# Isomer: Isomerous Transformer for Zero-shot Video Object Segmentation

Yichen Yuan[1] Yifan Wang[1*] Lijun Wang[1] Xiaoqi Zhao[1] Huchuan Lu[1]
Yu Wang[2] Weibo Su[2] Lei Zhang[2,3]
[1]School of Information and Communication Engineering, Dalian University of Technology, China
[2]OPPO Research Institute [3]The Hong Kong Polytechnic University

## Abstract

*Recent leading zero-shot video object segmentation (ZVOS) works devote to integrating appearance and motion information by elaborately designing feature fusion modules and identically applying them in multiple feature stages. Our preliminary experiments show that with the strong long-range dependency modeling capacity of Transformer, simply concatenating the two modality features and feeding them to vanilla Transformers for feature fusion can distinctly benefit the performance but at a cost of heavy computation. Through further empirical analysis, we find that* ***attention dependencies learned in Transformer in different stages exhibit completely different properties****: global query-independent dependency in the low-level stages and semantic-specific dependency in the high-level stages. Motivated by the observations, we propose two Transformer variants: i) Context-Sharing Transformer (CST) that learns the global-shared contextual information within image frames with a lightweight computation. ii) Semantic Gathering-Scattering Transformer (SGST) that models the semantic correlation separately for the foreground and background and reduces the computation cost with a soft token merging mechanism. We apply CST and SGST for low-level and high-level feature fusions, respectively, formulating a* ***level-isomerous Transformer*** *framework for ZVOS task. Compared with the baseline that uses vanilla Transformers for multi-stage fusion, ours significantly increase the speed by* $\mathbf{13}\times$ *and achieves new state-of-the-art ZVOS performance. Code is available at https://github.com/DLUT-yyc/Isomer.*

## 1. Introduction

Zero-shot Video Object Segmentation (ZVOS) aims at discovering the most visually attractive objects in a video sequence and serves as a fundamental computer vision technique. Different from image segmentation that mainly relies on static appearance features, ZVOS further explores temporal motion information to achieve reliable and temporally consistent results. One popular pipeline [39, 17, 59, 42] is integrating appearance and motion information by identically applying feature fusion modules in multiple stages as shown in Fig. 1 (a). While great efforts have been made, designing effective multi-stage appearance-motion fusion approaches for ZVOS is still an open problem.

Transformers [48] have made remarkable breakthroughs in many computer vision tasks [6, 3, 55, 28] due to its strong capability in modeling long-range dependencies and unique flexibility for cross-modal feature fusion. Nevertheless, their merits have not been fully explored in the ZVOS field. A straightforward way is adopting Transformer blocks as the appearance-motion fusion modules. In our preliminary experiments, for each feature level, we concatenate the extracted appearance and motion features and feed them to a vanilla Transformer block (Fig. 1 (b)). It shows that such a simple baseline achieves superior performance than all prior elaborate-designed approaches but at a cost of heavy computation. These motivate us to further investigate: 1) what Transformers exactly learn for performance gain, and 2) how to further relieve the computational burden without performance loss under this baseline framework.

To answer the above questions, we visualize the attention dependencies computed by the Multi-Head Self Attention (MHSA) step of Transformers in all feature fusion stages. Surprisingly, it finds that *vanilla Transformers in different levels characterize the attention dependencies from different perspectives to fit the ZVOS task*, which motivates our design of Transformer-based ZVOS framework as follows.

**First, Transformers in early fusion stages only capture global query-independent dependency.** As shown in Fig. 1 (d)(e), the attention maps of different query positions are almost the same for the low-level stages, which mainly highlight the foreground object and some background contours. It indicates that the network tends to understand the scenes via global context modeling in shallow layers and tries to distinguish the boundary between foreground and background under the ZVOS setting. Inspired by [2], we

*Corresponding author: wyfan@dlut.edu.cn

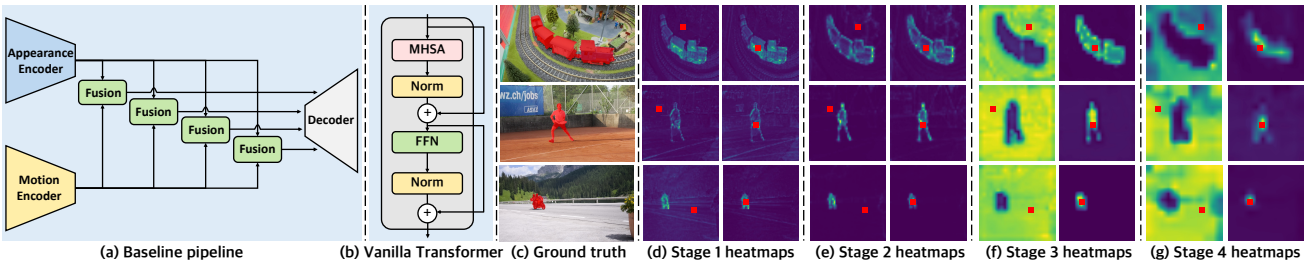


Figure 1. Visualization of the learned attention dependencies of the baseline pipeline (a) using vanilla Transformer (b) as fusion modules. (d)-(g) present the attention maps (heatmaps) of different queries (red points) computed by the vanilla Transformer in different fusion stages: for the same input image, different queries have almost the same attention maps in low-level stages (d) (e), while they only focus on image regions with the same semantics (*i.e.*, object or background) in high-level stages (f) (g).

propose to simplify vanilla Transformer by computing one query-independent dependence map for all query tokens, thus modeling the global-shared contextual information and largely reducing the computational cost. We term the simplified Transformer as Context-Sharing Transformer (CST).

