Video Salient Object Detection via Contrastive Features and Attention Modules

Yi-Wen Chen[1] Xiaojie Jin[2*] Xiaohui Shen[2] Ming-Hsuan Yang[1]
[1]University of California at Merced [2]ByteDance AI Lab

In this supplementary document, we provide additional analysis and experimental results, including 1) more implementation details, 2) ablation study for hard sample mining, 3) visual comparisons with state-of-the-art methods for video salient object detection on the FBMS [1] and ViSal [9] datasets, and 4) visual comparisons with state-of-the-art methods for unsupervised video object segmentation on the DAVIS [6] dataset.

1. Implementation Details

For the experiments of the unsupervised video object segmentation task, we use the fully-connected conditional random fields (CRFs) [3] to refine the response and generate the binary segmentation map.

2. Ablation Study

In Table 1, we present the ablation study of hard positive and negative mining. We first show the results of the baseline model and the improved ones of the model trained with the non-local self-attention and cross-level co-attention modules. When we train the model with the contrastive loss L_{cl}, the performance is improved by a limited amount. By considering hard negative samples, the results are further improved. When we take both hard positive and negative samples into account, our model achieves the best performance, which demonstrates the effectiveness of our hard sample mining technique.

3. Qualitative Results for Saliency and Segmentation

We provide more qualitative results compared with state-of-the-art video salient object detection approaches [7, 2, 4]. The results on the FBMS [1] and ViSal [9] datasets are shown in Figure 1 and Figure 2, respectively. Compared with previous methods, the proposed model generates more accurate results with more detailed information.

In Figure 3, we present the visual comparisons with state-of-the-art methods [5, 8, 10] for the unsupervised video object segmentation task. The results show that our model is able to segment the object more accurately, while the segmentation masks of other approaches tend to include background contents or lose detailed information near boundaries. We also show segmentation results of multiple continuous frames in Figure 4. The proposed method can distinguish the foreground object from the background under challenging conditions such as fast motion, occlusion and complex background.

*Corresponding author.

References

[1] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 1

[2] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019. 1, 2, 3

[3] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 1

[4] Haofeng Li, Guanqi Chen, Guanbin Li, and Yu Yizhou. Motion guided attention for video salient object detection. In *ICCV*, 2019. 1, 2, 3

[5] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 1, 3

[6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1

[7] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 1, 2, 3

[8] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 1, 3

[9] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. 1

[10] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip H. S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019. 1, 3

Table 1. **Ablation study of the proposed method.** We show the effectiveness of each component in the proposed framework, including the attention modules, contrastive loss, and hard positive and negative mining.

attn modules	L_{cl}	hard positives	hard negatives	DAVIS			FBMS			ViSal			DAVSOD		
				maxF ↑	S ↑	MAE ↓	maxF ↑	S ↑	MAE ↓	maxF ↑	S ↑	MAE ↓	maxF ↑	S ↑	MAE ↓
				86.8	88.4	3.1	86.9	84.2	5.8	92.4	91.8	2.9	62.9	72.1	9.6
✓				89.3	90.7	2.1	90.6	89.1	4.3	94.5	93.6	1.9	65.4	74.4	8.7
✓	✓			89.9	91.1	1.8	91.0	89.8	3.8	94.8	94.0	1.7	65.9	75.0	8.4
✓	✓		✓	90.5	91.5	1.6	91.3	90.5	3.2	95.0	94.4	1.5	66.1	75.2	8.3
✓	✓	✓	✓	**90.9**	**91.8**	**1.5**	**91.5**	**90.9**	**2.6**	**95.1**	**94.7**	**1.3**	**66.2**	**75.3**	**8.3**

