Towards Open-Vocabulary Video Semantic Segmentation

Xinhao Li, Yun Liu, Guolei Sun, Min Wu, Le Zhang and Ce Zhu, *Fellow, IEEE*

Abstract—Semantic segmentation in videos has been a focal point of recent research. However, existing models encounter challenges when faced with unfamiliar categories. To address this, we introduce the Open Vocabulary Video Semantic Segmentation (OV-VSS) task, designed to accurately segment every pixel across a wide range of open-vocabulary categories, including those that are novel or previously unexplored. To enhance OV-VSS performance, we propose a robust baseline, OV2VSS, which integrates a spatial-temporal fusion module, allowing the model to utilize temporal relationships across consecutive frames. Additionally, we incorporate a random frame enhancement module, broadening the model's understanding of semantic context throughout the entire video sequence. Our approach also includes video text encoding, which strengthens the model's capability to interpret textual information within the video context. Comprehensive evaluations on benchmark datasets such as VSPW and Cityscapes highlight OV-VSS's zero-shot generalization capabilities, especially in handling novel categories. The results validate OV2VSS's effectiveness, demonstrating improved performance in semantic segmentation tasks across diverse video datasets. The source code will be made available at: https://github.com/AVC2-UESTC/OV2VSS.

Index Terms—Open Vocabulary, Video Semantic Segmentation, Spatial-Temporal Fusion, Random Frame Enhancement, Video Text Encoding

I. INTRODUCTION

SEMANTIC segmentation, the task of classifying each pixel into distinct semantic categories [1, 2, 3, 4, 5, 6], is a core component of computer vision with substantial practical value. However, traditional segmentation models are constrained by being trained on a fixed set of predefined categories. Consequently, they can only recognize categories present in the training data, presenting a significant challenge for real-world applications where novel or unseen categories frequently emerge. This limitation hinders the model's adaptability and scalability in dynamic, diverse environments.

Recent progress has been made in addressing this issue, particularly in image-based semantic segmentation. Notably, studies have integrated the CLIP model [7] into semantic segmentation [8, 9] to achieve open-vocabulary image semantic segmentation. These efforts leverage vision-language knowledge to enhance generalization capabilities, particularly for unseen categories. During testing, the model is capable

Xinhao Li, Le Zhang and Ce Zhu are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. Corresponding author: Le Zhang.

Yun Liu is with the College of Computer Science, Nankai University, Tianjin 300350, China.

Guolei Sun is with Computer Vision Lab, ETH Zurich, Zurich, Switzerland.

Min Wu is with the Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore.
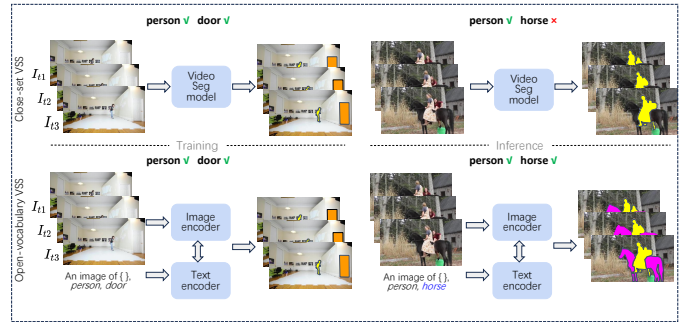

Fig. 1: Comparison of VSS and OV-VSS: In traditional VSS, the model is trained on a closed set of classes (e.g., "person" and "door") and fails to segment novel classes (e.g., "horse," as shown in the figure)(upper). In contrast, our Open-Vocabulary Segmentation model is trained on base classes but can simultaneously segment both base and novel categories(lower).

of segmenting unseen categories as long as their names are provided.

In contrast, the challenge of open-vocabulary video semantic segmentation (**OV-VSS**) has not been systematically explored, despite the fact that our real-world experiences are better represented by dynamic video content rather than static images. Video-based semantic segmentation presents additional complexities due to the temporal dimension, which captures evolving scenes, objects, and contextual information that cannot be encapsulated in a single image. Studying VSS is essential because videos are more prevalent in practical applications, such as autonomous driving, surveillance, and activity recognition, where recognizing objects and actions over time is critical.

While a straightforward solution might be to apply existing image-based methods, such as SAN [8], to video datasets, this approach neglects the rich temporal context inherent to video footage. To validate this hypothesis, we conducted a series of experiments by applying image-based methods to video datasets. The results, as shown in Fig. 2, reveal significant performance degradation, emphasizing the need for a more refined approach to VSS. These findings will be further elaborated upon in the subsequent sections.

Inspired by this, we introduce the challenge of open-vocabulary video semantic segmentation (OV-VSS), which differs from conventional VSS in its ability to segment and classify objects not predefined during training. The key distinctions are illustrated in Fig. 1. In OV-VSS, our goal is to classify every pixel within each video frame, including categories the model has not previously encountered. To address this task, we
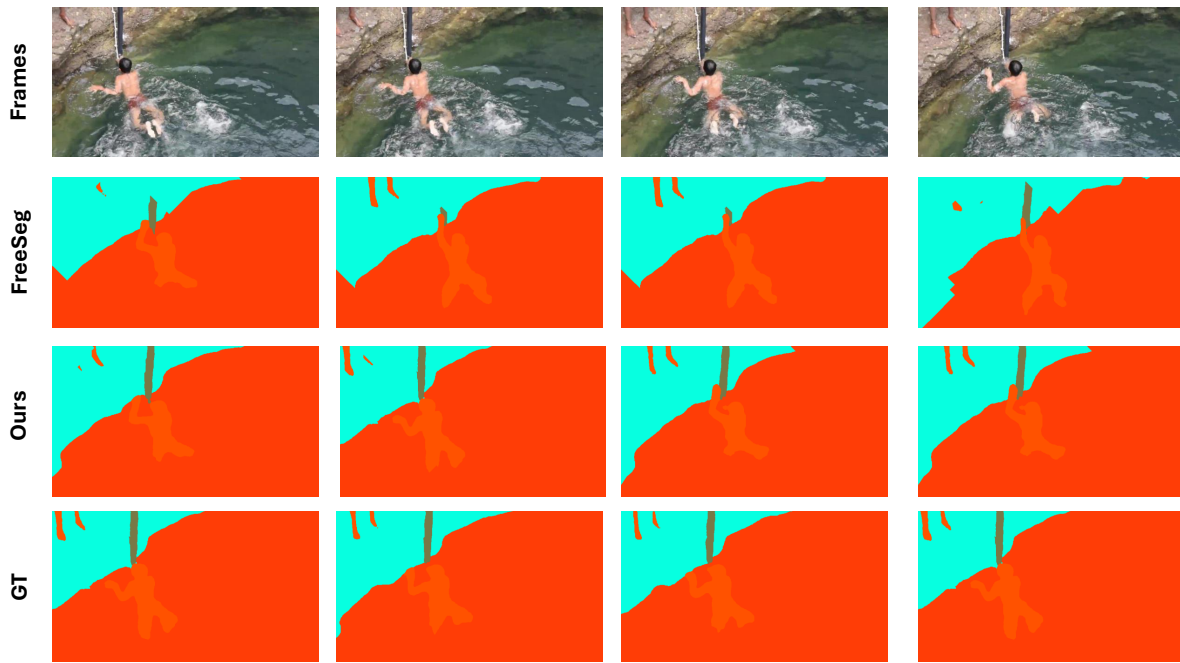
Fig. 2: The comparison between our method and image-based methods on the VSPW dataset, from top to bottom, includes the original frame, image method, our method, and ground truth. From the figure, it can be seen that image-based methods are prone to producing discontinuities in the image.

propose a robust baseline model, OV2VSS, which integrates a spatial-temporal fusion module that leverages temporal relationships across consecutive frames. Additionally, we include a random frame enhancement module to expand the model's understanding of semantic context throughout the entire video sequence. Our approach also incorporates video text encoding to enhance the model's ability to interpret textual information within the video, enriching its semantic comprehension.

