Deep Learning for Semantic Video Segmentation

Jianping Fan
Department of Computer Science
UNC-Charlotte

Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html

The approaches for semantic video segmentation can be categorized into two groups:

(a) frame-by-frame approaches by performing semantic image segmentation on every frame;

(b) segmentation+tracking approaches by performing semantic image segmentation on the first frame and following by temporal tracking along frames.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

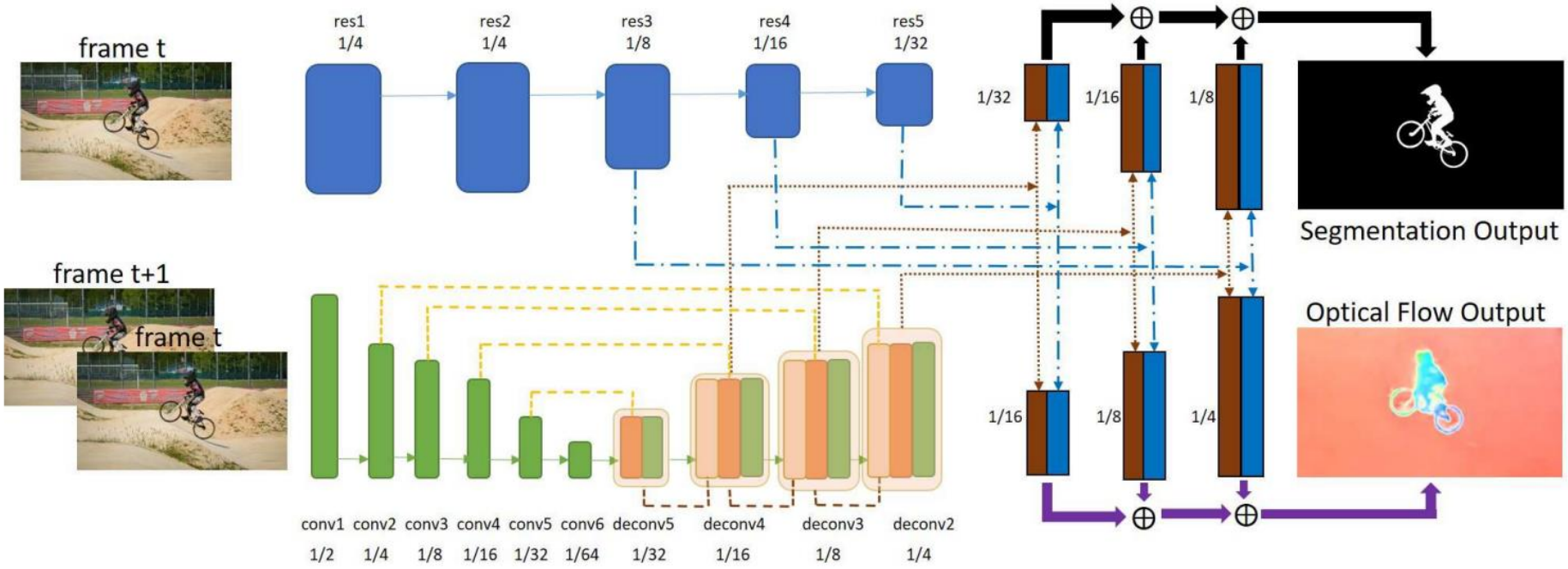
SegFlow has two branches: (a) segmentation branch; (b) optical flow branch, where useful information of object segmentation and optical flow is propagated bidirectionally in a unified framework.

The segmentation branch is based on a fully convolutional network, which has been proved effective in image segmentation task.

The optical flow branch takes advantage of the FlowNet model.

The unified framework is trained iteratively offline to learn a generic notion, and fine-tuned online for specific objects.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017



In SegFlow, the segmentation network based on a fully-convolutional ResNet-101 and the optical flow branch using the FlowNetS structure.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

In SegFlow, in order to construct communications between two branches (segmentation branch and optical flow estimation branch), an architecture is designed to bridge two networks (segmentation network and optical flow estimation network) during the up-sampling stage.

Specifically, feature maps are propagated bi-directionally through **concatenations** at different scales with proper operations (i.e., up-sampling or down-sampling) to match the size of different features.

