Simultaneously Short- and Long-Term Temporal Modeling for Semi-Supervised Video Semantic Segmentation

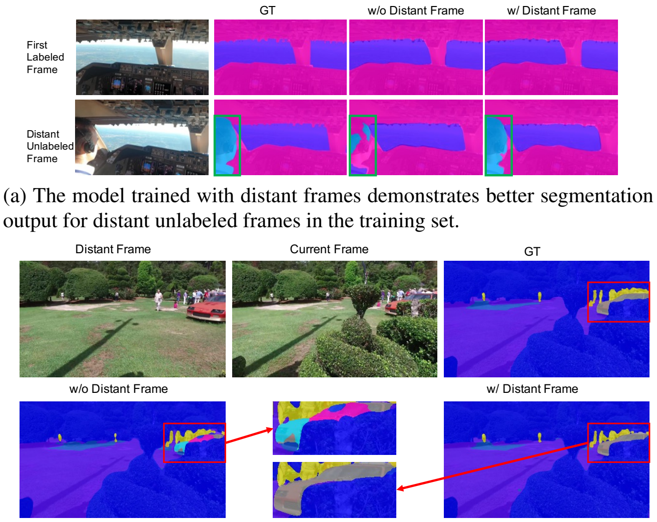
Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, Wei Chu
Ant Group
wenshuo.ljw@antgroup.com

Abstract

In order to tackle video semantic segmentation task at a lower cost, e.g.*, only one frame annotated per video, lots of efforts have been devoted to investigate the utilization of those unlabeled frames by either assigning pseudo labels or performing feature enhancement. In this work, we propose a novel feature enhancement network to simultaneously model short- and long-term temporal correlation. Compared with existing work that only leverage short-term correspondence, the long-term temporal correlation obtained from distant frames can effectively expand the temporal perception field and provide richer contextual prior. More importantly, modeling adjacent and distant frames together can alleviate the risk of over-fitting, hence produce high-quality feature representation for the distant unlabeled frames in training set and unseen videos in testing set. To this end, we term our method* ***SSLTM****, short for* ***S****imultaneously* ***S****hort- and* ***L****ong-Term* ***T****emporal* ***M****odeling. In the setting of only one frame annotated per video, SSLTM significantly outperforms the state-of-the-art methods by* $2\% \sim 3\%$ *mIoU on the challenging VSPW dataset. Furthermore, when working with a pseudo label based method such as MeanTeacher, our final model only exhibits* 0.13% *mIoU less than the ceiling performance (*i.e.*, all frames are manually annotated).*

1. Introduction

Deep neural networks have been the de-facto solution for many vision tasks such as image recognition [12], object detection [15] and semantic segmentation [21]. These state-of-the-art results are generally achieved by training very deep networks on large-scale labeled datasets, *e.g.*, ImageNet [27], COCO [16] and Cityscapes [4], *etc*. However, building such labeled datasets is labor-intensive and complicated. Hence, it is very appealing to explore less label-dependent alternatives that only requires a (small) portion of the dataset to be annotated [14, 24–26, 37].



(a) The model trained with distant frames demonstrates better segmentation output for distant unlabeled frames in the training set.

(b) The adjacent frames do not contain sufficient visual clues to segment the car in current frame, while the distant frame can provide richer context to guide the segmentation model.

Figure 1. **The importance of involving distant frames in training.** The exploitation of distant frames not only reduces the risk of over-fitting to the labeled frame and its adjacent ones, but also provides temporally long-term context to enhance the feature representation.

In this work, we aim to train the video semantic segmentation model under an extreme setting of annotation availability, *i.e.*, each video in the training set only has its first frame annotated. The significance of this problem is twofold: 1). Video semantic segmentation is a fundamental task in computer vision, with wide applications in many scenarios like autonomous driving [9], robot controlling [7]; 2). Compared with dense annotations, it takes much less (if not the least) cost to label one frame per video. More importantly, given the information redundancy [18, 41] within a video, it seems *intuitively unnecessary* to annotate every frame at all costs. Thus, it is of great practical and theoretical interests to explore the feasibility of conducting video semantic segmentation with one-frame-per-video-annotation.

Existing methods for this problem can be grouped into Pseudo Label based approaches and Feature Enhancement based ones, depending on whether explicit pseudo labels are generated for the unlabeled frames. For the former ones [1, 6, 40], a pseudo-label generator is often trained with the annotated frames, then the model is updated using both labeled and unlabeled data. As a comparison, the latter group of methods [18, 41] concentrates on obtaining high-quality representations based on the features from both labeled and unlabeled frames. Thus, these methods rely on feature enhancement modules that are specially designed for temporal feature fusion. Note, Pseudo Label based approaches and Feature Enhancement based ones are orthogonal, *i.e.*, they pay attention to different aspects of the semi-supervised video semantic segmentation task, and can usually work together to combine the best of two worlds as shown in Section 4.5. In this work, we will focus on the latter ones - designing innovative feature enhancement modules.