**Second, Transformers in late fusion stages capture long-range semantic-specific dependency.** As shown in Fig. 1(f)(g), the query tokens mainly pay attention to the image regions with the same semantic category, *i.e*. foreground or background. Besides, the attention maps of the query tokens with the same semantics share many similarities, indicating existing much attention redundancy that could be further pruned. Based on these observations, we propose another Transformer variant named Semantic Gathering-Scattering Transformer (SGST), which computes foreground and background attentions separately for the corresponding query tokens with some selected representative key/value tokens. In addition, a soft token merging mechanism is adopted to enable the token selection process differentiable. Compared to the vanilla Transformer, the proposed SGST is able to model the semantic-specific dependency more explicitly with less computation redundancy.

Upon the above findings and the two proposed Transformer blocks, a level-**Iso**merous Transfor**mer** (Isomer) ZVOS framework is formulated by applying CST and SGST blocks for early and late fusion stages, respectively. Compared with the baseline network that applies vanilla Transformer uniformly for all the feature fusion stages, ours treats the different fusion levels distinctively based on the observed Transformers properties, and achieves better segmentation results with $\mathbf{13}\times$ inference speed. Compared with the existing ZVOS works, our method equipped with Swin-Tiny[28] backbone (a comparable model size to ResNet50[14]) obtains significantly superior performance with real-time inference.

The main contributions of this work are as follows:

1) We analyze the properties of vanilla Transformers in terms of attention dependencies hierarchically learned from the ZVOS task, and propose two Transformer variants, *i.e*. Context-Sharing Transformer (CST) and Semantic Gathering-Scattering Transformer (SGST), to model the contextual dependencies from different levels effectively and efficiently.

2) We propose a level-isomerous Transformer paradigm for the ZVOS task, which applies the developed CST and SGST for low-level early fusion and high-level late fusion, respectively. Different from the prior works that fuse appearance-motion information for all stages in an identical way, ours performs different fusion levels differentially and better fits the properties of the ZVOS network.

3) Extensive experiments demonstrate the superiority of our method compared to the existing works as well as the strong vanilla Transformer-based baseline. To our best knowledge, this is the first successful attempt at developing a real-time Transformer-based work in the ZVOS field.

## 2. Related Work

### 2.1. Zero-shot Video Object Segmentatio n

Zero-shot video object segmentation (ZVOS) aims to automatically segment the salient objects from videos without any manual prompt. It has witnessed rapid progress with the development of deep learning techniques and the establishment of large-scale datasets [56, 41]. Early CNN-based methods [47, 51, 10] usually use recurrent neural networks to capture long-term dependencies. Inspired by the attention mechanism, several works [31, 50, 61] explore global context corrections between frames by designing cross-attention operations. Recent leading works [72, 17, 59, 39, 66] combine the appearance information with the motion cues extracted by the off-the-shelf optical flow methods [16, 45, 46] and have gained significant performance improvement. Among them, [17] designs a relational cross-attention module to achieve bi-directional message propagation in the appearance and motion subspaces. AMCNet [59] proposes an attentive multi-modality collaboration network to utilize appearance and motion informa-

tion uniformly. HFAN [39] proposes a sequential feature alignment module and a feature adaptation module for appearance and motion feature alignment. While promising performance has been achieved, the spirit of Transformers has not been fully explored in the existing ZVOS methods. In this work, we study the properties of Transformers in ZVOS setting in-depth, and propose two novel Transformer blocks and a level-isomerous Transformer framework, hoping to provide some new insights for this field.

### 2.2. Video Salient Object Detection

The task of Video Salient Object Detection (VSOD) is similar to ZVOS. The difference is that the ZVOS model predicts a binarized segmentation mask, while the VSOD model predicts a continuous-valued probabilistic saliency map[68, 69, 11, 37, 38]. Most works resort to capturing the temporal information by using recurrent neural networks [43, 21, 10] or the inter-frame motion cues [22, 64]. For example, [22] develops a two-branch network for appearance and motion feature extraction and introduces a motion-guided attention module to enhance appearance features with motion information. Similar to ZVOS, the prior VSOD works are also mainly based on convolutional neural networks and non-local networks, and our main contributions have not been explored in this field.

### 2.3. Lightweight Transformer in Vision Tasks

Transformer [48] is first proposed for sequence-to-sequence machine translation [5, 60]. Recently, it is successfully migrated into many computer vision tasks such as image classification [6, 28, 52], object detection [3, 73], image segmentation [71, 55, 44], *etc*. However, the enormous computational load and memory usage make Transformer difficult to be deployed, especially for video dense prediction tasks (*e.g.*, VOS and VSOD). To address this problem, many research efforts are taken for lightweight Transformers. [25, 33, 35, 28, 58, 12, 7, 23]. Among them, one direction is to combine lightweight CNN and attention mechanism to form a hybird architecture [12, 4, 33, 65]. Another track is to reduce the quadritic computation complexity of the attention mechanism [36, 58, 52, 53, 62, 35]. Wang *et al*. [52] design a spatial-reduction attention to reduce resource consumption. [23] reduces the computational complexity of Transformer by query/key selection strategy and local attention computation. [7] proposes a sparse Transformer formulation using grid attention and strided attention for video segmentation. As opposed to the above works that rely on hand-designed rules for Transformer lightening in a data-agnostic manner, we simplify vanilla Transformer based on the observed properties of Transformers under the ZVOS setting, which is able to learn task-aware attention in a more flexible data-driven way, yielding both accuracy and efficiency gain for ZVOS problem (see Sec. 4.3).

The most relevant work for ours is GCNet [2], which proposes a global context block to simplify the non-local network [54]. However, our work has significant differences compared to GCNet. First, GCNet only visualizes the high-level non-local neural networks and proposes to learn the query-independent attention map for all query positions. On the contrary, we analyze both low-level and high-level Transformers under the ZVOS task, and find that the attention dependencies modeled by Transformer are totally different along the network depth: global query-independent dependency in the low-level stages and semantic-specific dependency in the high-level stages. Second, it should be recognized that our CST is inspired by GCNet. Nevertheless, the design of SGST and our main insight about the level-isomerous framework are unique.

