Minimizing Labeled, Maximizing Unlabeled: An Image-Driven Approach for Video Instance Segmentation

Supplementary Material

A. More Implementation Details

Data Augmentation. As detailed in Section 3.3 of the main paper, instance association training is facilitated by augmenting an input image to create an image pair comprising the original image and its augmented counterpart. Table1 provides a summary of the data augmentation techniques.

Category	Augmentation	Value
Color	Brightness	$[-32, +32]$
	Contrast	$[0.5, 1.5]$
	Hue	$[-18, +18]$
	Saturation	$[0.5, 1.5]$
Affine Transformation	Rotation	$[-15°, +15°]$
	Translation	10%
	Scale	$[0.8, 1.2]$
	Shear	$[-5°, +5°]$

Table 1. Data augmentations applied for generating image pairs.

Hyper-Parameters. Table 2 provides an overview of the hyper-parameters used in MinMaxVIS.

Hyper-Parameter	Value
Number of Decoder Layers	9
Query Feature Dimension	256
Number of Attention Heads	8
FFN Dimension	2048
Number of Auxiliary Decoder Layers	6
Retrieval Threshold τ	0.99
Maximal Retrieval Number Per Category W	1,000
Truncation Weight β	0.5
Ratio of Labeled to Pseudo-Labeled	1:4
Classification Cost (Hungarian Matching)	2.0
Mask Loss Cost (Hungarian Matching)	5.0
Dice Loss Cost (Hungarian Matching)	2.0
Loss Weight (Classification)	2.0
Loss Weight (Mask)	5.0
Loss Weight (Dice)	2.0
Loss Weight (Association)	1.0
Inference Resolution	480p

Table 2. Summary of the hyper-parameters used in MinMaxVIS.

B. More Experiments

Consistent with the ablation studies in the main paper, we utilize MinMaxVIS with a Swin-L backbone on the YouTube-VIS 2019 dataset, employing only 2% labeled data. SA-1B acts as the unlabeled set.

Maximum Number of Pseudo-Labeled Images per Category. In Section 3.2 of the main paper, we propose a high-precision retrieval strategy to identify images containing instances of target categories. During implementation, we retain up to W pseudo-labeled images per category, prioritizing those with the highest confidence scores as predicted by our preliminary segmentation model. Additionally, instances within each pseudo-labeled image must meet a threshold condition: their confidence scores must exceed the predefined threshold τ. Table 3 examines the impact of varying W. Increasing the maximum number of pseudo-labeled images per category consistently improves the mAP, with the highest performance (62.2) achieved at $W = 1000$. We also observe that increasing W beyond this point does not yield further performance gains, indicating that the model effectively saturates in utilizing pseudo-labeled images when W reaches 1000. This suggests that retaining a larger number of pseudo-labeled images may introduce diminishing returns while potentially increasing computational overhead.

W	200	500	1,000
mAP	60.7	61.3	62.2

Table 3. Study on maximum number of pseudo-labeled images per Category (W).

Feature	mAP
Last Decoder Layer (Main)	62.2
Last Decoder Layer (Auxiliary)	61.1
MLP Projector (Auxiliary)	60.7

Table 4. Feature analysis for instance association.

Features for Instance Association. In Section 3.3, we introduce an auxiliary decoder designed to enhance intra-class feature differentiation through an instance association loss, which operates alongside the main decoder. This auxiliary decoder is supervised by both the instance association loss and segmentation loss, while the main decoder is supervised by the classification and segmentation losses. Additionally, gradients from the auxiliary decoder are propagated to the

