Supplement to " SSVOD: Semi-Supervised Video Object Detection with Sparse Annotations"

Tanvir Mahmud[1] Chun-Hao Liu[2] Burhaneddin Yaman[3] Diana Marculescu[1]
[1]University of Texas at Austin [2]Amazon Prime Video
[3]Bosch Research North America
{tanvirmahmud, dianam}@utexas.edu, chunhaol@amazon.com
burhaneddin.yaman@us.bosch.com

A. Overview

In this supplementary material, we provide additional details and experimental results for our proposed SSVOD. In summary, the following items are presented.

- Details of the mathematical notations used in the main paper.
- Details of the data augmentations used in model training.
- Additional ablation studies and performance analyses.
- Analysis of the training loss curves for SSVOD.
- Additional qualitative comparisons on different class of objects.

B. List of Notations

All the notations used in the main paper are summarized in Table 1.

C. Additional Implementation Details

Additional details on the implementations are summarized in Table 2.

D. Data Augmentations

In SSVOD, we use three sets of augmentations to process the labeled set, unlabeled set for the teacher, and unlabeled set for the student, respectively. The details of the augmentations used in SSVOD is presented in Table 3. For strong augmentation, we consider additional geometric augmentations and the cutout augmentation [1]. In contrast, weak augmentation only contains random flipping to reduce data variations for the teacher network to enable more confident pseudo-label generation. We primarily incorporate the augmentation schemes that are heavily used in semi-supervised image object detection [3, 4]. In SSVOD, the same augmentation parameters are maintained over each sequence of *key* and *reference* images in labeled and unlabeled sets. Since the *reference frames* are mostly used to enhance the *key* frame features, it is necessary to maintain the same augmentation choices on each set.

E. Additional Ablations

We perform ablation studies to compare the performance of the proposed SSVOD and baseline supervised approaches. We follow the single labeled *key frame* per video setting for the supervised baseline and the proposed SSVOD approach. All cases are evaluated on the ImageNet-VID [2] validation set and we report the mAP@0.5 score unless otherwise mentioned.

E.1. Per-Class Performance Analysis

We study the per-class performance to determine the improvement gain on the evaluation set. In Figure 1, we present per-class mAP scores obtained from the baseline supervised approach and the proposed SSVOD scheme. SSVOD significantly improves the mAP score for most of the classes. We observe large mAP improvements on the challenging classes such as *red panda* in which SSVOD achieves around 28 times higher mAP. We further present normalized confusion matrix on the class predictions of supervised and SSVOD approaches, which is shown in Figure 2. Our SSVOD approach achieves consistent performances on most classes by considerably improving the performance for the challenging classes. For example, the minimum accuracy on *lion* class with the supervised approach is nearly 0%, whereas with the SSVOD approach it rises to almost 20%. Thus, proper utilization of the unlabeled frames throughout the training video with the proposed SSVOD approach has great potential to improve the performance with scarce annotations.

E.2. Loss Curve Analysis

We present loss curves over training iterations as shown in Figure 3. In the supervised training, both the classification and bounding box losses gradually decrease until saturation. In the SSVOD training, we notice a similar behavior on the supervised losses. For unsupervised losses in SSVOD, the hard classification loss and bounding box loss start from zero since the model can't generate high confident pseudo-labels which leads to filtration of all the labels. On the other hand, the unsupervised soft loss initiates the training on the unlabeled sets. Gradually, the model generates highly confident pseudo-labels and the unsupervised losses continue to rise. Finally, both the supervised and unsupervised losses converge together as training progresses.

E.3. Performance Comparison on Objects of Different Sizes and Motions.

We study the performance on objects with different sizes and motion categories following [6]. The results are shown in Table 4. Recognizing smaller and faster objects is relatively more challenging. We observe that supervised accuracy is considerably low on small and fast objects. We achieve $+4.0$ and $+9.2$ mAP (@0.5:0.95) improvements on the middle and large objects, respectively. Accordingly, on the medium and slow objects, we notice consistent improvements of $+4.7$ and $+10.1$ points, respectively. However, the performance improvements are comparably smaller in the challenging small ($+2.1$ points) and fast ($+2.3$ points) objects.

