Multi-Granularity Context Network for Efficient Video Semantic Segmentation

Zhiyuan Liang, Xiangdong Dai, Yiqian Wu,
Xiaogang Jin, *Member, IEEE*, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*—Current video semantic segmentation tasks involve two main challenges: how to take full advantage of multi-frame context information, and how to improve computational efficiency in video processing. To tackle the two challenges simultaneously, we present a novel Multi-Granularity Context Network (MGCNet) by aggregating context information at multiple granularities in a more effective and efficient way. Our method first converts image features into semantic prototypes, and then conducts a non-local operation to aggregate the per-frame and short-term contexts jointly. An additional long-term context module is introduced to capture the video-level semantic information during training. By aggregating both local and global semantic information, a strong feature representation is obtained. The proposed pixel-to-prototype non-local operation requires less computational cost than traditional non-local ones, and is video-friendly since it reuses the semantic prototypes of previous frames. Moreover, we propose an uncertainty-aware and structural knowledge distillation strategy to boost the performance of our method. Experiments on Cityscapes and CamVid datasets with multiple backbones demonstrate that the proposed MGCNet outperforms other state-of-the-art methods with high speed and low latency.**

***Index Terms*—video semantic segmentation, light-weight networks, non-local operation.**

I. Introduction

Video semantic segmentation aims to assign pixel-wise semantic masks to video sequences. It is fundamental for many vision applications, such as autonomous driving [6], [28] and robot sensing. Some lightweight semantic segmentation algorithms [22], [33], [46], [24], [41] can satisfy the real-time requirement, but they ignore the context information among video frames, which hampers their performance on the video task.

Recent video semantic segmentation algorithms can be classified into two categories. As shown in Fig. 1, one category propagates semantic information to capture the global dependencies frame by frame [36], [26], [19], [38], [50], but the performance cannot be guaranteed due to the cumulative predicted error. The other category resorts to the local context aggregation. These algorithms aggregate deep features or semantic outputs of several nearby frames via optical flows [27], [30], attention modules [15], or recurrent networks [32]. However, the context interaction of local frames is not enough in the cases of motion blur or short-term occlusions. How to aggregate multi-frame context information to achieve high accuracy at a low inference cost is still a challenging problem in video semantic segmentation.

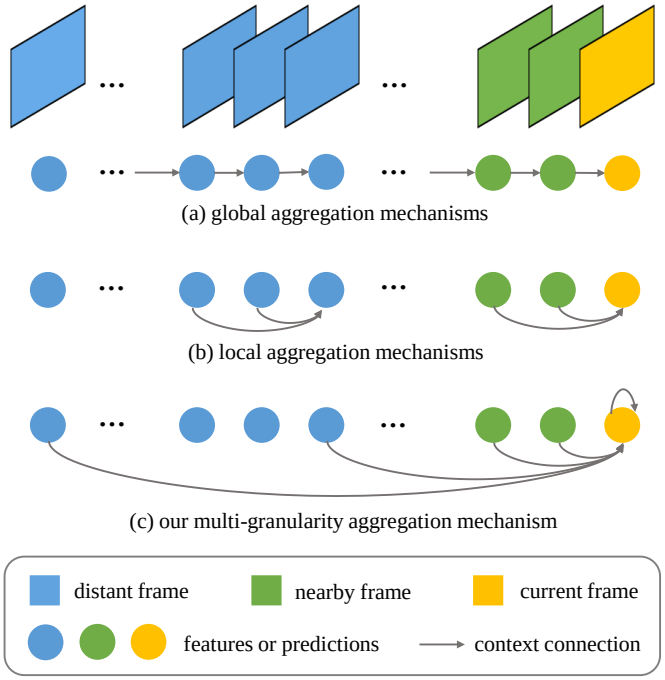


Fig. 1. Illustration of different context aggregation mechanisms in video semantic segmentation methods. (a) propagates the context information frame by frame. (b) extracts the multi-frame context only in nearby frames. (c) is the proposed method, which takes full consideration of both global and local information by aggregating contexts at different granularities.

To address this problem, we present a Multi-Granularity Context Network (MGCNet) and divide the video context information into three granularities, *i.e.*, the per-frame, the short-term, and the long-term contexts. As shown in Fig. 1 (c), these three granularities of context contain semantic information in the current frame, nearby frames, and distant frames, respectively. With the carefully-designed network architecture, our MGCNet can extract context information at different granularities efficiently. As a result, both the global semantic information and the local geometry information can be captured by our method.

Z. Liang is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P. R. China. (email: liangzhiyuan@bit.edu.cn)

X. Dai is with Guangdong OPPO Mobile Telecommunications Corp., Ltd. (email: daixiangdong@oppo.com)

Y. Wu and X. Jin are with State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, P. R. China. (email: onethousand1250@gmail.com, jin@cad.zju.edu.cn)

J. Shen is with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (email: shenjianbingcg@gmail.com)

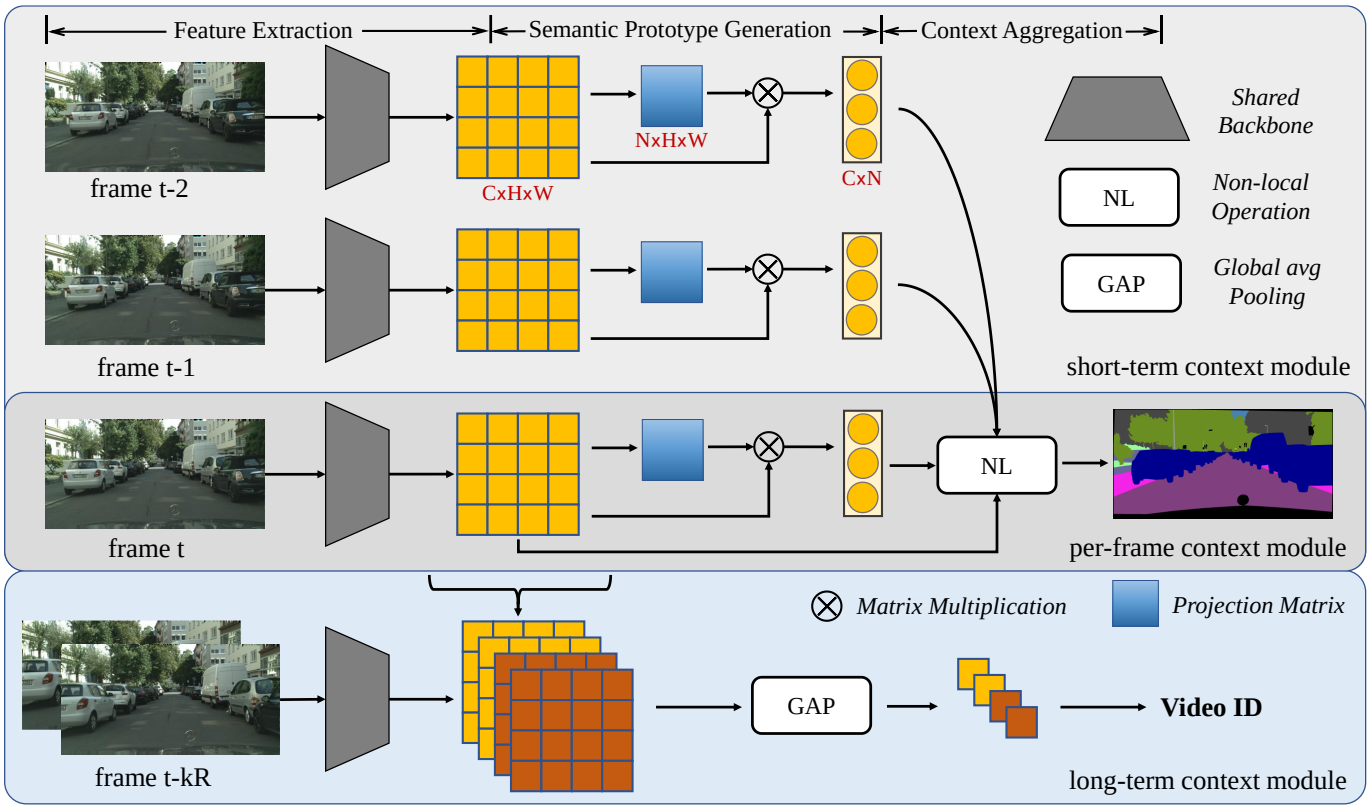


Fig. 2. The network architecture of the MGCNet. It adopts a shared backbone to extract image features at each frame and then leverages the per-frame, short-term, and long-term context modules to capture context information at different granularities during training. The pixel-to-prototype non-local operation is proposed for both per-frame and short-term context modules. The long-term context module is removed to maintain the high speed at the inference time. R is the frame interval for the distant frame selection.