## 3. Methodology

### 3.1. Vanilla Transformer Baseline

The baseline method is designed following the commonly adopted framework [17, 39] as shown in Fig. 1 (a). It consists of an appearance backbone, a motion backbone, multiple fusion modules, and a decoder.

Given the current video frame and the optical flow map that is computed between the current frame with its adjacent one, the appearance backbone and motion backbone extract four-stage *appearance features* $\mathbf{I}_l$ and *motion features* $\mathbf{M}_l$ ($l \in \{1, 2, 3, 4\}$), respectively. For each stage, one fusion module is applied to integrate appearance and motion features. Specifically, the extracted two modality features ($\mathbf{I}_l$ and $\mathbf{M}_l$) are first channel-wise combined to obtain a mixing representation $\mathbf{X}_l \in \mathbb{R}^{C\times H\times W}$, where $C, H, W$ denote the channel number, height, and width of $\mathbf{X}_l$, respectively. It can also be regarded as a list of tokens $\{\mathbf{x}_i|\mathbf{x}_i \in \mathbb{R}^C, i = 1, 2, \ldots, N\}$ with $N = H \times W$. Then, the mixing representation is fed into a vanilla Transformer block (see Fig. 1(b)) for cross-modality feature fusion. Following previous works [39, 59, 17], we use a feature pyramid decoder [55] to leverage the four-stage fused features for the final segmentation prediction.

Without bells and whistles, the vanilla Transformor-based baseline achieves outstanding performance but also brings about a heavy computational burden, which motivates us to further explore an effective solution to making a trade-off between performance and computation.

### 3.2. Analysis and Motivation

Visualizing the query-specific attention dependencies could help understanding what Transformers exactly learn during feature fusion. Therefore, we visualize the attention heatmaps computed by the Multi-Head Self Attention (MHSA) step of Transformers in all fusion stages of the baseline network in Fig. 1. The observations motivate us to

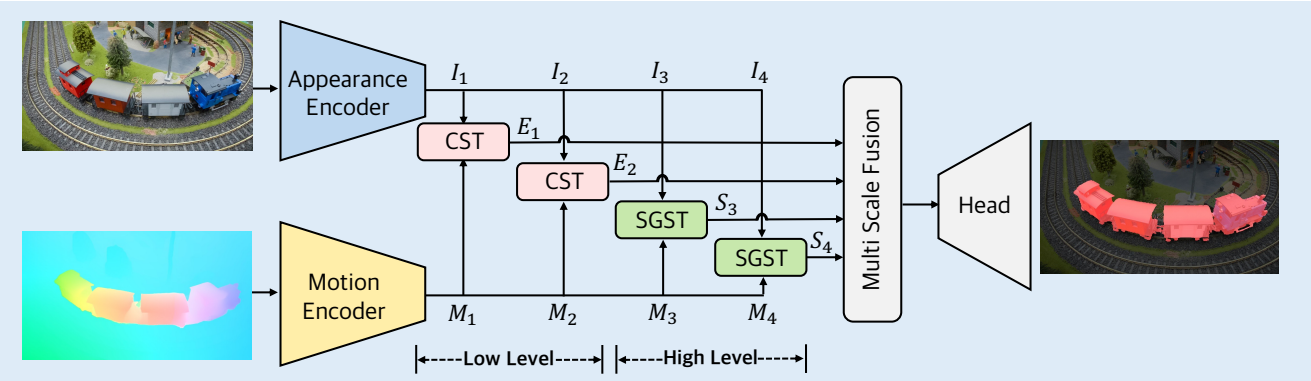


Figure 2. Overview of the proposed framework (Isomer). Given the current video frame and its corresponding optical flow map, Isomer extracts hierarchical appearance and motion features by two backbones. Then the proposed **CST (Context-Sharing Transformer)** and **SGST (Semantic Gathing-Scattering Transformer)** are adopted for cross-modality feature fusion in the low-level (the first two) and high-level (the last two) stages, respectively. The multi-stage fused features are fed into a segmentation head to obtain the final result.

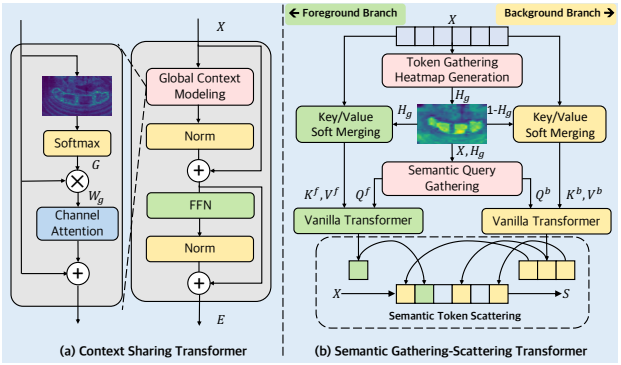


Figure 3. Illustration of the proposed CST and SGST modules.

rethink the baseline network in-depth. First, Transformers in early fusion layers only learn query-independent dependency and tend to understand the scenes via global context modeling. Inspired by [2], we claim that it can be simplified by computing global query-independent attention for all query positions, rather than computing one query-specific attention for one query position. Second, Transformers in late fusion layers learn to distinguish high-level semantics, and there exists much attention redundancy for both foreground and background queries. Therefore, we can select some representative queries from the foreground and background and model the semantic dependencies respectively, thus being more computationally effective.

## 3.3. Isomerous Transformer for ZVOS

Driven by the above findings, we reform the vanilla Transformer based fusion blocks in the baseline by developing two new Transformer blocks, named Context-Sharing Transformer (CST) and Semantic Gathering-Scattering Transformer (SGST). We apply them for multi-level feature fusion following the observed network properties and thus form a new level-isomerous Transformer (Isomer) framework. In the following, we will first illustrate the proposed blocks, followed by an overview of the whole network.

