Efficient Semantic Video Segmentation with Per-frame Inference

Yifan Liu[1], Chunhua Shen[1], Changqian Yu[2,1], and Jingdong Wang[3]

[1] The University of Adelaide, Australia
[2] Huazhong University of Science and Technology, China
[3] Microsoft Research

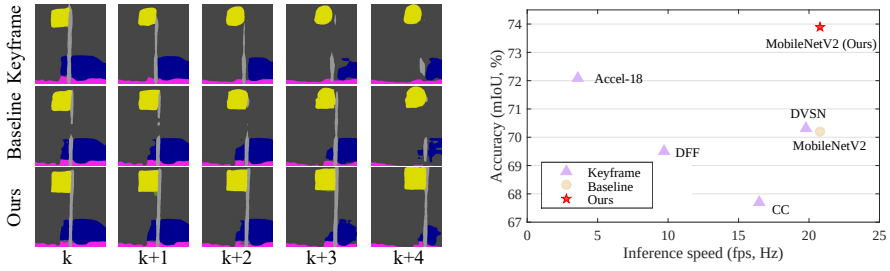
Abstract. For semantic segmentation, most existing real-time deep models trained with each frame independently may produce inconsistent results when tested on a video sequence. A few methods take the correlations in the video sequence into account, e.g., by propagating the results to the neighbouring frames using optical flow, or extracting frame representations using multi-frame information, which may lead to inaccurate results or unbalanced latency. In contrast, here we explicitly consider the temporal consistency among frames as extra constraints during training and process each frame independently in the inference phase. Thus no computation overhead is introduced for inference. Compact models are employed for real-time execution. To narrow the performance gap between compact models and large models, new temporal knowledge distillation methods are designed. Weighing among accuracy, temporal smoothness and efficiency, our proposed method outperforms previous keyframe based methods and corresponding baselines which are trained with each frame independently on benchmark datasets including Cityscapes and Camvid.

Keywords: Semantic video segmentation, temporal consistency

1 Introduction

Semantic segmentation, a fundamental task in computer vision, aims to assign a semantic label to each pixel in an image. In recent years, the development of deep learning has brought significant success to the task of image semantic segmentation [37, 31, 5] on benchmark datasets, but often with a high computational cost. This task becomes computationally more expensive when extending to video. For a few real-world applications, e.g., autonomous driving and robotics, it is challenging but crucial to build a fast and accurate video semantic segmentation system.

Previous works for semantic video segmentation can be categorized into two groups. The first group focuses on improving the performance for video segmentation by performing post-processing among frames [18], or employing extra modules to use multi-frames information during inference [8]. The high computational cost makes it difficult for mobile applications. The second group uses



(a) Temporal consistency (b) Accuracy vs. inference speed.

Fig. 1: (a) Visualization results on consecutive frames: *Keyframe*: Accel18 [13] propagates and fuses the results from the keyframe (k) to non-key frames ($k+1, \dots$), which may lead to poor results on non-key frames. *Baseline*: PSPNet18 [37] trains the model on single frames. Inference on single frames separately can produce temporally inconsistent results. *Ours*: training the model with the correlations among frames and inferring on single frames separately lead to high quality and smooth results. **(b)** Comparing our enhanced MobileNetV2 model with previous keyframe based methods: Accel [13], DVSN [32], DFF [39] and CC [27]. The inference speed is evaluated on a single GTX 1080Ti.

keyframes to avoid processing of each frame, and then propagate [39, 38, 32] the outputs or the feature maps to other frames (non-key frames) using optical flows. Keyframe based methods indeed accelerate inference. However, it requires different inference time for keyframes and non-key frames, leading to an unbalanced latency, thus being not friendly for real-time processing. Moreover, accuracy cannot be guaranteed for each frame due to the cumulative warping error, for example, the first row in Figure 1(a).

Efficient semantic segmentation methods on 2D images [20, 34, 23] have draw much attention recently. Clearly, applying compact networks to each frame of a video sequence independently may alleviate the latency and enable real-time execution. However, directly training the model on each frame independently often produces temporally inconsistent results on the video as shown in the second row of Figure 1(a). To address the above problems, we explicitly consider the temporal consistency among frames as extra constraints during the training process and employ compact networks with per-frame inference to ease the problem of latency and achieve real-time inference.

A motion guided *temporal loss* is employed with the motivation of assigning a consistent label to the same pixel along the time axis. A motion estimation network is introduced to predict the motion (e.g., optical-flow) of each pixel from the current frame to the next frame based on the input frame-pair. Predicted semantic labels are propagated to the next frame to supervise predictions of the next frame. Thus, the temporal consistency is encoded into the segmentation network through this constraint.

To narrow the performance gap between compact models and large models, we design a new *temporal consistency knowledge distillation* strategy to help the training of compact models. Distillation methods are widely used in image recognition tasks [19, 11, 16], and achieve great success. Different from previous distillation methods, which only consider the spatial correlations, we embed the temporal consistency into distillation items. We extract the pair-wise frames dependency by calculating the pair-wise similarities for different locations between two frames, and further encode the multi-frames dependency into a latent embedding by using a recurrent unit, ConvLSTM [28]. The new distillation methods not only improve temporal consistency but also boost segmentation accuracy. We also include the spatial knowledge distillation methods [19] of single frames in training to further improve the accuracy.