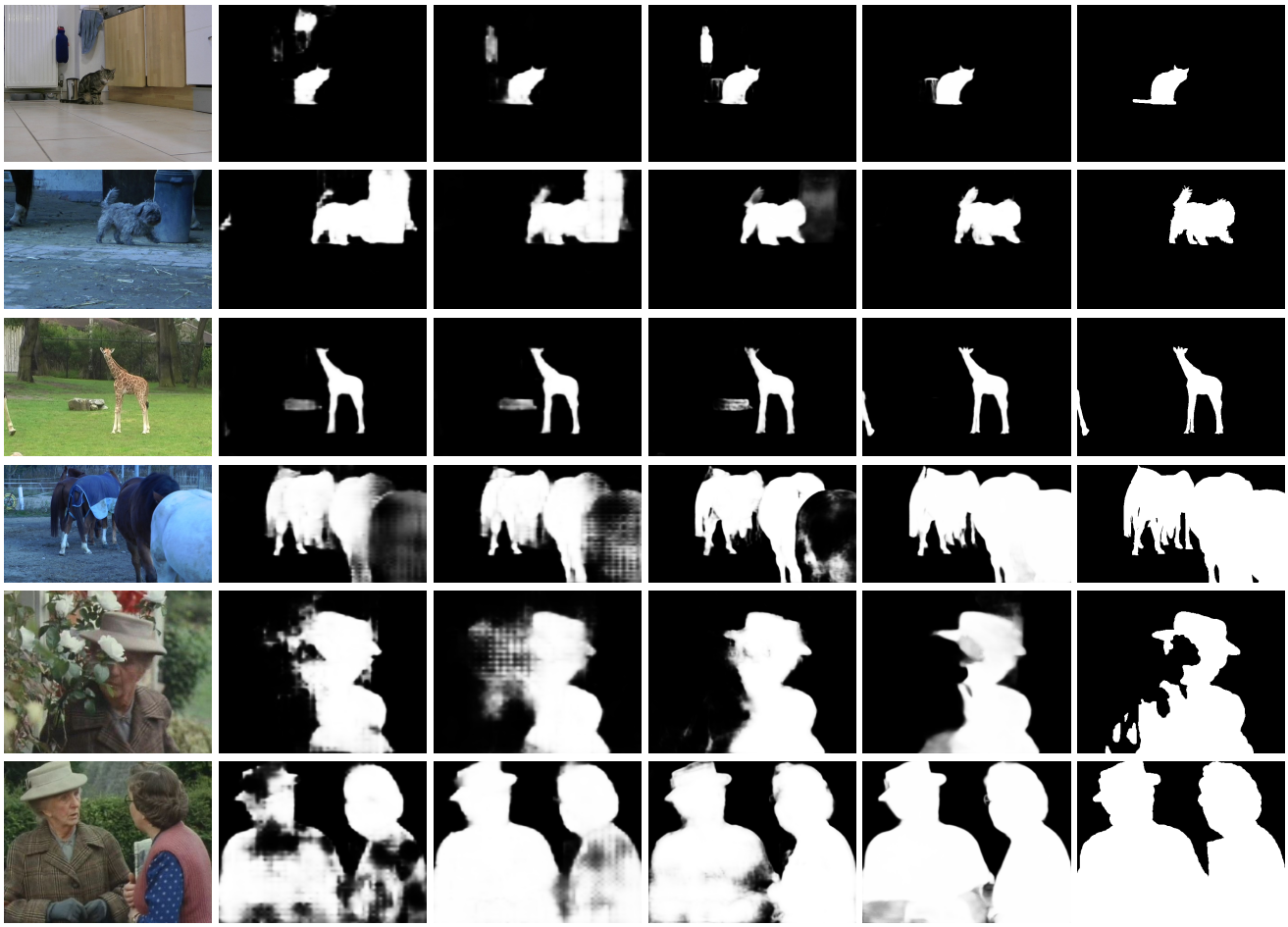

Figure 1. **Visual comparisons with the state-of-the-art video salient object detection methods on the FBMS dataset.** The ground truth masks (GT) are shown in the last column. The results by our method are more accurate and contain more details.

