Supplementary Material:
Joint Inductive and Transductive Learning for Video Object Segmentation

Yunyao Mao[1] Ning Wang[1] Wengang Zhou[1,2,*] Houqiang Li[1,2,*]
[1] CAS Key Laboratory of Technology in GIPAS, EEIS Department, University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{myy2016,wn6149}@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn

1. Label Encoder Structure

In Figure 1, we show the detailed architecture of our proposed two-head label encoder, which is a modified version of the label encoder proposed in LWL [1]. Note that the generated mask encodings are further forced to be disentangled by the cosine similarity loss.

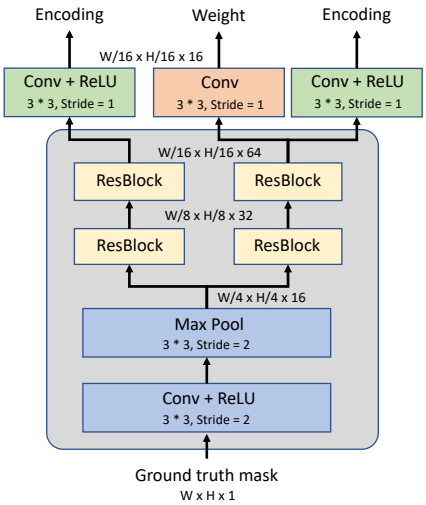

Figure 1. Detailed architecture of our two-head label encoder. For brevity, we omit the weight predictor for the induction branch.

2. Ablation Study

2.1. Hyper-parameters

In this section, we investigate the impact of some hyper-parameters like the template sampling interval T and memory size $N_{\max}$. Figure 2 shows the overall score for different sampling interval T. We can find that a sampling interval of 5 provides the best result with an overall score of 83.1% on YouTube-VOS 2018 [4] validation set.

Similarly, Figure 3 shows the overall score for different memory size $N_{\max}$. In these experiments, we fix the sampling interval T to 5. As we can see, larger memory sizes lead to higher overall scores until a memory size of 20. For memory sizes larger than 40, the performance does

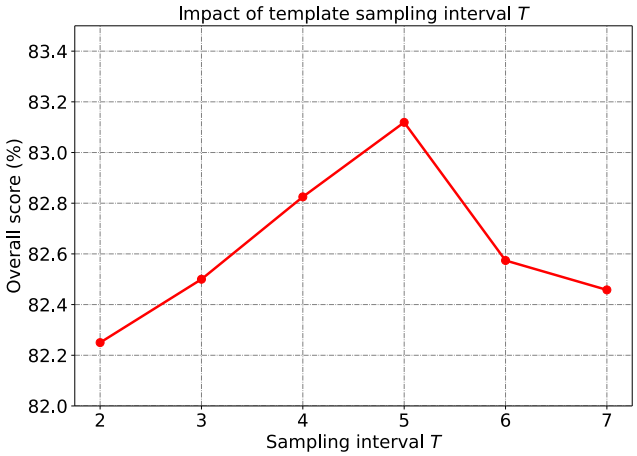

Figure 2. Impact of the template sampling interval. The performance is evaluated on the YouTube-VOS 2018 [4] validation set in terms of the overall score.

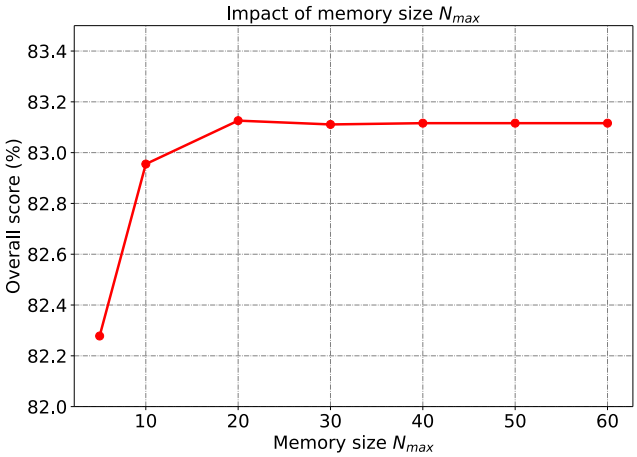

Figure 3. Impact of the memory size. The performance is evaluated on the YouTube-VOS 2018 [4] validation set in terms of the overall score.

