# Temporally-Consistent Video Semantic Segmentation with Bidirectional Occlusion-guided Feature Propagation

Razieh Kaviani Baghbaderani[1] Yuanxin Li[1] Shuangquan Wang[1] Hairong Qi[2]
[1]SOC R&D, Samsung Semiconductor, Inc.
[2]The University of Tennessee, Knoxville, TN, USA
{r.kaviani, yuanxin.li, shuangquan.w}@samsung.com hqi@utk.edu

## Abstract

*Despite recent progress in static image segmentation, video segmentation is still challenging due to the need for an accurate, fast, and temporally consistent model. Conducting per-frame static image segmentation on a video is not acceptable since it is computationally prohibitive and prone to temporal inconsistency. In this paper, we present bidirectional occlusion-guided feature propagation (BOFP) method with the goal of improving temporal consistency of segmentation results without sacrificing segmentation accuracy, while at the same time keeping the operations at a low computation cost. It leverages temporal coherence in the video by feature propagation from keyframes to other frames along the motion paths in both forward and backward directions. We propose an occlusion-based attention network to estimate the distorted areas based on bidirectional optical flows, and utilize them as cues for correcting and fusing the propagated features. Extensive experiments on benchmark datasets demonstrate that the proposed BOFP method achieves superior performance in terms of temporal consistency while maintaining comparable level of segmentation accuracy at a low computation cost, striking a great balance among the three performance metrics essential to evaluate video segmentation solutions.*

## 1. Introduction

Semantic segmentation is a fundamental problem in visual recognition systems. Regardless of tremendous progress for static image segmentation [3, 6, 8, 56, 63], video semantic segmentation (VSS) remains as a challenging problem mainly because of two reasons. First of all, processing a sheer amount of data in real time becomes non-trivial for some practical applications due to resource constraints. Secondly and more importantly, the segmentation predictions need to be temporally consistent in order to avoid the so-called "flickering" problem [31, 41].

To consider the temporal continuity in a video, an intuitive idea is to reuse the high-level features, extracted at sparse keyframes, by propagating them to non-keyframes for segmentation purpose [68], since high-level features change more slowly compared to shallower features extracted from video frames [49]. The feature propagation relies on a flow estimation which is much faster compared to deep feature extraction, and thus reduces the overall computational cost. However, the flow-based propagation can affect segmentation accuracy and create deteriorated predictions due to the imperfect optical flows, inevitable occlusions, or discontinuity across object boundaries.

To compensate for the aforementioned issues, it is adopted in [28, 69] to further update the features propagated from the sparse keyframes by features computed at the current frame. While approaches based on feature correction appear effective, their performance is bounded by the update branch. In particular, these methods may not be able to rectify all the distorted predictions using a light-weight network due to its weak feature extraction architecture. Note that utilizing a strong deep network as the update network contradicts the goal of reaching low computational cost. In addition, identifying the deteriorated regions is challenging since they have been accumulated through propagation over multiple frames. Hence the bottleneck in designing networks to achieve an accurate and temporally consistent VSS with low computation cost is the correction or update of high-level features propagated from keyframes.

In this paper, we propose a bidirectional occlusion-guided feature propagation (BOFP) framework for video semantic segmentation. It leverages the high-quality features extracted at the keyframes using a deep convolutional network, and propagates them toward the intermediate frames in bidirectional ways according to the flow fields. Unlike [67] that aggregates features of nearby frames with a similarity-based score for object detection, BOFP corrects the deteriorated features following the guidance from the learnable occlusion-based attention maps. The proposed attention network outputs maps that highlight the potentially

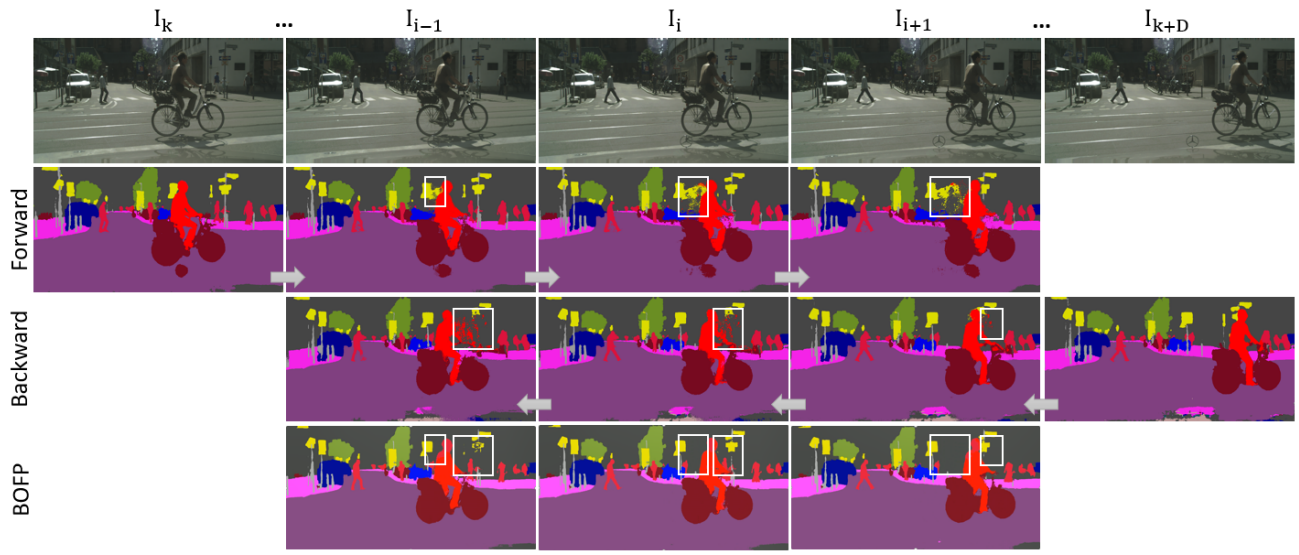


Figure 1. Illustration of the complementary relationship between the features propagated from the keyframes, $I_k$ and $I_{k+D}$, toward the intermediate non-keyframes, $I_{i-1}, I_i, I_{i+1}$, in forward and backward directions, respectively. It shows how distortions due to occlusion-disocclusion (highlighted with white boxes) can be mitigated through the proposed BOFP method.

occluded regions by the use of bidirectional optical flows embodying forward-backward movement information, in contrast to previous works [36, 39] that estimate occluded regions mathematically based on forward-backward consistency. The attention maps are then adopted as cues to compensate for the distortions from the features warped in the reverse direction.

