Supplementary Materials : Enhancing Self-supervised Monocular Depth Estimation via Piece-Wise Pose Estimation and Geometric Constraints

Pranjay Shyam
Faurecia IRYStec Inc.
Montreal, Canada
pranjay.shyam.psm@forvia.com

Alexandre Okon
Faurecia IRYStec Inc.
Montreal, Canada
alexandre.okon@forvia.com

HyunJin Yoo
Faurecia IRYStec Inc.
Montreal, Canada
hyunjin.yoo@forvia.com

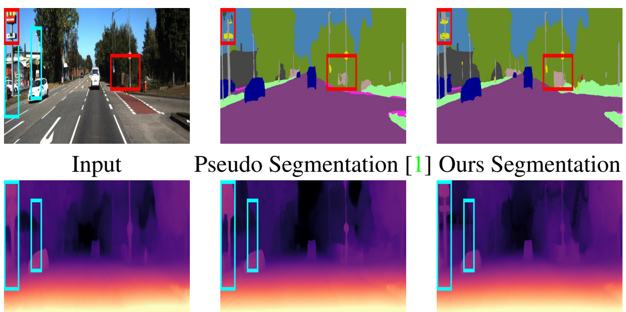

Figure 1. Performance of the proposed method compared to SoTA self-supervised algorithms. The proposed method demonstrates higher fidelity and sharper estimations around edges than SoTA. Red boxes in segmentation results highlight improved areas, while cyan-colored boxes highlight increased fidelity within estimated disparity.

1. Appendix-A

We follow the same architecture as that of HR Depth [8] in constructing the Depth estimation model while including YOSO [5] head for panoptic segmentation using the same encoder.

2. Appendix-C

We include qualitative results in Fig. 1 to demonstrate the effect of using segmentation branch results in poor occlusion handling wherein the object boundaries are thicker.

3. Appendix-F

We follow an incremental approach to perform ablation studies to determine different mechanisms' effects. Specifically, we first identify the effect of modifying the encoder network and summarize the performance of different backbone networks in Tab. 1. For our ablation, we vary the encoder within the Monodepth2 [3] network from lightweight MobileNetv3 [4] to heavy ResNet101. We measure the computational complexity (GMACs) and the total number of parameters to evaluate the performance of different encoders. For computing GMACs, we used the input size as 640 $\times$ 192. We observe that utilizing networks that tend to achieve higher performance on image classification tasks does not translate to higher performance on depth estimation. We validate this based on the observation that MobileNetv3 provides better performance to DenseNet-121 and ResNet-18 at a lower computational complexity. Furthermore, we also validate our initial motivation of using HRNet as an encoder, i.e., models designed for coarse prediction tasks cannot provide the necessary fidelity for dense prediction due to the loss of spatial correlation between pixels. Finally, we summarize that the high-performance gain achieved by HRNet is due to its architectural design of correlating features between different scales to ensure high-quality semantic features with good spatial properties.

4. Appendix-G

We include the ablation results of scale-Distillation parameter (α) sweep, Panoptic and Triplet Loss in Tab. 2.

5. Appendix-H

We include the ablation for single-frame and multi-frame MDE in Tab. 3.

6. Appendix-I

We include ablation on integration of panoptic segmentation branch in Tab. 2.

Table 1. Ablation studies on KITTI-2015 Eigen Split to examine the effect of varying the encoder architecture within the depth estimation network. We observe models having higher prediction accuracy for image classification to necessarily translate into higher performance for dense prediction tasks.

Backbone	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	GMACs	Params
ResNet-18	0.114	0.864	4.817	0.192	0.875	0.959	0.981	8.042	14.84
ResNet-50	0.110	0.831	4.642	0.187	0.883	0.962	0.982	16.643	34.57
SEResNet50 [6]	0.114	0.908	4.868	0.191	0.878	0.960	0.981	16.646	35.05
ResNest-14d [10]	0.113	0.860	4.738	0.189	0.878	0.960	0.982	13.339	17.57
ResNext-101 [9]	0.111	0.906	4.797	0.189	0.884	0.961	0.981	26.368	51.53
HRNet-30	0.105	0.877	4.736	0.185	0.892	0.963	0.982	22.622	32.46
DenseNet-121 [7]	0.111	0.883	4.866	0.191	0.882	0.960	0.981	13.251	13.60
Mobilenetv3 [4]	0.112	0.916	4.889	0.191	0.879	0.959	0.981	3.299	6.68

Table 2. Ablation studies on KITTI-2015 Eigen Split to examine the effect of varying different components within the decoder of the depth estimation network. The networks are trained using inputs of resolution 640 $\times$ 192.

	α	Triplet Loss	Panoptic	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				ResNet18 as Encoder						
1				0.114	0.864	4.817	0.192	0.875	0.959	0.981
2	0.5			0.112	0.916	4.889	0.191	0.879	0.959	0.981
3	1.0			0.109	0.862	4.809	0.190	0.890	0.962	0.982
4	1.0	✓		0.106	0.854	4.650	0.187	0.883	0.961	0.982
5	1.0	✓	✓	0.104	0.821	4.678	0.185	0.895	0.963	0.982
				HRNet as Encoder						
6				0.105	0.877	4.736	0.185	0.892	0.963	0.982
7	0.5			0.104	0.859	4.678	0.185	0.895	0.963	0.982
8	1.0			0.101	0.801	4.599	0.184	0.885	0.964	0.982
9	1.0	✓		0.100	0.793	4.544	0.184	0.885	0.966	0.984
10	1.0	✓	✓	0.098	0.713	4.397	0.181	0.899	0.966	0.984

Table 3. Qualitative results of SoTA on the NuScenes dataset.

Method	Res.	Backbone	Sem.	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [3]	640 $\times$ 192	ResNet-18		0.187	1.865	8.322	0.303	0.722	0.882	0.939
HRDepth [8]	640 $\times$ 192	ResNet-18		0.179	1.801	7.977	0.289	0.735	0.889	0.947
SAFENet [2]	640 $\times$ 192	ResNet-18	✓	0.172	1.652	7.776	0.277	0.752	0.895	0.950
Ours	640 $\times$ 192	HRNet		0.176	1.800	7.919	0.282	0.740	0.891	0.950
Ours	640 $\times$ 192	HRNet	✓	0.169	1.591	7.596	0.268	0.760	0.903	0.951
SAFENet [2]	1024 $\times$ 320	ResNet-18	✓	0.175	1.667	7.533	0.274	0.750	0.902	0.951
Ours	1024 $\times$ 320	HRNet		0.164	1.548	7.677	0.291	0.773	0.895	0.950
Ours	1024 $\times$ 320	HRNet	✓	0.149	1.299	5.914	0.195	0.874	0.939	0.980

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] Jaehoon Choi, Dongki Jung, Donghwan Lee, and Changick Kim. Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. *arXiv preprint arXiv:2010.02893*, 2020. 2

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019. 1, 2

[4] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. 1, 2

[5] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17819–17829, 2023. 1

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2

[7] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 2

[8] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *arXiv preprint arXiv:2012.07356*, 6, 2020. 1, 2

[9] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 2

[10] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi-Li Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alex Smola. Resnest: Split-attention networks. *ArXiv*, abs/2004.08955, 2020. 2