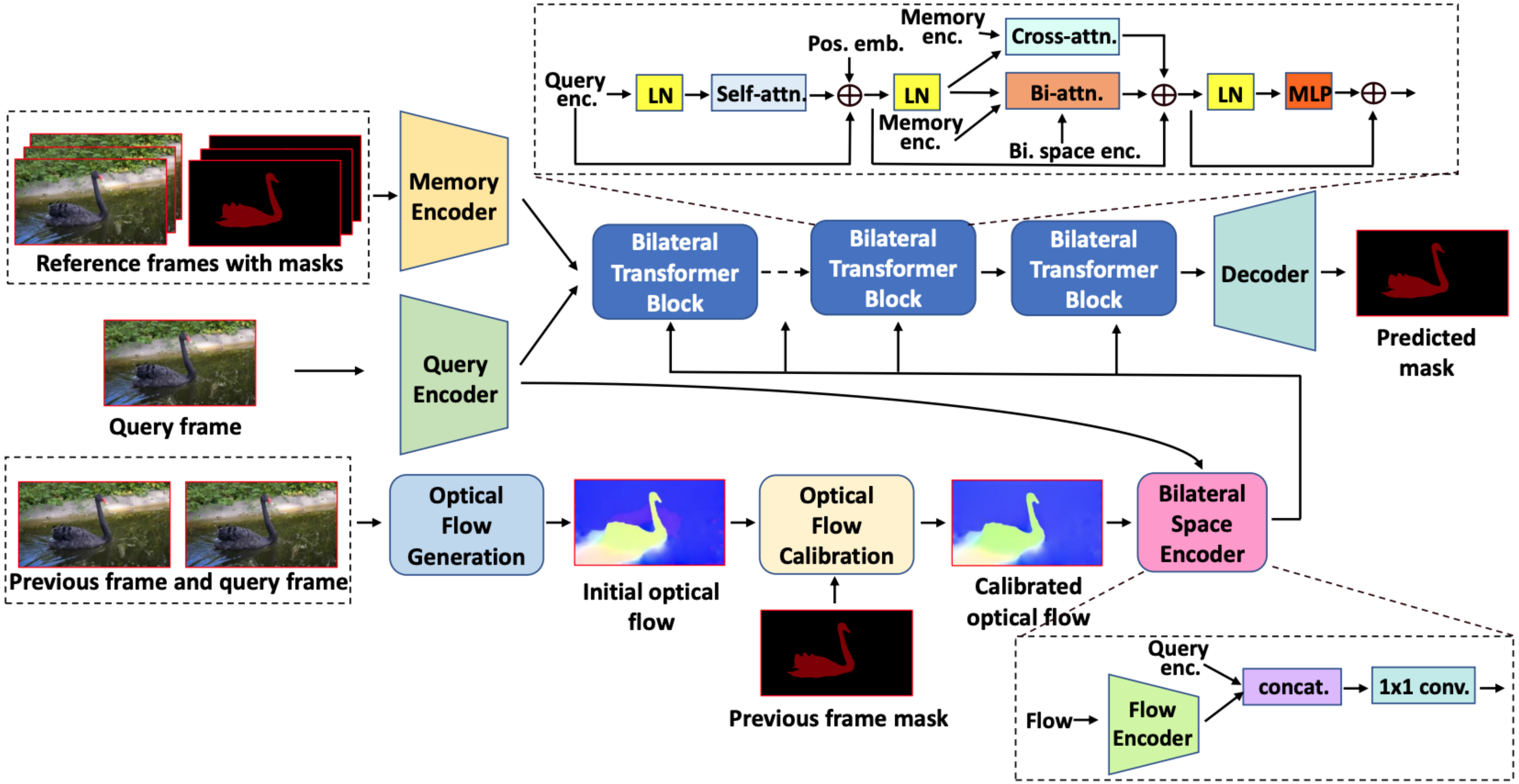
BATMAN: Bilateral Attention Transformer in Motion-Appearance Neighboring Space for Video Object Segmentation

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

Overall Pipeline

1. Extract frame-level feature
2. Compute optical flow
3. Optical flow calibration
4. Encode the optical flow and query
5. Bilateral transformer
6. Decode the final feature



