BATMAN: Bilateral Attention Transformer in Motion-Appearance Neighboring Space for Video Object Segmentation

Ye Yu[1], Jialin Yuan[2], Gaurav Mittal[1], Li Fuxin[2], and Mei Chen[1]

[1] Microsoft
{yu.ye,gaurav.mittal,mei.chen}@microsoft.com
[2] Oregon State University
{yuanjial,lif}@oregonstate.edu

Abstract. Video Object Segmentation (VOS) is fundamental to video understanding. Transformer-based methods show significant performance improvement on semi-supervised VOS. However, existing work faces challenges segmenting visually similar objects in close proximity of each other. In this paper, we propose a novel Bilateral Attention Transformer in Motion-Appearance Neighboring space (BATMAN) for semi-supervised VOS. It captures object motion in the video via a novel optical flow calibration module that fuses the segmentation mask with optical flow estimation to improve within-object optical flow smoothness and reduce noise at object boundaries. This calibrated optical flow is then employed in our novel bilateral attention, which computes the correspondence between the query and reference frames in the neighboring bilateral space considering both motion and appearance. Extensive experiments validate the effectiveness of BATMAN architecture by outperforming all existing state-of-the-art on all four popular VOS benchmarks: Youtube-VOS 2019 (85.0%), Youtube-VOS 2018 (85.3%), DAVIS 2017Val/Test-dev (86.2%/82.2%), and DAVIS 2016 (92.5%).

Keywords: Bilateral attention, Motion-appearance space, Optical flow calibration, Video object segmentation, Vision transformer

1 Introduction

Video Object Segmentation (VOS) is fundamental to video understanding with broad applications in content creation, content moderation, and autonomous driving. In this paper, we focus on the semi-supervised VOS task, where we segment target objects in each frame of the entire video sequence (query frames) given their segmentation masks in the first frame (reference frame) only. Moreover, the task is class-agnostic in that we do not have any class annotation for any object to be segmented in either training or testing phases. The key challenge

[2] At Oregon State University, Jialin Yuan and Li Fuxin are supported in part by NSF grant 1911232.

in semi-supervised VOS is how to propagate the mask from the reference frame to all the query frames in the rest of the sequence without any class annotation.

Due to the absence of class-specific features, VOS models need to match features of the reference frame to that of the query frames both spatially and temporally to capture the class-agnostic correspondence and propagate the segmentation masks. Previous methods attempt to store features from preceding frames in memory networks and match the query frame through a non-local attention mechanism [27,7], or compute a global-to-global attention through an encoder-decoder transformer [25], or propagate and calibrate features from the reference frame to the query frames using a propagation-correction scheme [47]. These methods employ a global attention mechanism to establish correspondence between the full reference frame and the full query frame. This can lead to failure in distinguishing the target object(s) from the background particularly when there are multiple objects with a similar visual appearance. A spatial local attention is proposed in [50] to mitigate this problem, where the attention is only computed between each query token and its surrounding key tokens within a spatial local window. However, it still suffers from incorrectly segmenting visually similar objects in close proximity of each other.

In addition to spatial correspondence, it is essential to match features temporally for optimal object segmentation across video frames. To this end, some VOS methods [45,8] leverage optical flow to capture object motion. [45] warps the memory frame mask using optical flow before performing local matching between memory and query features based on the warped mask, while [8] simultaneously trains the model for object segmentation and optical flow estimation by bidirectionally fusing feature maps from the two branches. However, these methods are not able to perform optimally as optical flow is usually noisy and warping features/masks to match objects across frames accumulates errors in both optical flow and segmentation mask along the video sequence.

To overcome the above challenges, we propose Bilateral Attention Transformer in Motion-Appearance Neighboring space (BATMAN). BATMAN introduces a novel bilateral attention module that computes the local attention map between the query frame and memory frames with both motion and appearance in consideration. Unlike the conventional spatial local attention mechanism (Fig. 1(a)) that computes the attention within a predefined fixed local window, our bilateral attention adaptively computes the local attention based on the tokens' spatial distance, appearance similarity, and optical flow smoothness, as shown in Fig. 1(b). Observing that optical flow may be especially noisy for fast-moving object(s), BATMAN introduces a novel optical flow calibration module that leverages the mask information from the memory frame to smooth the optical flow within the same object while reducing noise at the object boundary.

We conduct extensive experiments on four popular VOS benchmarks: Youtube-VOS 2019 [46], Youtube-VOS 2018 [46], DAVIS 2017 [32], and DAVIS 2016 [30] to validate the BATMAN architecture. We show that BATMAN achieves superior performance on all benchmarks and outperforms all previous state-of-the-art methods. We summarize the main contributions of our work below,

