Every Frame Counts: Joint Learning of Video Segmentation and Optical Flow

Mingyu Ding\textsuperscript{1,3} Zhe Wang\textsuperscript{5} Bolei Zhou\textsuperscript{4} Jianping Shi\textsuperscript{5} Zhiwu Lu\textsuperscript{1,2,*} Ping Luo\textsuperscript{3}

\textsuperscript{1}Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China
\textsuperscript{2}Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing 100872, China
\textsuperscript{3}The University of Hong Kong
\textsuperscript{4}The Chinese University of Hong Kong
\textsuperscript{5}SenseTime Research
dingmyu@gmail.com luzhiwu@ruc.edu.cn pluo@cs.hku.hk

Abstract

A major challenge for video semantic segmentation is the lack of labeled data. In most benchmark datasets, only one frame of a video clip is annotated, which makes most supervised methods fail to utilize information from the rest of the frames. To exploit the spatio-temporal information in videos, many previous works use pre-computed optical flows, which encode the temporal consistency to improve the video segmentation. However, the video segmentation and optical flow estimation are still considered as two separate tasks. In this paper, we propose a novel framework for joint video semantic segmentation and optical flow estimation. Semantic segmentation brings semantic information to handle occlusions for more robust optical flow estimation, while the non-occluded optical flow provides accurate pixel-level temporal correspondences to guarantee the temporal consistency of the segmentation. Moreover, our framework is able to utilize both labeled and unlabeled frames in the video through joint training, while no additional calculation is required in inference. Extensive experiments show that the proposed model makes the video semantic segmentation and optical flow estimation benefit from each other and outperforms existing methods under the same settings in both tasks.

Introduction

Video semantic segmentation, as an important research topic for applications such as robotics and autonomous driving, still remains largely unsolved. Current video segmentation methods mainly face two aspects of challenges: inefficiency and lack of labeled data. On the one hand, since frame-by-frame labeling of the video is time consuming, the existing data set contains only one annotated frame in each snippet, thus making the problem more challenging. On the other hand, to incorporate temporal information of the video, existing methods deploy feature aggregation modules to improve the segmentation accuracy, which leads to inefficiency during the inference phase.

Optical flow, which encodes the temporal consistency across frames in video, has been used to improve the segmentation accuracy or speed up the segmentation computation. For examples, the methods (Li, Shi, and Lin 2018; Zhu et al. 2017; Shelhamer et al. 2016) reuse the features in previous frames to accelerate computation. However, doing so will result in a decrease in the accuracy of the segmentation, and such methods are not considered in this paper. On the other hand, the methods (Fayyaz et al. 2016; Jin et al. 2017; Gadde, Jampani, and Gehler 2017; Nilsson and Sminchisescu 2018; Hur and Roth 2016) model multiple frames by flow-guided feature aggregation or a sequence module for better segmentation performance, which increases computational cost. Our motivation is to use optical flow to exploit temporal consistency in the semantic feature space for training better models, with no cost in inference time.

Current video segmentation datasets such as (Cordts et al. 2016) only annotate a small fraction of frames in videos. Existing methods focus on combining features of consecutive frames to achieve better segmentation performance. These methods can only use a small portion of frames in the video. Moreover, additional data is needed for training the feature aggregation module (FlowNet) in flow-guided methods (Nilsson and Sminchisescu 2018).

To address the two challenges of video semantic segmen-
tation, we propose a joint framework for semantic segmentation and optical flow estimation to fully utilize the unlabeled video data and overcome the problem of pre-computing optical flow. Semantic segmentation introduces semantic information that helps identify occlusion for more robust optical flow estimation. Meanwhile, non-occluded optical flow provides accurate pixel-level correspondences to guarantee the temporal consistency of the segmentation. These two tasks are related through temporal and spatial consistency in the designed network. Therefore, our model benefits from learning all the frames in the video without feature aggregation, which means that there is no extra calculation in inference. To the best of our knowledge, this is the first framework that joint learns these two tasks in an end-to-end manner.

We summarize our contributions as follow: (1) We design a novel framework for joint learning of video semantic segmentation and optical flow estimation with no extra calculation in inference. All the video frames can be used for training with the proposed temporally consistent constraints. (2) We design novel loss functions that handle flow occlusion in both two tasks, which improves the training robustness. (3) Our model makes the video semantic segmentation and optical flow estimation mutually beneficial and is superior to existing methods under the same setting in both tasks.