Prior arts on feature enhancement mostly focus on modeling short-term temporal correlation, under the assumption of temporal consistency [18, 41] among adjacent frames. Nevertheless, the distant frame is less exploited in existing work, due to its severe content changes and weak temporal consistency. However, in the setting of partial annotation, the absence of distant frames in training results in significant drawbacks: 1). As illustrated in Figure 1a, if the distant frame is not involved in the training phase, the model will be over adapted (or even over-fitted) to the labeled frame and its adjacent ones. Consequently, the generalization to distant frames and unseen videos in the testing set is severely hurt, leading to poor segmentation performance in the testing stage. 2). Since the distant frame can provide long-term temporal context, the representation quality of the current frame can be improved by leveraging the information from its distant frame. A qualitative sample is given in Figure 1b, where a severely occluded car is correctly segmented with the help of long-term context from the distant frame.

To address the aforementioned drawbacks, we propose a novel Simultaneously Short- and Long-Term Temporal Modeling (SSLTM) method to capture the temporal relationship from both adjacent and distant frames. To achieve this goal, we design three novel components in our model for representation learning: 1). We refer to the labeled frame as query frame, for its adjacent frames in the same video, we model the **short-term** inter-frame correlations by a Spatial-Temporal Transformer (STT). 2). For the purpose of **long-term** temporal modeling, we obtain a reference frame by randomly sampling a distant frame from the same video of the query frame, then feeding the reference frame's feature to our proposed Reference Frame Context Enhancement (RFCE) module, so as to enhance the representation of query frame. Meanwhile, as the reference frame is selected randomly from the entire video, our model is potentially trained with all data, rather than just the labeled frames and their adjacent ones. As such, we expect the model to be prevented from over-fitting to some extent. 3). To compensate for the semantic category representation from RFCE, we further propose a Global Category Context (GCC) module to model the global information across the whole dataset.

In summary, our method is a pioneer work to exploit both short- and long-term inter-frame correlations in the video semantic segmentation task. Thanks to the effective modeling of distant frames, our RFCE demonstrates outstanding performance, especially under the setting of partial annotation. Specifically, on the challenging VSPW dataset [22], the mIoU of our final model only decreases by 0.13% when switching from per-frame-annotation to the one-frame-per-video-annotation setting. To our knowledge, this is the first work that nearly closes the gap between dense annotations and one-frame-per-video ones, which is of great significance in practical applications. Compared with existing Feature Enhancement based video semantic segmentation methods [5, 18, 22, 23, 30, 41], our SSLTM demonstrates advantageous results (mIoU as 39.79%) by a large margin ($2\% \sim 3\%$ mIoU) on the VSPW dataset.

2. Related Work

2.1. Image Semantic Segmentation

Image semantic segmentation aims to assign a semantic class for each pixel in given images. Modern deep learning models for this task are mainly based on CNN [2, 10, 21, 31, 38, 39], with different emphases on multi-scale feature representation learning, object relationship modeling and context information aggregation, *etc*. With the recent development of transformer, a few specially designed transformer-based models [19, 29, 35] are proposed and outperform previous CNN-based methods. Unfortunately, it is computationally infeasible to naively extend these methods to the video domain, as the temporal correlation is neglected. We will treat image semantic segmentation methods as baselines in our experimental comparisons.

2.2. Semi-supervised Video Semantic Segmentation

As presented in the introduction, semi-supervised video semantic segmentation methods fall into two categories: Pseudo Label based methods and Feature Enhancement based ones.

Pseudo Label based methods focus on the generation of pseudo labels and subsequent model training using these pseudo labels. For example, Zhu *et al.* [40], Ganeshan *et al.* [6] and Naive-Student [1] first attain pseudo labels for unlabeled data in an offline manner, then train a model with those pseudo labels. In contrast, FixMatch [28], PseudoSeg [42] and CrossPseudo [3] produce pseudo labels in an online manner. It is worth noting that Zhu *et al*. [40] and Ganeshan *et al*. [6] design an independent pseudo labels generator for unlabeled frames, while Naive-Student [1], FixMatch [28], PseudoSeg [42] and CrossPseudo [3] directly harness the image segmentation model itself. As mentioned earlier, our work is orthogonal to this line of research and can be combined with any of the above Pseudo Label based methods, so as to further train our model with additional pseudo-labeled data.

Feature Enhancement based methods place particular emphasis on inter-frame correlations modeling. For example, NetWarp [5] employs optical flow to warp the features of the previous frames to that of the current frame, STGRU [23] leverages optical flow and a gated recurrent unit to adaptively propagate temporal information, SVP [17] fuses the semantic segmentation results of adjacent frames to refine prediction of the target frame, ETC [18] utilizes optical flow to warp the result of the adjacent frames to the target frame, and then adopts a temporal loss to narrow the difference between the results of these two frames, TMANet [33] uses the self-attention mechanism to integrate the temporal correlation between the current frame and adjacent frames, TCB [22] leverages a temporal context blending module to fuse the features of adjacent frames and the target frame, IFR [41] leverages the prototypes of unlabeled frames to reconstruct the feature of labeled frames, CFFM [30] utilizes a cross-frame Feature Mining module to fuse features between the query frame and adjacent frames. Additionally, some recently published works, such as TF-DL [11], build a fully supervised framework based on video tubes. They train the model with video tubes and supervise them with the labels of the whole video, making these methods difficult to apply to the semi-supervised tasks directly.

3. Methodology

In this section, we first present the overall framework for SSLTM, then explain each component in detail.

