Associating Objects with Transformers for Video Object Segmentation

Zongxin Yang[1,2], **Yunchao Wei**[3], **Yi Yang**[2]

[1] Baidu Research
[2] CCAI, College of Computer Science and Technology, Zhejiang University
[3] ReLER, Centre for Artificial Intelligence, University of Technology Sydney
{zongxinyang1996, wychao1987, yee.i.yang}@gmail.com

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 4 Jun 2021 (v1), last revised 22 Jun 2021 (this version, v2)]

Associating Objects with Transformers for Video Object Segmentation

Zongxin Yang, Yunchao Wei, Yi Yang

This paper investigates how to realize better and more efficient embedding learning to tackle the semi-supervised video object segmentation under challenging multi-object scenarios. The state-of-the-art methods learn to decode features with a single positive object and thus have to match and segment each target separately under multi-object scenarios, consuming multiple times computing resources. To solve the problem, we propose an Associating Objects with Transformers (AOT) approach to match and decode multiple objects uniformly. In detail, AOT employs an identification mechanism to associate multiple targets into the same high-dimensional embedding space. Thus, we can simultaneously process the matching and segmentation decoding of multiple objects as efficiently as processing a single object. For sufficiently modeling multi-object association, a Long Short-Term Transformer is designed for constructing hierarchical matching and propagation. We conduct extensive experiments on both multi-object and single-object benchmarks to examine AOT variant networks with different complexities. Particularly, our AOT-L outperforms all the state-of-the-art competitors on three popular benchmarks, i.e., YouTube-VOS (83.7% J&F), DAVIS 2017 (83.0%), and DAVIS 2016 (91.0%), while keeping more than 3X faster multi-object run-time. Meanwhile, our AOT-T can maintain real-time multi-object speed on the above benchmarks. We ranked 1st in the 3rd Large-scale Video Object Segmentation Challenge. The code will be publicly available at this https URL.

Comments: 18 pages, 9 figures, 5 tables

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**

Cite as: arXiv:2106.02638 **[cs.CV]**

(or arXiv:2106.02638v2 **[cs.CV]** for this version)

Submission history

From: Zongxin Yang [view email]

[v1] Fri, 4 Jun 2021 17:59:57 UTC (5,909 KB)

[v2] Tue, 22 Jun 2021 14:48:51 UTC (5,916 KB)

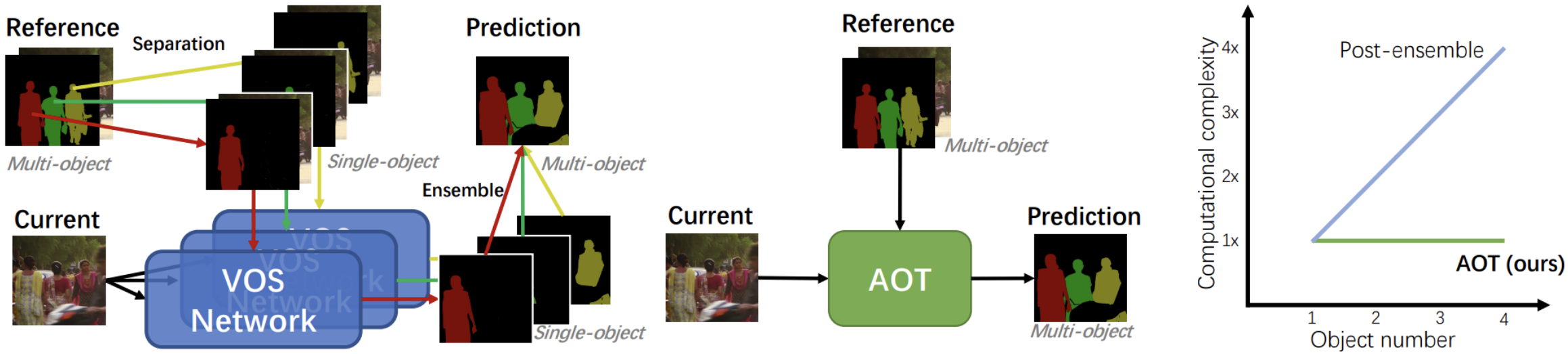


Figure 1: The state-of-the-art VOS methods (*e.g.*, [59, 40]) process multi-object scenarios in a post-ensemble manner (a). In contrast, our AOT associates and decodes multiple objects uniformly (b), leading to better efficiency (c).

