PanoVOS: Bridging Non-panoramic and Panoramic Views with Transformer for Video Segmentation

Shilin Yan[1], Xiaohao Xu[2], Renrui Zhang[3], Lingyi Hong[1], Wenchao Chen[1], Wenqiang Zhang[1], and Wei Zhang[1⋆]

[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2] University of Michigan, Ann Arbor
[3] MMLab CUHK
[4] tattoo.ysl@mail.com weizh@fudan.edu.cn

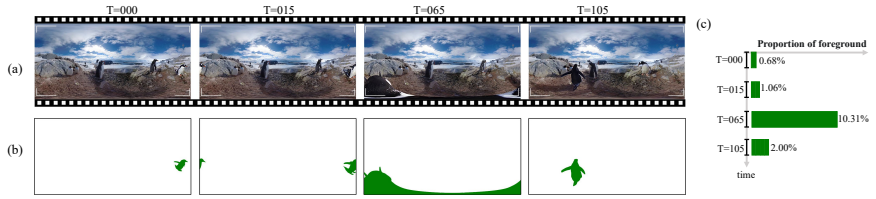

Fig. 1: Panoramic video object segmentation (PanoVOS). PanoVOS targets tracking and distinguishing the particular instances under content discontinuities (*e.g.* penguin in the image of $T = 15$) and serve distortion (*e.g.* penguin in the image of $T = 65$). We show the sample of (a) frames, (b) segmentation annotations, and (c) area proportion of foreground for the *Penguin* video in our dataset.

Abstract. Panoramic videos contain richer spatial information and have attracted tremendous amounts of attention due to their exceptional experience in some fields such as autonomous driving and virtual reality. However, existing datasets for video segmentation only focus on conventional planar images. To address the challenge, in this paper, we present a panoramic video dataset, *i.e.*, PanoVOS. The dataset provides 150 videos with high video resolutions and diverse motions. To quantify the domain gap between 2D planar videos and panoramic videos, we evaluate 15 off-the-shelf video object segmentation (VOS) models on PanoVOS. Through error analysis, we found that all of them fail to tackle pixel-level content discontinues of panoramic videos. Thus, we present a Panoramic Space Consistency Transformer (PSCFormer), which can effectively utilize the semantic boundary information of the previous frame for pixel-level matching with the current frame. Extensive experiments demonstrate that compared with the previous SOTA models, our PSCFormer network exhibits a great advantage in terms of segmentation results under the panoramic setting. Our dataset poses new challenges in panoramic VOS and we hope that

⋆ Corresponding author.

our PanoVOS can advance the development of panoramic segmentation/tracking. The dataset, codes, and pre-train models will be published at https://github.com/shilinyan99/PanoVOS.

1 Introduction

Semi-supervised video object segmentation (VOS) [52], which targets tracking and distinguishing the particular instances across the entire video sequence based on the first frame masks, plays an essential role in video understanding and editing. Conventionally, the images or videos studied in VOS are 2D planar data with a limited Field of View (FoV), which may lead to some ambiguities, especially when objects are out of view. Meanwhile, with the rapid development of VR/AR collection devices [12, 22], panoramic videos with a $360^\circ \times 180^\circ$ FoV are able to collect the entire viewing sphere and richer spatial information [1, 21, 27, 60]. To the best of our knowledge, we are the first to attempt to tackle the promising but challenging task of panoramic video object segmentation.

To foster the development of panoramic VOS, we propose a new dataset in this work, aiming at panoramic video object segmentation. The dataset contains a wide range of real-world scenarios in which scenes have a large magnitude of motion. The main characteristics of our dataset are three aspects. 1) Panoramic videos bring certain advantages (richer geometric information and wider FoV) in real-world applications as well as challenges (serve distortion and content discontinuities). 2) Compared to all existing VOS datasets, our dataset has longer video clips with an average length of 20 seconds. 3) Nearly half of the video resolutions in our dataset are $4K$, which may help facilitate broader video tracking/segmentation research under the high-resolution scenario.

In the proposed dataset, we annotated 150 videos with 19,145 annotated instance masks, including sports (*e.g.* parkour, skateboard), animals (*e.g.* elephant, monkey), and common objects (basketball, hot balloon). Since, annotating a pixel-level intensive task is very time-consuming and expensive, we proposed a semi-supervised human-computer joint annotation strategy. Concretely, we first annotated objects at selected keyframes (1 fps) Then we adopted the state-of-the-art video object segmentation model AOT [58] for mask propagation to the rest frames of videos and we manually refine parts of them.

Then, we conducted extensive experiments on PanoVOS to evaluate 15 off-the-shelf video object segmentation models. The results suggest that existing approaches can not handle several domain-unique challenges. The first is content discontinuities, which means the foreground object may be separated in the left and the right boundaries of the planar image, such as the case in the image of $T = 15$ in Fig. 1. The second is the severe distortions and deformations, such as the case in the image of $T = 65$ in Fig. 1.

To tackle these challenges of panoramic video segmentation, we proposed a PSCFormer model which consists of key component Panoramic Space Consistent (PSC) blocks. The PSC block is designed for constructing spatial-temporal class-agnostic correspondence and propagating the segmentation masks. Each PSC

block utilizes a cross-attention for matching with references' embeddings and a PSC-attention for modeling the boundary semantic relationship between the previous frame and the query frame. Hence, the network can effectively alleviate the problem that the left and the right boundaries are actually continuous in panoramic videos. Our method outperforms the SOTA models that are re-trained on PanoVOS train set in segmentation quality under the panoramic setting.
Our contributions are three-fold.

- We introduce a panoramic video object segmentation dataset (PanoVOS) with 150 videos and 19K annotated instance masks, which fills the gap of long-term instance-level annotated panoramic video segmentation datasets.
- Extensive experiments are conducted on 15 off-the-shelf VOS benchmarks and our baseline model on PanoVOS, which reveals that current methods could not tackle content discontinuities in panoramic videos well.
- We propose a Panoramic Space Consistency Transformer (PSCFormer) on PanoVOS that successfully resolves the challenges of discontinuity of pixel-level content segmentation.

