# Infer from What You Have Seen Before: Temporally-dependent Classifier for Semi-supervised Video Segmentation

Jiafan Zhuang
Shantou University
jfzhuang@stu.edu.cn

Zilei Wang* Yixin Zhang
University of Science and Technology of China
{zlwang, zhyx12}@ustc.edu.cn

Zhun Fan
Shantou University
zfan@stu.edu.cn

## Abstract

*Due to high expense of human labor, one major challenge for semantic segmentation in real-world scenarios is the lack of sufficient pixel-level labels, which is more serious when processing video data. To exploit unlabeled data for model training, semi-supervised learning methods attempt to construct pseudo labels or various auxiliary constraints as supervision signals. However, most of them just process video data as a set of independent images in a per-frame manner. The rich temporal relationships are ignored, which can serve as valuable clues for representation learning. Besides, this per-frame recognition paradigm is quite different from that of humans. Actually, benefited from the internal temporal relevance of video data, human would wisely use the distinguished semantic concepts in historical frames to aid the recognition of the current frame. Motivated by this observation, we propose a novel temporally-dependent classifier (TDC) to mimic the human-like recognition procedure. Comparing to the conventional classifier, TDC can guide the model to learn a group of temporally-consistent semantic concepts across frames, which essentially provides an implicit and effective constraint. We conduct extensive experiments on Cityscapes and CamVid, and the results demonstrate the superiority of our proposed method to previous state-of-the-art methods. The code is available at https://github.com/jfzhuang/TDC.*

## 1. Introduction

As a fundamental tool, semantic segmentation has profited many downstream applications, and deep learning further boosts this area with remarkable progress. However, training a promising segmentation network relies on sufficient finely annotated data. And pixel-wise labeling is time-consuming, *e.g.*, the annotation process costs more than 1.5 hours on average for a single image in Cityscapes [6].

In real-world scenarios, we can usually collect unlabeled data conveniently and economically, especially video data. Therefore, to alleviate the label-hungry problem, a growing attention is drown on semi-supervised learning to take advantage of unlabeled data to aid model training. Besides regular supervised learning on limited labeled images, semi-supervised image segmentation (SSIS) methods proposed to construct extra supervision signals for unlabeled images, *e.g.*, consistency constraint [21, 28] and pseudo label [4, 43]. It is shown that they can bring considerable performance improvement on the challenging datasets, *e.g.*, PASCAL VOC [10] and Cityscapes [6].

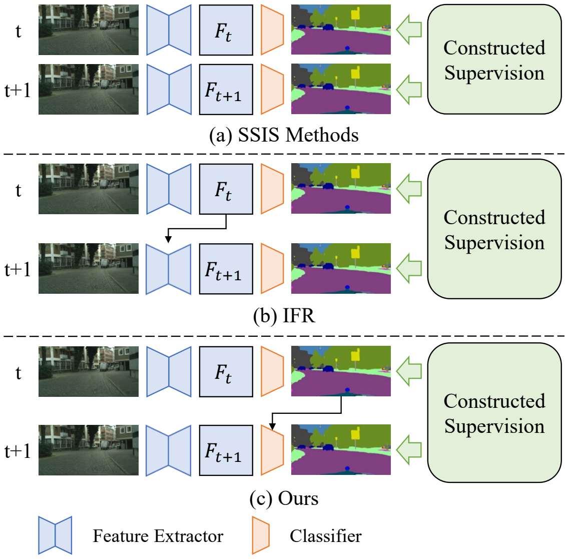

Figure 1. **Comparison between previous methods and ours.** To exploit unlabeled frames, SSIS methods attempted to construct different supervision signals. IFR [42] proposed to optimize feature extractor by inter-frame feature reconstruction. Differently, in this work, we propose to improve the recognition procedure by designing a temporal-dependent classifier.

However, when unlabeled data is scaled up by adding video data, SSIS methods can not obtain further improvement [42]. The primary reason is the homogenization of video data, *i.e.*, the contents of different frames within a

[*] Corresponding author

video are often similar, which contributes less information increment. Since SSIS methods are designed for image data and do not consider for video characteristics, they just regard videos as a set of independent images and thus fail to exploit unlabeled video data. To tackle this issue, as the first semi-supervised video segmentation (SSVS) method, IFR [42] proposed to optimize the feature extractor by inter-frame feature reconstruction. Specifically, by reconstructing the representation on labeled frame by features from other frames, the single label can indirectly provide accurate supervision for unlabeled frames within the same video. IFR essentially utilizes the content-relevance characteristic of video data and achieve a promising improvement.

In general, existing methods focus on designing supervision signals for unlabeled data and improving feature extraction, as shown in Figure 1. However, another critical component is ignored, *i.e.*, the classifier. Since video segmentation is derived from image segmentation, the final recognition process is naturally inherited in a per-frame manner. Specifically, after extracting the feature sequence for a video, the classifier is independently adopted on the features of each frame for recognition. However, we found that per-frame classification is quite different from human's recognition procedure, as shown in Figure 2. Specifically, when receiving a video stream, the recognition of the first frame is identical to image recognition. Since no prior information can be used, human would retrieve the global semantic concepts for classification, which is obtained from past learning experiences. After that, the situation becomes different. The distinguished objects and semantic concepts of historical frames are not immediately forgotten, but stored in the memory as context semantic concepts in a short period of time. Therefore, when recognizing the current frame, human would wisely retrieve both global concepts and updated context concepts. This is fundamentally different from per-frame classification. Actually, this temporally-dependent recognition procedure essentially utilizes the content relevance characteristic of video data.

