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

Shilin Yan[1] Xiaohao Xu[3] Lingyi Hong[1] Wenchao Chen[1]
Wenqiang Zhang[1,2] Wei Zhang[1†]
[1]Shanghai Key Laboratory of Intelligent Information Processing,
School of Computer Science, Fudan University
[2]Academy for Engineering and Technology, Fudan University
[3] University of Michigan, Ann Arbor
tattoo.ysl@gmail.com weizh@fudan.edu.cn

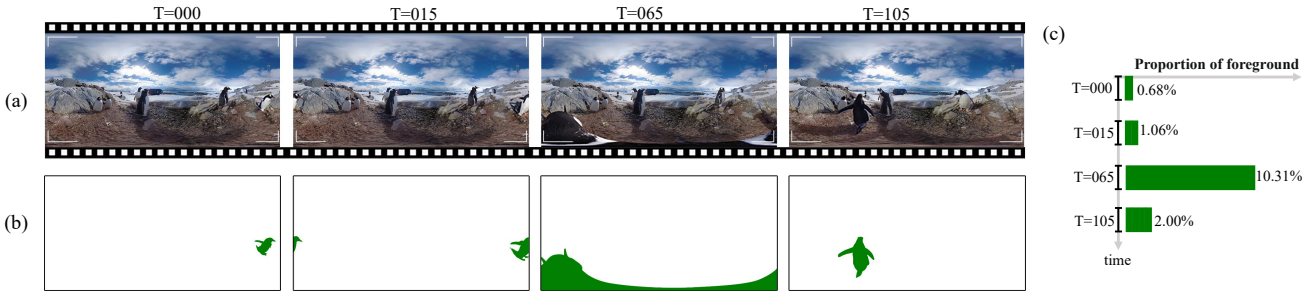

Figure 1. **Panoramic video object segmentation (PanoVOS).** PanoVOS targets tracking and distinguishing the particular instances under content discontinuities (*e.g.* penguin in the image of $T = 15$) and serve distortion (*e.g.* penguin in the image of $T = 65$). We show the sample of (a) frames, (b) 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 our PanoVOS can advance the development of panoramic segmentation/tracking. The proposed PanoVOS dataset has been released at https://github.com/shilinyan99/PanoVOS.

1. Introduction

Semi-supervised video object segmentation (VOS) [54], 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 [18, 12], panoramic videos with a 360° × 180° FoV are able to collect the entire viewing sphere and richer spatial information of the whole scene [1, 17, 62, 23]. To the best of our knowledge, we are the first to attempt to tackle the promising but challenging task of panoramic video object segmentation.

†Corresponding Author.

Datasets	Motion	#Videos	#Frames	Total #Masks	Avg. #Durations
SHD360 [65]	small	41	6,268	16,238	5s
SOD360 [66]	large	104	N/A	0	N/A
Wild360 [8]	large	85	N/A	0	N/A
PanoVOS	large	150	13,995	19,145	20s

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 [65, 66, 8] used for saliency detection, PanoVOS includes more diverse and larger motion, which makes it suitable to be used in the video tracking domain.

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 [60] 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 video object segmentation 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 [35], which only include wide but partial views of some range-view images collected from multi-cameras in. Existing panoramic datasets can be roughly classified into two categories: image-based and video-based.

Image-based panoramic datasets. Existing popular image-level panoramic segmentation datasets are Stanford2D3D [2] and DensePASS [32]. 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 [32] provides only 100 labeled panoramic images for testing and 2,000 unlabeled panoramic images for cross-domain transfer optimization.

Video-based panoramic datasets. Video-based benchmarks mainly include SHD360 [65], SOD360 [66] and Wild360 [8]. 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

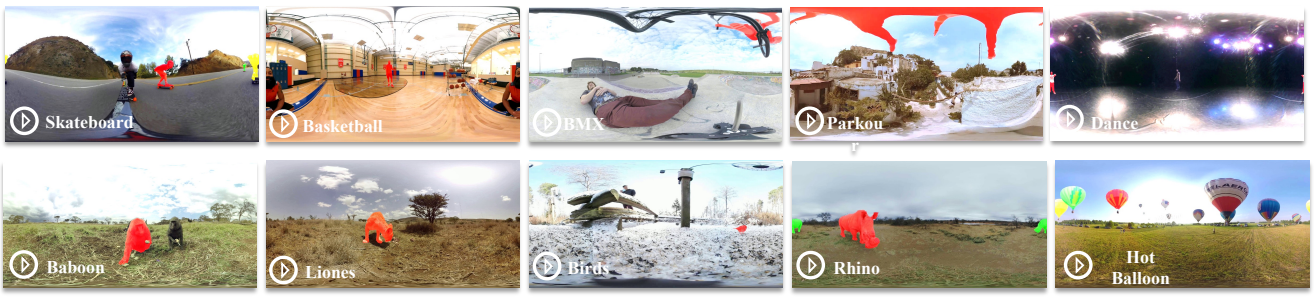

Figure 2. **PanoVOS dataset.** We select 10 samples from the dataset involving major scenes. For each video, there are high-quality instance-level pixel-wise masks, which display in color.

85 videos. Note that SOD360 [65] and Wild360 [8] 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 longer than SHD360 [65] (5s per video). By the way, the longer video is highlighted in a recent survey [51]. 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, 43]and YouTube-VOS [54] 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, 52, 33], which either train or fine-tune their networks with the first-frame ground truth at test time and are therefore a great wasteful of resources. OnAVOS [48] achieves promising results by introducing an online adaptation mechanism, but it still requires online fine-tuning. To a certain extent, it restricts networks' efficiency.

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,**

Propagation-based. Propagation-based models [4, 31, 40, 9, 37] get the target segmentation 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 [38, 7, 64, 34, 26, 55, 56] 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 [38], which has been extended to many works [45, 15, 53, 49, 29, 6]. AOT [60] is an excellent work that introduces an identification mechanism by encoding multiple targets into the same embedding space, which can simultaneously segment multiple objects. Despite the impressive results achieved by these methods, they still fail to address the challenges of the tremendous proportion of distortion and discontinuity under a panoramic input 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 [54] 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 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). 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 that are not present in the training set 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. 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.