An **iterative training** scheme is adopted to jointly optimize the loss functions for both segmentation and optical flow tasks.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

(a) Segmentation Branch

The segmentation branch is based on the ResNet-101 architecture, but modified for binary (foreground and background) segmentation predictions as follows: 1) the fully-connected layer for classification is removed, and 2) features of convolution modules in different levels are fused together for obtaining more details during up-sampling.

The ResNet-101 has five convolution modules, and each consists of several convolutional layers, Relu, skip links and pooling operations after the module. Specifically, we draw feature maps from the 3-th to 5-th convolution modules after pooling operations, where score maps are with sizes of 1/8, 1/16, 1/32 of the input image size, respectively. Then these score maps are up-sampled and summed together for predicting the final output.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

(b) Optical Flow Branch

The FlowNetS is selected for flow estimation. The optical flow branch uses an encoder-decoder architecture with additional skip links for feature fusions (feature concatenations between the encoder and decoder). In addition, a down-scaling operation is used at each step of the encoder, where each step of the decoder up-samples back the output.

Such structure shares similar properties with the segmentation branch and their feature representations are in similar scales, which enables plausible connections to the segmentation model, and vice versa.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

(c) Loss Functions

In segmentation network, a pixel-wise cross-entropy loss with the softmax function E is used during optimization.

$$\mathcal{L}_s(X_t) = -(1-w) \sum_{i,j \in fg} \log \mathbb{E}(y_{ij} = 1; \theta) \\ -w \sum_{i,j \in bg} \log \mathbb{E}(y_{ij} = 0; \theta), \qquad (1)$$

where i, j denotes the pixel location of foreground fg and background bg, y_{ij} denotes the binary prediction of each pixel of the input image X at frame t, and w is computed as the foreground-background pixel-number ratio.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

(c) Loss Functions

The optical flow branch uses an endpoint error (EPE) loss, which is defined as the following:

$$\mathcal{L}_f(X_t, X_{t+1}) = \sum_{i,j}((u_{ij} - \delta_{u_{ij}})^2 + (v_{ij} - \delta_{v_{ij}})^2), \quad (2)$$

where u_{ij}, v_{ij} denotes the motion at pixel (i, j) of input images from X_t to X_{t+1}, and $\delta_{u_{ij}}$ and $\delta_{v_{ij}}$ are network predictions. We use the images at frame t and $t + 1$ as the computed optical flow should align with the segmentation output (e.g., object boundaries) at frame t, so that their information can be combined later naturally.