E.4. Effect of Different Choices of Pseudo-Label Thresholds on performance.

We study the effect of different choices of thresholds in pseudo-label selection. We present the performance for combinations of threshold choices in Table 5. Best performance is achieved when the confidence threshold (γ_c) is set to 0.8 for soft class distillation, 0.9 for IoU threshold (ζ_{IoU}) in bounding box regression, and 0.005 for KL-divergence threshold (η_{div}) in hard-label classification. Higher values for these thresholds reduce the number of pseudo-labels for the unlabeled set, whereas lower values compromise the pseudo-label quality, resulting in confirmation bias [5]. We have carried out all other experiments with the best choice of these thresholds.

E.5. Additional Qualitative Results

We provide additional visualizations of the qualitative results obtained from SSVOD and the supervised scheme in Figure 4. As discussed earlier, SSVOD can learn the temporal variations of each object by utilizing sets of unlabeled images from different time steps in a video. We observe that SSVOD generates more stable class and bounding box predictions throughout the video without any post-processing. In contrast, the supervised approach generates inconsistent results which can not properly adapt to the temporal variations present in challenging conditions.

References

[1] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 3, 5

[3] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 4

[4] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 1, 4

[5] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 2

[6] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017. 2

Table 1. Different notations used to describe the operations of SSVOD. We categorize the notations into three types.

Type	Description	Notation
Scaler Parameters	number of videos	M
	number of frames per video	N
	number of *key frames* per video	n_k
	number of *reference frames* per video	n_r
	number of labeled *key frames* per video	n_k^l
	number of unlabeled *key frames* per video	n_k^u
	number of objects in the t^{th} frame	n^t
	cross-IoU threshold	ζ_{IoU}
	confidence threshold	γ_c
	cross-divergence threshold	η_{div}
Functions/Models	video object detection network	$Z_\theta(\cdot)$
	flow network	$\mathcal{F}(\cdot)$
	feature warping	$\mathcal{W}(\cdot)$
	cross-IoU estimator	$IoU(\cdot)$
	cross-KL divergence estimator	$D_{KL}(\cdot)$
Vectors/Matrix	*key frame* at timestamp t in the m^{th} video	K_m^t
	key frame at timestamp $t-i$ in the m^{th} video	R_m^{t-i}
	annotations of the t^{th} frame in the m^{th} video	y_m^t
	pseudo-label of the t^{th} frame in the m^{th} video	p_m^t
	key feature	f_k
	reference feature	f_r
	flow-warped *key* feature from the *reference* frame	$f_{r\to k}$
	Raw feature set at timestamp t	X_{raw}^t
	flow-warped feature set at timestamp $t+j$	X_{warped}^{t+j}
	predictions on raw feature set at timestamp t	P_{raw}^t
	class predictions on raw feature set at timestamp t	$P_{raw,cls}^t$
	bounding box predictions on raw feature set at timestamp t	$P_{raw,bbox}^t$
	filtered pseudo bounding boxes at timestamp t	P_{bbox}^t
	filtered pseudo hard-class labels at timestamp t	P_{cls}^t
	filtered pseudo soft-class labels at timestamp t	P_{soft}^t

Table 2. Implementation details on training and evaluation protocols of SSVOD. The provided values are chosen based on empirical study on ImageNet-VID dataset [2].

Variables		Value
Training	image size in pixels (height, width)	(1000, 600)
	number of *key* image/set	1
	number of *reference* images/set	2
	reference frame timestamp range shifted from *key frame*	[-9, 9]
	training iterations	40000
	labeled set/batch	1
	unlabeled set/batch	1
	optimizer	SGD
	learning rate	0.005
	EMA momentum	0.99
Evaluation	image size in pixels (height, width)	(1000, 600)
	number of *reference* images/set	30
	reference frame timestamp range shifted from *key* frame	[-15, 15]
	model	student

Table 3. Summary of the data augmentations used in SSVOD training. "-" denotes no augmentation is applied. We apply sequential augmentation on labeled and unlabeled sets where same augmentation parameters are used for each *key* and *reference frame* for a particular set. We extended the standard augmentations extensively used in semi-supervised image object detection [3, 4] to operate with videos.