3.1. Overview

The overall network architecture of our method is illustrated in Figure 2. Our goal is to exploit both short- and long-term inter-frame correlations properly. We construct our training sample as below:

$$\Big([I_q, I_{adj_1,\sim,adj_n}, I_{ref}], L_q\Big), \tag{1}$$

where:

- I_q represents query frame. L_q is the label of the query frame, if existed. During training, I_q is the frame with pixel-level annotation, while in inference phase, I_q is the frame being predicted.
- $I_{adj_1,\sim,adj_n}$ are n adjacent frames close to I_q, used to provide short-term temporal clues. In our experiments, we set n to 3.
- I_{ref} stands for a reference frame, which is a temporally distant frame from I_q in the same video and to contribute long-term temporal information. In training phase, we randomly select a frame in the same video with I_q, as the reference frame. While in inference phase, the reference frame is the temporally farthest frame in the same video with I_q.

As illustrated in Figure 2, given a training sample, we harness a backbone model (*e.g*., ResNet-101 [8]) to attain the feature maps for each frame. Since the backbone model normally produces representations of different strides (*e.g*., 4, 8, 16 and 32), we will potentially have multiple feature maps for each of the $n + 2$ frames in the training sample. To effectively exploit the short- and long-term temporal correlations, we propose two corresponding modules: 1). Spatial-Temporal Transformer (STT) module is used to handle short-term temporal modeling, given the feature maps of the query frame and its n adjacent frames. 2). Reference Frame Context Enhancement (RFCE) aims to model the long-term temporal correlation, and is optimized for both query frame and reference frame. In addition, since reference frame might lack certain categorical information from query frame, we develop a Global Category Context (GCC) module to further compensate RFCE. The details of each component are elaborated as follows.

3.2. Spatial-Temporal Transformer

Compared with image segmentation, the key of video semantic segmentation is to leverage the inter-frame temporal correlations. Here, we follow the design paradigm of transformer and propose Spatial-Temporal Transformer (STT) to facilitate temporal modeling. Since each frame has multiple feature maps of different strides, we use parallel STT blocks to process them as shown in Figure 3.

Specifically, for each stride, we first stack feature maps from query frame and its n adjacent frames, then feed them to the corresponding STT block. Next, we upsample only the query frame's features from STT's outputs, to the same scale and concatenate them as short-term temporal representation of query frame, denoted as F_q^{st}. As Equation 2, we call the STT module as f_{STT} and the Feature Extraction module as f_{FE}. Symbol $\circ$ is composition operator.

$$F_q^{st} = f_{STT} \circ f_{FE}(I_q, I_{adj_1,\sim,adj_n}) \tag{2}$$

The architecture of our STT base unit is demonstrated in the bottom of Figure 3. Each unit consists of a 3D Windows Multi-head Self Attention (3D W-MSA) [20], a Mix Feed-Forward Network (Mix-FFN) [36] and two Layer Norm

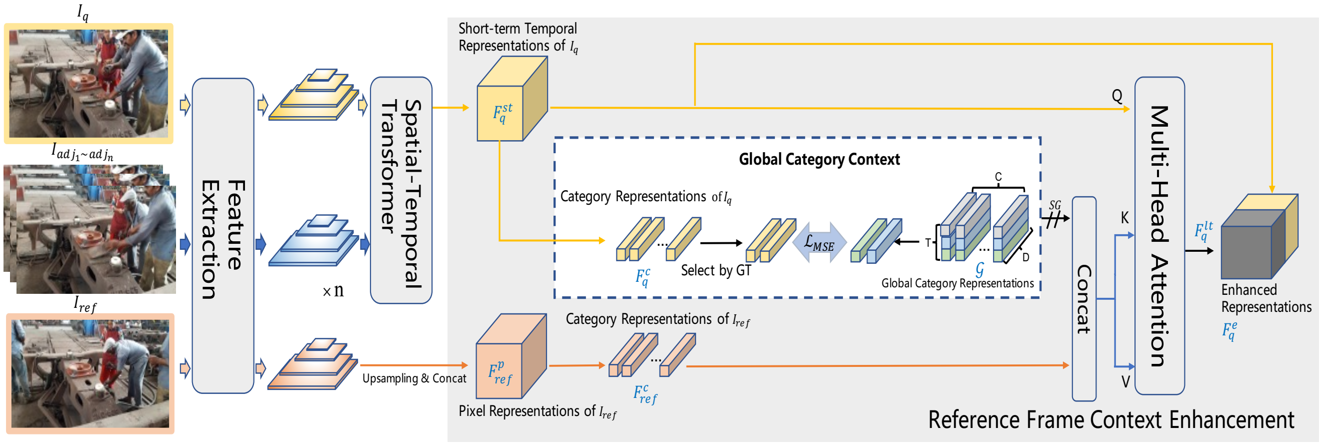


Figure 2. **Architecture Overview**. We simultaneously input a tuple of query frame, adjacent frames and reference frame to the model, and utilize several components (*e.g.*, STT, RFCE and GCC) to model the short- and long-term temporal correlations. Note, the Global Category Representations $\mathcal{G}$ is fixed in the phase of inference, and the process in the *blue dashed box* only runs in the training phase. *SG* stands for "Stop Gradient". GT stands for "Ground Truth".

(LN) layers. 3D W-MSA evenly partitions the 3D input feature map into a set of non-overlapping cubes and applies MSA on them. Mix-FFN introduces a depth-wise 3×3 convolution between the two MLPs to connect non-overlapping cubes. Details can be found in the supplementary.