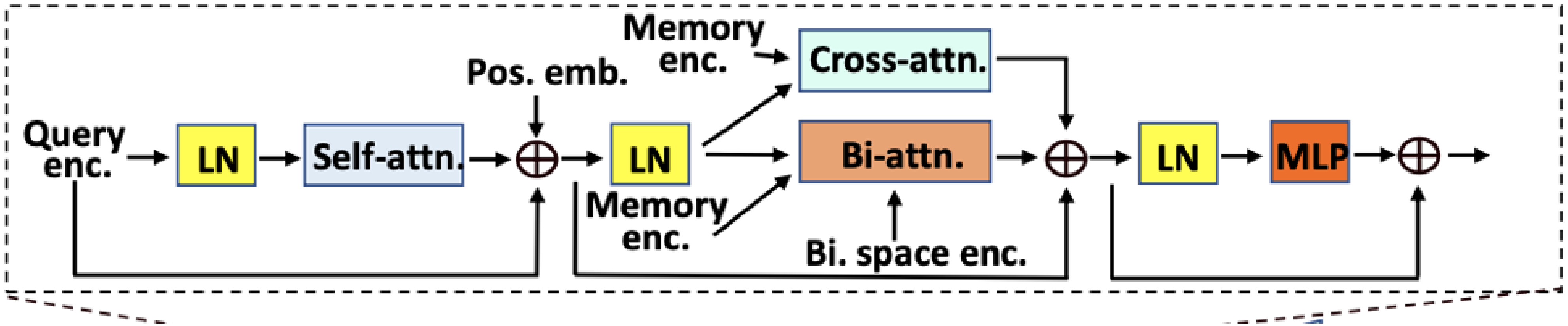
Bilateral space encoding

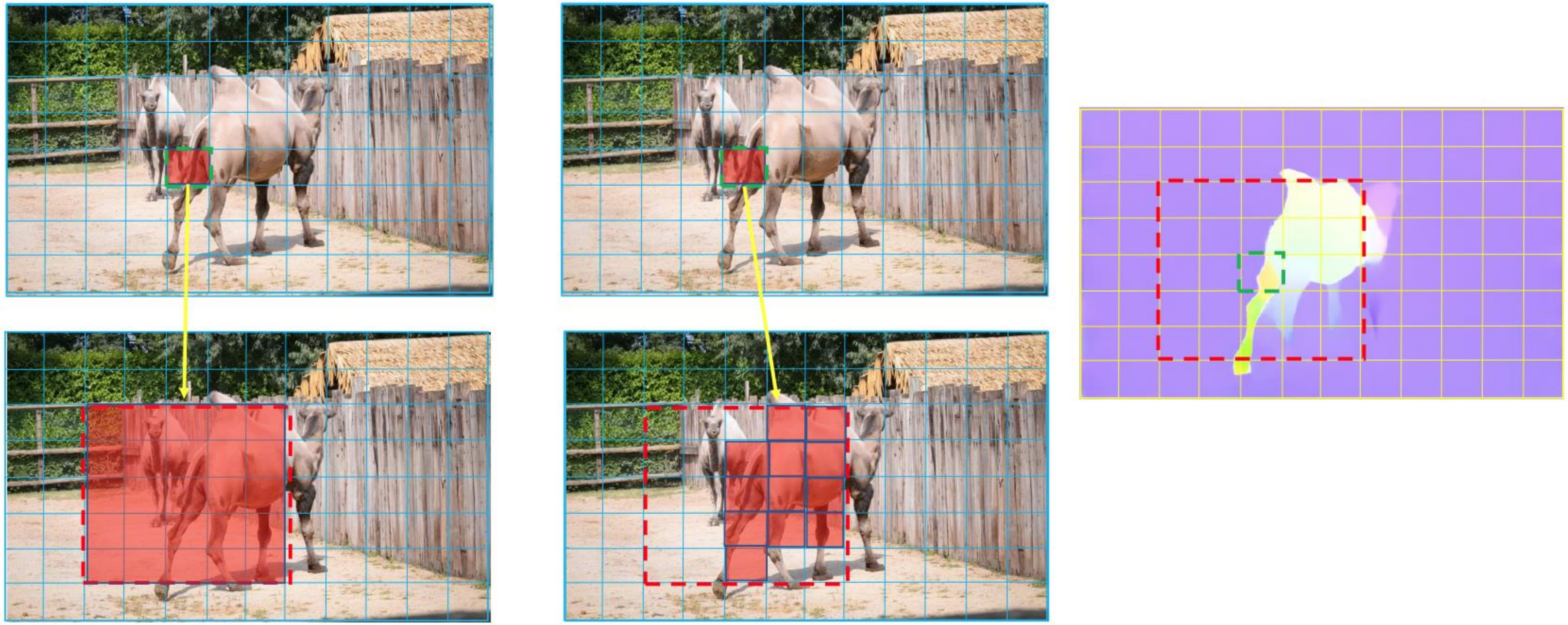
$$F = \text{optical flow}$$
$$F' = Flow\ encoder(F)$$
$$E = Conv_{1\times1}\big(Concate(F', Q)\big)$$