### 3.3.1 Context Sharing Transformer

The Context-Sharing Transformer (CST) is developed to simplify the MHSA step in vanilla Transformer with a global context modeling as shown in Fig. 3 (a), which computes global query-independent attention for all queries.

The global context modeling step first performs a query-shared spatial-wise attention followed by a channel-wise attention. Specifically, given a mixing representation $\mathbf{X}_l \in \mathbb{R}^{C\times H\times W}$ from one low-level stage $l(l \in \{1,2\})$, it first generates a one-channel attention weight map $\mathbf{G}_l \in \mathbb{R}^{H\times W}$ by a $1\times 1$ convolution with a Softmax function, which is used to weight $\mathbf{X}_l$ to obtain a query-shared weighted representation $\mathbf{W}_l^g \in \mathbb{R}^C$. Then two $1\times 1$ convolution layers interleaved by BN and ReLU are adopted to refine the global context features, which can be seen as a type of channel attention. A skip connection is adopted at the end of the global context modeling step to aggregate the global context information with $\mathbf{X}_l$. The aggregated result is sent to the remaining components of vanilla Transformer and produces the final fused feature $\mathbf{E}_l \in \mathbb{R}^{C\times H\times W}$.

While being embarrassingly simple, CST significantly speeds up the baseline inference (36 FPS *v.s.* 3 FPS) with negligible performance drop when replacing vanilla Transformers with CST for the first two fusion stages.

### 3.3.2 Semantic Gathering-Scattering Transformer

Semantic Gathering-Scattering Transformer (SGST) is to explicitly model the foreground/background semantic dependencies while reducing the attention redundancy. As shown in Fig. 3 (b), SGST consists of two parallel branches

to separately process foreground and background, mainly including semantic query gathering, key/value soft merging, dependencies calculation with a standard Transformer step, and semantic token scattering. As the two branches share a similar process, we simplify the description by illustrating only the foreground branch.

**Semantic Query Gathering.** Given a mixing representation $\mathbf{X}_l \in \mathbb{R}^{C\times H\times W}$ from one high-level stage $l(l \in \{3,4\})$, a one-channel token gathering heatmap $\mathbf{H}_l^g \in \mathbb{R}^{H\times W}$ is firstly generated using a $1\times 1$ convolution layer followed by a Sigmoid function. Upon the heatmap, the foreground tokens are identified via $\mathbf{X}_l[h_i >= 0.5]$, and gathered to form a list of foreground queries $\mathbf{Q}_l^f$. $h_i$ denotes the heatmap value at position $i$, and 0.5 is the threshold to distinguish foreground from background[1].

**Key/Value Soft Merging.** We begin by calculating the dot product between $\mathbf{X}_l$ and $\mathbf{H}_l^g$[2], producing a heatmap enhanced foreground token sequence $\mathbf{X}_l^e \in \mathbb{R}^{C\times N}$, where $N = H \times W$. To adaptively mine the representative information and remove redundancy, we merge the $N$ tokens of $\mathbf{X}_l^e$ into $K$ tokens $\mathbf{X}_l^c \in \mathbb{R}^{C\times K}$ ($K << N$) in a soft way, which is achieved using a learnable transformation matrix $\mathbf{W}_l^m \in \mathbb{R}^{N\times K}$ applied on $\mathbf{X}_l^e$. Then the compact foreground key and value sequences $\mathbf{K}_l^f, \mathbf{V}_l^f \in \mathbb{R}^{C\times K}$ are generated with linear transformations of $\mathbf{X}_l^c$.

**Dependencies Calculation.** With the compressed foreground query $\mathbf{Q}_l^f$, key $\mathbf{K}_l^f$, and value $\mathbf{K}_l^f$, a standard Transformer block is adopted to model the semantic dependencies and update the corresponding foreground representation tokens, where the attention computation can be largely reduced at the same time.

**Semantic Token Scattering.** With the original mixing representation $\mathbf{X}_l$, the updated foreground and background tokens are scattered back according to the indexes in the query gathering step and obtain the final fused feature $\mathbf{S}_l$.

In our experiments, $K$ is set to $\frac{1}{9}N$ in the key/value soft merging step, reducing $87\%$ of computation of MHSA in total compared to vanilla Transformer. SGST's unique ability of explicitly modeling semantic dependencies through foreground-background query separation, along with its efficient token merging mechanism, allows it to dramatically reduce computational complexity while maintaining top-notch performance.

#### 3.3.3 Isomerous Framework

With the two proposed Transformer variants, we take the vanilla Transformer baseline one step further. As shown in Fig. 2, we replace vanilla Transformers in the first two stages with our Context Sharing Transformer (CST) to model the global-shared contextual information within image frames, while in the last two stages with our Semantic Gathering-Scattering Transformer (SGST) to models the semantic correlation explicitly. Consequently, it formulates a level-isomerous Transformer (Isomer), which treats the different fusion levels distinctively based on the observed properties of Transformers to better fit the ZVOS task.

[1]The background queries $\mathbf{Q}_l^b$ are obtained using $\mathbf{X}_l[h_i < 0.5]$.

[2]The background branch adoptes $1 - \mathbf{H}_l^g$.

### 3.4. Implement Details

We use Swin-Tiny [28] as our backbone when reporting the final results for fair comparison, which has a comparable model size with ResNet50[14]. Other backbones are also compared in our experiments. Following [39, 59], the well-trained RAFT [46] is adopted to generate the optical flow maps for the video data. All the input images are resized to a spatial resolution of $512 \times 512$. Data augmentation including horizontal flipping and photometric distortion is adopted during training. We utilize a subset of the Youtube-VOS [56] training set (1 frame per every 30 frames sampled) to pre-train the network based on [72, 49, 32], followed by network fine-tuning with DAVIS-16 [41] and FBMS [34] training sets. The AdamW [29] optimizer is adopted with a fixed learning rate of 6e-5 throughout the training process. Our network is end-to-end trained on one NVIDIA 3090 GPU with a mini-batch size of 8, using binary cross-entropy loss for supervision.