3.3. Reference Frame Context Enhancement

STT uses a fixed-sized window to fuse temporal information within a small spatial area, which is incapable of modeling objects' notable motion changes. A naive extension for long-term modeling is to feed distant frame's feature map into STT and greatly expand window size to capture moving object, which inevitably leads to unaffordable computational cost. Thus, we propose Reference Frame Context Enhancement (RFCE) to model the long-term relationship, as shown in the right part of Figure 2.

We first use a shared Object-Contextual Representation (OCR) module [38] to extract I_q and I_{ref}'s category-level representations, F_q^c and F_{ref}^c, from F_q^{st} and F_{ref}^p respectively. F_{ref}^p is obtained via Equation 3.

$$F_{ref}^p = f_{FE}(I_{ref}) \tag{3}$$

Then, we regard the short-term temporal representation of query frame F_q^{st} as query, the concatenation of Global Category Representations $\mathcal{G}$ (will be elaborated in Section 3.4) and category-level representation of reference frame F_{ref}^c as key and value. Next, we run multi-head self-attention on such query, key and value to generate long-term temporal representation F_q^{lt} as Equation 4. D is feature dimension.

$$\begin{aligned} Q = F_q^{st}, K = V = \text{Concat}[F_{ref}^c, \mathcal{G}], \\ F_q^{lt} = \text{Softmax}(\frac{Q * K^T}{\sqrt{D}})V, \end{aligned} \tag{4}$$

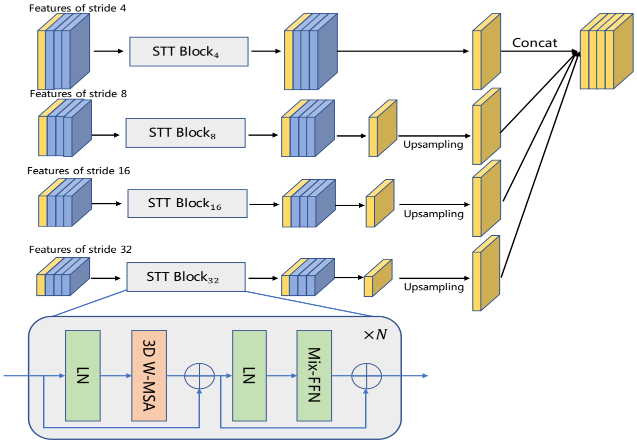


Figure 3. **Spatial-Temporal Transformer**. The feature maps of multiple scales are processed by STT blocks in parallel, then upsampled and concatenated. Features with different colors indicate that they are from different types of frames, *i.e.*, yellow and blue stand for the feature maps of query frame and adjacent frames.

Lastly, we concatenate the short-term temporal representations F_q^{st} and the long-term representations F_q^{lt} together to obtain the final enhanced representation F_q^e as Equation 5:

$$F_q^e = \text{Concat}([F_q^{lt}, F_q^{st}]), \tag{5}$$

We feed the enhanced representation F_q^e into a semantic segmentation head ϕ_{Seg} to produce segmentation result of the query frame. The result and query frame's label L_q are trained by a cross-entropy loss $\mathcal{L}_{sup}$ in Equation 6:

$$\mathcal{L}_{sup} = \text{CE}(\phi_{Seg}(F_q^e), L_q). \tag{6}$$

We briefly discuss the rationales behind such design: 1).

Long-term information from the reference frame can effectively boost the segmentation accuracy by providing richer context prior and implicitly alleviating over-fitting, as Figure 1a and 1b suggest. Performance details are shown in Table 4. 2). For F_q^{st} and F_{ref}^{p}, pixel-wise correlations modeling is extremely costly in computation, it's common to generate a compact representation for correlations modeling, either by region (*e.g.*, pooling) or category. 3). As has been shown in OCR [38], category-level compact representations usually demonstrate advantageous performance compared with region-based ones like PPM [39]. Thus OCR is exploited in our RFCE module. We note that OCR requires an auxiliary segmentation head ϕ_{Aux} to predict a coarse segmentation map. In this work, we use the representation of query frame F_q^{st} and query frame's label L_q to train ϕ_{Aux}, we refer to this training loss as $\mathcal{L}_{aux}$ in Equation 7:

$$\mathcal{L}_{aux} = \mathrm{CE}(\phi_{Aux}(F_q^{st}), L_q). \tag{7}$$

3.4. Global Category Context

Finally, we reach the question of global context modeling. As shown by our experiments (Figure 5), when query frame contains categories that are not in reference frame, self-attention mechanism might fail since the corresponding categories' information is missing from reference frame.

