Unsupervised High-Resolution Depth Learning From Videos With Dual Networks Supplementary Material

1. Network Detail

Our model mainly consists of three parts: LR-Net, HR-Net and PoseNet. The PoseNet is the same as SFM-Learner's which contains 8 convolutional layers. Relu activation and batch normalization are used except for the last convolutional layer.

The encoder of HR-Net is a simplified ResNet18 (shown in Table 1). The final features of ResNet50 have 2048 channels so we pass them to the first two conv layers (iconv5) in the decoder of LR-Net to reduce the dimensions. Then these features are passed to HR-Net (needed to be resized to align the shape).

LR-Net and HR-Net are typical encoder-decoder architecture with skip-connection. The encoder of LR-Net is a standard ResNet-50 pre-trained on ImageNet. The decoders of LR-Net and HR-Net are almost the same and their details are shown in Table 2.

layer	in-chns	out-chns	res	input
econv1	3	64	2	image
epool	64	64	4	econv1
resblock1_1	64	64	4	epool
resblock1_2(elayer1)	64	128	8	resblock1_1
resblock2_1	128	128	8	resblock1_2
resblock2_2(elayer2)	128	256	16	resblock2_1
resblock3_1	256	256	8	resblock2_2
resblock3_2(elayer3)	256	512	32	resblock3_1

Table 1. HR-Net encoder structure. resblock is the Bottleneck module for building resnet.

layer	in-chns	out-chns	res	input	activation
upconv6	2048	512	64	elayer4	BN+Relu
iconv6	1536	512	32	↑upconv6 elayer3	BN+Relu
upconv5	512	256	32	iconv6	BN+Relu
iconv5	768	256	16	↑upconv5 elayer2	BN+Relu
upconv4	256	128	16	iconv5	BN+Relu
iconv4	384	128	8	↑upconv4 elayer1	BN+Relu
disp4	128	1	8	iconv4	Sigmoid
upconv3	128	64	8	iconv4	BN+Relu
iconv3	129	64	4	↑upconv3 epool ↑disp4	BN+Relu
disp3	64	1	4	iconv3	Sigmoid
upconv2	64	32	4	iconv3	BN+Relu
iconv2	97	32	2	↑upconv2 econv1 ↑disp3	BN+Relu
disp2	32	1	2	iconv2	Sigmoid
upconv1	32	16	2	iconv2	BN+Relu
iconv1	17	16	1	↑upconv1 ↑disp2	BN+Relu
disp1	16	1	1	iconv1	Sigmoid

Table 2. Decoder structure, in-chns is # of input channels, out-chns is # of output channels, res is the downscaling factor relative to the input image. ↑ denotes $2\times$ nearest-neighbor upsampling. All the kernel sizes are 3 and strides are 1. "elayer*" denotes the features from encoder. Both decoders of LR-Net and HR-Net are almost the same except upconv6 due to the concatenation.