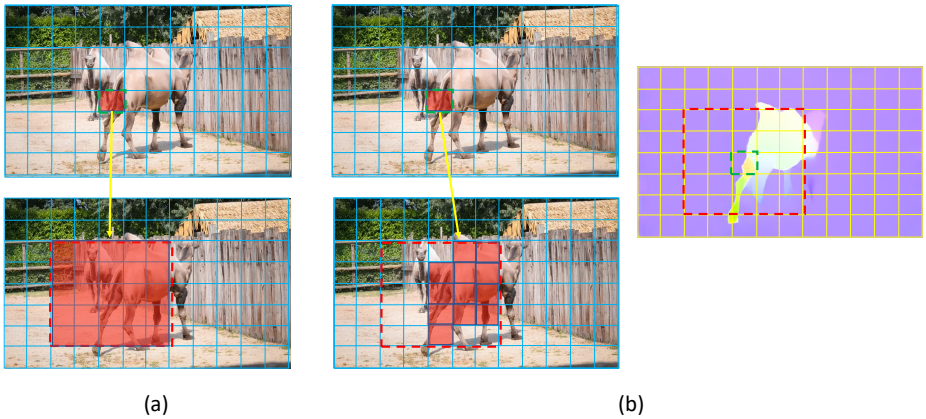


Fig. 1: Spatial local attention vs. bilateral attention. (a) Conventional spatial local attention. For any given token in the query frame (top), compute the attention with the neighboring tokens within a predefined fixed local window from the memory frame (bottom). (b) Our proposed bilateral attention. Given a token in the query frame (top), adaptively select the most relevant tokens (bottom), based on the distance in the bilateral space of appearance and motion (right), for cross attention computation

- A novel bilateral attention module that computes attention between query and memory features in the bilateral space of motion and appearance, which improves the correspondence matching by adaptively focusing on relevant object features while reducing the noise from the background.
- A novel optical flow calibration module that fuses the object segmentation mask and the initial optical flow estimation to smooth the within-object optical flow and reduce noise at the object boundary.
- Incorporating the optical flow calibration and bilateral attention mechanisms, we design a novel BATMAN architecture. BATMAN establishes new state-of-the-art performance on Youtube-VOS 2019 / 2018 and DAVIS 2017 / 2016 benchmarks. To the best of our knowledge, BATMAN is the first work to compute attention in the bilateral space of motion and appearance for VOS.

2 Related work

Semi-supervised VOS. The task aims to segment the particular object instances throughout the entire video sequence given one or more annotated frames (the first frame in general). Early DNN works [3,29,44] fine-tune the pre-trained networks on the first frame using multiple data augmentations on the given mask at test time to adapt to specific instances. Therefore, these methods are extremely slow during inference due to excessive fine-tuning. Later tracking-based works [40,17,5] adopt object tracking technologies to indicate the target

location of objects for segmentation to improve inference time. However, these approaches are not robust to occlusion and drifting with error accumulated during the propagation. "Tracking-by-detection" paradigm is introduced into VOS in [16] to take object segmentation as a subtask of tracking, in which the accuracy of tracking often limits the performance. To handle occlusion and drifting, matching-based methods [6,39] perform feature matching to find objects that are similar to the target objects in the reference frames. STM [27] and its following works [34,45] leverage an external memory to store past frames' features and then distinguish objects with a similar appearance by pixel-level attention-based matching from the memory.

Vision Transformer. Initially proposed for machine translation, Transformers [38] replace the recurrence and convolutions entirely with hierarchical attention-based mechanisms and achieve outstanding performance. Later, transformer networks became dominant models used in natural language processing (NLP) tasks [42,52]. Recently, with the observance of its strength in parallel modeling global correlation or attention, transformer blocks were introduced to computer vision tasks, such as image recognition [10], saliency prediction [53], object detection [54,4], and object segmentation [41], where vision transformers have achieved excellent performance compared to the CNN-based counterparts. Researchers then employed transformer architecture into the VOS task [11,23,25,50]. SST [11] adopts the transformer's encoder to compute attention based on the spatial-temporal information among multiple history frames. In [23], a transductive branch is used to capture the spatial-temporal information, which is integrated with an online inductive branch within a unified framework. TransVOS [25] introduces a transformer-based VOS framework with intuitive structure from the transformer networks in NLP. AOT [50] proposes an Identification Embedding to construct multi-object matching and computes attention for multiple objects simultaneously. In this paper, we introduce a novel bilateral attention transformer framework, where it computes the attention with both the encoded appearance features and the motion features in consideration. Therefore, it is robust to occlusion, drift, and ambiguity between objects with a similar appearance.

Optical Flow. Applying optical flow to VOS can encourage motion consistency through the entire video sequence. Early approaches [37,48,8] consider VOS and optical flow estimation simultaneously with the assumption that the two tasks are complementary. Recently, RMNet [45] introduces using optical flow generated with an offline model to warp object mask from the previous frame to the query frame and then performing regional matching. It avoids unnecessary matching in regions without target objects or mismatching of objects with a similar appearance. Instead of simply warping the object's mask to indicate the target area, our BATMAN computes the correlation of each pair of tokens considering their optical flow estimation, appearance similarity, and spatial distance simultaneously. Thus, it is more effective in removing irrelevant matching tokens