We evaluate the proposed methods on semantic video segmentation benchmarks: Cityscapes [6] and Camvid [3]. A few compact backbone networks, i.e., PSPNet18 [37], MobileNetV2 [26] and a lightweight HRNet [30], are included to verify that the proposed methods can empirically improve the segmentation accuracy and the temporal consistency, without any extra computation and post-processing during inference. The proposed methods also show superiority in the trade-off of accuracy and the inference speed. For example, with the per-frame inference fashion, our enhanced MobileNetV2 [26] can achieve higher accuracy with a faster inference speed compared with state-of-the-art keyframe based methods as shown in Figure 1(b). We summarize our main contributions as follows.

- We process semantic video segmentation with compact models by per-frame inference, without introducing post-processing and computation overhead, enabling real-time inference without latency.
- We explicitly consider the temporal consistency in the training process by using a temporal loss and newly designed temporal consistency knowledge distillation methods.
- Empirical experiment results on Cityscapes and Camvid show that with the help of proposed training methods, the compact models outperform previous state-of-the-art semantic video segmentation methods weighing among accuracy, temporal consistency and inference speed.

1.1 Related Work

Semantic Video Segmentation. Semantic video segmentation requires dense labeling for all pixels in each frame of a video sequence into a few semantic categories. Previous work can be summarized into two streams.

The first one focuses on improving the accuracy by exploiting the temporal relations and the unlabelled data in the video sequence. Nilsson and Sminchiesescu [22] employ a gated recurrent unit to propagate semantic labels to unlabeled frames. Other works like NetWarp [8], STFCN [7], and SVP [18] also employ optical-flow or recurrent units to fuse the results of several frames during inferring to improve the segmentation accuracy. Recently, Zhu *et al.* [40]

propose to use a motion estimation network to propagate labels to unlabeled frames as data augmentation and achieve state-of-the-art performance with the segmentation accuracy. These methods can achieve significant performance but can be difficult to be deployed on mobile devices.

The second line of works pay attention to reduce the computational cost by re-using the feature maps in the neighbouring frames. ClockNet [27] proposes to copy the feature map to the next frame directly, thus reducing the computational cost. DFF [39] employs the optical flow to warp the feature map between the keyframe and non-key frames. Xu *et al.* [32] further propose to use an adaptive keyframe selection policy while Zhu *et al.* [38] find out that propagating partial region in the feature map can get better performance. Li *et al.* [17] propose a low-latency video segmentation network by optimizing both the keyframe selection and the adaptive feature propagation. Accel [13] proposes a network fusion policy to use a large model to predict the keyframe and use a compact one in non-key frames. Keyframe based methods require different inference time and may produce different quantity results between keyframes and other frames. In this work, we solve the real-time video segmentation by per-frame inference with a compact network and propose a temporal loss and the temporal consistency knowledge distillation to ensure both good accuracy and temporal consistency.

Temporal Consistency. Applying image processing algorithms to each frame of a video may lead to inconsistent results. The temporal consistency problem has draw much attention in low-level and mid-level applications, such as task-specific methods including colorization [15], style transfer [9], and video depth estimation [2, 1] and task agnostic approaches [14, 33]. Temporal consistency is also essential in semantic video segmentation. Miksik *et al.* [21] employ a post-processing method that learns a similarity function between pixels of consecutive frames to propagate predictions across time. Nilsson and Sminchiesescu [22] insert the optical flow estimation network into the forward pass and employ a recurrent unit to make use of neighbouring predictions. Our method is more efficient than theirs as we employ per-frame inference. The warped previous predictions work as a constraint *only* during training.

Knowledge Distillation. The effectiveness of knowledge distillation has been verified in classification [12, 25, 35]. The output of the large teacher net, including the final logits and the intermediate feature maps, are treated as soft targets to supervise the compact student net. Previous knowledge distillation methods in semantic segmentation [11, 19] design distillation strategies only for improving the segmentation accuracy. To our knowledge, to date no distillation methods consider to improve temporal consistency. In this work, we focus on encoding the motion information into the distillation terms to make the segmentation networks more suitable for the semantic video segmentation tasks.