3.3.1 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.

We then use the state-of-the-art semi-supervised video object segmentation method [60] to propagate the instance masks frame by frame from the annotated keyframes to untagged intermediate frames and generate masks at 6 fps.

3.3.2 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

In this section, we introduce the PSCFormer adapted to panoramic VOS, which improves the effectiveness of performance due to the key component Panoramic Space Consistency (PSC) block.

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 [61, 7, 6] 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. 3(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 [59, 55, 56], 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 to be segmented.

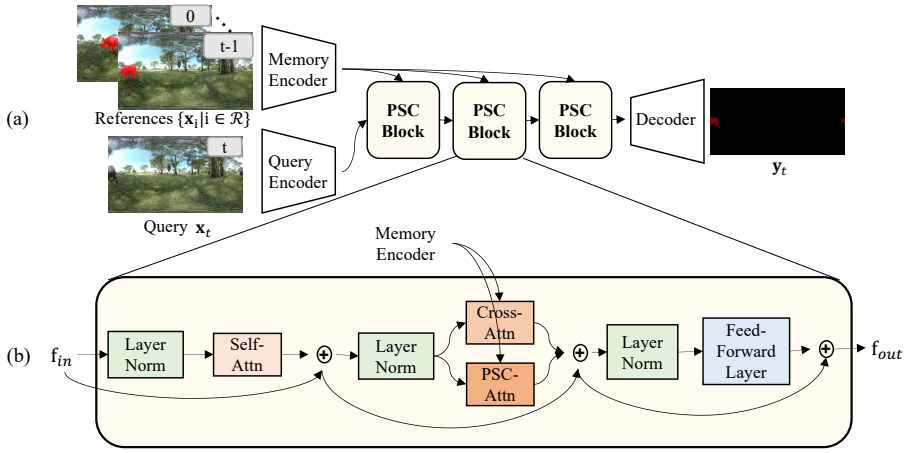

Figure 3. (a) **PSCFormer overview.** Given the query frame $\mathbf{x}_t$ and reference frames $\{\mathbf{x}_i|i \in \mathcal{R}\}$, the goal of VOS is to delineate objects from the background by generating mask $\mathbf{y}_t$ for query frame $\mathbf{x}_t$. References and the query frame are encoded 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) Detailed structure of a panoramic space consistency block.

4.2. Panoramic Space Consistency Block

Fig. 3(b) shows the structure of a PSC block. Motivated by the common transformer blocks [47], 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 [14] non-linearity activation function.

Panoramic Space Consistency Attention (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 become 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 left-most 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:

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

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,

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

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 follows:

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

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 [47], we implement the representational form of our PSCAttn module with multi-headed attention, de-

Methods	MF	YouTube-VOS	PanoVOS Validation					PanoVOS Test				
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
AOTT [60]		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 [60]		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 [60]		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[27]	✓	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 [7]	✓	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 [6]	✓	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 [60]	✓	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 [60]	✓	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 [60]	✓	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* [22]	✓	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* [7]	✓	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* [6]	✓	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 [54]. $*$ denotes a large-scale external dataset BL30K [7] dataset is used during training.

Methods	MF	YouTube-VOS	PanoVOS Validation					PanoVOS Test				
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
CFBI† [61]		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+† [61]		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 [60]		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 [60]		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 [60]		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[27]	✓	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 [22]	✓	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 [7]	✓	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 [6]	✓	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 [60]	✓	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 [60]	✓	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 [60]	✓	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* [22]	✓	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* [7]	✓	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* [6]	✓	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 [54]. $*$ denotes a large-scale external dataset BL30K [7] dataset is used during training. † denotes no synthetic data is used during the training stage.

fined 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 [47], we set the number of heads to $(h = C/d_{model})$ 8, where d_{model} is the projection dimension of each head.

5. Expriment

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-paranomic videos adapted to the panoramic world?

RQ2: Can the proposed PanoVOS datasets bring about a consistent performance gain to VOS methods?

RQ3: How does the proposed Panoramic Space Consistency Attention module contribute to VOS models?

RQ4: What are the remained problems for further research on panoramic VOS in academia?

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 [29, 60]. In our work, we set δ to 2 and 5 for training and testing respectively. For hyperparameters p and s in PSC block, we empirically set them as 2 and 7, respectively.

Training&Inference Details.

We adopt the two-stage training strategy used in most VOS methods [27, 6, 60, 7], where the training stage is divided into the pre-training and main-training. Following [60], during Pre-training, our models without the PSCAtten module are trained for 10,000 steps on a synthetic video sequence of 5 frames generated from static non-panoramic images [28, 24, 5, 63, 46, 50] by applying multiple data deformations. The initial learning rate for the pre-training stage is 4×10^{-4} and the weight decay is 0.03. During main-training, our models are trained on PanoVOS

Methods	MF	PanoVOS Validation					PanoVOS Test				
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
CFBI† [59]		35.8	34.6	44.8	24.2	39.7	19.1	18.2	26.1	12.2	19.8
CFBI+† [61]		41.3	38.0	47.9	32.5	46.9	30.9	30.8	42.7	21.4	28.5
AOTT [60]		65.6	59.4	68.3	59.7	75.0	53.4	49.3	61.6	47.5	55.1
AOTS [60]		67.7	61.2	70.0	62.4	77.1	55.9	**53.2**	**65.1**	48.6	57.0
AOTB [60]		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[27]	✓	34.3	34.8	42.8	24.9	34.5	34.2	28.2	38.8	32.9	36.8
RDE [22]	✓	50.5	49.7	58.4	39.2	54.9	42.5	36.9	46.6	38.5	48.2
STCN [7]	✓	52.0	51.2	60.8	41.5	54.5	50.8	43.6	56.5	49.3	53.7
XMem [6]	✓	55.7	54.8	63.3	45.2	59.7	53.5	49.5	62.6	47.1	54.8
AOTL [60]	✓	66.6	61.4	71.1	59.4	74.3	53.8	50.0	60.3	47.8	57.1
R50_AOTL [60]	✓	65.3	61.9	71.4	56.4	71.6	54.6	52.9	63.2	47.5	54.9
SwinB_AOTL [60]	✓	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* [22]	✓	54.3	52.8	61.6	44.6	58.2	52.2	44.5	56.0	49.3	59.1
STCN* [7]	✓	51.7	51.2	60.6	41.3	53.6	53.8	53.7	58.1	46.0	57.3
XMem* [6]	✓	57.7	55.6	64.6	48.6	61.9	57.9	51.3	64.5	53.2	62.7

Table 5. **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 [7] dataset is used during training.

