Supplementary Material: In Defense of Online Models for Video Instance Segmentation

Junfeng Wu[1]* , Qihao Liu[2]*, Yi Jiang[3], Song Bai[3]†, Alan Yuille[2], Xiang Bai[1]

[1] Huazhong University of Science and Technology
[2] Johns Hopkins University [3] ByteDance

Abstract. Sec. 1 provides implementation details about model settings, training and inference settings. Sec. 2 provides additional ablation experiments on OVIS. Sec. 3 gives additional qualitative results compared with other SOTA methods, demonstrating the efficiency of IDOL in three aspects. Sec. 4 and Sec. 5 show the improvement of our optimal transport and temporally weighted softmax, respectively.

1 Implementation Details

Model settings. We use ResNet-50 as our backbone unless otherwise specified. For a fair comparison with SOTA offline method, we use the same setting for Deformable DETR and the dynamic mask head following SeqFormer. For the transformer, we use 6 encoders, 6 decoder layers of width 256 with bounding box refinement mechanism, and the number of object queries is set to 300.
Training. We use AdamW optimizer with base learning rate of 1×10^{-4}, and weight decay of 10^{-4}. We first pre-train the model on COCO for instance segmentation following previous works. Then we train our model for 12000 iterations, on the corresponding training set and reduce learning rate by a factor of 10 at at the 8000 iterations. For the result with superscript "†", we randomly and independently crop the image from COCO twice to form a pseudo key-reference frame pair, which is used to pre-train the contrastive embedding of our models before training on video datasets. For YouTube-VIS 2019 and YouTube-VIS 2021, the input frames are downsampled and randomly cropped so that the longest side is at most 768 pixels. For OVIS, we use the same scale augmentation with COCO, resizing the input images so that the shortest side is at least 480 and at most 800 pixels while the longest is at most 1333. The model is trained on 8 V100 GPUs of 32G RAM, with 2 pairs of frames per GPU.
Inference. During inference, the input frames are downscaled to 360p for YouTube-VIS 2019 and YouTube-VIS 2021 following previous work, and 720p for OVIS as its videos has a higher resolution. For the hyper-parameters of temporally weighted softmax, we set $\tau = 0.5$ and $\mathrm{T} = 3$ by default.

* First two authors contributed equally. Work done during an internship at ByteDance.
† Corresponding author

2 Ablation Study

In this section, we provide extensive ablation experiments to study the importance of the core factors of our method on OVIS. As shown in Table 1, contrastive training increases AP from 11.0 to 18.4, an improvement of 67.3%. This indicates that embedding-based association is more robust in longer videos and complex scenarios. Compared with "multi-cues", our embedding association strategy improves the AP from 18.4 to 26.7. In addition, when temporally weighted softmax is added, it can be further improved by 1.9.

Table 1. Ablation study on contrastive learning and inference strategy on OVIS. Medium and heavy denote the AP and AR of objects moderately occluded, and heavily occluded, respectively.

Training			Inference		AP			AR	
ID Head	Contrastive	OT	Matching	Temporal	All	medium	heavy	medium	heavy
✓	-	-	multi-cues	-	11.0	11.9	2.3	16.4	7.6
-	✓	-	multi-cues	-	18.4	22.5	5.8	34.9	14.9
-	✓	-	embeddings	-	26.7	30.9	9.5	43.2	19.9
-	✓	✓	embeddings	-	28.3	34.0	9.8	44.5	20.3
-	✓	✓	embeddings	✓	30.2	36.5	10.3	46.9	20.5