2 Approach

In this section, we show how we exploit the temporal information during training. As shown in Figure 2(a), we introduce two terms: a simple temporal loss (Fig-

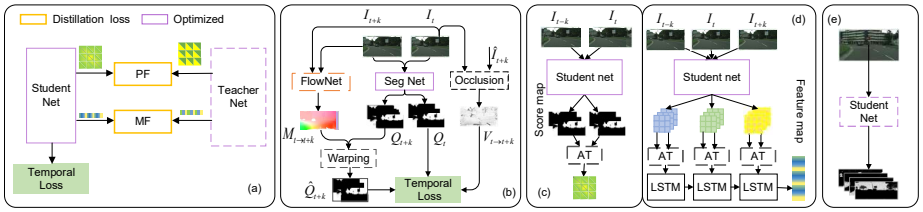


Fig. 2: (a) **Overall of proposed training scheme**: We consider the temporal information by the temporal consistency knowledge distillation (c and d) and the temporal loss (b) during training. (b) **Temporal loss (TL)** encode the temporal consistency through motion constraints. Both the teacher net and the student net are enhanced by the temporal loss. (c) **Pair-wise frame dependency (PF)**: encode the motion relations between two frames. (d) **multi-frame dependency (MF)**: extract the correlations of the intermediate feature maps among multi-frames. We only show the forward pass of the student net here and apply the same operations on the teacher net to get the dependency cross frames as soft targets. (e) **The inference process**. All the proposed methods are only applied during training. We can improve the temporal consistency as well as the segmentation accuracy without any extra parameters or post-processing during inference.

ure 2(b)) and newly designed temporal consistency knowledge distillation strategies (Figure 2(c) and Figure 2(d)). The temporal consistency of the single-frame models can be significantly improved by employing temporal loss. However, if compact models are employed for real-time execution, there is still a performance gap between large models and small ones. We design new temporal consistency knowledge distillation to transfer the temporal consistency from large models to small ones. With the help of temporal information, the segmentation accuracy can also be boosted.

2.1 Motion Guided Temporal Consistency

Training semantic segmentation networks independently on each frame of a video sequence often leads to undesired inconsistency. Conventional methods include previous predictions as an extra input, which introduces extra computational cost during inference. We employ previous predictions as supervised signals to assign consistent labels to each corresponding pixel along the time axis.

As shown in Figure 2(b), for two input frames $\mathbf{I}_t$, $\mathbf{I}_{t+k}$ from time t and $t+k$, we have:

$$\ell_{tl}(\mathbf{I}_t, \mathbf{I}_{t+k}) = \frac{1}{N}\sum_{i=1}^{N} V^{(i)}_{t\Rightarrow t+k} \left\| \mathbf{q}^i_t - \hat{\mathbf{q}}^i_{t+k\Rightarrow t} \right\|_2^2 \tag{1}$$

where $\mathbf{q}^i_t$ represents the predicted class probability at the position i of the segmentation map $\mathbf{Q_t}$, and $\hat{\mathbf{q}}^i_{t+k\Rightarrow t}$ is the warped class probability from frame $t+k$ to frame t, by using a motion estimation network(e.g., FlowNet) $f(\cdot)$. Such an

$f(\cdot)$ can predict the amount of motion changes in the x and y directions for each pixel: $f(\mathbf{I}_{t+k}, \mathbf{I}_t) = \mathbf{M}_{t \to t+k}$, where $\delta i = \mathbf{M}_{t \to t+k}(i)$, indicating the pixel on the position i of the frame t moves to the position $i + \delta i$ in the frame $t + k$. Therefore, the segmentation maps between two input frames are aligned by the motion guidance. An occlusion mask $\mathbf{V}_{t \Rightarrow t+k}$ is designed to remove the noise caused by the warping error: $\mathbf{V}_{t \Rightarrow t+k} = \exp(-\left|\mathbf{I}_t - \hat{\mathbf{I}}_{t+k}\right|)$, where $\hat{\mathbf{I}}_{t+k}$ is the warped input frame. We employ a pre-trained optical flow prediction network as the motion estimation net in implementation. We directly consider the temporal consistency during the training process through the motion guided temporal loss by constraining a moving pixel along the time steps to have a consistent semantic label. Similar constraints are proposed in image processing tasks [14, 33], but rarely discussed in semantic segmentation. We find that the straightforward temporal loss can improve the temporal consistency of single-frame models significantly.

2.2 Temporal Consistency Knowledge Distillation

Inspired by [19], we build a distillation mechanism to effectively train the compact student net S by making use of the cumbersome teacher net T. The teacher net T is already well trained with the cross-entropy loss and the temporal loss to achieve a high temporal consistency as well as the segmentation accuracy. Different from previous single frame distillation methods, two new distillation strategies are designed to transfer the temporal consistency from T to S: pair-wise-frames dependency (PF) and multi-frame dependency (MF).

Pair-wise-Frames Dependency. Following [19], we denote an attention (AT) operator to calculate the pair-wise similarity map $\mathbf{A}_{\mathbf{X}_1,\mathbf{X}_2}$ of two input tensors $\mathbf{X}_1, \mathbf{X}_2$, where $\mathbf{A}_{\mathbf{X}_1,\mathbf{X}_2} \in \mathbb{R}^{N \times N \times 1}$ and $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{N \times C}$. For the pixel a_{ij} in $\mathbf{A}$, we calculate the cosine similarity between $\mathbf{x}_1^i$ and $\mathbf{x}_2^j$ from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively: $a_{ij} = {\mathbf{x}_1^i}^\top \mathbf{x}_2^j / (\|\mathbf{x}_1^i\|_2 \|\mathbf{x}_2^j\|_2)$. It is an efficient and easy way to encode the correlations between two input tensors.