## 4. EXPERIMENT

### 4.1. Datasets and Evaluation Metrics

**Datasets.** We perform evaluation on three widely adopted ZVOS datasets: DAVIS-16 [41], FBMS [34], and Long-Videos [26]. DAVIS-16 is one of the most popular ZVOS benchmark datasets containing 50 high-quality video sequences (30 for training and 20 for validation). FBMS comprises 20 training video sequences and 30 test sequences. Long-Videos comprises three long-term videos, each of which has about 2500 frames. In addition, we also conduct experiments for the video salient object detection (VSOD) task to comprehensively evaluate our method using three datasets, including DAVIS-16, FBMS, and MCL [18].

**Evaluation metrics.** We report the standard evaluation metrics for ZVOS task, including mean of region similarity ($\mathcal{J}$ Mean), mean of contour accuracy ($\mathcal{F}$ Mean) [41], and $\mathcal{J}\&\mathcal{F}$ that is computed by averaging $\mathcal{J}$ Mean and $\mathcal{F}$ Mean. For VSOD task, we adopt four widely used metrics: MAE [40] ($\mathcal{M}$), maximum E-measure ($\mathcal{E}_\xi^{max}$) [9], maximum F-measure ($\mathcal{F}_\beta^{max}$, $\beta^2 = 0.3$) [1], and S-measure ($\mathcal{S}_\alpha$, $\alpha = 0.5$) [8]. Please note that the predicted maps are binarized with a threshold of 0.5 for ZVOS evaluation, but not for VSOD evaluation following [17, 64, 22].

Table 1. ZVOS performance on DAVIS-16 validation set. "CRF" means that conditional random field [19] is applied as post-processing. The inference speed is tested on one 3090 GPU. The best and second-best scores are indicated in red and blue, respectively.

| Method | Publication | Backbone | CRF | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ | $\mathcal{J}\&\mathcal{F}$ ↑ | FPS ↑ |
|---|---|---|---|---|---|---|---|
| PDB[43] | ECCV2018 | ResNet-50 | ✓ | 77.2 | 74.5 | 75.9 | 20.0 |
| AGS[51] | CVPR2019 | ResNet-101 | ✓ | 79.7 | 77.4 | 78.6 | 1.7 |
| AGNN[50] | ICCV2019 | ResNet-101 | ✓ | 80.7 | 79.1 | 79.9 | 1.9 |
| COSNet[31] | CVPR2019 | ResNet-101 | ✓ | 80.5 | 79.5 | 80.0 | 2.2 |
| AnDiff[61] | CVPR2019 | ResNet-101 | | 81.7 | 80.5 | 81.1 | 2.8 |
| MATNet[72] | AAAI2020 | ResNet-101 | ✓ | 82.4 | 80.7 | 81.5 | 1.3 |
| GraphMem[30] | ECCV2020 | ResNet-50 | ✓ | 82.5 | 81.2 | 81.9 | 5.0 |
| DFNet[70] | ECCV2020 | ResNet-101 | ✓ | 83.4 | 81.8 | 82.6 | 3.6 |
| F2Net[27] | AAAI2021 | ResNet-101 | | 83.1 | 84.4 | 83.7 | 10.0 |
| FSNet[17] | ICCV2021 | ResNet-50 | ✓ | 83.4 | 83.1 | 83.3 | 12.5 |
| AMCNet[59] | ICCV2021 | ResNet-101 | ✓ | 84.5 | 84.6 | 84.6 | 17.5 |
| TransportNet[63] | ICCV2021 | ResNet-101 | ✓ | 84.5 | 85.0 | 84.8 | 3.6 |
| RTNet[42] | CVPR2021 | ResNet-101 | ✓ | 85.6 | 84.7 | 85.2 | - |
| EFS[20] | AAAI2022 | ResNet-50 | | 84.5 | 86.7 | 85.6 | 2.0 |
| HFAN[39] | ECCV2022 | Swin-Tiny | | 86.0 | 87.3 | 86.7 | 26.7 |
| Ours | - | Swin-Tiny | | 88.8 | 91.1 | 90.0 | 24.6 |

Table 2. ZVOS performance on FBMS validation set.

| Method | OBN [24] | PDB [43] | COSNet [31] | MATNet [72] | AMCNet [59] | APS [67] | F2Net [27] | EFS [20] | TransportNet [63] | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{J}$ Mean ↑ | 73.9 | 74 | 75.6 | 76.1 | 76.5 | 76.7 | 77.5 | 77.5 | 78.7 | 87.6 |

Table 3. ZVOS performance on Long Videos dataset.

| Method | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ | $\mathcal{J}\&\mathcal{F}$ ↑ |
|---|---|---|---|
| 3DCSeg[32] | 34.2 | 33.1 | 33.7 |
| MATNet[72] | 66.4 | 69.3 | 67.9 |
| AGNN[50] | 68.3 | 68.6 | 68.5 |
| HFAN[39] | 80.2 | 83.2 | 81.7 |
| Ours | 81.4 | 84.9 | 83.2 |

### 4.2. Comparison with State-of-the-art

**Evaluation on ZVOS.** Tab. 1, 2, and 3 show the overall ZVOS performance on DAVIS-16, FBMS, Long-Videos dataset, respectively. We also report the inference speed of all the compared methods tested using a 3090 GPU. The proposed method Isomer consistently outperforms the compared methods on all datasets with nearly real-time inference (24.6 FPS). Specifically, on DAVIS-16 dataset, compared with the current leading method HFAN [39], Isomer achieves a significant improvement of 2.9% in terms of $\mathcal{J}\&\mathcal{F}$ with a comparable inference speed. Besides, Isomer reaches an improvement of 9.7% in terms of $\mathcal{J}$ compared with TransportNet [63] on FBMS dataset.