To address the problem, we propose a Global Category Context (GCC) module to model global category representations. The idea is to maintain a learnable set of cluster centers for each category, *i.e.*, a tensor of shape $T \times C \times D$, where T, C and D stand for the number of clusters in each category, the number of categories and the dimension of feature, respectively. We use $\mathcal{G}$ to represent this tensor and call it Global Category Representations. As shown in Figure 2, in the training phase, we first get the category representation of query frame F_q^c, then as shown in Equation 8, for each category j we extract the nearest cluster center v_j for $F_q^{c_j}$, among all clusters of that category in $\mathcal{G}$. i stands for index of clusters and $\mathcal{G}_{i,j}$ means the i^{th} cluster center of category j. Note, we use ground truth to filter out invalid categories. $\mathcal{C}_q$ is the category set of label L_q.

$$\begin{aligned} v_j &= \mathcal{G}_{k,j}, \\ where \quad j \in \mathcal{C}_q, &\quad k = \arg\min_i ||F_q^{c_j} - \mathcal{G}_{i,j}||_2. \end{aligned} \tag{8}$$

Then, we calculate the MSE loss for the category representation of the query frame for each category $F_q^{c_j}$ and its closest cluster center of the corresponding category v_j as shown in Equation 9.

$$\mathcal{L}_{MSE} = \frac{1}{C}\sum_j \mathrm{MSE}(F_q^{c_j} - v_j). \tag{9}$$

Note that $\mathcal{L}_{MSE}$ is only used to optimize the Global Category Representations $\mathcal{G}$ and does not propagate gradient to the category representations of query frame F_q^c. The Global Category Representations $\mathcal{G}$ are fixed during inference. The GCC can be viewed as a complement to RFCE, thus we concatenate their outputs before feeding into multi-head attention, as shown in Figure 2. Our final training loss is shown in Equation 10, where α and β are hyper-parameters.

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha\mathcal{L}_{MSE} + \beta\mathcal{L}_{aux}. \tag{10}$$

4. Experiments

In this section, we first present the experimental setup, then perform comparisons with existing work on two public benchmark datasets. Finally, extensive ablation studies are conducted to study each component of our model.

4.1. Experimental Setup

4.1.1 Dataset and Evaluation Metrics

We conduct our experiments on two popular datasets: Video Scene Parsing in the Wild (VSPW) [22] and CityScapes [4]. As a recently released dataset, VSPW [22] offers large-scale benchmark with well-trimmed long-temporal clips and dense annotation, hence is presumably considered as the most challenging dataset on video semantic segmentation task. Meanwhile, CityScapes [4] is another representative dataset in semantic segmentation field, and is one of the most popular benchmarks used by earlier work.

VSPW. The train, validation and test set of VSPW [22] contains 2,806/343/387 videos with 198,244/24,502/28,887 frames, respectively. Each video contains an average of 71 frames and a maximum of 482 frames. All the frames are resized to the shorter side as 480 for training and testing. Since the target of this paper is to train video semantic segmentation model in one-frame-per-video-annotation scenario, we come up with two versions of VSPW dataset:

- VSPW-SF: each training video only has its first frame annotated,
- VSPW-FULL: all frames have manual labels.

Note that all frames in the validation and test sets of both versions are annotated for performance evaluation.

CityScapes. Specifically, for each video, only the 20^{th} frame has pixel-level annotation. In total there are 5,000 labeled frames, which are divided into 2,975, 500, 1,525 images for training, validation and testing.

For VSPW dataset, we evaluate our method on four metrics: mean Intersection over Union (mIoU) [21], Weighted IoU (WIoU) [22], Temporal Consistency (TC) [13], mean Video Consistency (mVC including mVC_8 and mVC_{16}) [22].

However, since CityScapes dataset only annotates one frame per video in both training and validation sets, we cannot calculate temporal metrics such as TC and mVC. Also, WIoU is rarely used in Cityscapes dataset, thus, for the simplicity of comparison with other methods, we only report mIoU for Cityscapes experiments.

4.1.2 Implementation Details

Similar to existing work [34, 39], we adopt ResNet [8] + FPN [15] as backbone, and PPM [39] as the neck. The dimension of FPN and four STT blocks are 256 and $(32, 64, 128, 256)$, respectively. The window size of 3D W-MSA is $4 \times 7 \times 7$. We set the temporal distances from adjacent frames to the query frame as $3, 6$ and 9. The hyper-parameters α and β in Equation 10 are set as 1.0 and 0.5. The number of clusters T in global category representations $\mathcal{G}$ is set to 3, detailed experiments can be found in the supplementary.

We initialize the backbone ResNet with ImageNet pre-trained weights, and other parts of the model randomly. Then, the entire model is updated using the same training protocol as [22]. In detail, we employ SGD with momentum 0.9 to optimize our model and use the polynomial learning rate policy. During training, random scale and random crop data augmentation are adopted. On VSPW, we employ training crop size equal to 479×479 with batch size 8, and 120 training epochs. We set the initial learning rate as 0.002, weight decay as 0.0001. During testing, we conduct single-scale test and use the original image size of 480p for inference.

4.2. Results on VSPW Dataset

4.2.1 VSPW-FULL Experiments

In this setting, all frames are labeled in the training set. The purpose of this experiment is to obtain a *performance ceiling* for all methods. We regard image semantic segmentation methods as trivial baselines for our video semantic segmentation task, and compare our method with other Feature Enhancement based approaches. As shown in Table 1, Feature Enhancement based approaches demonstrate advantages in terms of all metrics (especially temporal metrics like TC and mVC), validating the necessity of exploiting temporal correlations. Our method not only surpasses all image semantic segmentation methods, but also outperforms the best Feature Enhancement based approaches IFR by 2.19% mIoU. Moreover, our method also demonstrates advantageous video stability (*i.e.*, TC and mVC) over all competing methods.