As shown in Figure 2(c), we encode the pair-wise dependency between the prediction of every two neighbouring frame pairs by using the AT operator, and get the similarity map $\mathbf{A}_{\mathbf{Q}_t,\mathbf{Q}_{t+k}}$, where $\mathbf{Q}_t$ is the segmentation map of frame t and a_{ij} of $\mathbf{A}_{\mathbf{Q}_t,\mathbf{Q}_{t+k}}$ denotes the similarity between the class probabilities on the location i of the frame t and the location j of the frame $t + k$. If a pixel on the location i of frame t moves to location j of frame $t+k$, the similarity a_{ij} may be higher. Therefore, the pair-wise dependency can reflect the motion correlation between two frames.

We align the pair-wise-frame (PF) dependency between the teacher net T and the student net S,

$$\ell_{PF}(\mathbf{Q_t}, \mathbf{Q_{t+k}}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_{ij}^{\mathsf{S}} - a_{ij}^{\mathsf{T}})^2, \tag{2}$$

where $\forall a_{ij}^{\mathsf{S}} \in \mathbf{A}_{\mathbf{Q}_t,\mathbf{Q}_{t+k}}^{\mathsf{S}}$ and $\forall a_{ij}^{\mathsf{T}} \in \mathbf{A}_{\mathbf{Q}_t,\mathbf{Q}_{t+k}}^{\mathsf{T}}$.

Multi-Frame Dependency. As shown in Figure 2(d), for a video sequence $\mathcal{I} = \{\dots \mathbf{I}_{t-1}, \mathbf{I}_t, \mathbf{I}_{t+1} \dots\}$, the corresponding feature maps $\mathcal{F} = \{\dots \mathbf{F}_{t-1}, \mathbf{F}_t, \mathbf{F}_{t+1} \dots\}$ are extracted from the output of the last convolutional block before the classification layer. Then, the self-similarity map, $\mathbf{A}_{\mathbf{F}_t, \mathbf{F}_t}$, for each frame are calculated by using AT operator in order to: 1) capture the structure information among pixels, and 2) align the different feature channels between the teacher net and student net.

We employ a ConvLSTM unit to encode the sequence of self-similarity maps into an embedding $\mathbf{E} \in \mathbb{R}^{1 \times D_e}$, where D_e is the length of the embedding space. For each time step, the ConvLSTM unit takes $\mathbf{A}_{\mathbf{F}_t, \mathbf{F}_t}$ and the hidden state which contains the information of previous $t-1$ frames as input and gives an output embedding $\mathbf{E}_t$ along with the hidden state of the current time step. We align the final output embedding [4] at the last time step, $\mathbf{E}^{\mathsf{T}}$ and $\mathbf{E}^{\mathsf{S}}$ from T and S, respectively. The output embedding encodes the relations of the whole input sequence, named multi-frame dependency (MF). The distillation loss based on multi-frame dependency is termed as: $\ell_{MF}(\mathcal{F}) = \left\|\mathbf{E}^{\mathsf{T}} - \mathbf{E}^{\mathsf{S}}\right\|_2^2$.

The parameters in the ConvLSTM unit are optimized together with the student net. To extract the multi-frame dependency, both the teacher net and the student net share the weight of the ConvLSTM unit. Note that there exists a model collapse point when the weights and bias in the ConvLSTM are all equal to zero. We clip the weights of ConvLSTM between a certain range and enlarges the $\mathbf{E}^{\mathsf{T}}$ as a regularization to prevent the model collapse.

2.3 Optimization

We pre-train the teacher net with the segmentation loss and the temporal loss to attain a segmentation network with a high semantic accuracy and temporal consistency. When optimizing the student net, we fix the weight of the motion estimation net (FlowNet) and the teacher net. These two parts are only used to calculate the temporal loss and the distillation terms, which can be seen as extra regularization terms during the training of the student net. During training, we also employ conventional cross-entropy loss, and the single frame distillation method (SF) proposed in [21] on every single frame to improve the segmentation accuracy. Details can be found in Section S1.1 in supplementary materials. The whole objective function for a sampled video sequence consists of the conventional cross-entropy loss ℓ_{ce}, the single-frame distillation loss ℓ_{SF}, temporal loss, and the temporal consistency distillation terms:

$$\ell = \sum_{t=1}^{T'} \ell_{ce}^{(t)} + \lambda\Big(\sum_{t=1}^{T} \ell_{SF}^{(t)} + \sum_{i=1}^{T-1} \ell_{tl}(\mathbf{Q}_t, \mathbf{Q}_{t+1}) + \sum_{i=1}^{T-1} \ell_{PF}(\mathbf{Q}_t, \mathbf{Q}_{t+1}) + \ell_{MF}\Big), \quad (3)$$

where T is the number of all the frames in one training sequence, and T' is the number of labeled frames. Due to the high labeling cost in semantic video

[4] The details of calculations in ConvLSTM is referred in [28], and we also include the key equations in Section S1.2 in supplementary materials.

segmentation tasks [6, 3], most of the datasets are only annotated with sparse frames. Our methods can be easily applied to the sparse-labeled dataset, because 1) we can make use of large teacher models to generate soft targets; and 2) we care about the temporal consistency between two frames, which can be self-supervised through motion. The loss weight for all regularization terms λ is set to 0.1.