Bilateral attention

$$BiAttn(Q, K, V) = softmax\left(\frac{QK^TM}{\sqrt{C}}\right)V, M \in [0,1]^{HW\times HW}, Q \in \mathbb{R}^{HW\times C},$$

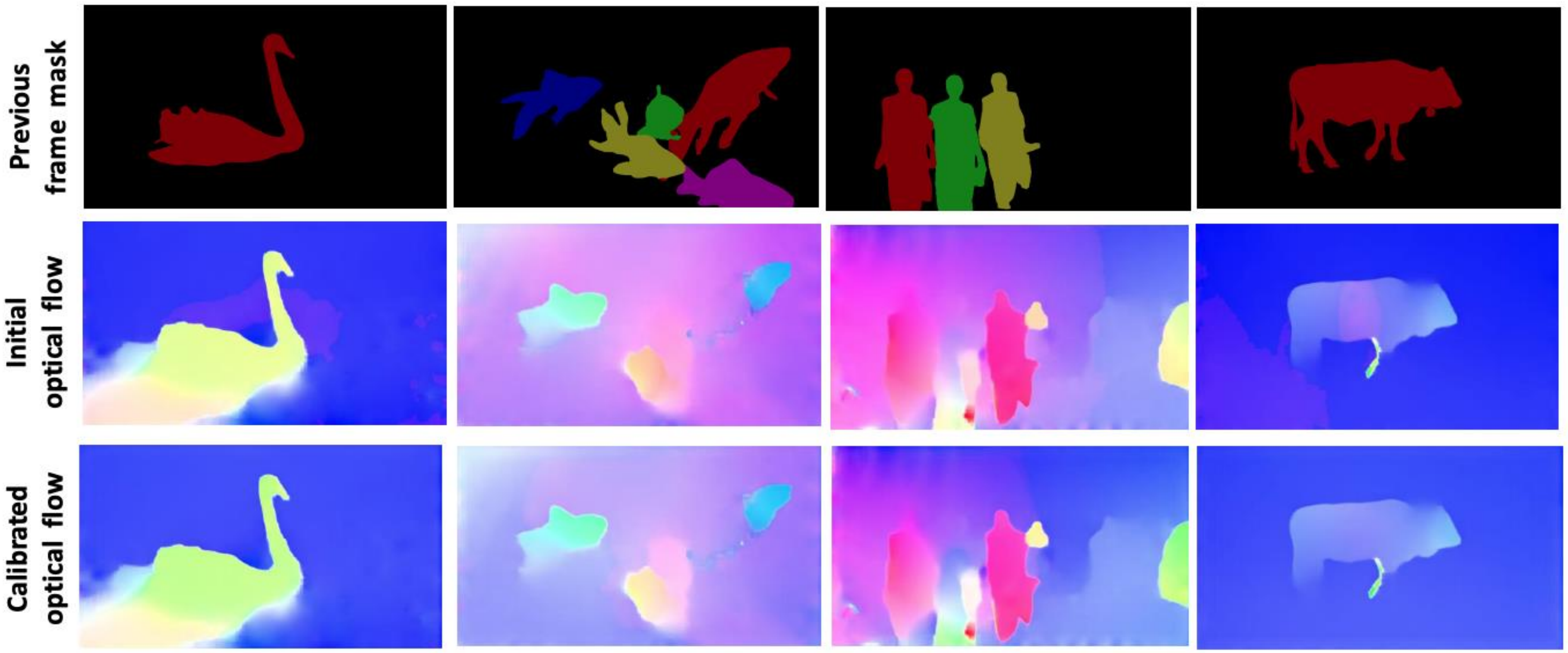
$$M_{h,w}(i,j,E) = \begin{cases} 1 & \text{if } |i-h| \leqslant W_d \text{ and } |j-w| \leqslant W_d \\ & \text{and } |argsort_{W_d}(E_{h,w}) - argsort_{W_d}(E_{i,j})| \leqslant W_b \\ 0 & \text{otherwise} \end{cases}$$