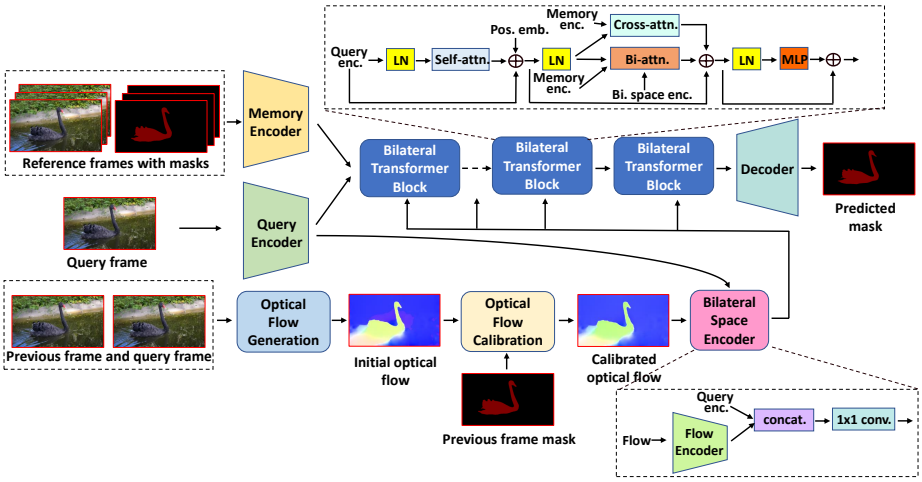


Fig. 2: Overview of the BATMAN architecture. Frame-level features of the reference frames and the query frame are extracted through the memory and query encoders, respectively. A pre-trained FlowNet is used to generate an initial optical flow estimation between the previous frame and the query frame, which is then improved by the optical flow calibration module. A bilateral space encoder is used to encode the query features and the calibrated optical flow into a bilateral space encoding, which is used by the bilateral attention. Multiple layers of bilateral transformer blocks are stacked for matching the correspondence between the reference and query features. Lastly, a decoder is used to predict the query frame segmentation mask

compared to [45]. Meanwhile, our method is more robust to the accumulated error in warping from the optical flow estimation.

3 Method

In this section, we first introduce the proposed BATMAN architecture, and then discuss in depth its core modules: bilateral attention and optical flow calibration.

3.1 Bilateral Attention Transformer in Motion-Appearance Neighboring space (BATMAN)

Fig. 2 provides an overview of the proposed BATMAN architecture. We first extract frame-level features through the memory and query encoders (details in Sec. 4.1) to capture the target object features for establishing correspondence in the later transformer layers. Meanwhile, we compute the initial optical flow between the query frame and its previous frame through a frozen pre-trained FlowNet [36]. Then, we feed the object mask from the previous frame, together

with the initial optical flow estimation, into our optical flow calibration module to improve the optical flow (Sec. 3.3). We then encode the calibrated optical flow and the query frame features into tokens in the bilateral space of motion and appearance. Following this, we stack multiple bilateral transformer blocks to model the spatial-temporal relationships among the reference and query frames at pixel-level, based on the bilateral space encoding tokens (Sec. 3.2). After aggregating the spatial-temporal information, the decoder predicts an object mask for the query frame.

3.2 Bilateral transformer and bilateral attention

As shown in Fig. 2, in each bilateral transformer block, the query frame features first go through a self-attention [38] to aggregate the information within the query frame followed by adding a sinusoidal position embedding [38] encoding the tokens' relative positions. Then we apply cross-attention and bilateral attention (described below) to it with the reference frame features and add the results. Following the common practice in vision transformers [50,25], we insert layer normalization [1] before and after each attention module. Finally, we employ a two-layer feed-forward MLP block before feeding the output to the next layer.

Bilateral space encoding (E) is used to index each position (token) of the query frame features in the bilateral space. As shown in Fig. 2, we first encode the calibrated optical flow using a flow encoder (details in Sec. 4.1). Then we concatenate the optical flow encoding and the query image encoding (from query encoder) in channel dimension. Finally, we use a 1×1 convolutional layer to project the concatenation to a 1-dimensional space (in channel) where each position (token) has a single scalar coordinate for the bilateral space of motion and appearance. Bilateral space encoding is employed in bilateral attention below.

Bilateral attention is used to aggregate spatial-temporal information between the query tokens and neighboring key tokens from the reference frames in the bilateral space of motion and appearance. Unlike global cross-attention where each query token computes attention with all key tokens from the reference frames, our bilateral attention adaptively selects the most relevant key tokens for each query token based on the bilateral space encoding. To formulate, we define query tokens $Q \in \mathbb{R}^{HW \times C}$, key tokens $K \in \mathbb{R}^{HW \times C}$, and value embedding tokens $V \in \mathbb{R}^{HW \times C}$, where Q is from the query frame and K and V are aggregated from multiple reference frames. H, W, and C represent the height, width, and channel dimensions of the tokens, respectively. Mathematically, we define bilateral attention as,

$$BiAttn(Q, K, V) = softmax(\frac{QK^T M}{\sqrt{C}})V \tag{1}$$

where $M \in [0,1]^{HW \times HW}$ is the bilateral space binary mask that defines the attention scope for each query token. For each query token $Q_{h,w}$ at (h,w) position, we define the corresponding bilateral space binary mask $M_{h,w}$ as,