*Corresponding authors: Wengang Zhou and Houqiang Li

not change, since the length of videos in the YouTube-VOS 2018 [4] validation set rarely exceeds 200.

2.2. Merging Strategy

We analyze different ways of merging the mask encodings provided by the two complementary branches. In our approach, the two mask encodings are element-wisely added and are forced to be decoupled during offline training. An alternative way to prevent the mask encodings from being coupled together is to directly concatenate them along the channel dimension. In Table 1, we report the performance of the aforementioned two strategies. Naively concatenating the encodings achieves an overall score of 81.7%, which is 1.4% lower than ours (83.1%). Note that for fair comparison, we modify the last layer of the label encoder to make sure that the outputs of the two merging strategies have the same channel dimension.

Table 1. Ablation study for merging strategy. The performance is evaluated on the YouTube-VOS 2018 [4] validation set in terms of mean Jaccard ($\mathcal{J}$) and boundary ($\mathcal{F}$) scores on both seen and unseen categories.

	$\mathcal{J}_{\text{seen}}$	$\mathcal{F}_{\text{seen}}$	$\mathcal{J}_{\text{unseen}}$	$\mathcal{F}_{\text{unseen}}$	Overall
Concatenated	80.9	85.4	76.3	84.1	81.7
Added	81.1	85.6	77.6	85.3	82.4
Added & decoupled	81.5	85.9	78.7	86.5	83.1

2.3. Self Attention

We also analyze the impact of the self-attention module adopted in the proposed transformer architecture. If we remove all the self-attention layers in the transformer, the transduction branch will be reduced into a non-local attention module. As we can see in Table 2, without the self-attention layers, the overall score drops from 83.1% to 81.7%. This verify the effectiveness of the self-attention layer for allowing template features to mutually reinforce to be more compact and representative.

Table 2. Ablation study for the self-attention module in our lightweight transformer. The performance is evaluated on the YouTube-VOS 2018 [4] validation set in terms of mean Jaccard ($\mathcal{J}$) and boundary ($\mathcal{F}$) scores on both seen and unseen categories.

	$\mathcal{J}_{\text{seen}}$	$\mathcal{F}_{\text{seen}}$	$\mathcal{J}_{\text{unseen}}$	$\mathcal{F}_{\text{unseen}}$	Overall
w/o self-attention	81.0	85.3	76.5	84.1	81.7
w/ self-attention	81.5	85.9	78.7	86.5	83.1

2.4. Induction branch v.s. Transductive branch

For the inductive branch, the online optimized model shows greater discrimination capability than the transductive branch when encountering *similar instances* (*e.g.* Fig. 6 in the paper). However, it struggles to explore the underlying context to produce temporal consistent results when the target undergoes *dramatic appearance changes* (*e.g.* Fig. 4 and Fig. 6 in the paper), as this model processes each frame and each local region inside a frame independently.

For the transductive branch, the non-local attention mechanism is naturally suitable for the spatio-temporal dependency modeling, thus shows better temporal coherence (Fig. 4). Whereas, it is less discriminative due to the fixed embedding for feature matching.

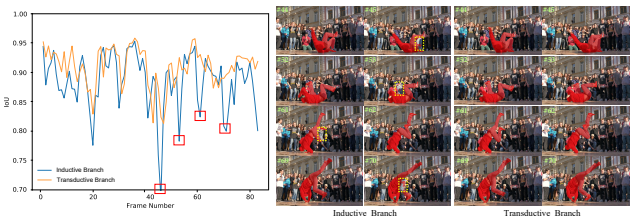

Figure 4. Per-frame IoU over time (left) and key frames with failures marked in yellow boxes (right). The transductive branch shows better spatio-temporal coherence. Please zoom in to view.