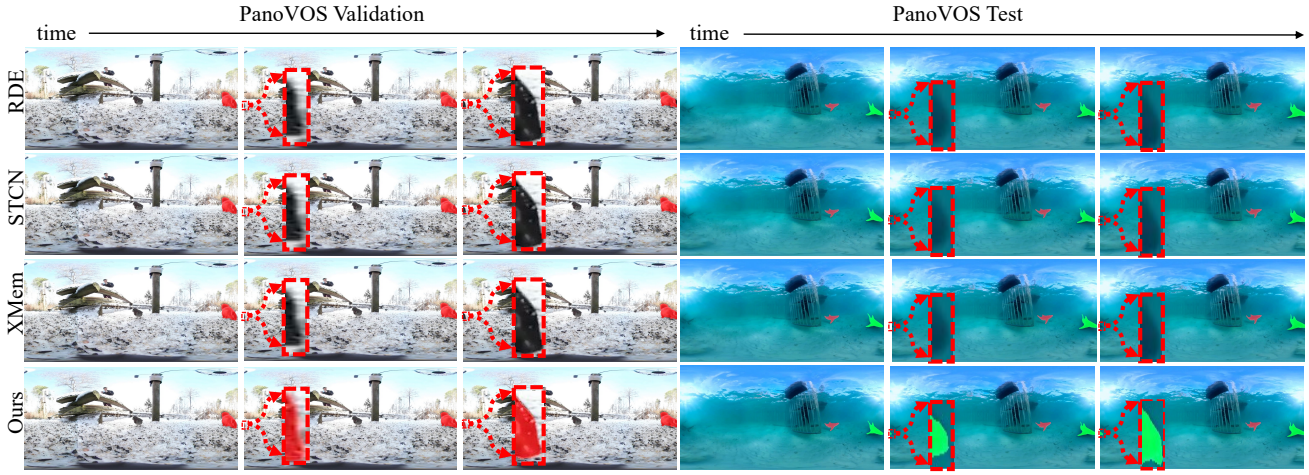

Figure 4. **Qualitative comparison to the state-of-the-art methods**, RDE[22], STCN[7], and XMem[6], on PanoVOS dataset. Our model performs better under the challenge of content discontinuities. Error regions are highlighted with zoom-in bounding boxes.

for 15,000 steps and we use an initial learning rate 2×10^{-4} with a weight decay of 0.07. Models are trained with four NVIDIA A100 GPUs using PyTorch [39] and the batch size is set as 16. All the models are optimized by AdamW [30] optimizer. MobileNet-V2 [44]is used as the backbone of memory encoder and query encoder, and the last-layer features with stride 16 are leveraged. One NVIDIA A100 GPU is used for evaluation.

Evaluation Metrics. Following the standard protocol [43, 41], we adopt the region accuracy $\mathcal{J}$ and boundary accuracy $\mathcal{F}$.1 $\mathcal{J}$ 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 models trained on planar video datasets, we evaluated 15 off-the-shelf VOS models, including [27, 59, 61, 22, 7, 6, 60], 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 [28] and ECSSD [46], 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 [28]) and video datasets (*e.g.* YouTube-VOS [54]), 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 in-

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

Table 6. **Ablation study** of *PSCAttn* module on PanoVOS.

sights. 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 [54] and BL30K [7] can slightly help mitigate the gap between planar and panoramic videos.

5.3. Main Results on PanoVOS (RQ2)

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 5, which demonstrates that all the previous VOS models perform worse on PanoVOS than on the traditional VOS benchmarks, *e.g.*, YouTube-VOS. And 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 to tackle panoramic videos. Fig.4 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.

5.4. Ablation Study (RQ3)

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 6 demonstrates the effectiveness of our PSCAttn module. Besides, Fig. 5 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 7, compared to the conventional cross-attention (CrossAttn) module, PSCAttn also achieves better performance.

In Table 8, 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 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 $\mathcal{J}\&\mathcal{F}$.

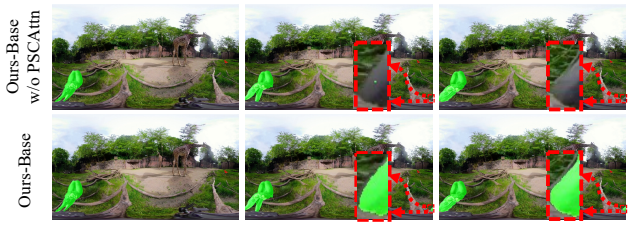

Figure 5. **Qualitative ablation study** of *PSCAttn* module.

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

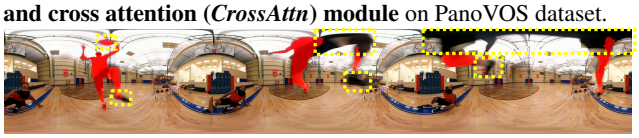

Figure 6. **Challenge.** Our model fails to segment some objects with strong distortion.

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 8. **Hyperparameter Analysis** of p, which enables the stitching mechanism, in PSCAttn for Ours-Large model on PanoVOS.