Sequential Augmentation	Labeled set training	Unlabeled set training (strong augmentation)	Unlabeled set training (weak augmentation)
Random flip	$p = 0.5$, ratio $\in (0, 1)$	$p = 0.5$, ratio $\in (0, 1)$	$p = 0.5$, ratio $\in (0, 1)$
Contrast jitter	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Equalize jitter	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Solarize jitter	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Brightness jitter	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Sharpness jitter	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Random posterize	$p = 0.1$, ratio $\in (0, 1)$	$p = 0.1$, ratio $\in (0, 1)$	-
Translation	-	$p = 0.3$, ratio $\in (-0.1, 0.1)$	-
Rotation	-	$p = 0.3$, ratio $\in (-30, 30)$	-
Shear	-	$p = 0.3$, ratio $\in (-30, 30)$	-
Cutout	-	num $\in (1, 5)$, ratio $\in (0, 0.2)$	-

Table 4. Performance comparison with objects of different sizes and motions. Here, mAP@0.5:0.95 score is reported.

Method	Object Size			Motion		
	Small	Middle	Large	Fast	Medium	Slow
Supervised	6.7	16.5	34.5	13.8	23.9	35.8
Ours	**8.8**	**20.5**	**43.7**	**16.1**	**28.6**	**45.9**

Table 5. Ablation study on the effect of different choice of thresholds for selecting pseudo bounding box, hard, and soft-class labels.

γ_c	ζ_{IoU}	η_{div}	mAP	mAP@0.5	mAP@0.75
0.8	0.8	0.005	38.2	61.4	41.3
0.8	0.9	0.005	**39.2**	63.8	**43.1**
0.9	0.9	0.005	38.5	61.7	42.6
0.8	0.9	0.01	38.4	62.0	42.3
0.8	0.85	0.005	38.7	**63.9**	42.4