Related Work

Video Segmentation. Video semantic segmentation considers temporal consistency of consecutive frames compared to semantic segmentation. Existing methods mainly fall into two categories. The first category aims to accelerate computation by reusing the features in previous frames. Shelhamer et al. proposed a Clockwork network (Shelhamer et al. 2016) that adapts multi-stage FCN and directly reuses the second or third stage features of preceding frames to save computation. (Zhu et al. 2017) presented the Deep Feature Flow that propagates the high level feature from the key frame to current frame by optical flow learned in FlowNet (Dosovitskiy et al. 2015). (Li, Shi, and Lin 2018) proposed a network using spatially variant convolution to propagate features adaptively and an adaptive scheduler to ensure low latency. However, doing so will result in a decrease of accuracy, which is not considered in this paper.

Another category focuses on improving accuracy of segmentation by flow-guided feature aggregation or some sequence module. Our model falls into this category. (Fayyaz et al. 2016) proposed to combine the CNN features of consecutive frames through a spatial-temporal LSTM module. (Gadde, Jampani, and Gehler 2017) proposed a Net-Warp module to combine the features wrapped from previous frames with flows and those from the current frame to predict the segmentation. (Nilsson and Sminchisescu 2018) proposed gated recurrent units to propagate semantic labels. (Jin et al. 2017) proposed to learn from unlabeled video data in an unsupervised way through a predictive feature learning model (PEARL). However, such methods require additional feature aggregation modules, such as flow warping modules and sequence modules, which greatly increase the computational costs during the inference phase. Moreover, the feature aggregation modules of these methods can only process the annotated frame and several frames around it, while the rest of the frames are largely discarded in the training. In contrast, our method has two parallel branches for semantic segmentation and optical flow estimation, which reinforce each other in training but adds no extra calculation in inference. Furthermore, we can also leverage all video frames to train our model, with our temporally consistent constraint.

There are also other video segmentation methods with different settings. (Kundu, Vineet, and Koltun 2016) applied a dense random field over an optimized feature space for video segmentation. (Chandra, Couprie, and Kokkinos 2018) introduced densely-connected spatio-temporal graph on deep Gaussian Conditional Random Fields. (Hur and Roth 2016) estimates optical flow and temporally consistent semantic segmentation based on an 8-DoF piecewise-parametric model with a superpixelization of the scene. However, the iterative method based on superpixel cannot benefit from unsupervised data nor be optimized end-to-end. Our model can benefit from unsupervised data and be trained in an end-to-end deep manner, making the two tasks mutually beneficial. (Cheng et al. 2017) proposed to learn video object segmentation and optical flow in a multi-task framework, which focuses on segmenting instance level object masks. Both optical flow and object segmentation is learned in a supervised manner. In comparison, our task is semantic segmentation for the entire image and our optical flow is learned unsupervised. The two tasks cannot be directly compared.

Optical Flow Estimation. Optical flow estimation requires finding correspondences between two input images. FlowNet and FlowNet2.0 (Dosovitskiy et al. 2015; Ilg et al. 2017) directly compute dense flow prediction on every pixel through fully convolutional neural networks. PWC-Net (Sun et al. 2018) uses the current optical flow estimate to warp the CNN features of the second image. (Patraucean, Handa, and Cipolla 2015) introduced a spatio-temporal video autoencoder based on an end-to-end architecture that allows unsupervised training for motion prediction. (Jason, Harley, and Derpanis 2016; Meister, Hur, and Roth 2018; Ren et al. 2017a) utilizes the Spatial Transformer Networks (Jaderberg et al. 2015) to warp current images and measures photometric constancy. (Wang et al. 2018; Janai et al. 2018) models occlusion explicitly during the unsupervised learning of optical flow. In this work, the occlusion mask is refined by introducing the semantic information in our proposed approach. Moreover, the unsupervised optical flow estimation framework can be further extended to estimate monocular depth, optical flow and ego-motion simultaneously in an end-to-end manner (Yin and Shi 2018). (Ren et al. 2017b) proposed a cascaded classification framework that accurately models 3D scenes by iteratively refining semantic segmentation masks, stereo correspondences, 3D rigid motion estimates, and optical flow fields.