Our hypothesis is that the high-level features propagated with bidirectional flows should be complementary to each other such that the ambiguous regions caused by inaccurate optical flows can be compensated by the features propagated from the opposite direction. As what could be observed in Fig. 1, the warped features from two keyframes are complementary, and the proposed BOFP method is capable of merging them effectively. In addition, feature corrections at non-keyframes are localized, that is, the level of corrections is determined by the severity of distortion calculated from the bidirectional flow.

The contribution of this work is three-fold: 1) We develop a novel occlusion-based attention estimation network and the corresponding occlusion-guided feature correction module such that feature correction is adapted to degree of distortion induced during feature propagation. 2) We apply the bidirectional feature propagation framework to the domain of video segmentation that takes full advantage of high-quality representations extracted at the keyframes when segmenting non-keyframes. 3) The proposed BOFP method achieves substantial improvements in the temporal consistency of the predictions while preserving the same level of segmentation accuracy at a low computation cost as compared to the state-of-the-art works.

## 2. Related Work

The proposed BOFP method is closely related to image semantic segmentation, bidirectional optical flow, and occlusion estimation, in addition to VSS. In the following, we discuss related works from these four areas.

**Image Semantic Segmentation**. As a seminal work, the fully convolutional network (FCN) [38] replaces fully-connected layers in a classification network [50] with convolutional layers to achieve dense predictions. Follow-up models [3, 35] extend FCN with explicit encoder-decoder architectures to obtain high-resolution outputs. Methods of [62, 63] utilize the dilated convolution to enlarge the receptive fields with a minimal computational cost. Different from these works, the spatial pyramid pooling module [18] presented in PSPNet [65] aggregates multi-scale contextual information obtained with different filters and pooling operations. DeepLabv3+ [8] combines the advantages from both encoder-decoder structure and atrous spatial pyramid pooling [6, 7]. The recent high performance network, HRNet [56], fuses various resolution representations in parallel to enhance the final prediction.

**Video Semantic Segmentation.** A number of works leverage cross-frame relations by applying the same image segmentation network to each video frame and aggregating the features over time. We refer to this group of VSS approaches as "non keyframe-based" VSS. For example, NetWarp [16] combines the intermediate features warped from the previous time step with the ones extracted from the current frame. STFCN [15] and GRFP [44] incorporate a spatial-temporal LSTM and a gated recurrent

unit, respectively, to temporally aggregate semantic labels. Consistency loss is incorporated in [37, 46, 64] as the extra constraint during training. TDNet [21] and its variants [33, 47, 55, 60] combine features extracted from shallower sub-networks by an attention-based propagation module. Recently, transformer-based methods, including STT [33] and CFFM [51], have improved segmentation accuracy with sacrificing the computation cost. Although these methods boost the segmentation accuracy, they suffer from high computation burden as all features are recalculated at each frame.

Another group of works aiming at efficient VSS take advantage of temporal continuity to reuse the deep features extracted at only sparse keyframes. We refer to this kind of methods as "keyframe-based" VSS. For example, Clockwork Net [49] directly reuses the features of preceding keyframe and updates them at the current frame according to their semantic stability. The features are propagated in [34] using spatial variant convolution. DFF [68] adopts optical flow to propagate high-level features to non-keyframes. Further, Accel [28] updates the propagated features with shallow features obtained from the current frame. DAVSS [69] proposes to correct the propagated features only on the distorted regions which is estimated by a light-weight network. Recently, GSVNet [32] introduces a guided spatially-varying convolution to fuse the segmentation outputs from nearby time steps. Although these methods decrease the overall computational cost compared to the non keyframe-based approaches, there is a drop in the segmentation accuracy due to potential errors in optical flows, large motions, and occlusions-disocclusion.

The proposed BOFP method is fundamentally a keyframe-based approach. However, unlike the works mentioned above, BOFP operates in a bidirectional framework and rectifies the features with the help of attention maps taking into consideration of occlusions-disocclusion in both forward and backward directions to largely mitigate ambiguities and uncertainties existed in the occluded areas.

**Bidirectional Optical Flow.** The bidirectional optical flow technique was initially proposed in [2]. It has been used for video frame interpolation [1, 4, 17, 19, 29, 43], motion detection [48], robust optical flow estimation [24, 25, 36, 39, 57], and object detection [67]. Recently, GRFP [27, 44] has incorporated it to a video semantic segmentation model which operates in a non keyframe-based manner. Unlike the methods mentioned above, we employ bidirectional flows in the keyframe-based framework which performs bidirectional propagation in a feature space and aggregates features with the guidance of the occlusion maps obtained from a learnable network.

**Occlusion Estimation.** Occlusion estimation plays a critical role in many vision tasks, including image instance/panoptic segmentation [10, 58], video frame interpolation [17, 43], object tracking [22, 61, 66], and optical flow estimation [24, 25, 36, 39, 57]. To the best of our knowledge, BOFP is the first to utilize occlusion estimation for video semantic segmentation such that the ambiguities and uncertainties in the propagated features caused by the object motions can be rectified by the guidance of occlusion-based attention maps while other high-quality propagated features are preserved.

## 3. Problem Statement

Given a video composed of $T$ frames $\{I_i\}, i = 1, 2, \dots T$, the goal is to design a model which outputs the corresponding segmentation of the video frames $\{S_i\}, i = 1, 2, \dots T$. To meet requirements of real-world applications, the segmentation model needs to maintain a reasonable balance among latency, accuracy, and temporal consistency.