Identity Assignment

- **Identity Embedding**

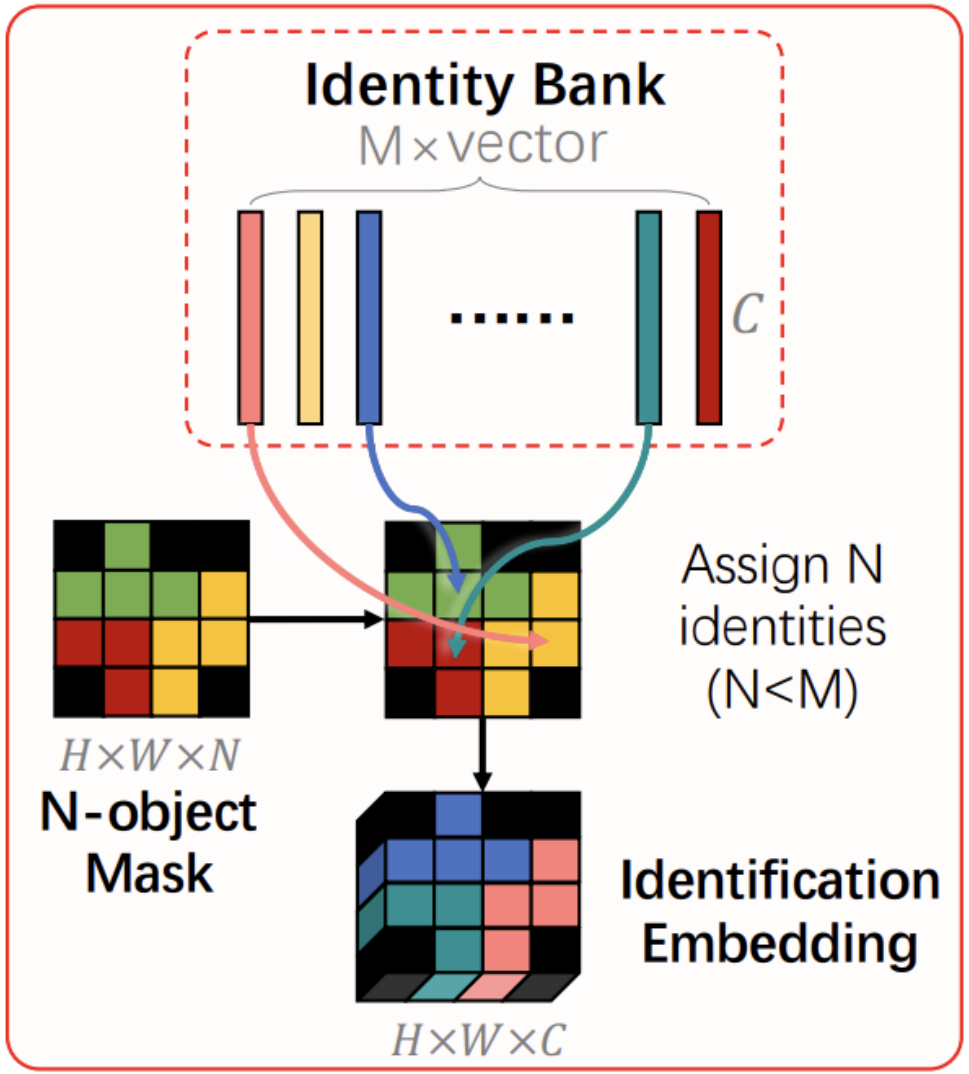
$$E = ID(Y, D) = YPD,$$

$$V' = AttID(Q, K, V, Y|D)$$

$$Att(Q, K, V + ID(Y, D)) = Att(Q, K, V + E),$$

- **Identity Decoding**

$$Y' = softmax(PF^{\mathcal{D}}(V')) = softmax(PL^{D}),$$



Long-short term transformer (LSTT)