5.5. Limitation and Future Work (RQ4)

To prompt greater progress in the academic world of panoramic VOS, we also analyze the limitation of our method. In specific, our method has no notion of severe distortion challenge since we do not employ a special design (such as deformable convolution [11]) to tackle deformations. That means our model may not be able to segment the objects with a large of distortions. One such failure case is shown in Fig 6. Besides, our panoramic dataset can be applied to other fields in video segmentation and tracking domains, such as referring video object segmentation [57], video object tracking [21], video instance segmentation [58], few-shot segmentation [16] and so on. Also, we believe that it would be valuable to investigate the zero-shot segmentation performance of visual foundation models [20] on our challenging panoramic dataset. We hope our work can shed light on the development of efficient adaptation from non-panoramic vision to panoramic vision perception.

6. Conclusion

In this paper, we introduce a high-quality dataset 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 further 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 also hope that our PanoVOS dataset can attract more researchers to pay attention to panoramic videos.

A. More Details about PanoVOS Dataset

A.1. Additional Dataset Statistics

To create our PanoVOS, we carefully selected videos with relatively large motion amplitudes and chose 35 video categories from three domains including person (*e.g.* parkour, dance, BMX, skateboard), animals (*e.g.* elephant, monkey, giraffe, rhino, birds) and common objects (*e.g.* basketball, hot balloon). Figure A shows the histogram of instance masks for all categories.

A.2. Detailed Annotation Pipeline

A.2.1 Problems of Previous Annotation Pipelines

The segmentation task is a pixel-level intensive task that requires the classification of each pixel in the image. Annotation is very time-consuming and expensive for a large-scale panoramic video dataset. Although there are many image annotation tools like LabelMe and EiSeg, annotating a video-level dataset at an image level can be difficult as follows. First, the state of the object is constantly changing between frames in the video. When annotating multiple images, you must use the same label for the same object, but there is a high risk of consistency errors. Second, for the vanishing and reappearance problem in some video clips, there is a high probability that the object will be incorrectly identified as another object when it reappears later. Third, losing the continuity of information in the video itself, like annotating a large number of images, is a burdensome project.

A.2.2 Overview of Our Human-computer Joint Annotation Pipeline

In order to obtain accurate large-scale video panoramic segmentation annotations and efficiently tackle the difficulties mentioned above, we propose a human-computer joint annotation pipeline as is shown in Figure B. First, key frames 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 key frames to those unlabeled intermediate frames using a method that has performed well in semi-supervised video object segmentation. Finally, due to the characteristics of the panoramic video such as distortions and discontinuities, we need to recalibrate the resulting annotations until all the pixels meet the requirements.

A.2.3 Annotation Propagation

To save manpower and finance, we propose a semi-automatic human-computer collaboration annotation strategy to annotate the panoramic segmentation dataset. 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 key frames with a speed of 1 fps are selected for manual annotation. In order 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. We then use the state-of-the-art semi-supervised video object segmentation method, AOT [60], to propagate the instance masks frame by frame from the annotated key frames to those untagged intermediate frames, and generate masks at 6 fps.

A.2.4 Annotation Refinement

After obtaining the masks from the first 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) The huge distortions and discontinuities present in the panoramic video are not used for the training of the propagation VOS method, so the quality of the masks obtained is relatively poor when this situation occurs in panoramic videos. The correction time for a video depends on the complexity of the video (the two main difficulties mentioned above) and usually takes between 30-60 minutes. Manual correction of the mask is checked by another annotator until the result is satisfactory before proceeding to the next video annotation.

A.3. Remark on PanoVOS Dataset

Difference with SHD360 [65, 19] dataset. PanoVOS and SHD360 are different in the target task: PanoVOS aims at object-agnostic video object segmentation; SHD360 aims at human-centric salient detection. Besides, we clarified the superiority of PanoVOS for VOS evaluation in three aspects: 1) **More diverse**: PanoVOS (150 videos with 36 subcategories); SHD360 (41 human-centric videos). 2) **Larger**

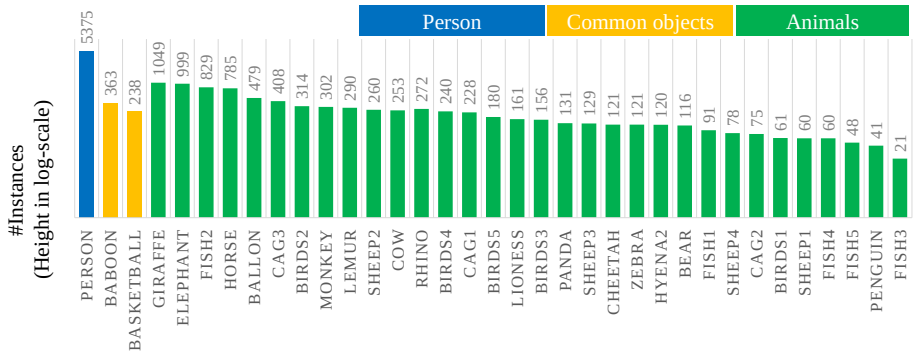

Figure A. Instance-level distribution of the proposed PanoVOS dataset. Our dataset mainly contains three major divisions: *person*, *animals*, and *common objects* with 35 minor sub-divisions.

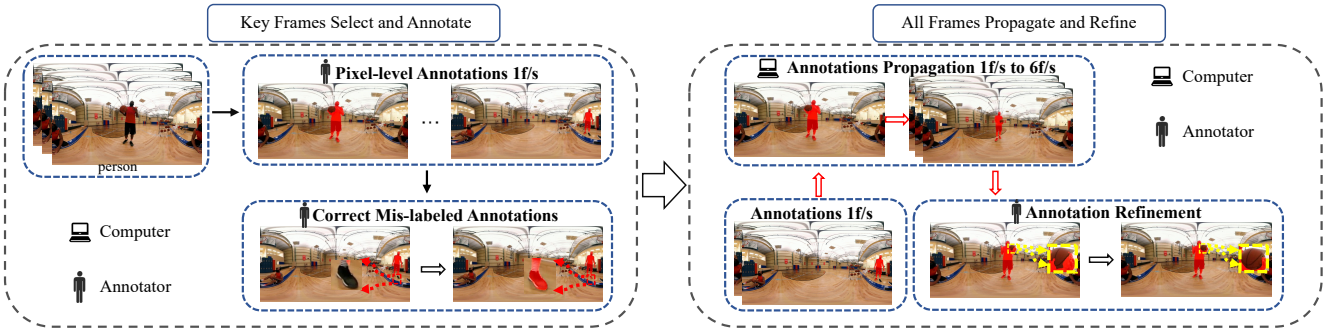

Figure B. Our annotation pipeline includes two phases. (1) The first phase is called *Key Frames Select and Annotate*. The annotator browses the video and picks out the object to be annotated. Then, instances are manually annotated at 1 fps and corrected by another annotator. (2) The second phase is 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.

object motion: PanoVOS (large motion); SHD360 (little motion). 3) **Longer video**: PanoVOS (20s); SHD360 (5s).