2 Related Work

2.1 Panoramic Datasets

In this paper, *panoramic videos* refers to complete (360°, full view) panoramic videos, which is different from the definition in [38], which only include wide but partial views of some range-view images collected from multiple cameras.
Image-based panoramic datasets. Existing popular image-level panoramic segmentation datasets are Stanford2D3D [2] and DensePASS [35]. The former one is mainly focused on indoor spaces including a total of 1,413 panoramic images with instance-level annotations in 13 categories. The latter targets driving scenes in cities. DensePASS [35] provides only 100 labeled panoramic images for testing and 2,000 unlabeled images for cross-domain transfer optimization.
Video-based panoramic datasets. Video-based benchmarks mainly include SHD360 [62], SOD360 [64] and Wild360 [7]. All of them are used for panoramic video saliency object detection. Specifically, 1) SHD360 only targets human-centric video scenes with little movement. It provides 6,268 object-level pixel-wise masks and 16,238 instance-level pixel-wise masks. 2) SOD360 focuses on the sports-centric scenario with 41 video clips (12 outdoor and 29 indoor). 3) Wild360 concentrates on natural scenes with 85 videos. Note that SOD360 and Wild360 have no object-level or instance-level annotations.
We make a comparison with the existing video panoramic datasets in Table 1. Specifically, our PanoVOS dataset contains 150 videos mainly from three different domains: person, animal, and common object, which makes the dataset more general for object-agnostic evaluations. Besides, videos in our dataset have a relatively large range of motion, making our PanoVOS dataset suitable for video tracking and segmentation evaluation tasks under panoramic scenes. Moreover, the average duration of each video in our dataset is $20s$, which is about 4 times

Datasets	Motion	#Videos	#Frames	#Total Masks	Average Duration
SHD360 [62]	Small	41	6,268	16,238	5s
SOD360 [64]	Large	104	N/A	0	N/A
Wild360 [7]	Large	85	N/A	0	N/A
PanoVOS	Large	**150**	**13,995**	**19,145**	**20**s

Table 1: Comparison of panoramic video datasets. Our PanoVOS is the first long-term panoramic video segmentation dataset with instance-level masks. Compared with existing panoramic video datasets [7, 62, 64] that are used for saliency detection, our panoramic video dataset for video segmentation, *i.e.*, PanoVOS, includes more diverse and larger motion, making it suitable for dense video tracking evaluation.

longer than SHD360 [62] (5s per video). By the way, the longer video is highlighted in a recent survey [49]. The longer the video, the more likely it is to introduce more panoramic video characteristics such as distortion and discontinuity, which is more challenging and more practical.

2.2 Video Object Segmentation Datasets

The establishment of DAVIS [41, 42]and YouTube-VOS [52] datasets pave the way for the boosting development of VOS methods. They are collected by traditional pinhole cameras and the duration of each video clip is very short, only 5s on average. In contrast, the average video length in the proposed PanoVOS dataset is 20s, which is 4 times longer than the existing video datasets. Our dataset includes more challenging scenes (*e.g.* distortion and discontinuity) that is non-negligible in real-world applications.

2.3 Video Object Segmentation Methods

Existing video object segmentation methods can be roughly classified into three subsets: online-learning-based, propagation-based, and matching-based.

Online learning-based. Online learning-based approaches [3, 36, 50], which either train or fine-tune their networks with the first-frame ground truth at test time and are therefore a great waste of resources. OnAVOS [47] achieves promising results by introducing an online adaptation mechanism, but it still requires online fine-tuning. To a certain extent, it restricts networks' efficiency.

Propagation-based. Propagation-based models [4, 8, 39] get the target masks in a frame-to-frame prorogation way. Although propagation-based methods improve efficiency, they lack long-term context and therefore are difficult to handle object disappearance and reappearance, severe obscuration, and distortion.

Matching-based. Matching-based methods [6, 10, 11, 14, 16, 25, 26, 31, 37, 40, 53, 54, 63] aim to learn an embedding space of target objects between query and memory. Recently state-of-the-art methods encode many frames into embeddings and store them as a feature memory bank. The most representative is STM [40],

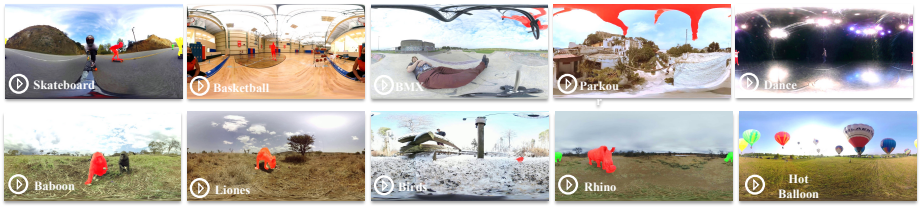

Fig. 2: PanoVOS dataset. We select 10 samples from the dataset involving major scenes. For each video, there are high-quality instance-level pixel-wise masks.

Splits	Train	Val	Test
#Videos	80	35 (10 unseen)	35 (10 unseen)
#Images	7,070 (50.5%)	3,464 (24.8%)	3,461 (24.7%)
#Masks	9,585 (50.1%)	4,957 (25.9%)	4,603 (24.0%)

Table 2: Statistics of PanoVOS dataset

which has been extended to many works [5, 19, 34, 44, 48, 51]. AOT [58] introduces an identification mechanism by encoding multiple targets into the same embedding space, which can simultaneously segment multiple objects. However, they fail to address the challenges of the tremendous proportion of distortion and discontinuity under panoramic setting.

3 PanoVOS Dataset

We introduced the proposed PanoVOS dataset in three parts, (1) collection process, (2) statistical summary, and (3) annotation pipeline.

3.1 Data collection

We built our PanoVOS dataset with the principle of diversity in mind. Moreover, the objects in the video should have a large amplitude of motion or camera movement. Based on the above viewpoint, we collected videos from the YouTube website for further annotation, respectively. The range of the video length is from 3 to 40 seconds. The average sequence length of each video in the dataset is approximately 20 seconds. We followed the settings of YouTube-VOS [52] to sample the frames at 6 fps.