Inspired by human's manner, we propose to design a novel temporal-dependent classifier (TDC), which is more suitable for video perception. Specifically, TDC contains two types of prototypes, *i.e.*, global and context prototypes. The former learns global semantic concepts during model training and fixed during inference, which is identical to the conventional classifier. Differently, the context prototypes are calculated on-the-fly based on the last processed frame. When recognizing the current frame, two types of prototypes are collaborated together to calculate similarity with the extracted feature. In this way, the distinguished context concepts from historical frames can assist the classification.

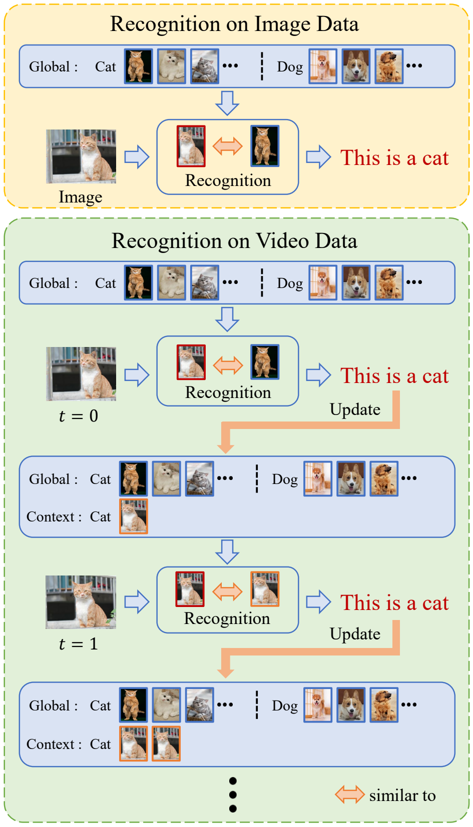

Figure 2. **Illustration of human-like recognition procedure**. When recognizing image data, human would retrieve the global semantic concepts for classification, which is obtained from past learning experiences. When recognizing the current frame of video data, besides global concepts, human would also retrieve distinguished objects from historical frames as temporal contexts.

From experimental results, we surprisingly found that TDC can achieve significant performance improvement by equipping a simple pseudo label supervision [11] in semi-supervised video segmentation task. The primary cause is that the temporal-dependent recognition paradigm implicitly imposes a temporal consistency constraint. Specifically, TDC actually requires the context prototypes calculated from historical frames to accurately represent the content of the current frame. In this way, the feature extractor would be guided to learn a group of temporally-consistent semantic concepts across frames, which indirectly introduces the content-relevance prior into the feature extraction procedure. This guidance effectively optimizes model training on unlabeled data and improve generalization ability.

We experimentally evaluate the proposed method on the Cityscapes and CamVid datasets. The results validate the effectiveness of TDC, and it can bring a significant improvement for mainstream video semantic segmentation methods. The contributions of this work are summarized as follows.

- We find that the commonly adopted per-frame classification paradigm is quite different from human's manner and sub-optimal for video perception task, which is ignored in previous studies.

- We propose a novel temporal-dependent classifier (TDC) for semi-supervised video segmentation, which can implicitly impose a temporal consistency constraint for unsupervised learning. TDC essentially utilizes the internal relevance of different frames within a video.
- We experimentally evaluate the effectiveness of our proposed methods, and the results on Cityscapes and CamVid demonstrate the superiority of our method to previous state-of-the-art methods.

## 2. Related Work

### 2.1. Image Semantic Segmentation

Image semantic segmentation aims to assign a class label to evey pixel in an image, which is a fundamental yet rather challenging task. Modern deep learning methods for semantic segmentation are mainly based on fully convolutional network (FCN) [25]. To further enhance segmentation results, the dilation convolution [2], pyramid pooling [37], attention mechanism [12, 17, 38], gating mechanism [7, 13, 22], and high-resolution architecture design [34] are proposed to model object relationships and aggregate context information. Recently, some transformer-based models [5, 24, 35] are proposed and exhibit promising performance. Despite the success of these models, they are often impeded in practical deployment due to requiring sufficient pixel-wise annotations for learning.

### 2.2. Video Semantic Segmentation

Video semantic segmentation aims to predict pixel-level semantics for each video frame. Different from static images, videos embody rich temporal information that can be exploited to improve segmentation performance. DFF [39] firstly proposes feature propagation to reuse key frame features under the guidance of estimated optical flows, which reduces the average computational cost. Inspired by DFF, Accel [19] proposes an adaptive fusion policy to effectively integrate the predictions from the key and current frames. DAVSS [41] proposes to correct the distorted features caused by inaccurate optical flow when propagating features. Besides the feature propagation paradigm, some recent works propose to improve the performance of lightweight models with temporal constraints [8, 23] and attention mechanism [16, 32, 33].