It is worth noting that while some recent attempts have addressed open-vocabulary segmentation in videos, these methods [10, 11] focus solely on instance segmentation. In contrast, our open-vocabulary semantic segmentation (VSS) task provides a broader and more comprehensive approach by classifying every pixel within each frame across a wide range of categories, including both foreground objects and the surrounding context. This distinction is crucial because our method goes beyond isolated instances to deliver a full scene understanding, capturing both object and non-object elements, such as "road","sky", or "vegetation", which may lack clear instance boundaries.

The capability of VSS to label both primary objects and contextual regions enriches video content representation, supporting applications where an in-depth understanding of the entire scene is essential. For instance, autonomous vehicles benefit from VSS's ability to identify drivable areas, sidewalks, and other critical scene elements that may not constitute distinct object instances but are vital for navigation. Similarly, in augmented reality, where overlaying information across entire scenes is important, VSS provides a more immersive and accurate experience by segmenting and labeling each part of the environment.

Methodologically, open-vocabulary instance segmentation (OV-VIS) typically employs CLIP as a classifier to map object queries (proposals) to open-vocabulary labels. This approach relies on object-centric queries to propose instances, which are then classified based on text embeddings from CLIP's text encoder. As a result, OV-VIS performance is constrained by the quality of its query proposal system, with segmentation accuracy heavily depending on how effectively the model can generate and classify object instances.

In contrast, our approach to OV-VSS is fundamentally different. We leverage both CLIP's image and text encoders to model pixel-level correlations between visual features and text prompts, allowing us to focus on pixel-wise semantic information across the entire video sequence rather than relying on object proposals. This enables us to utilize both spatial and temporal information, ensuring accurate labeling for all pixels—whether belonging to objects or background elements. The inclusion of our Spatio-Temporal Fusion and Random Frame Enhancement modules further strengthens our approach, enabling efficient capture of temporal dependencies between frames, which is critical for consistent segmentation over time.

In our experiments, we evaluate our methods using the VSPW datasets [12], comprising 124 classes. For clarity, we categorize the datasets into seen classes and unseen classes, maintaining as 80:44. On the challenging VSPW dataset [12], we achieve a mIoU of 17.99% on unseen classes, demonstrating a notable 4% improvement compared to the current state-of-the-art in image-based open-vocabulary segmentation.

To sum up, the contributions of this paper are as follows:
1) We introduce the novel task of Open Vocabulary Video Semantic Segmentation (OV-VSS), demonstrating its significance in addressing the limitations of traditional video semantic segmentation models that struggle with unseen categories.

2) We propose a new baseline method, OV2VSS, which integrates a Spatial-Temporal Context Fusion module to capture intra-frame correlations and enhance temporal consistency. Additionally, we introduce a Random Frame Enhancement module to leverage long-range frame dependencies, alongside a specialized Video-Text Encoding module to align video content with textual descriptions.

3) We conduct extensive evaluations on the VSPW dataset, showcasing the effectiveness and generalizability of our approach. Our method achieves state-of-the-art performance in segmenting unseen classes, significantly improving over previous image-based approaches.

II. RELATED WORK

A. Video Semantic Segmentation

Video Semantic Segmentation (VSS) aims to categorize pixels in each video frame according to predefined categories. Early VSS efforts were constrained by datasets like Cityscapes [13] and NYUDv2 [14], which provided annotations for only a few sporadic frames. Similarly, the CamVid dataset [15] suffers from limited scale and low frame rate. The introduction of the VSPW dataset [12], with 124 classes and a wide range of scenarios, significantly advanced the field. It features an average of over 70 annotated frames per video, providing rich data for training and evaluation.

On this challenging dataset, existing methods can be divided into two main categories. One category focuses on improving performance with minimal additional computation [16, 17, 18, 19, 20, 21, 22, 23, 24]. These methods often prioritize keyframes, applying high-computation networks to keyframes and using lower-computation methods, such as optical flow, to process non-keyframes and capture temporal correlations.

Conversely, another set of methods [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] prioritizes segmentation accuracy by deploying resource-intensive networks on each frame to achieve superior results. However, these methods rely heavily on segmenting classes present in the training datasets, limiting their adaptability to unseen classes—a significant challenge for VSS in real-world applications, where new and previously unencountered classes need accurate identification and segmentation.

B. Open-Vocabulary Segmentation

Vision-language models (VLMs) represent a significant leap in connecting visual and textual modalities. Pre-trained models like CLIP [7] and ALIGN [39] harness large datasets to enable impressive zero-shot object recognition capabilities, boosting performance in tasks such as classification, captioning, and segmentation.

In image segmentation, VLMs excel at zero-shot predictions. For example, Xu et al. [8] propose a side network for open-vocabulary segmentation, while Ding et al. [40] decouple segmentation and classification. Ghiasi et al. [9] align text queries with image regions.

However, applying VLMs directly to video data introduces new challenges, as models like CLIP are trained on static images. Efforts like Wang's [10] aim to extend VLMs to video instance segmentation, but further research is needed to address the complexities of dynamic visual content.

III. OUR METHODS

For a given video clip, a training sample can be represented as follows:

$$([I_{t-k_1}, \cdots, I_{t-k_n}, I_t, I_{random}], G_t), \tag{1}$$

where we define that

- $I_t \in \mathbb{R}^{H \times W}$ denotes the frame of interest, referred to as the target frame at timestamp t, and is accompanied by its associated ground-truth segmentation map $G_t \in \mathbb{R}^{H \times W}$.
- $I_{t-k_1}, \cdots, I_{t-k_n}$ represent the n preceding frames, each located $k_1, \cdots, k_n$ frames away from I_t ($k_1 > k_2 > \cdots > k_n$). In our experiments, $k_n - k_{n+1}$ is set to 3.
- I_{random} signifies a temporally distant frame extracted from the same video, contributing temporal correlations to facilitate the training process. During the training stage, a frame is randomly selected as I_{random}. In the inference stage, I_{random} will be the most temporally-distant frame in the same video.