3.2 Dataset Statistics

PanoVOS contains 150 videos, including 13,995 frames and 19,145 instance annotations from 35 categories. The average length of each video is 20 seconds. We believe that visual categories are representative of common life scenarios, and Fig.2 shows some samples of PanoVOS. To create our PanoVOS, in the

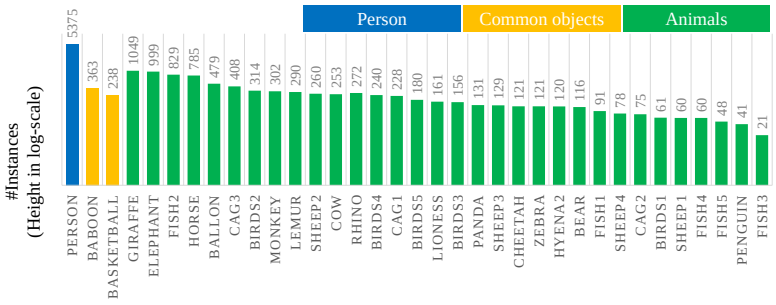

Fig. 3: Instance-level distribution of PanoVOS dataset. Our dataset contains three major divisions: *person*, *animals*, and *common objects* with 35 sub-divisions.

spirit of the video object segmentation task, we carefully selected videos with relatively large motion amplitudes and chose a set of video categories including person (*e.g.* parkour, dance, BMX, skateboard), animals (*e.g.* elephant, monkey, giraffe, rhino, birds) and common objects (*e.g.* basketball, hot balloon) as shown in Fig 3. PanoVOS dataset consists of 150 videos split into training (80), validation (35), and test (35) sets. Table 2 shows detailed division results. Both the validation and test sets have 35 videos (about 23% of the frames and the masks). For validation and test sets, we keep some unseen visual categories for generalization ability evaluation.

3.3 Annotation Pipeline

Annotation is very time-consuming and expensive for a pixel-level panoramic segmentation dataset. To obtain accurate large-scale video panoramic segmentation annotations and make the process more efficient, we propose a semi-automatic human-computer joint annotation strategy, as shown in Fig 4. First, keyframes are selected and manually annotated for each video, which are images with a speed of 1 fps. This is followed by a frame-by-frame propagation from the annotated keyframes to those unlabeled intermediate frames with a sophisticated semi-supervised VOS model. Then, to tackle the distortions and discontinuities in panoramic videos, we need to re-calibrate the resulting annotations via human refinement. More details will unfold below.

Annotation Propagation For the annotation of each video, we first need an expert to browse the current video and note down all objects that have a large amplitude of movement. Then, for each video, the recorded objects in keyframes with a speed of 1 fps are selected for manual annotation. To avoid consistency errors or the problem of objects being labeled as other instances when they disappear and reappear, another expert needs to double-check the annotations of all objects to improve the accuracy of the dataset annotation.

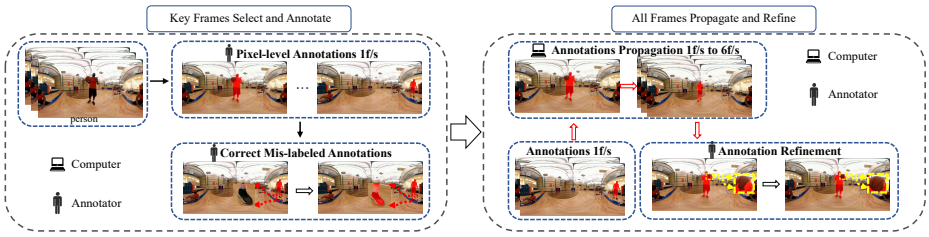

Fig. 4: PanoVOS annotation pipeline. Our annotation pipeline includes two phases. (1) The first phase is called *Key Frames Select and Annotate*. The annotator browses the video and picks out the object to be annotated. Then, instances are manually annotated at 1 fps and corrected by another annotator. (2) The second phase is called *All Frames Propagate and Refine*. In this phase, we apply a semi-supervised video object segmentation model to help propagate the annotated masks and the generated instances are refined by annotators.

We then use the off-the-shelf video labeling method [58] to propagate the instance masks frame by frame from the annotated keyframes to untagged intermediate frames and generate masks at 6 fps.

Annotation Refinement To present a new Panoramic dataset of high quality. After obtaining masks of the first propagation stage, annotators are asked to check the quality of the masks and refine them. The main amendments are in the following two areas. 1). Since our video resolution is generally relatively high, the propagation method will often fail when encountering complex videos with many small objects in a scene. 2) Due to the huge distortions and discontinuities present in the panoramic video, the quality of the masks obtained is relatively poor. Manual correction of the mask is checked by another annotator until the result is satisfactory before proceeding to the next video annotation.

4 Method

4.1 Overview

Video object segmentation targets assigning an instance label to every pixel in the given video sequence based on the first frame mask. Recent works [5,6,13,59] have demonstrated that the attention mechanism can significantly help improve the segmentation performance. However, for the challenge of content discontinuation in panoramic videos, only considering the original attention mechanism will not be able to fully utilize the semantic information on the left and right boundaries (pixel contiguity) in the spatial dimension and will lose valuable contextual information when segmenting objects. Therefore, in this work, our mission is to design an effective network architecture, which can help acquire valuable boundary relationships.

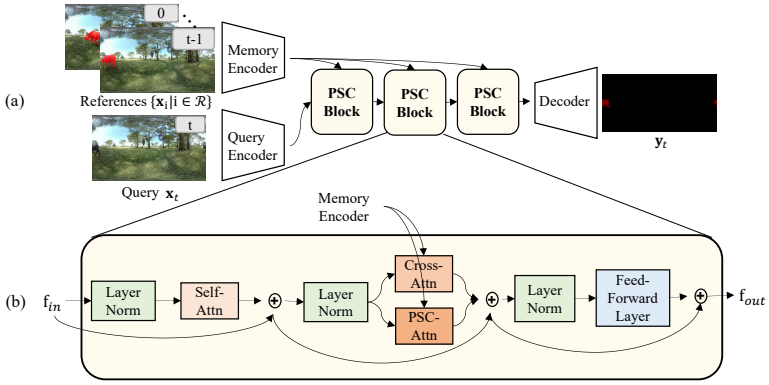

Fig. 5: (a) **PSCFormer overview.** Given the query frame $\mathbf{x}_t$ and reference frames $\{\mathbf{x}_i|i \in \mathcal{R}\}$, the goal of VOS is to delineate objects from the background by generating mask $\mathbf{y}_t$ for query frame $\mathbf{x}_t$. References and the query frame are encoded by the memory encoder and query encoder, respectively. Multiple stacking panoramic space consistency (PSC) blocks are used to leverage the correspondence in the panoramic space between references and the query frame. A decoder is used for generating the prediction of the query frame. (b) **Panoramic space consistency block architecture details.**

Fig. 5(a) illustrates the overall architecture of the proposed network. Given the query frame $\mathbf{x}_t$ and references $\{\mathbf{x}_i|i \in \mathcal{R}\}$, the goal of VOS is to delineate objects from the background by generating mask $\mathbf{y}_t$ for query frame $\mathbf{x}_t$. Following [57], our basic setting uses the first and previous frame as references $\mathcal{R} = \{1, t-1\}$. The memory encoder and query encoder are responsible for extracting frame-level features. After this, the panoramic space consistency block takes them as input and aggregates the spatial-temporal information between the reference frames and the query frame at the pixel level. Finally, the decoder uses the output of the sequence stacking PSC blocks to predict the mask of the object.