Methodology

Our framework, EFC model (Every Frame Counts), learns video semantic segmentation and optical flow estimation simultaneously in an end-to-end manner. In the following, we first give an overview of our framework and then describe each of its components in detail.
Framework Overview

An overview of our EFC model is illustrated in Figure 2. The input to our model is a pair of images \( I_i \) and \( I_{i+t} \), randomly selected from near-by video frames with \( t \in [1, 5] \). If either \( I_i \) or \( I_{i+t} \) has semantic labels, we can update weights of the network by supervised constraints with semantic labels as well as unsupervised constraints from near-by frame correspondence. It propagates semantic information across frames, and jointly optimize the semantic component and optical flow component to reinforce each other. Otherwise, only unsupervised consistency information can be utilized, and our network can benefit from the improvement in the optical flow component.

Specifically, our network consists of the following three parts, i.e., the shared encoder part, the segmentation decoder part and the flow decoder part. The shared encoder contains layers 1-3 of ResNet (He et al. 2016). It is helpful since semantic and flow information exchange among the representation, increasing the representation ability compared to (Zhao et al. 2017). The semantic decoder is adopted from layer 4 of ResNet if semantic label exists. The flow decoder combines intermediate feature from frame \( I_i \) and \( I_{i+t} \) via a correlation layer following (Ilg et al. 2017) to predict optical flow. A smoothness loss on flow result is applied to improve flow quality.

To enable end-to-end cross frame training without optical flow label, we design a temporal consistency module. Photometric consistency is usually adopted in optical flow estimation, where the first frame is warped to the next by optical flow and the warping loss can be used for training the network. In this work, we generalize the photometric loss to the feature domain. As the convolution neural network is translation invariant, the feature maps of adjacent frames should also follow the temporally consistent constraint.

Figure 2: The overall pipeline of our joint learning framework. The blocks represent the feature maps of our model, the gray dashed line represents the temporally consistent constraints. The gray solid line represents the occlusion handling module with the inconsistency of the segmentation maps.

Formally, \( S'_i \) can be obtained by

\[
S'_i = \text{Warp}(S_{i+t}, F_{i \rightarrow i+t}),
\]

(1)

where we adopt the differentiable bilinear interpolation for warping. Note that the warping direction is different from the flow direction. However, the flow can be invalid in occluded regions. So we estimate the occlusion maps \( O_{est}^{i+t} \) and \( O_{est}^{i} \) by checking if one pixel has a corresponding pixel in the adjacent frame. With the occlusion maps, we avoid penalizing the pixels in the occluded regions. The temporal consistency loss is thus defined as:

\[
L_{cons} = \sum_{x,y} (1 - O_{est}^{i+t}) \cdot \| S'_{i+t} - S'_{i} \|_2^2,
\]

(2)
produces finer masks by our loss function \(L\). F, fnels and a sigmoid layer for occlusion estimation. By back-branch, we add two convolutional layers with very few chan-
most of the parameters. For each block in non-occluded flow

The occlusion and the optical flow estimation network share

objects, in-and-out of image, change of view angle or so.
given frames, which can be caused by real occlusion by

refer to the pixels that are photometric inconsistent in two

Our model learns occlusion in a self-supervised manner. The

Occlusion Estimation

Our model learns occlusion in a self-supervised manner. The

occlusion defined here is a general term. By occlusion we refer to the pixels that are photometric inconsistent in two
given frames, which can be caused by real occlusion by objects, in-and-out of image, change of view angle or so.
The occlusion and the optical flow estimation network share

most of the parameters. For each block in non-occluded flow branch, we add two convolutional layers with very few channels and a sigmoid layer for occlusion estimation. By back-
ward optical flow \( F_{i+t \to i} \), we can calculate the correspondence between the two frames \( I_i, I_{i+t} \) in pixel-level. We de-
compose optical flow into vertical \( F_{i+t \to i}(y, x, 1) \) part and horizontal \( F_{i+t \to i}(y, x, 0) \) part. Then we have:

\[
y_{i+t} = y_i - F_{i+t \to i}(y_{i+t}, x_{i+t}, 1), \\
x_{i+t} = x_i - F_{i+t \to i}(y_{i+t}, x_{i+t}, 0).
\]

(3)