$$M_{h,w}(i,j,E) = \begin{cases} 1 & \text{if } |i-h| \leqslant W_d \text{ and } |j-w| \leqslant W_d \\ & \text{and } |argsort_{W_d}(E_{h,w}) - argsort_{W_d}(E_{i,j})| \leqslant W_b \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where (i, j) is the position for each key token, $E \in \mathbb{R}^{HW \times 1}$ is the bilateral space encoding of the queries discussed above, W_d and W_b are predefined local windows in spatial and bilateral domains, respectively. $argsort_{W_d}(E_{i,j})$ denotes sorting all bilateral space encoding E within the spatial local window W_d and finding the corresponding index at position (i,j). To train the bilateral space encoding E by stochastic gradient descent directly, in practice, instead of computing QK^TM as shown in Eq. 1, we compute $QK^T + E$ if $M = 1$, while computing $QK^T - L$ if $M = 0$, where $L \in \mathbb{R}$ is a large positive number. This approximates to QK^TM in Eq. 1 after using $softmax$. Eq. 2 shows that for each query token, it computes the attention with another key token only if they are close to each other spatially and share similar bilateral space encoding (similar motion and appearance). We further analyze the bilateral space binary mask with visualization in Sec. 4.3.

We implement the bilateral attention modules via a multi-headed formulation [38] where we linearly project queries, keys, and values multiple times with different learnable projections, and we feedforward the multiple heads of bilateral attention in parallel followed by concatenation and a linear projection. Mathematically, we define the multi-head bilateral attention as,

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1, ..., head_h)W^O \\ where\ head_i &= BiAttn(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (3)$$

where projection matrices are $W_i^Q \in \mathbb{R}^{C \times d_{hidden}}$, $W_i^K \in \mathbb{R}^{C \times d_{hidden}}$, $W_i^V \in \mathbb{R}^{C \times d_{hidden}}$, and $W^O \in \mathbb{R}^{C \times C}$. In this work, we set the number of heads ($h = C/d_{hidden}$) to 8 [38], where d_{hidden} is the hidden dimension of each head.

3.3 Optical flow calibration

As mentioned in the introduction, optical flow estimation can be noisy for objects with large motion and in texture-less areas. We introduce an optical flow calibration module to improve flow estimation by leveraging the segmentation mask from the previous frame. As shown in Fig. 3, the module employs an architecture similar to U-Net [33] with 11-layers total. To train this module to improve optical flow, we compute the Mean Square Error (MSE) between the initial optical flow and the output optical flow in training. Without the MSE loss, mask information can dominate the calibration module and thereby generate an embedding feature for the mask instead.

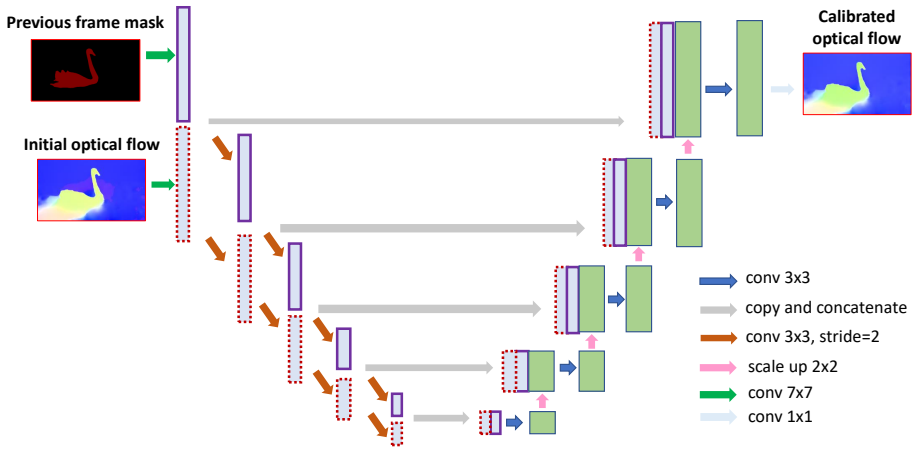


Fig. 3: The optical flow calibration module. A CNN in the U-Net architecture [33] is used to fuse the segmentation mask into the optical flow

4 Experiments

We validate BATMAN on popular benchmark datasets YouTube-VOS 2019/2018 and DAVIS 2017/2016. We first provide implementation details, followed by the experimental results. We then present the ablation study on our design.

4.1 Implementation details

We use ResNet50 [14] as the feature extractor for memory/query/flow encoder. We follow the identification embedding in [50] to encode multiple object masks in the memory encoding simultaneously. We use a RAFT [36] model pre-trained on FlyingThings3D [24] for optical flow generation. We use FPN [19] with Group Normalization [43] as the decoder. We employ 12 bilateral transformer blocks with W_d and W_b set to 7 [50] and 84 (details in supplementary), respectively.

We implement our model in PyTorch [28] and train with a batch size of 16 distributed on 8 V100 GPUs. Following previous works [22,50,25,45], we first pre-train our model on synthetic video sequences generated from static image datasets (COCO [20], ECSSD [35], MSRA10K [9], SBD [13], PASCALVOC2012 [12]) by applying random augmentations. We then train the model on the VOS benchmarks. The loss function is a combination of bootstrapped cross-entropy loss, soft Jaccard loss [26], and mean squared error loss. The training is optimized using AdamW [21] optimizer and Exponential Moving Average (EMA) [31]. The learning rate for training is set to 2×10^{-4} with a weight decay of 0.07. We train the model for 100,000 iterations.