We propose a pixel-to-prototype non-local operation to jointly aggregate the per-frame and short-term contexts. It first assigns pixels with similar features to the same semantic prototype via a learnable projection matrix and then conducts the non-local operation between original features and semantic prototypes. Compared to most previous non-local operations in semantic segmentation [47], [7], [44], [39], the pixel-to-prototype one dramatically reduces the computational costs and can capture the long-range dependencies beyond regular grids. To further reduce the redundant information among semantic prototypes, a diversity loss is designed to force them to focus on different spatial regions adaptively. We also propose a long-term context module to compensate for the long-term context. It learns to capture video-level semantic information, which helps improve the performance when motion blur or partial occlusion occurs in the local frames. At the inference time, the long-term context module is removed to avoid the cumulative predicted errors from the distant frames. Thanks to the per-frame, short-term, and long-term context modules, our method can generate a discriminative and complementary representation for video semantic segmentation.

To further boost the accuracy, an uncertainty-aware and structural knowledge distillation strategy is proposed. Different from the previous knowledge distillation strategies in video semantic segmentation which treat the correlations between complex and compact models equally at every location, we estimate the uncertainty of the complex model and then use it to guide the distillation. Thus the importance of uncertain predictions of the complex model is reduced to suppress additional errors. In addition to the pixel-wise correlations, we also calculate the prototype-wise correlations. Since the semantic prototypes aggregate pixels with similar features, they implicitly contain the semantic and geometry information. By modeling both pixel-wise and prototype-wise correlations, extra context and structural knowledge can be learned from the complex model. To verify the effectiveness of our approach, we adopt different lightweight networks (*i.e.*, MobileNetV2, ResNet18, and ResNet50) as the backbone of our compact model, and evaluate the performance on Cityscapes and CamVid datasets. At each inference step, the long-term context module is removed and the semantic prototypes of previous frames can be reused to further reduce the computational redundancy.

The main contributions of our paper are summarized as follows:

- We propose the MGCNet for efficient video semantic segmentation by capturing context information at multiple granularities. Our approach leverages the local and global semantic information from the complementary per-frame, short-term, and long-term contexts. By removing the long-term context module and reusing semantic prototypes of previous frames, the computational cost is further

reduced at the inference time.

- We propose a novel pixel-to-prototype non-local operation to extract the per-frame and the short-term contexts efficiently. It can capture long-range dependencies beyond regular grids with less computational complexity. Our approach reduces the redundancy among semantic prototypes by adopting a diversity loss. Besides, a long-term context module is proposed to extract video-level context information.
- We introduce an uncertainty-aware and structural knowledge distillation strategy to further boost the performance of our approach. With the uncertainty map of complex models, additional errors caused by the uncertain inference are suppressed. By modeling both pixel-wise and prototype-wise correlations, the context and structural knowledge from complex models can be propagated to compact ones.
- Experimental results on the Cityscapes and CamVid datasets indicate that the proposed MGCNet outperforms the state-of-the-art video semantic segmentation methods with a relatively high inference speed and a low latency.

II. Related Work

In this section, the previous works related to video semantic segmentation, non-local operation, and knowledge distillation are briefly reviewed.

A. Video Semantic Segmentation

Video semantic segmentation aims at pixel-wise dense labeling for all frames of a video sequence. Previous works focus on leveraging the temporal information from multiple frames [20], [15], [32], or improving the trade-off of accuracy and speed by selecting keyframes and reusing the previous high-level context information [36], [30], [26], [19]. CLK [36] introduces the clock signals to reuse the feature maps directly. The segmentation networks are manually divided into different stages which are skipped according to the clock signals during testing. PEARL [20] proposes a unified framework to address the semantic segmentation and the frame prediction jointly, and the additional semantic information is learned via the auxiliary task. LVS [26] adaptively propagates and fuses the semantic information with the spatially variant convolution. It can dynamically select the keyframes based on the predictive accuracy of segmentation models. Accel [19] presents a two-branch network to extract high-detail features at keyframes and warps these features to the rest frames with optical flow, leading to a high inference speed on average. However, these methods also require a heavy computational cost at the keyframe, which prevents them from being used in autonomous driving applications. Recently, some methods have been proposed to reduce the per-frame maximum latency in video processing. TDNet [15] leverages the inherent temporal continuity by distributing sub-networks over sequential frames. It achieves a balanced latency while the usage of multiple sub-networks results in a large number of parameters of segmentation models. ETC [30] presents an efficient temporal consistency loss to enhance the performance of image segmentation models with video sequences. Without additional computation overhead in the inference phase, this training framework can explicitly improve the consistency among frames. In this paper, we divide the semantic contexts into different granularities and leverage the multi-granularity context information to solve the video semantic segmentation problem. Furthermore, an uncertainty-aware and structural knowledge distillation strategy is proposed to further boost the accuracy while maintaining the speed of our method.

B. Non-local Operation

Many non-local operations [13], [37] have been proposed to capture the long-range dependencies in many computer vision tasks, such as video classification [37], [14], object detection [13], [10], [3], and semantic segmentation [7], [17], [51], [40]. The relation network [13] combines the semantic features with the localization features to capture the relation between object proposals. PSANet [47] and OCNet [44] apply non-local blocks to model spatial attention. DANet [7] further introduces self-attention modules at both spatial and channel dimensions. DNL [39] splits the attention computation into a whitened pairwise term and a unary term to facilitate learning for both terms. Although these pixel-to-pixel non-local operations are proved to be effective, the prohibitive computation prevents them from being widely used for real-time scenes directly. Recently, some works focus on simplifying and speeding up the traditional non-local operation. CCNet [17] presents a recurrent criss-cross non-local attention module. Each pixel just interacts with other ones within the same columns or rows of the feature maps. ANN [51] leverages a pyramid sampling module to reduce the computation complexity. It first aggregates the local contexts in regular grids then distributes them to every pixel. RGNN [40] learns to aggregate the context information beyond the regular grids. A spatial offset matrix is trained to adaptively select proper candidates for each pixel. OCRNet [43] aggregates category-level prototypes for feature augmentation. The number of prototypes is based on the category number of the dataset. Different from OCRNet, our prototype generation module is trained in an unsupervised way and more suitable for the video task, in which not all the training data are labeled. Besides, the total number of prototypes is more flexible than OCRNet, which can satisfy various cases in terms of computational costs. EMANet [25] formulates the non-local attention in an expectation-maximization manner to reduce the computational costs. Compared to EMANet, our method avoids iterative attention mechanisms and can be modulated with the proposed diversity loss. In this paper, we propose an efficient pixel-to-prototype non-local operation to aggregate the per-frame and short-term contexts among nearby frames.

C. Knowledge Distillation in Semantic Segmentation

Knowledge distillation [12] has been proved effective for performance improvement in classification tasks. In knowledge distillation, the teacher model often has a large volume of

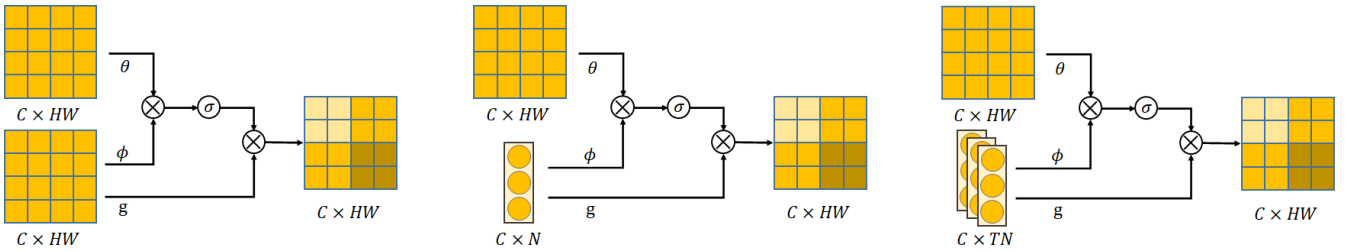


Fig. 3. The architectures of pixel-to-pixel non-local operation, pixel-to-prototype non-local operation, and multi-frame pixel-to-prototype non-local operation. θ, ϕ, and g are three linear transformations. $\otimes$ indicates the matrix multiplication. σ indicates the *Softmax* operation.

parameters to achieve accuracy saturation. The student model is more compact than the teacher model for higher inference speed. During training, the predicted masks or the intermediate features of the teacher model can be viewed as the soft guidance of the student model. Previous methods [11], [29] utilize knowledge distillation strategies to enhance the segmentation accuracy while ignoring the temporal information of video clips. To capture the temporal information, TDNet [15] proposes a grouped knowledge distillation strategy to provide the soft guidance in both full-feature space and sub-feature space. ETC [30] leverages an optical estimation method to model the pixel-wise motion information between two nearby frames and improves the temporal consistency via knowledge distillation. In this paper, we propose an uncertainty-aware and structural knowledge distillation strategy for video semantic segmentation. It estimates the uncertainty map to dynamically assign the loss weight of each pixel. In addition to learning the pixel-level knowledge, the prototype-level knowledge containing semantic and structural information is also considered during training. By utilizing multi-level knowledge jointly, the proposed distillation strategy can improve the accuracy of our method without extra computational costs during the inference phase.