4.2 Panoramic Space Consistency Block

Fig. 5(b) shows the structure of a PSC block. Motivated by the common transformer blocks [46], PSC firstly contains a self-attention layer, which is used to aggregate the target objects' correlation information within the query frame. Then, the middle module is composed of cross-attention and PSC-attention, in which cross-attention is responsible for learning the target objects' information from references $\mathcal{R}$ and the PSC-attention targets on exploring the boundary relationship between the query frame and previous frame. Finally, PSC employs a two-layer feed-forward MLP with GELU [17] non-linearity activation function.

Panoramic Space Consistency Attention (PSC-Attn). PSC-Attn is employed to model the spatial-temporal relationship between the query frame and reference frames considering the continuity of pixels of images in the panoramic space. How to establish a connection between the left and right boundaries be-

come especially important? The most intuitive solution would be to directly splice in length, but this would lead to a huge amount of computation. Therefore, we take the approach of moving a portion of the region in the length dimension from the right boundary to the leftmost boundary for stitching. Consequently, we only focus on the left and right boundaries between the query frame and the reference frame. Thus, unlike the original attention, where each query token is counted for attention along with all key tokens in the reference frame, our PSC attention takes care of the key tokens in a fixed window size. In particular, we define the reference frame feature embedding $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{H \times W \times C}$, which is extracted from the query encoder. H, W, and C represent the height, width, and channel dimensions, respectively. According to the solutions mentioned above, the new feature embedding $\mathbf{f}(\mathbf{x})'$ is calculated as follows:

$$\begin{aligned} \mathbf{f}(\mathbf{x})'[0:W/p] &= \mathbf{f}(\mathbf{x})[W/p:W] \\ \mathbf{f}(\mathbf{x})'[W/p:W] &= \mathbf{f}(\mathbf{x})[0:W/p] \\ \mathbf{f}(\mathbf{x})'[W/p:W-W/p] &= \mathbf{f}(\mathbf{x})[W/p:W-W/p], \end{aligned} \tag{1}$$

where $p \in \mathbb{Z}^+$. We define query embedding $Q \in \mathbb{R}^{HW \times C}$, key embedding $K \in \mathbb{R}^{HW \times C}$, value embedding $V \in \mathbb{R}^{HW \times C}$, where Q is from the query frame feature embedding, K and V are from $\mathbf{f}(\mathbf{x})'$ by performing dimensional transformations. Mathematically, we define the PSC attention as follows,

$$\text{PSCAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^T\mathbf{R}}{\sqrt{C}}\right)V, \tag{2}$$

where $\mathbf{R} \in [0,1]^{HW \times HW}$ means a window that represents the attention range of each query token. For query $Q_{(x,y)}$ at (x, y) position, we define the $\mathbf{R}_{(x,y)}$ as:

$$\mathbf{R}_{x,y}(i,j) = \begin{cases} 1 & \text{if } (x-i)^2 \leqslant s^2 \text{ and } (y-j)^2 \leqslant s^2 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where (i, j) is the position for each key token, s is the window size. For each query token, it calculates the attention with another key token only if they are spatially limited to a $(2 \times s + 1)$ size window, which significantly reduces the time complexity from $(h \times w)^2$ to $(2 \times s + 1)^2$.
Following [46], we implement the representational form of our PSCAttn module with multi-headed attention, defined mathematically as follows,

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \\ \text{head}_i &= \text{PSCAttn}\left(QW_i^Q, KW_i^K, VW_i^V\right), \end{aligned} \tag{4}$$

where $W_i^Q \in \mathbb{R}^{C \times d_{model}}$, $W_i^K \in \mathbb{R}^{C \times d_{model}}$, $W_i^V \in \mathbb{R}^{C \times d_{model}}$ and $W_i^O \in \mathbb{R}^{C \times C}$ are the linear projections. As [46], we set the number of heads to $(h = C/d_{model})$ 8, where d_{model} is the projection dimension of each head.

Methods		YouTube-VOS	PanoVOS Validation					PanoVOS Test				
	MF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
AOTT [58]		73.7	53.8$\downarrow_{19.9}$	44.4	58.3	46.9	65.7	43.7$\downarrow_{30.0}$	36.3	49.8	39.6	49.2
AOTS [58]		74.6	55.8 $\downarrow_{18.8}$	49.1	62.2	46.4	65.5	44.7$\downarrow_{29.9}$	32.4	43.3	46.5	56.8
AOTB [58]		75.2	53.7$\downarrow_{21.5}$	46.2	58.1	46.3	64.1	39.5$\downarrow_{35.7}$	34.4	44.4	35.0	44.4
AFB-URR [32]	✓	65.2	40.1$\downarrow_{25.1}$	31.1	41.5	35.8	51.8	30.4$\downarrow_{34.8}$	23.1	32.7	28.8	36.9
STCN [6]	✓	76.1	49.9$\downarrow_{26.2}$	42.7	53.4	45.1	58.4	48.0$\downarrow_{28.1}$	39.3	50.2	46.7	55.7
XMem [5]	✓	77.0	48.6$\downarrow_{28.4}$	40.7	50.1	44.8	58.6	40.2$\downarrow_{36.8}$	35.3	44.9	36.4	44.0
AOTL [58]	✓	74.7	49.2$\downarrow_{25.5}$	43.3	57.3	38.9	57.1	38.2$\downarrow_{36.5}$	32.3	43.7	32.7	44.1
R50_AOTL [58]	✓	76.5	50.1$\downarrow_{26.4}$	44.5	58.6	40.3	57.2	41.4$\downarrow_{35.1}$	33.7	45.0	38.3	48.4
SwinB_AOTL [58]	✓	74.4	44.8$\downarrow_{29.6}$	39.1	52.2	34.9	53.0	36.2$\downarrow_{38.2}$	31.1	42.0	31.0	40.6
RDE* [24]	✓	61.7	43.1$\downarrow_{18.6}$	36.0	48.4	35.2	52.7	41.3$\downarrow_{20.4}$	30.9	44.6	41.4	48.5
STCN* [6]	✓	56.3	43.2$\downarrow_{13.1}$	41.6	53.7	33.2	44.5	38.0$\downarrow_{18.3}$	32.8	43.2	35.5	40.4
XMem* [5]	✓	65.8	55.9$\downarrow_{9.9}$	52.2	64.0	47.2	60.0	49.6$\downarrow_{16.2}$	39.2	52.6	46.8	59.9