### 2.3. Semi-supervised Semantic Segmentation

To achieve good representation learning with limited annotations, semi-supervised image segmentation (SSIS) is studied by exploring unlabeled data. Existing methods can be roughly divided into three families. The adversarial based methods, *e.g.*, AdvSemiSeg [18] and S4GAN [26], utilize a discriminator to distinguish the confidence maps from labeled and unlabeled data predictions. The consistency based methods enforce the consistency of the predictions or intermediate features with various perturbations. The perturbations can be conducted on input images, *e.g.*, CutMix [11], ClassMix [27], and CAC [21], and feature space, *e.g.*, CCT [28]. The self-learning based methods generate pseudo segmentation maps on unlabeled data. [3, 14, 36] propose to generate pseudo labels in an offline manner and retrain the model iteratively. While PseudoSeg [43] and CPS [4] follow the FixMatch [31] scheme and design an online pseudo labeling mechanism.

However, SSIS methods are not designed for video data, which only regards videos as a collection of independent images. To tackle this issue, IFR [42] proposed to reconstruct feature on the labeled frame by that on other frames, which can provide accurate supervision signals from the single label for unlabeled frames. IFR essentially utilizes the content-relevance characteristic of video data.

Generally, existing methods mainly focus on supervision signal construction and improving feature extraction. Differently, in this work, we concentrate on the classifier design. We find that the per-frame recognition paradigm used in previous works is different from human's manner and thus propose a temporal-dependent classifier.

## 3. Method

### 3.1. Overview

Following IFR [42], in semi-supervised video segmentation task, we are provided with a small set of labeled videos with one frame annotated and a larger set of unlabeled videos. Let $V = \{x_1, \cdots, x_T\}$ represents $T$ frames in a video with $x_i$ as the $i^{th}$ frame with spatial resolution of $H \times W$. Let $\mathcal{D}_L = \{(V_1^L, \hat{x}_1, y_1), \cdots, (V_{n_L}^L, \hat{x}_{n_L}, y_{n_L})\}$ represent the $n_L$ labeled videos, where $\hat{x}_i$ is the annotated frame of the $i^{th}$ video, $y_i \in \mathbb{R}^{C \times H \times W}$ corresponds to the pixel-level one-hot label, and $C$ is the number of classes. Let $\mathcal{D}_U = \{V_1^U, \cdots, V_{n_U}^U\}$ represents the $n_U$ unlabeled videos. Besides, another set of labeled videos $\mathcal{D}_V = \{(\mathcal{V}_1^V, \hat{x}_1, y_1), \cdots, (\mathcal{V}_{n_V}^V, \hat{x}_{n_V}, y_{n_V})\}$ is used for performance evaluation.

As shown in Figure 3, our framework consists of three components, *i.e.*, a feature extractor $\texttt{Net}_B$, a conventional classifier $\texttt{Net}_C$ and the proposed temporal-dependent classifier $\texttt{Net}_{TDC}$. Two adopted classifiers are used in supervised and unsupervised learning branches, respectively. The reason for retaining the conventional classifier is to keep consistent with existing works [4, 11, 21, 28, 42] for fair comparison. In this way, the effectiveness of our TDC for optimizing unsupervised learning can be clearly revealed.

Our work aims to learn a segmentation model from $\mathcal{D}_L$ and $\mathcal{D}_U$, and generalize to $\mathcal{D}_V$. Following previous works, the objective of semi-supervised learning can generally be summarized as two loss functions. The first one is a regular

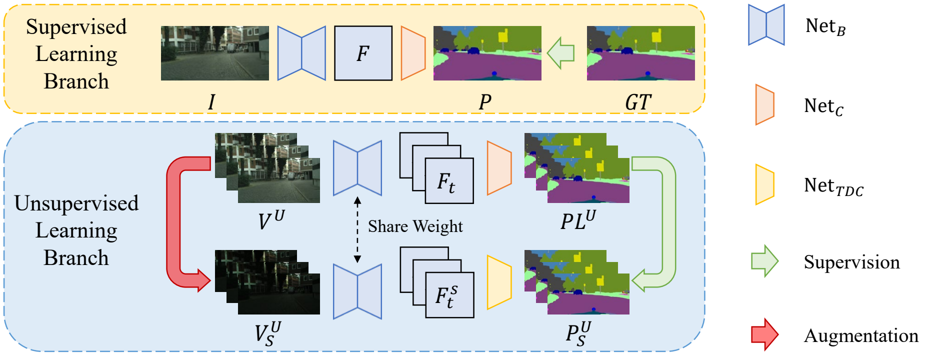

Figure 3. **Illustration of our SSVS framework**. Our framework consists of two learning branches. For supervised learning, we randomly sample an image $I$ and its corresponding ground-truth $GT$ for model training. For unsupervised learning, we randomly sample a video clip with several consecutive frames $V^U$. Following the previous work [11], we calculate pseudo labels $PL^U$ for $V^U$, which is used as supervision signals for predictions $P^U_S$ from the augmented video $V^U_S$.