4.2.2 VSPW-SF Experiments

In this setting, only the first frame of each video has pixel-level annotation. As shown in Table 2, existing methods suffer from drastic performance drops compared with Table 1, while the mIoU of ours only decreases slightly by 0.83%, validating the strong capability of our method for one-frame-per-video-annotation scenario. As a result, the advantages of our method are enlarged than in Table 1, *i.e.*, our method outperforms the best image semantic segmentation method Segmenter by 3.63% mIoU, and the best Feature Enhancement based method approach by 2.44% mIoU.

Table 1. **VSPW-FULL experiments**. Feature Enhancement based methods demonstrate superior results than Image Semantic Segmentation ones overall. Our proposed SSLTM achieves the best performance in terms of all metrics. All methods adopt ResNet-101 [8] as the backbone.

	Method	Params	Validation Set				
			mIoU	WIoU	TC	mVC_8	mVC_{16}
Image Semantic Segmentation Methods	DeepLabv3+ [2]	62.7M	34.67	58.81	65.45	83.24	78.24
	UperNet [34]	60.9M	36.46	58.60	63.10	82.55	76.08
	PSPNet [39]	70.5M	36.47	58.08	65.89	84.16	79.63
	OCRNet [38]	58.1M	36.68	59.24	66.21	83.97	79.04
	Segmenter [29]	59.6M	37.74	58.95	61.92	80.59	74.76
Feature Enhancement based Methods	ETC [18]	58.1M	37.46	59.13	68.99	84.10	79.10
	NetWarp [5]	58.1M	37.52	58.94	68.89	84.00	78.97
	TCB [22]	58.1M	37.82	59.49	73.63	87.86	83.99
	IFR [41]	65.8M	38.43	60.04	67.13	81.39	76.03
	CFFM [30]	58.6M	38.22	58.88	67.35	82.11	76.55
	SSLTM (Ours)	62.1M	**40.62**	**61.37**	**75.81**	**87.96**	**84.16**

Table 2. **VSPW-SF experiments**. The red down arrow indicates the performance drop between training on VSPW-FULL and VSPW-SF. The proposed SSLTM has the least degradation among all methods. All of them adopt ResNet-101 [8] as the backbone.

	Method	Params	Validation Set				
			mIoU	WIoU	TC	mVC_8	mVC_{16}
Image Semantic Segmentation Methods	DeepLabv3+ [2]	62.7M	$31.09_{\downarrow 3.58}$	55.97	61.30	81.73	76.09
	UperNet [34]	60.9M	$33.92_{\downarrow 2.54}$	57.06	64.21	82.09	76.24
	PSPNet [39]	70.5M	$34.15_{\downarrow 2.32}$	57.15	64.31	82.34	76.46
	OCRNet [38]	58.1M	$34.44_{\downarrow 2.24}$	57.32	64.32	83.30	78.04
	Segmenter [29]	59.6M	$36.16_{\downarrow 1.58}$	57.83	60.18	79.03	72.87
Feature Enhancement based Methods	ETC [18]	58.1M	$36.25_{\downarrow 1.21}$	58.95	65.89	84.00	78.91
	NetWarp [5]	58.1M	$36.55_{\downarrow 0.97}$	58.22	68.06	83.42	78.13
	TCB [22]	58.1M	$36.60_{\downarrow 1.22}$	59.01	71.71	87.03	82.98
	IFR [41]	65.8M	$37.35_{\downarrow 1.08}$	58.62	64.70	79.05	73.15
	CFFM [30]	58.6M	$36.12_{\downarrow 2.10}$	57.90	65.52	81.23	75.51
	SSLTM (Ours)	62.1M	$\mathbf{39.79}_{\downarrow 0.83}$	**60.75**	**72.14**	**87.25**	**83.10**

The main reason that our SSLTM works particularly well in one-frame-per-video-annotation scenario is that we not only use STT to model short-term temporal correlation, but also introduce RFCE and GCC to provide long-term and global prior. Especially for RFCE, this module brings improvements in two perspectives: 1) Intuitively, distant reference frames offer richer contextual prior, which has great benefits for query frame's representation learning; 2) Meanwhile, by randomly sampling a distant frame as the reference frame in the training phase, RFCE implicitly involves all frames into training and effectively prevents the model from over-fitting to the labeled frames and their adjacent ones, which leads to promising generalization performance on the unseen videos in testing set.

Table 3. Comparisons on the CityScapes validation set. Our proposed SSLTM demonstrates advantageous performance over all competitors. All methods adopt ResNet-50 [8] as the backbone.

Image Semantic Segmentation Methods	Params	mIoU	Feature Enhancement Based Methods	Params	mIoU
DeepLabv3+ [2]	43.7M	76.47	ETC [18]	39.1M	77.91
UperNet [34]	41.9M	76.72	TMANet [33]	32.1M	78.50
PSPNet [39]	51.5M	76.21	IFR [41]	46.7M	78.42
Segmenter [29]	40.6M	77.89	CFFM [30]	39.6M	78.11
OCRNet [38]	39.1M	77.12	SSLTM (Ours)	43.1M	**79.69**

4.3. Results on the CityScapes Dataset

The experimental results are shown in Table 3. For the fairness of comparison, we compare all methods with the same training and testing settings, *i.e.*, batch size as 8, rand crop size as 512 × 1024, number of epochs as 100, and evaluation with single-scale strategy. Similar to VSPW results, our proposed model still achieves advantageous performance over all other methods.