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

(d) Bi-directional Model

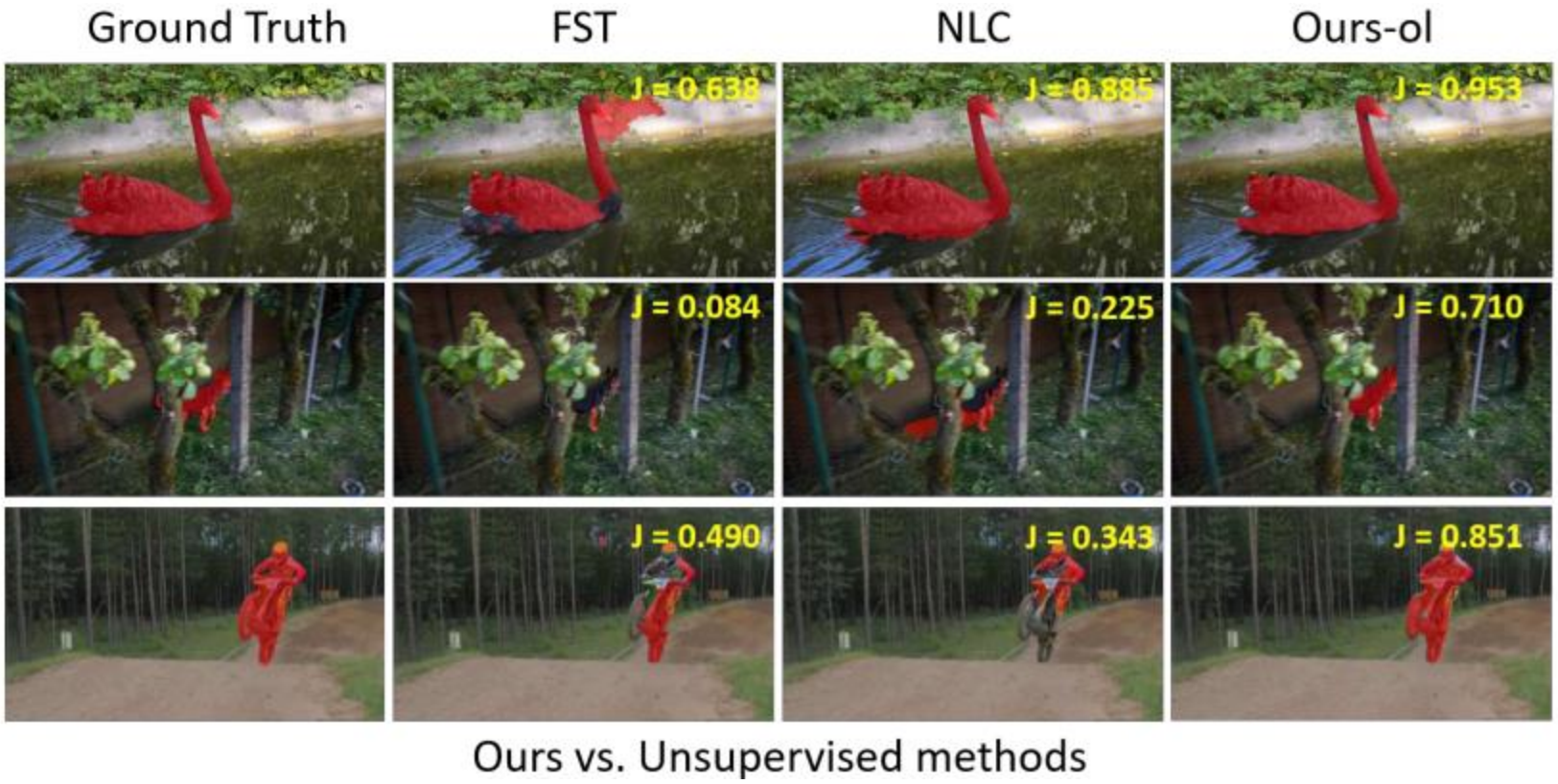
In order to make communications between two branches (segmentation and optical flow) to jointly predict their outputs. Therefore, the new optimization goal becomes to solve the following loss function that combines (1) and (2):

$$\mathcal{L}(X) = \mathcal{L}_s(X) + \lambda \mathcal{L}_f(X)$$

The bi-directional model propagates feature maps between two branches bi-directionally at different scales for the final prediction. For instance, features from each convolution module in the segmentation branch are first up-scaled (to match the size of optical flow features), and then concatenated to the optical flow branch. Similar operations are adopted when propagating features from segmentation to flow.