After training the compact network, all the motion-estimation net, teacher net, and the distillation modules can be removed. We only keep the student net as the semantic video segmentation network. Thus, both the segmentation accuracy and the temporal consistency can be improved with no extra computational cost in the per-frame inference process.

3 Implementation details

Dataset. We evaluate our proposed method on Camvid [3] and Cityscapes [6], which are standard benchmarks for semantic video segmentation [13, 27, 22]. More details of the training and evaluation can be found in Section S2 of the supplementary materials. **Network structures.** Different from the keyframe based method, which takes several frames as input during inferring, we apply our training methods to a compact segmentation model with per-frame inference. There are three main parts while training the system:

- A light-weight segmentation network. We conduct most of the experiments on ResNet18 with the architecture of PSPnet [37], namely PSPNet18. We also employ MobileNetV2 [26] and a light-weight HRNet-w18 [30] to verify the effectiveness and generalization ability.
- A motion estimation network. We use a pre-trained FlowNetV2 [24] to predict the motion between two frames. Because this module can be removed during inferring, we do not need to consider employing a lightweight flownet for acceleration, like in DFF [39] and GRFP [22].
- A teacher network. We adopt widely-used segmentation architecture PSPNet [37] with a ResNet101 [10] as the teacher network, namely PSPNet101, which is used to calculate the soft targets in distillation items. We train the teacher net with the temporal loss to enhance the temporal consistency of the teacher.

Random sampled policy. In order to reduce the computational cost while training video data, and make use of more unlabeled frames, we randomly sample frames in front of the labelled frame, named 'frame_f' and behind of the labelled frame, named 'frame_b' to form a training triplet (frame_f, labelled frame, frame_b), instead of only using the frames right next to the labelled ones. The random sampled policy can take both long term and short term correlations into consideration and achieve better performance. Training on a longer sequence may show better performance with more expensive computation.

Evaluation metrics. We evaluate our method on three aspects: accuracy, temporal consistency, and efficiency. The accuracy is evaluated by widely-used

Table 1: Accuracy and temporal consistency on Cityscapes validation set. SF: single-frame distillation methods, PF: our proposed pair-wise-frame dependency distillation method. MF: our proposed multi-frame dependency distillation method, TL: the temporal loss. The proposed distillation methods and temporal loss can improve both the temporal consistency and accuracy, and they are complementary to each other.

Scheme index	SF	PF	MF	TL	mIoU	Pixel accuracy	Temporal consistency
a					69.79	77.18	68.50
b	✓				70.85	78.41	69.20
c		✓			70.32	77.96	70.10
d			✓		70.38	77.99	69.78
e				✓	70.67	78.46	70.46
f		✓	✓		71.16	78.69	70.21
g	✓			✓	71.36	78.64	70.13
h		✓	✓	✓	71.57	78.94	70.61
i	✓	✓	✓		72.01	79.21	69.99
j	✓	✓	✓	✓	**73.06**	**80.75**	**70.56**

mean Intersection over Union (mIoU) and pixel accuracy for semantic segmentation [19]. We report the model parameters (#Param) and frames per second (fps) to show the efficiency of employed networks. We follow [14] to measure the temporal stability of a video based on the mean flow warping error between every two neighbouring frames. Different from [14], we use the mIoU score instead of the mean square error to evaluate the semantic segmentation results, and more details can be found in the supplementary materials.

4 Experiments

4.1 Ablations

All the ablation experiments are conducted on the Cityscapes dataset with the PSPNet18.

Effectiveness of proposed methods. In this section, we verify the effectiveness of the proposed training scheme. Both the accuracy and temporal consistency are shown in Table 1. We build the baseline scheme a, which is trained on every single labelled frame. Then, we apply three distillation terms: the single-frame dependency (SF), the pair-wise-frame dependency (PF) and multi-frame dependency (MF), separately, to get the scheme b, c and d. The temporal loss is employed in the scheme e. Compared with the baseline scheme, all the schemes can improve accuracy as well as temporal consistency. To compare scheme b with c and d, one can see that the newly designed distillation scheme across frames can improve the temporal consistency to a greater extent. From the scheme e, we can see the temporal loss is most effective for the improvement of temporal

Table 2: Impact of the random sample policy. RS: random sample policy, TC: temporal consistency, TL: temporal loss, Dis: distillation terms, ALL: combine TL with Dis. The proposed random sample policy can improve the accuracy and temporal consistency.

Method	RS	mIoU	TC
PSPNet18 + TL		70.04	70.21
PSPNet18 + TL	✓	70.67	70.46
PSPNet18 + Dis		71.24	69.48
PSPNet18 + Dis	✓	72.01	69.99
PSPNet18 + ALL		72.87	70.05
PSPNet18 + ALL	✓	73.06	70.56

consistency. To compare scheme f with i, we can see that single frame distillation methods [19] can improve the segmentation accuracy but may harm the temporal consistency.