For each frame I_{t_i}, we first pass it through a backbone network to extract intermediate feature maps $\{U_{t_i}^l \in \mathbb{R}^{H_l \times W_l \times C_l}\}_{l=1}^L$ across different scales from L layers, where H_l, W_l, C_l represent the height, width, and channels of the feature map, respectively. We simplify by omitting the superscript l. Next, the spatial-temporal context fusion module integrates the features across the video sequence, producing the refined feature O_t for the target frame.

The target frame I_t and class name text $T(n)$ for $n = 1, ..., N$ are input into the video-text encoding module to derive language-guided features, which are concatenated with U_t to produce position-aware features $\hat{X}$. Additionally, the random frame feature D_{random} is enhanced using O_t via the random frame enhancement module, yielding refined feature $\hat{O}_t$. Finally, the enhanced feature $\hat{O}_t$ and language-guided feature $\hat{X}$ are fed into the decoder. The detailed design is discussed in subsequent sections.

A. Spatial-Temporal Context Fusion

The key distinction between video and image semantic segmentation lies in capturing inter-frame temporal relationships. The Spatial-Temporal Context Fusion (STCF) module is specifically designed to integrate temporal information through an information-rich module and a temporal transfer module.

Traditional attention-based temporal fusion methods derive the query from the target frame and use reference frames for keys and values. However, this can cause discontinuities due to differences in categories and environments across frames. As the temporal distance increases, the likelihood of such variations also grows.

This insight motivates a gradual aggregation approach: instead of directly fusing I_{t-k_1} with I_t, we first aggregate the closest frames, such as I_{t-k_1} with I_{t-k_2}, then I_{t-k_2} with I_{t-k_3}, and so on until I_{t-k_n} is combined with I_t. This incremental fusion reduces computational overhead during
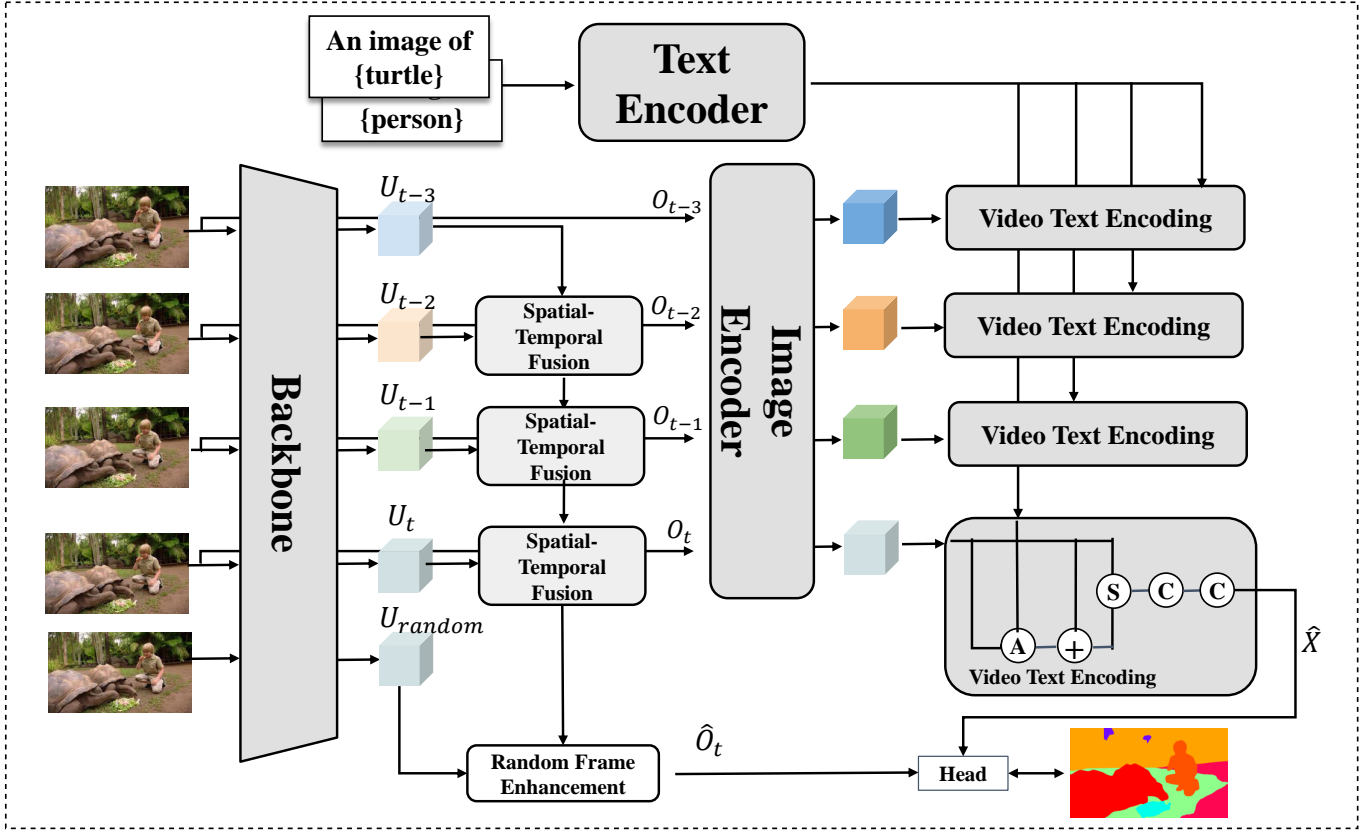
Fig. 3: **Overall structure of OV2VSS**. Our approach primarily leverages three modules: the spatial-temporal information fusion module, which integrates spatial-temporal details from the video; the random frame enhancement module, which acquires contextual information from a randomly selected frame; the video text encoding, which utilizes text supervision in the training process. A denotes attention, + denotes element-wise addition, S denotes establish a cost-volume, C denotes Concatenate. The text encoder and image encoder are obtained from CLIP. Detailed explanations are included in section III.
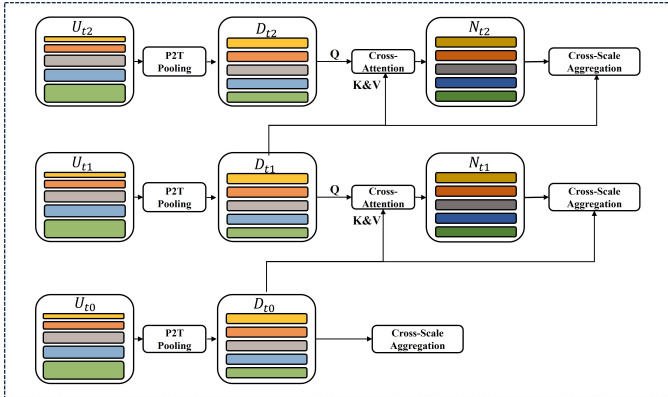

Fig. 4: The architecture of the Spatio-Temporal Context Fusion module, which includes a P2T pooling, cross-attention and a cross-scale Aggregation.

inference, as the intermediate results from closer frames can be reused for subsequent target frames, streamlining the process.