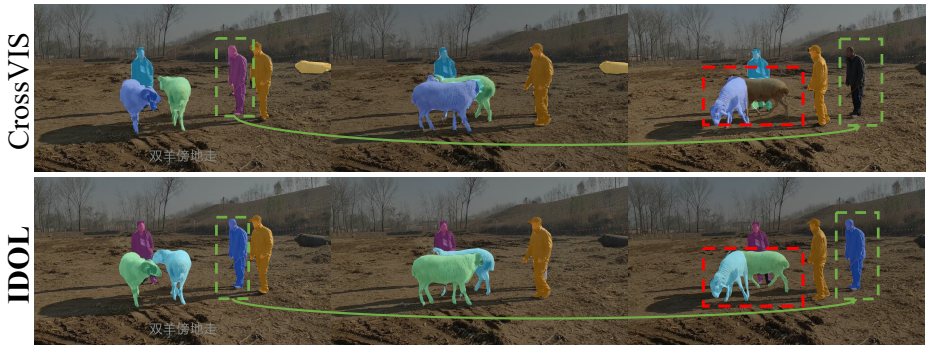
3 Qualitative Results

In this section, we show several qualitative results on the validation sets of YouTube-VIS and OVIS to demonstrate the following advantages of IDOL:

- For instances that belong to the same category and have very similar appearances, our contrastive learning enables IDOL to segment and track these instances more accurately. (*e.g.* Fig. 1)
- Our method learns embedding with better temporal consistency, benefiting the tracking in videos with high-speed, large, and/or complex motions. (*e.g.* Fig. 2)
- With the help of more stable and discriminative embeddings, as well as our one-to-many temporally weighted softmax during inference, IDOL is more robust when handling crowded scenes with heavy occlusions and frequent position exchanges. (*e.g.* Fig. 3)

4 Optimal Transport

Given a ground truth bounding box of an instance, the IoU-based method selects positive and negative samples by a hand-craft IoU threshold setting. A predicted box is defined as positive to an instance if they have an IoU higher than 0.7, or negative if they have an IoU lower than 0.3, which introduces false positives in



(a) YouTube-VIS 2019 / vid-28

(b) OVIS / vid-48

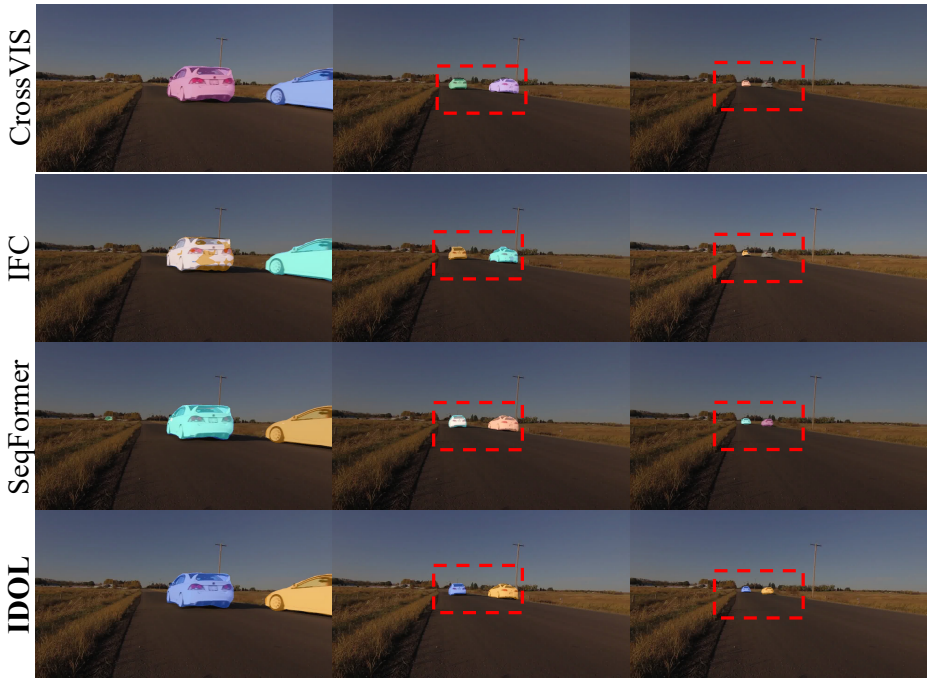
Fig. 1. Qualitative comparisons on videos with similar instances. Such kind of case is rare in YouTube-VIS, therefore we only select videos from OVIS. All methods use ResNet-50 backbone. Different color represents different instance id. Compare with the previous SOTA method, IDOL is able to segment and track instances with very similar appearances under complex motion and occlusions.