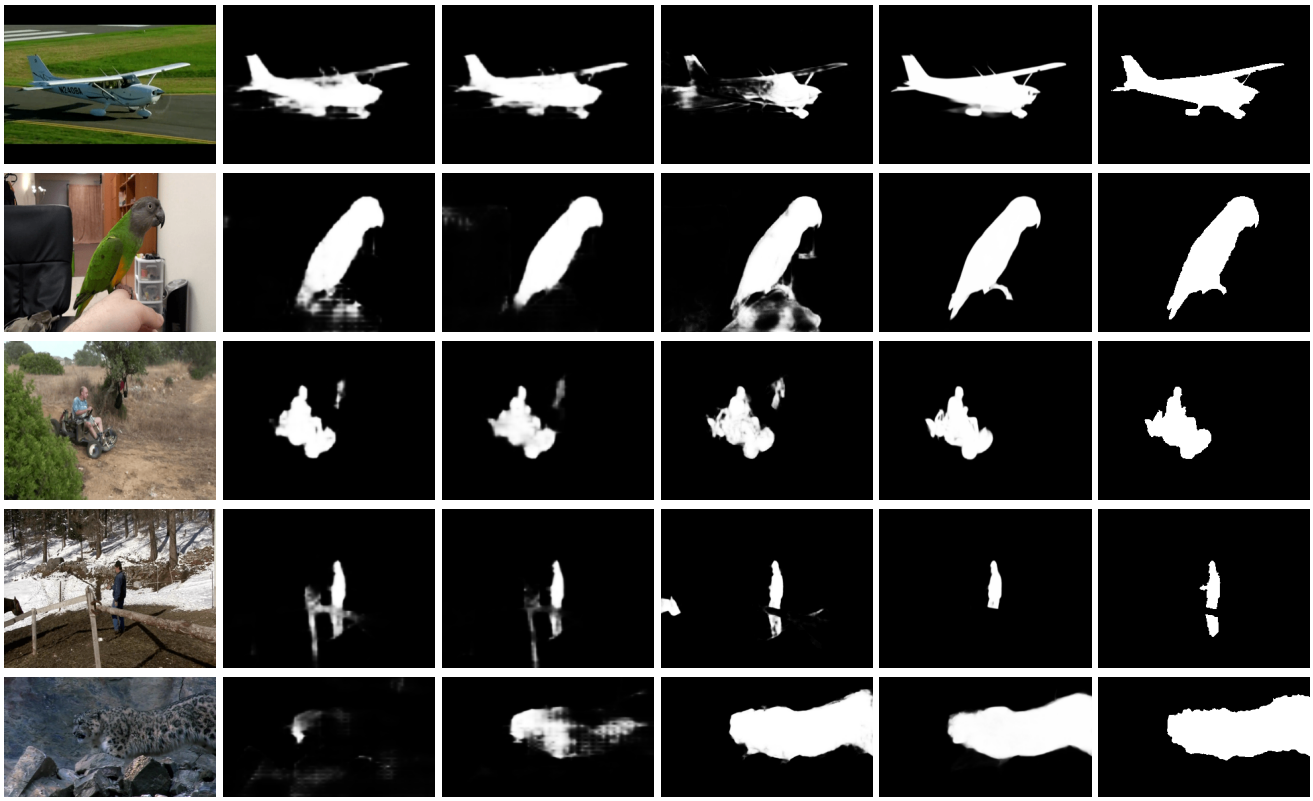

Figure 2. **Visual comparisons with the state-of-the-art video salient object detection methods on the ViSal dataset.** The ground truth masks (GT) are shown in the last column. The results by our method are more accurate and contain more details.

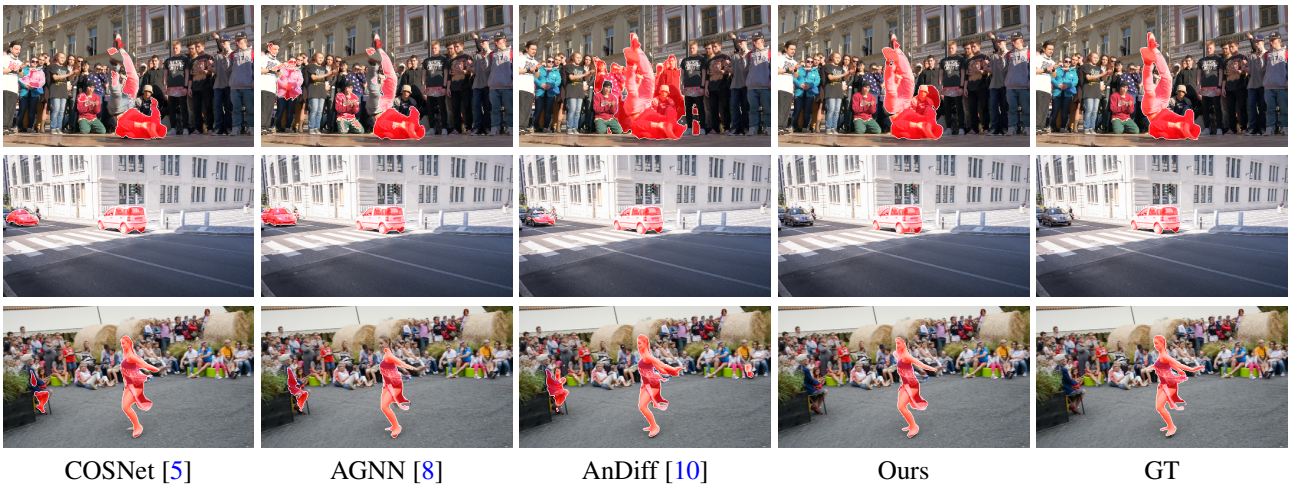

Figure 3. **Visual comparisons with the state-of-the-art unsupervised video object segmentation methods on the DAVIS dataset.** The ground truth masks (GT) are shown in the last column. The results by our method are more accurate and contain more details.