The occlusion mask \( \hat{O} \), for the backward flow \( F_{i+t \to i} \) can be formulated as:

\[ \hat{O}_i(y_i, x_i) = 0 \] if there is a corresponding pixel \( (y_{i+t}, x_{i+t}) \) in \( I_{i+t} \) \((0 \leq x_{i+t} < w \& 0 \leq y_{i+t} < h)\), otherwise \( \hat{O}_i(y_i, x_i) = 1 \). Then cross entropy with a penalty is used for occlusion estimation. The network mimics \( \hat{O} \), and produces finer masks by our loss function \( L_{occ} \):

\[
L_{occ} = - \sum_{x,y} \log p(O_{est}^{xy} = \hat{O}^{xy}) - \alpha e^{-O_{est}^{xy}}.
\]

(4)

Since we do not calculate the consistency loss of the oc-
clusion region, the network tends to predict more occlusion

regions. So the second penalty term is used to prevent exces-

sive occlusion prediction. The larger \( \alpha \) is, the greater penalty for the occlusion region, and the smaller the occlusion region predicted. We tried different \( \alpha \) values between 0 and 1, and found that 0.2 is the best.

Optical Flow Estimation

Similar to (Yin and Shi 2018; Jason, Harley, and Derpansis 2016; Wang et al. 2018), optical flow can be learned in a self-supervised manner. More specifically, the first frame can be warped to the next frame by the predicted optical flow, and the photometric consistency and motion smooth-

ness are exploited for training. Photometric consistency is to reconstruct the scene structure between two frames and motion smoothness is to filter out erroneous predictions and preserve sharp details. In this work, we observe that semantic information can be leveraged by joint training to help estimation of optical flow.

As shown in Figure 2, the semantic maps \( M \) introduce semantic information on the likely physical motion of the associated pixels. Besides, we generate error masks which point out the inaccurate regions of the optical flow for robust optical flow estimation. As illustrated in Figure 3, we first calculate an inconsistent mask \( O_{seg} = (M \neq M') \) between our two branches, where \( M' \) is the warped segmentation prediction with bilinear interpolation. Then we define the error mask \( O_{error} \) as:

\[
O_{error} = \max(O_{seg} - O_{est}, 0).
\]

(5)

The inconsistent mask of two segmentation maps should contain the occlusion mask and the offset due to inaccurate optical flow. To unify these two masks, we simply double the weight of the error mask region and ignore the occlusion mask region during optical flow learning. Our photometric
loss $L_{pm}$ can be calculated with the following equation:

$$L_{pm} = \sum_{x,y} (G(I, I')_{xy} \cdot (1 + O_{\text{error}}^{xy} - O_{\text{est}}^{xy})), $$

$$G(I, I') = \beta \frac{1 - SSIM(I, I')} {2} + (1 - \beta)\|I - I'\|_1,$$

where $I'$ is a warped image, $SSIM$ is the per pixel structural similarity index measurement (Wang et al. 2004), $G$ denotes the loss map, which indicates the weight to penalize at different locations. Here we adopt a linear combination of two common metrics for estimating similarity of the original image and the warped one. Intuitively, the pixels perfectly matched indicate the estimated flow is correct and get less penalized in the photometric loss. $\beta$ is taken to be 0.85 as in (Yin and Shi 2018). Following (Jason, Harley, and Derpanis 2016; Yin and Shi 2018), the smoothness loss is defined as:

$$L_{sm} = \sum_{x,y} |\Delta F(x, y)| \cdot (e^{-|\Delta I(x, y)|}),$$

where $\Delta$ is the vector differential operator. Note that both the photometric and smoothness losses are calculated on multiscale blocks and two directions.

**Joint Learning**

For the frames that have ground truths $M_{gt}$, we use the standard log-likelihood loss for semantic segmentation:

$$L_{seg} = -\sum_{x,y} \log p(M_{gt}^{xy} = M_{est}^{xy}).$$

To summarize, our final loss for the entire framework is:

$$L = L_{seg} + \lambda_{cons} L_{cons} + \lambda_{occ} L_{occ} + \lambda_{sm} L_{sm} + L_{pm},$$

where $\lambda_{cons}$, $\lambda_{occ}$, and $\lambda_{sm}$ denote the weights for multiple losses. Our entire framework is thus trained end-to-end.

