UVIS: Unsupervised Video Instance Segmentation
Supplementary Material

Shuaiyi Huang[1], Saksham Suri[1], Kamal Gupta[1*],
Sai Saketh Rambhatla[2*], Ser-nam Lim[3], Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Meta [3]University of Central Florida

In this supplementary material, we provide more detailed quantitative results and qualitative analysis of our method as follows: i) In Sec. A, we present F1-score statistics on the train set to assess the quality of pseudo-labels, in addition to the prototype memory filtering (PMF) ablation discussed in the main paper. ii) In Sec. B, we offer more insights into our implementation by providing details on class-agnostic mask generation, the prompts used for text-instance matching, and additional experimental details that complement the information provided in the main paper. iii) Sec. C showcases more qualitative results on Youtube-VIS 2019 [8], Youtube-VIS 2021 [8] and OVIS [5] validation set. For more qualitative video results, please refer to our uploaded video file in the supplementary material.

A. PMF Impact on Pseudo-Label Quality

To evaluate the impact of our PMF on pseudo-label quality, we conducted an analysis by computing the F1 score between the filtered pseudo-labels and the per-frame ground truth on the Youtube-VIS 2019 train set [8]. The F1 score provides insights into the removal of false positives while maintaining true positives. For evaluation, we consider a prediction as a true positive if its mask IoU with the corresponding ground truth mask is above 0.5. Table 1 summarizes the results, including per-class F1 scores and the averaged F1 score over all categories (mF1). Comparing the results to the case without filtering (28.6%), the score-based filtering (mask and CLIP score with a 0.7 threshold) improves the mF1 to 42.5%. With the integration of our PMF, we achieve a further improvement to 43.1%, obtaining the highest F1 score across the majority of classes. These findings indicate that our PMF enhances the quality of pseudo-labels, demonstrating its effectiveness in improving VIS results.

B. More Implementation Details

B1. Class-agnostic mask generation Our approach requires pseudo-labels that include both regions of interest and their corresponding labels. To generate possible object regions and their corresponding masks we use an existing off-the-shelf unsupervised approach called CutLER [6]. CutLER is trained in a unsupervised manner using coarse masks obtained from the self-supervised DINO [2] model for the ImageNet [3] dataset. These masks are then used to train a Cascade Mask R-CNN [1] backbone in a class agnostic manner. The trained detector referred to as CutLER shows good generalization in predicting masks and boxes around objects in our work. For each frame V_t, CutLER predicts a set of boxes $\{b_t^i\}$, masks $\{M_t^i\}$ and corresponding objectness scores $\{o_t^i\}$ where i corresponds to the i^{th} object instance in the frame. We use a threshold of 0.7 to filter out low confidence predictions for this step. More details about the training and generalizations of CutLER can be found in their paper [6].

B2. Prompts for CLIP-based Text-Instance Matching In CLIP-based Text-Instance Matching, an instance crop image is assigned a class label by computing the cosine similarity between the image embedding and a set of text prompts. The text prompts are generated given the dataset label set using simple string templates such as "a photo of $<$ class $>$". Multiple prompts per class are typically used to increase coverage. Specifically, the template "a photo of $<$ class $>$", along with the following six variations "a photo of $<$ class $>$ doing", "a photo of $<$ class $>$ moving", "a photo of $<$ class $>$ with", "a photo of $<$ class $>$ on", "a photo of $<$ class $>$ in", and "a photo of $<$ class $>$ at" are employed for each class. The model selects the closest matching prompt based on cosine similarity, thereby assigning the corresponding class label to the instance crop image.

B3. Architecture and Optimization In our implementation, we adopt Detectron2 [7] and adhere to the settings proposed in MinVIS [4] for video instance segmentation. Our chosen architecture consists of six multi-scale deformable attention Transformer (MSDeformAttn) [9] layers applied to feature maps at resolutions 1/8, 1/16, and 1/32. Additionally, we incorporate a simple upsampling layer with lateral connection to generate the final 1/4 resolution feature map, which serves as the per-pixel embedding. For the transformer decoder, we employ 9 layers and set the num-

*Work done while at UMD.