cross-entropy loss on the labeled data:

$$P = \texttt{Net}_C(\texttt{Net}_B(I)), \tag{1}$$

$$L_{sup} = CrossEntropy(P, GT), \tag{2}$$

where image $I$ and its corresponding ground-truth $GT$ are randomly sampled from $D_L$. The second loss aims to train model on unlabeled data with some constructed supervision signals, which is denoted by $L_{unsup}$. In our framework, we adopt a simple pseudo label strategy similar to FixMatch [11, 31] scheme:

$$V^U_S = T_s(V^U), \tag{3}$$

$$P^U_S = \texttt{Net}_{TDC}(\texttt{Net}_B(V^U_S)), \tag{4}$$

$$PL^U = \texttt{argmax}(\texttt{Net}_C(\texttt{Net}_B(V^U))), \tag{5}$$

$$L_{unsup} = CrossEntropy(P^U_S, PL^U), \tag{6}$$

where $V^U$ is randomly sampled from $D_U$ and $T_s$ represents the strong data augmentation like color jitter. Specifically, we calculate per-frame pseudo labels $PL^U$ for $V^U$, which is used as supervision signals for predictions $P^U_S$ from the augmented video $V^U_S$. Then the overall training objective can be presented as follows

$$L = L_{sup} + \lambda L_{unsup}, \tag{7}$$

where $\lambda$ is a trade-off parameter.

## 3.2. Conventional Classifier

To make it clear, we first elaborate on the process of conventional classifier, and then explain the core mechanism of our TDC.

A conventional classifier typically consists of a fully connected layer and a softmax layer. We use $\texttt{w}_i \in \mathbb{R}^n$ to denote the weight parameters for category $i$ in the fully connected layer and use $\texttt{f} \in \mathbb{R}^n$ to denote the feature for classification. Guided by the cross-entropy loss, the classifier is encouraged to produce a larger inner product between $\texttt{f}$ and $\texttt{w}_y$, where $y$ is the corresponding ground-truth category. After model training, the learned weight parameters would automatically align with the feature cluster centers, which is experimentally verified and visualized in previous works [9, 30]. Here, the cluster center is the mean of features belonging to the same category, which is usually denoted as prototype.

This tight relationship among features, prototypes and weight parameters is the basic of our TDC, which provides a solution for utilizing distinguished semantic concepts to aid recognition.

## 3.3. Temporal-dependent Classifier

In this work, we mainly focus on designing a classifier to mimic the temporal-dependent recognition paradigm on video data like human's manner. To this end, we optimize conventional classifier by introducing distinguished results from historical frames and then propose a novel classifier, *i.e.*, the temporal-dependent classifier (TDC).

### 3.3.1 Prototypes Construction

As shown in Figure 4, when recognizing $t{+}1^{th}$ frame, TDC would simultaneously utilize two types of prototypes, *i.e.*, global prototypes $PT^G$ and context prototypes $PT^C_t$ that calculated based on the prediction of $t^{th}$ frame. Here, similar to the conventional classifier, $PT^G = \{\texttt{w}_i \in \mathbb{R}^n | i = 0, 1, \cdots, C-1\}$ is implemented as a set of learnable weight parameters, where $C$ is the number of classes and $n$ is the feature dimension. $PT^G$ is updated by back-propagation during training and keeps fixed during inference. The key

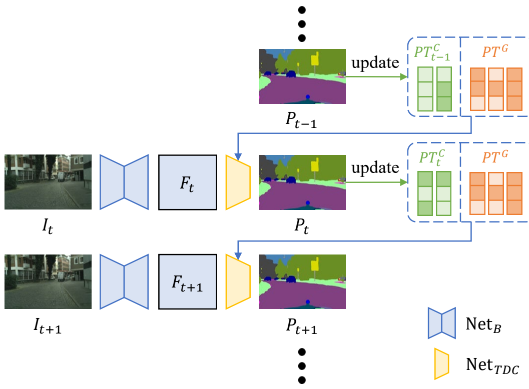

Figure 4. **Illustration of temporal-dependent recognition procedure**. The critical process is the update of context prototypes $PT^C$ based on the last frame prediction. In the classifier, context prototypes $PT^C$ and global prototypes $PT^G$ are utilized together for recognition.

of TDC is the design of context prototypes, which is calculated on-the-fly based on the prediction of last frame. Here, $PT^C_t = \{\mathtt{w}^c_k \in \mathbb{R}^n | k = 0, 1, \cdots, K-1\}$ represents $K$ prototypes calculated from the $t^{th}$ frame. Specifically, we calculate the pseudo label $PL_t$ of the extracted feature $F_t$ of the $t^{th}$ frame as

$$PL_t = \mathtt{argmax}(\mathtt{Net}_C(F_t)). \tag{8}$$

Then, we group the pixel-wise features in $F_t$ belonging to the same class $k$ according to $PL_t$, *i.e.*,

$$\mathtt{w}^c_k = \frac{\sum_i F^{(i)}_t * \mathbb{1}(PL^{(i,k)}_t == 1)}{\sum_i \mathbb{1}(PL^{(i,k)}_t == 1)}, \tag{9}$$

where $\mathbb{1}$ is an indicator function and $PL^{(i,k)}_t$ represents the one-hot pseudo label of pixel $i$ belonging to the $k^{th}$ class. In this way, context prototypes would be adaptively updated based on the last frame prediction.

