Proceedings of Machine Learning Research 176:308–313, 2022 NeurIPS 2021 Competition and Demonstration Track

Real-Time and Accurate Self-Supervised Monocular Depth Estimation on Mobile Device

Hong Cai HONGCAI@QTI.QUALCOMM.COM
Fei Yin FYIN@QTI.QUALCOMM.COM
Tushar Singhal TSINGHAL@QTI.QUALCOMM.COM
Sandeep Pendyam SPENDYAM@QTI.QUALCOMM.COM
Parham Noorzad PNOORZAD@QTI.QUALCOMM.COM
Yinhao Zhu YINHAOZ@QTI.QUALCOMM.COM
Khoi Nguyen KHOIN@QTI.QUALCOMM.COM
Janarbek Matai JMATAI@QTI.QUALCOMM.COM
Bharath Ramaswamy BHARATHR@QTI.QUALCOMM.COM
Frank Mayer FMAYER@QTI.QUALCOMM.COM
Chirag Patel CPATEL@QTI.QUALCOMM.COM
Abhijit Khobare AKHOBARE@QTI.QUALCOMM.COM
Fatih Porikli FPORIKLI@QTI.QUALCOMM.COM
*Qualcomm AI Research**

Editor: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

In this paper, we present our innovations on self-supervised monocular depth estimation. First, we enhance self-supervised monocular depth estimation with semantic information during training. This reduces the error by 12% and achieves state-of-the-art performance. Second, we enhance the backbone architecture using a scalable method for neural architecture search which optimizes directly for inference latency on a target device. This enables operation at more than 30 FPS. We demonstrate these techniques on a smartphone powered by a Snapdragon® Mobile Platform.[1]

Keywords: Monocular Depth Estimation, Self-Supervision, Neural Architecture Search

1. Introduction

Depth plays a key role in understanding the 3D world and is of great importance to a wide variety of applications, such as self-driving/ADAS, AR/VR, robotics, and mobile image processing. However, conventional learning-based depth estimation methods require a large amount of high-quality ground-truth annotations and/or stereo data, which are expensive to collect and pose considerable limitations.

Recently, self-supervised learning has been gaining increasing popularity for training deep neural networks, in areas such as classification (Chen et al., 2020), domain adaptation (Wang et al., 2020), and video segmentation (Xu and Wang, 2021). In fact, the number of papers on self-supervised or unsupervised learning has increased from 85 to 127 in CVPR 2021 as compared to the previous year, which is a nearly 50% jump.[2] Specifically,

* Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
1. Snapdragon is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.
2. CVPR 2021 paper statistics: https://github.com/hoya012/CVPR-2021-Paper-Statistics.

self-supervision has emerged as a new paradigm for training monocular depth estimation models (Godard et al., 2019), which makes it possible to do away with collecting massive ground-truth depth data and only train the network on unlabeled videos.

In this paper, we present our innovations on self-supervised monocular depth estimation, which utilizes semantic information and Neural Architecture Search (NAS) during training. More specifically, we exploit the commonalities between depth and segmentation and enable the depth network to digest key semantic information to significantly improve depth estimation accuracy with low complexity (Cai et al., 2021). To enable fast, real-time on-device inference, we further optimize the depth estimation neural network using Distilling Optimal Neural Networks (DONNA), a fast, efficient, and scalable NAS technique (Moons et al., 2021). Real-time depth estimation is demonstrated on a Snapdragon-powered smartphone.

2. Our Approach

In this section, we describe in more detail our technologies and innovations that enable an accurate and self-supervised monocular depth estimation algorithm running real-time on mobile device.

2.1. Enhancing Self-Supervised Monocular Depth Estimation with Semantic Information

We leverage our latest developed novel approach, X-Distill (Cai et al., 2021), to exploit semantic segmentation information to improve the self-supervised training of monocular depth estimation. More specifically, we allow the depth network to digest key, relevant semantic information during training, in addition to learning from the photometric matching of consecutive video frames. The accurate semantic segmentation information is provided by our state-of-the-art segmentation network (Borse et al., 2021). It is noteworthy that our method only modifies the training process and does not introduce additional computation during inference time.