4.4. Ablation Studies

We investigate the effectiveness of each component of the proposed framework on VSPW dataset and validate our design rationales. The performance achieved by different variants and parameter settings is reported as well.

4.4.1 Various number of annotated frames per video

We have demonstrated the superior performance of our model for one-frame-per-video-annotation scenario in previous experiments. A natural question to ask is: how is the trend segmentation performance w.r.t. the amount of annotated frames? We experimentally investigate this problem.

As in Figure 4a, the mIoU monotonously increases w.r.t. the number of annotated frames, yet the gains are dropping even with exponentially more labeled frames. Particularly, the mIoU tends to be saturated when 4 frames per video are labeled. Beyond a certain amount of labeled frames, the diversity of the scenes in the training data may matter more than the number of labeled frames. Thus, the strategy of one-frame-per-video labeling is a good balance between annotation cost and segmentation accuracy.

4.4.2 How much does STT contribute?

Our proposed Spatial-Temporal Transformer (STT) leverages attention mechanism to model short-term temporal correlation. As shown in Table 4, STT effectively boosts the base model's mIoU by 2.80% and mVC_8 by 4.97%. In addition, we compare STT with prior arts that concentrate on short-term modeling [5, 18, 22, 30, 41] as well.

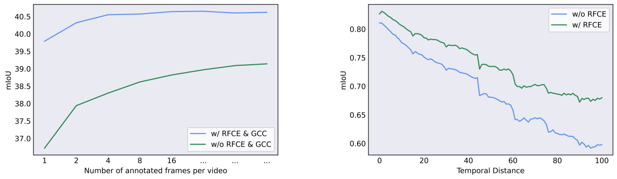
Particularly, recently proposed CFFM [41] is the most similar approach to our STT, but these two methods still differ in core motivation: CFFM applies a non-self attention mechanism for feature enhancement, *i.e.*, only query frame's feature is updated during training. Non-self attention works well with dense annotation, but suffers from the over-fitting risk in one-frame-labeled-per-video scenario [30]. As shown in Table 2, CFFM's mIoU drops the most among all Feature Enhancement based methods. However, our approach mainly targets at one-frame-labeled-per-video scenario, with feature updating both on the query and adjacent frames.

Table 4. **Ablation study on unlabeled frames of training set and validation set performance.** The performance on unlabeled frames of training set and the validation set monotonously increases as we add STT, RFCE and GCC.

STT	RFCE	GCC	Unlabeled Frames of Training Set	All Frames of Validation Set	
			mIoU	mIoU	mVC_8
			63.95	33.92	82.09
√			71.89	36.72	87.06
√	√		76.13	38.75	87.17
√	√	√	77.24	39.79	87.25

Table 5. **Compare STT with other short-term modeling modules.** UperNet [34] is adopted as the baseline framework, which obtains mIoU 33.92 and mVC_8 82.09. Our STT demonstrates superior performance.

Methods	mIoU	mVC_8	Methods	mIoU	mVC_8
+ NetWarp [5]	36.15	83.33	+ CFFM [30]	36.42	83.17
+ Temporal Loss [18]	36.04	83.54	+ IFR [41]	36.26	83.05
+ Spatial-Temporal OCR [22]	36.31	86.90	+ STT	**36.72**	**87.06**



(a) Performance with t evenly-selected annotation frames per video.

(b) Frame-level Performance on Training Set.

Figure 4. (a) Slight performance gain with exponentially increased labeling cost. (b) RFCE mitigates frame-level performance degradation on the training set.

The quantitative comparison between STT and NetWarp [5], Temporal Loss [18], Spatial-Temporal OCR [22], IFR [41], CFFM [30] demonstrates the advantages and effectiveness of STT , as shown in Table 5.

4.4.3 The Effectiveness of RFCE

Frame-level Performance on Training Set. In RFCE, we fuse the features extracted from reference and query frames to boost query frame's representation and train our SSLTM

Table 6. **Compare RFCE with QFCE**. Baseline is UperNet+STT. RFCE achieves superior performance than QFCE.

Method	Validation Set	
	mIoU	mVC_8
Baseline	36.72	87.06
+ QFCE	37.56	87.10
+ RFCE	**38.75**	**87.17**

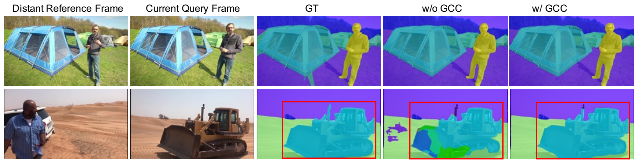


Figure 5. **Visual comparisons between w/ GCC and w/o GCC on the validation set of VSPW**. In the first row, the reference frame and the query frame are similar, hence GCC has a minor effect on the final result. While in the second row, the wheeled machine in the query frame does not appear in the reference frame. In this case, the GCC module can be viewed as a complement to the RFCE module, and boost the segmentation performance.

only with query frame's annotation. The feature fusion enables implicit optimization of features both for query and reference frames. Since reference frame is sampled across the whole video, the model is potentially trained with all frames in the training set, leading to a strong feature extractor for both labeled and unlabeled frames. As shown in Figure 4b, as the temporal distance from query (annotated) frame enlarges, the frame-level mIoU will decrease gradually. Nevertheless, the variant with RFCE demonstrates much lighter degradation, indicating that RFCE benefits the feature representation of distant frames.