Table 3: Domain transfer result of (static image datasets)→(PanoVOS Validation & Test). Subscript s and u denote scores in seen and unseen categories. MF denotes multiple historical frames as reference. $\downarrow$ represents the performance of the declining values compared to the YouTube-VOS dataset [52]. $*$ denotes a large-scale external dataset BL30K [6] dataset is used during training.

Methods		YouTube-VOS	PanoVOS Validation					PanoVOS Test				
	MF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
CFBI† [59]		81.4	60.9$\downarrow_{20.5}$	53.0	65.2	56.3	69.0	49.0$\downarrow_{32.4}$	49.4	47.6	46.2	52.6
CFBI+† [59]		82.8	57.6$\downarrow_{25.2}$	52.1	67.0	48.1	63.4	53.7$\downarrow_{29.1}$	51.6	59.3	46.6	57.5
AOTT [58]		80.2	61.5$\downarrow_{18.7}$	55.6	67.7	54.6	68.2	52.6$\downarrow_{27.6}$	44.8	55.3	51.5	58.8
AOTS [58]		82.6	66.7$\downarrow_{15.9}$	58.0	70.5	62.0	76.4	57.3$\downarrow_{25.3}$	50.2	61.0	54.6	63.5
AOTB [58]		83.5	70.5$\downarrow_{13.0}$	59.2	71.7	68.5	82.7	60.8$\downarrow_{22.7}$	53.0	64.4	57.8	68.2
AFB-URR [32]	✓	79.6	55.1$\downarrow_{24.5}$	44.7	55.6	53.4	66.7	52.4$\downarrow_{27.2}$	43.6	54.2	52.0	59.9
RDE [24]	✓	81.9	54.7$\downarrow_{27.2}$	50.3	63.9	44.6	60.1	55.4$\downarrow_{26.5}$	45.5	59.2	51.0	65.9
STCN [6]	✓	83.0	61.8$\downarrow_{21.2}$	50.3	63.5	61.3	72.1	53.4$\downarrow_{29.6}$	46.2	58.9	49.0	59.9
XMem [5]	✓	85.7	66.1$\downarrow_{19.6}$	56.6	68.7	62.0	77.2	62.5$\downarrow_{23.2}$	53.1	65.4	61.1	70.4
AOTL [58]	✓	83.8	71.9$\downarrow_{11.9}$	62.1	75.3	67.4	82.8	62.1$\downarrow_{21.7}$	57.1	69.0	56.2	66.1
R50_AOTL [58]	✓	84.1	69.2$\downarrow_{14.9}$	56.7	69.4	67.5	83.1	61.4$\downarrow_{22.7}$	57.5	69.0	53.3	65.7
SwinB_AOTL [58]	✓	84.5	67.5$\downarrow_{17.0}$	60.2	73.6	60.3	76.0	60.9$\downarrow_{23.6}$	53.9	63.7	58.7	67.4
RDE* [24]	✓	83.3	60.9$\downarrow_{22.4}$	51.4	64.7	56.0	71.6	55.6$\downarrow_{27.7}$	48.1	60.8	52.6	61.0
STCN* [6]	✓	84.3	61.7$\downarrow_{22.6}$	49.9	61.8	59.7	75.5	55.8$\downarrow_{28.5}$	48.2	59.8	52.7	62.5
XMem* [5]	✓	86.1	63.4$\downarrow_{22.7}$	53.5	64.4	61.5	74.1	61.0$\downarrow_{25.1}$	53.5	65.1	57.5	68.0

Table 4: Domain transfer result of (static image datasets & YouTubeVOS)→(PanoVOS Validation & Test). Subscript s and u denote scores in seen and unseen categories. MF denotes multiple historical frames as reference. $\downarrow$ represents the performance of the declining values compared to the YouTube-VOS dataset [52]. $*$ denotes a large-scale external dataset BL30K [6] dataset is used during training. † denotes no synthetic data is used during the training stage.

5 Experiment

In this section, we design a series of experiments to answer the following research questions related to how to tackle video object segmentation in panoramic scenes:
RQ1: How well are current VOS methods trained on non-panoramic videos adapted to the panoramic world?
RQ2: How well do variations of the foundation model Segment Anything Model [23] adapt to the panoramic world?
RQ3: Can the proposed PanoVOS datasets bring about a consistent performance gain to VOS methods?
RQ4: How well does Panoramic Space Consistency Attention contribute?
RQ5: What are the remained problems for panoramic-related research?

Methods	PanoVOS Validation					PanoVOS Test				
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
PerSAM [61]	19.1	12.3	19.8	17.4	27.1	19.5	7.4	14.9	23.8	31.7
SAM-PT [43]	47.5	36.7	48.6	46.0	58.7	41.0	31.1	40.5	40.2	52.3
SAM-PT-reinit [43]	43.7	34.3	44.3	41.3	54.9	43.6	35.0	42.7	43.5	53.0

Table 5: Quantitative comparison on PanoVOS for variations of foundation model Segment Anything Model [23]. Subscript s and u denote scores in seen and unseen categories.