III. Approach

In this section, we elaborate on the proposed MGCNet for efficient video semantic segmentation, which is illustrated in Fig. 2. We first introduce the per-frame context module to aggregate the intra-frame context by the pixel-to-prototype non-local operation (see § III-A), then we extend the aggregation size of the non-local operation to nearby frames and introduce the short-term context module for inter-frame context aggregation (see § III-B). To capture video-level semantic contexts, the long-term context module is applied (see § III-C). Finally, the uncertainty-aware and structural knowledge distillation strategy is employed to further boost the accuracy of our approach (see § III-D).

A. Per-frame Context Module

Following the FCN-based algorithms, we feed the image into basic networks to extract the per-frame features, and then introduce a global context module for the intra-frame feature augmentation. Existing algorithms [44], [7], [17] adopt the non-local attention module to generate the global context by computing the pixel-to-pixel dependencies. However, as shown in Fig. 3 (a), the pixel-to-pixel non-local operation is time-consuming and contains redundant computations, because only a certain part of the pixels are highly corresponding to each other [40]. To reduce the redundant computation, we propose a pixel-to-prototype non-local operation. It first converts the pixel-wise feature maps into semantic prototypes with a learnable soft projection matrix, then performs the non-local operation between pixel features and semantic prototypes. Since the number of semantic prototypes is much less than the number of pixels, the computational cost can be significantly reduced.

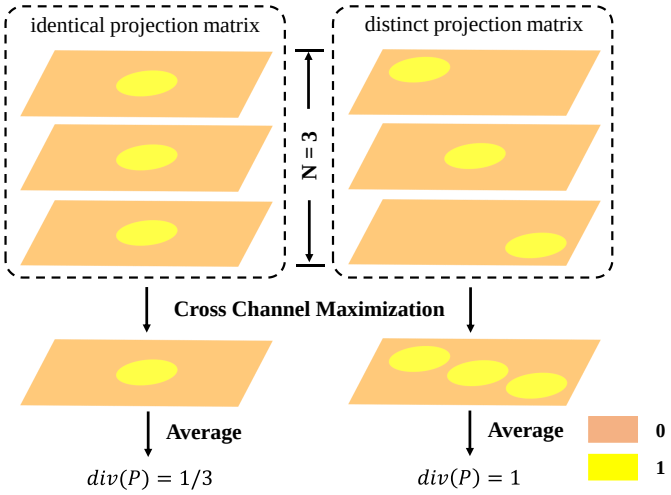


Fig. 4. The calculation process of the diversity value. Assuming that the projection matrix is binary, the left case shows that each semantic prototype aggregates the same pixel features. The right case shows that the prototypes focus on totally different regions, and reach the upper bound of the diversity value.

Pixel-to-prototype non-local operation: Given the image features $X_t \in \mathbb{R}^{C\times H\times W}$ of the t-th frame I_t, where C, H and W represent the channel, height, and width, the soft projection matrix is computed as follows:

$$P_t = \sigma(Conv(X_t)) \in \mathbb{R}^{N\times H\times W}, \tag{1}$$

where N is the number of the semantic prototypes and it is much smaller than the total number of pixels, $Conv$ denotes the learnable 1×1 convolutional layer, and σ denotes the

Softmax operation along the spatial dimensions. Then X_t can be converted into semantic prototypes N_t with matrix multiplication:

$$N_t = \gamma(X_t)\gamma(P_t)^T \in \mathbb{R}^{C\times N}, \tag{2}$$

where $\gamma(*)$ is the reshaping operation, and $\gamma(X_t) \in \mathbb{R}^{C\times HW}$. For brevity, we denote $\gamma(X_t)$ as X_t in the following of this section.

Each semantic prototype contains a C dimensional feature embedding. Then the non-local operation between pixel-wise features X_t and semantic prototypes N_t is performed to obtain the per-frame context augmented features $\widehat{X}_t$:

$$\widetilde{X}_t = \sigma(\theta(X_t)\phi(N_t)^T)g(N_t) \in \mathbb{R}^{C\times HW}, \tag{3}$$

$$\widehat{X}_t = F_{out}(X_t||\widetilde{X}_t) \in \mathbb{R}^{C\times HW}, \tag{4}$$

where θ, ϕ, and g denote three linear transformations, $||$ is the concatenating operation, and F_{out} is implemented with the 1×1 convolutional layer to reduce the channel dimension of augmented features. As shown in Fig. 3 (b), the computational complexity is reduced from $\mathcal{O}(CH^2W^2)$ to $\mathcal{O}(CHWN)$ by replacing the traditional pixel-to-pixel non-local operation with the pixel-to-prototype one. Note that usually $N \ll HW$. For example, for the 769×769 input images, the resolution of feature maps $HW = 97\times 97 = 9409$, and $N = 32$ in our experimental implementations.

Diversity loss for prototype generation: Benefiting from the end-to-end training process, the per-frame context module can learn to aggregate semantic prototypes for feature augmentation adaptively. However, it may aggregate pixels in similar regions for different prototypes, which results in information redundancy. To solve this problem, we propose a diversity loss to reduce the redundancy among semantic prototypes by forcing them to focus on different semantically discriminative regions. We first compute the diversity value of semantic prototypes based on the soft projection matrix P_t in Eq. 1 as follows:

$$div(P_t) = \frac{1}{N}\sum_{m=1}^{HW}\max_{n=1,2,...,N}[\gamma(P_t)]_{n,m}, \tag{5}$$

where N is the number of the semantic prototypes and γ is the reshaping operation. Fig. 4 illustrates the calculation process of the diversity value $div(P_t) \in [\frac{1}{N}, 1]$. When $div(P_t)$ reaches the minimum value, all semantic prototypes degenerate into one prototype. Note that the greater value of $div(P_t)$, the more diversity of the prototypes. Then the diversity loss is employed to punish the low diversity situation of P_t:

$$L_{div} = max(\theta_{div} - div(P_t),\ 0), \tag{6}$$

where θ_{div} is the diversity threshold of the soft projection matrix. Instead of adopting the $L1$ or $L2$ losses which constrain the $div(P_t)$ to converge to a specific scalar value, we add a one-side constraint on P_t, so the non-local operation module can learn an optimal P_t with higher diversity.

B. Short-term Context Module

The short-term context module is proposed to transform the context information from previous frames to the current frame. Given a frame pair $\{I_{t-\tau}, I_t\}$, where t indicates the index of current frame and τ indicates a small positive integer, the short-term context module is applied to generate the inter-frame context augmented features. Similar to Eq. 1 and Eq. 2, the semantic prototypes of the previous frame $N_{t-\tau}$ are used and a non-local operation is performed between $N_{t-\tau}$ and X_t to capture inter-frame context information:

$$\widetilde{X}_{t-\tau,t} = \sigma(\theta(X_t)\phi(N_{t-\tau})^T)g(N_{t-\tau}) \in \mathbb{R}^{C\times HW}, \tag{7}$$

$$\widehat{X}_{t-\tau,t} = F_{out}(X_t||\widetilde{X}_{t-\tau,t}) \in \mathbb{R}^{C\times HW}. \tag{8}$$

Although $\widehat{X}_{t-\tau,t}$ contains spatial-temporal context information of the frame pair, it is not sufficient to handle challenging situations like motion blur, illumination variation, and object occlusion. To achieve better performance, the context information of multiple local frames is leveraged. Notice that the semantic prototypes can be reused in the process of both the per-frame context aggregation and the short-term context aggregation, the prototype generation only needs to be carried out once for each frame at the inference time. The per-frame and the short-term context can be aggregated jointly via the same non-local operation (see Fig. 3 (c)). Considering the trade-off between accuracy and speed, we propose two types of short-term context modules.

