Learning to Extract a Video Sequence from a Single Motion-Blurred Image

Meiguang Jin Givi Meishvili Paolo Favaro
University of Bern, Switzerland
{jin, meishvili, favaro}@inf.unibe.ch

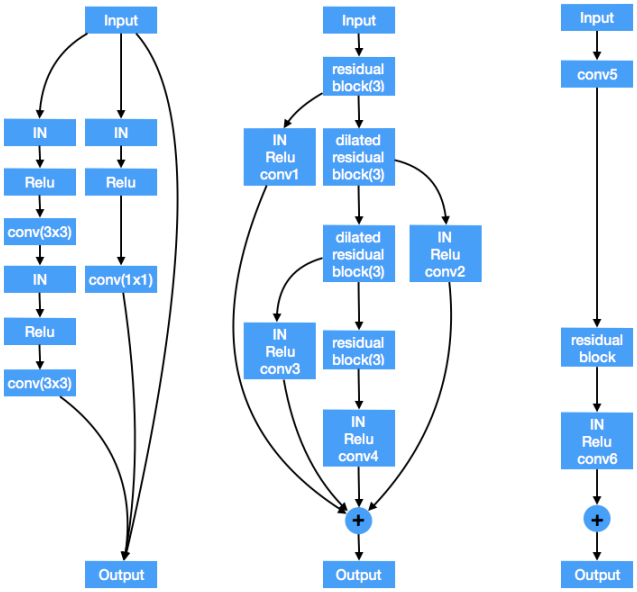

Figure 1. Details of our architecture. Left to right: the residual block, the feature refinement block and the feature fusion block (see Fig. 2).

1. Overview

In this supplementary material, we give the detailed structure of our proposed networks. The overall structure of the middle frame prediction network is shown in Fig. 2. Our middle frame prediction network consists of three parts: feature extraction, feature refinement and feature fusion. A blurry image is first split into three color channels. Resampling with factor 4 is applied to each color channel separately. Resampling creates $\texttt{factor}^2$ sub-sampled images. Each sub-sampled image is obtained by sampling the original image one pixel every `factor` pixels (along both axes). Every sub-sampled image differs by the initial sampled pixel on the original input (up to $\texttt{factor}^2$ possible initial positions). Moreover, the 16 sub-sampled images are generated for each color channel. Resampling can also be seen as the inverse process of the sub-pixel convolution proposed in [1]. Feature extraction

Table 1. The layer architecture of the middle frame prediction network. RB and IN in the table indicate a residual block and instance normalization.

layer	normaliz.	non-lin.	convolution	dilation
conv0			$144 \times 16 \times 5 \times 5$	1×1
RB(1-3)	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(4)	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 3 \times 3$	2×2
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(5)	IN	ReLU	$144 \times 144 \times 3 \times 3$	2×2
	IN	ReLU	$144 \times 144 \times 3 \times 3$	4×4
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(6)	IN	ReLU	$144 \times 144 \times 3 \times 3$	4×4
	IN	ReLU	$144 \times 144 \times 3 \times 3$	8×8
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(7)	IN	ReLU	$144 \times 144 \times 3 \times 3$	8×8
	IN	ReLU	$144 \times 144 \times 3 \times 3$	4×4
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(8)	IN	ReLU	$144 \times 144 \times 3 \times 3$	4×4
	IN	ReLU	$144 \times 144 \times 3 \times 3$	2×2
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(9)	IN	ReLU	$144 \times 144 \times 3 \times 3$	2×2
	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
RB(10-12)	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 3 \times 3$	1×1
	IN	ReLU	$144 \times 144 \times 1 \times 1$	1×1
conv1-4			$16 \times 144 \times 3 \times 3$	1×1
conv5			$64 \times 3 \times 3 \times 3$	1×1
RB	IN	ReLU	$64 \times 64 \times 3 \times 3$	1×1
	IN	ReLU	$64 \times 64 \times 3 \times 3$	1×1
	IN	ReLU	$64 \times 64 \times 1 \times 1$	1×1
conv6			$3 \times 64 \times 3 \times 3$	1×1

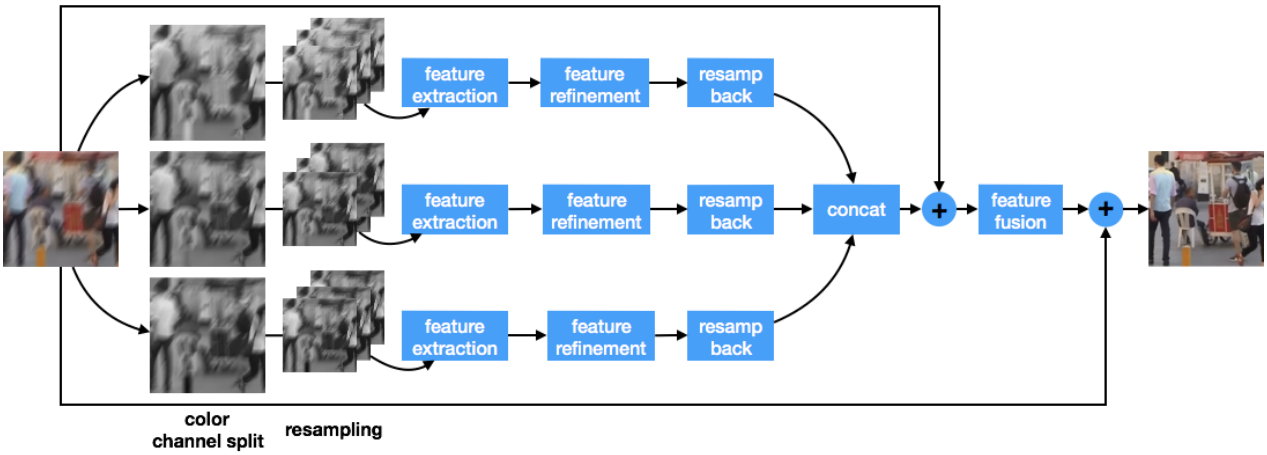

Figure 2. Middle frame prediction network architecture.

is a convolutional layer (conv0 in Table 1), where filters with size 5×5 elements are used. The architectures of the feature refinement and feature fusion blocks are shown in Fig. 1. In the feature refinement part, 12 residual blocks are used and architecture of each residual block is shown on the left column of Fig. 1. To further increase the receptive field, dilated convolutions are applied to the middle 6 residual blocks. The structure of our proposed middle frame prediction network is also described in Table 1. For the non-middle frame prediction networks, similar architectures are also used. The main differences are the number of channels (128 instead of 144), the resampling `factor` (5 instead of 4), and the feature extraction layers. For networks with two inputs, *e.g.*, $\phi_i(B, \phi_4(B))$, 64 features are extracted from B and $\phi_4(B)$ respectively, and concatenated.

References

[1] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 1