In spite of the accuracy achieved by recent single-frame segmentation networks, these approaches are either too slow or could not produce smooth segmentation results over time. Other works exploiting temporal continuity of video through warping high-level features generate deteriorated segmentation due to inaccurate optical flows. Instead, we aim to develop a model that benefits from the temporal information in a video while being able to compensate for the inherent spatial misalignment between frames.

## 4. Methodology

We propose the BOFP framework as occlusion/disocclusion is the main source of distortion in the flow-based spatial warping. This framework contains three major components, as shown in Fig. 2: 1) bidirectional feature propagation leveraging the high-level features of the keyframes for segmenting the non-keyframes, 2) occlusion-based attention estimation identifying ambiguities occurred during feature propagation via flow fields, and 3) occlusion-guided feature correction.

### 4.1. Bidirectional Feature Propagation

Assuming there is a distance $D$ between the consecutive keyframes, $I_k$ and $I_{k+D}$, a deep image segmentation network, $\text{SegNet}_{\text{deep}}$, is executed only on the keyframes. The extracted keyframe feature, $h_k = \text{SegNet}_{\text{deep}}(I_k)$, is forward-propagated to subsequent frames in a frame-by-frame fashion using the optical flow, $F^f_{i,i-1} = \text{FlowNet}(I_{i-1}, I_i)$ where FlowNet represents the flow estimation network. Meanwhile, the feature obtained from the incoming keyframe, $h_{k+D} = \text{SegNet}_{\text{deep}}(I_{k+D})$, is warped backward to previous frames in a frame-by-frame fashion along the optical flow, $F^b_{i,i+1} = \text{FlowNet}(I_{i+1}, I_i)$. The feature warping in forward and backward directions can be represented, respectively, as:

$$h^f_i = \text{W}(h^f_{i-1}, F^f_{i,i-1}), \quad h^b_i = \text{W}(h^b_{i+1}, F^b_{i,i+1}), \quad (1)$$

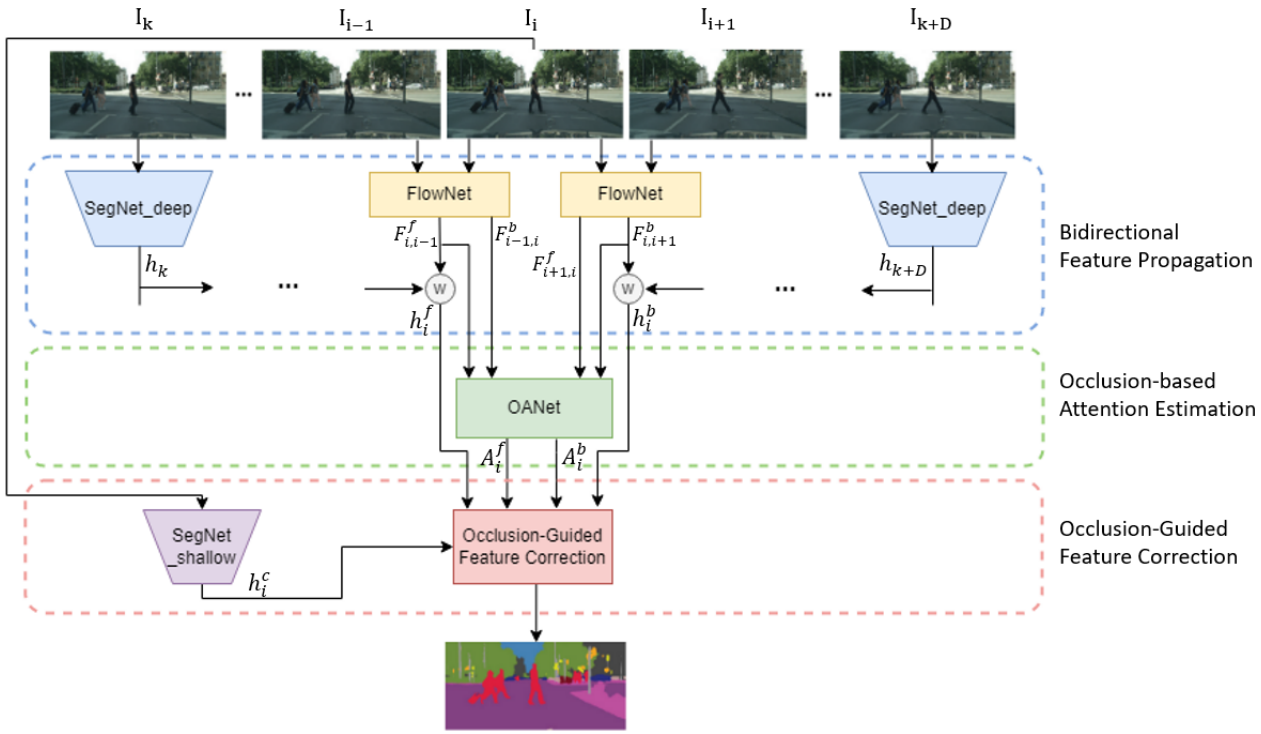


Figure 2. An overview of the proposed BOFP framework that comprises of a bidirectional feature propagation component (Sec. 4.1), occlusion-based attention estimation component (Sec. 4.2), and an occlusion-guided feature correction component (Sec. 4.3). It leverages the high-quality deep features extracted from two keyframes ($I_k$ and $I_{k+D}$) using a static deep segmentation network (SegNet$_{\text{deep}}$) and propagates them toward the non-keyframe ($I_i$) using optical flows. The potential ambiguities in the propagated features are rectified with the help of attention maps estimated by an occlusion-based attention network (OANet). The ambiguities are further corrected by a shallow segmentation network (SegNet$_{\text{shallow}}$) executed on the current frame to save computation cost.

where W denotes the warping operation which is the bilinear interpolation as in [69].