Variants of RFCE. As listed in Table 4, RFCE drastically boosts the mIoU performance on validation set by 2.03%. Here we further study a variant of RFCE, which enhances the pixel representation of the input image by its own category representation. Specifically, we replace the reference frame with the query frame following the spirit of [38], thus, we term this variant as Query Frame Context Enhancement (QFCE) and investigate its performance in Table 6. Obviously, RFCE achieves significantly better performance than QFCE, validating the necessity of bringing in the distant context from the reference frame.

4.4.4 Details about Global Category Context

We illustrate the motivation of our Global Category Context (GCC) in Figure 5. As shown by the second row, when the reference frame contains irrelevant semantic categories with the query frame, GCC drastically boosts the segmentation result with the help of the global context prior. Thus, GCC can be viewed as a complement to the RFCE module.

Table 7. **Combining SSLTM with common Pseudo Label based methods.** With the help of MeanTeacher [32], the mIoU reaches as high as 40.49%, which is only 0.13% lower than the ceiling performance (*i.e.*, trained with dense annotations).

Method	Validation Set				
	mIoU	WIoU	TC	mVC_8	mVC_{16}
Ours	$39.79_{\downarrow 0.83}$	60.75	72.14	87.25	83.10
+ Naive-Student [1]	$40.15_{\downarrow 0.47}$	61.08	72.03	87.10	82.88
+ CrossPseudo [3]	$40.20_{\downarrow 0.42}$	60.97	73.19	87.32	83.49
+ MeanTeacher [32]	$40.49_{\downarrow 0.13}$	61.20	74.69	87.40	83.72

In Table 4, we demonstrate that GCC further improves the mIoU of the validation set by 1.04%.

4.5. Working with Pseudo Label based Methods

As stated in Section 2.2, our proposed SSLTM is orthogonal to Pseudo Label based methods, hence it is feasible for our SSLTM to work with existing Pseudo Label based methods as base model. As shown in Table 7, our method demonstrates better performance when working with any Pseudo Label based approach. Particularly, SSLTM with Mean Teacher [32] produces the highest mIoU at 40.49%, which is only 0.13% mIoU less than the ceiling performance (40.62% using full annotations).

4.6. Future Work

In the supplementary, we include a few examples of our method's failure modes, which are mainly caused by the very long temporal distance between reference frame and query frame, indicating our model's potential limitation on real-world untrimmed videos. How to elegantly segment extremely long videos could be an interesting direction to explore later. Furthermore, annotating one frame per video is still costly under certain circumstances, finding minimal labeling demand given certain segmentation performance requirement is worthy of future investigation as well.

5. Conclusions

In this paper, we address the video semantic segmentation problem under the setting that each video only has one frame labeled. We propose a powerful network architecture to simultaneously exploit both short- and long-term inter-frame correlations, so as to obtain high-quality representations for labeled, unlabeled and unseen frames. Particularly, the proposed SSLTM method achieves 39.79% mIoU and outperforms other state-of-the-art feature enhancement based approaches by a large margin (2% $\sim$ 3% mIoU) on the challenging VSPW dataset. When working with MeanTeacher [32], our model yields a high mIoU at 40.49%, which exhibits only 0.13% less in mIoU than the ceiling performance (*i.e.*, trained with dense annotations).

References

[1] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, 2020. 2, 3, 8

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 6, 7

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 3, 8

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 5

[5] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In *ICCV*, 2017. 2, 3, 6, 7

[6] Aditya Ganeshan, Alexis Vallet, Yasunori Kudo, Shin-ichi Maeda, Tommi Kerola, Rares Ambrus, Dennis Park, and Adrien Gaidon. Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency. In *ICCV*, 2021. 2, 3

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6, 7

[9] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *CVPR*, 2018. 1

[10] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[11] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *CVPR*, pages 13904–13914. IEEE, 2022. 3

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[13] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pages 179–195, 2018. 5

[14] Li-Jia Li and Li Fei-Fei. OPTIMOL: automatic online picture collection via incremental model learning. *IJCV*, 2010. 1

[15] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 6

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[17] Si Liu, Changhu Wang, Ruihe Qian, Han Yu, Renda Bao, and Yao Sun. Surveillance video parsing with single frame supervision. In *CVPR*, 2017. 3

[18] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, 2020. 1, 2, 3, 6, 7

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021. 3

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2, 5

[22] 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 *CVPR*, 2021. 2, 3, 5, 6, 7

[23] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *CVPR*, 2018. 2, 3

[24] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 1

[25] Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 1

[26] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *WACV/MOTION*, 2005. 1

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[28] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 3

[29] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2, 6, 7

[30] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *CVPR*, 2022. 2, 3, 6, 7

[31] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 8

[33] Hao Wang, Weining Wang, and Jing Liu. Temporal memory attention for video semantic segmentation. In *ICIP*, 2021. 3, 7

[34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 6, 7

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2

[36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3

[37] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. 1

[38] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2, 4, 5, 6, 7, 8

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 5, 6, 7

[40] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn D. Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *CVPR*, 2019. 2, 3

[41] Jiafan Zhuang, Zilei Wang, and Yuan Gao. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In *CVPR*, 2022. 1, 2, 3, 6, 7

[42] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *ICLR*, 2021. 3