Methods		PanoVOS Validation					PanoVOS Test				
	MF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
CFBI† [57]		35.8	34.6	44.8	24.2	39.7	19.1	18.2	26.1	12.2	19.8
CFBI+† [59]		41.3	38.0	47.9	32.5	46.9	30.9	30.8	42.7	21.4	28.5
AOTT [58]		65.6	59.4	68.3	59.7	75.0	53.4	49.3	61.6	47.5	55.1
AOTS [58]		67.7	61.2	70.0	62.4	77.1	55.9	**53.2**	**65.1**	48.6	57.0
AOTB [58]		67.6	62.3	72.0	61.5	74.8	55.4	53.5	64.2	47.7	56.0
Ours-Base		**74.0**	**66.4**	**80.4**	**66.2**	**83.0**	**56.8**	49.4	62.7	**52.4**	**62.5**
AFB-URR [32]	✓	34.3	34.8	42.8	24.9	34.5	34.2	28.2	38.8	32.9	36.8
RDE [24]	✓	50.5	49.7	58.4	39.2	54.9	42.5	36.9	46.6	38.5	48.2
STCN [6]	✓	52.0	51.2	60.8	41.5	54.5	50.8	43.6	56.5	49.3	53.7
XMem [5]	✓	55.7	54.8	63.3	45.2	59.7	53.5	49.5	62.6	47.1	54.8
AOTL [58]	✓	66.6	61.4	71.1	59.4	74.3	53.8	50.0	60.3	47.8	57.1
R50_AOTL [58]	✓	65.3	61.9	71.4	56.4	71.6	54.6	52.9	63.2	47.5	54.9
SwinB_AOTL [58]	✓	62.1	58.9	66.5	54.3	68.8	53.1	49.0	57.8	49.0	56.6
Ours-Large	✓	**77.9**	**70.5**	**85.2**	**69.5**	**86.4**	**59.9**	**54.9**	**69.2**	**53.0**	**62.4**
RDE* [24]	✓	54.3	52.8	61.6	44.6	58.2	52.2	44.5	56.0	49.3	59.1
STCN* [6]	✓	51.7	51.2	60.6	41.3	53.6	53.8	53.7	58.1	46.0	57.3
XMem* [5]	✓	57.7	55.6	64.6	48.6	61.9	57.9	51.3	64.5	53.2	62.7

Table 6: Quantitative comparison on PanoVOS for models with pretraining on static image datasets. Subscript s and u denote scores in seen and unseen categories. MF denotes multiple historical frames as reference. $*$ denotes a large-scale external dataset BL30K [6] dataset is used during training.

5.1 Implementation Details

Model Architecture. We build two variants of our method with different reference bank sizes $\mathcal{R}$ for a fair comparison with previous methods. **Ours-Base** uses only the first frame and the previous frame as reference ($\mathcal{R} = \{1, t-1\}$), which are for the sake of high inference speed and low memory consumption. **Ours-Large** uses multiple historical frames as reference ($\mathcal{R} = \{1+2\delta, 1+2\delta, 1+3\delta...\}$), which follows [34, 58]. In our work, we set δ to 2 and 5 for training and testing respectively. For p and s in PSC block, we set them as 2 and 7.

Evaluation Metrics. Following the standard protocol [41, 42], we adopt the region accuracy $\mathcal{J}$ and boundary accuracy $\mathcal{F}$. $\mathcal{J}$ means the Jaccard Index/Intersection over Union (IoU), which is the ratio of intersection and the joint area between predicted masks and ground truths. And $\mathcal{F}$ evaluates the accuracy of the segmentation boundary, which is computed by transforming it into a bipartite graph matching problem with predicted masks and ground truths.

5.2 Domain Transfer Results (RQ1)

We evaluate previous SOTA methods, which are trained on conventional datasets that are captured by pinhole cameras, on PanoVOS datasets to evaluate the domain transfer performance. To quantify the transfer performance of advanced

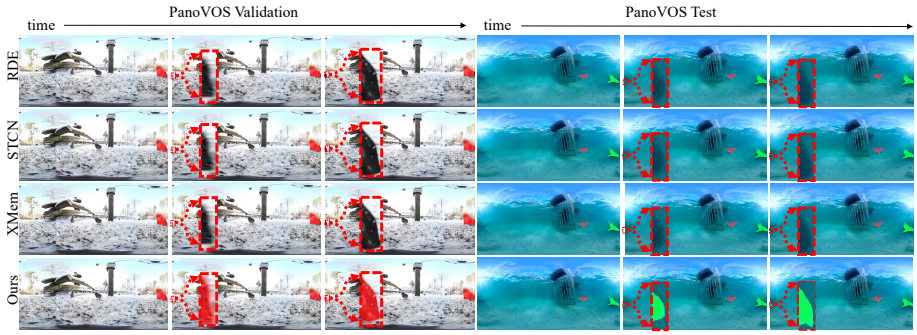

Fig. 6: Qualitative comparison to the state-of-the-art methods, RDE [24], STCN [6], and XMem [5], on PanoVOS dataset. Our model performs better under the challenge of content discontinuities. Error regions are bounded.

Methods		Validation	Test
	PSCAttn	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Ours-Base	✓	72.8 **74.0**	55.4 **56.8**
Ours-Large	✓	74.8 **77.9**	59.5 **59.9**

Table 7: Ablation study of *PSCAttn* module on PanoVOS.

models trained on planar video datasets, we evaluated 15 off-the-shelf VOS models, including [5, 6, 24, 32, 57–59], and we follow official implementations and training strategies details of them. Table 3 summarizes the domain transfer results of methods that are only trained on synthetic datasets, such as COCO [33] and ECSSD [45], on PanoVOS dataset. Table 4 shows the domain transfer results of state-of-the-art methods, that are trained on synthetic datasets (*e.g.* COCO [33]) and video datasets (*e.g.* YouTube-VOS [52]), on our PanoVOS validation and test sets. By analyzing the performance of advanced VOS methods that target conventional planar videos on panoramic videos, we provide the following insights. Firstly, the performance of current sophisticated VOS models will largely degrade when employed to tackle panoramic videos. Secondly, we can observe a trend that training on larger VOS datasets, *i.e.*, YouTube-VOS [52] and BL30K [6] can help mitigate the gap between planar and panoramic videos.

5.3 Results via Visual Foundation Model (RQ2)

To quantity the segmentation performance of different variations of the foundation model Segment Anything Model [23] on PanoVOS, we evaluate the latest top performing models PerSAM [61] and SAM-PT [43], as shown in Table 5. The performance of these models on our challenging PanoVOS dataset is still unsatisfactory, which leaves space for further exploration.

Methods	Attention Type	Validation $\mathcal{J}\&\mathcal{F}$	Test $\mathcal{J}\&\mathcal{F}$
Ours-Base	*CrossAttn*	72.5	54.8
	PSCAttn	**74.0**	**56.8**
Ours-Large	*CrossAttn*	76.8	59.1
	PSCAttn	**77.9**	**59.9**

Table 8: Comparison between our PSC attention (*PSCAttn*) and cross attention (*CrossAttn*) module on PanoVOS dataset.