**Evaluation on VSOD.** Tab. 4 provides the quantitative comparison for VSOD task, showing that Isomer achieves the best results across all evaluation metrics on the three datasets. Compared with the second best methods, Isomer outperforms FSNet [17] by 5.3% and 3.2% in terms of $\mathcal{F}_\beta^{max}$ and $\mathcal{S}_\alpha$ on DAVIS-16 dataset, and exceeds MGA [22] by 2.9% and 3.7% on FBMS dataset. While MCL dataset has blurry boundaries in the low-resolution frames, Isomer surpasses FSNet [17] by 4.7% and 7.4% for $\mathcal{E}_\xi^{max}$ and $\mathcal{F}_\beta^{max}$, respectively. These results indicate that our method generalizes well across both ZVOS and VSOD tasks.

### 4.3. Ablation Study

To analyze the impact of our key components, we conduct several ablation studies on the DAVIS-16 validation set with ZVOS evaluation metrics. To conserve computing resources, we use a lightweight backbone MiT-b0 [55] for the following experiments unless otherwise stated.

**Effectiveness of CST, SGST, and Level-Isomerous Scheme.** We begin by studying the importance of leveraging Transformers for effective feature fusion. Tab. 5 shows that a basic vanilla Transformer (VT) implementation can surpass leading CNN-based fusion modules [22, 72, 39] with the same backbone (MiT-b0) and experimental setup. However, the VT baseline incurs substantial computation and can only achieve a speed of 3 FPS. Then, we replace the VT blocks with our CST in the low-level (*i.e.* the first two) feature fusion stages, and observe a significant speed improvement (36 FPS) with almost no performance drop. Based on this, we then replace VT with our SGST in the high-level (*i.e.* the last two) feature fusion stages and obtain further improvements in both speed and performance. To further investigate the superiority of the proposed level-isomerous framework, we also apply CST

Table 4. Overall VSOD performance on three benchmark datasets. The best and second-best scores are indicated in red and blue, respectively.

| Methods | DAVIS-16[41] | | | | FBMS[34] | | | | MCL[18] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}\downarrow$ | $\mathcal{E}_\xi^{max}\uparrow$ | $\mathcal{F}_\beta^{max}\uparrow$ | $\mathcal{S}_\alpha\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{E}_\xi^{max}\uparrow$ | $\mathcal{F}_\beta^{max}\uparrow$ | $\mathcal{S}_\alpha\uparrow$ | $\mathcal{M}\downarrow$ | $\mathcal{E}_\xi^{max}\uparrow$ | $\mathcal{F}_\beta^{max}\uparrow$ | $\mathcal{S}_\alpha\uparrow$ |
| RCR[57] | 0.027 | 0.947 | 0.848 | 0.886 | 0.053 | 0.905 | 0.859 | 0.872 | 0.028 | 0.895 | 0.742 | 0.864 |
| SSAV[10] | 0.028 | 0.948 | 0.861 | 0.893 | 0.040 | 0.926 | 0.865 | 0.879 | 0.026 | 0.889 | 0.773 | 0.819 |
| MGA[22] | 0.022 | 0.961 | 0.902 | 0.913 | 0.027 | 0.949 | 0.910 | 0.907 | 0.031 | 0.901 | 0.798 | 0.845 |
| PCSA[13] | 0.022 | 0.961 | 0.880 | 0.902 | 0.041 | 0.914 | 0.831 | 0.866 | N/A | N/A | N/A | N/A |
| DCFNet[64] | 0.016 | 0.969 | 0.900 | 0.914 | 0.037 | 0.916 | 0.849 | 0.877 | 0.029 | 0.875 | 0.716 | 0.762 |
| FSNet[17] | 0.020 | 0.970 | 0.902 | 0.920 | 0.041 | 0.935 | 0.888 | 0.890 | 0.023 | 0.924 | 0.821 | 0.864 |
| Ours | 0.010 | 0.987 | 0.946 | 0.950 | 0.019 | 0.974 | 0.944 | 0.934 | 0.015 | 0.967 | 0.882 | 0.893 |

Table 5. Ablation study for the proposed CST and SGST blocks, and our level-isomerous fusion scheme. **Bold** font indicates the best trade off between accuracy and speed.

| Low Level | High Level | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ | FPS ↑ |
|---|---|---|---|---|
| MGA[22] | MGA[22] | 79.7 | 80.7 | 27 |
| MAT[72] | MAT[72] | 80.0 | 80.9 | 16 |
| HFAN[39] | HFAN[39] | 81.5 | 80.8 | 42 |
| VT | VT | 84.2 | 85.4 | 3 |
| CST | VT | 84.2 | 85.2 | 36 |
| **CST** | **SGST** | **84.6** | **85.6** | **39** |
| CST | CST | 82.9 | 83.6 | 44 |
| SGST | SGST | 84.5 | 85.8 | 29 |

Table 6. Performance comparison with lightweight Transformers.

| Method | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ | FPS ↑ |
|---|---|---|---|
| AxialNet[15] | 83.6 | 84.4 | 32 |
| PoolFormer[62] | 82.0 | 83.8 | 41 |
| EdgeViT[35] | 84.1 | 83.9 | 36 |
| PVT V2[53] | 83.0 | 84.6 | 30 |
| PVT[52] | 81.0 | 83.6 | 28 |
| **Ours** | **84.6** | **85.6** | **39** |

Table 7. Performance comparison with the state-of-the-art method HFAN[39] on different backbones.

| Backbone | $\mathcal{J}\&\mathcal{F}\uparrow$ (HFAN[39]) | $\mathcal{J}\&\mathcal{F}\uparrow$ (Ours) |
|---|---|---|
| MiT-b0 | 81.2 | **85.1** |
| ResNet101 | 87.0 | **87.5** |
| Swin-Tiny | 86.7 | **89.2** |

or SGST identically for all the stages. It shows that the proposed level-isomerous Transformer framework can achieve the best trade-off between speed and accuracy.

