

# Shift and matching queries for video semantic segmentation

Tsubasa Mizuno<sup>1</sup> and Toru Tamaki<sup>1</sup>[0000–0001–9712–7777]

Nagoya Institute of Technology, Nagoya, Japan  
 t.mizuno.384@nitech.jp, tamaki.toru@nitech.ac.jp

**Abstract.** Video segmentation is a popular task, but applying image segmentation models frame-by-frame to videos does not preserve temporal consistency. In this paper, we propose a method to extend a query-based image segmentation model to video using feature shift and query matching. The method uses a query-based architecture, where decoded queries represent segmentation masks. These queries should be matched before performing the feature shift to ensure that the shifted queries represent the same mask across different frames. Experimental results on CityScapes-VPS and VSPW show significant improvements from the baselines, highlighting the method’s effectiveness in enhancing segmentation quality while efficiently reusing pre-trained weights.

**Keywords:** semantic segmentation · vision transformer · feature shift · adaptation

## 1 Introduction

Segmentation [22] involves predicting the labels for objects or scene elements for each pixel in an image. This is useful in many applications, such as understanding urban landscapes in automated driving technology, and research has attracted much attention. Hence, various *image* segmentation models have been proposed [3, 4, 8, 10, 13, 22]. However, in the case of *video* segmentation, applying image segmentation models frame-by-frame does not preserve temporal consistency. Therefore, capturing the temporal relationship between frames is necessary, and many segmentation methods have been developed specifically for videos.

There are two major problems in the development of video segmentation models. The first is the increased computational complexity and time required to process videos. Unlike image segmentation, which deals with a single image, video segmentation requires handling many frames efficiently. Therefore, methods suitable for real-time processing [23] and online processing [28, 29] have been proposed. The second is the effective use of image segmentation models. Currently, many pre-trained image models are available through platforms such as Hugging Face [27]<sup>1</sup>. However, designing new models for large video datasets and training them is an expensive process. Therefore, many tasks have been

---

<sup>1</sup> <https://huggingface.co>

studied by reusing weights from image models and applying them to video models [19, 24, 30]. Although methods for video segmentation have been proposed [17, 29], they do not efficiently reuse pre-trained weights.

In this study, we propose a video segmentation method that effectively uses an image segmentation model pre-trained on image datasets. The proposed model not only applies the image segmentation model to each frame, but also models temporal information using feature shift. This feature shift is widely used in action recognition [7, 18, 30], but it is the first time it has been introduced in a segmentation model. The segmentation model that we based on is a query-based method [4], which represents a segmentation mask by a single trainable *query*. However, simple feature shift and query-based methods are incompatible. This is because a feature shift exchanges features of the previous and next frames, but if applied to queries, it might shift queries representing different segmentation masks. In this study, we propose a query matching before performing the feature shift and demonstrate its effectiveness.

## 2 Related Work

### 2.1 Image Segmentation

Image segmentation models using CNNs, like as U-Net [25], typically include an encoder that takes an input image and reduces its spatial resolution, and a decoder that increases the spatial resolution and outputs a label image of the same size as the input image. However, after the success of DETR [2], a query-based object detection method using Transformer [26], methods that incorporate the idea of object query in DETR have been proposed for tasks beyond object detection, leading to many new query-based segmentation methods [3, 4, 10, 15]. In this study, we adopt MaskFormer [4] as the backbone and extend it to video using feature shift.

### 2.2 Video Segmentation

Various methods have been proposed for modeling temporal information in video segmentation. These include aggregating information from previous frames [11], using multiresolution information by changing the frame sampling stride [14], using pseudolabels for frames, and employing attention to past frames [9, 16].

However, all of these models are specific to video segmentation and no study has efficiently reused image segmentation models as in this study. Previous studies that extended query-based methods to video include Tube-Link [17] and TarViS [1] for universal video segmentation. These require matching between multiple queries across frames, making the mechanism very complex. In contrast, this study aims for an extension method that is easy to implement and analyze, by matching queries and applying temporal feature shift.

