Hindawi
Complexity
Volume 2021, Article ID 6680509, 10 pages
https://doi.org/10.1155/2021/6680509

Research Article

Dynamic Warping Network for Semantic Video Segmentation

Jiangyun Li,[1,2] **Yikai Zhao**,[1] **Xingjian He**,[3,4] **Xinxin Zhu**,[3,4] **and Jing Liu**[4]

[1]*School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China*
[2]*Shunde Graduate School of University of Science and Technology Beijing, Foshan 528300, China*
[3]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100083, China*
[4]*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100083, China*

Correspondence should be addressed to Jiangyun Li; leejy@ustb.edu.cn

Received 6 December 2020; Revised 3 January 2021; Accepted 24 January 2021; Published 8 February 2021

Academic Editor: Ning Cai

A major challenge for semantic video segmentation is how to exploit the spatiotemporal information and produce consistent results for a video sequence. Many previous works utilize the precomputed optical flow to warp the feature maps across adjacent frames. However, the imprecise optical flow and the warping operation without any learnable parameters may not achieve accurate feature warping and only bring a slight improvement. In this paper, we propose a novel framework named Dynamic Warping Network (DWNet) to adaptively warp the interframe features for improving the accuracy of warping-based models. Firstly, we design a flow refinement module (FRM) to optimize the precomputed optical flow. Then, we propose a flow-guided convolution (FG-Conv) to achieve the adaptive feature warping based on the refined optical flow. Furthermore, we introduce the temporal consistency loss including the feature consistency loss and prediction consistency loss to explicitly supervise the warped features instead of simple feature propagation and fusion, which guarantees the temporal consistency of video segmentation. Note that our DWNet adopts extra constraints to improve the temporal consistency in the training phase, while no additional calculation and postprocessing are required during inference. Extensive experiments show that our DWNet can achieve consistent improvement over various strong baselines and achieves state-of-the-art accuracy on the Cityscapes and CamVid benchmark datasets.

1. Introduction

Semantic segmentation aims to assign a specific semantic label to each pixel for a given image. In recent years, the models based on deep learning [1–5] have brought the performance of the task to a new level. However, most existing methods are only designed for parsing images and may produce inconsistent results to video frames, due to lack of temporal information.

To address the problem, many methods tend to incorporate temporal information of the video to improve the accuracy of video segmentation. And optical flow, which encodes the temporal consistency across frames in the video, has been widely used for semantic video segmentation. Gaddel et al. [6] propose to combine the features wrapped from previous frames with optical flow and those from the current frame to enhance the features. Studies of [7–9] use feature warping for acceleration.

However, there are two main problems with existing warping-based methods. Firstly, the optical flow obtained by the traditional algorithms or optical flow estimation networks [10–12] cannot accurately estimate the motion of all pixels across adjacent frames. Second, the warping operation adopted by previous methods [6, 7, 13] is implemented with standard bilinear interpolation and does not contain any learnable parameters. Therefore, warping features relying on the imprecise optical flow may result in feature misalignment between the warped features and expected ones. TWNet [9] introduces a correction stage after warping to refine the warped features. However, the method has some limitations, because it needs motion vectors and residuals in the compressed video according to a specific compression standard.

In this paper, we propose a novel framework named Dynamic Warping Network (DWNet) to adaptively warp the interframe features for improving the accuracy of

warping-based models. First, we design a flow refinement module (FRM) to optimize the precomputed optical flow and produce more accurate pixel displacement for every pixel location. Besides, we propose a flow-guided convolution (FG-Conv) to achieve the adaptive feature alignment based on the refined optical flow instead of the original warping operation. Furthermore, we introduce the temporal consistency loss including the feature consistency loss and prediction consistency loss to explicitly supervise the warped features and guarantee the temporal consistency of video segmentation, as shown in Figure 1. Our DWNet adopts extra constraints to improve the temporal consistency instead of simple feature fusion and feature propagation [6, 7], which makes the network explicitly model the temporal consistency of the video in the training phase. And, in the inference phase, the optical flow network, the flow refinement module, and the flow-guided convolution can be removed. Hence, the final network can be regarded as a semantic image segmentation network with no post-processing during inference.

We evaluate our DWNet on two semantic video segmentation benchmarks: Cityscapes and CamVid. Extensive experiments show that our DWNet can significantly outperform existing warping-based methods and achieve state-of-the-art accuracy on the two benchmark datasets. In particular, our DWNet can achieve consistent improvement over various strong baselines, which demonstrates the generalization ability of our method.

To conclude, our main contributions are five-fold:

(i) We propose a novel framework named Dynamic Warping Network (DWNet) to adaptively warp the interframe features

(ii) We design a flow refinement module (FRM) to optimize the optical flow and propose a flow-guided convolution (FG-Conv) to adaptively align features across adjacent frames according to the refined optical flow

(iii) We explicitly model the temporal consistency of the video and introduce the temporal consistency loss to supervise the warped features

(iv) Our DWNet needs no additional parameters and calculation during inference because the optical flow network, the flow refinement module, and the flow-guided convolution can be removed in the inference phase

(v) The experimental results demonstrate that our DWNet can outperform previous warping-based methods and achieve state-of-the-art accuracy on the Cityscapes and CamVid datasets