Multi-stream short-term context module (MSC). Given the current frame $\{I_t\}$ and a sequence of previous frames $\{I_{t-1}, I_{t-2}, ..., I_{t-T}\}$, the corresponding features augmented with the short-term context information $\{\widehat{X}_{t-1,t}, \widehat{X}_{t-2,t}, ..., \widehat{X}_{t-T,t}\}$ are generated with the pixel-to-prototype non-local operation. T is the number of previous frames, and the computational complexity is $\mathcal{O}(TCHWN)$. After the non-local operation, these augmented features are fed into different segmentation streams to separately generate the predictions for the current frame $\{S_{t-1,t}, S_{t-2,t}, ..., S_{t-T,t}\}$. These segmentation streams have the same architecture but with different parameters. Each stream consists of a 3×3 convolutional layer and a 1×1 convolutional layer. The final output of our method is the average of all predictions:

$$\bar{S} = \frac{1}{T+1}\sum_{\tau=0}^{T} S_{t-\tau,t}, \tag{9}$$

where $S_{t,t}$ denotes the prediction obtained from the per-frame context module. With the multiple context information from both the current and nearby frames, our method is more robust to short-term challenging scenarios.

Inspired by the co-training assumption [35], [48], we regard each frame pair $\{I_{t-\tau}, I_t\}$ as a view of the video sequence, and adopt a co-training loss $\mathcal{L}_{cot}$ to encourage the segmentation streams to learn different and complementary knowledge among frame pairs:

$$\mathcal{L}_{cot} = H(\frac{1}{T+1}\sum_{\tau=0}^{T} S_{t-\tau,t}) - \frac{1}{T+1}\sum_{\tau=0}^{T} H(S_{t-\tau,t}), \tag{10}$$

where $H(*) = \mathbb{E}[-\log(*)]$. This co-training loss helps the segmentation streams learn from each other, which leads to better ensemble performance.

Single-stream short-term context module (SSC). To further reduce the computation cost of our method, we propose the single-stream short-term context module. Instead of performing the non-local operation multiple times when a new frame comes, it only calls for the non-local operation once. The semantic prototypes of multiple frames are first concatenated together to form the local prototype set $N_{t-T:t}$, where $N_{t-T:t} = (N_{t-T} \,||\, ... \,||\, N_t) \in \mathbb{R}^{C\times(T+1)N}$, then a non-local operation is carried out between X_t and $N_{t-T:t}$ to generate the augmented features. Finally, these augmented features are fed into a single segmentation head to obtain the segmentation mask S. Compared to the MSC, the SSC just feeds augmented features to the segmentation head once to obtain the final predicted result, leading to a faster inference speed. By reusing the semantic prototype features of previous frames, the short-term context module can achieve fast inference speed with a balanced low latency.

Benefiting from the multi-stream architecture and the extra co-training loss, the MSC achieves better performance in terms of accuracy, while the SSC requires a lower computation cost. To further boost performance, we employ the MSC/SSC in the Teacher/Student models, and propose a novel knowledge distillation strategy in § III-D.

Algorithm 1 Inference pipeline of the proposed MGCNet

Input: Video sequence with the total frame number N; Maximum frame number of short-term context module T.
Output: Segmentation results $\{S_t\}_{t=1}^{N}$.

```
Initialize the pool of semantic prototypes P = ∅
Initialize the current frame index t = 0
while t < N do
    Generate the per-frame features X_t
    Transform X_t into the semantic prototypes N_t
    Push N_t into P
    if t < T then
        Generate S_t with the per-frame module
    else
        Pop out the prototypes of the oldest frame in P
        Generate the S_t with the short-term module
    end if
    t = t + 1
end while
```

C. Long-term Context Module

Although the spatial information of distant frames is not effective due to the object and camera movement, it is still useful to capture the video-level context information. When the motion blur or partial occlusion occurs in a short-term time span, looking out of the local frames gives the models additional information to handle these problems. To this end, we introduce the long-term context module during training. As shown in Fig. 2, the image features from nearby frames are fed into the long-term context module, together with the same number of features from distant frames. Since the input carries the global information of a video, the long-term context can be aggregated. The long-term context module consists of a global pooling layer and a multilayer perceptron. During training, we assign each training video a separate ID, and a classification loss $\mathcal{L}_{cls}$ is utilized for supervision. The long-term context module is introduced to learn video-level fine-grained scene information during training. At the inference time, the long-term module is removed to avoid additional computational costs. The inference pipeline of the proposed method is summarized in Alg. 1.

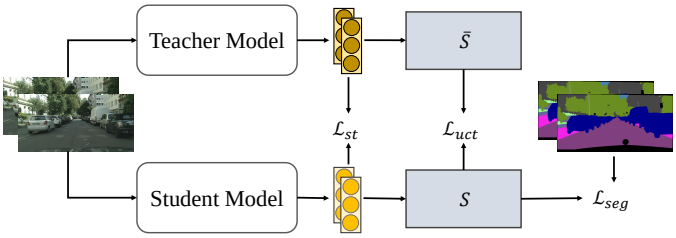


Fig. 5. The overall training framework with the proposed knowledge distillation strategy. During training, the parameters of Teacher Model are fixed and the lightweight Student Model is trained with L_{seg}, L_{st}, and L_{uct} jointly.

The proposed MGNet is trained to jointly aggregate the per-frame, the short-term, and the long-term context information in an end-to-end manner. The total segmentation loss $\mathcal{L}_{seg}$ of MCGNet equipped with the MSC is as follows:

$$\mathcal{L}_{seg} = \mathcal{L}_{mask} + \alpha_1\mathcal{L}_{div} + \alpha_2\mathcal{L}_{cls} + \alpha_3\mathcal{L}_{cot}, \tag{11}$$

where $\mathcal{L}_{mask}$ is supervised by the segmentation annotations, and the binary cross-entropy loss is adopted as $\mathcal{L}_{cls}$. We set $\alpha_1 = 1.0$, $\alpha_2 = 0.4$, and $\alpha_3 = 0.5$ in our experiments. When the SSC is introduced as the short-term context module, the co-training loss $\mathcal{L}_{cot}$ is removed during training.

D. Uncertainty-aware and Structural Knowledge Distillation

We build an effective distillation mechanism to improve the performance of the compact model while maintaining the high inference speed. As shown in Fig. 5, the teacher model is first trained with large backbone networks to reach performance saturation. Then the teacher model is fixed and the student model is trained. Our distillation method adopts different network architectures of the teacher and student models, which not only helps transfer the structural and semantic knowledge but also reduces the computation cost of the student model at the same time.

Model Architecture: For the teacher model, we choose ResNet101 as the backbone network, and introduce the MSC as the short-term context module. For the student model, we choose MobileNetV2, ResNet18 and ResNet50 as the backbone networks, and introduce the SSC as the short-term context module.

Uncertainty-aware Knowledge Distillation: For uncertainty estimation of the teacher module, the Monte Carlo Dropout [9], [21], [23] is applied at each segmentation stream in the MSC. Similar to [2], [16], [18], the final prediction of

the teacher model is generated by the mean probability map $\bar{S}$ according to Eq. 9, and the uncertainty map is approximated as:

$$U = \frac{1}{T+1}\sum_{\tau=0}^{T}(S_{t-\tau,t}{}^2 - \bar{S}^2), \tag{12}$$

where U calculates the variance of multi-steam predictions at each pixel. U and $\bar{S}$ have the same spatial size as ground-truth labels. Following the assumption of uncertainty estimation, the uncertainty map measures the pixel-wise predicted confidence of the models, so we regard U as a soft weight on the whole predictions:

$$\mathcal{L}_{uct} = \frac{1}{HW}\sum_{i=1}^{HW}(1-U)KL(S_{\mathcal{S}}(i), S_{\mathcal{T}}(i)), \tag{13}$$

where KL denotes the Kullback-Leibler divergence, $S_{\mathcal{S}}$ and $S_{\mathcal{T}}$ are the segmentation masks of the student model and the teacher model, respectively.

Structural Knowledge Distillation: In addition to the above pixel-level distillation, we also design a prototype-level distillation strategy to learn structural information from the teacher model. Since the soft projection matrix in the pixel-to-prototype non-local operation is used to aggregate pixels into semantic prototypes, it contains semantically structural information of images. To implicitly transfer the structural knowledge, we propose a structural distillation loss as follows:

$$\mathcal{L}_{st} = \frac{1}{HW}\sum_{i=1}^{HW}KL(P_{\mathcal{S}}(i), P_{\mathcal{T}}(i)), \tag{14}$$

where $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ are the soft projection matrix of the student and the teacher models, respectively. The total loss of the student model is a combination of the segmentation loss and the knowledge distillation losses:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \beta_1\mathcal{L}_{uct} + \beta_2\mathcal{L}_{st}, \tag{15}$$

where we empirically set $\beta_1 = 0.5$ and $\beta_2 = 0.4$ in our implementations. The knowledge distillation has been applied in previous semantic segmentation methods [15], [30]. Different from these methods, we estimate the uncertainty with spatial-temporal information, and both pixel-wise and prototype-wise knowledge are utilized. Our uncertainty-aware and structural knowledge distillation strategy is proposed to boost the performance of the lightweight models, which can further improve the efficiency of our method.