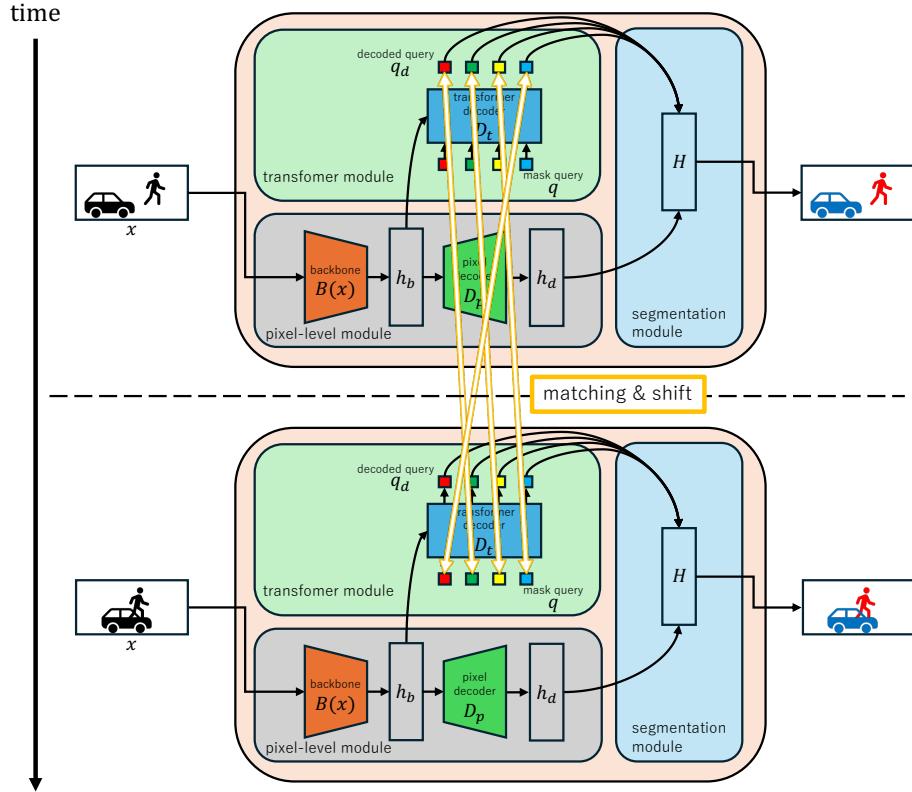


Fig. 1: The architecture of our proposed method includes feature shift and query matching. It utilizes an image segmentation model (the orange plate) frame-by-frame, which has a backbone for feature extraction, a pixel decoder, a transformer decoder for processing mask queries, and a segmentation module for prediction. The proposed query matching ensures temporal consistency even when feature shift is applied to decoded queries of different frames.

### 3 Method

In this section, we first describe the architecture common to query-based segmentation models and then explain how we propose to extend it to videos by using feature shift and query matching.

#### 3.1 Segmentation architecture

Semantic segmentation models for images often consist of a backbone  $B(x)$  that receives an input image  $x$ , a pixel decoder  $D_p(h_b)$  that transforms the backbone features  $h_b = B(x)$  back to the input image size, a transformer decoder  $D_t(h_b, q)$

that transforms *mask queries*  $q$  based on  $h_b$ , and a header  $H(h_d, q_d)$  that predicts segmentation masks and classes using the *decoded queries*  $q_d = D_t(h_d, q)$  and the pixel decoder output  $h_d$ .

When applied to video, each frame  $v(t), t = 1, \dots, T$  of an input video  $v$  is fed  $B$  (Fig.1). However, this naive frame-by-frame approach does not preserve temporal coherency.

### 3.2 Feature shift

In this study, the feature shift is used to exchange temporal information in a video, by shifting the features of specific channels of a model in the temporal direction [18].