occlusions and crowded scenes. As shown in Fig. 4 (a), in the case of occlusion between two pandas, IoU-based method would take the boxes belonging to the panda in the back as the positive samples of the front one, which causes false positives. To address it, we formulate the problem of sample selection as an Optimal Transport problem in Optimization Theory, which reduces false positives and further improves the quality of the embedding. For each ground truth, we sum the top 10 IoU values to get $m1$ and the top 100 IoU values to get $m2$. Then we take top $m1$ predictions with the lowest cost as positive and top $300 - m2$ predictions with the highest cost as negatives. As shown in Fig. 4 (c), the optimal transport provides a better selection of positive embeddings during training, and thus improves the quality of the embedding.

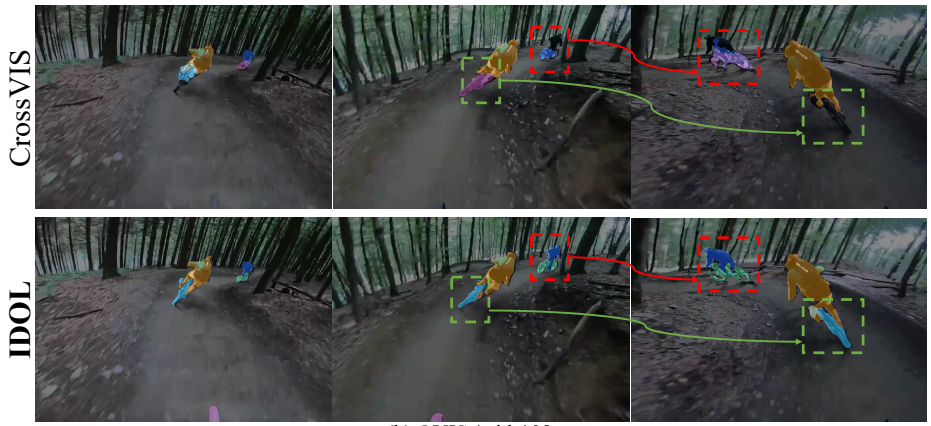
5 Temporally Weighted Softmax

In Fig. 5, we show qualitative results of the temporally weighted softmax in our association strategy. As shown in Fig. 5 (a) and (b), the bear with 'id:1' in (a)

is occluded by another bear in some frames, and without temporally weighted softmax, it is assigned a new id when it reappears. As shown in Fig. 5 (c), the people with 'id:3' and elephant with 'id:0' disappear in the corner of the video, but they swap ids when they reappear after several frames, and this leads to classification errors. However, in Fig. 5 (d), temporally weighted softmax helps maintain temporal consistency of id for the sampe people and elephant.