To further improve the performance, we combine the distillation terms with the temporal loss and achieve the mIoU of 73.06% and temporal consistency of 70.56%. We do not increase any parameters or extra computational cost with per-frame inference. Both the distillation terms and the temporal loss can be seen as regularization terms, which can help the training process. Such regularization terms introduce extra knowledge from the pre-trained teacher net and the motion estimation network. Besides, performance improvement also benefits from temporal information and unlabelled data from the video.

Impact of the random sample policy. We apply the random sample (RS) policy when training with video sequence in order to make use of more unlabelled images, and capture the long-term dependency. Experiment results are shown in Table 2. By employing the random sampled policy, both the temporal loss and distillation terms can benefit from more sufficient training data in the video sequences, and obtain an improvement on mIoU from 0.24% to 0.69% as well as the temporal consistency from 0.19% to 0.63%. We employ such a random sampled policy considering the memory cost during training.

Table 3: Influence of the teacher net. TL: temporal loss. TC: temporal consistency. We use the pair-wise-frame distillation to show our design can transfer the temporal consistency from the teacher net.

Method	Teacher Model	mIoU	TC
PSPNet101	None	78.84	69.71
PSPNet101 + TL	None	79.53	71.68
PSPNet18	None	69.79	68.50
PSPNet18	PSPNet101	70.26	69.27
PSPNet18	PSPNet101 + TL	**70.32**	**70.10**

Table 4: We compare our methods with recent efficient image/video semantic segmentation networks on three aspects: accuracy (mIoU,%), smoothness (TC, %) and inference speed (fps, Hz). TL: temporal loss, ALL: all proposed terms, TC: temporal consistency, #Param: parameters of the networks.

Method	Backbone	#Params	Cityscapes			Camvid		
			mIoU	TC	fps	mIoU	TC	fps
Video-based methods: Train and infer on multi frames								
CC [27]	VGG16	-	67.7	71.2	16.5	-	-	-
DFF [39]	ResNet101	-	68.7	71.4	9.7	66.0	78.0	16.1
GRFP [22]	ResNet101	-	69.4	-	3.2	66.1	-	6.4
DVSN [32]	ResNet101	-	70.3	-	19.8	-	-	-
Accel [13]	ResNet101/18	-	72.1	70.3	3.6	66.7	76.2	7.1
Single frame methods: Train and infer on each frame independently								
PSPNet [37]	ResNet101	68.1	78.8	69.7	1.7	77.6	77.1	4.1
SKD-MV2 [19]	MobileNetV2	8.3	74.5	68.2	14.4	-	-	-
SKD-R18 [19]	ResNet18	15.2	72.7	67.6	8.0	72.3	75.4	13.3
PSPNet18 [37]	ResNet18	13.2	69.8	68.5	9.5	-	-	-
HRNet-w18 [29, 30]	HRNet	3.9	75.6	69.1	18.9	-	-	-
MobileNetV2 [26]	MobileNetV2	3.2	70.2	68.4	20.8	74.4	76.8	27.8
Ours: Train on multi frames and infer on each frame independently								
Teacher Net	ResNet101	68.1	**79.5**	**71.7**	1.7	**79.4**	**78.6**	4.1
PSPNet18+TL	ResNet18	13.2	71.1	70.0	9.5	-	-	-
PSPNet18+ALL	ResNet18	13.2	73.1	70.6	9.5	-	-	-
HRNet-w18+TL	HRNet	3.9	76.4	69.6	18.9	-	-	-
HRNet-w18+ALL	HRNet	3.9	76.6	70.1	18.9	-	-	-
MobileNetV2+TL	MobileNetV2	3.2	70.7	70.4	20.8	76.3	77.6	27.8
MobileNetV2+ALL	MobileNetV2	**3.2**	73.9	69.9	**20.8**	78.2	77.9	**27.8**

Impact of the teacher net. The temporal loss can improve the temporal consistency of both cumbersome models and compact models. We compare the performance of the student net training with different teacher net (i.e., with and without the proposed temporal loss) to verify that the temporal consistency can be transferred with our designed distillation term. The results are shown in Table 3. The temporal consistency of the teacher net (PSPNet101) can be enhanced by training with temporal loss by 1.97%. Meanwhile, the mIoU can also be improved by 0.69%. By using the enhanced teacher net in the distillation framework, the segmentation accuracy is comparable (70.26 **vs.** 70.32), but the temporal consistency has a significant improvement (69.27 **vs.** 70.10), indicating that the proposed distillation methods can transfer the temporal consistency from the teacher net.

Discussions. We focus on improving the accuracy and temporal consistency for real-time models by making use of temporal correlations. Thus, we do not introduce extra parameters during inference. A series of work [36, 34, 23] focus on designing network structures for fast segmentation on single images and achieve promising results. They do not contradict to our work. We will verify

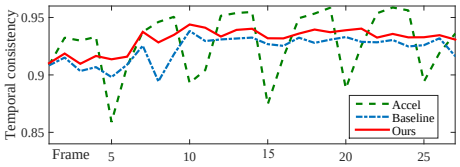


Fig. 3: The temporal consistency between neighbouring frames in one sampled sequence on Cityscapes. The keyframe based method Accel shows severe jitters between keyframes and others.

that our methods can generalize to different network structures, e.g. ResNet18, MobileNetV2 and HRNet in the next session. Besides, large models [37, 40] can achieve high segmentation accuracy but have low inference speed. The temporal loss is also effective when applying to large models, e.g., our teacher net.