### 4.2. Occlusion-based Attention Estimation

Inspired by the forward-backward consistency assumption [53], we design an occlusion-based attention network. It is built upon the assumption that the forward flow $F^f$ negates the backward flow $F^b$ on the pixels where there is no error in the flow field. This observation does not hold true for pixels which become occluded/dis-occluded.

The binary occlusion mask can be calculated using the well-known occlusion reasoning approach [24, 39]. However, it has two hyper-parameters which are difficult to tune for each individual video. One of our major contributions is to design an occlusion-based attention estimation network, dubbed OANet, that takes four optical flows, $[F^f_{i,i-1}, F^b_{i-1,i}, F^f_{i+1,i}, F^b_{i,i+1}]$, among three consecutive frames ($I_{i-1}$, $I_i$, $I_{i+1}$) as input and estimates the occlusion-based attention maps. See Fig. 3. The attention maps, as compared to the binary occlusion masks, are the key enabler to a more effective feature correction approach that attends to the occluded/dis-occluded areas more than the relatively static regions such that the bidirectionally propagated features can be rectified and fused more efficiently and effectively.

Denote the binary occlusion masks between each pair of consecutive frames ($I_{i-1}$, $I_i$, $I_{i+1}$) as $[O^f_{i,i-1}, O^b_{i-1,i}]$ and $[O^f_{i+1,i}, O^b_{i,i+1}]$, respectively, then the bidirectionally propagated features can be aggregated as follows:

$$h_{i_1} = (1 - O^b_{i-1,i}) \times h^f_i + O^b_{i-1,i} \times h^b_i, \tag{2}$$

$$h_{i_2} = (1 - O^f_{i+1,i}) \times h^b_i + O^f_{i+1,i} \times h^f_i, \tag{3}$$

$$h_i = \frac{1}{2}(h_{i_1} + h_{i_2}) = A^f_i \times h^f_i + A^b_i \times h^b_i, \tag{4}$$

where

$$\begin{aligned} A^f_i &= \frac{1}{2}(1 - O^b_{i-1,i} + O^f_{i+1,i}), \\ A^b_i &= \frac{1}{2}(1 + O^b_{i-1,i} - O^f_{i+1,i}), \end{aligned} \tag{5}$$

can be viewed as the attention maps that are learned by the proposed occlusion-based attention estimation network automatically. To be specific, the complementary optical

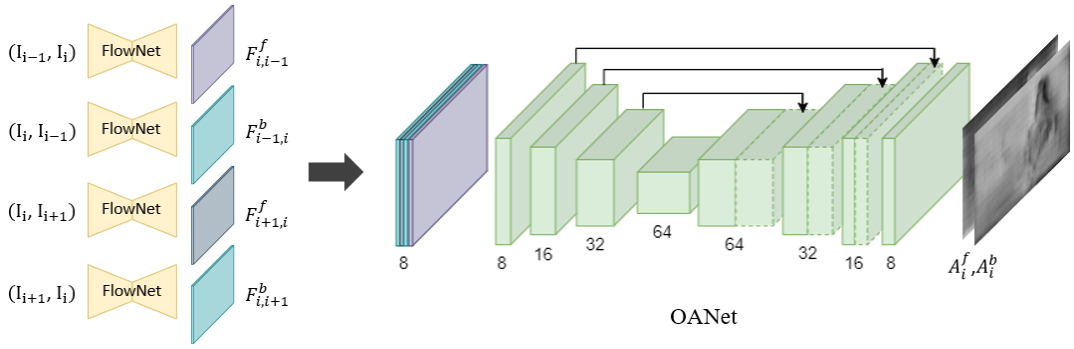


Figure 3. Occlusion-based attention estimation network. The Convolution blocks in the encoder and Deconvolution blocks in decoder are interlaced with BatchNorm and ReLU.

flows predict attention maps in forward direction, $A^f_i$, and in backward direction, $A^b_i$, that highlight the distorted areas in the bidirectional propagation directions.

Note that since calculating $O^b_{i-1,i}$ and $O^f_{i+1,i}$ requires knowing $[F^f_{i-1,i}, F^b_{i,i-1}]$ and $[F^f_{i,i+1}, F^b_{i+1,i}]$, respectively, the OANet takes all the four bidirectional flows as input to produce $A^f_i$ and $A^b_i$.

### 4.3. Occlusion-Guided Feature Correction

The attention maps, $A^f$ and $A^b$, which are the attention maps associated with the warped features in the forward and backward directions, respectively, yield higher values for areas with intact features while smaller ones for the occluded/disoccluded areas. Therefore, the attention maps can serve as the weights indicating the "quality" of the features propagated from the keyframes. Then the corrected features at the current frame $I_i$ can be obtained by,

$$h_i = h^f_i \times A^f_i + h^b_i \times A^b_i + h^c_i \times (1 - A^f_i - A^b_i) \quad (6)$$

where $\times$ represents the spatially element-wise multiplication, $h^c_i$ is the extracted feature from the current frame using the shallow image segmentation network $\text{SegNet}_{\text{shallow}}$, and $1 - A^f_i - A^b_i$ highlights pixels where the attention network is not certain, which usually happens on the object boundaries. For these ambiguous regions, the proposed BOFP framework gives more weight to features extracted from the current frame. Later, the *argmax* operation is applied on the rectified features to output the final segmentation result $S_i = argmax(h_i)$.

### 4.4. Training and Inference Strategy

**Training.** The proposed framework is trained in two phases. In phase one, the bidirectional propagation only comprises of $\text{SegNet}_{\text{deep}}$ and FlowNet. It fine-tunes FlowNet while the weights of $\text{SegNet}_{\text{deep}}$ are kept fixed. In parallel, the shallow static segmentation network $\text{SegNet}_{\text{shallow}}$ is trained on the particular target dataset (e.g. Cityscapes) from scratch. In phase two, OANet in the full framework is trained from scratch while the weights of other networks are kept frozen.