Suppose  $z_{in} \in \mathbb{R}^{T \times D}$  is an input feature of dimension  $D$ , which is a stack of frame features of a video clip of  $T$  frames. In the following,  $z_{in,t,d}$  denotes the element at  $(t, d)$  in  $z_{in}$ . It is fed to shifting modules [7] to produce output  $z_{out} \in \mathbb{R}^{T \times D}$ , a temporally shifted version of the input as follows;

$$z_{out,t,d} = \begin{cases} z_{in,t-1,d}, & \text{if } 1 < t \leq T, 1 \leq d \leq D_f \\ z_{in,t+1,d}, & \text{if } 1 \leq t < T, D - D_b \leq d \leq D \\ z_{in,t,d}, & \text{if } \forall t, D_f + 1 \leq d \leq D - D_b - 1. \end{cases} \quad (1)$$

This means that first  $D_f$  channels are shifted forward to the next frame  $t + 1$ , the last  $D_b$  channels are shifted backward to the previous frame  $t - 1$ , and the remaining channels are left untouched. Apparently, this feature shift only works when the channel indexes have the same meaning. However, queries in query-based methods may refer to different things or stuff (i.e., segmentation masks) for each frame. Therefore, a naive shift does not maintain feature compatibility between frames.

In this experiment, we have choices for the features that are shifted in the backbone  $B$ , pixel decoder  $D_p$ , transformer decoder  $D_t$ , or queries  $q$  and  $q_d$ . Among these, shifting the decoded mask queries  $q_d$  is reasonable because it directly affects the quality of the output segmentation mask, and our preliminary experiments have shown a small impact on performance when we apply the feature shift to other modules.

### 3.3 Query matching

Since each decoded mask query corresponds to a segmentation mask, simply shifting the feature of the queries with the same index across frames might mix up features of queries that actually represent different segmentation masks in different frames. To address this, we propose a method to match queries across frames that are likely to correspond to the same segmentation mask (see Fig.1).

First, we compute cosine similarities for each pair of  $N$  queries in adjacent frames. Next, we match the queries by solving a bipartite matching problem using the Hungarian algorithm to find the optimal permutation  $\hat{\sigma}$  as follows:

$$\hat{\sigma} = \operatorname{argmin}_{\sigma \in \Omega} \sum_i \operatorname{sim}(q_i^t, q_{\hat{\sigma}(i)}^{t+1}), \quad (2)$$

where  $\text{sim}()$  is cosine similarity between the  $i$ -th query  $q_i^t$  at time  $t$  and the query  $q_{\hat{\sigma}(i)}^{t+1}$  of index  $\hat{\sigma}(i)$  at time  $t+1$ , and  $\Omega$  is a set of permutations. This query matching is performed on all adjacent frames, and then feature shift is applied to the matched queries.

## 4 Experiments

### 4.1 Datasets

**Cityscapes-VPS** [12] is a dataset for Video Panoptic Segmentation (VPS) built upon Cityscapes [5], which is a well-known dataset for image panoptic segmentation but only has 5,000 finely annotated frames from 5,000 videos. In contrast, Cityscapes-VPS sampled every five frames from 500 Cityscapes videos, making it suitable for our purpose because the temporal feature shift is suitable for the denser frame sampling. The resulting dataset contains 2,400 finely annotated frames for training and 500 for validation, with 19 categories at a resolution of  $1,024 \times 2,048$ . There are six annotated frames in each video, with approximately 0.29 seconds between each frame, since the frame rate of the videos is 17 fps. Although instance IDs are also associated, we do not use them because this paper focuses on video semantic segmentation.

**VSPW** (Video Scene Parsing in the Wild) [21] is a dataset that includes full annotations for 3,337 videos, each approximately 3 to 10 seconds (5 seconds on average), recorded at 15 fps with a frame interval of approximately 0.067 seconds. Pixel-level annotations cover 124 categories for 239,934 frames, with resolutions ranging from 720P to 4K. Because the length varies by video, in this experiment, we used only videos with a minimum length of 32 frames and excluded shorter videos. Consequently, 161,984 frames were used for training and 18,944 frames for validation.

### 4.2 Model

In this experiment, we use MaskFormer [4], a simple query-based architecture for images. It consists of a backbone that encodes the input image, a pixel decoder, a transformer decoder that transforms mask queries, and a segmentation head, as mentioned in Section 3.1. The pixel-level module, which includes the backbone and the pixel decoder, extracts visual embeddings used to generate binary mask predictions. The transformer decoder computes segment-level embeddings from mask queries and generates decoded queries. The segmentation head generates predictions from the visual embeddings and the decoded queries. We applied the feature shift with or without query matching for the decoded queries. The dimension of a query is  $D = 256$ , and we specify the amount of feature shift as a fraction [18] ranging from  $1/128$  (two channels are shifted; one channel forward and one channel backward) to  $1/4$  (64 channels are shifted) of the dimension  $D$ .