### 3.3.2 Recognition Procedure

After constructing the global and context prototypes, here we introduce how to conduct recognition based on theses prototypes. The core process of recognition is to calculate similarity between the input feature and prototypes with known categories. After that, the category corresponding to the prototype that generates the maximum similarity will be regarded as the semantic prediction. Therefore, a naive solution is to calculate similarity with global and context prototypes simultaneously.

However, the input feature would naturally be similar to context prototypes since they are calculated from consecutive frames and thus global prototypes are always ignored, especially in the early stage of model training. This shortcut behavior would make the model only learning to capture similarity across frames and not being able to learn semantic concepts on unlabeled data. The cause of this issue is that two types of prototypes are updated in different mechanism and contains diverse statistical distribution like magnitudes. Therefore, we first conduct normalization on each prototype and the input feature respectively before classification, which can effectively narrow their the distribution discrepancy and achieve a fair learning process. Thus, the predicted probability of the $i^{th}$ category is computed as follows:

$$p_i = \frac{e^{\eta \cdot \langle \overline{PT}_i, \overline{F}_t \rangle}}{\sum_j e^{\eta \cdot \langle \overline{PT}_j, \overline{F}_t \rangle}}, \tag{10}$$

where $\overline{v} = v/||v||_2$ denotes the L2 normalized vector, and $\langle \overline{v}_1, \overline{v}_2 \rangle = \overline{v}^T_1 \overline{v}_2$ measures the cosine similarity between two normalized vectors. Here, $PT = \{PT^G, PT^C_t\}$ is the combination of two types of prototypes. Inspired by a previous work [15], since the range of $\langle \overline{v}_1, \overline{v}_2 \rangle$ is restricted to $[-1, 1]$, a learnable scalar $\eta$ is introduced to control the peakiness of softmax distribution. Differently, we set two independent scalars, *i.e.*, $\eta_G$ and $\eta_C$, for global prototypes $PT^G$ and context prototypes $PT^C_t$ respectively.

### 3.3.3 Joint-prototypes Label Relaxation

After calculating the probability of input feature based on constructed prototypes, a cross-entropy loss is applied according to a pseudo label, as described in Section 3.1. However, there is a mismatch between the probability vector $p \in \mathbb{R}^{C+K}$ and the one-hot label $y \in \mathbb{R}^C$, where $C$ and $K$ are the number of categories and prototypes predicted in the last frame. Intuitively, for a specific category $i$, its corresponding context prototype $PT^C_{t,i}$ and global prototype $PT^G_i$ are different descriptions for an identical semantic concept. Therefore, for a feature belonging to category $i$, it should not be penalized by loss function regardless of which prototype it is more similar to.

Based on this consideration, we propose a joint-prototypes label relaxation, which is also inspired by a previous work [40]. Suppose we are classifying a feature belonging to class $i$ and class $i$ is related to two prototypes $PT^G_i$ and $PT^C_{t,i}$. To achieve label relaxation, we propose to maximize the likelihood of both prototypes together instead of any single one. Since two prototypes are mutually exclusive in cross-entropy loss, we aim to maximize the union of them:

$$P(PT^G_i \cup PT^C_{t,i}) = P(PT^G_i) + P(PT^C_{t,i}), \tag{11}$$

where $P$ is the softmax probability of each prototype. Therefore, for each pixel $i$ in the $t^{th}$ frame in the video clip, the loss function is as follows:

$$L_{TDC} = \frac{1}{H*W} \frac{1}{T} \sum_i^{H*W} \sum_t^{T} L_{TDC}(i, t), \tag{12}$$

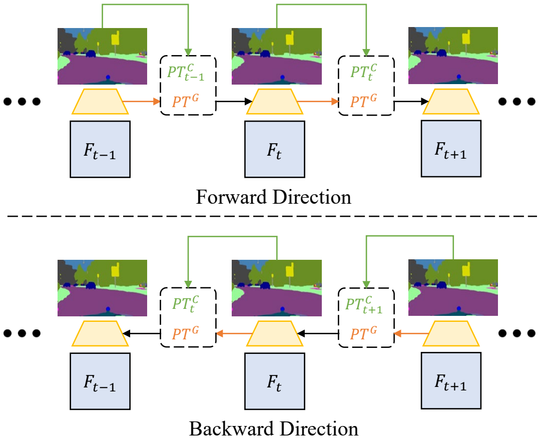

Figure 5. **Illustration of bidirectional prediction**. We simultaneously considering forward ($t \rightarrow t+1$) and backward ($t+1 \rightarrow t$) directions when constructing context prototypes.

$$L_{TDC}(i,t) = -logP(PT_{y_{i,t}}^G \cup \mathcal{X}(PT_{t-1,y_{i,t}}^C)), \quad (13)$$

$$\mathcal{X}(x) = \begin{cases} x, & if\ x\ exists \\ 0, & else \end{cases} \quad (14)$$

To be noticed, when recognizing the first frame, only global prototypes are adopted since context prototypes has not been constructed. This is reasonable and similar to human's manner. When obtaining the first frame of a video, human can only conduct image recognition by recalling learned experience, *i.e.*, global prototypes, since there is no prior information about this video.