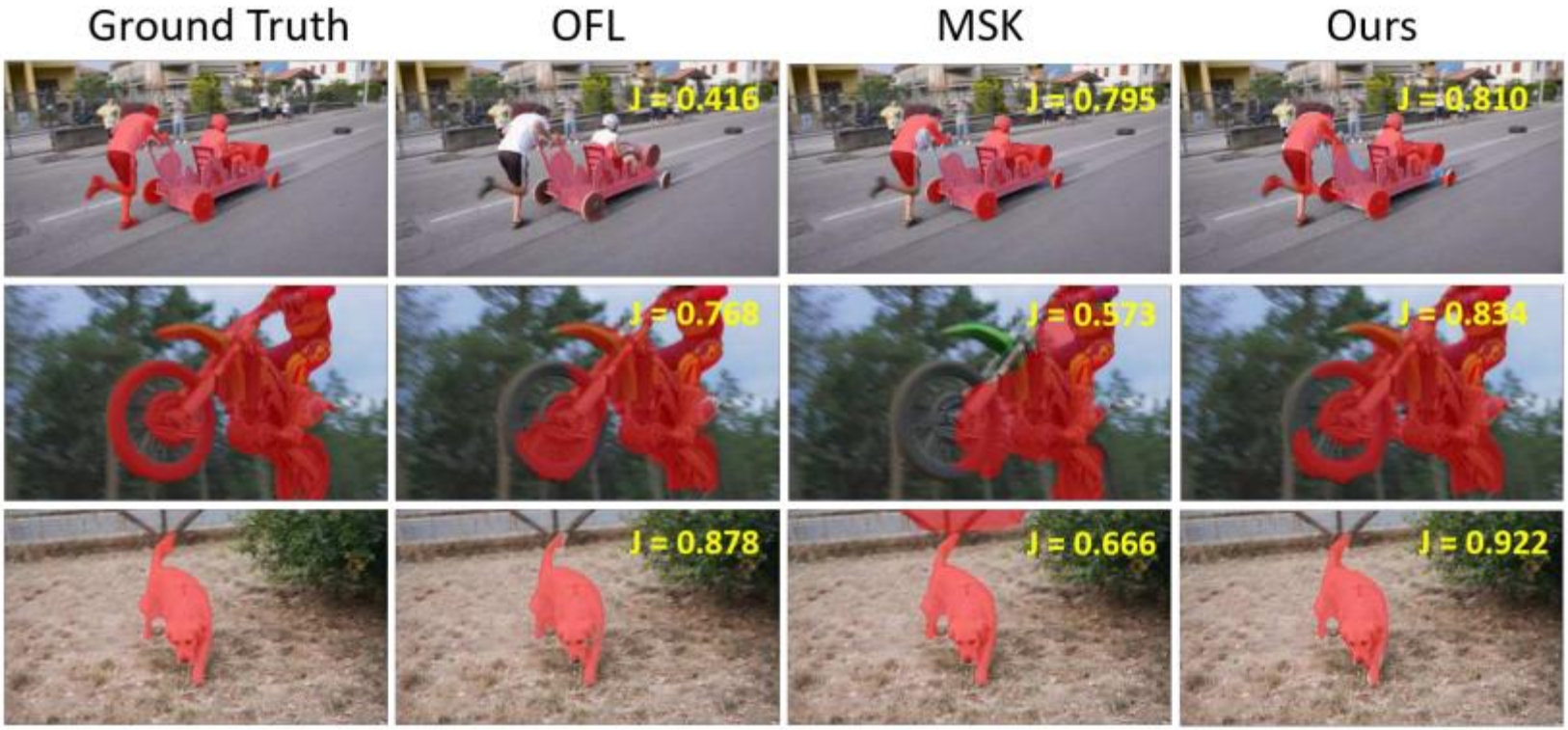
SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

The code and model are available at **https://github.com/JingchunCheng/SegFlow**.



SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

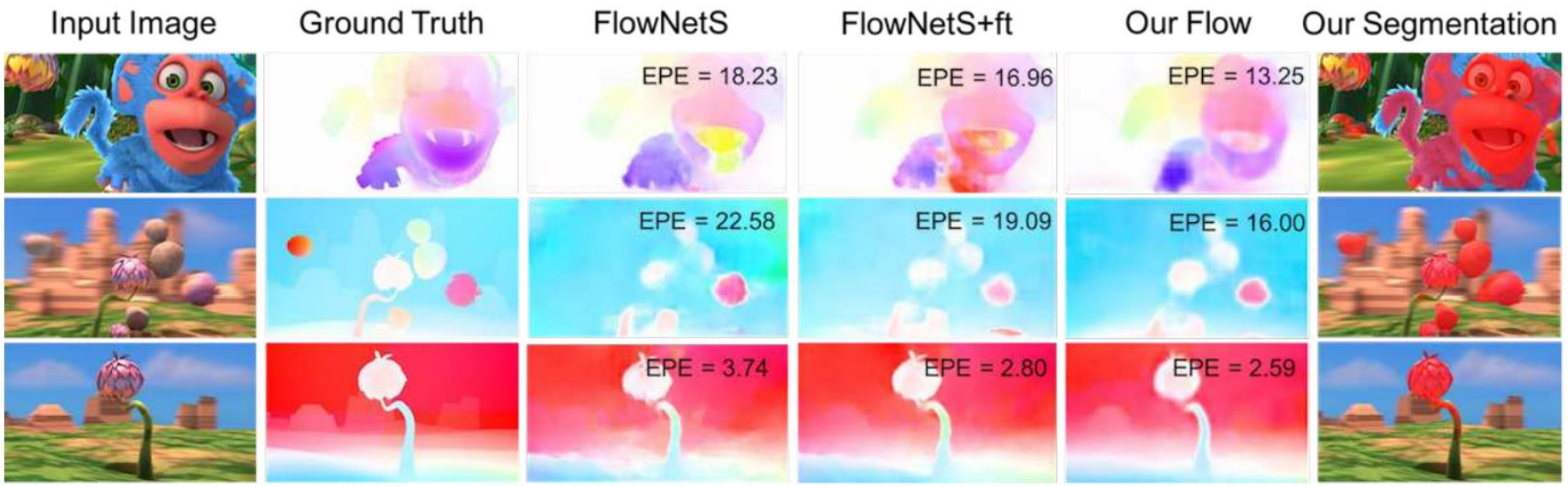
The code and model are available at **https://github.com/JingchunCheng/SegFlow**.