Table 1: Performance comparison on CityScapes-VPS for different shift channels and query-matching configurations. Performance differences from the baseline (no shifts) is shown in parentheses.

shift				
fraction	channels	matching	mIoU	pix. acc.
0	0		55.66	92.91
$\frac{1}{128}$	2	✓	53.40 (-2.26)	92.28 (-0.63)
			53.26 (-2.40)	92.71 (-0.20)
$\frac{1}{64}$	4	✓	50.69 (-4.97)	91.35 (-1.56)
			56.61 (+0.95)	92.68 (-0.23)
$\frac{1}{32}$	8	✓	51.49 (-4.17)	91.86 (-1.05)
			<b>57.68 (+2.02)</b>	<b>93.23 (+0.32)</b>
$\frac{1}{16}$	16	✓	54.22 (-1.44)	92.03 (-0.88)
			56.23 (+0.63)	92.57 (-0.34)
$\frac{1}{8}$	32	✓	54.10 (-1.56)	92.21 (-0.70)
			56.88 (+1.22)	92.71 (-0.20)
$\frac{1}{4}$	64	✓	55.92 (+0.26)	92.42 (-0.49)
			53.73 (-1.93)	92.56 (-0.35)

**Training setup.** In this experiment, we used MaskFormer [4] pre-trained on ADE20K [31] as the image model and applied the proposed method to the video datasets. The experimental settings for Cityscapes-VPS [12] are a crop size of  $256 \times 512$ , 50 epochs, a learning rate of 1e-3, and a batch size of 6 frames (i.e., one video). For VSPW [21], the crop size is  $288 \times 512$ , 10 epochs, a learning rate of 5e-4, and a batch size of 16 frames. The evaluation metrics used are mean Intersection over Union (mIoU) [6] and pixel accuracy [20].

### 4.3 Quantitative results

Table 1 shows the performance on CityScapes-VPS for different configurations. The top row shows the baseline configuration (no shifts, no query matching) as a reference, with differences from the baseline shown in parentheses in the following rows. Without query matching, the mIoU generally decreases as the amount of shift increases from  $1/128$  to  $1/32$ . Pixel accuracy also shows a slight decline. However, for fractions lower than  $1/32$  (or more channels are shifted), both indicators rise to the same level as the baseline. Introducing the query matching with a shift between  $1/64$  to  $1/8$  consistently outperforms the baseline and their non-matching counterparts in both mIoU and pixel accuracy. In particular, the  $1/32$  shift with query matching outperforms the baseline by about 2%, compared to a 6% decrease without query matching. This indicates that the introduction of feature shift and query matching is effective and can significantly improve performance when the amount of feature shift is appropriately selected for the dataset, since making the shift fraction very small or very large decreases performance in both cases.

Table 2: Performance comparison on VSPW for different shift channels and query-matching configurations.

shift			mIoU	pix. acc.
fraction	channels	matching		
0	0		56.85	87.66
$1/128$	2	✓	54.27 ( $-2.57$ ) 58.22 ( $+2.72$ )	87.66 ( $\pm 0.00$ ) 88.35 ( $+0.98$ )
$1/64$	4	✓	57.26 ( $+1.21$ ) 57.20 ( $+1.70$ )	87.99 ( $+0.62$ ) 88.97 ( $+1.60$ )
$1/32$	8	✓	57.66 ( $+0.81$ ) 57.36 ( $+1.86$ )	88.79 ( $+1.13$ ) 87.77 ( $+0.40$ )
$1/16$	16	✓	58.44 ( $+1.59$ ) <b>58.68</b> ( $+3.18$ )	87.96 ( $+0.30$ ) 88.37 ( $+1.00$ )
$1/8$	32	✓	57.59 ( $+0.74$ ) 58.29 ( $+2.79$ )	88.17 ( $+0.80$ ) <b>89.08</b> ( $+1.71$ )
$1/4$	64	✓	57.82 ( $+2.32$ ) 55.52 ( $-1.33$ )	88.18 ( $+0.81$ ) 87.99 ( $+0.33$ )