- **Long Term Attention**

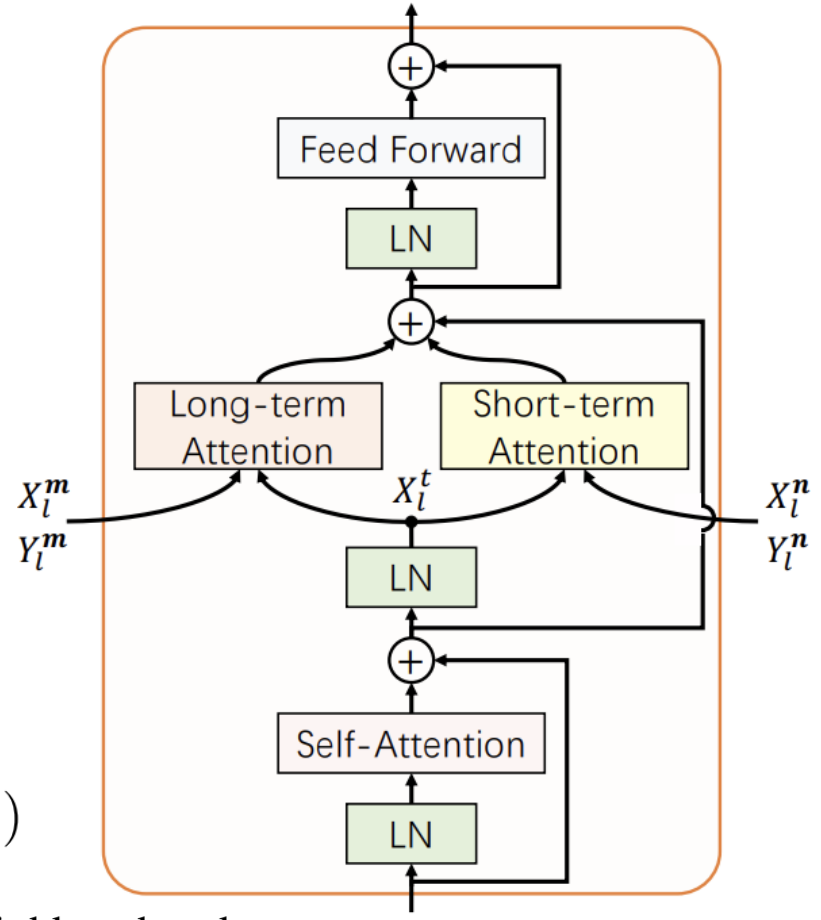
$$AttLT(X_l^t, X_l^{\mathbf{m}}, Y^{\mathbf{m}}) = AttID(X_l^t W_l^K, X_l^{\mathbf{m}} W_l^K, X_l^{\mathbf{m}} W_l^V, Y^{\mathbf{m}}|D),$$

$$X_l^{\mathbf{m}} = Concat(X_l^{m_1}, ..., X_l^{m_T}) \text{ and } Y^{\mathbf{m}} = Concat(Y^{m_1}, ..., Y^{m_T})$$