**Experiments**

**Dataset and Setting**

**Datasets** We evaluate our framework for video semantic segmentation on the Cityscapes (Cords et al. 2016) and CamVid datasets (Brostow, Fauqueur, and Cipolla 2009). We also report our competitive results for optical flow estimation on the KITTI dataset (Geiger, Lenz, and Urtasun 2012).

Cityscapes (Cords et al. 2016) contains 5,000 sparsely labeled snippets collected from 50 cities in different seasons, which are divided into sets with numbers 2,975, 500, and 1,525 for training, validation and testing. Each snippet contains 30 frames, and only the 20th frame is finely annotated in pixel-level. 20,000 coarsely annotated images are also provided.

CamVid (Brostow, Fauqueur, and Cipolla 2009) is the first collection of videos with object class semantic labels, it contains 701 color images with annotations of 11 semantic classes. We follow the same split in (Kundu, Vineet, and Kolten 2016; Nilsson and Sminchisescu 2018) with 367 training images, 100 validation images and 233 test images.

| Table 1: Ablation study for video semantic segmentation on the Cityscapes validation set. ResNet50-based PSPNet (single scale testing) is used as a baseline model. |
|-------------------------|-----------------|
| Model                 | IoU (%)         |
| ResNet50 + PSPNet     | 76.20           |
| + TCC$_{fix}$         | 77.02           |
| + TCC$_{single}$      | 77.58           |
| + TCC$_{single} + OM$ | 77.79           |
| + TCC$_{multi} + OM$  | 78.07           |
| + TCC$_{multi} + OM + UD$ | **78.44**       |

KITTI (Geiger, Lenz, and Urtasun 2012) is a real-world computer vision benchmark dataset with multiple tasks. The training data we use here is similar to (Yin and Shi 2018), where the official training images are adopted as testing set. All the related images in the 28 scenes covered by testing data are excluded. Since there are no segmentation labels on our training set, we generate some coarse segmentation results as the segmentation ground truths through a model trained on Cityscapes.

**Evaluation Metrics** We report mean Intersection-over-Union (mIoU) scores for semantic segmentation task on Cityscapes and CamVid datasets. The optical flow performance for the KITTI dataset is measured by the average end-point-error (EPE) score.

**Implementation Details** Our framework is not limited to specific CNN architectures. In our experiments, we use the original PSPNet (Zhao et al. 2017) and the modified FlowNetS (Dosovitskiy et al. 2015) as the baseline network unless otherwise specified. The FlowNetS is modified as follows: (1) share the encoder with PSPNet. (2) add two $3 \times 3$ convolution layers for occlusion estimation with 32 and 1 channels, respectively. The loss weights are set to be $\lambda_{cons} = 10, \lambda_{occ} = 0.4$ and $\lambda_{sm} = 0.5$ for all experiments.

During training, we randomly choose ten pairs of images with $\Delta t \in [1, 5]$ from one snippet, five of which contain images with ground truths. The training images are randomly cropped to $713 \times 713$. We also perform random scaling, rotation, flip and other color augmentations for data augmentation. The network is optimized by SGD, where momentum and weight decay are set to 0.9 and 0.0001 respectively. We take a mini-batch size of 16 on 16 TITAN Xp GPUs with synchronous Batch Normalization. We use the ‘poly’ learning rate policy and set base learning rate to 0.01 and power to 0.9, as in (Zhao et al. 2017). The iteration number for training process is set to 120K.

**Ablation Study**

To further evaluate the effectiveness of the proposed components, i.e., the joint learning, the temporally consistent constraint, the occlusion masks, and the unlabeled data, we conduct ablation studies on both the segmentation and optical flow tasks. All experiments use the same training setting.

For video segmentation, we make comparisons to five simplified versions on the Cityscapes validation set: (1) TCC$_{fix}$ – temporally consistent constraint on a single pair of
Table 2: Ablation study for optical flow estimation on the KITTI Dataset.

| Method           | Noc   | All   |
|------------------|-------|-------|
| UL               | 7.53  | 11.03 |
| UL + OE          | 7.23  | 8.72  |
| UL + TC          | 4.94  | 8.84  |
| UL + OE + TC     | 4.51  | 7.79  |
| EFC_full         | 3.93  | 7.05  |