2. Related Work

2.1. Semantic Video Segmentation. Semantic video segmentation aims to carry out dense labeling for all pixels in each frame of a video sequence. Compared with semantic image segmentation, semantic video segmentation needs to focus more on the temporal consistency of consecutive frames and produces a more consistent interframe prediction. Therefore, many works tend to incorporate temporal information of the video to improve the video segmentation accuracy, including optical flow-based feature warping [6, 8, 9, 13–17], propagation-based [18, 19], LSTM-based [15, 20], 3D CNN-based method [21], and the weakly supervised method [22]. And optical flow, which encodes the temporal consistency across frames in the video, has been most widely used for semantic video segmentation. The optical flow-based methods first compute the optical flow between the current frame and the previous frame and then enhance the features of the current frame by warping the features of the previous frame or utilize the warped features from the keyframe as the features of the current frame for acceleration. Despite its relative strength, the optical flow-based feature warping contains two main problems as discussed above. TWNet [9] and DMNet [23] propose to correct the warped features by utilizing the postprocessing, which only brings a slight improvement. To our best knowledge, we are the first to directly optimize the warping operation and propose the learnable dynamic warping operation instead of the original one.

2.2. Dynamic Convolution. The study [24] proposes dynamic filters or kernels to generate context-aware filters which are adaptive to the input and are predicted by the network. Many works [25, 26] have adopted the predicted dynamic filters to obtain better feature representations. Deformable convolution [27, 28] utilizes the input features to generate different offsets and weights for each sample position. Motivated by deformable convolution, we observe that the optical flow can be regarded as the offset and we can utilize the offset to adaptively align interframe features. Different from the deformable convolution whose offsets are generated by the input features, we utilize the flow refinement module to optimize the optical flow and obtain more accurate pixel displacement. Furthermore, we propose a flow-guide convolution to dynamically warp the features based on the refined optical flow and achieve better feature warping.

3. Methods

In the section, we first give an overview of our DWNet framework and then describe each of its components in detail. Finally, we describe how to optimize the whole network for improving semantic video segmentation.

3.1. Overview. The overall structure of our DWNet framework is illustrated in Figure 2. The inputs of our DWNet are a pair of RGB images I_t and I_{t+k}, where I_t represents the labeled frame and I_{t+k} represents the unlabeled frame randomly selected from the near-by frames of I_t with $k \in [-5, 5]$. The two images are first sent to the shared segmentation network to extract the semantic features F_t and F_{t+k}. Meanwhile, the two images are also sent to the optical flow estimation network to predict the coarse optical flow $O_{t+k \longrightarrow t}$. Then, we utilize the flow refinement module to optimize the optical flow and produce more accurate optical

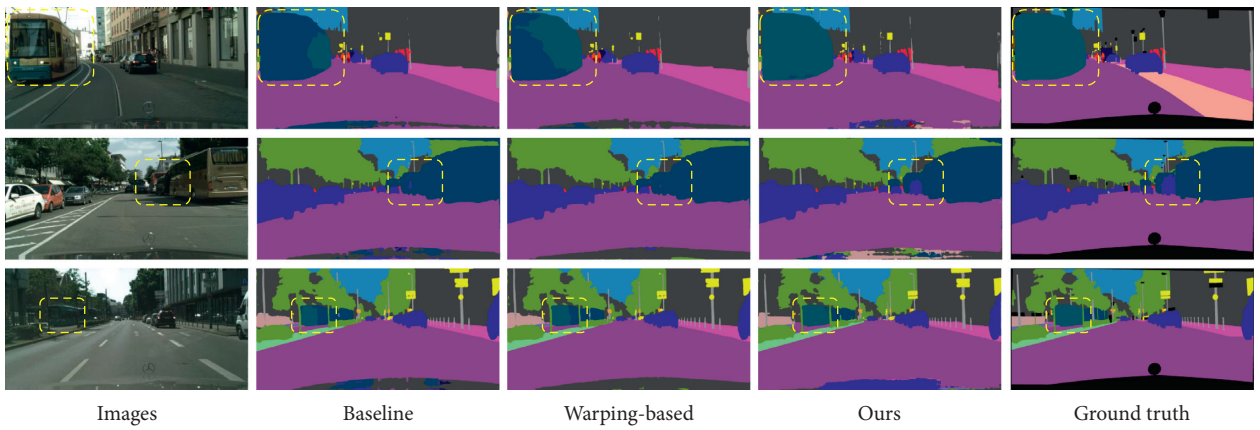

Figure 1: Qualitative results from the Cityscapes dataset. Baseline method: training the model on single frames and inferring the segmentation maps on single frames. Warping-based method: adopting the original warping operation implemented with standard bilinear interpolation to propagate and fuse the features brings a slight improvement. Our method: utilizing the flow-guided convolution to adaptively warp the interframe features and introducing temporal consistency loss to explicitly supervise the warped features instead of simple feature propagation and fusion, hence producing more accurate segmentation results.