Table 1: Results on Youtube-VOS 2019/2018 validation split. Subscript s and u denote scores in seen and unseen categories, respectively. BATMAN outperforms all state-of-the-art methods on both benchmarks

Method	Youtube-VOS 2019					Youtube-VOS 2018				
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{J}_u$	$\mathcal{F}_s$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{J}_u$	$\mathcal{F}_s$	$\mathcal{F}_u$
STM[27]	-	-	-	-	-	79.4	79.7	72.8	84.2	80.9
AFB-URR[18]	-	-	-	-	-	79.6	78.8	74.1	83.1	82.6
KMN[34]	-	-	-	-	-	81.4	81.4	75.3	85.6	83.3
CFBI[49]	81.0	80.6	75.2	85.1	83.0	81.4	81.1	75.3	85.8	83.4
LWL[2]	-	-	-	-	-	81.5	80.4	76.4	84.9	84.4
RMN[45]	-	-	-	-	-	81.5	82.1	75.7	85.7	82.4
SST[11]	81.8	80.9	76.6	-	-	81.7	81.2	76.0	-	-
TransVOS[25]	-	-	-	-	-	81.8	82.0	75.0	86.7	83.4
LCM[15]	-	-	-	-	-	82.0	82.2	75.7	86.7	83.4
CFBI+[51]	82.6	81.7	77.1	86.2	85.2	82.8	81.8	77.1	86.6	85.6
STCN[7]	82.7	81.1	78.2	85.4	85.9	83.0	81.9	77.9	86.5	85.7
RPCMVOS[47]	83.9	82.6	**79.1**	86.9	87.1	84.0	83.1	78.5	87.7	86.7
AOT[50]	84.1	83.5	78.4	88.1	86.3	84.1	83.7	78.1	88.5	86.1
BATMAN	**85.0**	**84.5**	79.0	**89.3**	**87.2**	**85.3**	**84.7**	**79.2**	**89.8**	**87.4**

4.2 Experimental results

We present validation results on the popular Youtube 2019/2018 and DAVIS 2017/2016 benchmarks compared to existing state-of-the-art methods.

Metrics. The region similarity ($\mathcal{J}$) and the boundary accuracy ($\mathcal{F}$) are computed following the standard evaluation setting proposed in [30]. On DAVIS, we report the two metrics and their mean value ($\mathcal{J}\&\mathcal{F}$). On YouTube-VOS, we report all the metrics on seen categories and unseen categories separately as generated by the evaluation server at CodaLab.

Youtube-VOS [46] is a large-scale dataset for multi-object video segmentation with objects in multiple categories. In YouTube-VOS 2018, the *training* set contains $3,471$ videos with $5,945$ unique objects in 65 categories and the *validation* set has 474 videos containing of 894 unique objects in 65 seen categories and additional 26 unseen categories. YouTube-VOS 2019 expands the YouTube-VOS 2018 dataset with more videos and object annotations. Its *training* set contains the same $3,471$ videos but has $6,459$ objects. Its *validation* set has 507 videos containing of $1,063$ objects. With the existence of the unseen object categories, the YouTube-VOS is useful to evaluate the generalization capability of the VOS model on unseen object categories. We evaluate all the results on the official YouTube-VOS evaluation servers on CodaLab.

Table 1 shows that BATMAN outperforms all state-of-the-art on Youtube-VOS 2019 and 2018 benchmarks. The higher region similarity ($\mathcal{J}$) and better

Table 2: Comparisons to the state-of-the-art methods on DAVIS benchmarks. (**Y**) indicates including Youtube-VOS dataset in training. BATMAN outperforms all state-of-the-art methods on all three DAVIS benchmarks

Method	DAVIS 2017 val			DAVIS 2017 test-dev			DAVIS 2016 val		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
AFB-URR[18]	74.6	73.0	76.1	-	-	-	-	-	-
LWL[2]	81.6	79.1	84.1	-	-	-	-	-	-
STM[27](**Y**)	-	79.2	84.3	-	-	-	-	88.7	89.9
CFBI[49](**Y**)	81.9	79.3	84.5	75.0	71.4	78.7	89.4	88.3	90.5
SST[11](**Y**)	82.5	79.9	85.1	-	-	-	-	-	-
KMN[34](**Y**)	82.8	80.0	85.6	77.2	74.1	80.3	90.5	89.5	91.5
CFBI+[51](**Y**)	82.9	80.1	85.7	75.6	71.6	79.6	89.9	88.7	91.1
RMN[45](**Y**)	83.5	81.0	86.0	75.0	71.9	78.1	88.8	88.9	88.7
LCM[15](**Y**)	83.5	80.5	86.5	78.1	74.4	81.8	90.7	89.9	91.4
RPCMVOS[47](**Y**)	83.7	81.3	86.0	79.2	75.8	82.6	90.6	87.1	94.0
TransVOS[25](**Y**)	83.9	81.4	86.4	76.9	73.0	80.9	90.5	89.8	91.2
AOT[50](**Y**)	84.9	82.3	87.5	79.6	75.9	83.3	91.1	90.1	92.1
STCN[7](**Y**)	85.4	82.2	88.6	76.1	72.7	79.6	91.6	**90.8**	92.5
BATMAN(Y)	**86.2**	**83.2**	**89.3**	**82.2**	**78.4**	**86.1**	**92.5**	90.7	**94.2**

boundary accuracy ($\mathcal{F}$) validate that bilateral attention is able to learn the most informative features from the reference frames and match the query frames.