Methods	PanoVOS Validation				
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
w/o	73.7	68.8	82.6	63.1	80.3
p=3	76.3	70.4	85.0	65.8	84.1
p=5	74.2	65.3	79.7	67.5	84.2
p=10	75.9	68.1	81.9	68.6	85.2
p=15	75.0	66.7	81.0	67.0	85.4
p=2 (Ours)	**77.9**	**70.5**	**85.2**	**69.5**	**86.4**

Table 9: Hyperparameter Analysis of p, which enables the stitching mechanism, in PSCAttn for Ours-Large model.

5.4 Main Results on PanoVOS (RQ3)

To evaluate the performance of previous methods on the proposed panoramic VOS dataset, we re-trained them on the training set of PanoVOS for the sake of fairness. We report the performance in Table 6,
which demonstrates that all the previous VOS models perform worse on PanoVOS than on the traditional VOS benchmarks, *e.g.*, YouTube-VOS. Our model substantially outperforms all these methods and achieves state-of-the-art on all evaluation metrics on PanoVOS, which verifies the effectiveness of our model in tackling panoramic videos. Fig.6 visualizes some qualitative comparisons between our model and previous state-of-the-art methods on PanoVOS dataset, which shows that previous benchmarks fail to cope with content discontinuities while our model tackles them well.

5.5 Ablation Study (RQ4)

In this section, we conduct ablation studies to demonstrate the effectiveness of the main component, *i.e.*, Panoramic Space Consistency Attention (PSCAttn), of our model, with all the experiments performed based on our two model variants, *i.e.*, Ours-Base and Ours-Large. For training, static image datasets are used for pre-training and PanoVOS is used for main training. Table 7 demonstrates the effectiveness of our PSCAttn module. Besides, Fig. 7 illustrates the qualitative comparison between our default model (Ours-Base) and the setting without PSCAttn module. Our model performs better when coping with the pixel discontinuity problem. Moreover, as is shown in Table 8, compared to the conventional cross-attention (CrossAttn) module, PSCAttn also achieves better performance. In Table 9, we analyze the hyperparameter p, which influences the stitching mechanism in PSCAttn, of our model (Ours-Large) on the PanoVOS validation set. Specifically, the highest overall performance ($\mathcal{J}\&\mathcal{F}$) is achieved

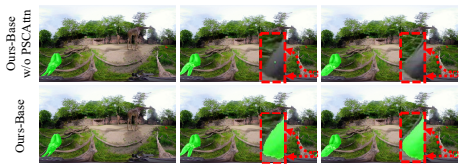

Fig. 7: Qualitative ablation study of *PSCAttn* module.

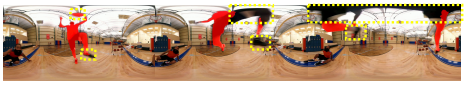

Fig. 8: Challenge. Our model fails to segment some objects with strong distortion.

when setting p as 2. Compared to the setting without using the stitching mechanism (w/o), our model can achieve much better performance. Specifically, our final model (Ours-Large, $p = 2$) achieves more than 4% gain in $\mathcal{J}\&\mathcal{F}$.

5.6 Limitation and Future Work (RQ5)

To prompt greater progress of panoramic VOS, we also analyze the limitations of our method. Specifically, our method has no notion of severe distortion challenge since we do not employ a special design (such as deformable convolution [9]) to tackle deformations. That means our model may not segment the objects with large distortions. One such failure case is shown in Fig. 8. Besides, our panoramic dataset can be applied to broader video segmentation and tracking domains, such as referring video object segmentation [28–30, 56], video object tracking [18], video instance segmentation [15], few-shot segmentation [20], and more broader embodied navigation tasks [55]. Also, it would be valuable to investigate the zero-shot segmentation performance of visual foundation models [23] on our challenging panoramic dataset. We hope our work can shed light on efficient adaptation from non-panoramic to panoramic perception.

6 Conclusion

In this paper, we introduce a high-quality dataset, *i.e.*, PanoVOS, for panoramic video object segmentation. Our PanoVOS dataset provides pixel-level instance annotations with diverse scenarios and significant motions. Based on this dataset, we evaluate 15 off-the-shelf VOS models and carefully analyze their limitations. Then, we further present our model, *i.e.*, PSCFormer, which is equipped with the proposed panoramic space consistency transformer block. Our preliminary experiment demonstrates the effectiveness of our proposed model to enhance the segmentation performance and consistency in panoramic scenes. In conclusion, this provides a new challenge for video understanding, and we hope our PanoVOS dataset can attract more researchers to pay attention to panoramic videos.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No.62072112), and Scientific and Technological Innovation Action Plan of Shanghai Science and Technology Committee (No.22511101502, No.22511102202 and No.21DZ2203300).

References

1. Ai, H., Cao, Z., Zhu, J., Bai, H., Chen, Y., Wang, L.: Deep learning for omnidirectional vision: A survey and new perspectives. arXiv preprint arXiv:2205.10468 (2022)
2. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
3. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 221–230 (2017)
4. Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D.: State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9384–9393 (2020)
5. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision. pp. 640–658. Springer (2022)
6. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in Neural Information Processing Systems **34**, 11781–11794 (2021)
7. Cheng, H.T., Chao, C.H., Dong, J.D., Wen, H.K., Liu, T.L., Sun, M.: Cube padding for weakly-supervised saliency prediction in 360 videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1420–1429 (2018)
8. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7415–7424 (2018)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
10. Dang, J., Zheng, H., Xu, X., Guo, Y.: Unified spatio-temporal dynamic routing for efficient video object segmentation. IEEE Transactions on Intelligent Transportation Systems (2023)
11. Dang, J., Zheng, H., Xu, X., Wang, L., Hu, Q., Guo, Y.: Adaptive sparse memory networks for efficient and robust video object segmentation. IEEE Transactions on Neural Networks and Learning Systems (2024)
12. Eger Passos, D., Jung, B.: Measuring the accuracy of inside-out tracking in xr devices using a high-precision robotic arm. In: International Conference on Human-Computer Interaction. pp. 19–26. Springer (2020)
13. Fang, R., Yan, S., Huang, Z., Zhou, J., Tian, H., Dai, J., Li, H.: Instructseq: Unifying vision tasks with instruction-conditioned multi-modal sequence generation. arXiv preprint arXiv:2311.18835 (2023)
14. Guo, P., Hong, L., Zhou, X., Gao, S., Li, W., Li, J., Chen, Z., Li, X., Zhang, W., Zhang, W.: Clickvos: Click video object segmentation. arXiv preprint arXiv:2403.06130 (2024)