Table 3: Comparative results of video segmentation on the Cityscapes test set. Notation: ‘PSP’ – the PSPNet trained with only finely annotated data, ‘PSP_CRS’ – the PSPNet trained with both finely and coarsely annotated data, ‘C’ – whether coarsely annotated data is used, ‘IoU cls’ – average class IoU (%), ‘IoU cat’ – average category IoU (%).

| Method            | C     | IoU cls | IoU cat |
|-------------------|-------|---------|---------|
| Clockwork (2016)  |       | 66.4    | 88.6    |
| PEARL (Jin et al. 2017) |       | 75.4    | 89.2    |
| LLVSS (Li, Shi, and Lin 2018) |       | 76.8    | 89.8    |
| Accel (2019)      |       | 75.5    | –       |
| DFANet (Li et al. 2019) |       | 71.3    | –       |
| Dilation10 (2015) |       | 67.1    | 86.5    |
| Dilation10 + GRFP (2018) |       | 67.8    | 86.7    |
| Dilation10 + EFC (Ours) |       | 68.7    | 87.3    |
| PSP (Zhao et al. 2017) |       | 78.4    | 90.6    |
| PSP + EFC (Ours)  |       | 80.2    | 90.9    |
| PSP_CRS (Zhao et al. 2017) | ✓    | 80.2    | 90.6    |
| PSP_CRS + NetWarp (2017) | ✓    | 80.5    | 91.0    |
| PSP_CRS + GRFP (2018) | ✓    | 80.6    | 90.8    |
| PSP_CRS + EFC (Ours) | ✓    | 81.0    | 91.2    |
| DeepLabv3+ (Chen et al. 2018) | ✓    | 82.1    | 92.0    |
| + EFC (Ours)      | ✓     | 82.7    | 92.1    |
| + VPLR (2019)     | ✓     | 83.5    | 92.2    |

images with the fixed pre-trained FlowNetS. (2) TCC \textsubscript{single} – temporally consistent constraint without the occlusion mask on a single pair of images. (3) TCC \textsubscript{single} + OM – temporally consistent constraint with the occlusion mask on a single pair of images. (4) TCC \textsubscript{multi} + OM – temporally consistent constraint with the occlusion mask on randomly selected five pairs of images. (5) TCC \textsubscript{multi} + OM + UD – our full EFC model with unlabeled data.

The ablation study results for segmentation are presented in Table 1. It can be seen that: (1) The performance continuously increases when more components are used for video segmentation, showing the contribution of each part. (2) Compared with the fixed FlowNetS, joint learning with the optical flow benefits the video segmentation, which shows the close relationship between these two tasks. (3) The temporally consistent constraint has made huge improvements (a percentage of 1.3) to video segmentation, even without the use of occlusion mask. (4) The improvements achieved by occlusion mask show that modeling of occlusion regions benefits the video segmentation. (5) Both the use of more labeled data pairs and unlabeled data clearly lead to performance improvements, which provides evidence that our EFC model takes full advantage of video information.

For optical flow estimation, we make comparisons to five versions of our model: (1) UL – unsupervised learning of only the flow branch with the smooth loss and the photometric loss; (2) UL + OE – adding occlusion estimation ($O_{est}$) without the occlusion mask $O_{seg}$; (3) UL + TC – adding the segmentation branch and the temporal consistency module; (4) UL + OE + TC – our model without the occlusion handling module; (5) EFC\textsubscript{full} – our full model.

From Table 2, we can observe that: (1) Our model can learn in an unsupervised manner using only the optical flow branch. (2) The segmentation branch and temporal consistent constraints greatly facilitate the learning of optical flow. (3) A better occlusion estimation can further improve the performance of optical flow estimation.

**Comparative Results**

**Video Semantic Segmentation** We compare our video semantic segmentation model to the state-of-the-art alternatives on the challenging Cityscapes and CamVid datasets.

**Cityscapes** To validate the robustness of the proposed method on different network architectures, we used Dilation10 (Yu and Koltun 2015), PSPNet (Zhao et al. 2017) and DeepLabv3+ (Chen et al. 2018) as backbone network for the segmentation branch, respectively. In Table 3 we show the quantitative comparison with a number of state-of-the-art video segmentation models.