In particular, we first apply a P2T pooling layer [41], denoted as $f(\cdot)$, to enhance each intermediate feature map $U_{t_i}^l$, formulated as:

$$D_{t_i}^l = f(U_{t_i}^l). \tag{2}$$

After obtaining the enhanced feature set $\{D_{t_i}^l\}_{l=1}^L$, we introduce an attention mechanism to extract important information

from frames and propagate it through the temporal sequence.

For simplicity, we omit the superscript l in $D_{t_i}^l$. The first frame feature, D_{t_0}, is directly treated as the representative feature for the final decoder input, where $D_{\text{past}} = D_{t_0}$. For the second frame feature, D_{t_1}, we derive three feature maps using fully-connected layers:

$$Q_{t_i} = \text{FC}(U_{t_i}), \quad K_{\text{past}} = \text{FC}(D_{\text{past}}), \quad V_{\text{past}} = \text{FC}(D_{\text{past}}), \tag{3}$$

where $\text{FC}(\cdot)$ represents the fully-connected layer. We then compute the attention map between two frames, along with the intermediate feature N_{t_i}, as follows:

$$\begin{aligned} A_{t_i} &= Q_{t_i} \times K_{\text{past}}, \\ N_{t_i} &= A_{t_i} \times V_{\text{past}}. \end{aligned} \tag{4}$$

Here, $A_{t_i} \in \mathbb{R}^{HW \times HW}$ and $N_{t_i} \in \mathbb{R}^{HW}$. The feature D_{past} is updated as N_{t_i}, and for subsequent frames, we follow the same process to generate an attention map. At each scale, every frame produces an attention map using the previous historical feature D_{past}.

However, these attention maps are generated without exchanging information across different scales. For instance, shallow layer features offer fine-grained but less semantic information, while deeper layers capture more semantic but

coarser details. This highlights the need for a method to enable cross-scale information communication.

Inspired by [34], we introduce a multi-scale affinity aggregation module for frame t_i, which aggregates attention maps from deeper layers to shallower ones. This can be formulated as:

$$
\begin{aligned}
B_{t_i}^L &= A_{t_i}^L, \\
B_{t_i}^l &= \text{Conv}(\text{Upsample}(B_{t_i}^{l+1} + A_{t_i}^l)), l = L-1, ..., 1.
\end{aligned}
\tag{5}
$$

where L denotes the layer depth, and $B_{t_i}^l$ represents the refined attention map. Since the attention map $B_{t_i}^l$ is refined across multiple scales, it encapsulates the informative affinity among multi-scale features. Using this refined attention B_{t_i} along with the corresponding value V_{past}, the fused features can be computed as:

$$
O_t = B_{t_i} \times V_{past}.
\tag{6}
$$

Finally, we apply an auxiliary convolution head ϕ_{seg} to compute a cross-entropy loss:

$$
\mathcal{L}_{aux} = \text{CE}(\phi_{seg}(O_t), G_t),
\tag{7}
$$

in which $\text{CE}(\cdot)$ is the standard cross-entropy loss function.

B. Random Frame Enhancement

To capture long-term dependencies among video frames, a straightforward approach would be to input all frames into the model. However, this would be computationally prohibitive. Instead, we adopt a strategy of randomly selecting a temporally distant frame relative to the target frame. By aggregating information from this randomly selected frame into the target frame's features, we enable the model to develop a more comprehensive understanding of the environment, thereby enhancing model training.

It is worth noting that our STCF module from section III-A integrates features from two closely positioned frames during each operation, gradually incorporating information from neighboring frames. In this context, the random frame selection complements the STCF module. Since the environment depicted in the randomly selected frame may differ from that of the target frame, we employ an attention mechanism to automatically emphasize relevant information while reducing noise.

Suppose that the backbone features of the random frame I_{random} from a P2T [41] pooling layer are denoted as D_{random}^l for the scale l. We aggregate these multi-scale features for the scales from l to L using

$$
\begin{aligned}
D_{random} &= \text{Concat}[D_{random}^l, D_{random}^{l+1}...D_{random}^L], \\
D_{random} &= \text{Linear}(D_{random}).
\end{aligned}
\tag{8}
$$

Inspired by the Object Contextual Representation (OCR) module [42], we propose a contextual representation method that characterizes each pixel by utilizing its corresponding contextual information.

We begin by applying a simple convolution to extract the region-level representation. Next, we sum the pixel representations, weighted by their association with the object region.
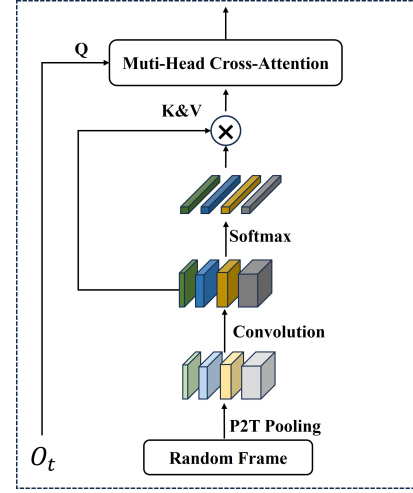

Fig. 5: The architecture of the Random Frame Enhancement module utilizes cross-attention to aggregate contextual information from the randomly selected frame.

Using an aggregation module, we compute the relationships between pixels and object regions to derive enhanced context-level information, as follows:

$$
\begin{aligned}
L_d &= \text{Conv}(D_{random}), \\
L_r &= \text{Softmax}(L_d) \times D_{random}.
\end{aligned}
\tag{9}
$$

Then, for the target frame, we have the short-term temporal representation O_t and the long-term context representation L_r. Next, we apply a multi-head self-attention mechanism to aggregate these features:

$$
\begin{aligned}
Q &= \text{FC}(O_t), K = \text{FC}(L_r), V = \text{FC}(L_r), \\
\hat{O}_t &= \text{Softmax}\left(\frac{Q \times K^T}{\sqrt{C}}\right) \times V,
\end{aligned}
\tag{10}
$$

Now, we have the random frame enhanced feature $\hat{O}_t$, where C denotes the feature dimension.

C. Video Text Encoding Module

Given an image $I \in \mathbb{R}^{H \times W \times 3}$ and a set of candidate class category prompts $\{T(n)\}$ for $n = 1, \ldots, N$, instead of directly using the pretrained text encoder from the CLIP model [7], we enhance its performance for our task by introducing a video-specific prompt scheme.

The core idea is to leverage video context to enrich text representations within the video. A common approach is to directly use the text representation. However, we demonstrate that this can reduce segmentation performance in video contexts. To address this, we introduce a video-specific text encoding module.

In human video comprehension, key terms often help distinguish visual content. For example, including the prompt "water" makes it easier to differentiate between "fish" and "horse" within the visual context.

Specifically, we propose the utilization of a multi-head self-attention (MHSA) mechanism [43] to model the correlation between video clips and text. Given the CLIP image encoder Φ^V and the text encoder Φ^T, we initially extract dense image

TABLE I: The results on the VSPW [12]dataset.

