AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching
–Supplementary Material

Xiao Song1, Guorun Yang1,3, Xinge Zhu2, Hui Zhou1, Zhe Wang1,4, Jianping Shi1,5
1SenseTime Research 2The Chinese University of Hong Kong
3Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
4Shanghai AI Laboratory 5Qing Yuan Research Institute, Shanghai Jiao Tong University
{songxiao,yangguorun,zhouhui,wangzhe,shijianping}@sensetime.com zx018@ie.cuhk.edu.hk

1. Implementation Details

1.1. Specifications of Stereo Models

Here we introduce the network structures of our domain-adaptive stereo models: Ada-ResNetCorr and Ada-PSMNet. Ada-ResNetCorr is extended based on the ResNetCorr [5, 6, 4], which is a baseline model among correlation-based 2-D disparity networks. We extend the ResNetCorr, where “conv1_1” to “conv1_3” from the ImageNet-pretrained ResNet-50 [2] are adopted as the shallow feature extractor. We also replace the single-channel output convolutional layer with the soft-argmin block in 3-D disparity networks [3, 1] for disparity regression. The proposed cost normalization layer is directly applied on the extracted lower-layer features of two views before correlation, which are of 1/2 spatial size to the input image. The maximum displacement in the correlation layer is set to 128. Ada-PSMNet is extended based on the PSMNet [1], which is a baseline model among cost-volume based 3-D disparity networks. We follow the network structure of the PSMNet except the maximum disparity range is set to 256. The cost normalization layer is directly applied on the extracted lower-layer features of two views before constructing the 4-D cost volume.

1.2. Specifications of Self-Supervised Occlusion-Aware Reconstruction

Here we introduce the network structure of our occlusion prediction network (only adopted during training), which is a small three-layer fully-convolutional network. The first and second convolutional layers both use the $3 \times 3$ kernel and the stride of 1 with 32 output channels, followed by a batch normalization layer and a ReLU layer each. The last convolutional layer uses the $1 \times 1$ kernel and the stride of 1 with 1 output channel, afterwards the sigmoid layer is adopted for activation. Finally, a full-size occlusion mask is produced, whose element denotes the per-pixel occlusion probability from 0 to 1.

2. Qualitative Results of Color Transfer

As shown in Fig. 1, we provide qualitative results of our color transfer algorithm, from a synthetic dataset (source domain) to varied real-world datasets (target domain). As can be seen, our method enables highly effective image style translations, meanwhile ensuring the semantic invariance without geometrical distortions, which is of vital importance for the low-level stereo matching task.

References


* indicates equal contribution.
Figure 1. Qualitative results of color transfer from SceneFlow to real-world datasets. Top-down: transfer to KITTI, Middlebury, and ETH3D.