Supplementary Material: Video Instance Segmentation with a Propose-Reduce Paradigm

1. Overview

We provide additional details in this supplementary file. Sec. 2 describes the details of the sequence proposals reduction. In Sec. 3, we describe more details regarding the Seq-Prop head. In Sec. 4, we clarify more details of the implementation. More discussions are presented in Sec. 5. More visual results are shown in Sec. 6.

2. Sequence Proposals Reduction

With the defined input sequence set S, sequence score (Eq. (1)) and sequences IoU (Eq. (2)) described in our main paper (Sec. 3.2), we apply the traditional NMS algorithm on the sequence set to reduce the redundant sequences. The algorithm for sequence proposal reduction is illustrated in Alg. 1. The IoU threshold θ is set to 0.5 in our experiments.

Algorithm 1: Sequence Proposals Reduction

Input: Input sequence set $S = \{S_k^o\}$;
Its classification score $C(S) = \{C(S_k^o)\}$;
Its mask sequence set $M(S) = \{M(S_k^o)\}$;
where $k = 0, 1, ..., K-1$ and $o = 0, 1, ..., O-1$.
IoU threshold θ.
Output: Final sequence set $\widehat{S}$
$\widehat{S} \leftarrow \{\}$;
while $S \neq \varnothing$ **do**
  $(k', o') \leftarrow$ argmax $C(S)$;
  $V \leftarrow S_{k'}^{o'}$;
  $\widehat{S} \leftarrow \widehat{S} \cup V$; $S \leftarrow S - V$;
  for S_k^o *in* S **do**
    if *IoU*$(M(V), M(S_k^o)) \geq \theta$ **then**
      $S \leftarrow S - S_k^o$;
      $C(S) \leftarrow C(S) - C(S_k^o)$;
      $M(S) \leftarrow M(S) - M(S_k^o)$;
    end
  end
end
return $\widehat{S}$;

3. Seq-Prop Head

Architecture The detailed architecture of the Seq-Prop head is shown in Fig. 1.

Soft-Agg The soft aggregation [6] of estimated mask $\tilde{M}_q(p)$ is defined as

$$M_q^o(p) = \frac{\tilde{M}_q^o(p)/(1-\tilde{M}_q^o(p))}{\sum_{i=0}^{O} \tilde{M}_q^i(p)/(1-\tilde{M}_q^i(p))}, \tag{1}$$

where $\tilde{M}_q^0(p) = \prod_{i=1}^{O}(1-\tilde{M}_q^i(p))$ denotes the background prediction.

Training Loss With $\hat{M}_q$ and M_q denoting the ground-truth and predicted masks, the scale-balanced soft IoU loss [4] is defined as