IV. Experiments

In this section, the datasets and metrics are introduced in §IV-A. The implementation details are provided in §IV-B. Then the ablation studies are carried out in §IV-C to evaluate the effectiveness of the proposed modules. Finally, the experimental results in comparison to other state-of-the-arts are shown in §IV-D.

A. Datasets & Metrics

Datasets. We evaluate our method on Cityscapes [4] and CamVid [1] datasets. The Cityscapes dataset is built for urban scene understanding. It consists of 2,975, 500, and 1,525 frame snippets for training, validation, and testing, respectively. There are 30 annotated classes in the dataset, and 19 of them are used for semantic segmentation. The resolution of frames is 1024×2048, with a ground truth annotation of the 20th frame in each snippet. The CamVid dataset contains 467, 100, and 233 frame snippets with annotations of 11 semantic classes. The resolution of frames is 720×960, and the ground truth annotation of the 30th frame is provided in each snippet.

Metrics. Following the previous works, we use the mean Intersection-over-Union (mIoU) to evaluate the accuracy. Mathematically, the mIoU can be formulated as follows:

$$IoU = \frac{tp}{tp + fp + fn}, \tag{16}$$

$$mIoU = \frac{1}{N}\sum_{n=1}^{N} IoU, \tag{17}$$

where the IoU score is leveraged to evaluate the accuracy of binary classification. The tp, fp, and fn are the number of pixels belonging to the true-positive, false-positive, and false-negative sets, respectively. N is the total number of semantic categories. In addition to mIoU, the average inference time and the per-frame maximum latency are also used for efficiency analysis.

B. Implementation Details

Our model is initialized with ImageNet [5] pre-trained parameters. For the training data augmentation, we perform the mean subtraction, random horizontal flipping, random scaling in $[0.5, 2.0]$, random cropping, and random brightness in $[-10, 10]$. The mini-batch stochastic gradient descent (SGD) is employed. The momentum is 0.9 with weight decay 0.0005. The learning rate is initialized as 0.01 and decayed by $(1 - \frac{iter}{max\_iter})^{0.9}$ at each iteration. We train our models with a batch size of 8 for 120k iterations. The Sync-BN [45] is applied. The auxiliary supervision and the OHEM cross-entropy loss are applied in $\mathcal{L}_{mask}$. For the Teacher Model, the ResNet101 is selected as the backbone network. For the Student Model, the ResNet50, ResNet18, and MobileNetV2 are selected as backbones. The input resolution of our method is 769×769 and 769×1537 for the Cityscapes dataset. The input resolution is 560×560 for the CamVid dataset. During testing, the Student Models are used for performance comparison and the long-term context module in §III-C is removed to improve inference speed. The model predictions are resized to the original resolution for evaluation. The number of per-frame semantic prototypes N for the pixel-to-prototype non-local operation is 32. The diversity threshold θ_{div} in Eq. 6 is 0.25. The number of previous frames in the short-term context module is 3. The frame interval for the distant frame selection is 5 in the long-term context module. To obtain the reported results in the testing split, both the training and validation sets are used during training.

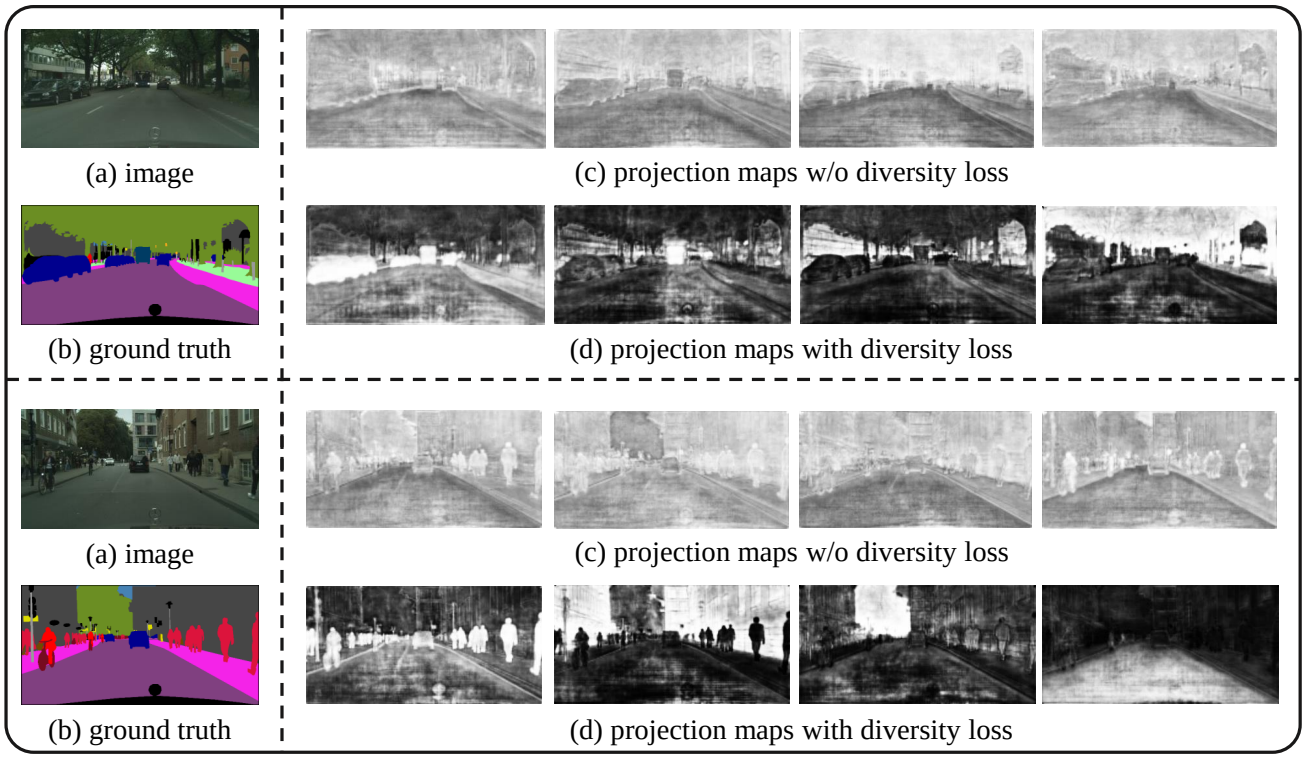


Fig. 6. Visualized results for the soft projection matrix P_t in Eq. 1 on Cityscapes dataset. Four channels of P_t are randomly selected and normalized with *Sigmoid* operation for better visualization quality. Given the (a) input image and the (b) corresponding ground truth, the projection maps with and without diversity loss are illustrated in (c) and (d), respectively. By introducing diversity loss, the projection matrix can focus on different and semantic regions effectively, leading to less information redundancy within the semantic prototypes.