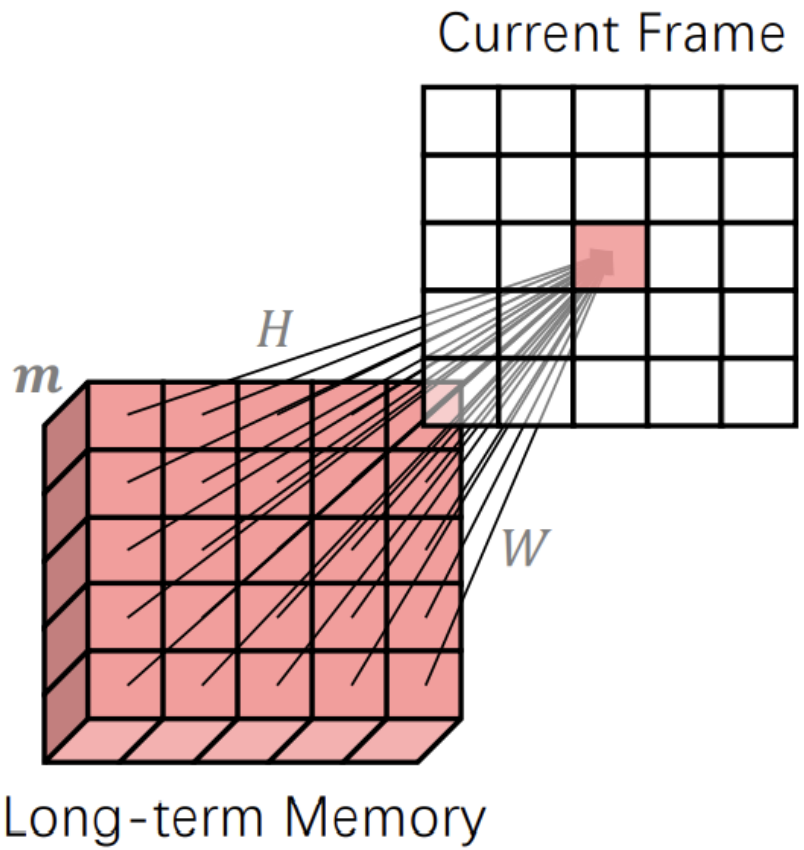
- **Short Term Attention**

$$AttST(X_l^t, X_l^{\mathbf{n}}, Y^{\mathbf{n}}|p) = AttLT(X_{l,p}^t, X_{l,\mathcal{N}(p)}^{\mathbf{n}}, Y_{l,\mathcal{N}(p)}^{\mathbf{n}}),$$

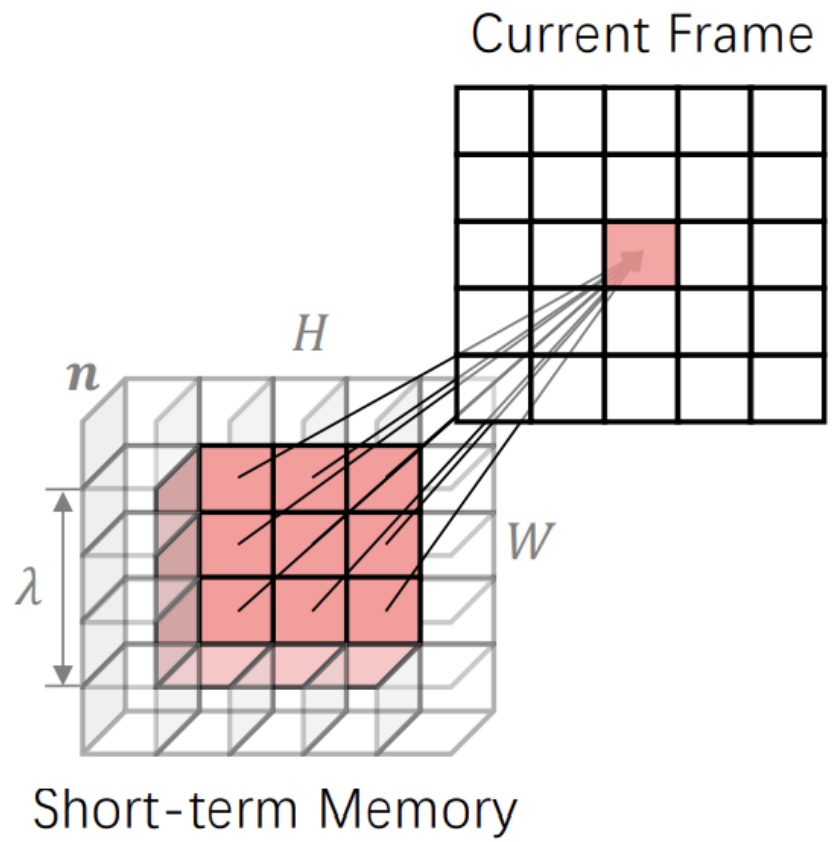
$$X_l^{\mathbf{n}} = Concat(X_l^{t-1}, ..., X_l^{t-n}) \text{ and } Y^{\mathbf{n}} = Concat(Y^{t-1}, ..., Y^{t-n})$$



