Progressive Sparse Local Attention for Video Object Detection —Supplementary Material—

Chaoxu Guo[1,2] Bin Fan[1*] Jie Gu[1] Qian Zhang[3] Shiming Xiang[1,2]
Véronique Prinet[1] Chunhong Pan[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Horizon Robotics
{*chaoxu.guo,bfan,smxiang,prinet,chpan*}*@nlpr.ia.ac.cn*, {*qian01.zhang*}*@horizon.ai*

1. Visualization

We show the qualitative results of feature alignment (Fig. 1), video object detection (Fig. 2) and video semantic segmentation (Fig. 3) as below.

1.1. Feature Alignment

We show and compare results of feature alignment based on Progressive Sparse Local Attention (PSLA) and optical flow in Fig. 1, where the optical flow is produced by FlowNet [4] in DFF. They are visualized and compared on the following challenging cases:

1. Two objects interact with each other such as moving close or apart; as shown in Fig. 1(a) and 1(b), PSLA can still capture the motion information and align the features accurately under these situations. In Fig. 1(b), PSLA separates two objects totally although the motion blur deteriorates the appearance of two cars. On the contrary, optical flow is influenced by the motion blur and the aligned feature becomes obscure on the boundary of feature parts of two cars.

2. Movement of small and large objects; detection for objects of both small and large scales has been a challenging problem for a long time due to the insufficient context or appearance details in features. Thus it is crucial to obtain a good feature used for detection of these objects. PSLA can also handle the movement of large objects (Fig. 1(c)) and small objects (Fig. 1(d)) well and produce accurate aligned features for detection. Specifically, for small object in Fig. 1(d), PSLA captures more motion than DFF and generates feature that is aligned with the image better.

3. Objects enlarging or entering the view; in these cases, PSLA performs in par with optical flow and is capable of producing the aligned features that contains detailed information about the enlarged object (Fig. 1(e)) or feature points about the new part of objects (Fig. 1(f)).

All the results of feature alignment in Fig. 1 demonstrate that PSLA is a competitive alternative to optical flow for feature propagation.

1.2. Detection Results

Fig. 2 shows qualitative comparisons of detection results between R-FCN [3], DFF [6] and our framework on ImageNet VID validation dataset [5]. Green, red and yellow boxes indicate correct, missed and misclassified detections respectively. As shown in Fig. 2(a), our framework is robust to weird poses that are frequently encountered in videos. Comparing to R-FCN and DFF, our framework produces consistent detections of the panda on the right by considering temporal context of the video. When the discriminative part of the target is occluded sometimes (see Fig. 2(b)), R-FCN and DFF may miss the detection of that target because R-FCN ignores the information from previous frame and DFF relies heavily on the selected key frames without incorporating long-term temporal information. On the contrary, our framework aggregates and updates feature of sparse key frames across time and can produce more robust predictions. Due to the proper use of temporal information, our framework can detect objects robustly when motion blur occurs. Nevertheless, R-FCN and DFF are frequently confused by the deteriorated appearance, leading to miss detections (see red boxes in Fig. 2(c)) or incorrect detections (see yellow box in Fig. 2(c)).

1.3. Segmentation Results

The example results of our framework for video semantic segmentation on CityScapes validation dataset [2] is shown in Fig. 3. It can be seen that PSLA can capture the temporal evolvement of a video and propagate the semantic information throughout the video, which proves the effectiveness of PSLA.

[*] Bin Fan is the corresponding author

Components	runtime cost (ms)
ResNet101 †	43.5
Quality Net †	1.3
Res4b3 ‡	16.3
Transform Net ‡	1.6
PSLA †‡	1.7
rfcn head †‡	5.1

Table 1. Runtime cost of each component in our framework. † means the operation is run on key frames and ‡ means it is run on non-key frames.

2. Implementation Details of Video Semantic Segmentation

Following DFF [6], the task-specific sub-network of our framework for video semantic segmentation is replaced with deeplab-v2 [1], where PSLA is exploited to propagate feature maps across frames. The framework is trained and evaluated on training split and validation split of CityScapes dataset, respectively. The segmentation model is trained on 8 GPUs with a batch of 3 images in each GPU. During training, we sample the batch by firstly selecting the 19^{th} frame of a snippet as the non-key frame. Then another two images are sampled in $[-l+19, l+19]$ as key frames, where we set l as 5. The learning rate are 5×10^{-4} and 5×10^{-5} for the first 30k iterations and the last 7k iterations, respectively. During testing, since only the 19^{th} frame of a snippet is provided with pixel-wise annotations, we propagate nearby frames to 19^{th} frame to obtain multiple segmentation results and take the average of those results as the final result, the same as DFF did[1].