DAVIS is one of the most popular benchmarks for video object segmentation with high-quality masks for salient objects. As part of DAVIS, DAVIS 2016 [30] is a single-object benchmark and DAVIS 2017 [32] is a multi-object extension of DAVIS 2016. In DAVIS 2016, the *training* and *validation* sets contain 30 and 20 videos, respectively. In DAVIS 2017, the *training* set consists of 60 videos, and the *validation* set consists of 30 videos, and the *test-dev* set consists of 30 videos with only the first frame annotated.

Table 2 compares BATMAN with existing state-of-the-art methods on DAVIS 2017 validation set, test-dev set, and DAVIS 2016 validation set. Note that KMN [34] only reports the results of DAVIS 2017 test-dev split with images resized to 600p. We follow the standard practice of most previous works and keep the images in the original 480p resolution in evaluation. On both multi-object datasets (DAVIS2017 val/test) and single-object dataset (DAVIS 2016), BATMAN outperforms all existing state-of-the-art methods. Moreover, BATMAN achieves the largest absolute accuracy improvement (2.6%) on the hardest DAVIS 2017 test-dev split, which validates the robustness of our model for VOS.

4.3 Ablation study

In this section, we analyze the effectiveness of the bilateral attention and compare it to the conventional spatial local attention, as well as the efficacy of the

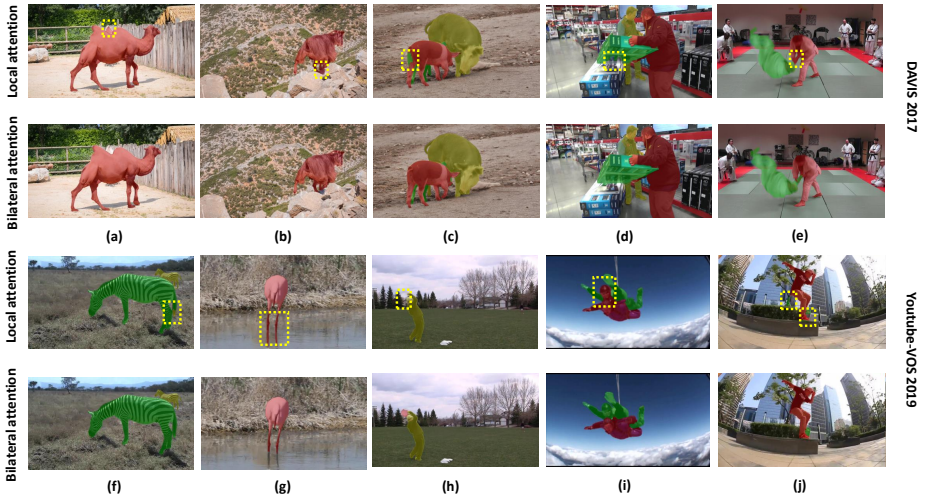


Fig. 4: Qualitative results. Compared to spatial local attention, bilateral attention segments objects better especially when background shares similar appearance with the target object

Table 3: Ablation on bilateral attention. The model with bilateral attention outperforms that with spatial local attention on all benchmarks

Attention type	DAVIS 2017 val	DAVIS 2017 test-dev	DAVIS 2016 val	Youtube-VOS 2019	Youtube-VOS 2018
Spatial local	84.9	77.5	91.6	84.1	83.8
Bilateral	**86.2**	**82.2**	**92.5**	**85.0**	**85.3**

calibrated optical flow. For qualitative analysis, we visualize the bilateral space binary mask generated by the bilateral attention, and the optical flow output from our calibration module.

Bilateral Attention. Table 3 compares the accuracy ($\mathcal{J}\&\mathcal{F}$) between our proposed bilateral attention and the conventional spatial local attention, and validates that the bilateral attention achieves superior performance on all benchmarks. We also visualize the segmentation masks from the two attention mechanisms in Fig. 4 for both DAVIS 2017 and Youtube-VOS 2019. We can see that with spatial local attention, the model tends to fail to segment objects with similar appearances (e.g., the second camel is included in the mask of the first camel (Fig. 4a); part of the red pig is segmented as the green pig (Fig. 4c); the right hand of the man in green is mistakenly segmented as part of the man in red (Fig. 4e); the tail of the zebra in the green mask is mistakenly segmented as that of the zebra in yellow (Fig. 4f)). Besides, when the appearance features (especially at the object boundary) are fuzzy (e.g., the shade of the goat (Fig. 4b),