During training, a batch of three frames $[I_{i-p}, I_i, I_{i+q}]$ are utilized, where only the image $I_i$ has ground truth on semantic annotation. The indices of the images in the batch are generated based on the relation as $q = D - p + 1$ where $1 \leq p \leq D$. It enables the framework to propagate the features with various propagation distances while the distance between the keyframes is kept fixed as $D$. Following [28, 68, 69], keyframes are selected at regular intervals, starting with the first frame in the video.

**Inference.** The inference of the proposed BOFP framework is demonstrated in Algorithm 1. Given a video of frames $\{I_i\}$ and the specified keyframe interval $D$, the proposed method bidirectionally rectifies the features of the intermediate frames with the help of learnt attention maps.

**Algorithm 1** Inference of the proposed BOFP framework for video semantic segmentation

**Input**: video frames $\{I_i\}$, $D$
**Output**: segmentation results $\{S_i\}$

$k = 0$
$h_0 = \text{SegNet}_{\text{deep}}(I_0)$
**for** $i = 0$ to $T$ **do**
  **if** $isKeyframe(i)$ **then**
    $h^f_i = h_i$
    $k = i$
    $h_{k+D} = \text{SegNet}_{\text{deep}}(I_{k+D})$
    $h^b_{k+D} = h_{k+D}$
  **else**
    $h^f_i = \text{W}(h^f_{i-1}, F^f_{i,i-1})$
    $h^b_i = \text{W}(h^b_{i+1}, F^b_{i,i+1})$
    $h^c_i = \text{SegNet}_{\text{shallow}}(I_i)$
    $h_i = h^f_i \times A^f_i + h^b_i \times A^b_i + h^c_i \times (1 - A^f_i - A^b_i)$
  **end if**
  $S_i = argmax(h_i)$
**end for**

## 5. Experiments and Results

### 5.1. Experimental Setup

**Datasets.** Cityscapes [11] consists of 5000 snippets recorded at 17 FPS of street scenes in 50 different cities. The dataset is divided into 2975/500/1525 snippets corresponding to training/validation/test sets. Each frame is in resolution of 1024×2048, and the 20th frame of each snippet is carefully annotated with 19 semantic labels. CamVid [5] contains 4 video clips captured at 30 FPS with resolution of 720×960. The 11-class high-quality dense pixel annotation is provided for each 30th frame. Following [69], the trainval and test sets have 468 and 233 samples, respectively. VSPW [40], the current largest VSS benchmark, contains 2806 training clips, 343 validation clips, and 387 test clips, respectively, of which the pixel-level annotations are provided at 15 FPS. During experiments, we resize all images in VSPW to size of 480×853.

**Evaluation Metrics.** The standard mean Intersection over Union (mIoU) [14] is calculated as the metric for segmentation accuracy. More importantly, we utilize the mean temporal consistency (mTC) metric [54] as another essential metric to measure the smoothness and continuation of segmentation predictions over time. The metric frames per second (FPS) is reported to compare the latency of different models. We also report the amount of giga floating point operations per second (GFLOPS) [20, 42] which is a more intrinsic description of the computation cost of a model. Following [40], we calculate the mean video concictency (mVC) for the experiments on the VSPW dataset.

**Implementation Details.** We use DeepLabv3+ [8] or HRNetV2 [56] as the segmentation model for keyframes, i.e. $\text{SegNet}_{\text{deep}}$, due to their superior performance in terms of accuracy and efficiency in segmenting static images. They are pretrained on the ImageNet [12] dataset and then fine-tuned on the target segmentation dataset. The segmentation model for non-keyframes, i.e. $\text{SegNet}_{\text{shallow}}$, has a U-Net shape structure comprising of a few convolutional layers, interlaced with BatchNorm and Leaky-ReLU layers in the encoder and a few deconvolutional layers interlaced with Leaky-ReLU in the decoder.

We adopt the modified FlowNet2-S [26] which has less complexity compared to the original one. FlowNet is pretrained on the synthetic Flying Chairs dataset [13] and then fine-tuned on particular dataset during the bidirectional propagation learning. We use Adam optimizer [30] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The optimization is performed for 100 epochs with a learning rate of $10^{-4}$. We set the keyframe interval as 5. Note that we use a fixed keyframe interval to follow the same protocol as in previous works [28, 68, 69]. In Sec. 5, we study the effect of different $D$.

### 5.2. Comparison with State-of-the-Arts

We compare our proposed method with other state-of-the-art approaches in Tab. 1, where none of the comparison methods uses bi-direction technique. As shown, our method with Xception-71 and HRNetV2-W18 backbones outperform the corresponding baselines (i.e. DeepLabV3+ and HRNetV2) in terms of mTC and FPS, with a subtle drop in mIoU. It can be observed that with Xception-71 as backbone, our method achieves higher mIoU compared to the keyframe-based methods, including DFF, Accel, DAVSS, while having comparable FPS. As our target is to have reasonable balance between accuracy and efficiency, we reimplemented GRFP and TDNet with HRNetV2-W18 backbone to have comparable computation costs. It can be seen that our method surpasses these methods significantly in terms of mTC and FPS, with comparable mIoU. Also, our method outperforms AuxAdapt significantly in terms of mTC while having comparable mIoU and also achieves better mIoU than CFFM and MRCFA. It is to be noted that the methods with more expensive backbones, like NetWarp, STT, GRFP, TDNet, DDRNet, CAA with ResNet101 backbone, and TMANet have boosted the accuracy while suffering from high computation cost resulting in low FPS.

The comparison with other methods on VSPW dataset is reported in Tab. 2. It can be observed that our BOFP achieves better performance compared to the baseline, i.e. DeepLabv3+, and better TC compared to TCB, NetWarp, and ETC methods.

In order to demonstrate how much accuracy drops for each class after our feature propagation and rectification steps, the class-wise IoUs are presented in the Supplementary Material.