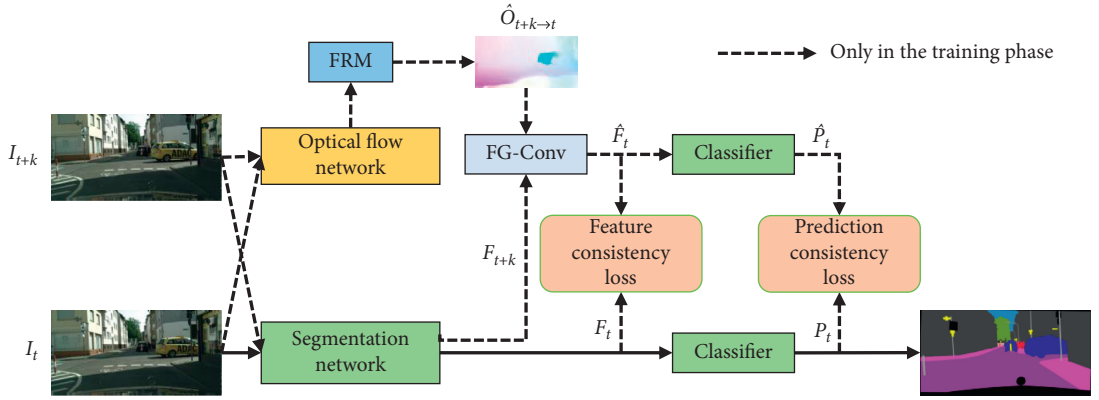

Figure 2: The overall structure of our DWNet framework. FRM denotes the flow refinement module. FG-Conv denotes the flow-guided convolution. Feature consistency loss and prediction consistency loss are both the temporal consistency loss, which improves the temporal consistency of video segmentation. The dotted lines denote that the components are only used in the training phase and will be removed in the inference phase.

flow $\widehat{O}_{t+k\longrightarrow t}$ for every pixel position. After that, we adopt the flow-guided convolution to dynamically warp F_{t+k} to $\widehat{F}_t$ according to the refined optical flow $\widehat{O}_{t+k\longrightarrow t}$. Finally, F_t and $\widehat{F}_t$ are sent to the shared classifier to produce the segmentation map P_t and $\widehat{P}_t$ respectively, and we introduce two kinds of temporal consistency losses as extra constraints to supervise the warped features $\widehat{F}_t$ and $\widehat{P}_t$, respectively. In the following, we will introduce each key component of our DWNet in detail.

3.2. Flow Refinement Module. We first utilize the existing optical flow estimation network to obtain the optical flow $O_{t+k\longrightarrow t}$. The optical flow network computes the pixel displacement $(\Delta x, \Delta y)$ for every pixel location (x, y) in I_t to the spatial locations (x', y') in I_{t+k}, which means that $(x', y') = (x + \Delta x, y + \Delta y)$. And Δx and Δy are floating point numbers and denote pixel displacements in horizontal and vertical directions, respectively [6]. However, the optical flow estimated by the optical flow network may not be enough accurate due to occlusion and new objects. Therefore, we propose the flow refinement module to optimize the coarse optical flow. We concatenate the two input images, the difference of the two images, and the coarse optical flow, resulting in an 11 channel tensor as the input to the flow refinement module. The flow refinement module consists of 4 convolution layers. The first 3 layers are made up of 3×3 kernels with stride 2 following BatchNorm and ReLU, and the number of the output channels is set to 64, 128, and 256, respectively. The output of the third layer is then passed on to the last 3×3 convolution layer with $2s^2$ output channels to attain the refined optical flow $\widehat{O}_{t+k\longrightarrow t}$, whose spatial size is corresponding to the feature F_t and F_{t+k}. s represents the kernel size of the flow-guided convolution which will be discussed in Section 3.3 and is set to 1 as default. We visualize the original optical flow and the refined optical flow, as shown in Figure 3. The refined optical flow has sharper motion boundaries for moving objects and semantics, such as humans and cars, which demonstrates the effectiveness of the flow refinement module. Next, we will introduce how to

Figure 3: Visual comparison on the Cityscapes dataset for the original optical flow. The first column denotes the input frame. The middle column denotes the coarse optical flow produced by the optical flow network. The last column denotes the refined optical flow optimized by the flow refinement module. The refined optical flow has sharper motion boundaries than the original optical flow.

use the refined optical flow to achieve better features warping.

3.3. Flow-Guided Convolution. The flow refinement module utilizes the original optical flow and images to produce more precise optical flow estimation. Given the optical flow, previous methods utilize the warping operation to transform the feature F_{t+k} to the feature of the current frame $\widehat{F}_t$:

$$\widehat{F}_t = \text{warp}\left(F_{t+k}, O_{t+k\longrightarrow t}\right). \tag{1}$$

However, it cannot accurately align the warped feature and the feature of the current frame due to the imprecise optical flow and the original warping operation without any learnable parameters. Hence, we firstly utilize the flow refinement module to optimize the optical flow as discussed in Section 3.2. Besides, we propose the flow-guided convolution to adaptively warp the interframe features. The standard convolution samples the input feature map at fixed locations, and the DCNv1 [27] adds 2D offsets to the regular grid sampling locations to enable free form deformation of the sampling grid. Motivated by this work, we observe that the optical flow which encodes the pixel displacement across frames can be regarded as a specific offset, and we can utilize the optical flow to dynamically warp the interframe features. Formally, the standard 2D convolution can be written as

$$y[i] = \sum_{p}^{P} w[p] \cdot x[i + p], \tag{2}$$

where y denotes the output after the convolution, i denotes the location, x denotes the input features, w denotes the convolution filters with a length of P, and p enumerates P. p is usually the regular sampling locations in a $s \times s$ kernel, and we propose the flow-guided convolution by adding the location offsets into p as follows:

$$y[i] = \sum_{p}^{P} w[p] \cdot x[i + p - \Delta p], \tag{3}$$

where $\Delta p \in \widehat{O}_{t+k\longrightarrow t}$. The refined optical flow is regarded as the offsets for the flow-guided convolution to adaptively sample more corresponding pixel locations between interframe features. The kernel size s is the key parameter for the flow-guided convolution, and we will discuss the parameter in 4.2.2. Compared with the DCNv1 [27], we obtain the

offsets from the flow refinement module instead of applying a convolution layer to the input feature. Hence, we can attain more accurate offsets and achieve better feature warping.