15. Guo, P., Huang, T., He, P., Liu, X., Xiao, T., Chen, Z., Zhang, W.: Openvis: Open-vocabulary video instance segmentation. arXiv preprint arXiv:2305.16835 (2023)
16. Guo, P., Zhang, W., Li, X., Zhang, W.: Adaptive online mutual learning bi-decoders for video object segmentation. IEEE Transactions on Image Processing **31**, 7063–7077 (2022)
17. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
18. Hong, L., Yan, S., Zhang, R., Li, W., Zhou, X., Guo, P., Jiang, K., Chen, Y., Li, J., Chen, Z., et al.: Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19079–19091 (2024)
19. Hu, L., Zhang, P., Zhang, B., Pan, P., Xu, Y., Jin, R.: Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4144–4154 (2021)
20. Iqbal, E., Safarov, S., Bang, S.: Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. arXiv preprint arXiv:2206.09667 (2022)
21. Jiang, H., Jiang, G., Yu, M., Zhang, Y., Yang, Y., Peng, Z., Chen, F., Zhang, Q.: Cubemap-based perception-driven blind quality assessment for 360-degree images. IEEE Transactions on Image Processing **30**, 2364–2377 (2021)
22. Jost, T.A., Nelson, B., Rylander, J.: Quantitative analysis of the oculus rift s in controlled movement. Disability and Rehabilitation: Assistive Technology **16**(6), 632–636 (2021)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
24. Li, M., Hu, L., Xiong, Z., Zhang, B., Pan, P., Liu, D.: Recurrent dynamic embedding for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1332–1341 (2022)
25. Li, W., Fan, J., Guo, P., Hong, L., Zhang, W.: Hfvos: History-future integrated dynamic memory for video object segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2024)
26. Li, W., Guo, P., Zhou, X., Hong, L., He, Y., Zheng, X., Zhang, W., Zhang, W.: Onevos: Unifying video object segmentation with all-in-one transformer framework. arXiv preprint arXiv:2403.08682 (2024)
27. Li, X., Cao, H., Zhao, S., Li, J., Zhang, L., Raj, B.: Panoramic video salient object detection with ambisonic audio guidance. arXiv preprint arXiv:2211.14419 (2022)
28. Li, X., Wang, J., Xu, X., Li, X., Raj, B., Lu, Y.: Robust referring video object segmentation with cyclic structural consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22236–22245 (2023)
29. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3402–3413 (2024)
30. Li, X., Wang, J., Xu, X., Yang, M., Yang, F., Zhao, Y., Singh, R., Raj, B.: Towards noise-tolerant speech-referring video object segmentation: Bridging speech and text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2283–2296 (2023)
31. Liang, S., Shen, X., Huang, J., Hua, X.S.: Video object segmentation with dynamic memory networks and adaptive object alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8065–8074 (2021)

32. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
34. Liu, Y., Yu, R., Wang, J., Zhao, X., Wang, Y., Tang, Y., Yang, Y.: Global spectral filter memory network for video object segmentation. In: European Conference on Computer Vision. pp. 648–665. Springer (2022)
35. Ma, C., Zhang, J., Yang, K., Roitberg, A., Stiefelhagen, R.: Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 2766–2772. IEEE (2021)
36. Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: Video object segmentation without temporal information. IEEE transactions on pattern analysis and machine intelligence **41**(6), 1515–1530 (2018)
37. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9670–9679 (2021)
38. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX. pp. 53–72. Springer (2022)
39. Oh, S.W., Lee, J.Y., Sunkavalli, K., Kim, S.J.: Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7376–7385 (2018)
40. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9226–9235 (2019)
41. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
42. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
43. Rajič, F., Ke, L., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Segment anything meets point tracking. arXiv preprint arXiv:2307.01197 (2023)
44. Seong, H., Hyun, J., Kim, E.: Kernelized memory network for video object segmentation. In: European Conference on Computer Vision. pp. 629–645. Springer (2020)
45. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended cssd. IEEE transactions on pattern analysis and machine intelligence **38**(4), 717–729 (2015)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
47. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. arXiv preprint arXiv:1706.09364 (2017)

48. Wang, H., Jiang, X., Ren, H., Hu, Y., Bai, S.: Swiftnet: Real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1296–1305 (2021)
49. Wang, W., Zhou, T., Porikli, F., Crandall, D., Van Gool, L.: A survey on deep learning technique for video segmentation. arXiv preprint arXiv:2107.01153 (2021)
50. Xiao, H., Feng, J., Lin, G., Liu, Y., Zhang, M.: Monet: Deep motion exploitation for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1140–1148 (2018)
51. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1286–1295 (2021)
52. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
53. Xu, X., Wang, J., Li, X., Lu, Y.: Reliable propagation-correction modulation for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2946–2954 (2022)
54. Xu, X., Wang, J., Ming, X., Lu, Y.: Towards robust video object segmentation with adaptive object calibration. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2709–2718 (2022)
55. Xu, X., Zhang, T., Wang, S., Li, X., Chen, Y., Li, Y., Raj, B., Johnson-Roberson, M., Huang, X.: Customizable perturbation synthesis for robust slam benchmarking. arXiv preprint arXiv:2402.08125 (2024)
56. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., Dong, H., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6449–6457 (2024)
57. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: European Conference on Computer Vision. pp. 332–348. Springer (2020)
58. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
59. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by multi-scale foreground-background integration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
60. Yuan, M., Richardt, C.: 360 optical flow using tangent images. In: British Machine Vision Conference (BMVC) (2021)
61. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
62. Zhang, Y., Zhang, L., Wang, K., Hamidouche, W., Deforges, O.: Shd360: A benchmark dataset for salient human detection in 360 videos. arXiv preprint arXiv:2105.11578 (2021)
63. Zhang, Y., Wu, Z., Peng, H., Lin, S.: A transductive approach for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6949–6958 (2020)
64. Zhang, Z., Xu, Y., Yu, J., Gao, S.: Saliency detection in 360 videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 488–503 (2018)