Methods	Backbone	mIoU↑	fwIoU↑	mAcc↑	pAcc↑
Zegformer [40]	ResNet-101	1.91	4.91	6.93	12.21
DeOP [45]	ResNet-101	2.58	6.29	9.33	17.68
SAN [8]	ResNet-101	6.56	12.9	16.45	24.33
zsseg.baseline [46]	ResNet-101	9.46	16.89	20.97	30.21
FreeSeg [47]	ResNet-101	13.13	22.73	29.66	33.11
Ours	ResNet-101	**17.22**	**31.89**	**53.22**	**50.91**
SAM [48]	ViT-B2 [49]	12.67	22.43	**47.32**	33.11
Ours	ViT-B2	**17.99**	**25.22**	34.07	**41.92**

TABLE II: The results on the Cityscapes[13] dataset. The results are obtained by directly testing on the dataset using the model pretrained on VSPW [12] dataset.

Methods	Backbone	mIoU↑	fwIoU↑	mAcc↑	pAcc↑
DeOP [45]	ResNet-101	3.47	5.94	10.04	22.22
FreeSeg [47]	ResNet-101	8.37	3.19	16.96	6.07
SAN [8]	ResNet-101	14.68	28.5	27.59	37.28
Ours	ResNet-101	**27.65**	**41.12**	**71.29**	**64.88**
SAM [48]	ViT-B2 [49]	11.09	20.02	43.13	29.77
Ours	ViT-B2	**26.56**	**39.31**	**63.77**	**57.91**

features $F^V = \Phi^V(I) \in \mathbb{R}^{H \times W \times C}$ for each input frame $I \in [I_{t-k_1}, I_t]$ and text features $F^T = \Phi^T(T_n) \in \mathbb{R}^{N \times C}$ for $n \in \{1, \cdots, N\}$. We refine the text feature F^T with the visual feature F^V using

$$\overline{F^T} = F^T + \text{MHSA}(F^T, F^V). \quad (11)$$

After obtaining refined text feature $\overline{F^T} \in \mathbb{R}^{N \times C}$ and image feature $F^V \in \mathbb{R}^{H \times W \times C}$, inspired by [44], we compute a cost volume $X \in \mathbb{R}^{N \times H \times W}$ by cosine similarity. Formally, it can be defined as:

$$X = \frac{\overline{F^T} \cdot F^V}{||\overline{F^T}|| \cdot ||F^V||}. \quad (12)$$

After constructing the cost volume, we apply a simple convolution on X to refine it, resulting in the enhanced cost volume $\tilde{X}$.

Since CLIP[7] is pretrained on a single image but our model is working with a pixel-level task, it is vital to lead the CLIP to learn the position information. In order to provide corresponding positional information for cost volume, we used shallow features $U_t^{l=1}$ from the backbone, as it contains a large amount of object positional information. We integrated cost volume location data by concatenate it with the existing location information, like

$$\hat{X} = \text{Linear}((\text{Concat}[\tilde{X}, \text{Upsample}(U_t^{l=1})]). \quad (13)$$

After obtaining the position-acquired cost volume $\hat{X}$, we first perform a upsample operation on cost volume $\hat{X}$. We then concatenate the cost volume $\hat{X}$ and refined features $\hat{O}_t$ from section III-B on the channel dimension for final prediction. This process can be expressed as

$$\check{X} = \text{Concat}[\text{Upsample}(\hat{X}), \hat{O}_t]. \quad (14)$$

After that, we use a convolution head to obtain the prediction result. Finally, the training loss is calculated by a standard CE loss and an auxiliary loss from Eq. (7), like:

$$\begin{aligned} \mathcal{L}_{main} &= \text{CE}(\phi_{seg}(\check{X}), G_t), \\ \mathcal{L} &= \alpha \mathcal{L}_{main} + \beta \mathcal{L}_{aux}. \end{aligned} \quad (15)$$

where α and β are trade-off parameters.

D. Complexity Analysis

Here we analyze the complexity of our methods for processing video clips $([I_{t-k_1}, \cdots, I_{t-k_n}, I_{random}])$. The Spatial-Temporal Context Fusion Module has a complexity of $\mathcal{O}((n-1)(hwc^2 + h^2w^2c))$, where h, w, and c represent the height, width, and number of channels in the input frames, respectively. The complexity of the Random Frame Enhancement module has a complexity of $\mathcal{O}(hwc^2 + h^2w^2c)$. For the Video Text Encoding Module, it has a complexity of $\mathcal{O}(n(hwc^2 + h^2w^2c))$.

IV. EXPERIMENTS

A. Datasets

We conduct out experiments on two popular datasets: Video Scene Parsing in the Wild (VSPW) [12] and CityScapes [13].

VSPW. VSPW [12] provides a large-scale benchmark with dense annotations for well-trimmed, long-term video clips, making it one of the most challenging datasets for video semantic segmentation tasks. Its diversity of 124 classes makes it ideal for evaluating open-vocabulary video segmentation. The dataset contains 2806/343/387 videos in the training, validation, and test sets, respectively, with 198244/24502/28887 frames. Each video has an average of 71 frames, with a maximum of 482, resized to 480×853. Our OV-VSS model will be trained on a subset of classes and tested on another.

CityScapes. CityScapes [13] is a widely recognized dataset in the field of video semantic segmentation and has been a key benchmark in earlier research. Notably, for each video, only the 20th frame is annotated at the pixel level. The dataset consists of 5,000 labeled frames, divided into 2975 for training, 500 for validation, and 1525 for testing. This setup makes CityScapes particularly suitable for evaluating models with sparse annotations across video sequences.

B. Experimental Setting

Similar to existing work, we adopt ResNet [50] as the backbone. We set the temporal distances between adjacent frames and the target frame to 3, 6, and 9. We initialize the backbone ResNet with ImageNet pretrained weights, and other parts of the model randomly. We train OV-VSS on VSPW for 10,0000 iterations with a batch size of 4, learning rate initialized as 3e-6. We employ AdamW [51] with weight decay as 1e-2. We first train the model with a linear warm-up for 1500 iterations. We use 1 NVIDIA RTX 3090 GPU for training, costing 20 hours.

For the pretrained CLIP text encoder, the ViT-B/16 [49] backbone is used as default if not specified. We also use prompt templates with class names for all experiments. We follow [40] to generate multiple text embeddings for each class name based on several text prompt templates. Then, we use the mean of these embeddings for final classification.

During training, we adopt random scaling and random cropping for data augmentation. In testing, we perform a single-scale test and use the original 480p image size for inference.

C. Evaluation Protocol.

Following the literature of video semantic segmentation [34, 33], we mainly conduct experiments on VSPW datasets [12], which containing 124 classes. We evaluate our model on their validation set containing 24392 images.

We first compare the State-of-the-Art results on the VSPW dataset. As there are no existing video-based open vocabulary semantic segmentation methods for comparison, we conduct a comparison between image-based open vocabulary methods and our approach. By splitting video clips into continuous frames, we train all models and evaluate them on the test set.