TABLE I
EFFECTIVENESS ANALYSIS OF THE PROPOSED PER-FRAME (PF.), SHORT-TERM (ST.), LONG-TERM (LT.) MODULES, AND UNCERTAINTY-AWARE AND STRUCTURAL KNOWLEDGE DISTILLATION (KD.) STRATEGY ON THE CITYSCAPES VALIDATION SET.

pf.	st.	lt.	kd.	mIoU (%)	speed (ms)	max latency (ms)
				68.8	42.7	42.7
✓				71.5	46.3	46.3
✓		✓		71.8	46.3	46.3
✓	✓			73.0	48.6	48.6
✓	✓	✓		73.7	48.6	48.6
✓	✓	✓	✓	**75.1**	48.6	48.6

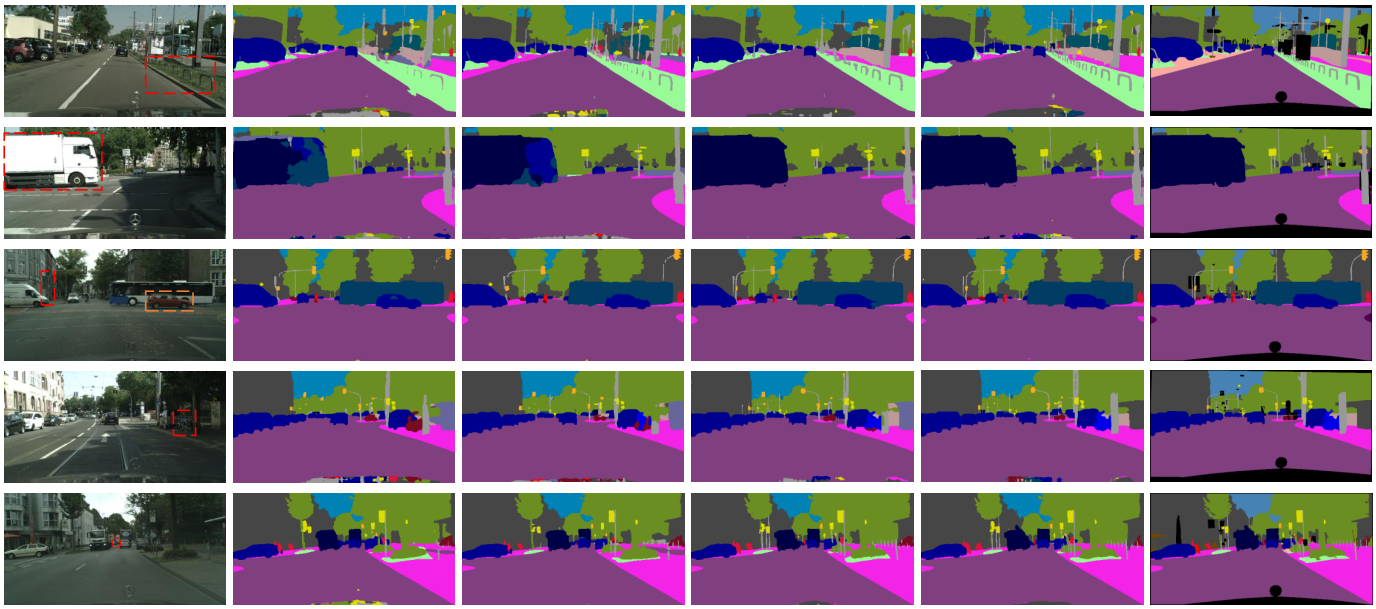
The proposed method is implemented on PyTorch [34]. All the networks are trained on 8 TESLA V100 GPUs, and the speed is evaluated on the GTX 1080Ti GPUs.

C. Ablations

We conduct the ablation studies on Cityscapes validation set. The resolution of input images is 769×769. Most of the experiments are carried out on ResNet18. For the knowledge distill evaluation, we also report results on MobileNetV2 to verify the generalization ability.

Overall Comparison. In this part, we evaluate the effectiveness of the proposed context modules and the knowledge distillation strategy. The accuracy and the speed metrics are reported in Table I. Our baseline model is built with FCN8s [31] and trained with every single labeled frame. As shown in the first row of Table I, it achieves 68.8% mIoU. We first build our image segmentation model by adding the per-frame context module into the baseline and report the results in the second row. The aggregation of intra-frame contexts brings 2.7% mIoU improvement. To capture the inter-frame context, the short-term and long-term context modules are further introduced, and the mIoU is improved from 71.5% to 73.0% and 71.8%, respectively. As shown in the fifth row, by applying the per-frame, short-term, and long-term context modules jointly, better performance is achieved, which demonstrates that the multi-granularity context information is complementary and discriminative to video semantic segmentation. The mIoU can be improved from 68.8% to 73.7%. To further boost the performance of our method, we introduce the proposed distillation strategy during training and obtain an extra 1.4% mIoU improvement. Our complete method achieves 6.3% mIoU improvement in comparison to the result of baseline model in the first row of Table I.

We also evaluate the average inference time and the per-frame max latency for the proposed modules. The average inference time is increased from 42.7 ms to 46.3 ms by introducing the per-frame module. An additional 2.3 ms is required by further introducing the short-time context module. The long-term context module is just utilized during training and can be removed during testing, avoiding the improvement of inference time. Besides, the proposed knowledge distillation strategy can also enhance the performance of our



(a) img (b) pf. (c) pf.+st. (d) pf.+st.+lt. (e) pf.+st.+lt.+kd. (f) ground truth

Fig. 7. Qualitative results of our method on the Cityscapes dataset. The pf., st., and lt. denote the per-frame, short-term, and long-term modules, respectively. The kd. denotes the proposed knowledge distillation strategy.

TABLE II
COMPARISON OF THE SHORT-TERM CONTEXT MODULES ON THE CITYSCAPES VALIDATION SET. N IS THE NUMBER OF SEMANTIC PROTOTYPES IN EACH FRAME.

module	$\mathcal{L}_{div}$	N=8	N=16	N=32	N=64	N=128
MSC		72.5	72.6	73.0	73.3	73.4
	✓	72.9	73.1	73.4	73.6	73.6
SSC		72.2	72.3	72.4	72.7	72.8
	✓	72.5	72.8	73.0	72.8	73.2

method without the extra computational costs during inferring. Compared to the baseline model, the average inference time is only increased by 6.0 ms due to the efficiency of our proposed modules. Different from the keyframe-based video semantic segmentation methods [36], [19], [26] which require more inference time at a certain frame, our method regards each frame equally and leverages the semantic prototypes for efficient context aggregation. The per-frame max latency is the same as the average inference time in video processing in all variants of our method, which is more friendly to the video segmentation applications.

Evaluation on Pixel-to-prototype Non-local Operation. The pixel-to-prototype non-local operation is proposed to aggregate both the per-frame and the short-term context information in a unified manner. We perform ablation experiments about the diversity loss $\mathcal{L}_{div}$ and the per-frame prototype number N on the MSC and the SSC modules proposed in § III-B. The mIoU results are reported in Table II. As shown in the table, the MSC achieves higher accuracy while the SSC has a faster inference speed under the same hyper-parameter settings. The diversity loss is designed to reduce the redundancy among per-frame semantic prototypes. When equipped with the diversity loss, both the MSC and the SSC modules obtain performance improvements, which demonstrates the effectiveness. In detail, when $N = 32$, the mIoU of MSC is improved from 73.0% to 73.4%, and the mIoU of SSC is improved from 72.4% to 73.0%. The per-frame prototype number N is another significant hyper-parameter for our method. Different from OCRNet [43], our method learns to construct semantic prototypes in an unsupervised way, so each semantic prototype not just focuses on a certain semantic category. OCRNet requires the model to learn a fixed number of prototypes such as 19 for Cityscapes [4] and 150 for ADE20k [49], while our prototype generation strategy is more flexible than OCRNet since the per-frame prototypes can be any number such as 32 in our main experiments. We report the results of MSC and SSC with a different number of prototypes including 8, 16, 32, 64, and 128 for each frame. The accuracy can be improved by adopting a larger number of semantic prototypes and the improvement of accuracy tends to saturation with 128 prototypes. Especially for SSC, our model without diversity loss achieves 72.8% mIoU when N equals 128. With the help of diversity loss, the redundant information among prototypes is reduced and our model can reach comparable performance with only a quarter of N. Considering the trade-off between accuracy and speed, we set the per-frame prototype number as 32 in our main experiments. Moreover, we employ the MSC for the complex model to get higher accuracy and the SSC for the lightweight model to maintain the high speed, and propose a distillation strategy to narrow the performance gap between the two models.

Comparison to Pixel-to-pixel Non-local Operation. The non-local operation is utilized to aggregate the context information for semantic segmentation by capturing the long-range

TABLE III
EFFECTIVENESS EVALUATION OF THE NON-LOCAL OPERATIONS. FOR THE PIXEL-TO-PROTOTYPE NON-LOCAL OPERATION, THE PER-FRAME PROTOTYPE NUMBER N IS 32.

non-local operation	$\mathcal{L}_{div}$	mIoU	speed
pixel-to-pixel		72.5	55.4
pixel-to-prototype		72.4	48.6
pixel-to-prototype	✓	**73.0**	48.6

relation among pixels. We compare our pixel-to-prototype non-local operation with the traditional pixel-to-pixel one and report the results in Table III. The SSC module is adopted for the two non-local operations. The pixel-to-pixel non-local operation calculates the pixel-wise affinity matrix directly and achieves 72.5% mIoU. Our pixel-to-prototype one first transforms the feature maps into semantic prototypes then calculates the affinity matrix between image features and semantic prototypes. It achieves 72.4% mIoU when leveraging 32 prototypes at each frame. By further introducing the diversity loss during training, the proposed non-local operation obtains 0.6% mIoU improvement, which even outperforms the pixel-to-pixel one. The experimental results in Tabel III demonstrate that the noise existing in the traditional non-local operation can be reduced by our method. The diversity loss helps the models learn proper feature representation for video semantic segmentation. For the speed comparison, our method requires 48.6 ms/frame while the pixel-to-pixel non-local operation requires 55.4 ms/frame. Our method can achieve higher accuracy with lower computational costs than the pixel-to-pixel one.