The performance on VSPW is shown in Table 2, where the proposed method demonstrated improvements compared to Table 1. One notable observation is that the performance on VSPW improves even without query matching. However, using query matching further enhances the performance, with mIoU consistently better for a shift between  $1/128$  and  $1/8$ . The significant improvement on VSPW may be attributed to its nature as a denser video dataset. While CityScapes-VPS videos have a frame interval of 0.29 seconds, VSPW videos have a much shorter interval of 0.067 seconds. This results in very little difference in appearance between frames, making the feature shift highly effective.

#### 4.4 Qualitative results

Figure 2 shows examples of segmentation results for CityScapes-VPS with and without query matching for a shift of  $1/32$  and  $1/16$ . As seen in Table 1, applying the proposed query matching improves performance, particularly in the “sidewalk” category. Without query matching, a large part of the sidewalk is missing in the fourth frame of Fig. 2(g) and the third frame of Fig. 2(e). These missing parts are improved with query matching, as shown in Fig. 2(f) and Fig. 2(d). In the first frame of the shift of  $1/16$ , the small “person” region (in red), visible far ahead (or on the right side in the image) of the car (blue) in the center, was misidentified as a “tree” (green) without matching (Fig. 2(g)). This was improved with matching, as shown in Fig. 2(f), which demonstrates the effect of the feature shift between the same objects across the frames by the proposed query matching.

Figure 3 shows examples of segmentation results for VSPW with and without query matching for a shift of  $1/16$  and  $1/8$ . In frames 4 to 6, an incorrect segmentation of “road” (dark red) was observed as “path” (dark yellow) without

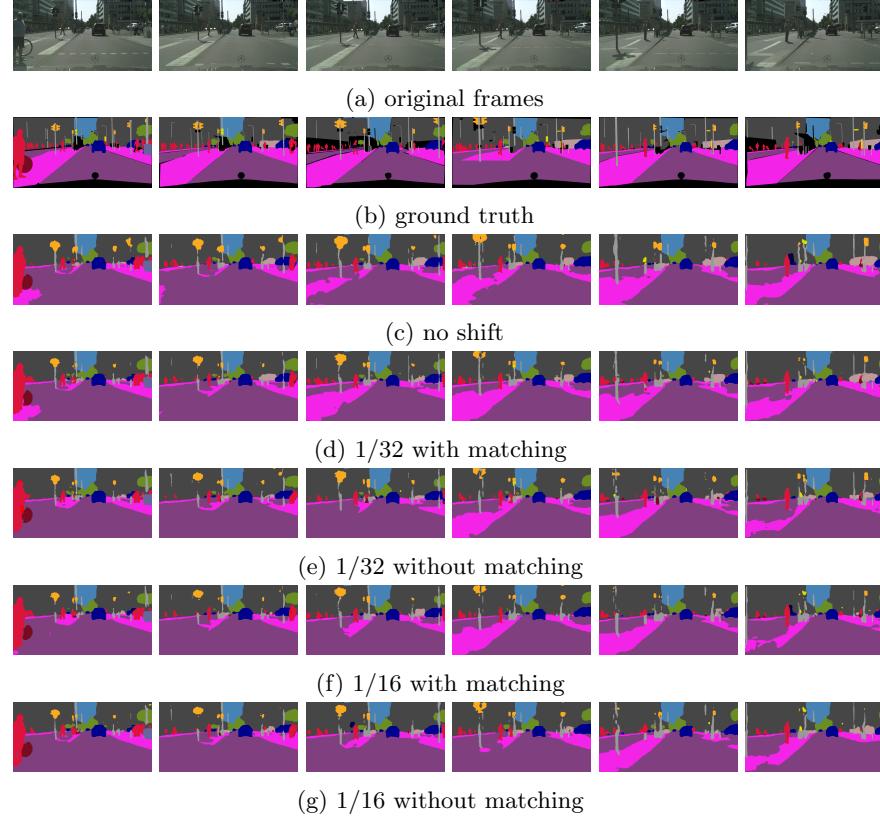


Fig. 2: Segmentation results for CityScapes-VPS. First two rows show (a) original frames, and (b) ground truth segmentation labels. The following rows show results of (c) the baseline (no shifts, no query matching), (d) 1/32 shift with and (e) without matching, (f) 1/16 shift with and (g) without matching.

shift or with a small shift. This was corrected by increasing the shift fraction to 1/16. Even with a shift of 1/16, performance was still poor without matching (for example, “building” was confused with “wall”). However, improvement was observed with the proposed query matching, which is also seen in the category “billboard\_or\_Bulletin\_Board,” suggesting that appropriate matching of queries leads to better results.