When evaluating on the KITTI Eigen split (Eigen and Fergus, 2015) benchmark, our trained model achieves significantly smaller errors as compared to the state-of-the-art, e.g., reducing the squared relative error from 0.903 to 0.791 when using the same network architecture of Godard et al. (2019). Furthermore, our trained model achieves similar performance as compared to other much heavier state-of-the-art models while using significantly less computation. For instance, as compared to PackNet (Guizilini et al., 2020), our model has a similar squared relative error (ours: 0.791 vs. PackNet: 0.785) while using 96% less computation (in GMACs).[3]

2.2. Enhancing Backbone Architecture Using Neural Architecture Search

Neural networks for dense prediction tasks like segmentation and depth estimation are generally too complex to run efficiently on memory-, compute-, and power-constrained edge devices. Together with conventional approaches like quantization and model compression, Neural Architecture Search is gaining popularity to optimize models for efficient edge inference (Tan et al., 2019). While NAS research has made good progress (Cai et al., 2019; Liu et al., 2018; Tan et al., 2019), existing solutions still fail to address all challenges, notably lacking diverse search spaces, requiring high compute cost, not scaling efficiently, or

3. See Cai et al. (2021) for more detailed descriptions and evaluation of X-Distill.

Table 1: Performance evaluation on the KITTI Eigen split. The encoder backbones are indicated in the parentheses (e.g., ResNet50, DONNA). The frames-per-second (FPS) numbers are measured when the model runs on the smartphone. The depth estimation quality is measured with the squared relative error metric.

Model	#Param	On-Device FPS	Sq Rel
Monodepth2 (ResNet50)	32M	–	0.83
8-bit quantized	32M	23	0.83
X-Distill (ResNet50)	32M	–	0.69
8-bit quantized	32M	23	0.71
X-Distill (DONNA)	3.7M	–	0.75
8-bit quantized	3.7M	35	0.75

not providing reliable hardware performance estimates. Here, we utilize our latest NAS research – DONNA, Distilling Optimal Neural Network Architectures (Moons et al., 2021), that addresses these challenges. DONNA is a scalable method that finds optimal network architectures in terms of accuracy and latency at low cost by making use of diverse search space and direct hardware measurements. Architectures obtained by DONNA have been shown to provide 20%+ lower latency compared to models obtained with other state-of-the-art NAS techniques. We use DONNA to optimize the backbone of the depth estimation neural network, thereby further lowering the model compute and memory requirements, and enabling real-time depth estimation on a smartphone.

3. Implementation and Results

In this part, we provide implementation details on deploying the depth network on a smartphone and evaluate on-device performance in terms of both accuracy and inference speed.

3.1. Implementation

We run the monocular depth estimation network on a commercial mobile phone powered by Qualcomm® AI Engine.[4] For on-device processing, the trained depth estimation network is quantized using the AI Model Efficiency Toolkit (AIMET).[5]

3.2. Results

Table 3.1 shows our evaluation on the KITTI Eigen split (Eigen and Fergus, 2015). We compare three networks: 1) Monodepth2 (Godard et al., 2019) with ResNet50 (He et al., 2016) backbone, 2) X-Distill (Cai et al., 2021) with ResNet50 backbone, and 3) X-Distill with a backbone optimized via DONNA (Moons et al., 2021). By applying X-Distill, we significantly reduce the depth estimation error from 0.83 to 0.69 (in squared relative error) while using the same model architecture, as compared to the widely used Monodepth2. By additionally leveraging a backbone found via DONNA as the depth network encoder, we significantly reduce the model size by 88% while preserving the estimation quality.

4. Qualcomm AI Engine is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.
5. AIMET is a product of Qualcomm Innovation Center, Inc. It is available on https://quic.github.io/aimet-pages/index.html.