3. Qualitative Results

In Figure 5, we show some qualitative comparisons between our approach and recently proposed methods like EGMN [2], CFBI [5], STM [3], and LWL [1]. We select six representative video sequences from DAVIS 2017 validation set, including *breakdance*, *dance-twirl*, *india*, *pigs*, *soapbox*, and *bike-packing*. These sequences contain many challenging scenarios like similar distractors, occlusions, and appearance changes. In the *breakdance* sequence, EGMN [2] mistakenly regards the person squatting behind as part of the target. CFBI [5], STM [3], and LWL [1] fail to segment one of the legs. In the *pigs* sequence, EGMN [2], CFBI [5], and LWL [1] fail to segment the piglet in green after severe occlusion. In general, our approach performs well on the first five sequences compared with other methods. In the *bike-packing* sequence, our approach fails to segment the legs of the person.

References

[1] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, 2020. 1, 2

[2] Xiankai Lu, Wenguan Wang, Danelljan Martin, Tianfei Zhou, Jianbing Shen, and Van Gool Luc. Video object segmentation with episodic graph memory networks. In *ECCV*, 2020. 2

[3] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2

[4] N. Xu, L. Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and T. Huang. Youtube-vos: A large-

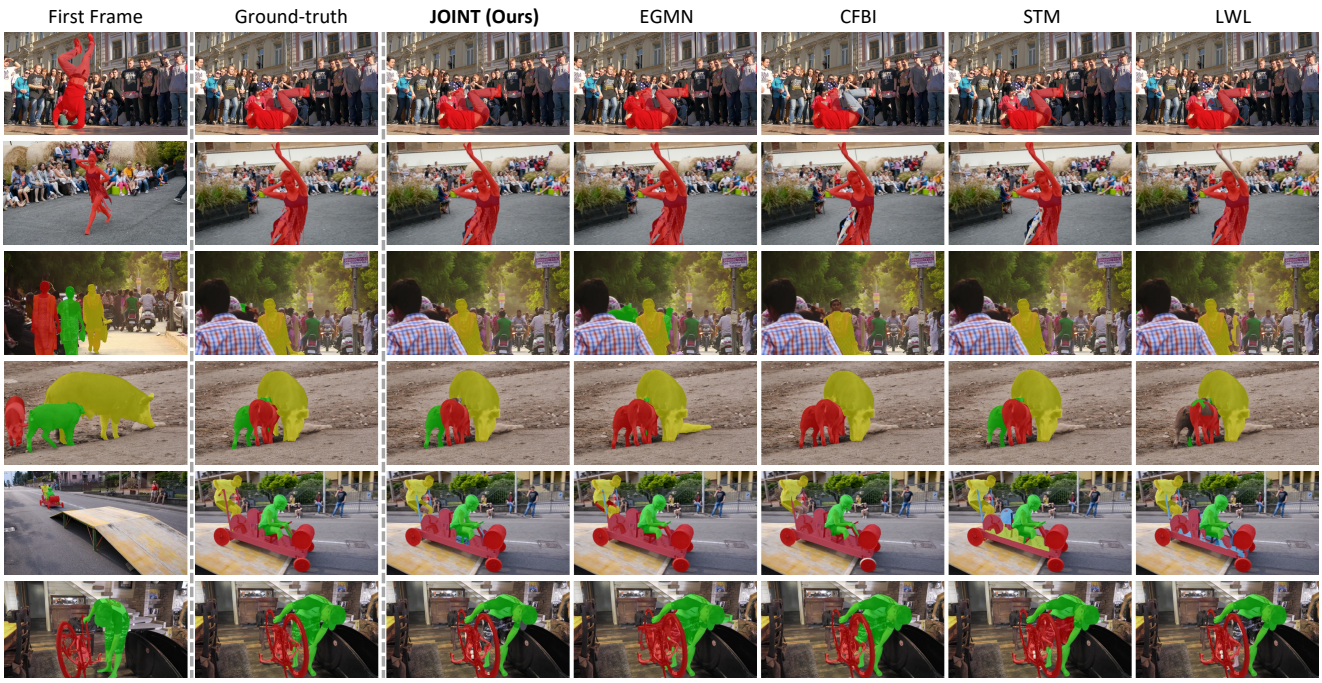

Figure 5. Qualitative comparisons between our approach and recently proposed methods on the DAVIS 2017 validation set. The selected frames experience similar distractors (1^{st}), occlusions (3^{rd} and 4^{th} rows), or tremendous appearance changes (2^{nd} and 5^{th} rows). The last row shows a failure case, the algorithm fails to segment the legs of the person.

scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 2

[5] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 2