### 3.3.4 Bidirectional Constraint

Actually, TDC works in a RNN-like manner, where the output from the previous step can affect the prediction on the current step. In TDC, the continuously updated prototypes serve like the hidden state vectors in RNN structure. According to successfully practices [20, 29] in NLP field, comparing to unidirectional RNN, bidirectional RNN can help the model to better exploit contextual information and effectively improve the model performance.

Therefore, in this work, we further propose a bidirectional constraint, *i.e.*, simultaneously considering forward ($t \rightarrow t+1$) and backward ($t+1 \rightarrow t$) directions, as shown in Figure 5. Specifically, context prototypes are calculated based on the $t+1^{th}$ frame to aid the current frame recognition:

$$L_{Bi-TDC} = \frac{1}{H*W}\frac{1}{T}\sum_{i}^{H*W}\sum_{t}^{T} L_{Bi-TDC}(i,t), \quad (15)$$

$$L_{Bi-TDC}(i,t) = -logP(PT_{y_{i,t}}^G \cup \mathcal{X}(PT_{t+1,y_{i,t}}^C)). \quad (16)$$

Actually, the primary cause behind the effectiveness of TDC on semi-supervised learning is that TDC can implicitly impose a temporal consistency constraint across frames. Intuitively, the bidirectional calculation can further enhance the constraint by simultaneously considering both forward and backward consistency. Finally, in this work, the adopted unsupervised learning loss is as follows:

$$L_{unsup} = L_{TDC} + L_{Bi-TDC}. \quad (17)$$

## 4. Experiments

### 4.1. Dataset

**Cityscapes** [6] is a representative dataset in semantic segmentation and autonomous driving domain. It focuses on semantic understanding to urban street scenes. The training and validation subsets contain $2,975$ and $500$ videos, respectively, and each video contains 30 frames at a resolution of $1024 \times 2048$. The $20^{th}$ frame in each is annotated by pixel-level semantic labels with 19 categories.

**CamVid** [1] also focuses on the semantic understanding to urban street scenes, but it contains less data than Cityscapes. It has four driving videos and each video contains frames ranging from $3,600$ to $11,000$ at a resolution of $720 \times 960$. Every $30^{th}$ frame of videos is annotated with 11 semantic classes, which results in a total of 701 samples. Similar to Cityscapes, we split videos into 701 videos and each video contains 30 frames. All videos are divided into the trainval set with 468 videos and test set with 233 videos.

We follow the partition protocols of IFR [42] and divide the whole training set via randomly sub-sampling 1/2, 1/4, 1/8, 1/16, and 1/30 of all training videos, *i.e.*, 2975 videos in Cityscapes and 468 videos in CamVid, as the labeled set and regard the remaining videos as the unlabeled set. Following previous work [11, 18, 26–28, 42], we apply bilinear interpolation to resize every video frame in Cityscapes and CamVid to $512 * 1024$ and $360 * 480$ for the efficiency of training and inference.

### 4.2. Implementation Details

Following IFR [42], we adopt Accel [19] as video semantic segmentation architecture due to its good performance. It consists of two segmentation branches, *i.e.*, a heavy reference branch and a light-weight update branch, an optical flow network, and a score fusion layer. Similar to IFR, we adopt semi-supervised methods for two segmentation branches to improve representation learning with limited annotations. The segmentation model is trained using a SGD optimizer with a momentum of 0.9 and a weight decay of $10^{-4}$. The learning rate is set at $10^{-3}$ for the backbone parameters and $10^{-2}$ for others, which is annealed following the poly learning rate policy.

We evaluate the segmentation performance on the validation videos, *i.e.*, validation subset in Cityscapes and test

| Method | 1/30 (100) | 1/16 (186) | 1/8 (372) | 1/4 (744) | 1/2 (1488) |
|---|---|---|---|---|---|
| Accel | 45.73 | 52.10 | 57.12 | 60.55 | 62.83 |
| + CCT [28] | 48.05 (+2.32) | 53.25 (+1.15) | 58.88 (+1.76) | 62.00 (+1.45) | 64.02 (+1.19) |
| + CAC [21] | 48.83 (+3.10) | 54.56 (+2.46) | 58.55 (+1.43) | 62.78 (+2.23) | 63.87 (+1.04) |
| + CPS [4] | 48.97 (+3.24) | 54.69 (+2.59) | 58.97 (+1.85) | 62.43 (+1.88) | 63.74 (+0.91) |
| + IFR [42] | 52.86 (+7.13) | 56.39 (+4.29) | 60.08 (+2.96) | 63.45 (+2.90) | 64.53 (+1.70) |
| + Ours | **55.81 (+10.08)** | **60.97 (+8.87)** | **63.12 (+6.00)** | **65.83 (+5.28)** | **66.67 (+3.84)** |

Table 1. **Comparison with state-of-the-art methods on the Cityscapes validation subset** under different partition protocols. Accel is adopted as the supervised baseline that is only trained on labeled data. Our method gets more gains especially for few labeled training data.