where $X_{l,p}^t \in \mathbb{R}^{1\times C}$ is the feature of X_l^t at location p, $\mathcal{N}(p)$ is a $\lambda \times \lambda$ spatial neighbourhood centered at location p, and thus $X_{l,\mathcal{N}(p)}^{\mathbf{n}}$ and $Y_{l,\mathcal{N}(p)}^{\mathbf{n}}$ are the features and masks of the spatial-temporal neighbourhood, respectively, with a shape of $n\lambda^2 \times C$ or $n\lambda^2 \times N$.



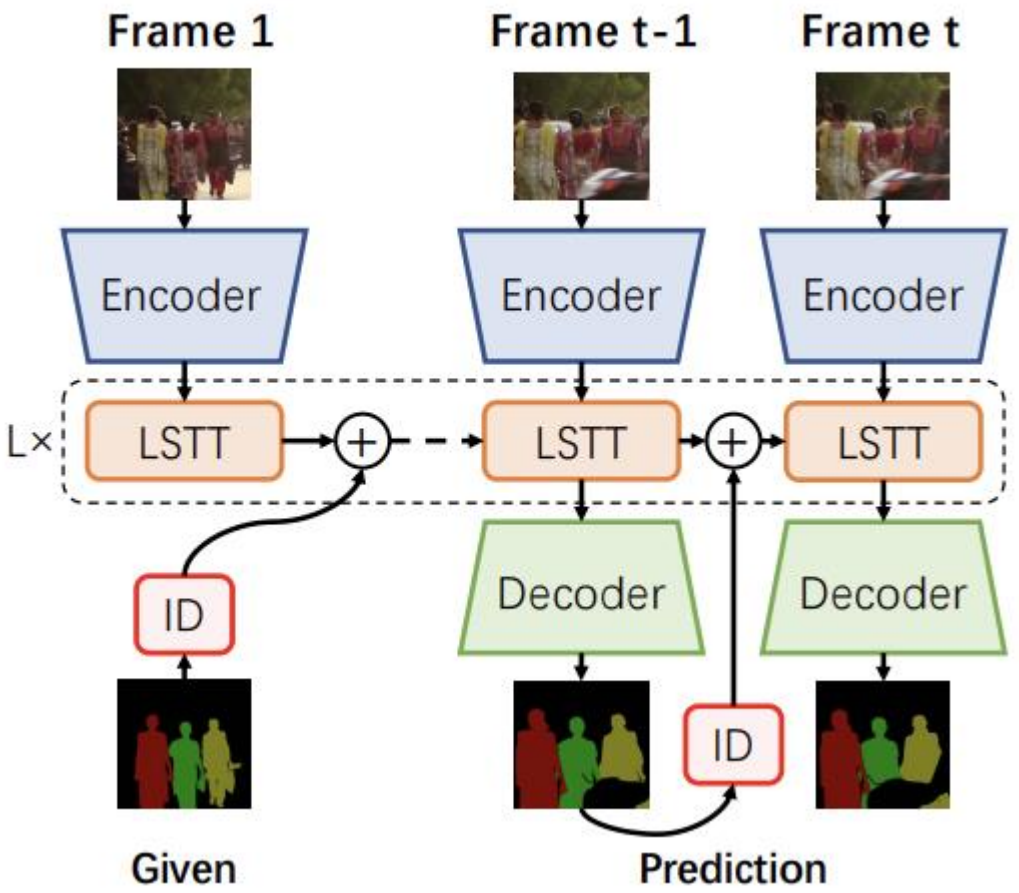
(a) Long-term Attention



(b) Short-term Attention

Overview Architecture

- Encoder
  - MobileNet V2
- Decoder
  - FPN



(a) YouTube-VOS

Methods	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
	Seen			Unseen		
Validation 2018 Split						
AG[CVPR19] [21]	66.1	67.8	-	60.8	-	-
PReM[ACCV18] [27]	66.9	71.4	75.9	56.5	63.7	0.17
BoLT[arXiv19] [48]	71.1	71.6	-	64.3	-	0.74
STM[ICCV19] [32]	79.4	79.7	84.2	72.8	80.9	-
EGMN[ECCV20] [26]	80.2	80.7	85.1	74.0	80.9	-
KMN[ECCV20] [40]	81.4	81.4	85.6	75.3	83.3	-
CFBI[ECCV20] [59]	81.4	81.1	85.8	75.3	83.4	3.4
LWL[ECCV20] [7]	81.5	80.4	84.9	76.4	84.4	-
SST[CVPR21] [15]	81.7	81.2	-	76.0	-	-
CFBI+[TPAMI21] [60]	82.8	81.8	86.6	77.1	85.6	4.0
AOT-T	80.2	80.1	84.5	74.0	82.2	**32.2**
AOT-S	82.6	82.0	86.7	76.6	85.0	22.1
AOT-B	83.2	**82.6**	87.4	77.3	85.6	17.0
AOT-L	**83.7**	82.5	**87.5**	**77.9**	**86.7**	15.2
Validation 2019 Split						
CFBI[ECCV20] [59]	81.0	80.6	85.1	75.2	83.0	3.4
SST[CVPR21] [15]	81.8	80.9	-	76.6	-	-
CFBI+[TPAMI21] [60]	82.6	81.7	86.2	77.1	85.2	4.0
AOT-T	79.7	79.6	83.8	73.7	81.8	**32.2**