4.2 Results on Cityscapes

Comparison with single-frame based methods. Single-frame methods are trained and inferred on each frame independently. Directly apply such methods to video sequences will produce inconsistent results. We apply our training schemes to several efficient single-frame semantic segmentation networks: PSPNet18 [37], MobileNetV2 [26] and HRNet-w18 [30, 29]. Metrics of mIoU, temporal consistency, inference speed, and model parameters are shown in Table 4. As Table 4 shows, the proposed training scheme works well with a few compact backbone networks (e.g., PSPNet18, HRNet-w18 and MobileNetV2). Both temporal consistency and segmentation accuracy can be improved using the temporal information among frames.

We also compare our training methods with the single-frame distillation method [19]. According to our observation, GAN based distillation methods proposed in [19]can produce inconsistent results. For example, with the same backbone ResNet18, training with the GAN based distillation methods (SKD-R18) achieves higher mIoU: 72.7 vs. 69.8, and a lower temporal consistency: 67.6 vs. 68.5 compared with the baseline PSPNet18, which is trained with cross-entropy loss on each single frame. We replace the GAN based distillation term with our temporal consistency distillation terms and the temporal loss, denoted as "PSPNet18+ALL". Both accuracy and smoothness are improved. Note that we also employ a smaller structure of the PSPNet with half channels than in [19].

Comparison with video-based methods. Video-based methods are trained and inferred on multi frames, we list current methods including keyframe based methods: CC [27], DFF [39], DVSN [32], Accel [13] and multi-frame input method: GRFP [22] in Table 4. The compact networks with per-frame inference can be more efficient than video-based methods. Besides, with per-frame inference, semantic segmentation networks have no unbalanced latency and can handle every frame independently. Table 4 shows the proposed training schemes can achieve a better trade-off between the accuracy and the inference speed compared

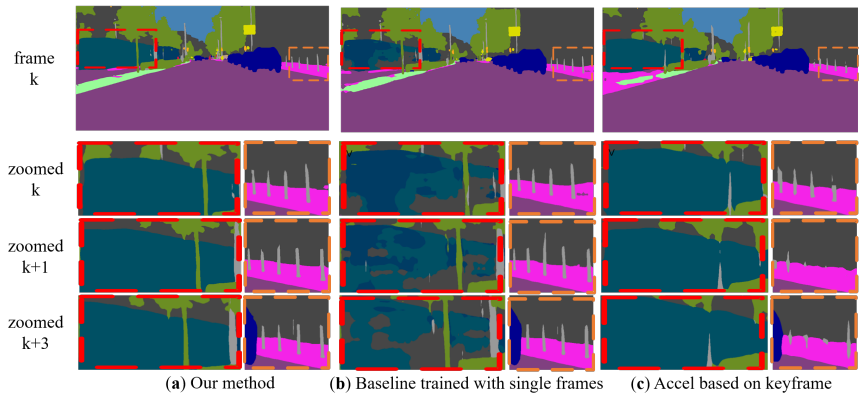


Fig. 4: Qualitative outputs. (**a**): PSPNet18, training on multi frames and inferring on each frame. (**b**): PSPNet18, training and inferring on each frame. (**c**): Accel-18 [13], training and inferring on multiple frames. The keyframe is selected in every five frames. For better visualization, we zoom the region in the red and orange box. The proposed method can give more consistent labels to the moving train and the trees in the red box. In the orange boxes, we can see our methods have similar quantity results in each frame while the keyframe based methods may generate worse results in the frame (e.g., $k+3$) which is far from the keyframe (i.e., k).

with other state-of-the-art semantic video segmentation methods, especially the MobileNetV2 with the fps of 20.8 and mIoU of 73.9. Although keyframe methods can achieve a high average temporal consistency score, the predictions beyond the keyframe are in low quality. Thus, the temporal consistency will be quiet low between keyframe and non-key frames, as shown in Figure 3. The high average temporal consistency score is mainly from the low-quality predictions on non-key frames. In contrast, our method can produce stable segmentation results on each frame.

Qualitative visualization. Qualitative visualization results are shown in Figure 4, in which, we can see, the keyframe-based method Accel-18 will produce unbalanced quality segmentation results between the keyframe (e.g., the orange box of k) and non-key frames (e.g., the orange box of $k+1$ and $k+3$), due to the different forward-networks it chooses. By contrast, ours can produce stable results on the video sequence because we use the same enhanced network on all frames. Compared with the baseline method trained on single frames, we can see our proposed method can produce more smooth results, e.g., the region in red boxes. *Video results* can be found in the supplementary materials. The improvement of temporal consistency is more clearly shown in the video comparison results. Moreover, we show a case of the temporal consistency between neighbouring frames in a sampled frame sequence in Figure 3. Temporal consistency between two frames is evaluated by the warping pixel accuracy. The higher, the better. The keyframe based method will produce jitters between keyframe and non-key frames, while our training methods can improve the temporal consis-

tency for every frame. The temporal consistency between non-key frames are higher than our methods, but the segmentation performance is lower than ours.