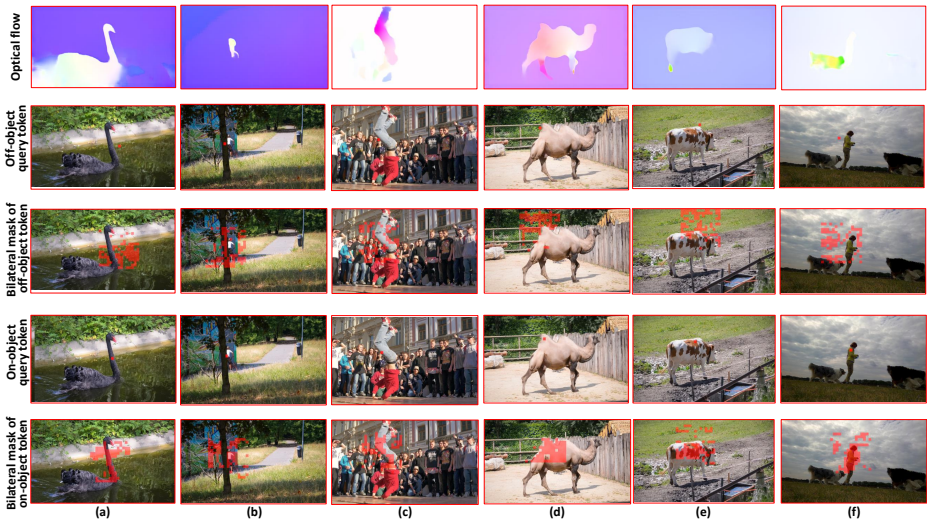
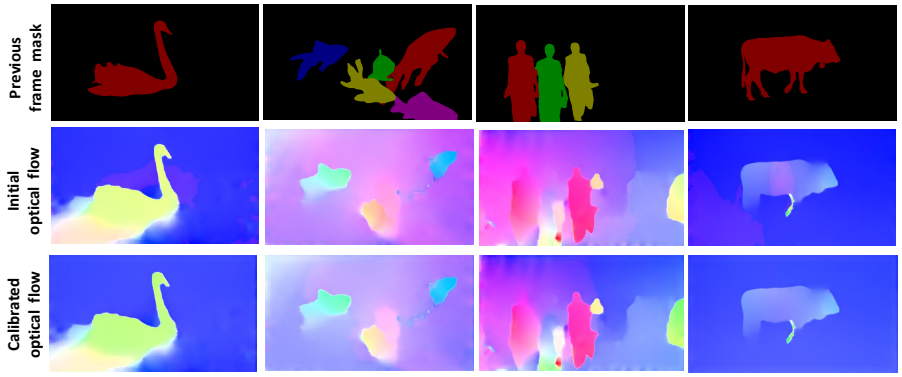


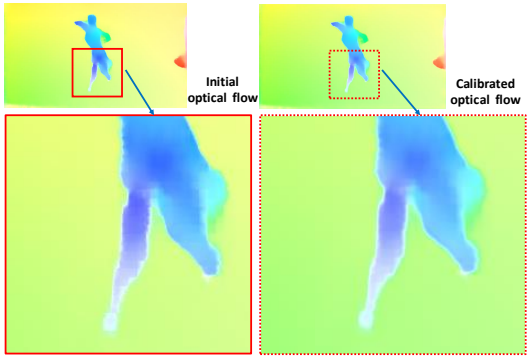
Fig. 5: Visualization of bilateral space binary masks from the bilateral attention. The bilateral attention adaptively generates binary masks for on and off object query tokens. Better view in color version

the reflection on the TV box (Fig. 4d), and the reflection of the bird's legs in the water (Fig. 4g)), the model with spatial local attention finds it difficult to segment the object properly. In contrast, the bilateral attention and the resultant adaptive bilateral space binary masks enables our model to segment target objects correctly, especially when the target object exhibits salient motion (e.g., the Frisbee (Fig. 4h) and the skydiving men (Fig. 4i)). We provide additional visualizations for segmentation in the supplementary.

Fig. 5 shows some examples of the binary masks generated from the bilateral attention. The first row shows the optical flow of the query frames. One off-object (background) query token is highlighted (in red) in the second row for each scene. The corresponding bilateral space binary mask is highlighted in the third row. In comparison, we also show an on-object query token in the fourth row, and the corresponding binary mask is given in the last row. We can see that for an off-object query token, the bilateral attention module tends to focus on the background locations (e.g., the water around the swan neck (Fig. 5a) or the sky around the woman with dogs (Fig. 5f)). On the other hand, when the query token is on the object, it tends to select the neighboring on-object tokens (e.g., the leg of the dancing man (Fig. 5c) or the camel hump (Fig. 5d)) for the attention computation. This qualitatively validates that adaptive attention computation enables propagating segmentation masks from the reference frames to the query frame more accurately.



(a) Comparison of the initial optical flow (middle) and the calibrated optical flow (bottom). The calibrated optical flow is smoother within the same object, and sharper at object boundary



(b) Blocky artifact on the initial optical flow is decreased on the calibrated optical flow

Fig. 6: Visualization of optical flow on Davis 2017 *val.* set. The calibrated optical flow is smoother within the object and sharper at the boundary

Optical flow calibration. The optical flow calibration module leverages the predicted previous frame mask to improve the optical flow estimation for the current frame. Table 4 compares the bilateral attention w/ and w/o calibrated optical flow. With the calibrated optical flow, BATMAN achieves higher accuracy on all benchmarks, validating that optical flow is improved with the help of the previous frame segmentation mask. As shown in Fig. 6, the calibrated optical flow is smoother, both within the same object and within the background. Meanwhile, the object boundary is sharper. Specifically, the blocky artifacts along the object boundary, which exists in the initial optical flow, are reduced effectively without affecting the object boundary sharpness.