**Effect of Keyframe Interval.** To evaluate how the segmentation accuracy changes with respect to the keyframe interval, $D$, the proposed method with Xception-71 backbone is compared with other keyframe-based methods for various keyframe intervals ranging from 1 to 9. Fig. 4 depicts the results of different models on the Cityscapes validation set. The result is in line with intuition that as the keyframe interval increases, the segmentation accuracy decreases due to increasing inaccuracy in optical flow estimation and the growing ambiguities caused by occlusion-disocclusion. Meanwhile, our model is able to achieve a more stable performance demonstrating the importance of bidirectional framework on rectifying the ambiguities aggregated over the frames.

**Visual Comparison.** The qualitative comparison of the proposed BOFP method and the state-of-the-art keyframe-based method (i.e., DAVSS) is shown in Figs. 5 and 6 on the Cityscapes and CamVid datasets, respectively. In particular, our approach is able to better segment the moving objects where ambiguities occur due to object motion. In order to demonstrate the superiority of BOFP over the keyframe-

Table 1. Comparison with state-of-the-art methods on the Cityscapes [11] and CamVid [5].

| Method | Backbone | Cityscapes | | | | CamVid | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | mIoU ↑ | mTC ↑ | FPS ↑ | GFLOPs ↓ | mIoU ↑ | mTC ↑ | FPS ↑ | GFLOPs ↓ |
| PSPNet50 [18] | ResNet50 | 75.5 | - | 4.2 | - | 71.0 | - | 2.8 | - |
| PSPNet101 [18] | ResNet101 | 79.4 | 69.7 | 2.1 | 2048 | 77.6 | 77.1 | 4.1 | - |
| NetWarp [16] | ResNet101 | 80.6 | - | 0.3 | - | 67.1 | - | 2.8 | - |
| STT [33] | ResNet101 | 82.5 | 73.9 | 2.2 | - | 80.2 | 82.3 | 4.2 | - |
| DDRNet [45] | ResNet101 | 80.4 | - | 22 | 281 | 76.3 | - | 94 | - |
| CAA [23] | ResNet101 | 82.6 | - | - | 224 | - | - | - | - |
| TMANet [55] | FCN-50 | 80.3 | - | - | 754 | 76.5 | - | - | - |
| SegFormer [59] | MiT-B0 | 76.2 | - | 15 | 125 | - | - | - | - |
| CFFM [51] | MiT-B1 | 74.8 | 84.4 | 17.2 | - | - | - | - | - |
| MRCFA [52] | MiT-B1 | 75.1 | - | 21.5 | - | - | - | - | - |
| DFF [68] | Xception-71 | 68.7 | - | 9.9 | 308 | 66.0 | - | 18.5 | 102 |
| Accel [28] | Xception-71 | 72.1 | 70.3 | 3.6 | 572 | 66.7 | - | 7.6 | 190 |
| DAVSS [69] | Xception-71 | 75.4 | 84.5 | 9.7 | 334 | 71.1 | 85.0 | 18.2 | 112 |
| GSVNet [32] | BiSeNet-18 | 72.0 | - | 123.4 | - | 64.8 | - | 250 | - |
| GRFP [44] | ResNet101 | 80.2 | - | 3.2 | - | 66.1 | - | 4.4 | - |
| GRFP [44] | HRNetV2-W18 | 76.6 | 83.8 | 4.5 | 468 | 74.6 | 87.2 | 9.3 | 156 |
| TDNet [21] | ResNet50 | 79.9 | 71.1 | 5.6 | - | 72.6 | 77.4 | 11.1 | - |
| TDNet [21] | HRNetV2-W18 | 76.5 | 81.6 | 18.6 | 161 | 72.6 | 84.7 | 33.0 | 54 |
| AuxAdapt [64] | HRNetV2-W18 | 76.6 | 75.3 | - | - | 73.2 | 79.1 | - | - |
| DeepLabv3+ [8] | Xception-71 | 76.6 | 76.6 | 1.2 | 820 | 72.0 | 83.2 | 2.2 | 270 |
| HRNetV2 [56] | HRNetV2-W18 | 75.9 | 81.0 | 18.9 | 156 | 75.0 | 83.9 | 37.7 | 52 |
| BOFP (Ours) | Xception-71 | 76.5 | 84.8 | 9.6 | 349 | 71.8 | 85.5 | 18.1 | 116 |
| BOFP (Ours) | HRNetV2-W18 | 75.7 | 86.5 | 19.7 | 216 | 74.4 | 88.4 | 38.3 | 73 |

Table 2. Comparison with SOTA works on the VSPW dataset.

| Method | Backbone | mIoU | TC | $mVC_8$ |
|---|---|---|---|---|
| DeepLabv3+ [9] | ResNet-101 | 34.7 | 65.4 | 83.2 |
| PSPNet [18] | ResNet-101 | 36.5 | 65.9 | 84.2 |
| ETC [37] | PSPNet | 36.5 | 67.9 | 84.1 |
| NetWarp [16] | PSPNet | 36.9 | 67.8 | 84.1 |
| $TCB_{st-ppm}$ [40] | ResNet-101 | 37.5 | 70.3 | 86.9 |
| BOFP (Ours) | Xception-71 | 35.5 | 70.6 | 84.5 |

and non keyframe-based methods in terms of temporal stability, two videos showing the results over the course of time are provided in the Supplementary Material.