(a) (b)

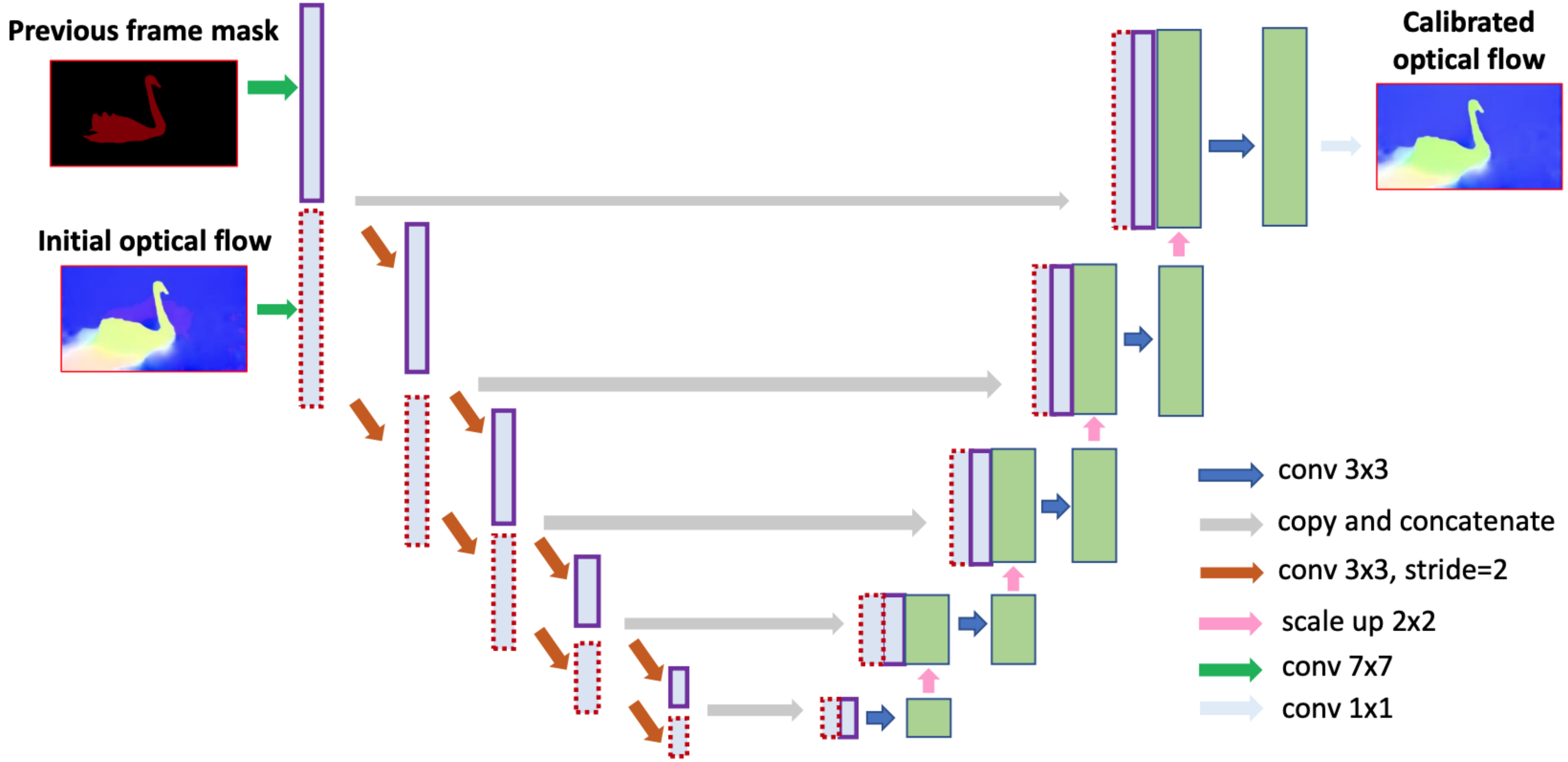
Optical flow calibration



(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

Optical flow calibration

Use MSE between the initial optical flow and the output



Results

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**

Results

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**