| Method | 1/30 (15) | 1/16 (29) | 1/8 (58) | 1/4 (117) | 1/2 (234) |
|---|---|---|---|---|---|
| Accel | 42.37 | 47.57 | 50.78 | 56.40 | 59.37 |
| + CCT [28] | 47.09 (+4.72) | 52.45 (+4.88) | 54.50 (+3.72) | 58.69 (+2.29) | 61.58 (+2.21) |
| + CAC [21] | 46.85 (+4.48) | 52.16 (+4.59) | 54.03 (+3.25) | 59.67 (+3.27) | 63.17 (+3.80) |
| + CPS [4] | 46.05 (+3.68) | 52.04 (+4.47) | 55.30 (+4.52) | 59.02 (+2.62) | 62.49 (+3.12) |
| + IFR [42] | 49.50 (+7.13) | 53.71 (+6.14) | 57.37 (+6.59) | 61.27 (+4.87) | 63.86 (+4.49) |
| + Ours | **50.95 (+8.58)** | **55.36 (+7.79)** | **59.88 (+9.10)** | **62.63 (+6.23)** | **64.59 (+5.22)** |

Table 2. **Comparison with state-of-the-art methods on the CamVid test** subset under different partition protocols. Accel is adopted as the supervised baseline that is only trained on labeled data. Our method gets more gains especially for few labeled training data.

set in CamVid. Following Accel, for each test video, we conduct the reference branch on a selected key frame and update branch on the annotated frame. The segmentation results are predicted via the feature propagation and score fusion. To be notice, for fair comparison, we adopt the conventional classifier for semantic prediction while TDC is only involved in the training period, since in this work we mainly investigate the effectiveness of TDC on representation learning. We evaluate different methods with mean Intersection-over-Union (mIoU) as metric. The key frame interval is set as 5 throughout the experiments. For our TDC, there is a trade-off parameter $\lambda$ in Eq. 7, which is set $\lambda = 0.2$ for all experiments.

## 4.3. Performance Comparison

To demonstrate the superiority of our method, we make a comparison with recent state-of-the-art methods, *i.e.*, three SSIS methods including CCT [28], CAC [21] and CPS [4], and a SSVS method IFR [42]. Following IFR, for fair comparison, all critical components are keep consistent among different methods, including base segmentation model, data splits, data augmentation and training settings (*i.e.*, optimizer and hyperparameters). In this way, we can fairly compare the improvement brought by different methods on the basis of the same supervised learning baseline.

The comparison on Cityscapes and CamVid is shown in Table 1 and Table 2. From the results, we have the following two observations. First, our TDC can bring significant improvement under all partition protocols comparing to the supervised learning baseline. A larger gain is achieved for the cases with less labeled data, *e.g.*, $10.08\%$ mIoU gain with 100 samples in Cityscapes. Such experimental results well verify that TDC can effectively improve the generalization of models. Second, our method surpasses other state-of-the-art methods by a large margin. For example, it outperforms IFR by $4.58\%$ under 1/16 partition protocol on Cityscapes and $2.51\%$ under 1/8 partition protocol on CamVid, respectively. It shows that TDC can better utilize the unlabeled video data for training model.

| $L_{sup}$ | $L_{TDC}$ | | $L_{Bi-TDC}$ | mIoU (%) |
|---|---|---|---|---|
| | $PT^G$ | $PT_t^C$ | | |
| √ | | | | 43.68 |
| √ | √ | | | 46.78 (+3.10) |
| √ | √ | √ | | 50.14 (+6.46) |
| √ | √ | √ | √ | **51.29 (+7.61)** |

Table 3. **Ablation study on our proposed components**. Each component can bring performance improvement compared to the baseline, and their combination performs best.

## 4.4. Ablation Study

In this subsection, we conduct experiments to reveal the effectiveness of our proposed method. All experiments are particularly conducted with 1/30 labeled data on Cityscapes. For efficient training, we adopt PSPNet with ResNet18 as the segmentation network by default.

**Effect of Components.** To reveal the contribution of our proposed components, *i.e.*, global prototypes $PT^G$, context prototypes $PT_t^C$ and bidirectional constraint $L_{Bi-TDC}$, we conduct an extensive study by evaluating their combina-

| $\lambda$ | 0.0 | 0.05 | 0.1 | 0.2 | 0.5 | 1.0 |
|---|---|---|---|---|---|---|
| mIoU (%) | 43.68 | 46.87 | 49.62 | **51.29** | 50.74 | 48.54 |

Table 4. **Effect of trade-off parameter** $\lambda$. Based on the results, we set $\lambda = 0.2$ for all experiments.

| $T$ | 0 | 1 | 2 | 3 | 5 | 7 | 10 |
|---|---|---|---|---|---|---|---|
| mIoU (%) | 43.68 | 46.87 | 49.24 | 50.18 | **51.29** | 50.93 | 50.64 |

Table 5. **Effect of Number of frames**. Based on the results, we sample 5 frames per unlabeled video clip for all experiments.

tions, and the results are shown in Table 3. It can be seen that each component can bring performance improvement compared with the baseline. In particular, based on the supervised learning baseline, TDC equipped with only global prototypes serves like a pseudo-label scheme similar to previous works [4, 43]. After introducing context prototypes, TDC can obtain significant improvement. Finally, TDC performs best with bidirectional constraint.