### 5.3. Method Analysis

**Ablation Study.** The contribution of each component in the proposed framework is illustrated in Tab. 3, where DeepLabv3+ and forward propagation is considered as baseline (see 1st row). It is observed that incorporating the backward propagation (Bkwd) improves mIoU by 3.9% while introducing 25 ms of latency (see 3rd row). Adding the features extracted from the current frame (Curr) slightly improves the baseline (see 2nd row), but reduces the performance of the forward-backward framework (see 4th row). Adding OANet to the bidirectional framework further enhances mIoU by rectifying only the distorted areas indicated in the occlusion maps (see 5th row). Our full framework outperforms the other variants and achieves 76.5% mIoU

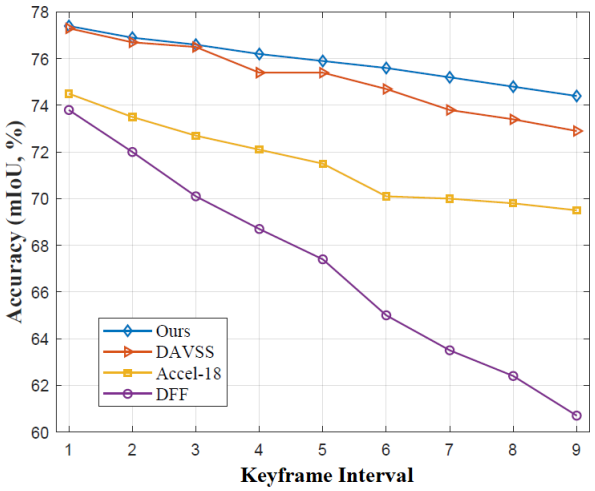


Figure 4. Comparison of accuracy-interval on the Cityscapes dataset.

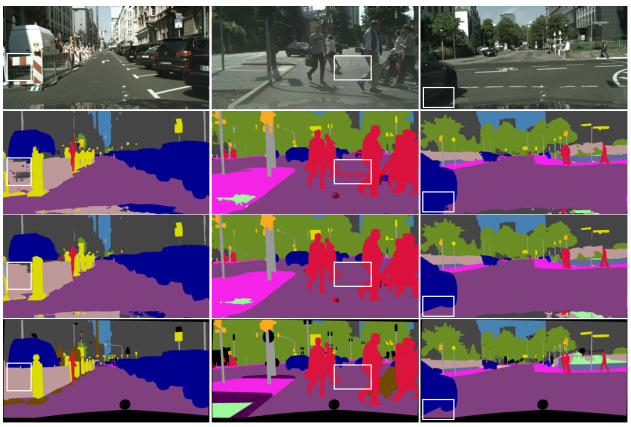

Figure 5. Semantic segmentation results on Cityscapes validation dataset. From top to bottom: the image, video segmentation by DAVSS [69], video segmentation by BOFP, and the ground truth.

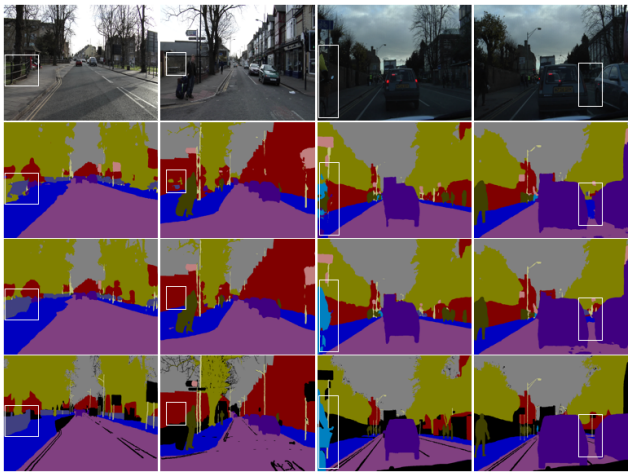

Figure 6. Semantic segmentation results on CamVid dataset. From top to bottom: the image, video segmentation by DAVSS [69], video segmentation by BOFP, and the ground truth.

with a reasonable overhead latency.

**Effect of the Feature Correction Module.** The impact of the occlusion-guided feature correction is compared with other fusion methods including directly adding feature maps (Add) and adopting a $1\times1$ convolutional layer (Conv$1\times1$). As shown in Tab. 4, our "OANet" approach outperforms the "Add" module which treats every pixel equally, and also, fusion with a $1\times1$ convolutional layer which leads to the worst accuracy.

**Occlusion Visualization.** The intermediate results including the propagated features in forward and backward directions, feature extracted from the current frame and the attention maps, are illustrated in Fig. 7. It can be observed that the proposed method estimates the occlusion areas for both forward and backward directions, and highlights the regions where both propagated features are uncertain about the predictions which is used for extra refinement with the help of shallow features extracted from the current frame.

Table 3. Contribution of different components in BOFP, including backward propagation (Bkwd), Occlusion-based attention network (OANet), and the use of update branch at the current frame (Curr). The mIoU scores and runtimes on a non-keyframe for the Cityscapes dataset are provided.

| Bkwd | OANet | Curr | mIoU (%) | Time ($ms/f$) |
|---|---|---|---|---|
| | | | 72.0 | 171 |
| | | ✓ | 72.1 | 184 |
| ✓ | | | 75.9 | 196 |
| ✓ | | ✓ | 75.2 | 213 |
| ✓ | ✓ | | 76.3 | 253 |
| ✓ | ✓ | ✓ | 76.5 | 265 |

Table 4. Effect of different feature correction modules in our framework on the Cityscapes dataset. "OANet" represents the proposed feature correction based on occlusion maps. "Add" and "Conv$1\times1$" denote adding features maps and fusion using a $1\times1$ convolution layer, respectively.

| Method | mIoU (%) | Time ($ms/f$) |
|---|---|---|
| OANet | 76.5 | 265 |
| Add | 75.2 | 213 |
| Conv$1\times1$ | 61.4 | 244 |

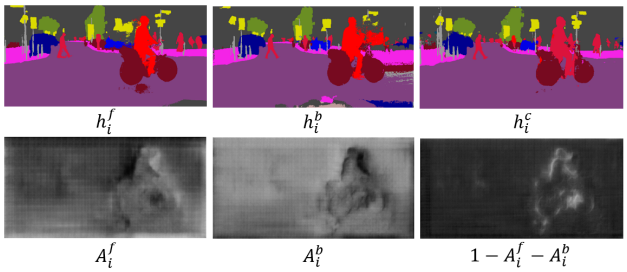


Figure 7. Intermediate result examples from BOFP on Cityscapes. Top row, from left to right: the propagated features in forward and backward directions, and the one obtained from the current frame. Bottom row, from left to right: the occlusion maps in forward and backward directions, and the remaining distortion map.