3.4. Temporal Consistency Loss. The flow-guided convolution can dynamically warp the feature F_{t+k} and produce the estimated feature $\widehat{F}_t$ of the current frame. Previous methods concatenate or do the weighted sum of the warped feature $\widehat{F}_t$ and the original feature F_t to achieve feature fusion and propagation. However, we argue that the warped feature $\widehat{F}_t$ is expected to be consistent with the original feature F_t, and the two features should be the same ideally. Hence, we propose the temporal consistency loss to explicitly supervise the feature $\widehat{F}_t$ and the segmentation map $\widehat{P}_t$ respectively. Compared with the previous methods using feature fusion or fusion propagation, we utilize extra constraints to improve the temporal consistency of video segmentation, which is more reasonable and does not introduce additional calculation or postprocessing in the inference phase. The temporal consistency loss contains the feature consistency loss and the prediction consistency loss, which are related to the feature $\widehat{F}_t$ and the segmentation map $\widehat{P}_t$, respectively.

3.4.1. Feature Consistency Loss. We attempt to constraint both features of F_t and $\widehat{F}_t$ to be similar enough by designing a feature consistency loss. Instead of per-pixel similarity calculation, we measure the similarity between the self-attention maps A_t and $\widehat{A}_t$ of both features. Since the self-attention maps present high-order relationships among pixels, such a similarity measurement is more robust than the typical per-pixel one. Let $a_{i,j}$ denote the similarity between the ith pixel and the jth pixel of the original feature F_t, and let $\widehat{a}_{i,j}$ denote the similarity between the ith pixel and the jth pixel of the warped feature $\widehat{F}_t$, where $a_{i,j} \in A_t$ and $\widehat{a}_{i,j} \in \widehat{A}_t$. The $a_{i,j}$ is computed from the feature $F_{t,i}$ and $F_{t,j}$ as

$$a_{i,j} = \frac{F_{t,i}^T F_{t,j}}{\left(\left\|F_{t,i}\right\|_2 \left\|F_{t,j}\right\|_2\right)}. \tag{4}$$

And, we adopt the squared difference to formulate the feature consistency loss:

$$\ell_{\text{fc}}\left(F_t, \widehat{F}_t\right) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(a_{ij} - \widehat{a}_{ij}\right)^2, \tag{5}$$

where N denotes the total number of the pixels. The warped feature and the original feature should produce a similar attention map that encodes the pixel correlations. Hence, this loss can strengthen the feature consistency by explicitly supervising the attention maps.

3.4.2. Prediction Consistency Loss. The segmentation map $\widehat{P}_t$ produced by the feature $\widehat{F}_t$ should be also consistent with the segmentation map P_t of the current frame. Hence, we introduce the prediction consistency loss [17] to improve the temporal consistency of video segmentation as follows:

$$\ell_{\text{pc}}\left(I_t, I_{t+k}\right) = \frac{1}{N} \sum_{i=1}^{N} M_{t+k\longrightarrow t,i} \left\|P_{t,i} - \widehat{P}_{t+k\longrightarrow t,i}\right\|_2^2. \tag{6}$$

Due to the occlusion and new objects across frames, we predict a mask $M_{t+k\longrightarrow t}$ to assign different weights to each pixel according to the warping error $E_{t+k\longrightarrow t}$, where $E_{t+k\longrightarrow t} = |I_t - \widehat{I}_t|$ and $\widehat{I}_t$ denotes the warped input frame from I_{t+k}. Then, $M_{t+k\longrightarrow t}$ is denoted as

$$M_{t+k\longrightarrow t} = \exp\left(-\frac{E_{t+k\longrightarrow t}}{\delta}\right), \tag{7}$$

where δ is a hyperparameter which controls the amplitude of the difference between high error and low error. The pixels with higher warping errors are assigned to lower weights and vice versa, because higher warping error represents that the optical flow and the warped feature are more inaccurate. $M_{t+k\longrightarrow t}$ can speed up the convergence of the prediction consistency loss and improve the accuracy of video segmentation by considering the pixels with more precise optical flow and ignoring the noise produced by occlusion and new objects.

3.5. Optimization. The loss of our DWNet consists of the conventional cross-entropy loss ℓ_{ce} and the temporal consistency loss including the feature consistency loss ℓ_{fc} and the prediction consistency loss ℓ_{pc}. Hence, our final objective function is

$$\ell = \ell_{\text{ce}} + \lambda_1 \ell_{\text{fc}} + \lambda_2 \ell_{\text{pc}}, \tag{8}$$

where λ_1 and λ_2 denote the weights for multiple losses. As illustrated in Figure 2, our DWNet can be trained in an end-to-end fashion. And in the inference phase, the optical flow network, the flow refinement module, and the flow-guided convolution in the dotted line can be removed. Hence, the final network can be regarded as a semantic image segmentation network with no additional calculation or postprocessing during inference.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets. We evaluate our proposed DWNet on two semantic video segmentation benchmarks datasets Cityscapes [29] and CamVid [30].

