

Knowledge Distillation for Fast and Accurate Monocular Depth Estimation on Mobile Devices

Yiran Wang, Xingyi Li, Min Shi, Ke Xian, and Zhiguo Cao

Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

Fwangyi ran, zgcaoG@hust.edu.cn

Abstract

Fast and accurate monocular depth estimation on mobile devices is a challenging task as one should always trade off the accuracy against the inference time. Most monocular depth methods adopt models with large computation overhead, which are not applicable on mobile devices. However, directly training a light-weight neural network to estimate depth can yield poor performance. To remedy this, we utilize knowledge distillation, transferring the knowledge and representation ability of a stronger teacher network to a light-weight student network. Experiments on Mobile AI 2021 (MAI2021) dataset demonstrate that our solution helps increase the fidelity of the output depth map and maintain fast inference speed. Specifically, with 94.7% less parameters than teacher network, the si-RMSE of student network only decrease by 10%. Moreover, our method ranks second in the MAI2021 Monocular Depth Estimation Challenge, with a si-RMSE of 0.2602, a RMSE of 3.25, and the inference time is 1197 ms tested on the Raspberry Pi 4.

1. Introduction

Real-time monocular depth estimation on mobile devices is a task in great demand. For example, accurate depth estimation helps robots sense the surroundings. Meanwhile, depth estimation is a preliminary task for many applications, such as semantic segmentation [7], bokeh effect rendering [31], and relighting [15]. Many other applications also need to be deployed on mobile devices like smartphones such as image enhancement [22] and super resolution [43]. However, most depth estimation methods adopt convolutional neural networks (CNNs) with complex architectures and large computation overhead, which makes them infeasible for real-time depth estimation on mobile devices due to the limited computing power.*

*Corresponding author.

Figure 1. Comparison of teacher and student networks. with 94.7% less parameters than teacher network, the RMSE of student network only decrease by 10%, which can show the efficiency of our knowledge distillation strategy. These metrics are defined in section 4.1. (a) Number of parameters; (b) RMSE; (c) si-RMSE; (d) rel.

A straightforward way to implement real-time depth estimation is to reduce the complexity of CNN. For dense prediction tasks, most methods adopt a fully convolutional network with an encoder and a decoder. One can apply light-weight encoder and decoder, e.g., MobileNet [18] and its variants [36], to extract features from input images. Model complexity can also be reduced by pruning redundant parameters [14]. Nonetheless, trade-off always exists between the accuracy and the efficiency. Directly training a light-weight network can not obtain depth maps with high accuracy and fidelity. Naturally, we come to the question that how we can take advantage of a complex and deep model with strong capability to enhance the performance of a light-weight network. To address this, in this paper, we transfer the strong representation ability of a teacher network to a light-weight student network via knowledge distillation [33]. Specifically, during training, the student network learns to extract similar feature maps at different scales as

