RPM-Net: Robust Pixel-Level Matching Networks for Self-Supervised Video Object Segmentation (Supply)

A. Network Architecture

In Table 1, we present the detailed architecture of RPM-Net. As we mentioned in the main paper, the embedding module is based on FCN-Resnet101 model [3] with modified upsampling layers. Also the matching module consists of 4 deformable convolution layers [2]. Moreover, we illustrate the training and inference scheme of our RPM-Net in Fig. 1 with layer numbers. Please see Section 3 in the main paper for discussion of our model.

B. Results of Ablation Studies

Table 2 shows the performance of different network settings as we discussed in Section 4.5 in the main paper. For experiments, we adopt FCN-ResNet18, FCN-ResNet34, FCN-ResNet50, and FCN-ResNet101 for the embedding module. And we also use standard convolution and dilated convolution [1] for the matching module. The matching module with deformable convolution shows the highest performance in our experiments. Please see Section 4.4 in the main paper for discussion of our model.

C. Visualization Results from Two Modules

We visualize embedding feature maps and sampling locations in Fig. 2. The target point is marked in green and sampling points are enclosed in a black border.

References

[1] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[2] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*. 2017.

[3] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*. 2015.

Number	Layer Description	Output Tensor Dimension
Embedding module		
I_{E1}	target frame	(3, H, W)
I_{E2}	reference frame	(3, H, W)
1	7×7 conv, 32, stride 2	(32, H/2, W/2)
2	3×3 maxpool, stride 2	(32, H/4, W/4)
3-11	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	(256, H/8, W/8)
11-23	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	(512, H/16, W/16)
24-92	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 23$	(1024, H/32, W/32)
93-101	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	(2048, H/64, W/64)
102	3×3 transposed conv, 1024, stride 2 add features from layer 102 and features from layer 92	(1024, H/32, W/32)
103	3×3 transposed conv, 512, stride 2 add features from layer 103 and features from layer 23	(512, H/16, W/16)
104	3×3 transposed conv, 256, stride 2 add features from layer 104 and features from layer 11	(256, H/8, W/8)
105	3×3 transposed conv, 32, stride 2 add features from layer 105 and features from layer 2	(32, H/4, W/4)
Matching module		
I_{M1}	concatenated features	(64, H/4, W/4)
I_{M2}	reference image (training) or $t-1$ segmentation mask (inference)	(3, H, W) or (1, H, W)
106	3×3 deformable conv, 32, stride 1 offset conv : 3×3 conv, 18, stride 1	(32, H/4, W/4)
107	3×3 deformable conv, 16, stride 1 offset conv : 3×3 conv, 18, stride 1	(16, H/4, W/4)
108	3×3 deformable conv, 2, stride 1 offset conv : 3×3 conv, 18, stride 1	(2, H/4, W/4)
109	1×1 deformable conv, 3, stride 1 offset : feature maps from layer 108	(3, H/4, W/4) or (1, H/4, W/4)

Table 1: The layer-by-layer definition of RPM-Net, which is end-to-end trainable.

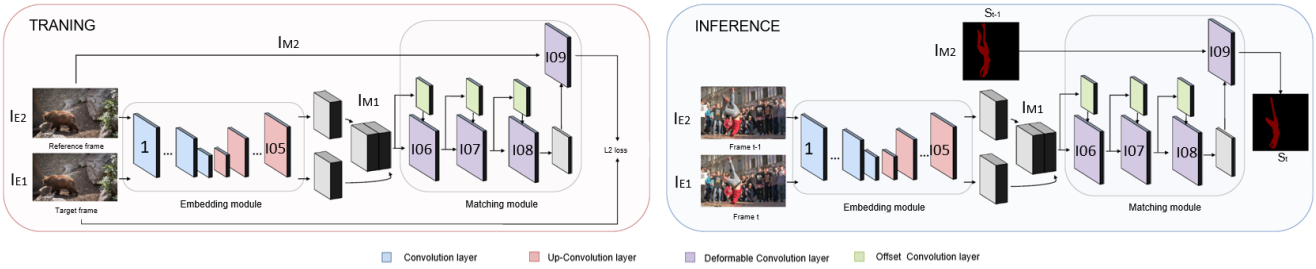

Figure 1: Illustration of RPM-Net architecture. We overlay the layer numbers on the figure.