Comparisons to panoptic segmentation datasets [35]. we differentiate our work from them as follows. 1) There is a slight difference between the definition of *Panoramic videos* between our work and [35]. *Panoramic videos* refers to **complete (360°, full view)** panoramic videos in our work, while it refers to range-view images collected from multi-cameras in [35], which only include **wide but partial views**. 2) Our PanoVOS task aims at pixel-level object-agnostic video segmentation in panoramic videos, while [35, 19] aim at panoptic segmentation. We will supplement the comparisons in the revision.

Imaging way of the panoramic lens for PanoVOS. The imaging ways of panoramic videos that we collected online are not guaranteed to be the same. This is because the info about the imaging way of these videos is missing and not provided. We argue that it's not trivial to explore the influence of imaging ways but we hope there could be some future work study on this.

B. More Results

Domain transfer result of PSCFormer. The motivation for conducting domain transfer studies (Tab. 3 and Tab. 4 of the main paper) is to answer the research question of *how well are current VOS methods trained on conventional non-panoramic video datasets adapted to the panoramic world*. Thus, we considered it improper to add the proposed method, which is tailored for panoramic scenes, for comparisons in Tab. 3 and Tab. 4. However, we supplement domain transfer results of the proposed methods (Tab. A and Tab. B) here for future potential comparisons.

C. More Implementation Details

For our models, flipping, scaling, and resizing are applied for data augmentations. And the window size is 465×465 after resizing. For speeding up training, we use Automatic Mixed Precision (AMP) [39] with automatic gradient scaling. The bootstrapped cross-entropy loss and soft Jaccard loss [36] are used and the proportion is 1:1. Besides, we apply Exponential Moving Average (EMA) [42] for stabilizing the training, which is similar to [60]. For the detailed static image datasets for pre-training of vari-

Methods	MF	YouTube-VOS $\mathcal{J}\&\mathcal{F}$	PanoVOS Validation $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	PanoVOS Test $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Ours-Base	✓	72.9	58.2$\downarrow_{14.7}$	50.8	63.6	51.1	67.4	48.4 $\downarrow_{22.4}$	34.7	45.8	52.3	61.0
Ours-Large	✓	71.9	55.8$\downarrow_{16.1}$	49.2	64.0	45.5	64.6	41.2$\downarrow_{30.7}$	31.7	42.3	41.1	49.7

Table A. **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 [54].

wo

Methods	MF	YouTube-VOS $\mathcal{J}\&\mathcal{F}$	PanoVOS Validation $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	PanoVOS Test $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Ours-Base	✓	83.0	72.2$\downarrow_{10.8}$	62.7	75.8	68.5	81.8	60.6$\downarrow_{22.4}$	57.5	64.2	56.0	64.8
Ours-Large	✓	83.7	73.1$\downarrow_{10.6}$	63.9	75.7	69.4	83.3	61.6$\downarrow_{22.1}$	54.5	65.6	58.8	67.5

Table B. **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 [54].

ous models, following official implementations, STCN [7], RDE [22], and XMem [6] use ECSSD [46], FSS1000 [24], HRSOD [63], BIG [5], DUTS [50]; AFB-URR [27] and AOT [60] use COCO [28], PASCAL(-S) [25], PASCAL VOC2012 [13], ECCSD [46], MSRA10K [10].

D. Qualitative Ablation Study of Panoramic Space Consistency Attention

We provide more cases to illustrate the effect of our Panoramic Space Consistency Attention (PSCAttn) module in Figure C and Figure D. As shown, our full model performs better in terms of handling the challenge of pixel-level content discontinuity.

E. More Qualitative Comparisons

We present more qualitative cases to show the performance of previous state-of-the-art models, including RDE [22], STCN [7], XMem [6] and our model, under panoramic scenarios on the PanoVOS Validation set and Test set. As demonstrated, our model is robust to many challenging cases, including similar objects (Figure E), background interference (Figure F), and content discontinuities (Figure G, Figure H, Figure I, and Figure J).

