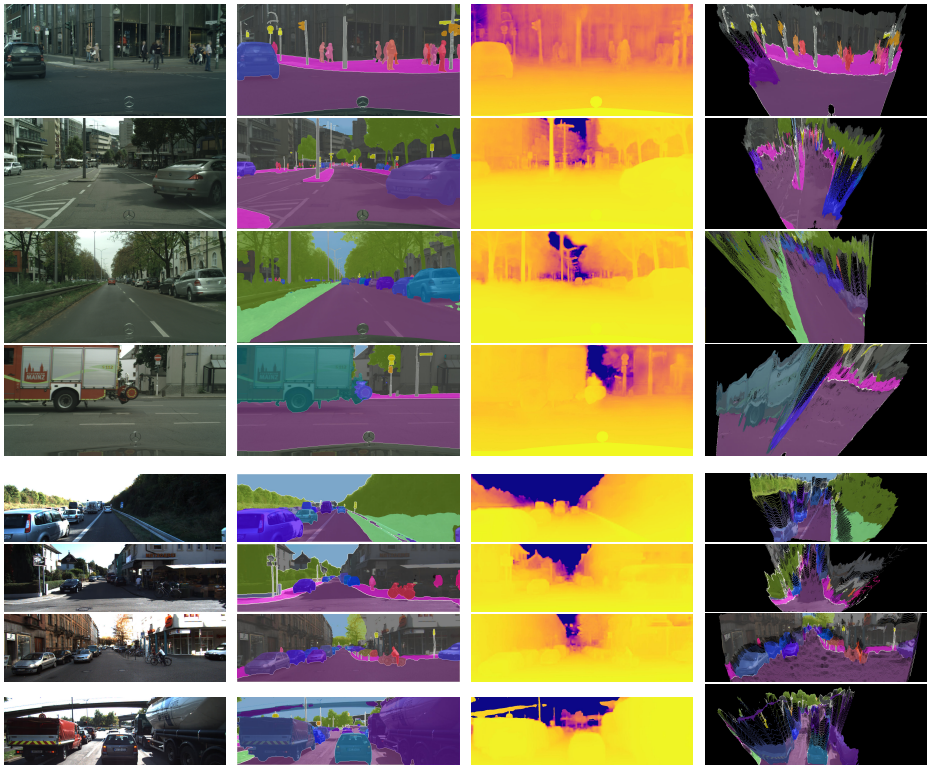

Fig. 1: Qualitative results of MGNiceNet on Cityscapes and KITTI. The columns (from left to right) show the input image, the panoptic prediction overlay, the monocular depth estimation, and the 3D panoptic point cloud prediction, respectively. The top block shows predictions on the Cityscapes [2] dataset, while the bottom block shows predictions on the KITTI [6] dataset. In each block, the last row shows an example of inaccurate predictions.

such as pedestrians or bicyclists, and large amorph regions, such as sidewalks or roads, are segmented more accurately. For depth estimation, MGNiceNet especially shows improvements in capturing fine-grained details, such as poles. We argue that the improvements are mainly due to our unified approach to monocular geometric scene understanding.

D Implementation Details

We implement our method in Pytorch [13] using the detectron2 framework [18]. Since the ego-car region of Cityscapes does not adhere to the Lambertian assumptions in the photometric loss, we load pre-calculated ego-car masks and omit this region in the photometric loss calculation and during inference. For panoptic segmentation post-processing, we follow RT-K-Net [15] settings. In particular, we

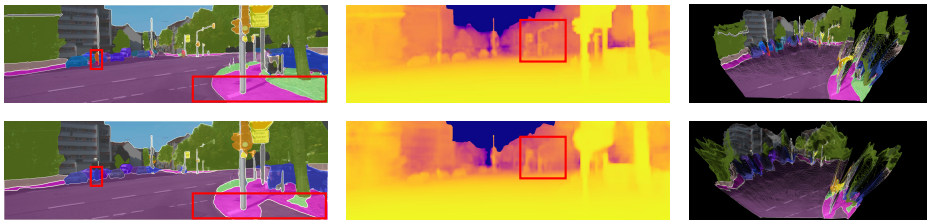

Fig. 2: Comparison of our MGNiceNet (top) with MGNet [14] (bottom). MGNiceNet achieves more accurate segmentation of small objects, such as bicyclists, and large areas, such as sidewalks, as well as a more fine-grained depth prediction.

filter out masks with confidences below $\delta_s = 0.3$ and an overlap below $\delta_o = 0.6$. Inference times are measured as an average of 500 forward passes through our model, including data-loading and post-processing. Inference times are measured on a single NVIDIA Titan RTX GPU without TensorRT [12] optimization. For comparison to state-of-the-art methods, we use inference times as stated in the respective publication or calculate inference times based on official code releases if available.

E Potential Negative Impact

While our technical innovations do not appear to have ethical biases, trained models can reflect the inherent biases of the dataset used. Hence, trained models should undergo an ethical review to ensure that predictions are not biased toward certain ethnic groups and that our method is not misused for applications such as illegal surveillance. The two used datasets, Cityscapes and KITTI, both contain sensitive personal data, such as the faces of human subjects. We used sensitive data carefully and according to the respective privacy protection agreement, *e.g.*, using anonymized images for our visualizations and demo videos if available.

References

1. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1290–1299 (2022)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3213–3223 (2016)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 248–255 (2009)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information Processing Systems (NIPS) **27** (2014)

5. Gao, N., He, F., Jia, J., Shan, Y., Zhang, H., Zhao, X., Huang, K.: Panopticdepth: A unified framework for depth-aware panoptic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1632–1642 (2022)
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) **32**, 1231–1237 (2013)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
8. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023)
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10012–10022 (2021)
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11976–11986 (2022)
11. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 5000–5009 (2017)
12. NVIDIA: Tensorrt library, `https://developer.nvidia.com/tensorrt`
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS) pp. 8024–8035 (2019)
14. Schön, M., Buchholz, M., Dietmayer, K.: Mgnet: Monocular geometric scene understanding for autonomous driving. Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 15784–15795 (2021)
15. Schön, M., Buchholz, M., Dietmayer, K.: Rt-k-net: Revisiting k-net for real-time panoptic segmentation. IEEE Intelligent Vehicles Symposium (IV) pp. 1–7 (2023)
16. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation, vol. 4, pp. 108–126. Springer (2020)
17. Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E., Wang, J.: Rtformer: Efficient design for real-time semantic segmentation with transformer. Advances in Neural Information Processing Systems (NeurIPS) pp. 7423–7436 (2022)
18. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
19. Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: k-means Mask Transformer, vol. 29, pp. 288–307. Springer (2022)
20. Yuan, H., Li, X., Yang, Y., Cheng, G., Zhang, J., Tong, Y., Zhang, L., Tao, D.: PolyphonicFormer: Unified Query Learning for Depth-aware Video Panoptic Segmentation, vol. 27, pp. 582–599. Springer (2022)
21. Zhang, J., Huang, J., Lu, S.: Unidaformer: Unified domain adaptive panoptic segmentation transformer via hierarchical mask calibration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11227–11237 (2023)