Ours vs. Semi-supervised methods

SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, ICCV 2017

The code and model are available at **https://github.com/JingchunCheng/SegFlow**.



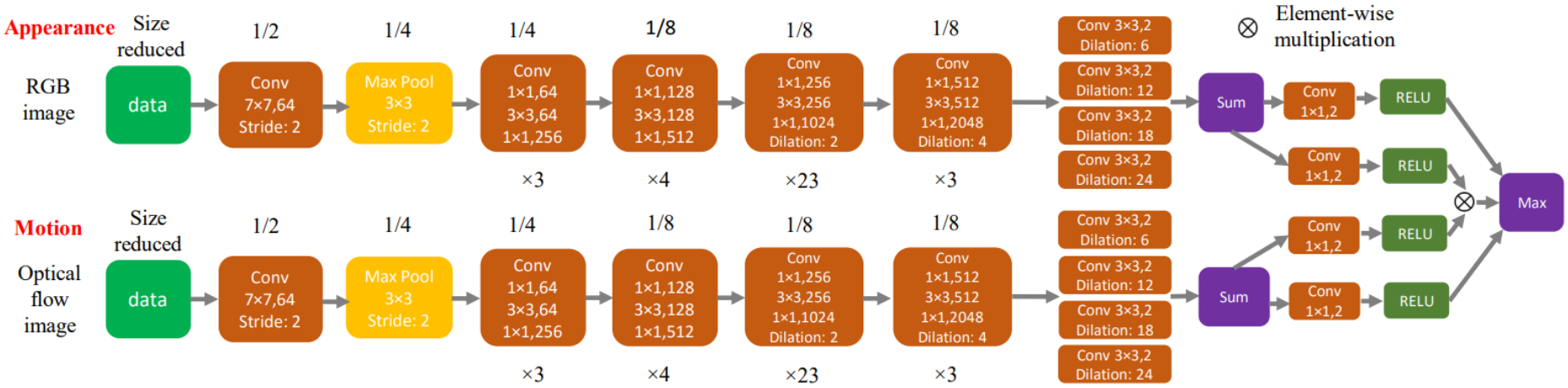
FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

FusionSeg learns to combine **appearance** and **motion information** to produce pixel level segmentation masks for all prominent objects, where a two-stream fully convolutional neural network is designed to fuse together motion and appearance in a unified framework.

Two individual streams encode generic appearance and motion cues derived from a video frame and its corresponding optical flow. These individual cues are fused in the network to produce a final object versus background pixel-level binary segmentation for each video frame.

FusionSeg can segment both static and moving objects in new videos without any human involvement.

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017



FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

(a) Appearance Stream

The image classification model ResNet-101 is adopted by replacing the last two groups of convolution layers with dilated convolution layers to increase feature resolution. This results in only an 8× reduction in the output resolution instead of a 32× reduction in the output resolution in the original ResNet model.