## 6. Conclusion

This work presents the novel bidirectional occlusion-guided feature propagation framework for video semantic segmentation. It leverages temporal coherence in the video by feature propagation along bidirectional flows and rectifies the deteriorated predictions according to the learnt attention maps from the proposed occlusion-based attention network. The extensive experiments on Cityscapes, CamVid, and VSPW benchmarks demonstrate the superiority of the proposed method with a better balance among accuracy, temporal consistency, and computation cost.

## References


[1] Alexander Alshin and Elena Alshina. Bi-directional pptical flow for future video codec. In *2016 Data Compression Conference (DCC)*, pages 83–90. IEEE, 2016. 3

[2] Alexander Alshin, Elena Alshina, and Tammy Lee. Bi-directional optical flow for improving motion compensation. In *28th Picture Coding Symposium*, pages 422–425. IEEE, 2010. 3

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1, 2

[4] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 3

[5] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 6, 7

[6] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 2

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 2, 6, 7

[9] L. et al. Chen. Encoder-decoder with atrous separable convolution for semantic image seg. In *ECCV*, 2018. 7

[10] Yifeng Chen, Guangchen Lin, Songyuan Li, Omar Bourahla, Yiming Wu, Fangfang Wang, Junyi Feng, Mingliang Xu, and Xi Li. Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2020. 3

[11] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 6

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 6

[15] Mohsen Fayyaz, Mohammad Hajizadeh Saffar, Mohammad Sabokrou, Mahmood Fathy, Fay Huang, and Reinhard Klette. Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes. In *Asian Conference on Computer Vision*, pages 493–509. Springer, 2016. 2

[16] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4462, 2017. 2, 7

[17] Donghao Gu, Zhaojing Wen, Wenxue Cui, Rui Wang, Feng Jiang, and Shaohui Liu. Continuous bidirectional optical flow for video frame sequence interpolation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1768–1773. IEEE, 2019. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 2, 7

[19] Evan Herbst, Steve Seitz, and Simon Baker. Occlusion reasoning for temporal interpolation using optical flow. *Department of Computer Science and Engineering, University of Washington, Tech. Rep. UW-CSE-09-08-01*, 2009. 3

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[21] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020. 3, 7

[22] Yan Huang and Irfan Essa. Tracking multiple objects through occlusions. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1051–1058. IEEE, 2005. 3

[23] Ye Huang, Di Kang, Wenjing Jia, Liu Liu, and Xiangjian He. Channelized axial attention–considering channel relation within spatial attention for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1016–1025, 2022. 7

[24] Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 312–321, 2017. 3, 4

[25] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 3

[26] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 6

[27] Samvit Jain and Joseph E Gonzalez. Fast semantic segmentation on video using block motion-based feature interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[28] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019. 1, 3, 5, 6, 7

[29] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 3

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 1

[32] Shih-Po Lee, Si-Cun Chen, and Wen-Hsiao Peng. Gsvnet: Guided spatially-varying convolution for fast semantic segmentation on video. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3, 7

[33] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 59–68, 2021. 3, 7

[34] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5997–6005, 2018. 3

[35] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2

[36] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 2, 3

[37] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, pages 352–368. Springer, 2020. 3, 7

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[39] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3, 4

[40] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 7

[41] Ondrej Miksik, Daniel Munoz, J Andrew Bagnell, and Martial Hebert. Efficient temporal consistency for streaming video scene analysis. In *2013 IEEE International Conference on Robotics and Automation*, pages 133–139. IEEE, 2013. 1

[42] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016. 6

[43] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1710, 2018. 3

[44] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018. 2, 3, 7

[45] Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3448–3460, 2022. 7

[46] Hyojin Park, Alan Yessenbayev, Tushar Singhal, Navin Kumar Adhikari, Yizhe Zhang, Shubhankar Mangesh Borse, Hong Cai, Nilesh Prasad Pandey, Fei Yin, Frank Mayer, et al. Real-time, accurate, and consistent video semantic segmentation via unsupervised adaptation and cross-unit deployment on mobile device. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21431–21438, 2022. 3

[47] Matthieu Paul, Martin Danelljan, Luc Van Gool, and Radu Timofte. Local memory attention for fast video semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1102–1109. IEEE, 2021. 3

[48] Sandeep Singh Sengar and Susanta Mukhopadhyay. Motion detection using block based bi-directional optical flow method. *Journal of Visual Communication and Image Representation*, 49:89–103, 2017. 3

[49] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision*, pages 852–868. Springer, 2016. 1, 3

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[51] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3126–3137, 2022. 3, 7

[52] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *European Conference on Computer Vision*, pages 522–539. Springer, 2022. 7

[53] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. 4

[54] Serin Varghese, Yasin Bayzidi, Andreas Bar, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico M Schmidt, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 336–337, 2020. 6

[55] Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258. IEEE, 2021. 3, 7

[56] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 6, 7

[57] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018. 3

[58] Zhengyang Wu, Fuxin Li, Rahul Sukthankar, and James M Rehg. Robust video segment proposals with painless occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4203, 2015. 3

[59] E. et al. Xie. Segformer: Simple and efficient design for semantic segmentation with transformers. *NerIPS*, 2021. 7

[60] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2946–2954, 2022. 3

[61] Alper Yilmaz, Xin Li, and Mubarak Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1531–1536, 2004. 3

[62] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[63] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 1, 2

[64] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2339–2348, 2022. 3, 7

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[66] Yue Zhou and Hai Tao. A background layer model for object tracking through occlusion. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1079–1085. IEEE, 2003. 3

[67] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017. 1, 3

[68] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017. 1, 3, 5, 6, 7

[69] Jiafan Zhuang, Zilei Wang, and Bingke Wang. Video semantic segmentation with distortion-aware feature correction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1, 3, 4, 5, 6, 7, 8