In order to verify the open-vocabulary capability of the model, previous work [40] requires careful selection of visible classes during training and invisible classes during testing according to super categories. We believe that such a setting limits the ability to fully validate the model. To demonstrate the superiority of our model's performance, we directly select the first 80 classes as visible classes during training and the remaining 44 classes as invisible classes during testing, following the category order of VSPW. To prevent the model from learning information about invisible classes during training, we mask out the invisible classes based on their classification, ensuring that the model only learns the features of the predefined visible classes. Furthermore, we mask out the invisible classes in the video frames before feeding them into the model to ensure that the model solely focuses on the visible classes we set.

To verify our cross-dataset performance, we also conducted additional experiments. We directly validated the model trained on the top 80 visible classes in VSPW [12] on the CityScapes [13] dataset. After expert confirmation, 8 categories in CityScapes were new categories that the model had not encountered before.

For evaluation, we adopt four metrics to evaluate the results.
1) mIoU: Mean Intersection over Union (mIoU). mIoU is a standard metric for evaluating image segmentation models. It calculates the average overlap between predicted and ground truth segmentation regions across all classes, balancing precision and recall.
2) FWIoU: Frequency Weighted Intersection over Union. FWIoU evaluates segmentation performance by weighting the IoU of each category by its pixel count, emphasizing classes with more pixels and accounting for class distribution.
3) mAcc: Mean Accuracy. mAcc reports the average accuracy for each class in a dataset, ensuring equal contribution from each class. It is useful for datasets with imbalanced class distributions.
4) pAcc: Pixel-wise Classification Accuracy. pAcc measures the proportion of correctly classified pixels in an image. It provides an overall accuracy assessment but can be biased toward larger classes.

D. Open-Vocabulary Semantic Segmentation Results

Image-based methods. For the image-based methods, Zegformer [40] decouples segmentation and classification by generating class-agnostic segment masks. DeOP [45] proposes a decoupled one-pass structure for computational efficiency. SAN [8] frames the segmentation task as a region recognition problem by attaching a side network. zsseg.baseline [46] presents a two-stage approach to address the discrepancy in processing granularity. FreeSeg [47] introduces a generic framework for universal open-vocabulary image segmentation.

Segment-Anything-Model. Given the lack of text prompt functionality in SAM [48], we designed a novel approach that combines CLIP [7] with the SAM model. This approach aims to establish a baseline for SAM's performance by leveraging CLIP's ability to understand both text and images. Specifically, our method utilizes SAM models to identify regions of interest in video images, after which CLIP is used to classify each identified region into a predefined class. This allows us to adapt SAM for Open-Vocabulary Video Segmentation, enabling the segmentation of novel classes without additional training.

E. Segmentation Results

In this experimental analysis, we begin by training the model using the initial 80 classes for training purposes. During testing, we evaluate its performance on the remaining 44 novel classes. We categorize the methods into two groups based on their backbone architectures.

Regarding segment IoU, our approach outperforms other methods, surpassing the second-best methods by 4.09% and 5.32%, respectively. The state-of-the-art comparisons on the VSPW dataset [12] are presented in Table I.

To validate the zero-shot capability of our model on novel categories, we conduct direct testing on the CityScapes dataset [13] without any modifications. The results, showcased in Table II, underscore the superior performance of our method across both ResNet-101 and ViT backbones.

In addition to the quantitative comparisons mentioned above, we also perform qualitative assessments by juxtaposing our proposed method with the baseline on sampled video clips, as illustrated in Fig. 6. Upon closer inspection, our method consistently generates more accurate segmentation masks across the sampled video clips, reaffirming its efficacy and robustness in real-world scenarios.

F. Ablation Study

We conduct ablation studies on the large-scale VSPW [12] dataset to validate the key designs of OV-VSS and report the results in Table III. For fairness, we maintain the same settings as in section IV-E, unless otherwise specified. The ablation studies are conducted on the ViT-B2 backbone.

1) Ablation Study on Key Components: Results shows the ablation study for each component of OV-VSS. First, we establish a simple baseline that aligns image features with text features. Initially, we train the model in a fully-supervised manner, which is GPU-intensive and requires per-image labels. By integrating a random frame enhancement
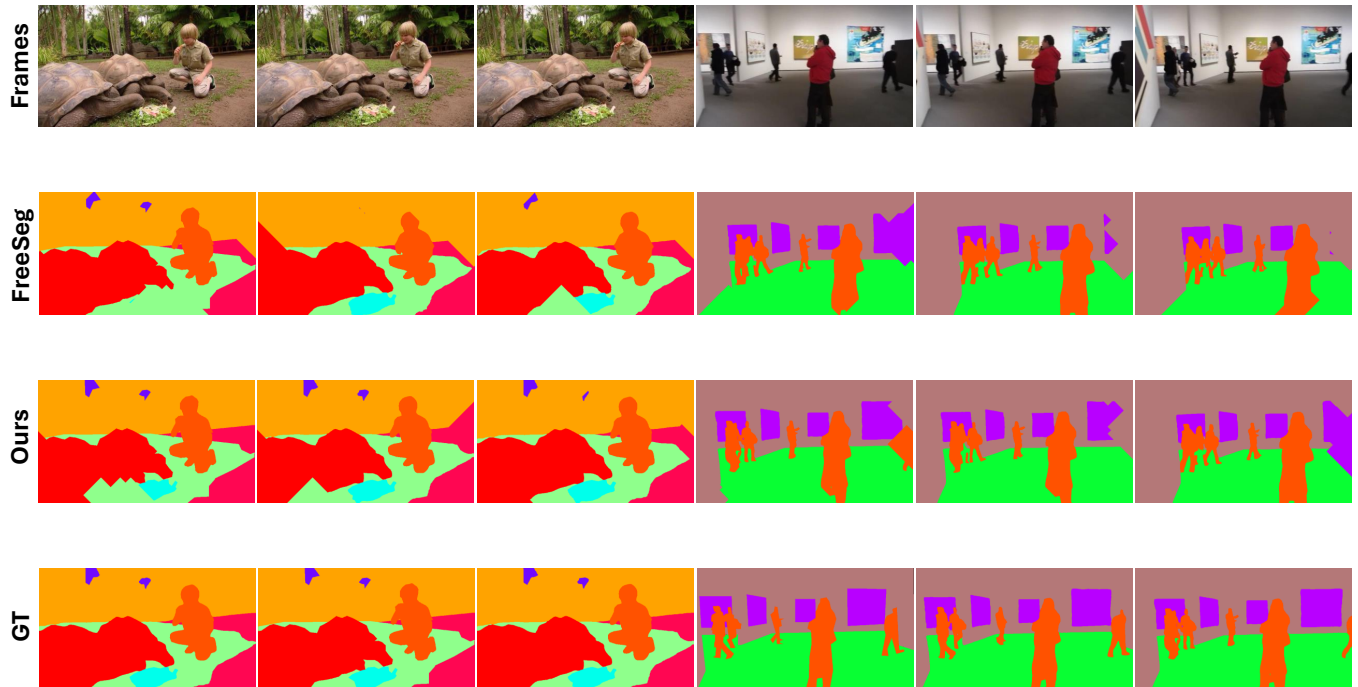
Fig. 6: Qualitative results on the VSPW dataset. From top to bottom: Original frames, segmentation results from the image-based method (FreeSeg), segmentation results from our method (ViT-B), and the ground truth mask. Clearly, our model generates better segmentation results.
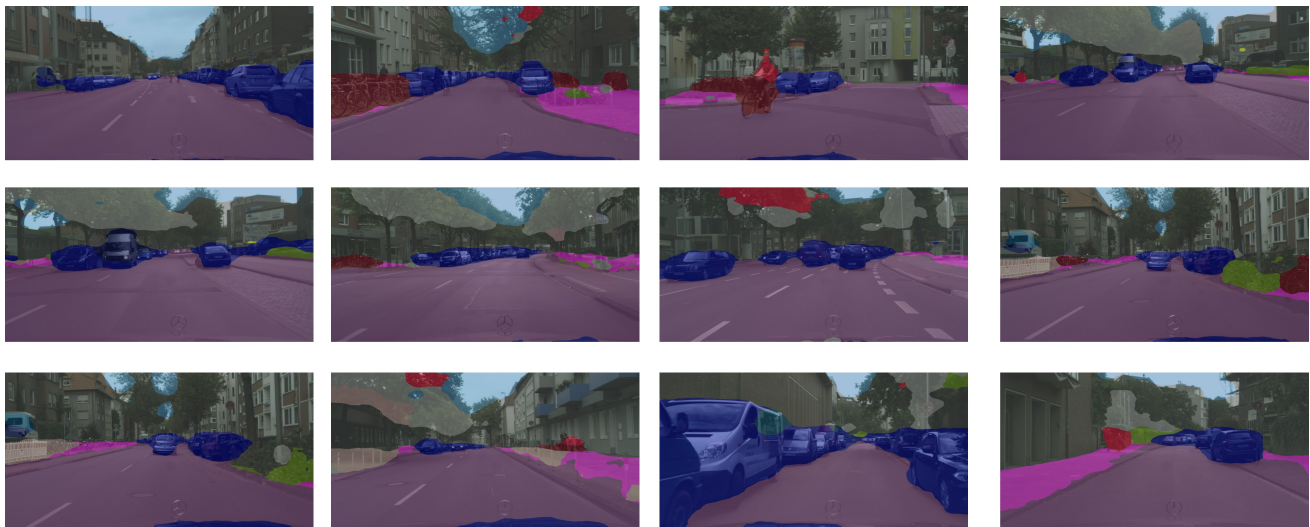