4.3 CamVid

We provide additional experiments on CamVid. We use MobileNetV2 as the backbone in the PSPNet. In Table 4, the segmentation accuracy, and the temporal consistency are improved compared with the baseline method. We also outperform current state-of-the-art semantic video segmentation methods with a better trade-off between the accuracy and the inference speed. We use the pre-trained weight from cityscapes following VideoGCRF [4], and achieve better segmentation results of 78.2 vs. 75.2. VideoGCRF [4] can achieve 22 fps with 321×321 resolution on a GTX 1080 card. We can achieve 78 fps with the same resolution. The consistent improvements on both datasets verify the value of our training schemes for real-time semantic video segmentation.

5 Conclusions

In this work, we have developed real-time video segmentation methods that consider not only accuracy but also temporal consistency. To this end, we have proposed to use compact networks with per-frame inference. We explicitly consider the temporal correlation during training by using: the temporal loss and the new temporal consistency knowledge distillation. For inference, the model processes each frame separately, which does not introduce latency and avoids post-processing. The compact networks achieve considerably better temporal consistency and semantic accuracy, without introducing extra computational cost during inference. Our experiments have verified the effectiveness of each component that we have designed. They can improve the performance individually and are complement to each other.

Acknowledgements Correspondence should be addressed to CS. CS was in part supported by ARC DP 'Deep learning that scales'.

References

1. Bian, J.W., Zhan, H., Wang, N., Chin, T.J., Shen, C., Reid, I.: Unsupervised depth learning in challenging indoor video: Weak rectification to rescue. arXiv: Comp. Res. Repository **abs/2006.02708** (2020)
2. Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: Proc. Advances in Neural Inf. Process. Syst. pp. 35–45 (2019)
3. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Proc. Eur. Conf. Comp. Vis. pp. 44–57. Springer (2008)
4. Chandra, S., Couprie, C., Kokkinos, I.: Deep spatio-temporal random fields for efficient video segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 8915–8924 (2018)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
7. Fayyaz, M., Saffar, M.H., Sabokrou, M., Fathy, M., Huang, F., Klette, R.: Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes. In: Proc. Asian Conf. Comp. Vis. pp. 493–509. Springer (2016)
8. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video cnns through representation warping. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 4453–4462 (2017)
9. Gupta, A., Johnson, J., Alahi, A., Fei-Fei, L.: Characterizing and improving stability in neural style transfer. In: Proc. IEEE Int. Conf. Comp. Vis. pp. 4067–4076 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 770–778 (2016)
11. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 578–587 (2019)
12. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv: Comp. Res. Repository **abs/1503.02531** (2015)
13. Jain, S., Wang, X., Gonzalez, J.E.: Accel: A corrective fusion network for efficient semantic segmentation on video. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 8866–8875 (2019)
14. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proc. Eur. Conf. Comp. Vis. pp. 170–185 (2018)
15. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. **23**(3), 689–694 (2004)
16. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 7341–7349 (2017)
17. Li, Y., Shi, J., Lin, D.: Low-latency video semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5997–6005 (2018)
18. Liu, S., Wang, C., Qian, R., Yu, H., Bao, R., Sun, Y.: Surveillance video parsing with single frame supervision. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 413–421 (2017)

19. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2604–2613 (2019)
20. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. Proc. Eur. Conf. Comp. Vis. (2018)
21. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: Proc. IEEE Int. Conf. Robotics and Automation. pp. 133–139. IEEE (2013)
22. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 6819–6828 (2018)
23. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
24. Reda, F., Pottorff, R., Barker, J., Catanzaro, B.: flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks. https://github.com/NVIDIA/flownet2-pytorch (2017)
25. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv: Comp. Res. Repository **abs/1412.6550** (2014)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)
27. Shelhamer, E., Rakelly, K., Hoffman, J., Darrell, T.: Clockwork convnets for video semantic segmentation. In: Proc. Eur. Conf. Comp. Vis. pp. 852–868. Springer (2016)
28. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Proc. Advances in Neural Inf. Process. Syst. pp. 802–810 (2015)
29. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
30. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv: Comp. Res. Repository **abs/1904.04514** (2019)
31. Tian, Z., He, T., Shen, C., Yan, Y.: Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3126–3135 (2019)
32. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 6556–6565 (2018)
33. Yao, C.H., Chang, C.Y., Chien, S.Y.: Occlusion-aware video temporal consistency. In: Proc. ACM Int. Conf. Multimedia. pp. 777–785. ACM (2017)
34. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proc. Eur. Conf. Comp. Vis. pp. 325–341 (2018)
35. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. Proc. Int. Conf. Learn. Representations (2017)
36. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. Proc. Eur. Conf. Comp. Vis. (2018)
37. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2881–2890 (2017)

38. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 7210–7218 (2018)
39. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2349–2358 (2017)
40. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 8856–8865 (2019)