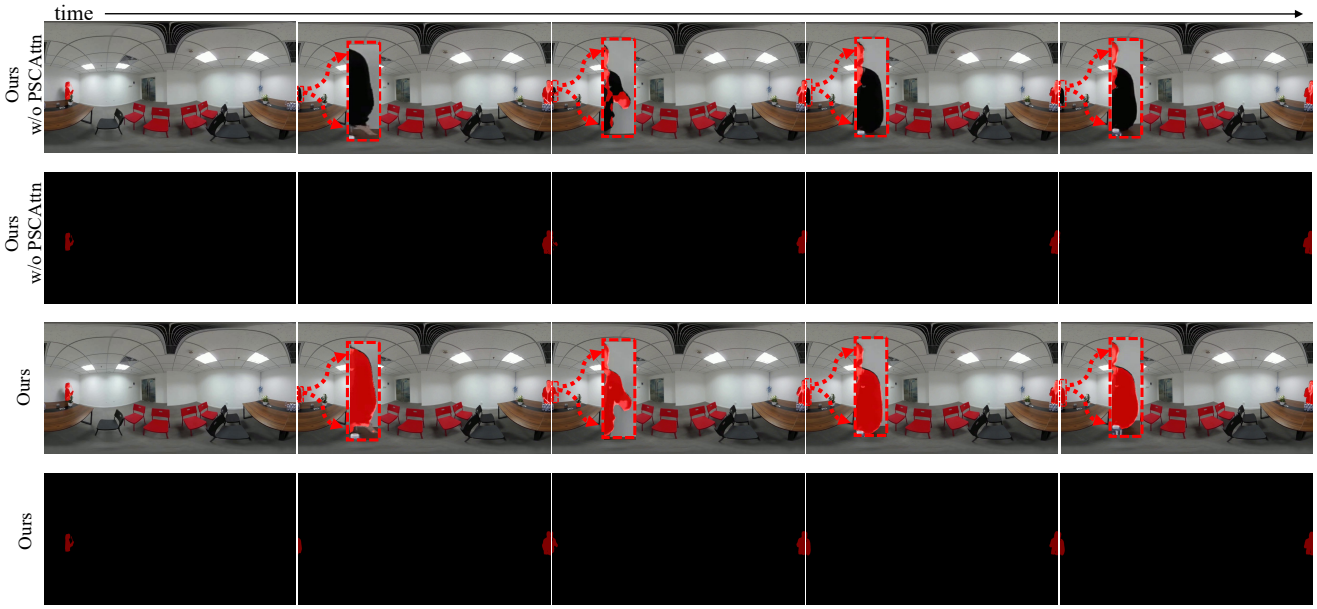

Figure C. Qualitative ablation study of Panoramic Space Consistency Attention (*PSCAttn*) module. Our model with *PSCAttn* can effectively help keep consistency in the panoramic space. With *PSCAttn*, our model successfully segments the *man* who disappears on the left boundary of the image and reappears on the right boundary of the image. Error regions are zoomed in and highlighted.

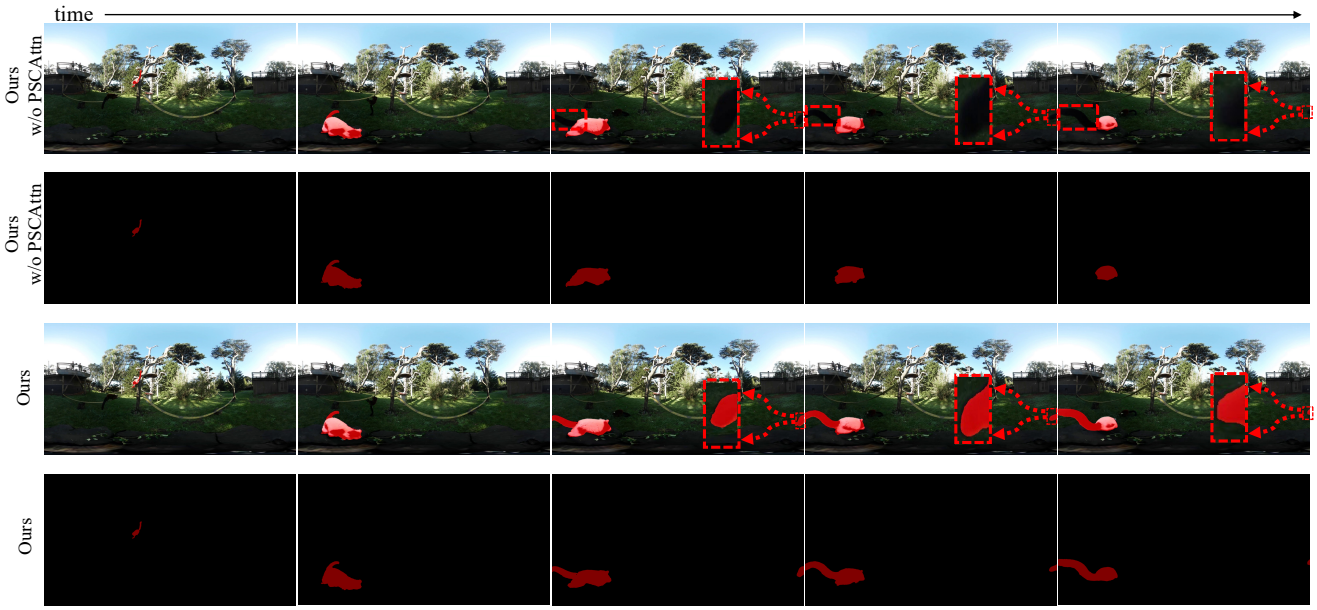

Figure D. Qualitative ablation study of Panoramic Space Consistency Attention (*PSCAttn*) module. Our model with *PSCAttn* can effectively help keep consistency in the panoramic space. With *PSCAttn*, our model successfully segments the *lemur* that disappears on the right boundary of the image and reappears on the left boundary of the image. Error regions are zoomed in and highlighted.