Fig. 7: Qualitative results on CityScapes datasets. We conduct inference on Cityscapes using the model trained on VSPW without further training. Although we have achieved good results on this dataset, our model still cannot accurately segment complex image regions, such as when trees and sky overlap.

module, we achieve better results (+0.91%) with a resource-friendly training mechanism. This method processes frames one by one but loses the temporal information between frames. By utilizing a spatial-temporal context fusion module, we observe an accuracy improvement of 0.43%. Lastly, since CLIP is trained on image-text pairs, which may degrade performance on video-text tasks, we propose a video-based text encoding mechanism that yields an impressive 1.67% performance boost.

2) Ablation Study on Fusing Strategy: We also evaluate the

performance of our method with respect to different temporal fusion strategies. We make the following observations: First, using a simple element-wise addition results in 5.42% mIoU, possibly because simple addition neglects crucial information and diminishes the informative features. Next, by employing a concatenation-based fusion strategy, we achieve 17.99% mIoU.

3) Ablation Study on Random Frame Enhancement: We also study the random frame enhancement module through several methodologies. Initially, we conducted an experiment

wherein the introduction of the random frame was deliberately omitted. This adjustment yielded a segmentation result of 13.55% mIoU. Following this, we proceeded to train the model under a condition referred to as "a-label-per-image", where the model was trained with a single label per image. Under this condition, loss computation was performed individually on each frame, resulting in a segmentation performance of 12.55% mIoU.

4) Ablation Study on the Video Text Encoding Module: We further delve into the efficacy of our video text encoding module through a series of evaluations. Initially, we conducted an experiment wherein we retained solely a multi-head self-attention structure to model the intricate relationship between text and images. This initial configuration yielded a segmentation performance of 14.88% mIoU. Subsequently, we integrated the FFN structure into the encoding module, aiming to enhance the model's capability in capturing text-image interactions. With this enhancement, we observed a notable improvement in segmentation performance, achieving an mIoU of 17.99%.

5) Ablation Study on the loss ratio: As detailed in Table III, we assessed the effectiveness of the loss ratio through a series of experiments. Initially, we focused on adjusting the ratio between the primary and auxiliary losses. In the first experiment, we set α to 0.1 and β to 1, achieving an mIoU of 14.88% on the test set. Increasing α to 1 resulted in a notable performance boost, with an mIoU of 17.99%. However, further increasing α to 10 led to a decline in accuracy, reducing the mIoU to 15.43%. In addition to this, we have further investigated the impact of modifying β alone on the segmentation accuracy, while keeping α at 1 and reducing β to 0.1, the segmentation result was 15.38%. Keeping α at 1 and reducing β to 0.1, the segmentation result was 13.78%.

V. FUTURE WORK

While this work presents a strong foundation for Open Vocabulary Video Semantic Segmentation (OV-VSS), there are several avenues for future research and improvements that could further advance the field:

- **Exploring other Multi-Modal Learning:** Future work could explore integrating additional modalities, such as infrared images [52], depth images [53], or point clouds [54], to enhance segmentation performance. Leveraging these diverse data sources could improve the model's robustness and accuracy, especially in challenging environments where visual information alone is insufficient.
- **Addressing Label Noise in Training and Testing:** In real-world applications, not only might the categories of test samples be unclear, but training labels can also be inaccurate due to errors or inconsistencies in the annotation process [55, 56]. Future work should investigate how our method handles such label noise, particularly when training labels are noisy. While our current approach demonstrates robust performance with clean labels, it is important to explore how the model performs when the training data contains noisy labels, as this is a common scenario in practical deployments.

TABLE III: Various ablation studies on the large-scale VSPW [12] dataset.