3. Runtime analysis

To further analyze our framework, the detail runtime of each component in our framework for video object detection is illustrated in Table 1. Obviously, the proposed PSLA only costs 1.7 ms and it is extremely time-saving to align the features across time. Contrary to the key frames that rely on a ResNet101 to extract features, a more efficient feature extractor Res4b3 is employed to extract the features of non-key frames, which are used by PSLA to align the features from key frames. Res4b3 only accounts for nearly $\frac{1}{3}$ computation cost of ResNet101(16.3 ms vs 43.5 ms) thus the overall computation cost of feature extraction is reduced. Besides, the proposed Recursive Feature Updating (RFU) that includes PSLA and Update Net, and Dense Feature Transforming (DFT) that includes PSLA and Transform Net only require a little cost of runtime (*i.e.* 3 ms and 3.3 ms respectively). Therefore, these fast runtime components guarantee the high efficiency of the proposed method.

[1] It is known from the github issue https://github.com/msracver/Deep-Feature-Flow/issues/25.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *T-PAMI*, 40(4):834–848, 2018. 2

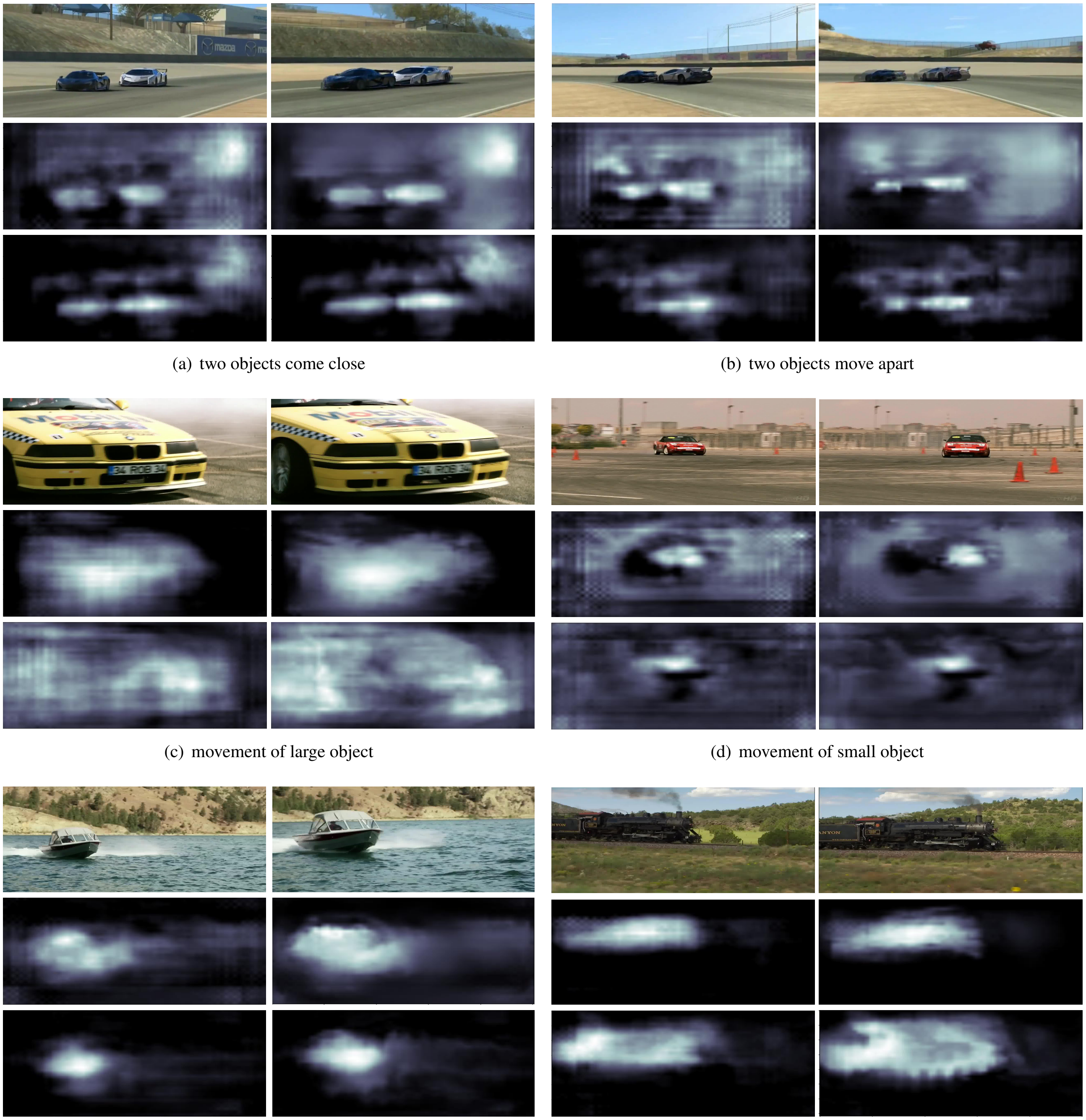
[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 1