Limitations. The bilateral space binary mask generation is influenced by the motion in the scene. Therefore, if the target objects do not exhibit salient motion,

Table 4: Comparisons of bilateral attention w/ and w/o optical flow calibration. Calibrating the optical flow leads to higher accuracy on all benchmarks

Optical flow type	DAVIS 2017 val	DAVIS 2017 test-dev	DAVIS 2016 val	Youtube-VOS 2019	Youtube-VOS 2018
w/o calibration	86.0	81.7	92.4	84.6	84.8
w/ calibration	**86.2**	**82.2**	**92.5**	**85.0**	**85.3**

Fig. 7: Failure cases of the bilateral binary mask generation. The bilateral attention may lose focus when a background object exhibits dominant motion and/or the target object does not exhibit salient motion

or some background object(s) exhibit salient motion and/or share(s) a similar appearance to the target object(s), the bilateral mask can be noisy and the bilateral attention may lose focus. Fig. 7 shows two failure cases: in the upper row, a man on a motorcycle moves quickly across the scene, which overwhelms the motion of the target woman. Hence, the bilateral attention fails to focus on the target object. Similarly, in the bottom row, the motion of the target woman is not salient (especially on the boundary) so the bilateral mask scatters. We plan to extend our method to better handle such scenarios.

5 Conclusions

This paper proposes a novel architecture, BATMAN, for semi-supervised VOS by adaptively computing attention between the query frame and reference frames based on the bilateral encoding of motion and appearance. Compared to conventional spatial local attention, bilateral attention adaptively selects the most relevant tokens to compute the correlation attention which helps to match the object correspondence spatially and temporally with the help of calibrated optical flow. Extensive experiments validate that BATMAN outperforms all existing state-of-the-art on all popular Youtube-VOS and DAVIS benchmarks.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bhat, G., Lawin, F.J., Danelljan, M., Robinson, A., Felsberg, M., Gool, L.V., Timofte, R.: Learning what to learn for video object segmentation. In: European Conference on Computer Vision. pp. 777–794. Springer (2020)
3. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 221–230 (2017)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
5. Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D.: State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9384–9393 (2020)
6. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1189–1198 (2018)
7. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in Neural Information Processing Systems **34** (2021)
8. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 686–695 (2017)
9. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE transactions on pattern analysis and machine intelligence **37**(3), 569–582 (2014)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5912–5921 (2021)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
13. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 international conference on computer vision. pp. 991–998. IEEE (2011)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hu, L., Zhang, P., Zhang, B., Pan, P., Xu, Y., Jin, R.: Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4144–4154 (2021)

16. Huang, X., Xu, J., Tai, Y.W., Tang, C.K.: Fast video object segmentation with temporal aggregation network and dynamic template matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8879–8889 (2020)
17. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for video object segmentation. International Journal of Computer Vision **127**(9), 1175–1197 (2019)
18. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
19. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Gool, L.V.: Video object segmentation with episodic graph memory networks. In: European Conference on Computer Vision. pp. 661–679. Springer (2020)
23. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9670–9679 (2021)
24. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Mei, J., Wang, M., Lin, Y., Yuan, Y., Liu, Y.: Transvos: Video object segmentation with transformers. arXiv preprint arXiv:2106.00588 (2021)
26. Nowozin, S.: Optimal decisions from probabilistic models: The intersection-over-union case. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 548–555 (2014)
27. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9225–9234 (2019)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
29. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2663–2672 (2017)
30. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
31. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM journal on control and optimization **30**(4), 838–855 (1992)
32. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)

33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
34. Seong, H., Hyun, J., Kim, E.: Kernelized memory network for video object segmentation. In: European Conference on Computer Vision. pp. 629–645. Springer (2020)
35. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended cssd. IEEE transactions on pattern analysis and machine intelligence **38**(4), 717–729 (2015)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
37. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3899–3908 (2016)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9481–9490 (2019)
40. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 1328–1338 (2019)
41. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021)
42. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. pp. 38–45 (2020)
43. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
44. Xiao, H., Feng, J., Lin, G., Liu, Y., Zhang, M.: Monet: Deep motion exploitation for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1140–1148 (2018)
45. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1286–1295 (2021)
46. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
47. Xu, X., Wang, J., Li, X., Lu, Y.: Reliable propagation-correction modulation for video object segmentation. arXiv preprint arXiv:2112.02853 (2021)
48. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6556–6565 (2018)
49. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: European Conference on Computer Vision. pp. 332–348. Springer (2020)

50. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34** (2021)
51. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by multi-scale foreground-background integration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
52. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al.: Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems **33**, 17283–17297 (2020)
53. Zhang, J., Xie, J., Barnes, N., Li, P.: Learning generative vision transformer with energy-based latent space for saliency prediction. Advances in Neural Information Processing Systems **34** (2021)
54. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)