## 5 Conclusion

In this paper, we propose a video segmentation method that leverages pre-trained image segmentation models and incorporates feature shift to model temporal information. The feature shift, commonly used in action recognition, is applied to segmentation for the first time. The proposed method uses a query-based

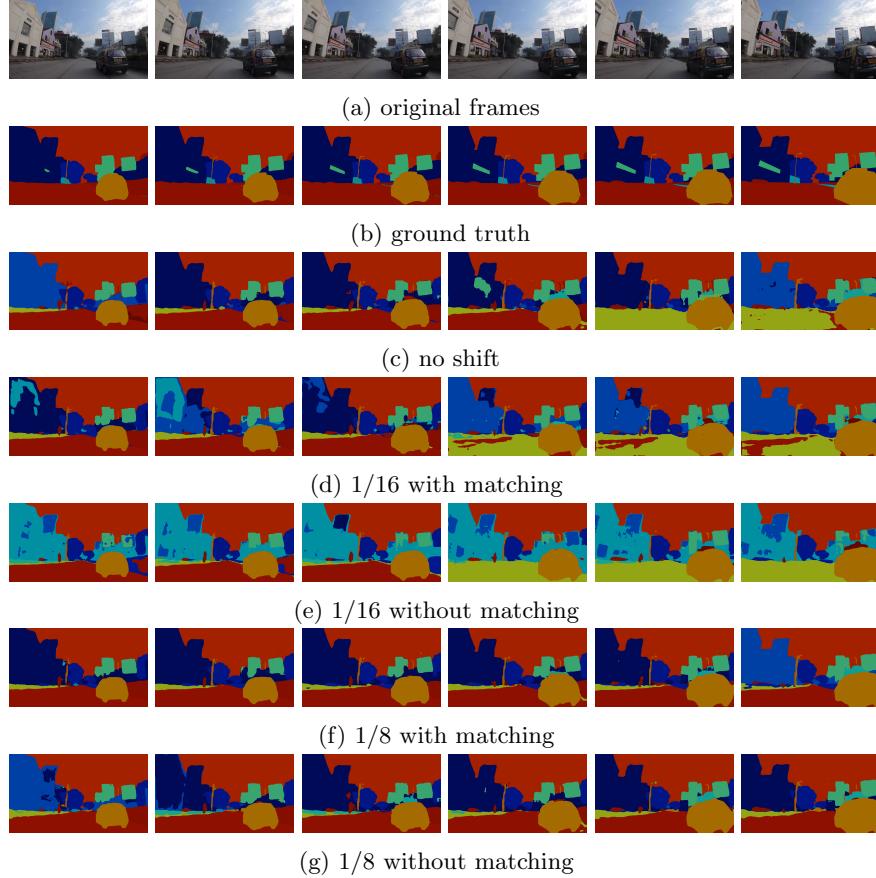


Fig. 3: Segmentation results for VSPW. First two rows show (a) original frames, and (b) ground truth segmentation labels. The following rows show results of (c) the baseline (no shifts, no query matching), (d) 1/16 shift with and (e) without matching, (f) 1/8 shift with and (g) without matching.

architecture, where we find correspondences between decoded queries representing segmentation masks in different frames. Using the proposed query matching, we perform feature shift to maintain coherency across frames. Our results demonstrate significant improvements in performance, especially on dense video datasets like VSPW, highlighting the effectiveness of the proposed method in enhancing segmentation quality while efficiently reusing pre-trained weights.

Future work includes improving the performance of the proposed method and comparing it with other video segmentation methods. Instead of developing an entirely new video model, we will explore approaches that retain the core architecture, applying a segmentation model to images frame by frame, with a detailed yet straightforward extension.