In order to improve the model's ability to handle both large and small objects, FusionSeg replaces the classification layer of ResNet101 with four parallel dilated convolutional layers with different sampling rates to explicitly account for object scale. Then we fuse the prediction from all four parallel layers by summing all the outputs.

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

(b) Motion Stream

FusionSeg also wants to train a strong generic motion model that can segment foreground objects purely based on motion, where the same network architecture as the appearance model is exactly used for the motion stream.

The motion model takes only optical flow as the input and is trained with automatically generated pixel level ground truth segmentations. In particular, the raw optical flow is converted to a 3-channel (RGB) color-coded optical flow image. This color-coded optical flow image is used as the input to the motion network.

By feeding the motion model with only optical flow, FusionSeg ensures the motion stream learns to segment objects from motion.

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

© Fusion Model

The final processing in FusionSeg pipeline joins the outputs of the appearance and motion streams, and aims to leverage a whole that is greater than the sum of its parts.

An object segmentation prediction is reliable if:

(1) either appearance or motion model alone predicts the object segmentation with very strong confidence

(2) their combination together predicts the segmentation with high confidence. This motivates the structure of the joint model.

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

© Fusion Model

FusionSeg implements the idea of fusion model by creating three independent parallel branches:

(1) a 1×1 convolution layer is applied followed by a RELU to the output of the appearance model

(2) a 1×1 convolution layer is applied followed by a RELU to the output of the motion model

(3) the structure of first and second branches are replicated and element-wise multiplication is applied on their outputs.

The element-wise multiplication ensures the third branch outputs confident predictions of object segmentation if and only if both appearance model and motion model have strong predictions. A layer is used to take the element-wise maximum to obtain the final prediction.

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

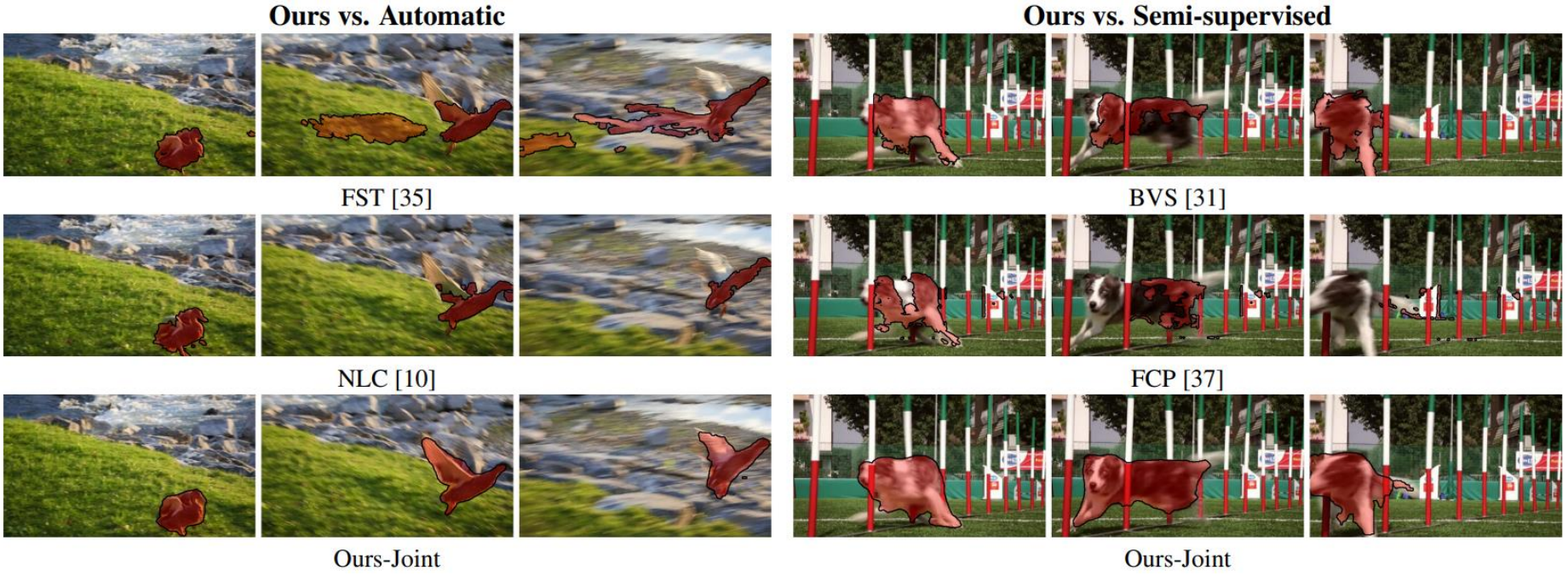
Appearance model (Ours-A)