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[6] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017. 1, 2



(a) two objects come close

(b) two objects move apart

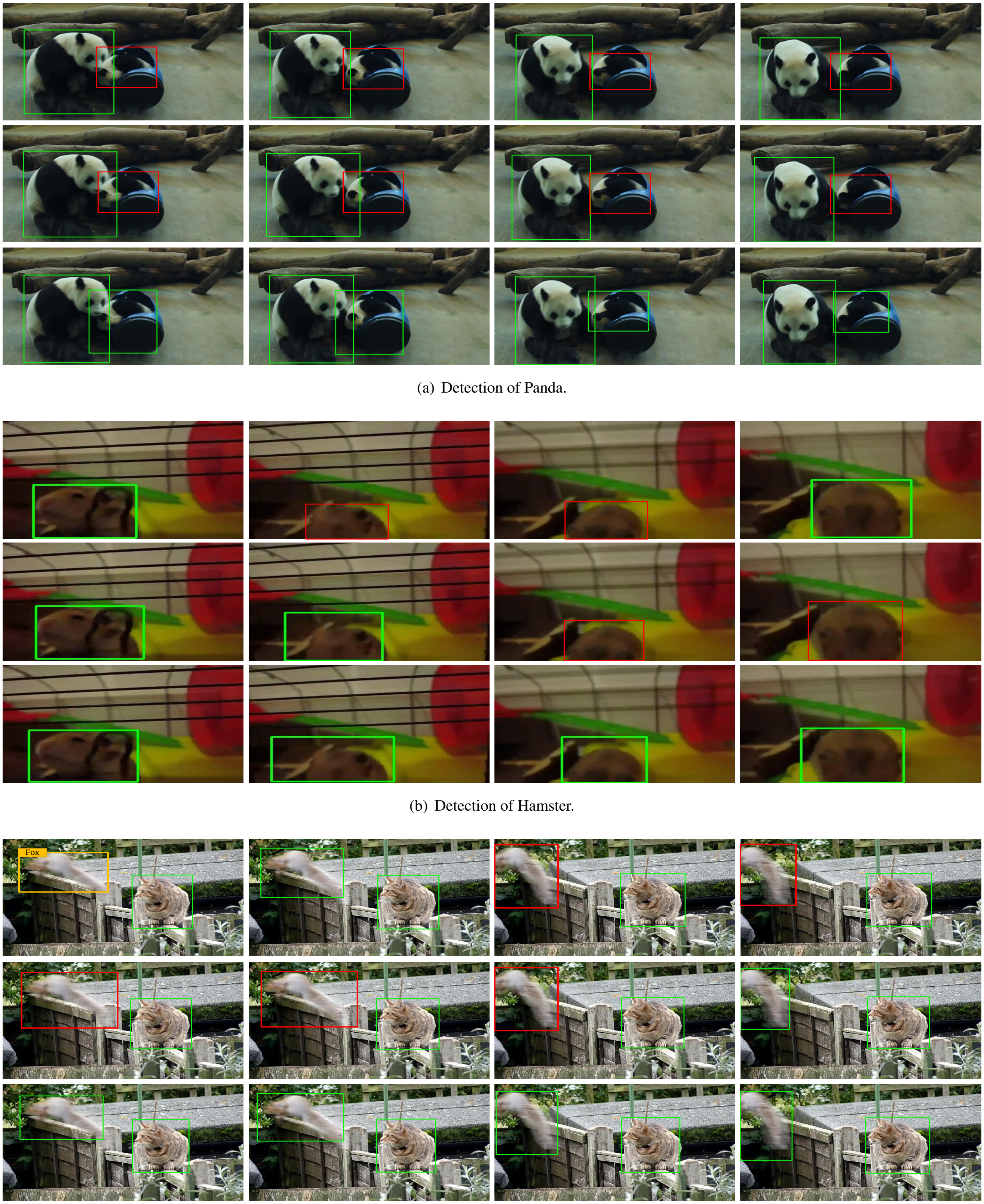
(c) movement of large object

(d) movement of small object

(e) object enlarges gradually

(f) object comes out gradually

Figure 1. Visualization results of feature alignment produced by PSLA and optical flow, where optical flow is generated by FlowNet in DFF. In each sub-figure, the first row represents the original images, *i.e.* key frame(left) and non-key frame(right), and the second and third row are feature maps of the corresponding frames produced with two different methods respectively, where feature map of non-key frame is obtained by aligning feature map of key frame with that of non-key frame using PSLA (second row) or warped from that of key frame using optical flow (third row). Best viewed in color.