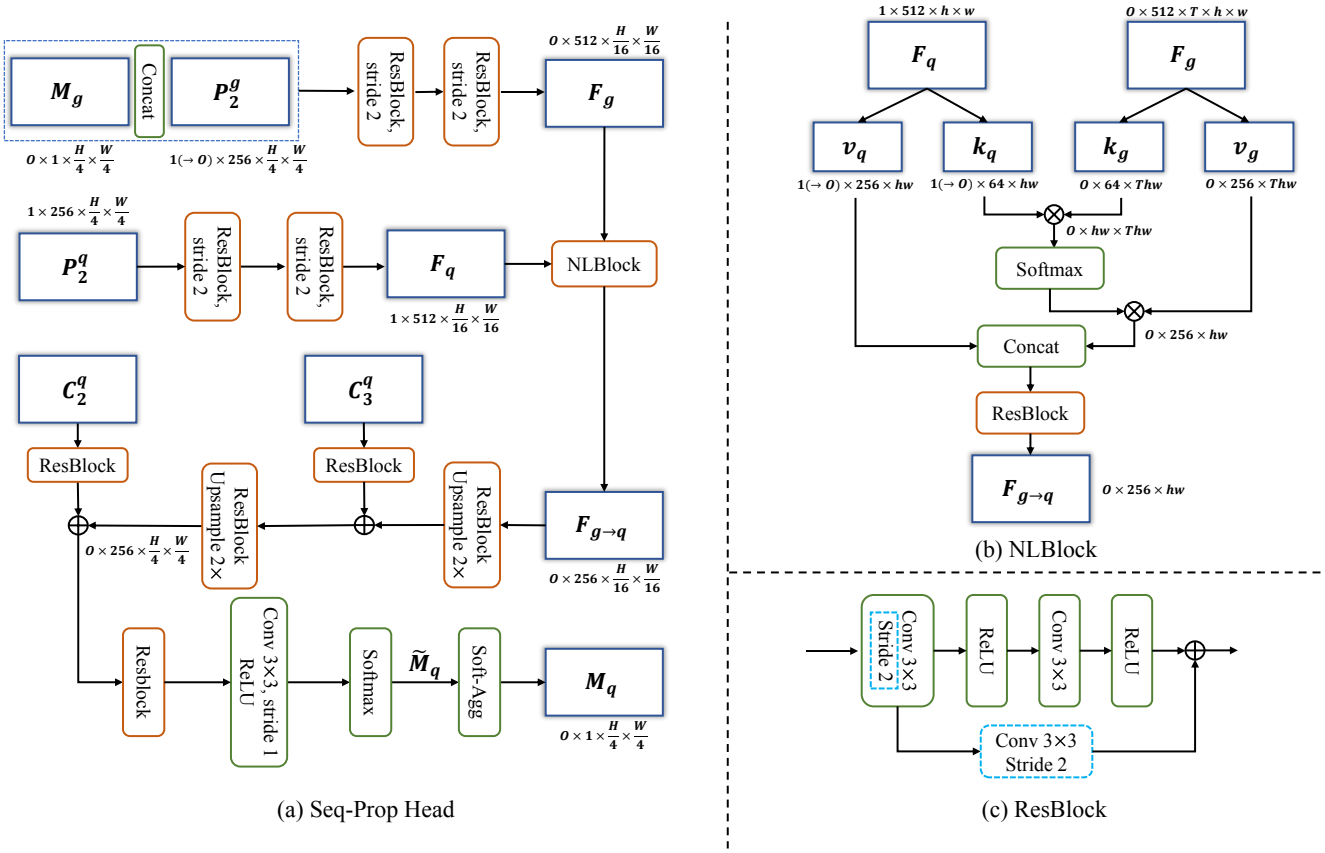
$$\mathcal{L}(\hat{M}_q, M_q) = 1 - \frac{1}{O}\sum_{o=1}^{O} \frac{\sum_p min(\hat{M}_q^o(p), M_q^o(p))}{\sum_p max(\hat{M}_q^o(p), M_q^o(p)))}, \tag{2}$$

where $M_q^o(p)$ denotes the value of the o^{th} instance in query mask M_q at pixel p and so as $\hat{M}_q^o(p)$.

4. Implementation Details

Training We follow the training setup as in [2]. We train our model for 4 epochs in the main-training stage and 5 epochs for the finetuning stage. In the main-training stage, we adopt the SGD optimizer with an initial learning rate of $5e$-3. The learning rate decays by a factor of 10 at the 3 and 4 epochs. In the finetuning stage, the learning rate is fixed at $5e$-5. The batch size is set to the maximum possible magnitude for different backbones. Our model is initialized with the pre-trained weight of Mask R-CNN [2] on COCO, while the additional propagation head is initialized randomly.

Inference During the testing stage, RPN generates 200 proposals for each key frame. For a key frame, the detected instances are sorted by score and the top 10 (*i.e.*, O) ones with scores higher than 0.2 are used for generating sequence proposals. The memory pool is updated every 5 frames in the Seq-Prop head.



(a) Seq-Prop Head

(b) NLBlock

(c) ResBlock

Figure 1. Architectures of the (a) **Seq-Prop head**, including the (b) NLBlock [5] and the (c) ResBlock [3]. O, T, H and W indicate instance number, frame number, height and width respectively. '$(\rightarrow O)$' denotes expanding the tensor along the specific dimension. The 'Soft-Agg' operation refers to Eq. (1).

5. Discussion

Comparison with MaskProp Since MaskProp [1] does not release codes or pre-computed results, a qualitative comparison is infeasible. Nevertheless, Tab. 1 (in the main paper) can give some hints about the difference between MaskProp and our method. In Tab. 1, when the $\mathcal{AP}$ is close (47.6 *vs*. 46.6), our method has a better $\mathcal{AR}$@10 than MaskProp (56.0 *vs*. 52.6). It indicates our method has more true positives in the top-10 scoring instances. However, with higher recall, they have similar $\mathcal{AP}$. This suggests that MaskProp has the better scoring (according to the rules of mAP), which may be because of the stronger backbone employed (i.e., STSN-ResNeXt-101). When using the same backbone (ResNeXt-101), our method is better than MaskProp by 3.3% in terms of $\mathcal{AP}$.