Motion model (Ours-M)

Joint model (Ours-Joint)

Optical Flow Image

FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos, CVPR 2017

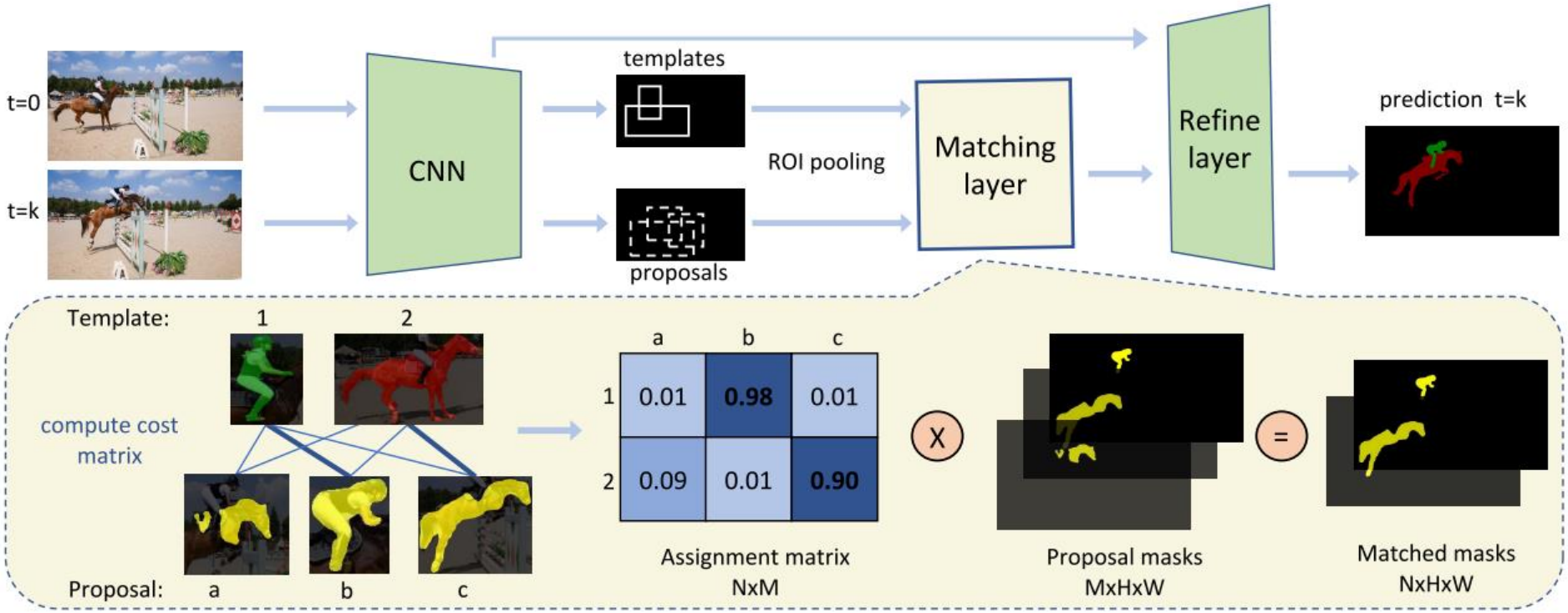


DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation, ICCV 2019

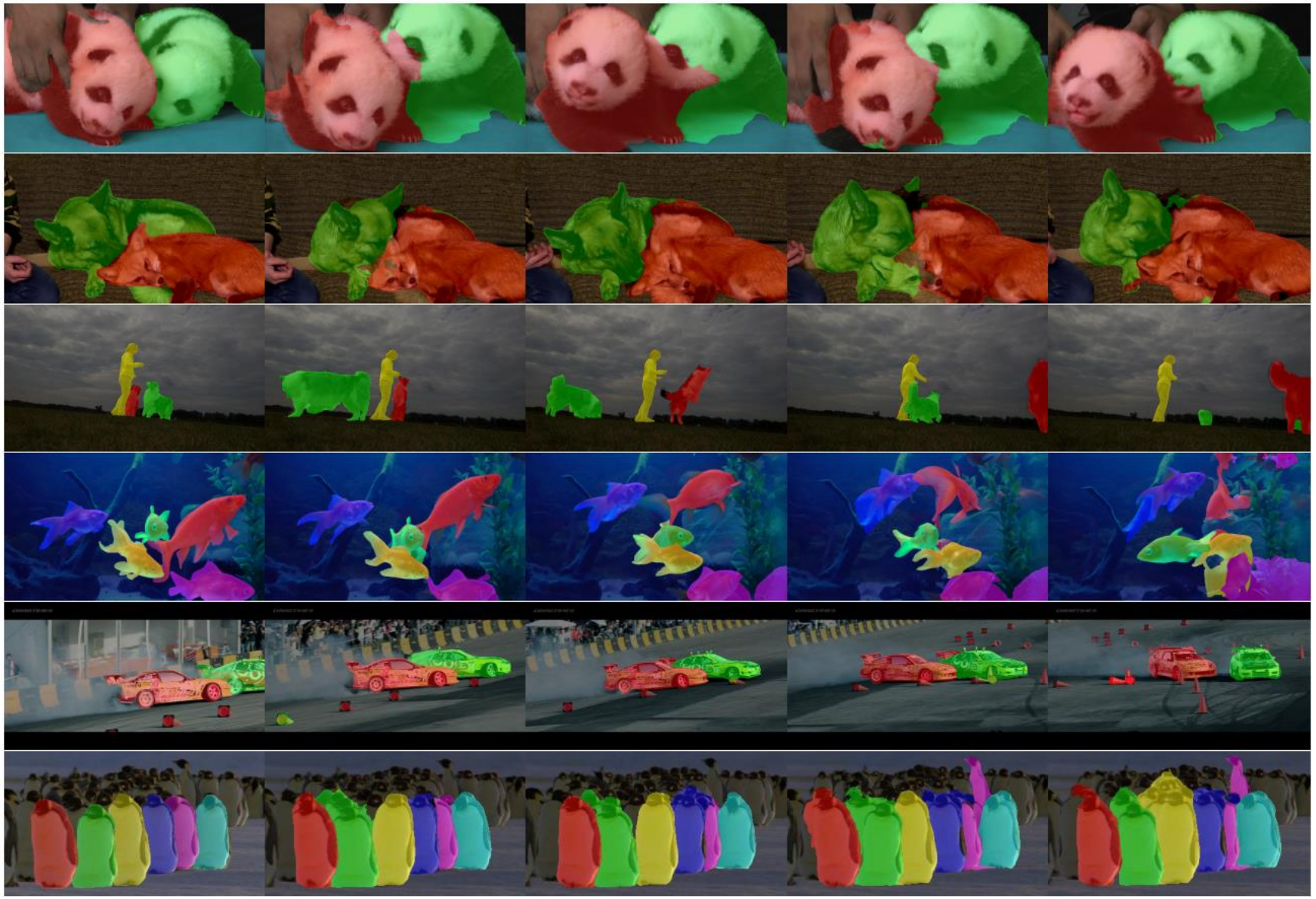
Relying on the Mask R-CNN backbone, DMM-Net extracts mask proposals per frame and formulate the matching between object templates and proposals at one time step as a linear assignment problem where the cost matrix is predicted by a CNN.

A differentiable matching layer is developed by unrolling a projected gradient descent algorithm in which the projection exploits the Dykstra's algorithm. The matching is guaranteed to converge to the optimum. In practice, it performs similarly to the Hungarian algorithm during inference. After matching, a refinement head is leveraged to improve the quality of the matched mask.

DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation, ICCV 2019



DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation, ICCV 2019



(a) 0% (b) 25% (c) 50% (d) 75% (e) 100%