Visualization of Soft Projection Matrix. The visualization of soft projection matrix of Eq. 1 is illustrated in Fig. 6. For each input image, we randomly select 4 projection maps from the soft projection matrix, and each projection map can be leveraged to form a certain semantic prototype. The projection maps are normalized with *Sigmoid* operation. The lighter the pixel, the more contribution to the prototype. As shown in Fig. 6 (c), the difference between projection maps is insignificant when trained without the diversity loss. To increase the difference, the diversity loss is adopted. As shown in Fig. 6 (d), projection maps are able to adaptively focus on different semantic regions, which reduces the information redundancy among semantic prototypes. Moreover, we observe that the diversity loss has the potential to help the segmentation model aggregate the hierarchical semantic information. As shown in the upper example of Fig. 6, the first projection map in (d) highlights pixels belonging to the "car" and the "bus" category, which means this prototype focuses more on the "vehicle" parent category. And the second projection map in (d) just focuses on the "bus" category. Such hierarchical information may increase the interpretability of semantic prototypes in the case when the per-frame prototype number is larger than the semantic categories of the dataset.

Evaluation on Co-training Loss. The co-training loss $\mathcal{L}_{cot}$ is adopted to improve the accuracy of MSC. It encourages the multiple segmentation streams to learn the discriminate and complementary segmentation information. Since the MSC is only used for the Teacher Model, we evaluate the effectiveness of $\mathcal{L}_{cot}$ on the ResNet101 backbone with different loss weight α_3 of Eq. 11. As shown in Table V, the $\alpha_3 = 0.0$ denotes the case that trained without the co-training loss. It achieves 78.8% mIoU. The performance can be improved by adopting the $\mathcal{L}_{cot}$. The α_3 is empirically set at 0.5 and 0.6% mIoU improvement can be achieved.

TABLE IV
EFFECTIVENESS ANALYSIS ON THE NUMBER OF PREVIOUS FRAMES T IN THE SHORT-TERM CONTEXT MODULE.

T	0	1	2	3
mIoU	71.8	73.0	73.4	**73.7**
speed	46.3	46.8	47.5	48.6

TABLE V
EFFECTIVENESS ANALYSIS ON THE CO-TRAINING LOSS WITH THE MSC. α_3 IS THE LOSS WEIGHT IN EQ. 10.

α_3	0.0	0.1	0.3	0.5	1.0
mIoU	78.8	79.2	79.2	**79.4**	79.1

Evaluation on Number of Previous Frames. To evaluate the effectiveness of our spatial-temporal context aggregation strategy, experiments about the number of previous frames are carried out and the results are shown in Table IV. The backbone network is ResNet18 and the SSC is selected as the short-term context module. The model is trained without the proposed knowledge distillation strategy. When the number of previous frames comes to 0, the SSC degenerates into the per-frame context module and only the current context information is utilized for video semantic segmentation. It achieves 71.8% mIoU. By adding just one previous frame, the mIoU is improved from 71.8% to 73.0%, which demonstrates the significance of the temporal information for video semantic segmentation. As shown in the table, the more previous frames, the better segmentation performance in terms of the mIoU metric. Our context aggregation modules is effective and the mIoU is improved from 71.8% to 73.7%. For the speed evaluation, the per-frame context module requires 46.3 ms/frame inference speed. Since the proposed non-local operation can reuse the semantic prototypes of previous frames during inferring, the semantic prototypes are just calculated once for each frame. The improvement of computational costs is acceptable when incorporating more previous frames for context aggregation. When incorporating 3 previous frames, the average inference speed is only increased from 46.3 ms/frame to 48.6 ms/frame in our method. In our main experiments, the number of previous frame is set as 3 considering the tradeoff between accuracy and speed.

Evaluation on Knowledge Distillation Strategy. Traditional knowledge distillation leverages the pixel-wise predic-

TABLE VI
INFLUENCE OF THE PROPOSED KNOWLEDGE DISTILLATION STRATEGY IN TERMS OF THE MIOU (%) METRIC. THE EFFECTIVENESS OF $\mathcal{L}_{uct}$ AND $\mathcal{L}_{st}$ IS ANALYZED.

Model	Backbone	$\mathcal{L}_{uct}$	$\mathcal{L}_{st}$	mIoU
T-Net	ResNet101	-	-	79.4
S-Net	ResNet18			73.7
		✓		74.6
			✓	74.4
		✓	✓	**75.1**
S-Net	MobileNetV2			73.4
		✓		74.4
			✓	73.8
		✓	✓	**74.7**

tions of the Teacher Networks (T-Net) to guide the learning process of the Student Networks (S-Net). However, it regards each pixel equally and ignores the predicted confidence of T-Net. The predicted error of the T-Net may affect the learning of S-Net. To suppress such noise during training, we leverage an asymmetric Teacher-Student architecture and propose the uncertainty-aware knowledge distillation to detect the unconfident predictions of the T-Net. Specifically, the MSC and the SSC are applied in the T-Net and the S-Net, respectively. We select ResNet101 as the backbone of T-Net to obtain higher accuracy while selecting ResNet18 and MobileNetV2 as the backbones of S-Net for higher inference speed. As shown in Table VI, the T-Net achieves 79.4% mIoU, and the S-Nets without any distillation strategy separately achieve 73.7% and 73.4% mIoU with Res18 and MobV2 backbones. By introducing the uncertainty-aware knowledge distillation, the S-Nets obtain 0.9% and 1.0% mIoU improvements, respectively. In addition to providing pixel-level semantic guidance from the predictions of T-Net, the structural knowledge distillation is also proposed to train the S-Net at the prototype level. It can implicitly capture structural dependencies between the T-Net and the S-Net. As shown in Table VI, by introducing the uncertainty-aware knowledge distillation, the S-Nets obtain 0.7% and 0.4% mIoU improvements, respectively. These two distillation approaches work by providing the pixel-level and prototype-level similarity during training. By introducing the uncertainty-aware and the structural knowledge distillation jointly, better performance is obtained. Our knowledge distillation strategy is complementary and robust to different networks. The ResNet18 and the MobileNetV2 backbones obtain 1.4% and 1.3% mIoU improvement in total on the Cityscapes dataset.

Qualitative Analysis. Fig. 7 shows the qualitative results with the proposed mechanisms including the per-frame (pf.), the short-term (st.), the long-term (lt.) modules, and the knowledge distillation (kd.) strategy. More significant differences are highlighted within the bounding boxes in the figure. In the first row of Fig. 7, the image segmentation model fails to separate the slender objects. By introducing multi-frame contexts, more delicate semantic information is preserved to improve the segmentation quality of slender objects. The second row shows the scenario when the short-term context module fails due to the object occlusion within the nearby frames. By introducing the long-term context module, the video-level semantic information can be learned during training, and the category-level error is corrected successfully. The third row illustrates the effectiveness of the proposed knowledge distillation strategy. By equipped with our knowledge distillation strategy during training, the segmentation model can separate the objects from the background much better. The last two rows show segmentation performance under the challenging scenarios. Especially, the fourth and fifth rows are the scenarios about illumination variation and small objects, respectively. The complete method (e) handles these challenging scenarios better in comparison to our per-frame baseline (b).

D. Compared to the state-of-the-arts

Cityscapes dataset. We compare our MGCNet to the state-of-the-art video semantic segmentation methods including CLK [36], DFF [50], GRFP [32], Accel [19], ETC [30], PEARL [20], LVS [26], and TDNet [15]. Since the input resolutions are different among these methods, we report the results of ours with various resolutions (i.e., 769×769 and 769×1537). The results of mIoU, average inference speed, and per-frame max latency metrics are reported in Table VII. By adjusting the backbone networks and modifying the input sizes, the proposed MGCNet can satisfy the different computational requirements, especially the ResNet50 backbone achieves 80.6% and 80.2% mIoU with 186 ms/frame inference speed on Cityscapes val and test set, respectively. Compared to keyframe-based methods like CLK, DFF, Accel, and LVS, our method achieves the best results by reusing the semantic prototypes of previous frames. Besides, our method avoids additional computational costs at the keyframe so it has the same per-frame max latency as the average inference speed. Among the previous methods, TDNet ranks first with 79.9% and 79.4% mIoU on val and test set, respectively. It presents a temporally distributed network based on multiple sub-networks and introduces cross attention modules for spatial-temporal context aggregation. However, TDNet may generate fluctuating accuracy when initialized with a different order of sub-networks, and the computational cost of the pixel-to-pixel cross attention is quite expensive. THe MGCNet leverages a shared network for per-frame feature extraction so our segmentation performance is more stable than TDNet. The proposed pixel-to-prototype non-local operation can capture the structural and semantic information with lower computational complexity than the cross attention modules adopted by TDNet. compared with TDNet, our method requires about half the total number of parameters and outperforms it in terms of accuracy with comparable efficiency.