(a) Detection of Panda.

(b) Detection of Hamster.

(c) Detection of Cat and Squirrel.

Figure 2. Example detection results of R-FCN, DFF and our framework on ImageNet VID validation dataset. In each sub-figure, the first, second and third row are the results of R-FCN, DFF and our framework respectively. Green, red and yellow boxes indicate correct, missed and misclassified detections respectively. For misclassified detection the predicted label is shown at the top-left corner of the yellow box. Best viewed in color.

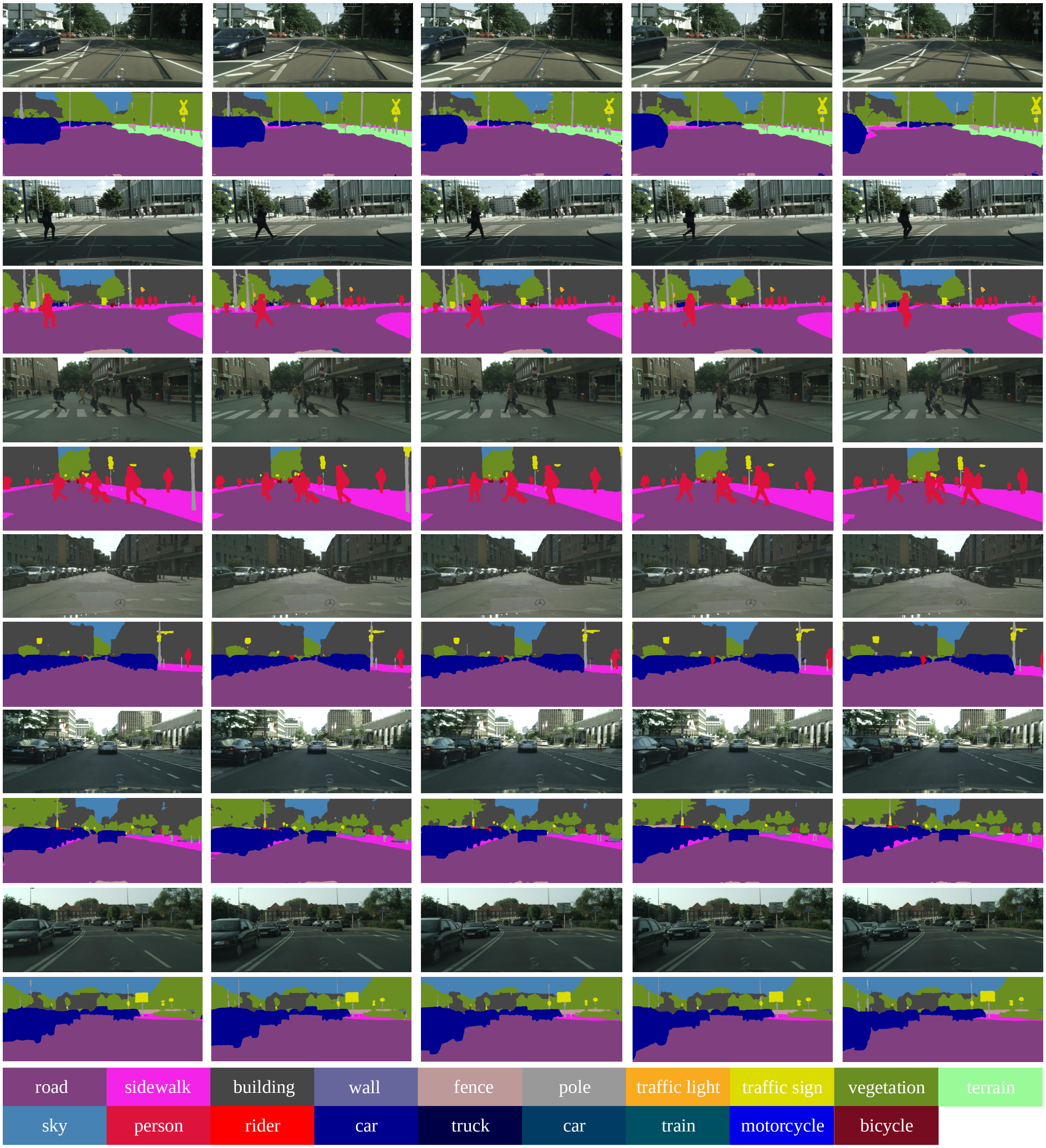


Figure 3. Example results of our framework for video semantic segmentation on CityScapes validation dataset. Best viewed in color.