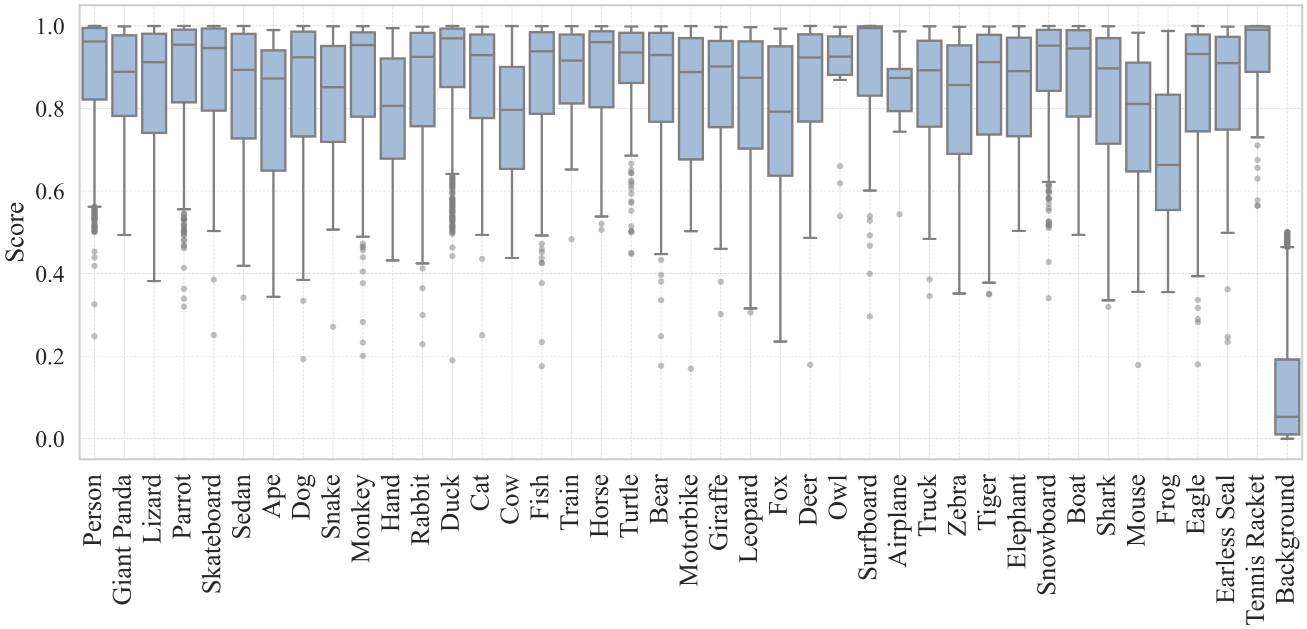

Figure 1. Score distribution of low-confidence background queries. The analysis is performed on all pseudo-labeled images from SA-1B. Each data point represents the maximum classification score of a specific low-confidence background query. For each category in YouTube-VIS 2019, we display the median, upper bound, upper quartile, lower bound, lower quartile, and outliers.

shallower layers of the main decoder. During inference, the auxiliary decoder is removed, and the features generated by the main decoder are used for instance association.

Here, we explore alternative feature strategies for instance association beyond the default approach. Specifically, the auxiliary decoder comprises multiple decoder layers followed by an MLP projector. We investigate the use of features from the last decoder layer and those produced by the MLP projector. A comparison of these strategies is presented in Table 4.

Instance association aims to track and link the same instance across frames. This process involves two key requirements. First, the same instance must belong to the same category. In the embedding space, this translates to instances of the same category being closer together than those of different categories. This is achieved through the classification and segmentation losses applied to the features from the main decoder. Second, intra-class feature differentiation is necessary to distinguish between different instances within the same category. This is facilitated by the gradients propagated from the instance association loss, which encourage variability among features within the same class. Together, these mechanisms enable robust and accurate instance association across frames. Consequently, features generated by the main decoder produce the best results, as the main decoder benefits from supervision provided by the classification loss, segmentation loss, and instance association loss.

Color	Affine Transformation	mAP
		59.8
✓		61.1
	✓	60.7
✓	✓	62.2

Table 5. Analysis of the effects of color and affine augmentations on the final performance.

Data Augmentations for Instance Association. Training for instance association relies on augmenting input images to generate an image pair consisting of the original image and its augmented version. Table 1 categorizes the applied data augmentations into two groups: color augmentations and affine transformations. Table 5 presents an analysis of the effects of color and affine augmentations on the final results.

C. Visualizations

Score Distribution of Low-Confidence Background Queries. In Section 3.3 of the main paper, we propose a selective gradient backpropagation technique to mitigate noise in pseudo-labeled images. Specifically, for each pseudo-labeled image, background queries are first identified, and only high-confidence background queries contribute to the training of background classification. Low-confidence background queries, defined as those with background scores below 0.5, may represent either true negatives or false negatives. To minimize training noise, gradients from these low-confidence queries are detached during training. The score distribution of these low-confidence background queries is visualized in Figure 1.