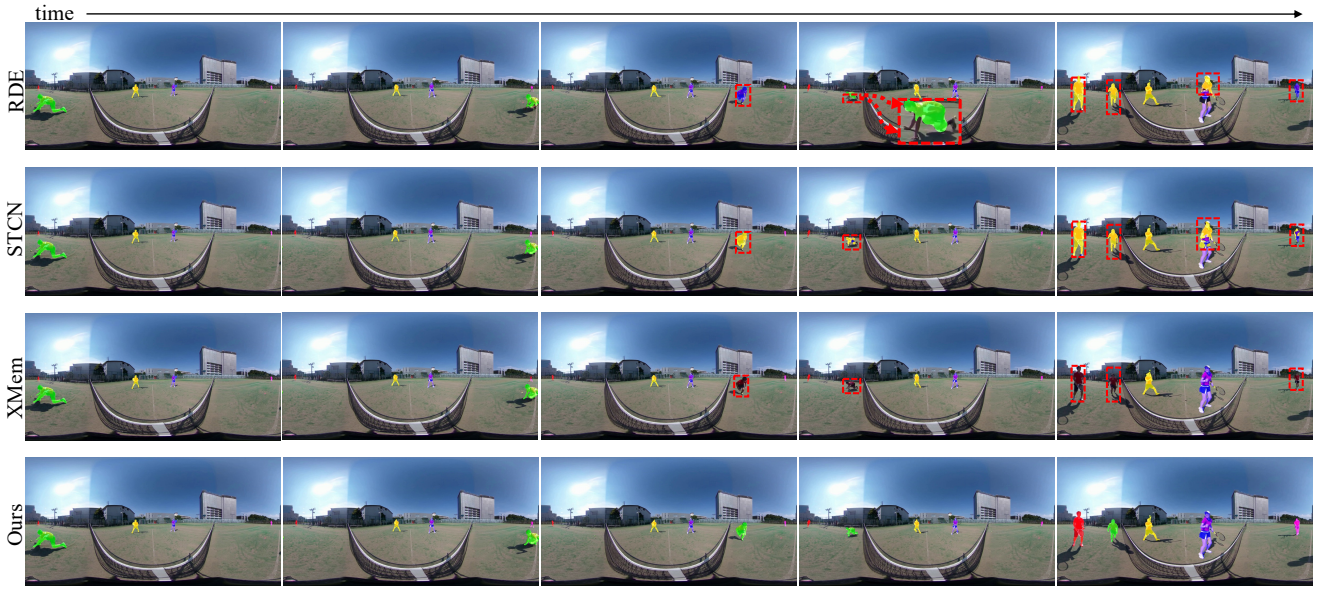

Figure E. Qualitative comparisons on PanoVOS Validation set and test set. Compared with the state-of-the-art methods, including RDE[22], STCN[7], and XMem[6], our model (Ours) performs better under the challenge of multiple similar objects (people). Error regions are highlighted with bounding boxes.

Figure F. Qualitative comparisons on PanoVOS Validation set and test set. Compared with the state-of-the-art methods, including RDE[22], STCN[7], and XMem[6], our model (Ours) performs better under the challenge of background interference and large image distortions. Error regions are highlighted with zoom-in bounding boxes.

Figure G. Qualitative comparisons on PanoVOS Validation set and test set. Compared with the state-of-the-art methods, including RDE[22], STCN[7], and XMem[6], our model (Ours) performs better under the challenge of content discontinuities. Error regions are highlighted with zoom-in bounding boxes.

Figure H. Qualitative comparison to the state-of-the-art methods, RDE[22], STCN[7], and XMem[6] on PanoVOS Validation set and test set. Compared with others, our model (Ours) performs better under the challenge of content discontinuities. Error regions are highlighted with zoom-in bounding boxes.

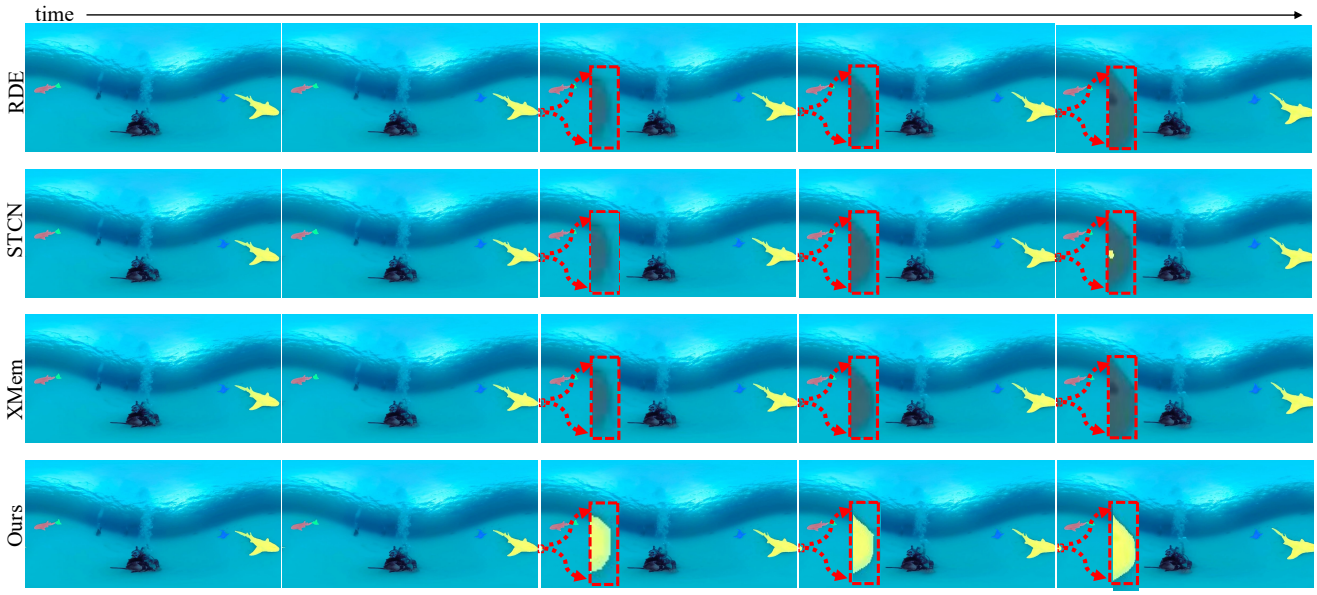

Figure I. Qualitative comparisons on PanoVOS Validation set and test set. Compared with the state-of-the-art methods, including RDE[22], STCN[7], and XMem[6], our model (Ours) performs better under the challenge of content discontinuities. Error regions are highlighted with zoom-in bounding boxes.