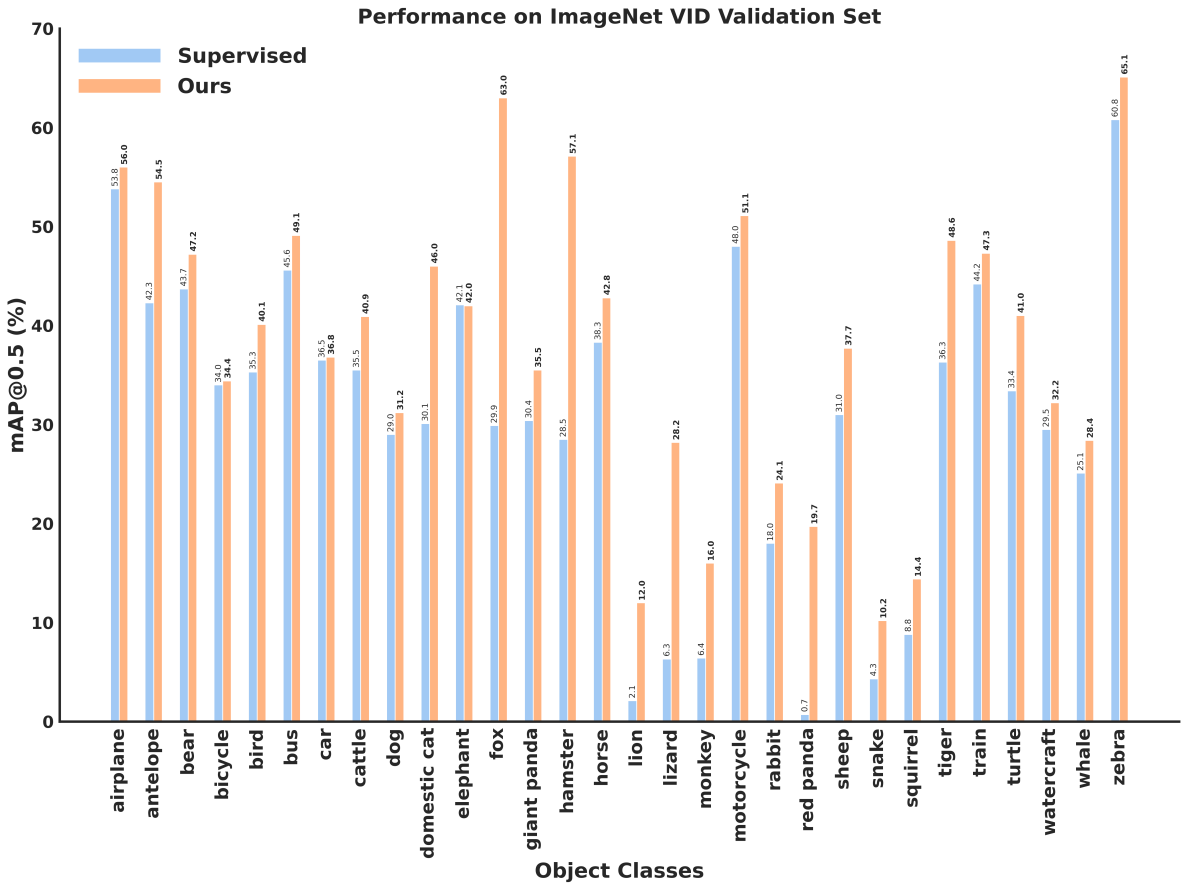
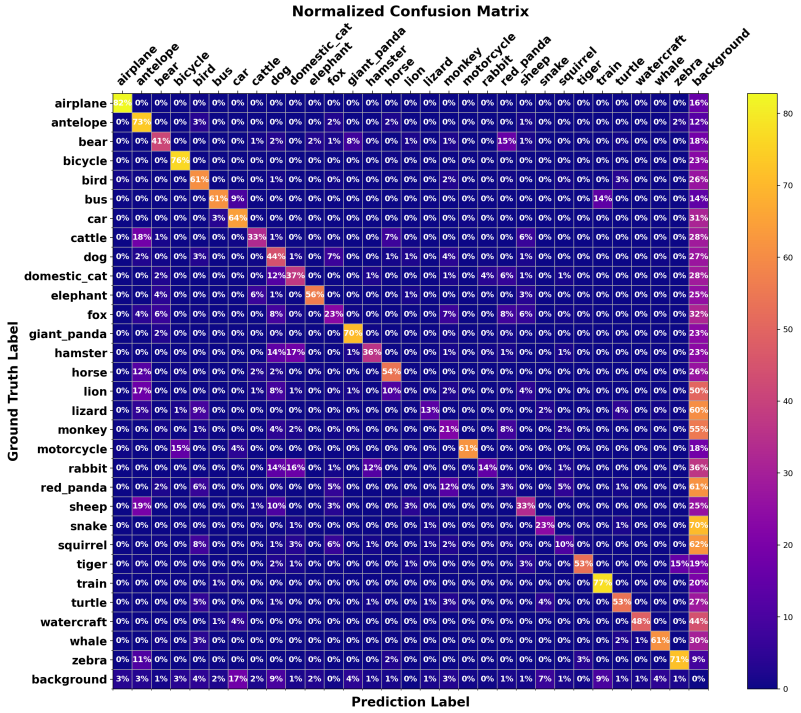


Figure 1. Per-class mAP performance comparison on the ImageNet-VID validation set [2] between the supervised and our proposed SSVOD approaches. In general, SSVOD significantly improves the supervised performance. Performance gain is more prominent on the challenging classes.