CamVid dataset. The experimental results on the CamVid dataset are shown in Table VIII. We compare our method to state-of-the-art video semantic segmentation methods including DFF [50], GRFP [32], Accel [19], Netwarp [8],

TABLE VII
EXPERIMENTAL RESULTS ON THE CITYSCAPES VAL AND TEST SETS. THE RESOLUTION OF INPUT IMAGES FOR MGCNET AND MGCNET† ARE 769×769 AND 769×1537, RESPECTIVELY.

Method	Publication	Input Resolution	Backbone	mIoU (%) val	mIoU (%) test	speed (ms)	max latency (ms)
CLK [36]	ECCV'16	500×500	VGG16	64.4	-	-	-
DFF [50]	CVPR'17	512×1024	ResNet101	69.2	-	178	-
GRFP [32]	CVPR'18	1024×2048	ResNet101	73.6	72.9	535	535
Accel [19]	CVPR'19	-	ResNet101/18	-	72.1	440	740
ETC [30]	ECCV'20	768×768	MobileNetV2	73.9	-	48	48
PEARL [20]	ICCV'17	1024×2048	ResNet101	76.5	75.2	-	-
LVS [26]	CVPR'18	713×713	ResNet101	76.8	-	171	380
TDNet [15]	CVPR'20	769×1537	ResNet50	79.9	79.4	178	178
MGCNet	-	769×769	MobileNetV2	74.7	-	45	45
			ResNet18	75.1	-	49	49
			ResNet50	76.5	76.2	98	98
MGCNet†	-	769×1537	MobileNetV2	77.8	77.2	62	62
			ResNet18	78.5	77.9	80	80
			ResNet50	**80.6**	**80.2**	186	186

TABLE VIII
EXPERIMENTAL RESULTS ON THE CAMVID DATASET. THE RESOLUTION OF INPUT IMAGES FOR MGCNET IS 560×560.

Method	Backbone	mIoU (%)	speed (ms)
DFF [50]	ResNet101	66.0	-
GRFP [32]	ResNet101	66.1	-
Accel [19]	ResNet101/18	66.7	179
Netwarp [8]	Dilation-CNN [42]	67.1	395
ETC [30]	MobileNetV2	75.2	36
TDNet [15]	ResNet50	76.0	90
MGCNet	ResNet18	74.8	35
	ResNet50	**76.5**	51

TDNet [15], and ETC [30]. We report the mIoU and average inference speed of our method with the ResNet18 and the ResNet50 backbones. For the ResNet18 backbone, our method achieves 74.8% mIoU, with only 35 ms/frame inference speed. For the ResNet50 backbone, our method achieves 76.5% mIoU and outperforms other state-of-the-art methods.

V. CONCLUSION

In this paper, we have presented an efficient video semantic segmentation method by aggregating multi-granularity context information. A pixel-to-prototype non-local operation is introduced to aggregate the per-frame and short-term contexts jointly. Compared to traditional pixel-to-pixel non-local operations, it has less computational complexity and is more suitable for video tasks via reusing the semantic prototypes of previous frames. We also propose a long-term context module to aggregate video-level semantic information during training. We further improved the performance of our method by employing an uncertainty-aware and structural knowledge distillation strategy, learning the pixel-wise and the prototype-wise correlations from the large model. Our method outperforms the state-of-the-art ones in terms of both accuracy and efficiency.

REFERENCES

[1] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *European Conference on Computer Vision*, 2008.

[2] M. Cai, F. Lu, and Y. Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[3] Y. Chen, Y. Cao, H. Hu, and L. Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[6] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, 2017.

[7] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[8] R. Gadde, V. Jampani, and P. V. Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[9] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 2016.

[10] J. Gu, H. Hu, L. Wang, Y. Wei, and J. Dai. Learning region features for object detection. In *European Conference on Computer Vision*, 2018.

[11] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[12] G. Hinton, O. Vinyals, J. Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[14] H. Hu, Z. Zhang, Z. Xie, and S. Lin. Local relation networks for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[15] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[16] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, and M. Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *European Conference on Computer Vision*, 2018.

[17] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[18] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision*, 2018.
[19] S. Jain, X. Wang, and J. E. Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[20] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, et al. Video scene parsing with predictive feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
[21] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.
[22] I. Kreso, S. Segvic, and J. Krapac. Ladder-style densenets for semantic segmentation of large natural images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.
[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.
[24] H. Li, P. Xiong, H. Fan, and J. Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[25] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.
[26] Y. Li, J. Shi, and D. Lin. Low-latency video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
[27] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun. Surveillance video parsing with single frame supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.
[28] X. Liu, W. Ji, J. You, G. E. Fakhri, and J. Woo. Severity-aware semantic segmentation with reinforced wasserstein training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
[29] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[30] Y. Liu, C. Shen, C. Yu, and J. Wang. Efficient semantic video segmentation with per-frame inference. In *European Conference on Computer Vision*, 2020.
[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.
[32] D. Nilsson and C. Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
[33] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
[35] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille. Deep co-training for semi-supervised image recognition. In *European Conference on Computer Vision*, 2018.
[36] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell. Clockwork convnets for video semantic segmentation. In *European Conference on Computer Vision*, 2016.
[37] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
[38] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee. Dynamic video segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6556–6565, 2018.
[39] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, 2020.
[40] C. Yu, Y. Liu, C. Gao, C. Shen, and N. Sang. Representative graph neural network. In *European Conference on Computer Vision*, 2020.
[41] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European*
[42] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
Conference on Computer Vision, 2018.
[43] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision*, pages 173–190, 2020.
[44] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.
[45] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
[46] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *European Conference on Computer Vision*, 2018.
[47] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia. Psanet: Point-wise spatial attention network for scene parsing. In *European Conference on Computer Vision*, 2018.
[48] S. Zhao, L. Zhou, W. Wang, D. Cai, T. L. Lam, and Y. Xu. Splitnet: Divide and co-training. *arXiv preprint arXiv:2011.14660*, 2020.
[49] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.
[50] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.
[51] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Zhiyuan Liang is working toward the Ph. D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include video segmentation and object tracking.

Xiangdong Dai is currently working on camera image processing at OPPO Mobile Communications Corp., Ltd., Guangdong Province, China. He received the MS degree from University of Electronic Science and Technology of China in 2014. His current research interests include portrait segmentation, depth map estimation, and portrait bokeh rendering.

Yiqian Wu received a BSc degree from Zhejiang University in 2021. She is currently a PhD candidate in the State Key Laboratory of CAD&CG, Zhejiang University. Her current research interests include artificial intelligence, computer vision and portrait editing.

Xiaogang Jin (M'04) is a Professor in the State Key Laboratory of CAD&CG, Zhejiang University. He received the B.Sc. degree in computer science and the M.Sc. and Ph.D degrees in applied mathematics from Zhejiang University, P. R. China, in 1989, 1992, and 1995, respectively. His current research interests include image processing, digital human, traffic simulation, collective behavior simulation, cloth animation, virtual try-on, digital face, implicit surface modeling and applications, creative modeling, computer-generated marbling, sketch-based modeling, and virtual reality. He received an ACM Recognition of Service Award in 2015 and two Best Paper Awards from CASA 2017 and CASA 2018. He is a member of the IEEE and ACM.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. He is also an adjunct Professor with the State Key Laboratory of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau, and also with the School of Computer Science, Beijing Institute of Technology, Beijing, China. He received his PhD degree from the Department of Computer Science, Zhejiang University in 2007. He published more than 100 top journal and conference papers, and his Google scholar citations are about 14,600 times with H-index 66.

He was rewarded as the Highly Cited Researcher by the Web of Science in 2020 and 2021, and also the most cited Chinese researchers by the Elsevier Scopus in 2020 and 2021. His research interests include computer vision, deep learning, self-driving cars, medical image analysis and smart city. He is/was an Associate Editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *Pattern Recognition*, and other journals.