Cityscapes is an urban scene dataset and contains 5000 video snippets collected from 50 cities in different seasons. Each snippet contains 30 frames and only the 20th frame is pixel-level finely annotated, leading to the dataset containing 5000 images which are divided into 2975, 500, and 1525 images for training, validation, and testing respectively. Besides, the dataset also contains 20000 coarsely annotated images, but we do not utilize these data in all experiments except otherwise stated.

CamVid is composed of 701 densely annotated images from five video sequences. The images are labeled every 30 frames with 11 semantic classes. Following the previous

work [6], the dataset is split into 367 training, 101 validation, and 233 testing images.

4.1.2. Models. To validate the effectiveness of our proposed method, we conduct extensive experiments with different network configurations. We adopt the ResNet50 [31], ResNet101 [31], and MobileNetv2 [32] as the backbone to extract the high-level features. And we choose the PSPNet [33], DeeplabV3+ [3], and DANet [5] as the segmentation model. The segmentation network is combined with different backbones and segmentation models. We conduct the ablation experiments on ResNet50 with the structure of PSPNet, namely, PSPNet50. Because the optical flow network can be removed in the inference phase, we adopt the more powerful optical flow estimation network FlowNetV2 [11] to extract the more accurate optical flow, though it is slower and with more parameters during training compared with the lightweight FlowNet, like [10, 12].

4.1.3. Implementation Details. We implement our method based on PyTorch. We employ an SGD optimizer and a poly learning rate policy, where the initial learning rate is multiplied by $(1 - (\text{epoch}/\text{max\_epoch}))^{\text{power}}$ with power = 0.9 after each iteration. The base learning rate is set to 0.01 for both datasets. Momentum and weight decay are set to 0.9 and 0.0001, respectively. We utilize the synchronized batch normalization [4] with a batch size of 8 for both datasets. For data augmentation, we apply random scaling of the input images (from 0.5 to 2.2 on Cityscapes, from 0.5 to 2.0 on CamVid), random cropping (768×768 for Cityscapes, 384×384 for CamVid), and random left-right flipping during training. Note that the optical flow network FlowNetV2 is also joint optimized with the base learning rate 0.00001. We employ the standard pixel-wise cross-entropy loss function as the main loss to train the whole network with 8 cards of NVIDIA TITAN RTX. The loss weights are set to be $\lambda_1 = 10$ and $\lambda_2 = 0.1$ for all experiments. After training, we utilize the original images to inference unless otherwise stated. Following the previous works [6, 8], we apply mean intersection-over-union (mIoU) as the evaluation metric to validate our method.

4.2. Ablation Study. We build the DWNet based on the single-frame segmentation model. And, we adopt the PSPNet50 as the single-frame model to conduct all the ablation experiments on the Cityscapes dataset.

4.2.1. Effectiveness of the Proposed Method. In this section, we evaluate the different components of our DWNet with different settings, and the results are shown in Table 1. The baseline model is the PSPNet50 with single-frame training and inference. When we utilize the original warping operation and adopt the feature consistency loss as a constraint, the performance is only improved by 0.55%. However, when we replace the original warping operation with our proposed flow-guided convolution, it brings a further improvement by 0.57%, which demonstrates that the dynamic warping is better than the original warping operation. Besides, the flow refinement module and the prediction consistency loss can improve the performance by 0.47% and 0.38%, respectively. And introducing the two components simultaneously can further improve the accuracy to 75.62%. We also verify whether the two components are beneficial to the warping-based method, and the results show that the accuracy can be improved from 74.3% to 74.76%, whose improvement is lower than our proposed method (from 74.87% to 75.62%).

Table 1: Ablation study of our DWNet on the Cityscapes validation set.

Warp	ℓ_{fc}	FG-Conv	FRM	ℓ_{pc}	mIoU %
					73.75
√	√				74.30
	√	√			74.87
	√	√	√		75.34
	√	√		√	75.25
√	√		√	√	74.76
	√	√	√	√	**75.62**

Warp denotes the original warping operation. ℓ_{fc} and ℓ_{pc} denote the feature consistency loss and prediction consistency loss, respectively. FG-Conv denotes the flow-guided convolution. FRM denotes flow refinement module. The bold values denote our method can achieve the best accuracy compared with other methods.

4.2.2. Flow-Guided Convolution. The flow-guided convolution is the core operation of our DWNet, which utilizes the refined optical flow to adaptively warp the interframe features. The kernel size s is the key parameter for the flow-guided convolution. According to the original warping operation, each pixel corresponds to a specific offset, and we can utilize the offset to warp each pixel independently. However, we argue that we can consider more adjacent pixels to judge the warped result of each pixel. Hence, we can adjust s to achieve more precise feature warping. When s is equal to 1, the flow-guided convolution is similar to the original warping operation which treats each pixel independently. However, our flow-guided convolution contains the learnable parameters and can adaptively adjust the warped features. As shown in Table 2, when we set s to 3, the flow-guided convolution yields the best performance. Besides, the flow-guided convolution with different values of s all outperforms the original warping operation, which demonstrates that our proposed method can achieve better feature warping. When s is set to 5, the accuracy gets worse. We think that the larger s may bring more noise and influence the stable training of the whole model.