(a) Supervised



(b) SSVOD

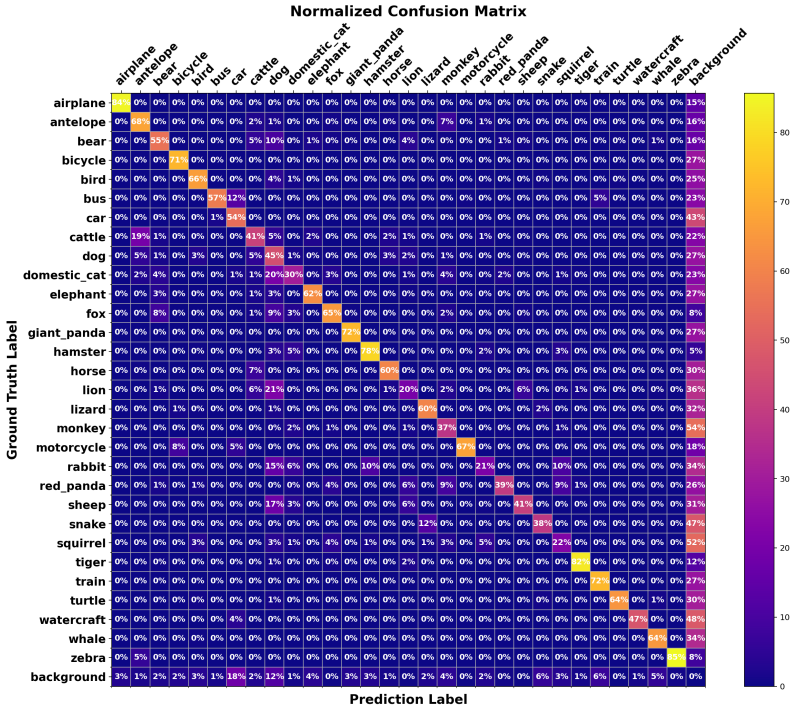
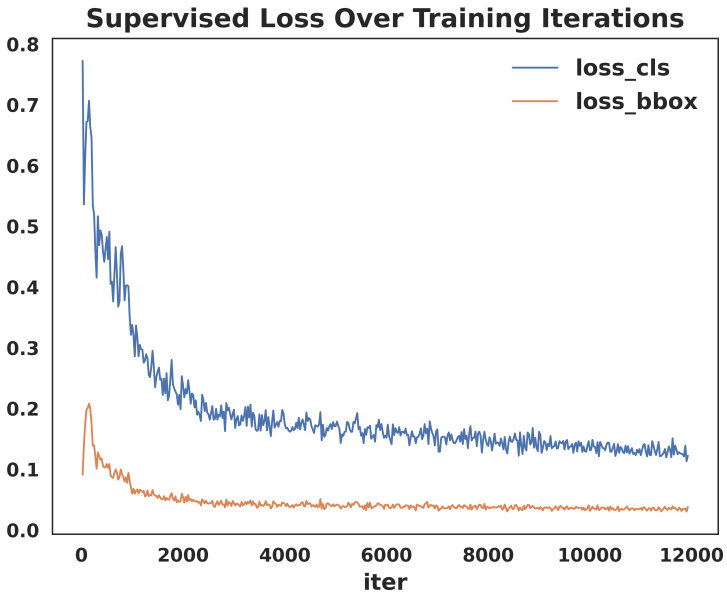


Figure 2. The normalized confusion matrix for the per-class classification performance with the (a) supervised, and (b) our proposed SSVOD approaches. SSVOD significantly improves the classification performance on the challenging classes in which the supervised approach performs poorly.

(a)



(b)

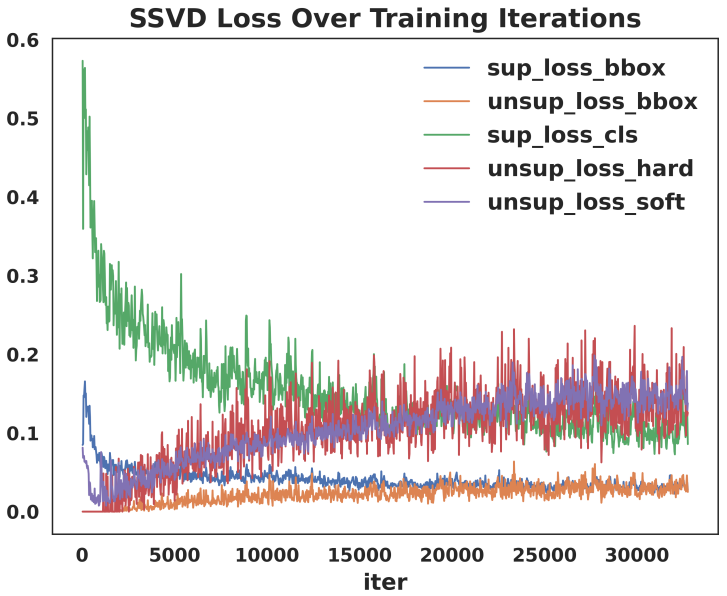


Figure 3. Analysis of the loss curves with (a) supervised approach, and (b) our proposed SSVOD approach. Supervised losses gradually decreases and saturates. In SSVOD, unsupervised losses rises gradually with improved pseudo-label generation. Finally it converges with the supervised loss.