Distant Frame Pairs During the inference stage, the Seq-Prop head propagates segmented masks (in key frames) to other frames. It is worth investigating the effectiveness of propagating to distant frames. To this end, we group sequence proposals into three types with different initial quality (*i.e.*, IoU) at the starting key frame (Fig. 2). The IoU of propagated masks drops by around 20% with high-quality initial masks. As the initial mask quality lowers, the IoU at distant frames drops less and even rises. This may be due to the reason that the propagation head learned shape information for objects.

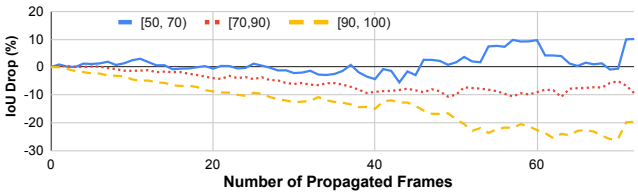


Figure 2. IoU (*i.e.*, $\mathcal{J}$-Mean) drops regarding propagated distance on DAVIS with ResNeXt-101. '$[a, b)$' indicates the group of sequence proposals where the IoU between the initial mask and corresponding ground-truth is in the range of $[a\%, b\%)$.

6. Visual Results

We provide more visual results in Fig. 3. The last row is a failure case, where the deer is misclassified as a 'fox'.

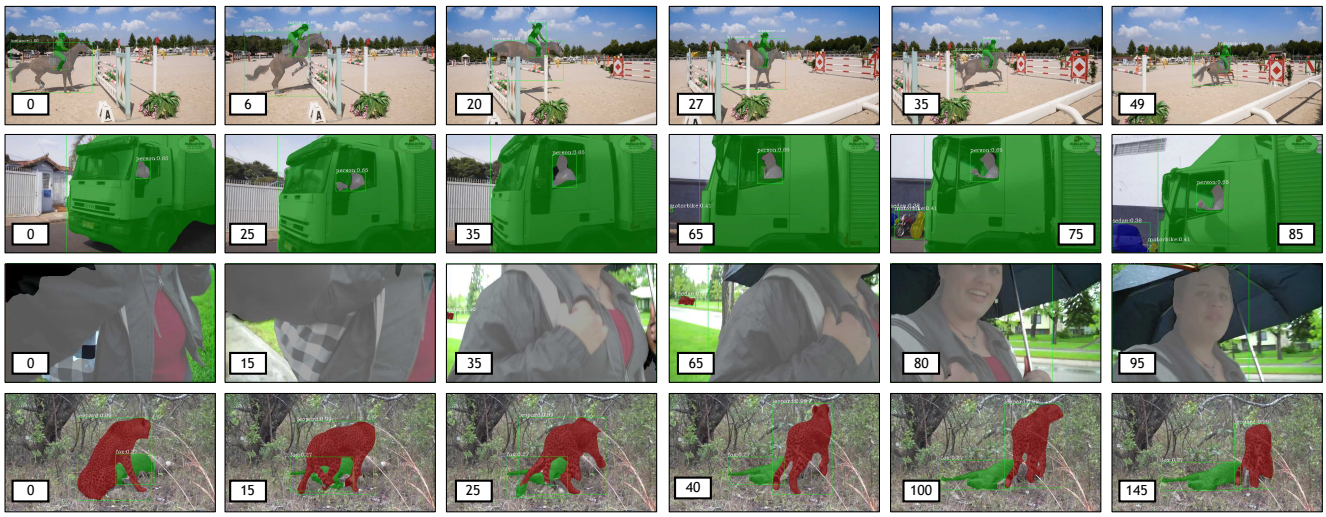


Figure 3. Visual results in various scenarios on DAVIS-UVOS and YouTube-VIS validation set. Category 'Instance' in DAVIS-UVOS denotes the salient generic object. The last row is a failure case. Zoom in for details.

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[4] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *ICCV*, 2019. 1

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[6] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018. 1