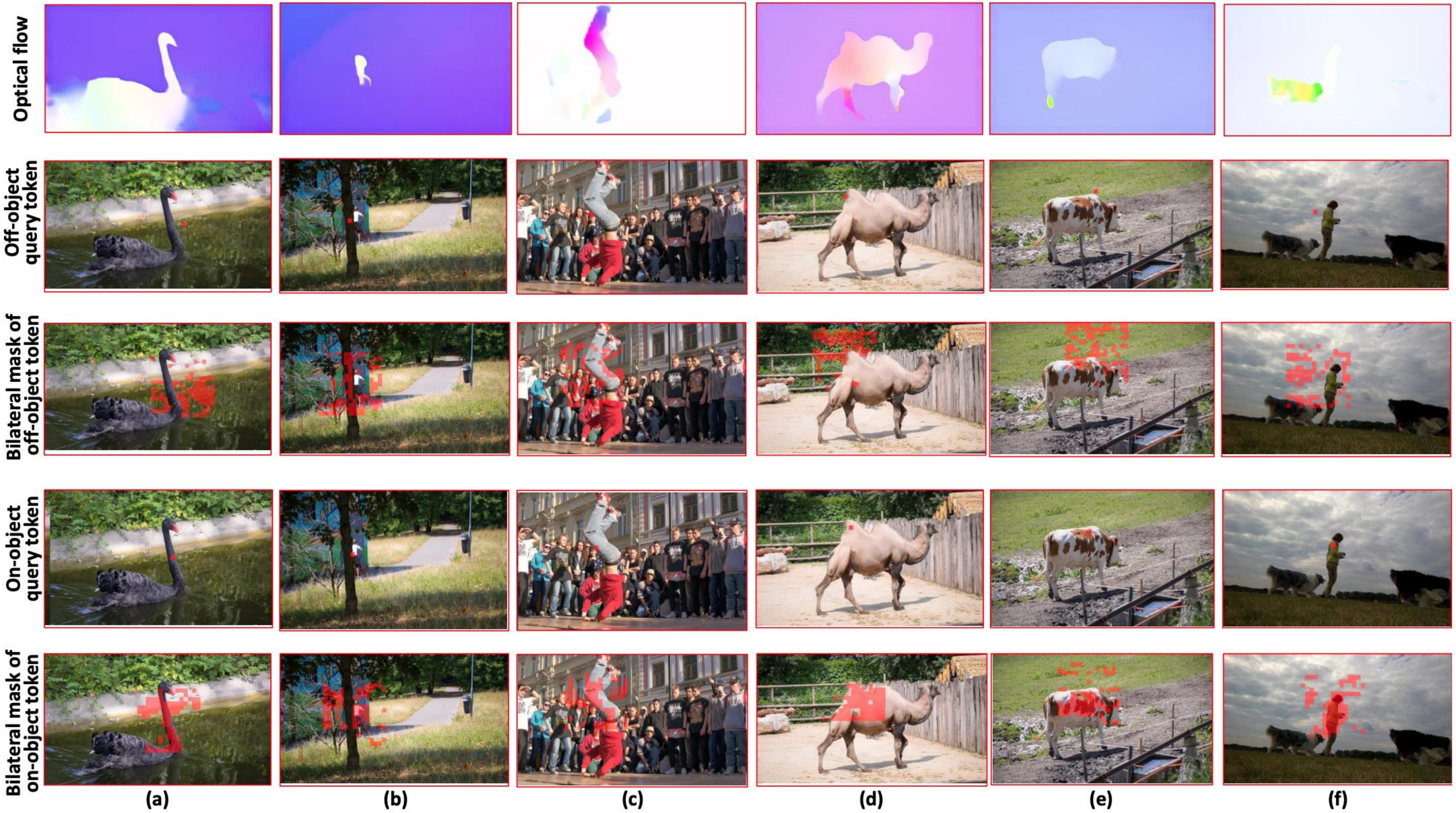
Ablation study

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**

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**


Optical flow
Off-object query token
Bilateral mask of off-object token
On-object query token
Bilateral mask of on-object token
(a)
(b)
(c)
(d)
(e)
(f)