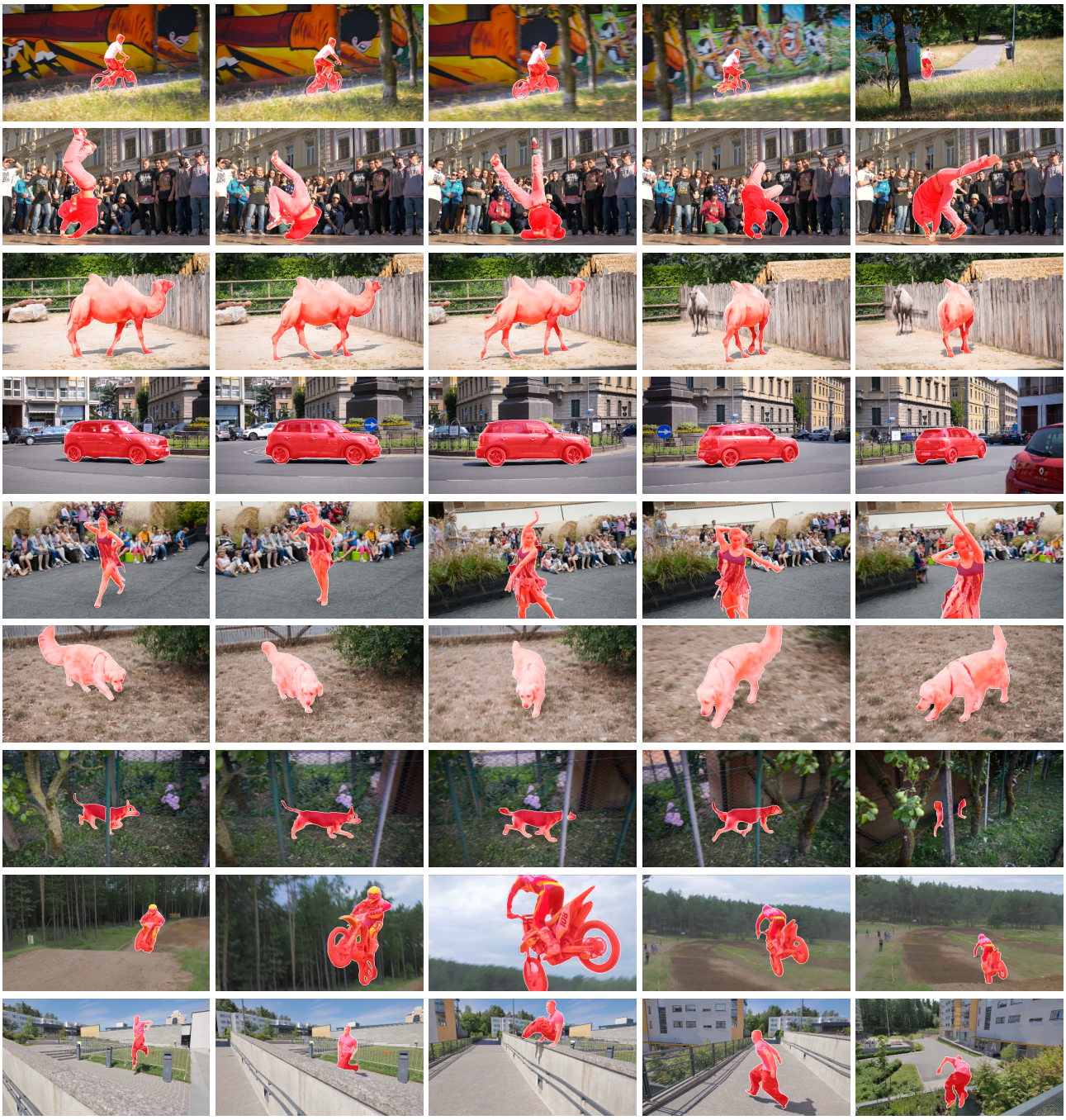

Figure 4. **Qualitative results of unsupervised video object segmentation on the DAVIS dataset.** The proposed model is able to generate accurate segmentation masks under challenging conditions such as fast motion, occlusion and complex background.