4.2.3. Prediction Consistency Loss. The prediction consistency loss aims to improve segmentation stability. We calculate the occlusion mask to speed up the convergence and improve the accuracy of video segmentation by considering the pixels with more precise optical flow and ignoring the noise produced by occlusion and new objects. And the δ is a hyperparameter that controls the amplitude of the difference between high error and low error. Hence, we provide a discussion about the δ, and the results are shown in

Table 2: Ablation study of the flow-guided convolution on the Cityscapes validation set.

Method	mIoU %
Baseline	73.75
Baseline + warp	74.30
Baseline + FG-Conv ($s = 1$)	75.02
Baseline + FG-Conv ($s = 3$)	**75.34**
Baseline + FG-Conv ($s = 5$)	75.16

s denotes the kernel size of the flow-guided convolution. The bold values denote our method can achieve the best accuracy when s is set to 3.

Table 3. We first try the prediction consistency loss without the occlusion mask, and we find the performance decrease by 0.22% compared with the baseline, which demonstrates the importance of the occlusion mask. If we treat all pixels equally, the pixels with high warping errors will seriously affect the training and the final segmentation accuracy. And when we introduce the mask and set δ to 2, it can obtain the best performance.

In fact, the first designs for both temporal consistency losses consider the occlusion and new objects. However, the impact on the feature consistency loss is slight (from 74.87% to 74.89%). The occlusion and new objects usually reflect some small and local changes across different frames, and the feature consistency loss aims to model the long-range and high-order relationships and is more robust to such small changes, while the prediction consistency loss aims to model the per-pixel similarity and is susceptible to the occlusion and new objects. Hence, we only add the occlusion mask in the prediction consistency loss.

4.2.4. Feature Fusion and Propagation. To mask the use of the warped features, previous methods try to do weighted sum or concatenate the warped features and the original features for feature fusion and propagation. We compare the previous methods with our proposed method in Table 4. The results show that our proposed method is obviously better than the previous methods, which demonstrates our conjecture to the warped feature reuse.

4.3. Comparative Results on Cityscapes Dataset

4.3.1. Effectiveness of Different Network Structures. To validate the effectiveness of our DWNet, we apply different network configurations. The results are shown in Table 5. SWarp (Static Warping) denotes the original warping operation and DWarp (Dynamic Warping) denotes our proposed DWNet. The results demonstrate that our DWNet has a strong generalization ability for different network structures and can significantly improve the accuracy compared with the SWarp.

4.3.2. Comparison with State-of-the-Art. We compare our DWNet with existing methods on the Cityscapes test dataset. The results are shown in Table 6, and our DWNet can outperform the existing methods with a significant advantage. In particular, with the PSPNet as the backbone, our method with the only fine set for the train can improve the mIoU score by 0.9%, which is superior to previous methods with both fine and coarse sets for the train, like [6, 13, 15]. And when we also utilize both fine and coarse images for the train, our method can bring a further improvement by 0.7%, which demonstrates the effectiveness of our method. Besides, we utilize the DANet as the segmentation network and the accuracy is improved to 82.1%, which shows that our method has a strong generalization for different segmentation networks.

4.3.3. Qualitative Results. The qualitative comparison is shown in Figure 4. Existing warping-based methods adopt the standard bilinear interpolation without any learnable parameters to warp the interframe features based on imprecise precomputed optical flow and produce the negative

Table 3: Ablation study of the prediction consistency loss on the Cityscapes validation set.

Method	mIoU %
Baseline	75.34
Baseline + ℓ_{pc} + w/o mask	75.13
Baseline + ℓ_{pc} + w/mask ($\delta = 1$)	75.46
Baseline + ℓ_{pc} + w/mask ($\delta = 2$)	**75.62**
Baseline + ℓ_{pc} + w/mask ($\delta = 5$)	75.44

δ denotes the amplitude of the difference between high warping error and low warping error. The bold values denote our method can achieve the best accuracy when delta is set to 2.

Table 4: Ablation study of feature fusion and propagation on the Cityscapes validation set.

Method	mIoU %
Baseline	73.75
Baseline + sum	74.30
Baseline + concatenate	74.25
Baseline + TCLoss (ℓ_{fc})	74.87
Baseline + TCLoss ($\ell_{fc} + \ell_{pc}$)	**75.25**

Sum and Concatenate denote the weighted sum and concatenation of the warped features and the original features for feature fusion, respectively. TCLoss denotes the temporal consistency loss, including feature consistency loss and prediction consistency loss. The bold values denote our method can achieve the best accuracy using both the feature consistency loss and prediction consistency loss.

Table 5: Comparison of our DWNet with different network structures on the Cityscapes validation set.

Method	Backbone	SWarp	DWarp	mIoU %
PSPNet	MNV2			72.34
PSPNet	MNV2	√		73.52
PSPNet	MNV2		√	**74.46**
PSPNet	ResNet101			78.90
PSPNet	ResNet101	√		79.32
PSPNet	ResNet101		√	**79.85**
DeeplabV3+	ResNet101			80.15
DeeplabV3+	ResNet101	√		80.32
DeeplabV3+	ResNet101		√	**80.78**
DANet	ResNet101			79.94
DANet	ResNet101	√		80.21
DANet	ResNet101		√	**80.67**

SWarp (Static Warping) denotes the original warping operation. DWarp (Dynamic Warping) denotes our proposed DWNet. MNV2 denotes the MobileNetV2. The bold values denote our method can achieve the higher accuracy than the static warp with different baseline models.