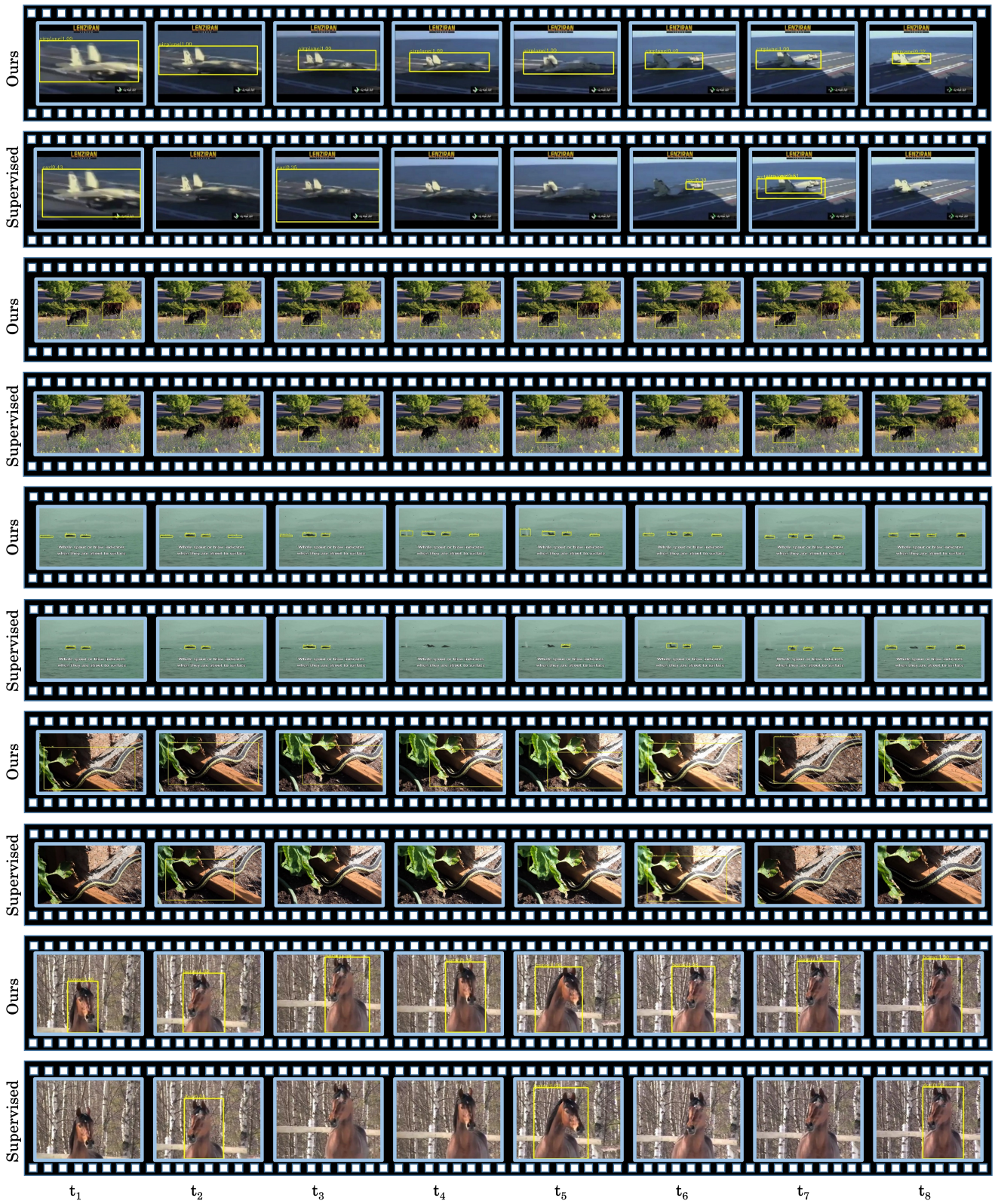


Figure 4. Qualitative performance analysis between the supervised and our proposed SSVOD approach. SSVOD generates more consistent predictions over various time steps across the video.