AOT-S	82.2	81.3	85.9	76.6	84.9	22.1
AOT-B	83.3	**82.5**	**87.0**	77.8	86.0	17.0
AOT-L	**83.6**	82.2	86.9	**78.3**	**86.9**	15.2

(b) DAVIS 2017

Methods	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
Validation 2017 Split				
STM [32] (**Y**)	81.8	79.2	84.3	3.1‡
CFBI [59] (**Y**)	81.9	79.3	84.5	5.9
SST [15] (**Y**)	82.5	79.9	85.1	-
EGMN [26] (**Y**)	82.8	80.2	85.2	2.5‡
KMN [40]	76.0	74.2	77.8	4.2‡
KMN [40] (**Y**)	82.8	80.0	85.6	4.2‡
CFBI+ [60] (**Y**)	82.9	80.1	85.7	5.6
AOT-T (**Y**)	78.2	75.8	80.6	**39.1**
AOT-S	79.2	76.4	82.0	29.0
AOT-S (**Y**)	81.0	78.5	83.4	29.0
AOT-B (**Y**)	82.1	79.4	84.8	22.7
AOT-L (**Y**)	**83.0**	**80.3**	**85.7**	18.9
Testing 2017 Split				
STM* [32] (**Y**)	72.2	69.3	75.2	-
CFBI [59] (**Y**)	75.0	71.4	78.7	5.3
CFBI* [59] (**Y**)	76.6	73.0	80.1	2.9
KMN* [40] (**Y**)	77.2	74.1	80.3	-
CFBI+* [60] (**Y**)	78.0	74.4	81.6	3.4
AOT-T (**Y**)	69.3	66.0	72.5	**39.1**
AOT-S (**Y**)	73.6	69.7	77.4	29.0
AOT-B (**Y**)	75.5	71.8	79.1	22.7
AOT-L (**Y**)	78.4	74.8	82.1	18.9
AOT-L* (**Y**)	**78.8**	**75.3**	**82.3**	12.7

AOT-Tiny:L=1, m=1

AOT-Small:L=2, m=1

AOT-Base:L=3, m=1

AOT-Large:L=3, m={1,7,13,……}

AOT-Base 5 times faster than CFBI (15.2fps vs 3.4fps)

Ablation study

Table 3: Ablation study. The experiments are based on AOT-S and conducted on the validation 2018 split of YouTube-VOS [55] without pre-training on synthetic videos. Self: the position embedding type used in the self-attention. Rel: use relative positional embedding [41] on the local attention.