Ablation Setting		mIoU(%)
Key Component	Baseline w/o text	14.98
	+ Random Frame Enhancement	15.89
	+ Spatial-Temporal Context Fusion	16.32
	+ Video Text Encoding	**17.99**
Fusing Strategy	Element-Wise Addition	5.42
	Concat on Channel	**17.99**
Random Frame Enhancement	4 frames, 4 annotations	12.25
	4 frames, 1 annotation	13.55
	4 frames + 1 random frame, 1 annotation	**17.99**
Video Text Encoding	Self-Attention	14.88
	Self-attention & FFN	**17.99**
Loss Hyperparameters	$\alpha = 0.1$, $\beta = 1$	14.88
	$\alpha = 1$, $\beta = 1$	**17.99**
	$\alpha = 10$, $\beta = 1$	15.43
	$\alpha = 1$, $\beta = 0.1$	15.38
	$\alpha = 1$, $\beta = 10$	13.78

- **Improving Performance with Low-Quality Input Data:** In scenarios where the input data quality is low, applying image enhancement techniques could improve the model's performance. Future work should explore how pre-processing with enhancement methods [57, 58] affects segmentation accuracy and investigate joint optimization approaches to improve both data quality and segmentation results simultaneously.
- **Few-Shot Segmentation Capabilities:** While our method performs well in the open vocabulary setting, the ability to handle few-shot learning [59, 60] remains an important avenue for future work. Investigating whether the model can further improve performance in scenarios with limited annotated data is crucial.

By addressing these challenges, we believe that OV-VSS could evolve into a more powerful and versatile tool for video analysis, with a broad range of real-world applications, from autonomous systems to interactive media editing.

VI. CONCLUSION

In this study, we explore the task of Open-Vocabulary Video Semantic Segmentation (OV-VSS) and propose a comprehensive end-to-end framework. Our approach introduces several novel components to address the complexities of OV-VSS. First, we develop a Spatial-Temporal Context Fusion module to effectively capture intra-frame correlations. Building on this, a Random Frame Enhancement module refines the target frame based on these correlations, improving segmentation accuracy. Additionally, we incorporate a video-specific text encoding module. Our framework not only addresses open-vocabulary challenges but also establishes a strong baseline for OV-VSS tasks.

ACKNOWLEDGMENT

REFERENCES

[1] Z. Jin, T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, and J. Shao, "Mining contextual information beyond image for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7231–7241.

[2] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7189–7198.

[3] Q. Zhou, L. Wang, G. Gao, K. Bin, W. Ou, and H. Lu, "Boundary-guided lightweight semantic segmentation with multi-scale semantic context," *IEEE Transactions on Multimedia*, 2024.

[4] G. Gao, G. Xu, J. Li, Y. Yu, H. Lu, and J. Yang, "Fbsnet: A fast bilateral symmetrical network for real-time semantic segmentation," *IEEE Transactions on Multimedia*, vol. 25, pp. 3273–3283, 2022.

[5] B. Yan, X. Niu, B. Bare, and W. Tan, "Semantic segmentation guided pixel fusion for image retargeting," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 676–687, 2019.

[6] J. Gou, X. Zhou, L. Du, Y. Zhan, W. Chen, and Z. Yi, "Difference-aware distillation for semantic segmentation," *IEEE Transactions on Multimedia*, 2024.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[8] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.

[9] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 540–557.

[10] H. Wang, C. Yan, S. Wang, X. Jiang, X. Tang, Y. Hu, W. Xie, and E. Gavves, "Towards open-vocabulary video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4057–4066.

[11] P. Guo, T. Huang, P. He, X. Liu, T. Xiao, Z. Chen, and W. Zhang, "Openvis: Open-vocabulary video instance segmentation," *arXiv preprint arXiv:2305.16835*, 2023.

[12] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "VSPW: A large-scale dataset for video scene parsing in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4133–4143.

[13] 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, pp. 3213–3223.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 746–760.

[15] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 44–57.

[16] 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, pp. 5997–6005.

[17] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 852–868.

[18] B. Mahasseni, S. Todorovic, and A. Fern, "Budget-aware deep semantic video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1029–1038.

[19] 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, pp. 2349–2358.

[20] 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*, 2018, pp. 6556–6565.

[21] 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, pp. 8866–8875.

[22] 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, pp. 8818–8827.

[23] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 352–368.

[24] M. Paul, M. Danelljan, L. Van Gool, and R. Timofte, "Local memory attention for fast video semantic segmentation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 1102–1109.

[25] A. Kundu, V. Vineet, and V. Koltun, "Feature space optimization for semantic video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3168–3175.

[26] J. Li, W. Wang, J. Chen, L. Niu, J. Si, C. Qian, and L. Zhang, "Video semantic segmentation via sparse temporal transformer," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 59–68.

[27] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video CNNs through representation warping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 4453–4462.

[28] 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/CVF International Conference on Computer Vision*, 2017, pp. 5580–5588.

[29] 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, pp. 413–421.

[30] 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, pp. 6819–6828.

[31] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865.

[32] Y. Zhang, S. Borse, H. Cai, and F. Porikli, "AuxAdapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 2339–2348.

[33] G. Sun, Y. Liu, H. Ding, T. Probst, and L. Van Gool, "Coarse-to-fine feature mining for video semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3126–3137.

[34] G. Sun, Y. Liu, H. Tang, A. Chhatkuli, L. Zhang, and L. Van Gool, "Mining relations among cross-frame affinities for video semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 522–539.

[35] G. Sun, Y. Liu, H. Ding, M. Wu, and L. Van Gool, "Learning local and global temporal contexts for video semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, "Video K-Net: A simple, strong, and unified baseline for video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 847–18 857.

[37] X. Li, H. Yuan, W. Zhang, G. Cheng, J. Pang, and C. C. Loy, "Tube-Link: A flexible cross tube baseline for universal video segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 923–13 933.

[38] Y. Weng, M. Han, H. He, M. Li, L. Yao, X. Chang, and B. Zhuang, "Mask propagation for efficient video semantic segmentation," in *Proceedings of the Conference on Neural Information Processing Systems*, 2023, pp. 7170–7183.

[39] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 4904–4916.

[40] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 583–11 592.

[41] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 760–12 771, 2022.

[42] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 173–190.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[44] S. Cho, H. Shin, S. Hong, S. An, S. Lee, A. Arnab, P. H. Seo, and S. Kim, "CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation," *arXiv preprint arXiv:2303.11797*, 2023.

[45] C. Han, Y. Zhong, D. Li, K. Han, and L. Ma, "Open-vocabulary semantic segmentation with decoupled one-pass network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1086–1096.

[46] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 736–753.

[47] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "FreeSeg: Unified, universal and open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 446–19 455.

[48] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[51] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International*

Conference on Learning Representations, 2014, pp. 1–15.

[52] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-irtarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Transactions on Multimedia*, vol. 24, pp. 1735–1749, 2021.

[53] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.

[54] L. Tang, Y. Zhan, Z. Chen, B. Yu, and D. Tao, "Contrastive boundary learning for point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8489–8499.

[55] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[56] R. Zhang, Z. Cao, S. Yang, L. Si, H. Sun, L. Xu, and F. Sun, "Cognition-driven structural prior for instance-dependent label transition matrix estimation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[57] A. Li, L. Zhang, Y. Liu, and C. Zhu, "Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12514–12524.

[58] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17452–17462.

[59] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9197–9206.

[60] R. Zhang, J. Tan, Z. Cao, L. Xu, Y. Liu, L. Si, and F. Sun, "Part-aware correlation networks for few-shot learning," *IEEE Transactions on Multimedia*, 2024.