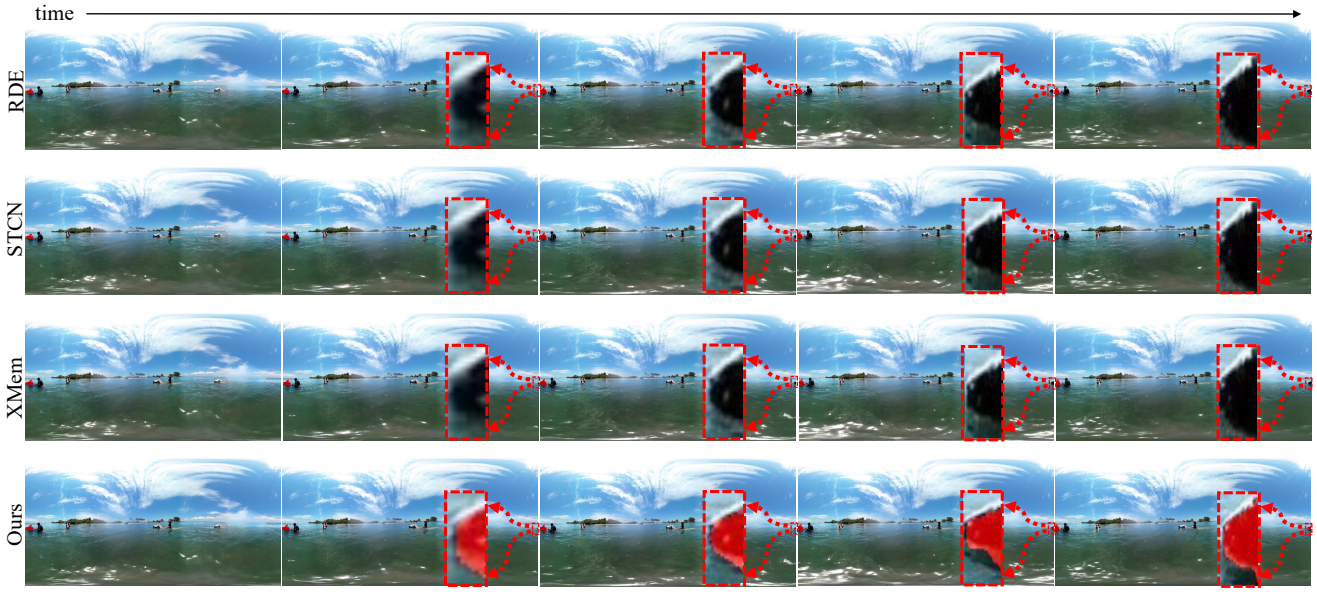

Figure J. Qualitative comparisons on PanoVOS Validation set and test set. Compared with the state-of-the-art methods, including RDE[22], STCN[7], and XMem[6], our model (Ours) performs better in the case of object reappear. Error regions are highlighted with zoom-in bounding boxes.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 1

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2

[3] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 3

[4] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2020. 3

[5] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8890–8899, 2020. 6, 11

[6] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 3, 4, 6, 7, 11, 13, 14, 15

[7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 3, 4, 6, 7, 8, 11, 13, 14, 15

[8] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2018. 2, 3

[9] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7415–7424, 2018. 3

[10] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 11

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 8

[12] Daniel Eger Passos and Bernhard Jung. Measuring the accuracy of inside-out tracking in xr devices using a high-precision robotic arm. In *International Conference on Human-Computer Interaction*, pages 19–26. Springer, 2020. 1

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 11

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[15] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021. 3

[16] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*, 2022. 8

[17] Hao Jiang, Gangyi Jiang, Mei Yu, Yun Zhang, You Yang, Zongju Peng, Fen Chen, and Qingbo Zhang. Cubemap-based perception-driven blind quality assessment for 360-degree images. *IEEE Transactions on Image Processing*, 30:2364–2377, 2021. 1

[18] Tyler A Jost, Bradley Nelson, and Jonathan Rylander. Quantitative analysis of the oculus rift s in controlled movement. *Disability and Rehabilitation: Assistive Technology*, 16(6):632–636, 2021. 1

[19] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020. 9, 10

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 8

[21] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2711–2738, 2021. 8

[22] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1332–1341, 2022. 6, 7, 11, 13, 14, 15

[23] Xiang Li, Haoyuan Cao, Shijie Zhao, Junlin Li, Li Zhang, and Bhiksha Raj. Panoramic video salient object detection with ambisonic audio guidance. *arXiv preprint arXiv:2211.14419*, 2022. 1

[24] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020. 6, 11

[25] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 280–287, 2014. 11

[26] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. Video object segmentation with dynamic memory networks and adaptive object alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8065–8074, 2021. 3

[27] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020. 6, 7, 11

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 7, 11

[29] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *European Conference on Computer Vision*, pages 648–665. Springer, 2022. 3, 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[31] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018. 3

[32] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2766–2772. IEEE, 2021. 2

[33] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018. 3

[34] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9670–9679, 2021. 3

[35] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. 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*, pages 53–72. Springer, 2022. 2, 10

[36] Sebastian Nowozin. Optimal decisions from probabilistic models: the intersection-over-union case. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 548–555, 2014. 10

[37] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385, 2018. 3

[38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 3

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7, 10

[40] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017. 3

[41] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 3, 7

[42] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 10

[43] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3, 7

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[45] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *European Conference on Computer Vision*, pages 629–645. Springer, 2020. 3

[46] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 6, 7, 11

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 6

[48] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 3

[49] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1296–1305, 2021. 3

[50] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 6, 11

[51] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation. *arXiv preprint arXiv:2107.01153*, 2021. 3

[52] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1140–1148, 2018. 3

[53] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021. 3

[54] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 3, 4, 6, 7, 8, 11

[55] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2946–2954, 2022. 3, 4

[56] Xiaohao Xu, Jinglu Wang, Xiang Ming, and Yan Lu. Towards robust video object segmentation with adaptive object calibration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2709–2718, 2022. 3, 4

[57] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. *arXiv preprint arXiv:2305.16318*, 2023. 8

[58] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 8

[59] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 4, 7

[60] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 2, 3, 4, 6, 7, 9, 10, 11

[61] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4, 6, 7

[62] Mingze Yuan and Christian Richardt. 360 optical flow using tangent images. In *British Machine Vision Conference (BMVC)*, 2021. 1

[63] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7234–7243, 2019. 6, 11

[64] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2020. 3

[65] Yi Zhang, Lu Zhang, Kang Wang, Wassim Hamidouche, and Olivier Deforges. Shd360: A benchmark dataset for salient human detection in 360 videos. *arXiv preprint arXiv:2105.11578*, 2021. 2, 3, 9

[66] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 488–503, 2018. 2