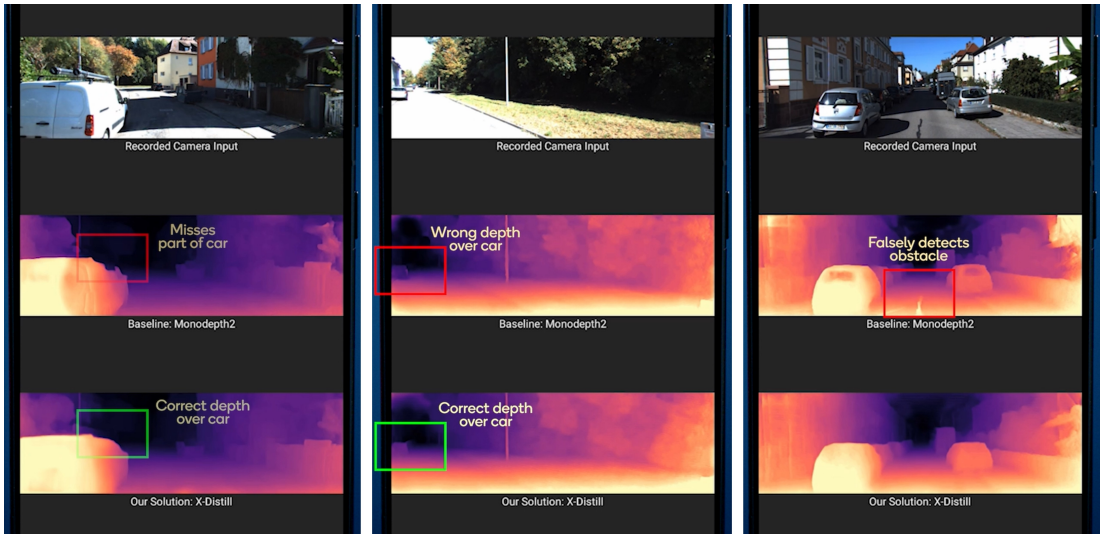

Figure 1: Sample visual results of monocular depth estimation captured on-device. The second and third rows show the estimated depth maps of Monodepth2 (Godard et al., 2019) and our approach, respectively. Sample regions where our method considerably improves the estimation quality are highlighted. Note that the network weights are quantized when running on the smartphone.

To efficiently deploy on device, the network weights are quantized to 8 bits using AIMET. As shown in Table 3.1, our improvements on depth estimation quality over Monodepth2 are well preserved through quantization. Sample visual results and comparisons captured from the phone screen are shown in Fig. 1. Furthermore, when running on the smartphone, our model with the DONNA backbone achieves a real-time inference speed of 35 FPS, while the ResNet50-based network runs at a considerably slower rate of 23 FPS.

4. Conclusions

In this paper, we presented our real-time and accurate self-supervised monocular depth estimation algorithm running on a commercial smartphone. To enable such a system, we leveraged our latest innovations. First, we applied our novel self-supervised training technique, X-Distill, which effectively utilized semantic information during training to considerably enhance the network's accuracy. Second, we adopted an enhanced backbone architecture optimized via our novel NAS solution, DONNA, which significantly reduced model size and improved inference speed. By using these technologies, we can efficiently run a self-supervised monocular depth estimation network on a Snapdragon-powered smartphone with state-of-the-art accuracy and a real-time inference speed of 35 FPS.

Acknowledgments

We thank Andrii Skliar, Tijmen Blankevoort, Joseph Soriaga, Murali Akula, Pat Lawlor, Armina Stepan, Kristine Chow, and Ron Tindall for their valuable help and support for this system demonstration.

References

Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-All: train one network and specialize it for efficient deployment. In *Proceedings of the International Conference on Learning Representations*, 2019.

Hong Cai, Janarbek Matai, Shubhankar Borse, Yizhe Zhang, Amin Ansari, and Fatih Porikli. X-Distill: Improving self-supervised monocular depth via cross-task distillation. In *Proceedings of the British Machine Vision Conference*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *Proceedings of the International Conference on Learning Representations*, 2018.

Bert Moons, Parham Noorzad, Andrii Skliar, Giovanni Mariani, Dushyant Mehta, Chris Lott, and Tijmen Blankevoort. Distilling optimal neural networks: Rapid search in diverse spaces. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. To appear.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations*, 2020.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.