	Convolution	Kernel Size / Dilation Rate	Number of Layer	Embedding Module			
				FCN-ResNet18	FCN-ResNet34	FCN-ResNet50	FCN-ResNet101
Matching Module	**Deformable**	**3 / 1**	**3**	**32.9 / 36.2**	**34.6 / 37.5**	**35.3 / 37.9**	**35.7 / 38.8**
	Standard	3 / 1	1	27.1 / 27.9	26.8 / 27.5	27.0 / 27.9	28.6 / 29.2
	Standard	3 / 1	3	31.8 / 34.1	31.3 / 32.7	30.4 / 31.8	31.7 / 33.5
	Standard	3 / 1	5	29.6 / 31.0	31.6 / 33.0	31.2 / 33.3	31.6 / 33.8
	Dilated	3 / 3	3	29.7 / 31.4	30.8 / 33.6	29.3 / 30.9	30.8 / 33.1
	Dilated	3 / 6	3	29.5 / 31.0	30.5 / 32.1	30.5 / 32.6	29.8 / 31.3
	Dilated	3 / 9	3	29.1 / 30.5	28.1 / 30.1	29.5 / 31.5	29.7 / 31.1

Table 2: Performance of our self-supervised training scheme with different network settings. We report the mIOU performance and contour accuracy ($\mathcal{J}/\mathcal{F}$) on the DAVIS-2017 validation set. Note that we maintain the last 1×1 deformable convolution layer for pixel-level matching.

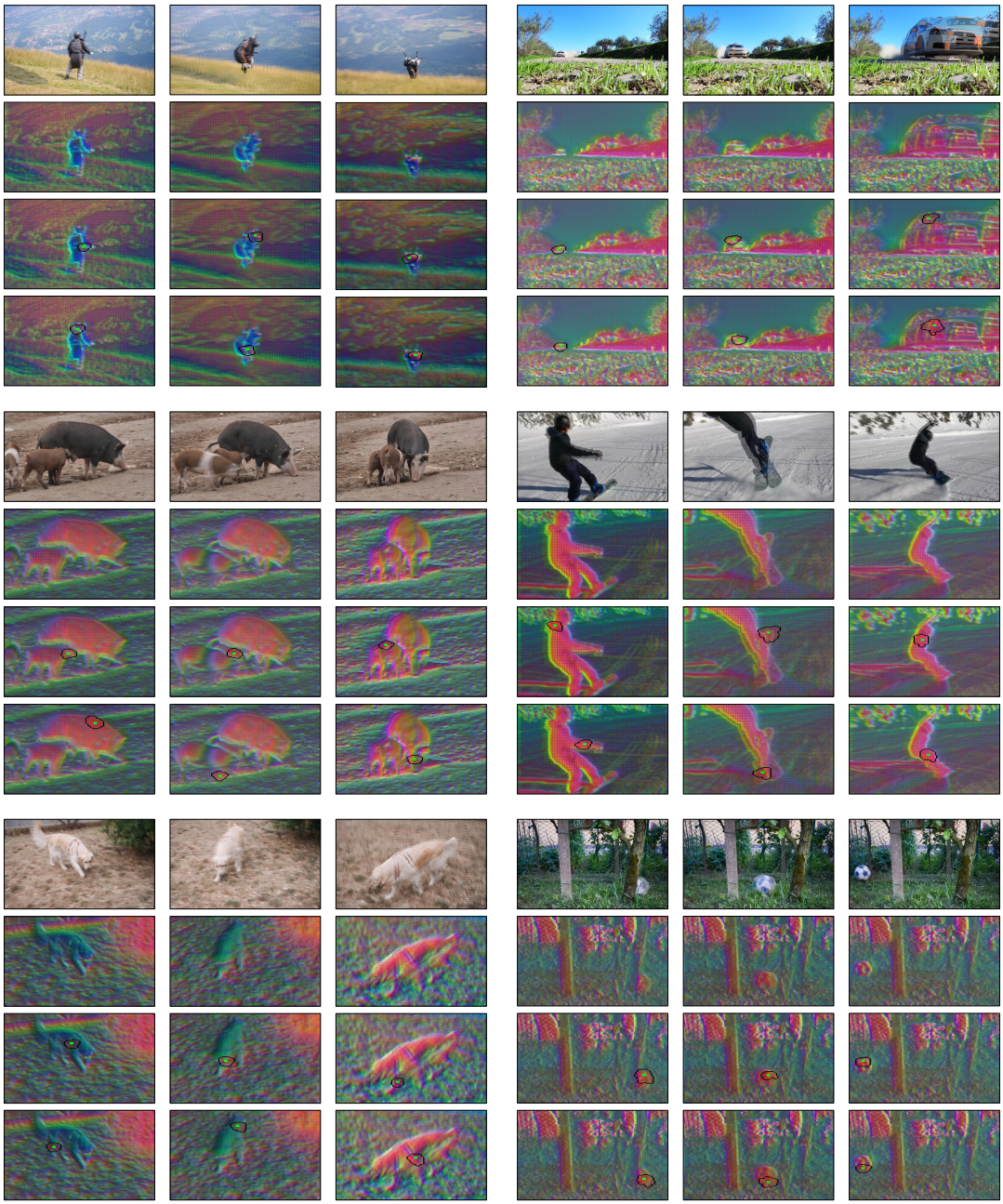

Figure 2: Visualization of the contributions of embedding and matching modules. For given frame $t-1$ and frame t (two images are overlaid in the first row), we show the examples of embedding features (second row) and sampling locations (third and fourth rows) for target pixel along the video sequence. The best view is in color and zoomed in.