Filtering Methods	person	panda	lizard	parrot	skateboard	sedan	ape	dog	snake	monkey	hand	rabbit	duck	cat	cow	fish	train	horse	turtle	bear	**mF1**(%)
None	**18.3**	36.4	41.1	43.0	**1.7**	28.1	34.2	20.5	14.3	36.4	8.8	39.4	39.0	22.3	31.6	23.9	20.7	30.4	46.8	28.7	-
score-based	8.2	47.5	56.3	58.5	1.6	**35.8**	47.0	37.2	35.3	46.4	9.9	53.8	**52.3**	50.2	54.0	42.0	43.3	52.5	**63.9**	56.3	-
score-based + PMF	5.9	**47.9**	**57.6**	**59.8**	1.6	34.7	**47.2**	**40.0**	**39.7**	**47.0**	**10.0**	**53.8**	52.0	**51.9**	**55.6**	**43.1**	**45.7**	**53.2**	63.5	**57.3**	-

Filtering Methods	motorbike	giraffe	leopard	fox	deer	owl	surfboard	airplane	truck	zebra	tiger	elephant	snowboard	boat	shark	mouse	frog	eagle	seal	tennis racket	**mF1**(%)
None	18.7	43.2	44.0	44.6	20.4	51.0	1.5	22.6	30.0	35.2	44.4	43.9	**0.2**	20.0	22.7	26.0	37.5	33.4	31.5	6.4	28.6
score-based	35.4	**59.8**	**52.0**	**55.7**	**15.0**	**65.9**	**2.5**	46.5	48.8	58.2	**57.7**	63.0	0.1	29.2	40.8	50.1	**51.0**	58.2	49.7	8.5	42.5
score-based + PMF	**36.9**	59.6	51.4	55.2	14.7	65.4	2.4	**48.9**	**50.0**	**60.4**	57.2	**64.9**	0.1	**30.3**	**41.4**	**50.2**	49.6	**58.3**	**49.7**	**9.0**	**43.1**

Table 1. **Per-class and overall F1-score results for pseudo-labels filtering on the Youtube-VIS 2019 train set [8].** F1-scores are obtained using three different filtering methods: without any filtering (row 1), filtering by mask and clip threshold (row 2), and our prototype memory filtering (PMF) method (row 3). The best-performing results are highlighted in bold. Among the methods, our PMF approach achieves the highest mean F1-score across multiple classes, indicating its effectiveness in reducing false positives while preserving true positives.

ber of queries to 100 by default. During optimization, we assign a weight of 2.0 to the classification loss ($\mathcal{L}_{\text{cls}}$) and 5.0 to the segmentation loss ($\mathcal{L}_{\text{seg}}$). We utilize the AdamW optimizer with an initial learning rate of 0.0001 and employ a step learning rate schedule. In our unsupervised setup, we keep the backbone fixed. During inference, we retain the top 10 predictions for each video sequence.

C. More Qualitative Results

More qualitative results from the predictions of our UVIS on Youtube-VIS 2019 [8], Youtube-VIS 2021 [8] and OVIS [5] validation set, are shown in Figure 2, 1 and 3, respectively. For more qualitative video results, please refer to our uploaded video file in the supplementary material.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43 (5):1483–1498, 2019. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*, 2022. 1

[5] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *arXiv preprint arXiv:2102.01558*, 2021. 1, 2, 5

[6] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. *arXiv preprint arXiv:2301.11320*, 2023. 1

[7] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[8] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1, 2, 3, 4