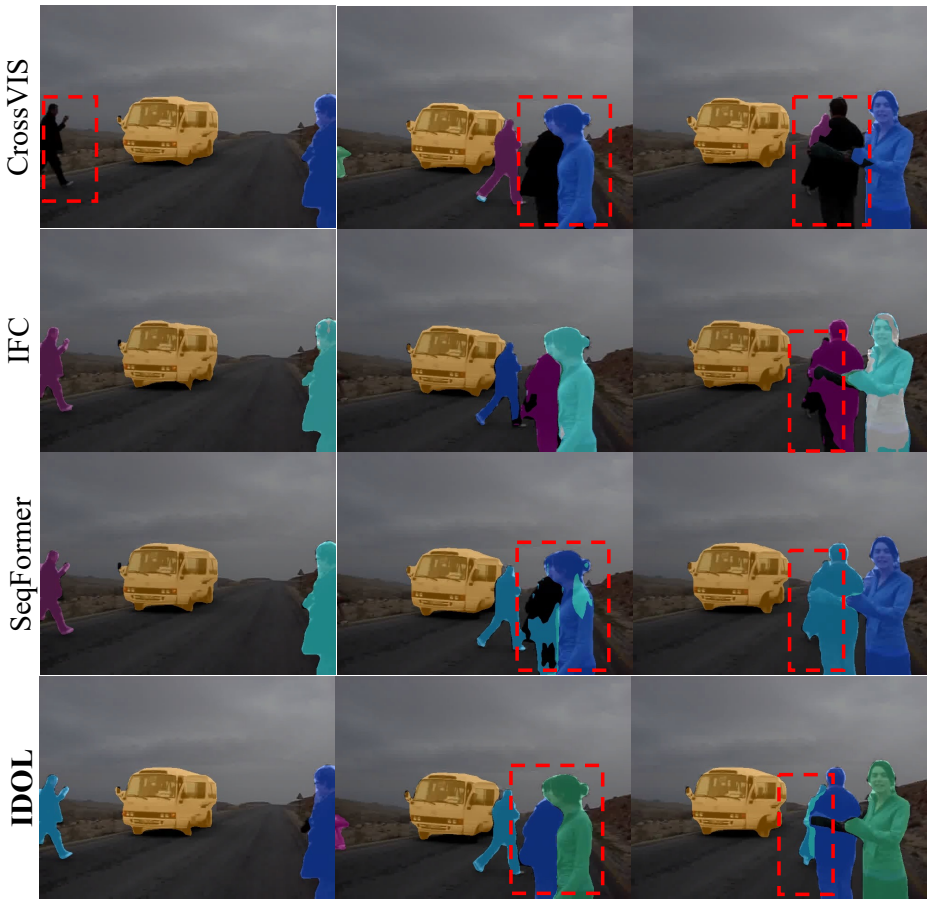


(a) YouTube-VIS 2019 / vid-109

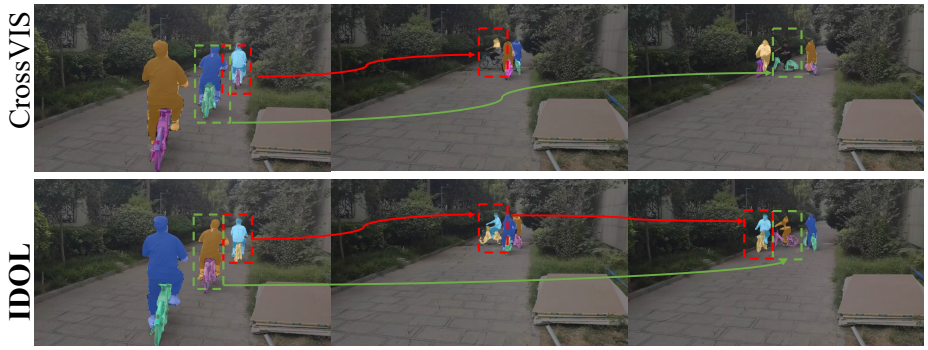


(b) OVIS / vid-109

Fig. 2. Qualitative comparisons on videos with complex motions. We don't show the results of offline methods (IFC, SeqFormer) on OVIS since they do not provide official code/models on OVIS and the clip matching method provided by IFC fails in complex cases. All methods use ResNet-50 backbone. Different color represents different instance id. Compare with the previous SOTA methods, IDOL performs much better on videos with high-speed and large motions (a), and complex motions (b).