**Comparison with Existing Lightweight Transformers.** We also conduct experiments using current general lightweight Transformers as the fusion modules to further verify the superiority of our CST and SGST blocks for ZVOS task. Results are reported in Tab. 6. Compared with the VT block (fourth row in Tab. 5), these lightweight Transformers all achieve notable acceleration while sacrificing accuracy. In contrast, Our method is designed based on Transformers' behavior under ZVOS settings to learn task-specific attention in a flexible data-driven manner, and provides both improved accuracy and efficiency.

Table 8. Effectiveness of foreground-background separate modeling in SGST. "S" denotes foreground and background separation. "F" denotes foreground.

| Method | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ |
|---|---|---|
| Ours (w/o S) | 84.3 | 85.4 |
| only F | 81.8 | 82.2 |
| **Ours** | **84.6** | **85.6** |

Table 9. Ablations on the ratio of key/value soft merging in SGST.

| $K/N$ | $\mathcal{J}$ Mean ↑ | $\mathcal{F}$ Mean ↑ |
|---|---|---|
| 1 | 83.6 | 84.1 |
| 4/9 | 84.1 | 85.2 |
| 1/4 | 84.1 | 85.4 |
| **1/9** | **84.6** | **85.6** |
| 1/36 | 83.7 | 84.4 |

**Performance with Different Backbones.** We adopt different backbones to verify the generality of the proposed feature fusion method. Tab. 7 shows that our method consistency outperforms the recent leading method HFAN [39], and the improvement is particularly prominent for the lightweight backbones MiT-b0 [55] and Swin-Tiny [28].

**Foreground-background Separate Modeling in SGST.** In Tab. 8, we ablate our method by removing the foreground-background separation in SGST (top row), which calculates attention by indiscriminately mixing the foreground and background together as vanilla Transformer. Results verify the effectiveness of the main concept of SGST that modeling the semantic dependencies explicitly for the foreground and background. Furthermore, we explore the impact of modeling only the foreground dependencies, which aligns with our final goal. However, as evidenced by the second row of Tab. 8, a significant decrease in performance is observed, suggesting that both foreground and background are crucial in aiding the comprehension of semantics.

**Merging Ratio in Soft Token Selection.** SGST uses token soft merging to eliminate foreground/background redundancy. We study the impact of different merging ratios (*i.e.* $K/N$) on performance. Tab. 9 shows that $1/9$ merging ratio obtains superior performance with $87\%$ reduction of computational cost compared with vanilla Transformer.

**Visualization.** Fig. 4 shows some qualitative results of Iso-

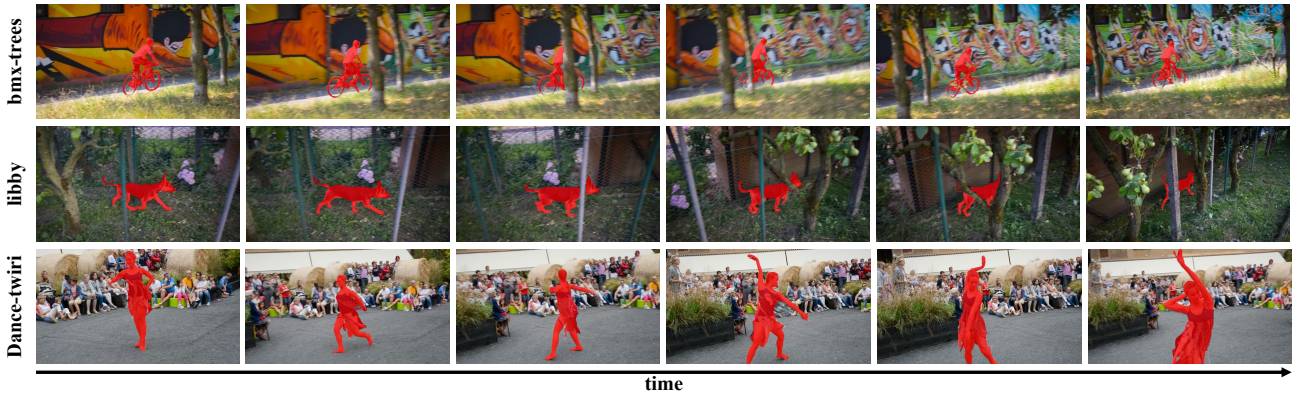


Figure 4. Qualitative results on three challenging video clips from DAVIS-16[41].

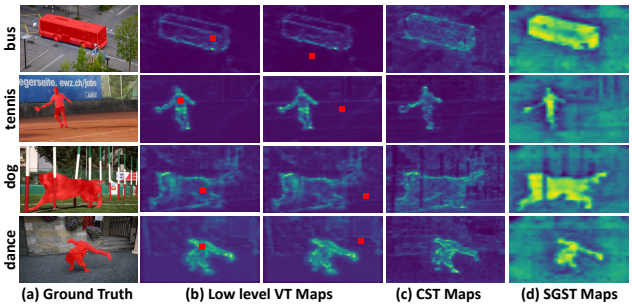


Figure 5. Illustration of the attention maps from VT and CST and token gathering heatmaps from SGST.

mer, showing its promising prediction ability across various challenging situations. Fig. 5 visualizes several attention maps (heatmaps) of VT (b) and CST (c), showing that CST captures mostly similar query-independent dependency as VT but with less computational effort. Fig. 5(d) shows the token gathering heatmaps learned in SGST, which effectively distinguishes between foreground and background to help the network explicitly model semantic dependencies.

