Supplementary Material: Learning to Adapt Multi-View Stereo by Self-Supervision

Arijit Mallick
arijit.mallick@uni-tuebingen.de
Jörg Stückler
joerg.stueckler@tuebingen.mpg.de
Hendrik Lensch
hendrik.lensch@uni-tuebingen.de

1 Computer Graphics Group
University of Tübingen
Tübingen, Germany

2 Embodied Vision Group
Max Planck Institute for Intelligent Systems
Tübingen, Germany

1 Introduction

In the following we provide additional details and results for our approach.

2 Network Architecture

We base our network architecture on MVSNet [6]. We do not use the depth refinement module, but extend the network with a subnetwork which predicts a confidence mask for the self-supervised loss. We provide a comparison of the two architectures in Fig. 1. The confidence mask subnetwork is a 4-layer CNN with a sigmoid activation unit at the end to generate values between 0 to 1. The confidence mask prediction network comprises of a combination of two basic sub-blocks. The first sub-block consists of a 2D convolutional layer (kernel size=3, stride=1) followed by a BatchNorm layer and ReLU as it’s activation function. This sub-block layer is used 3 times successively and then is followed by a final sub-block which consists of a 2D convolutional layer (kernel size=3, stride=1) followed by sigmoid activation function. The subnetwork receives as input the out-of-image projection masks and a photometric error maps for each neighbouring view. The photometric error maps are determined by warping the neighbouring views to the reference view and taking the difference.

3 Additional Quantitative Results

We also provide evaluation results on the DTU Buddha scan (see Table 1). Results of several classical and supervised methods are taken from [4]. Supervised MVSNet [6] fares best, while our self-supervised method ranks second and outperforms supervised and classical methods, highlighting the efficacy of our meta-learning approach.

© 2020. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
Figure 1: Network architecture difference between MVSNet (top) and our model (bottom). Our model builds on the initial stages of MVSNet: Deep features are extracted from the reference frame and the neighbouring frames. A plane-sweep cost volume is determined by homographic warping of the neighbouring feature maps to the reference view in a set of depth planes. This cost volume is refined in an encoder-decoder architecture and a depth map is obtained using a soft argmin operation. In case of MVSNet, this initial depth map is further improved by a refinement network. Supervised losses are determined that compare the refined depth with ground truth. In our model, we do not use the refinement branch. Instead, we determine photometric error maps by warping the neighbouring frames to the reference view and comparing them with the reference image. These photometric error maps are input to a confidence mask prediction network which also receives out-of-image projection masks. Finally, a self-supervised loss is computed by utilising the confidence mask on the photometric error maps.
Figure 2: Examples of predicted confidence masks. From left to right: reference frame, predicted confidence mask for first view (red: 1, black: 0 confidence), first view, predicted confidence mask for second view, second view, predicted depth maps on the DTU dataset.

Figure 3: Point-cloud reconstruction results on the DTU Buddha dataset. The qualitative results of our meta-learning approach appear superior to supervised SurfaceNet [3] and fairly close to MVSNet [6].

Table 1: Ranking of several methods by overall metric on the DTU Buddha dataset. Lower is better (best as bold). C: classical, Sup: supervised, Self: self-supervised, Meta: meta-learning. Our self-supervised meta-learning approach performs better than several supervised and classical methods. PT bMVS, FT DTU denotes pre-trained with Blended MVS dataset and fine-tuned with DTU dataset. DTU denotes trained on DTU dataset.
Figure 4: Point-cloud reconstruction results of our meta-learning approach on ETH3D low resolution multiview dataset reconstruction.

4 Additional Qualitative Results

ETH3D. Qualitative evaluation was performed on the ETH3D test dataset for low resolution multiview stereo. The model was meta-trained on BlendedMVS and fine-tuned on ETH3D low resolution training dataset. The test point-clouds show clear reconstruction results (see Fig. 4).

DTU. We provide additional point-cloud reconstruction results on the DTU dataset in Fig. 6. We also show depth maps predicted by our approach in Fig. 2).

References


Figure 5: Point-cloud reconstruction of DTU evaluation scans using our approach.
Figure 6: Depth maps predicted on the DTU test set. From left to right: reference frame, depth maps predicted by our network trained self-supervised on DTU only, depth maps predicted by our network pretrained self-supervised on BlendedMVS and fine-tuned self-supervised on DTU, our meta-learning approach pretrained on BlendedMVS and fine-tuned on DTU. Our approach predicts smoother depth maps at homogeneous surfaces and provides better completeness.