We observe that: (1) With DeepLabv3+, PSPNet and Dilation10 as our backbones, our model are able to improve the mIoU score by 0.6, 1.8 and 2.1 respectively. Notice that our approach can be applied to any image semantic segmentation model for more accurate semantic segmentation. (2) VPLR (Zhu et al. 2019) first pre-trained on the Mapillary dataset, which contains 18,000 street-level scenes anno-
Table 4: Comparative results on the test set of CamVid for different video segmentation methods. All the methods are based on Dilation8 Network. Our model performs best and improve the mIoU score by 2.1 percentage.

| Method                  | mIoU (%) |
|-------------------------|----------|
| Dilation8 (Yu and Koltun 2015) | 65.3     |
| + STFCN (2016)          | 65.9     |
| + GRFP (2018)           | 66.1     |
| + FSO (2016)            | 66.1     |
| + VPN (2017)            | 66.7     |
| + NetWarp (2017)        | 67.1     |
| + EFC (ours)            | 67.4     |

Table 5: Average end-point error (EPE) on KITTI 2015 flow training set over non-occluded regions (Noc) and overall regions (All). Notation: ‘C’ – the FlyingChairs dataset, ‘S’ – the Sintel dataset, ‘T’ – the FlyingThings3D dataset, ‘K’ – the KITTI dataset, ‘R’ – the RoamingImages dataset.

| Method                | Data      | Noc   | All   |
|-----------------------|-----------|-------|-------|
| EpicFlow (2015)       | -         | 4.45  | 9.57  |
| FlowNetS (2015)       | C+S       | 8.12  | 14.19 |
| FlowNet2 (2017)       | C+T       | 4.93  | 10.06 |
| FlowNet2+ft (2017)    | C+T+K     | -     | 2.3   |
| PW-C-Net (2018)       | C+T       | -     | 10.35 |
| PW-C-Net+ft (2018)    | C+T+K     | -     | 2.16  |
| DSTFlow (2017a)       | K         | 6.96  | 16.79 |
| GeoNet (2018)         | K         | 8.05  | 10.81 |
| OAULFlow (2018)       | K         | -     | 8.88  |
| Unflow (2018)         | K         | -     | 8.80  |
| SC (2019)             | K         | 4.30  | 7.13  |
| EFC (ours)            | K         | 3.93  | 7.05  |
| Back2Future (2018)    | R + K     | 3.22  | 6.59  |
| SelFlow (Liu et al. 2019) | S + K | -     | 4.84  |

CamVid We evaluate our method on the CamVid dataset and compare it with multiple video semantic segmentation methods. The comparative results are given in Table 4. Our model achieves the best result under the same setting.

Optical Flow Estimation To quantify how optical flow estimation benefits from the semantic segmentation, we evaluate the estimated flow on the KITTI dataset. Both supervised and unsupervised methods are included. As shown in Table 5, our model not only outperforms the existing un-supervised learning methods, but also yields comparable results with the Flownet2 (Ilg et al. 2017) which is trained on FlyingChairs and FlyingThings3D datasets. Following the common practice in (Ren et al. 2017a; Yin and Shi 2018; Wang et al. 2018; Meister, Hur, and Roth 2018; Lai, Tsai, and Chiu 2019), we use no additional data and discard the whole sequence as long as it contains any test frames, while (Janai et al. 2018; Liu et al. 2019) use the RoamingImages dataset and the Sintel dataset for pre-training, respectively. Besides, they use PWC-Net (Sun et al. 2018) as the base model, which is powerful than FlowNetS.

The semantic segmentation brings semantic information to the optical flow estimation, which facilitates recovering sharp motion boundaries in the estimated flow. As shown in Figure 5, our model fixes large regions of errors compared to (Yin and Shi 2018).

Conclusion

In this paper, we propose a novel framework (EFC) for joint estimation of video semantic segmentation and optical flow. We observe that semantic segmentation introduces semantic information and helps model occlusion for more robust optical flow estimation. Meanwhile, non-occluded optical flow provides accurate pixel-level temporal correspondences to guarantee the temporal consistency of the segmentation. Moreover, we address the insufficient data utilization and the inefficiency issues through our framework. Extensive experiments have shown that our approach outperforms the state-of-the-art alternatives under the same settings in both tasks.

Acknowledgements

Zhiwu Lu is partially supported by National Natural Science Foundation of China (61976220, 61832017, and 61573363), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098). Ping Luo is partially supported by the HKU Seed Funding for Basic Research and SenseTime’s Donation for Basic Research.