**Limitation and Future Work.** This work leverages the power of Transformers to integrate dependable appearance-motion information for ZVOS. However, our method is still limited by the following factors: (1) Since videos often contain multiple moving objects, optical flow maps tend to have a lot of noise. Therefore, directly fusing optical flow information at the pixel level may not be the best approach for VOS. Instead, a promising solution could be a simple coarse-to-fine pipeline. This involves using a detection network to roughly locate the target area (such as a bounding box) through the optical flow map, and then using an image segmentation network to segment the foreground and background in the detected region. This pipeline not only helps locate moving objects with the aid of optical flow maps, but also avoids the influence of optical flow noise during segmentation, resulting in more accurate segmentation results. (2) As the generation of optical flow maps requires additional optical flow networks and time, there may be room for optimization in the strategy of using optical flow. For instance, it may be possible to design a better flow-based training method that retains some of the performance improvements brought by optical flow without requiring input of optical flow maps during the test phase. This would save the time needed for optical flow generation and improve the efficiency of the inference process. (3) The current ZVOS datasets lack uniformity in the definition and labeling rules of video segmentation targets, which can cause models that perform well on some datasets to perform poorly on others. This brings uncertainty and difficulty in selecting datasets for model training and validation. Therefore, ZVOS urgently requires the proposal of a new dataset with clearly defined video segmentation targets. We intend to explore these avenues in our future work.

## 5. Conclusion

This paper proposes a new ZVOS framework named Isomer (level-**iso**merous Transfor**mer**), which treats the different feature fusion levels distinctively. We develop two core components Context-Sharing Transformer (CST) and Semantic Gathering-Scattering Transformer (SGST) to model the contextual dependencies from different levels effectively and efficiently. In early fusion stages, CST models the global-shared contextual information by computing one query-independent dependence map for all query tokens. In late fusion stages, SGST captures long-range semantic-specific dependency by computing the foreground and background attentions separately with a soft token merging mechanism. Experimental results show that the proposed Isomer achieves state-of-the-art performance in both ZVOS and VSOD tasks with real-time inference. To our best knowledge, this work is the first successful application of Transformers in ZVOS task, which could provide a new paradigm for the dense prediction vision tasks in exploring Transformer based architecture.

## References


[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 5

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 1, 3, 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1, 3

[4] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022. 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[7] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *CVPR*, pages 5912–5921, 2021. 3

[8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 5

[9] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 6:6, 2021. 5

[10] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 2, 3, 7

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 3

[12] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 3

[13] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, pages 10869–10876, 2020. 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5

[15] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 7

[16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 2

[17] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, pages 4922–4933, 2021. 1, 2, 3, 5, 6, 7

[18] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 5, 7

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 6

[20] Youngjo Lee, Hongje Seong, and Euntai Kim. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1245–1253, 2022. 6

[21] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, pages 3243–3252, 2018. 3

[22] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 3, 5, 6, 7

[23] Jiangtong Li, Wentao Wang, Junjie Chen, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Video semantic segmentation via sparse temporal transformer. In *ACM MM*, pages 59–68, 2021. 3

[24] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 207–223, 2018. 6

[25] Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.01191*, 2022. 3

[26] 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. 5

[27] Daizong Liu, Dongdong Yu, Changhu Wang, and Pan Zhou. F2net: Learning to focus on the foreground for unsupervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2109–2117, 2021. 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1, 2, 3, 5, 7

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *European Conference on Computer Vision*, pages 661–679. Springer, 2020. 6

[31] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, pages 3623–3632, 2019. 2, 6

[32] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516*, 2020. 5, 6

[33] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 3

[34] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1187–1200, 2013. 5, 7

[35] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 294–311. Springer, 2022. 3, 7

[36] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *arXiv preprint arXiv:2205.13213*, 2022. 3

[37] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pages 235–252, 2020. 3

[38] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE TIP*, 2023. 3

[39] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022. 1, 2, 3, 5, 6, 7

[40] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 5

[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 *CVPR*, pages 724–732, 2016. 2, 5, 7, 8

[42] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15455–15464, 2021. 1, 6

[43] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018. 3, 6

[44] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021. 3

[45] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 2

[46] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 2, 5

[47] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4490, 2017. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 5998–6008, 2017. 1, 3

[49] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2019. 5

[50] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, pages 9236–9245, 2019. 2, 6

[51] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, pages 3064–3074, 2019. 2, 6

[52] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 3, 7

[53] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3, 7

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021. 1, 3, 6, 7

[56] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 2, 5

[57] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video

salient object detection using pseudo-labels. In *ICCV*, pages 7284–7293, 2019. 7

[58] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11998–12008, 2022. 3

[59] Shu Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang. Learning motion-appearance co-attention for zero-shot video object segmentation. In *ICCV*, pages 1564–1573, 2021. 1, 2, 3, 5, 6

[60] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 3

[61] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, pages 931–940, 2019. 2, 6

[62] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 3, 7

[63] Kaihua Zhang, Zicheng Zhao, Dong Liu, Qingshan Liu, and Bo Liu. Deep transport network for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8781–8790, 2021. 6

[64] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563, 2021. 3, 5, 7

[65] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022. 3

[66] Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *arXiv preprint arXiv:2303.10383*, 2023. 2

[67] Xiaoqi Zhao, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *ACM MM*, pages 2645–2653, 2021. 6

[68] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020. 3

[69] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *arXiv preprint arXiv:2303.10396*, 2023. 3

[70] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. Learning discriminative feature with crf for unsupervised video object segmentation. In *ECCV*, pages 445–462, 2020. 6

[71] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 3

[72] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, pages 13066–13073, 2020. 2, 5, 6, 7

[73] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3