**Effect of Trade-off Parameter $\lambda$.** We investigate the influence of trade-off parameter $\lambda$ used to control the unsupervised learning loss in Eq 7, as shown in Table 4. Here, $\lambda = 0.0$ represents the supervised learning baseline. The results show that: with the increase of $\lambda$, the unsupervised loss help model training on unlabeled data and effectively overcome the overfitting issue due to limited labeled training data, which performs best when $\lambda = 0.2$. When $\lambda$ continuously increases, the performance degrades since the supervised learning branch is interfered. Therefore, based on the results, we set $\lambda = 0.2$ for all experiments.

**Effect of the Number of frames.** In our framework, we randomly sample a unlabeled video clip with $T$ frames for unsupervised learning. Here, we investigate the influence of number of frames, which relates to the temporal modeling in our proposed TDC. Table 5 shows the results. Here, $T = 0$ represents the supervised learning baseline and $T = 1$ represents using TDC with only global prototypes. From the results we can see that: as $T$ increases, TDC can impose a stronger temporal consistency constraint and improve model training on unlabeled video data. However, if $T$ gets too large, TDC would suffer from training unstability and lead to a slight performance degradation. Therefore, based on the results, we set $T = 5$ for all experiments.

**Performance with Different VSS Architectures.** To study the generalization ability of our method, We apply it to different video semantic segmentation architectures. We particularly adopt three widely adopted video segmentation architectures, *i.e.*, DFF [39], Accel [19] and DAVSS [41]. As shown in Table 6, our proposed method can bring significant performance improvement consistently.

| Method | Baseline | +IFR | + Ours |
|---|---|---|---|
| DFF [39] | 44.19 | 49.97 (+5.78) | **52.63 (+8.44)** |
| Accel [19] | 45.73 | 52.86 (+7.13) | **55.81 (+10.08)** |
| DAVSS [41] | 46.51 | 51.97 (+5.46) | **55.25 (+8.74)** |

Table 6. **Performance on different VSS architectures**. Our method can bring significant improvement for different video semantic segmentation architectures consistently.

| Method | Accel | + CCT | + CAC | + CPS | + IFR | + Ours |
|---|---|---|---|---|---|---|
| TC (%) | 70.04 | 71.43 | 71.47 | 70.74 | 73.88 | **75.38** |

Table 7. **Comparison on Temporal Consistency**. We evaluate different methods with temporal consistency (TC) score [23]. Our proposed method can bring a significantly improvement.

**Effectiveness on Temporal Consistency.** It is important for VSS methods to produce temporally stable predictions. Thus, we further verify the effectiveness of our method on improving temporal consistency. Particularly, we follow previous works [23, 42] and adopt temporal consistency (TC) score as the evaluation metric. As shown in Table 7, our method can bring a significant improvement, since consistent predictions among different frames can be achieved.

## 5. Conclusion

In this paper, we focus on the semi-supervised video semantic segmentation problem, and propose a novel temporal-dependent classifier (TDC). Motivated by the observation of human's behavior, TDC is designed to utilize distinguished semantic concepts of historical frames when recognizing the current frame, which is more suitable for video data. We found that TDC can achieve significant performance improvement in semi-supervised learning since its recognition mechanism implicitly imposes a temporal-consistency constraint across frame, which is valuable especially in label-scarce scenarios. To further enhance the effect of TDC, we propose a bidirectional constraint by simultaneously considering both forward and backward calculation. Extensive experiments on Cityscapes and CamVid validated the effectiveness of our method, which outperforms previous state-of-the-art methods.

## Acknowledge

This work is supported in part by the National Key R&D Program of China (grant numbers 2021ZD0111501, 2021ZD0111502), the National Natural Science Foundation of China (grant number 62176246), the Anhui Province Key Research and Development Plan (grant number 202304a05020045), and the STU Scientific Research Foundation for Talents (grant number NTF22030).

## References

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2009. 6

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 3

[3] 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. 3

[4] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 1, 3, 7, 8

[5] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 3

[6] 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, 6

[7] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 3

[8] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every frame counts: joint learning of video segmentation and optical flow. In *AAAI*, 2020. 3

[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 4

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1

[11] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2019. 2, 3, 4, 6

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 3

[13] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *ICCV*, 2019. 3

[14] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, 2021. 3

[15] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 5

[16] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *CVPR*, 2020. 3

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 3

[18] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018. 3, 6

[19] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *CVPR*, 2019. 3, 6, 8

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 6

[21] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021. 1, 3, 7

[22] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. Gated fully fusion for semantic segmentation. In *AAAI*, 2020. 3

[23] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, 2020. 3, 8

[24] 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. *ICCV*, 2021. 3

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[26] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *TPAMI*, 2019. 3, 6

[27] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021. 3

[28] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 1, 3, 6, 7

[29] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018. 6

[30] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018. 4

[31] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 3, 4

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

[33] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. In *ECCV*, 2022. 3

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 3

[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. *arXiv preprint arXiv:2105.15203*, 2021. 3

[36] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *ICCV*, 2021. 3

[37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 3

[38] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 3

[39] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017. 3, 8

[40] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *CVPR*, 2019. 5

[41] Jiafan Zhuang, Zilei Wang, and Bingke Wang. Video semantic segmentation with distortion-aware feature correction. *TCSVT*, 2020. 3, 8

[42] 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, 8

[43] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 1, 3, 8