Table 6: Comparison of state-of-the-art semantic video segmentation models on the Cityscapes test set.

Method	Source	mIoU %
Clockwork [34]	ECCV2016	66.4
PEARL [35]	ICCV2017	75.4
LLVSS [18]	CVPR2018	76.8
Accel [8]	CVPR2019	75.5
TDNet [19]	CVPR2020	79.9
ESVS [17]	ECCV2020	76.6
PSPNet [33]	CVPR2017	80.2
PSPNet + NetWarp ‡ [6]	ICCV2017	80.5
PSPNet + GRFP ‡ [15]	CVPR2018	80.6
PSPNet + EFC ‡ [13]	AAAI2020	81.0
PSPNet + ours		**81.1**
PSPNet + ours‡		**81.8**
DANet [5]	CVPR2019	81.5
DANet + ours		**82.1**

Methods trained using both fine and coarse sets are marked with "‡." The bold values denote our method can achieve the best accuracy compared with other state-of-the-art methods.

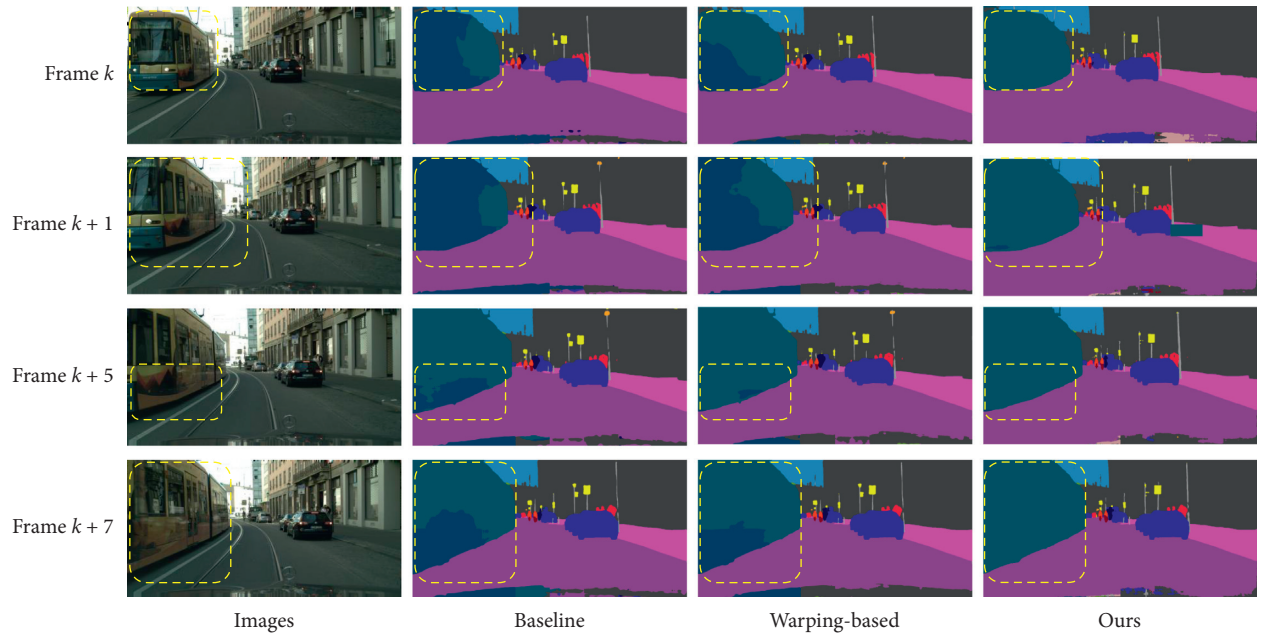

Figure 4: Qualitative results of consecutive frames on the Cityscapes dataset. Baseline methods: training and inferring on single frames. Warping-based method: adopting the original warping operation to enhance the feature. Our method: utilizing the flow-guided convolution to adaptively warp the interframe features. Compared with the baseline, the warping-based method brings a slight improvement in the moving objects, and our method can produce more accurate and consistent segmentation results.

Table 7: Comparison of state-of-the-art semantic video segmentation models on the CamVid test set.

Method	Source	mIoU %
STFCN [20]	arXiv2016	65.9
DFF [7]	CVPR2017	66.0
NetWarp [6]	ICCV2017	70.3
GRFP [15]	CVPR2018	66.1
Accel [8]	CVPR2019	69.3
EFC [13]	AAAI2020	67.4
TDNet [19]	CVPR2020	76.0
ESVS [17]	ECCV2020	76.3
PSPNet [33]	CVPR2017	75.4
PSPNet + ours		**76.5**

The bold values denote our method can achieve the best accuracy compared with other methods on the CamVid test set.

results in the highlighted regions. Compared with the existing warping-based methods, our method adopts the dynamic warping operation to achieve more precise feature alignment based on the refined optical flow and improve temporal consistency of video segmentation.

4.4. Comparative Results on CamVid Dataset. To evaluate the generalization of our method on different datasets, we conduct experiments on the CamVid dataset. We use the ResNet101 as the backbone with the architecture of PSPNet. The results are shown in Table 7, and our method outperforms the current state-of-the-art methods, which demonstrates the generalization for different datasets.