## Acknowledgements

The authors thank Shimon Hori for his help in implementing the shift to a query-based architecture.

This work was supported in part by JSPS KAKENHI Grant Number JP22K12090.

## References

1. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18738–18748 (June 2023) [2](#)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020) [2](#)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1290–1299 (June 2022) [1](#), [2](#)
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 17864–17875. Curran Associates, Inc. (2021), [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/950a4152c2b4aa3ad78bdd6b366cc179-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/950a4152c2b4aa3ad78bdd6b366cc179-Paper.pdf) [1](#), [2](#), [5](#), [6](#)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) [5](#)
6. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (Jan 2015) [6](#)
7. Hashiguchi, R., Tamaki, T.: Temporal cross-attention for action recognition. In: Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops. pp. 276–288 (December 2022), [https://openaccess.thecvf.com/content/ACCV2022W/TCV/html/Hashiguchi\\_Temporal\\_Cross-attention\\_for\\_Action\\_Recognition\\_ACCVW\\_2022\\_paper.html](https://openaccess.thecvf.com/content/ACCV2022W/TCV/html/Hashiguchi_Temporal_Cross-attention_for_Action_Recognition_ACCVW_2022_paper.html) [2](#), [4](#)
8. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) [1](#)
9. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) [2](#)
10. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2989–2998 (June 2023) [1](#), [2](#)
11. Jain, S., Wang, X., Gonzalez, J.E.: Accel: A corrective fusion network for efficient semantic segmentation on video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) [2](#)

12. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) [5](#), [6](#)
13. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) [1](#)
14. Lao, J., Hong, W., Guo, X., Zhang, Y., Wang, J., Chen, J., Chu, W.: Simultaneously short- and long-term temporal modeling for semi-supervised video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14763–14772 (June 2023) [2](#)
15. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3041–3050 (2023). <https://doi.org/10.1109/CVPR52729.2023.00297> [2](#)
16. Li, J., Wang, W., Chen, J., Niu, L., Si, J., Qian, C., Zhang, L.: Video semantic segmentation via sparse temporal transformer. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 59–68. MM ’21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3474085.3475409>, <https://doi.org/10.1145/3474085.3475409> [2](#)
17. Li, X., Yuan, H., Zhang, W., Cheng, G., Pang, J., Loy, C.C.: Tube-link: A flexible cross tube framework for universal video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13923–13933 (October 2023) [2](#)
18. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) [2](#), [4](#), [5](#)
19. Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 388–404. Springer Nature Switzerland, Cham (2022) [2](#)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015) [6](#)
21. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: 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 (CVPR). pp. 4133–4143 (June 2021) [5](#), [6](#)
22. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3523–3542 (2022). <https://doi.org/10.1109/TPAMI.2021.3059968> [1](#)
23. Park, H., Yessenbayev, A., Singhal, T., Adhikari, N.K., Zhang, Y., Borse, S.M., Cai, H., Pandey, N.P., Yin, F., Mayer, F., Calidas, B., Porikli, F.: Real-time, accurate, and consistent video semantic segmentation via unsupervised adaptation and cross-unit deployment on mobile device. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21431–21438 (June 2022) [1](#)
24. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: 2023 IEEE/CVF Conference on

- Computer Vision and Pattern Recognition (CVPR). pp. 6545–6554 (2023). <https://doi.org/10.1109/CVPR52729.2023.00633> 2
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015) 2
  26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), [https://proceedings.neurips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html) 2
  27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>, <https://aclanthology.org/2020.emnlp-demos.6> 1
  28. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2761–2770 (October 2023) 1
  29. Ying, K., Zhong, Q., Mao, W., Wang, Z., Chen, H., Wu, L.Y., Liu, Y., Fan, C., Zhuge, Y., Shen, C.: Ctviz: Consistent training for online video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 899–908 (October 2023) 1, 2
  30. Zhang, H., Hao, Y., Ngo, C.W.: Token shift transformer for video classification. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 917–925. MM '21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3474085.3475272>, <https://doi.org/10.1145/3474085.3475272> 2
  31. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019) 6