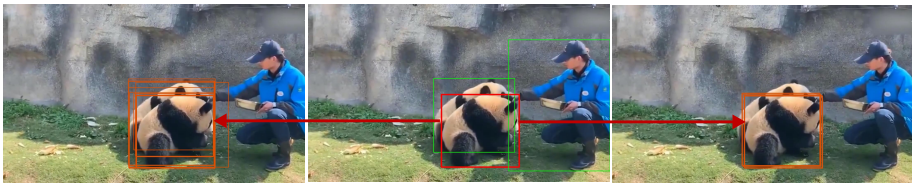


(a) YouTube-VIS2019 / vid-22



(b) OVIS / vid-23

Fig. 3. Qualitative comparisons on videos with severe occlusions. All methods use ResNet-50 backbone. Different color represents different instance id. Compare with the previous SOTA methods, IDOL is more robust when handling crowded scenes with severe occlusions and frequent position exchanges.



(a) IoU-Based (b) GT Box (c) Optimal Transport

Fig. 4. Visualization of positive samples selected by IoU-based method (a) and our optimal transport method (c). The panda with red bounding box in (b) is the key instance. The positive samples selected by the IoU-based method are shown in (a), which causes false positives (*i.e.*, the orange bounding box belonging to the panda behind the key instance). The positive samples selected by our method are shown in (c). It gives more accurate samples for positive embeddings and reduces false positives, further improving the quality of the embedding and the performance.

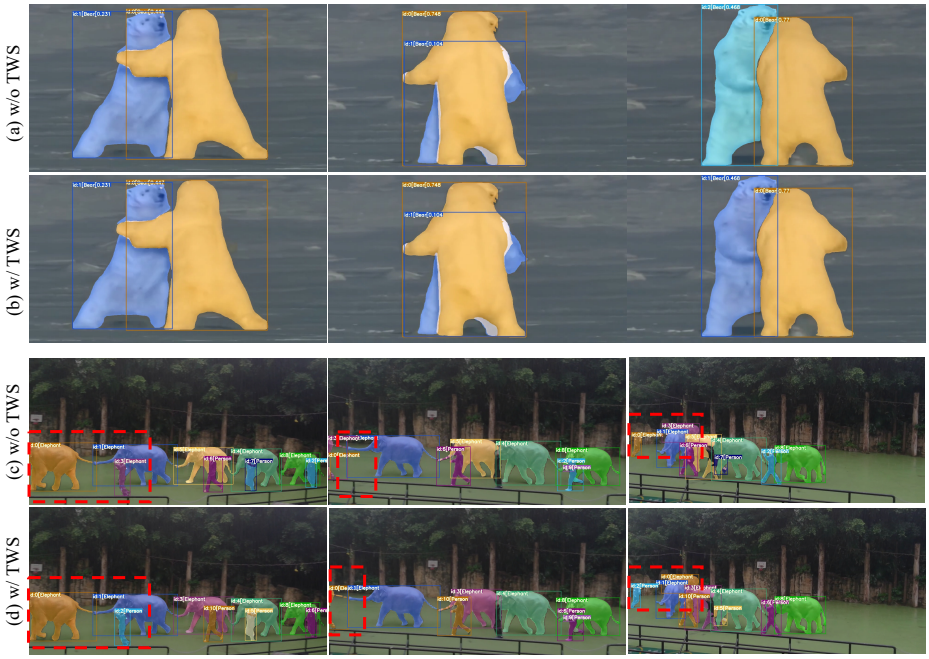


Fig. 5. Visualization of association quality with/without temporally weighted softmax (TWS). Each row shows three adjacent frames from the same video. (a) and (c) show the association quality without temporally weighted softmax. (b) and (d) show the association quality with temporally weighted softmax. The bear with 'id:1' in (a) is occluded by another bear in some frames, and it is assigned a new id when it reappears. When the people with 'id:3' and elephant with 'id:0' in (c) disappear in the corner of the video and reappear after several frames, they are also assigned new ids. However, this problem is solved in (b) and (d) by our one-to-many temporally weighted softmax during inference.