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

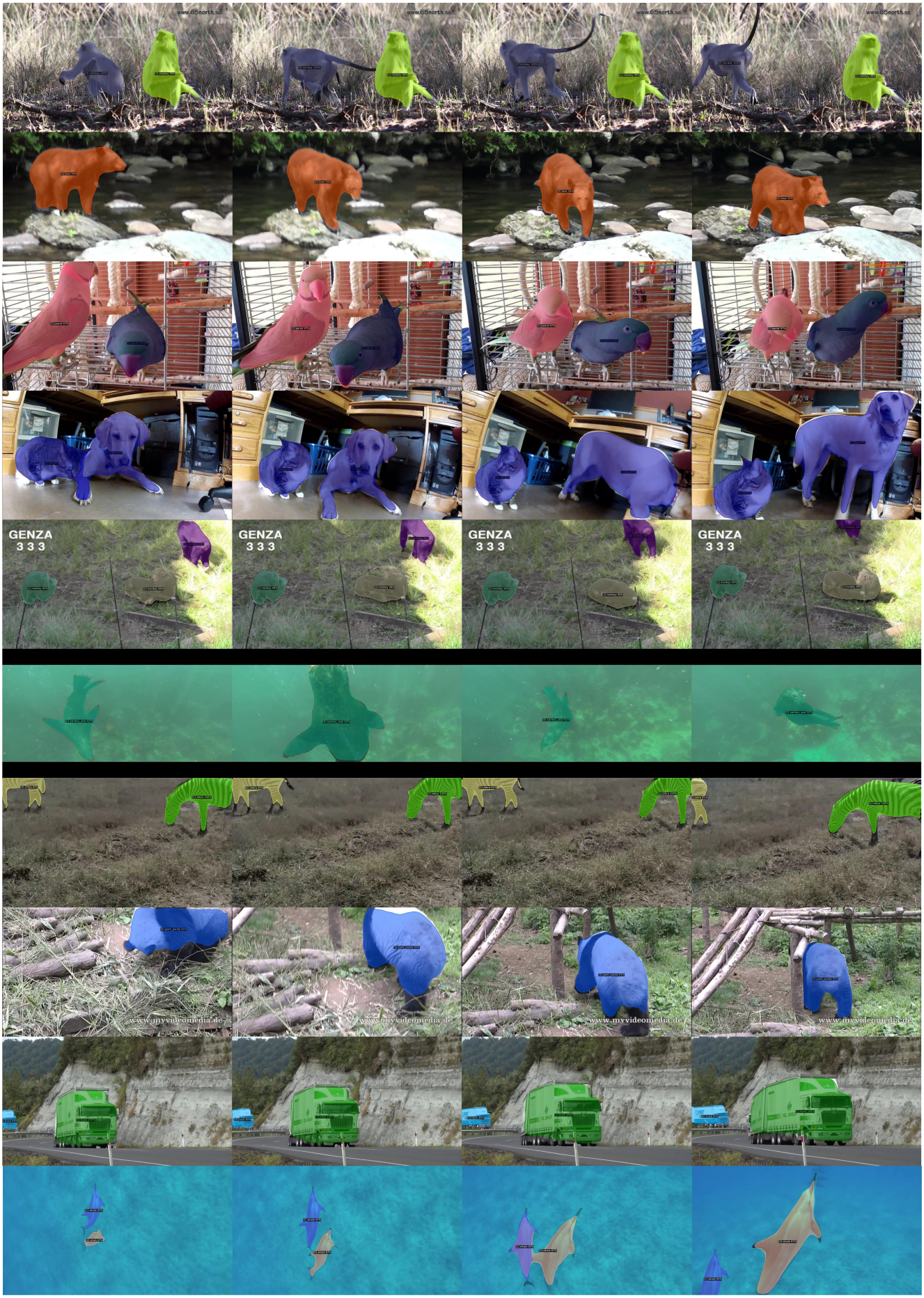

Figure 1. **Qualitative results of our UVIS on Youtube-VIS 2021 [8] validation set.**

Figure 2. **Qualitative results of our UVIS on Youtube-VIS 2019 [8] validation set.**

Figure 3. **Qualitative results of our UVIS on OVIS [5] validation set.**