5. Conclusion

In this paper, we propose a novel framework named DWNet to adaptively warp the interframe features. We design the flow refinement module to optimize the optical flow and propose the flow-guide convolution to achieve adaptive feature alignment. Besides, we introduce the temporal consistency loss to explicitly supervise the warped features to guarantee the temporal consistency of video segmentation. Extensive experiments have shown that our method outperforms existing warping-based methods and achieves state-of-the-art on the Cityscapes and CamVid benchmark datasets.

Data Availability

The Cityscapes and CamVid data can be downloaded freely at https://www.cityscapes-dataset.com/file-handling/?packageID=3 and http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Fundamental Research Funds for the China Central Universities of USTB (FRF-DF-19-002), Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (BK20BE014).

References

[1] J. Long, E. Shelhamer, and T. Darrell, *Fully Convolutional Networks for Semantic Segmentation*, CVPR, London, UK, 2015.

[2] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[3] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, *Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation*, ECCV, London, UK, 2018.

[4] H. Zhang, K. Dana, J. Shi et al., *Context Encoding for Semantic Segmentation*, CVPR, London, UK, 2018.

[5] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 22, pp. 3146–3154, 2019.

[6] R. Gadde, V. Jampani, and P. V. Gehler, *Semantic Video Cnns through Representation Warping*, ICCV, London, UK, 2017.

[7] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, *Deep Feature Flow for Video Recognition*, CVPR, London, UK, 2017.

[8] S. Jain, X. Wang, and J. E. Gonzalez, *Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video*, CVPR, London, UK, 2019.

[9] J. Feng, S. Li, Y. Chen, F. Huang, J. Cui, and X. Li, *How to Train Your Dragon: Tamed Warping Network for Semantic Video Segmentation*, 2020.

[10] A. Dosovitskiy, P. Fischer, E. Ilg et al., *Flownet: Learning Optical Flow with Convolutional Networks*, ICCV, London, UK, 2015.

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: evolution of optical flow estimation with deep networks," CVPR, London, UK, 2017.

[12] D. Sun, X. Yang, M. Y. Liu, and J. Kautz, *Pwc-net: Cnns for Optical Flow Using Pyramid, Warping, and Cost Volume*, CVPR, London, UK, 2018.

[13] M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu, and P. Luo, *Every Frame Counts: Joint Learning of Video Segmentation and Optical Flow*, AAAI, London, UK, 2020.

[14] S. Chandra, C. Couprie, and I. Kokkinos, *Deep Spatio-Temporal Random Fields for Efficient Video Segmentation*, CVPR, London, UK, 2018.

[15] D. Nilsson and C. Sminchisescu, *Semantic Video Segmentation by Gated Recurrent Flow Propagation*, CVPR, London, UK, 2018.

[16] Y. S. Xu, T. J. Fu, H. K. Yang, and C. Y. Lee, *Dynamic Video Segmentation Network*, CVPR, London, UK, 2018.

[17] Y. Liu, C. Shen, C. Yu, and J. Wang, *Efficient Semantic Video Segmentation with Per-Frame Inference*, ECCV, London, UK, 2020.

[18] Y. Li, J. Shi, and D. Lin, *Low-Latency Video Semantic Segmentation*, CVPR, London, UK, 2018.

[19] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, *Temporally Distributed Networks for Fast Video Semantic Segmentation*, CVPR, London, UK, 2020.

[20] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, R. Klette, and F. Huang, "STFCN: spatio-temporal FCN for semantic video segmentation," 2016.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Deep End2end Voxel2voxel Prediction*, CVPR, London, UK, 2016.

[22] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, *Bringing Background into the Foreground: Making All Classes Equal in Weakly-Supervised Video Semantic Segmentation*, ICCV, London, UK, 2017.

[23] J. Zhuang, Z. Wang, and B. Wang, "Video semantic segmentation with distortion-aware feature correction," 2020.

[24] D. B. Bert, J. Xu, T. Tinne, and V. G. Luc, *Dynamic Filter Networks*, NIPS, London, UK, 2016.

[25] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, *Pixel-Adaptive Convolutional Neural Networks*, CVPR, London, UK, 2019.

[26] J. Liu, J. He, S. R. Jimmy, Y. Qiao, and H. Li, *Learning to Predict Context-Adaptive Convolution for Semantic Segmentation*, ECCV, London, UK, 2020.

[27] J. Dai, H. Qi, Y. Xiong et al., *Deformable Convolutional Networks*, ICCV, London, UK, 2017.

[28] X. Zhu, H. Hu, S. Lin, and J. Dai, *Deformable ConvNets V2: More Deformable, Better Results*, CVPR, London, UK, 2019.

[29] M. Cordts, M. Omran, S. Ramos et al., *The Cityscapes Dataset for Semantic Urban Scene Understanding*, CVPR, London, UK, 2016.
[30] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, *Segmentation and Recognition Using Structure from Motion Point Clouds*, ICCV, London, UK, 2008.
[31] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, CVPR, London, UK, 2016.
[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, CVPR, London, UK, 2018.
[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, *Pyramid Scene Parsing Network*, CVPR, London, UK, 2017.
[34] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, *Clockwork Convnets for Video Semantic Segmentation*, ECCV, London, UK, 2016.
[35] X. Jin, X. Li, H. Xiao et al., *Video Scene Parsing with Predictive Feature Learning*, ICCV, London, UK, 2017.