Supplementary Materials for "Deep Bilateral Learning for Stereo Image Super-Resolution"

In the supplemental material, we describe the details of recurrent refinement module in Fig. 2. The recurrent refinement module consists of one RDB, one splat block and three bilateral filters.

I. RDB

In the RDB, we use 6 convolutions with a growth rate of 16. Then, the output of our RDB (I^0_{res}) is added to I^m_{res} in each bilateral filter.

II. Splat Block

Our splat block consists of a RDB block, two 3×3 convolutions with stride 2 and two upscalers. Given input features (e.g., $F^{'}_{right}$), it is first passed to the RDB block for multi-scale feature extraction. Then, the resulting features are fed to two cascaded convolutions with stride 2. Then we upsample the features with two cascaded upscalers. Then, the output $F_{splat} \in R^{h \times w \times 128}$ is fed to each bilateral filter for the generation of bilateral grid.

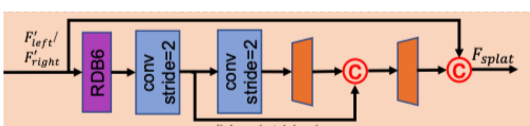

Fig. 1. The structure of the splat block.

III. Bilateral Grid

After splat block, the output features F_{splat} is fed to a 3×3 convolution for channel fusion. Then, we produce bilateral grid $\Gamma \in R^{w \times h \times d \times g}$ by reshape the fused feature F_{map}. Our bilateral grid delivers a compact set of transformation from LR images to HR images. Then, we use another 3×3 convolution to predict a guidance map Z from I_{in}, resulting in $Z \in R^{H \times W \times d}$. For each pixel (x, y) in an HR image $I^{HR} \in R^{H \times W \times 3}$, its coordinates are first transformed as $\hat{x} = x/H * h, \hat{y} = y/W * w$. Then, bilinear interpolation is used to obtain a guidance vector $z \in R^{h \times w \times d}$ from Z at location $(\hat{x}, \hat{y})$. Next, trilinear interpolation is performed on bilateral grid at location $(\hat{x}, \hat{y}, z(\hat{x}, \hat{y}))$ to produce a kernel $K_c \in R^{81 \times 1}$, which is then reshaped to $R^{3 \times 3 \times k \times k}$. The kernel K_c learns a mapping from a mapping from a 3×3 neighborhood centered at (x, y) in I_{in}. In our experiment, we set $k = 3, g = 81$ and $d = 8$.

IV. Bilateral Filter

For each pixel (x, y) in I^m, the sliced K_c is convolved with the neighborhood patch $I_{patch} \in R^{3 \times 3 \times 3}$ centered at (x, y). Then, the residual image I^m_{res} and I^m are added to the resulting to produce $I_{out} \in R^{H \times W \times 3}$. Next, I_{out} is passed to a space-to-depth layer [1] to produce feature $R^{h \times w \times 3r^2} (r = H/h)$, which is then concatenated with I^0_{res} to feed to the next bilateral filter to produce I^{m+1}_{res}. For gray image $I_{gray} \in R^{H \times W \times 1}$, we process the image $I^{'}_{gray} \in R^{H \times W \times 3}$, which is generated by replicating I_{gray} in the third dimension.

References

[1] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using hr optical flow estimation," *IEEE Transactions on Image Processing*, vol. 29, pp. 4323–4336, 2020.