(a) Identity number

M	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}^{seen}$	$\mathcal{J}^{unseen}$
10	**80.3**	**80.6**	**73.7**
15	79.0	79.4	72.1
20	78.3	79.4	70.8
30	77.2	78.5	70.2

(b) Local window size

λ	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}^{seen}$	$\mathcal{J}^{unseen}$
7	**80.3**	**80.6**	**73.7**
5	78.8	79.5	71.9
3	78.3	79.3	70.9
0	74.3	74.9	67.6

(c) Local frame number

n	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}^{seen}$	$\mathcal{J}^{unseen}$
1	**80.3**	**80.6**	**73.7**
2	80.0	79.8	73.7
3	79.1	80.0	72.2
0	74.3	74.9	67.6

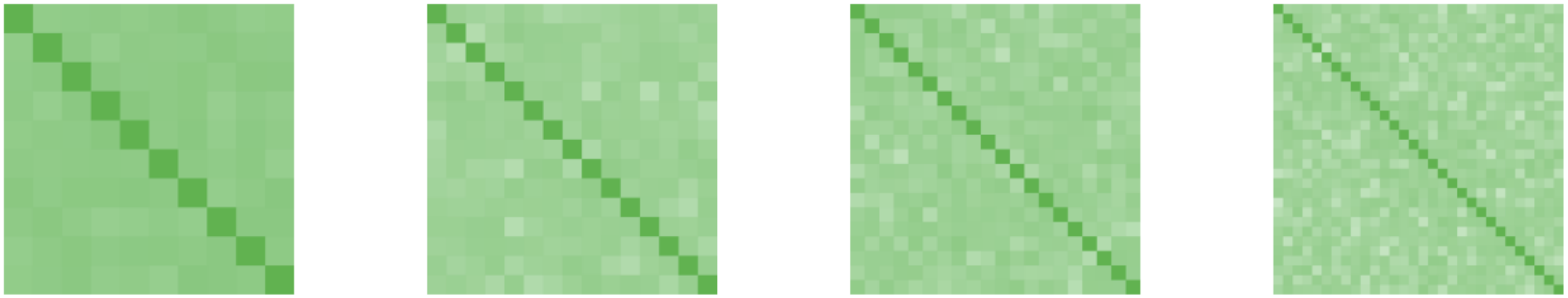
(d) LSTT block number

L	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}^{seen}$	$\mathcal{J}^{unseen}$	FPS	Param
2	80.3	80.6	73.7	22.1	7.0M
3	**80.9**	**81.1**	**74.0**	17.0	8.3M
1	77.9	78.8	71.0	**32.2**	**5.7M**

(e) Positional embedding

Self	Rel	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}^{seen}$	$\mathcal{J}^{unseen}$
sine	✓	**80.3**	**80.6**	**73.7**
none	✓	80.1	80.4	73.5
sine	-	79.7	80.1	72.9

Interpretability — Identity Bank



(a) $M = 10$ (default) (b) $M = 15$ (c) $M = 20$ (d) $M = 30$

Figure 4: Visualization of the cosine similarity between every two of M identification vectors in the identity bank. We use the form of a $M \times M$ symmetric matrix to visualize all the cosine similarities, and the values on the diagonal are all equal to 1. The darker the green color, the higher the similarity. In the case of $M = 10$, the similarities are stable and balanced. As the vector number M increases, The visualized matrix becomes less and less smooth, which means the similarities become unstable.

Interpretability — Long term & Short term Memory

Thanks for watching!