Pseudo-Labeled Instances. In Sections 3.1 and 3.2 of the main paper, we detail the process of training a preliminary segmentation model on a small labeled dataset (e.g., YouTube-VIS 2019 with 2% labeled data) to retrieve relevant instances from a large unlabeled dataset (e.g., SA-1B). For each retrieved instance, the model generates a pseudo-label comprising both a class label and a mask label. Figure 2 showcases two pseudo-labeled instances per category from YouTube-VIS 2019, with all instances sourced from the SA-1B dataset. The proposed high-precision retrieval strategy ensures the accuracy of both class labels and mask labels for the majority of pseudo-labeled instances.

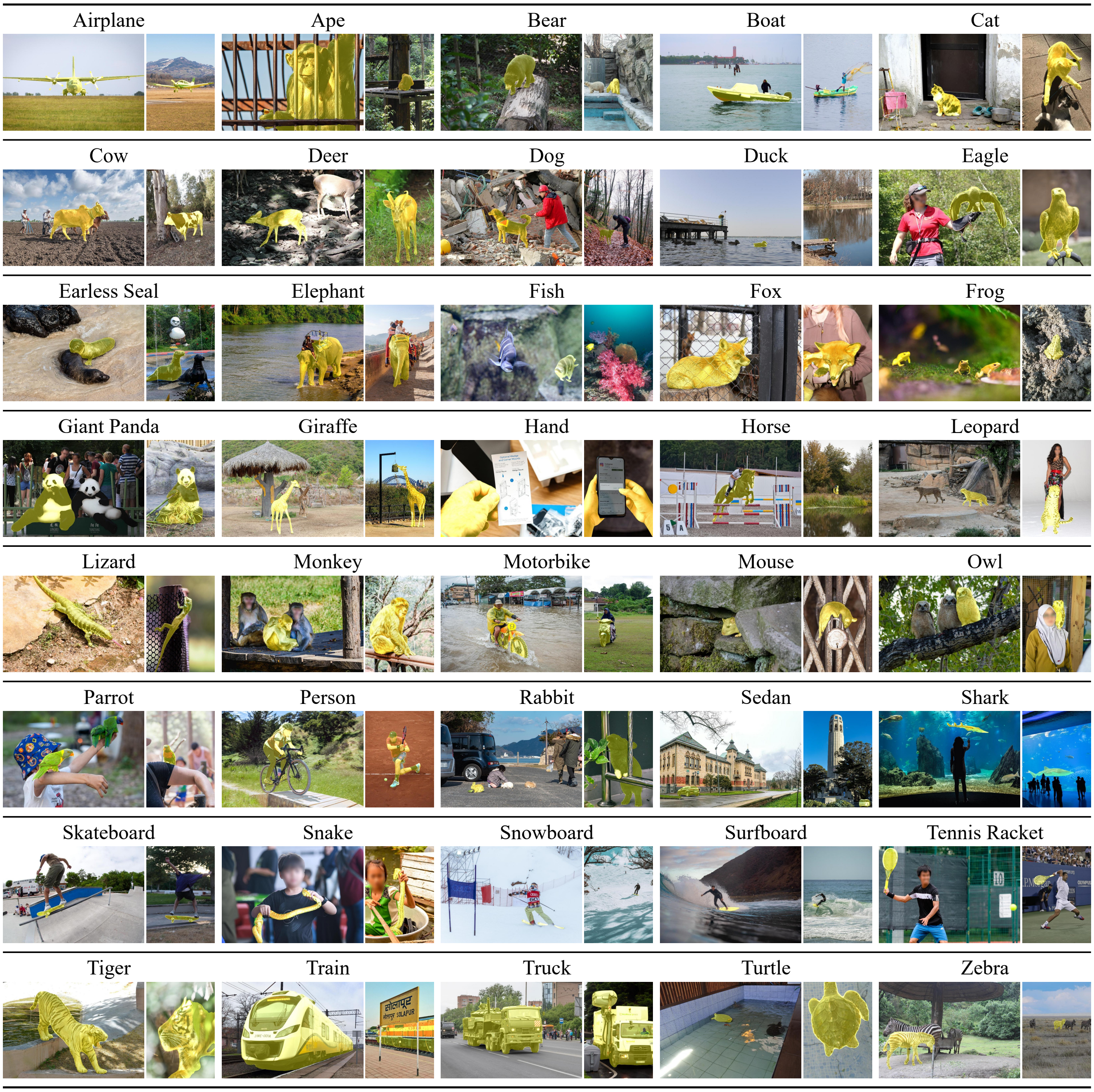

Figure 2. Visualization of the retrieved pseudo-labeled instances from the SA-1B dataset.