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

Abstract

Despite significant progress in semi-supervised learning for image object detection, several key issues are yet to be addressed for video object detection: (1) Achieving good performance for supervised video object detection greatly depends on the availability of annotated frames. (2) Despite having large inter-frame correlations in a video, collecting annotations for a large number of frames per video is expensive, time-consuming, and often redundant. (3) Existing semi-supervised techniques on static images can hardly exploit the temporal motion dynamics inherently present in videos. In this paper, we introduce SSVOD, an end-to-end semi-supervised video object detection framework that exploits motion dynamics of videos to utilize large-scale unlabeled frames with sparse annotations. To selectively assemble robust pseudo-labels across groups of frames, we introduce flow-warped predictions from nearby frames for temporal-consistency estimation. In particular, we introduce cross-IoU and cross-divergence based selection methods over a set of estimated predictions to include robust pseudo-labels for bounding boxes and class labels, respectively. To strike a balance between confirmation bias and uncertainty noise in pseudo-labels, we propose confidence threshold based combination of hard and soft pseudo-labels. Our method achieves significant performance improvements over existing methods on ImageNet-VID, Epic-KITCHENS, and YouTube-VIS datasets. Codes are available at https://github.com/enyac-group/SSVOD.git.

1. Introduction

Human annotations are expensive, time-consuming, and hard to collect for large-scale datasets [35]. In this regard, semi-supervised learning has great potential to utilize large-scale unlabeled data with limited annotations [1, 28, 34]. In recent years, researchers have shown significant progress in semi-supervised learning (SSL) on many applications, such as image classification [1, 29], semantic segmentation [2, 39], and object detection [12, 37]. However, prior work for semi-supervised object detection has mostly focused on image-based methods which opens up several key issues to be addressed particularly for videos. In this paper, we aim to fill this gap by redesigning the SSL architecture for video object detection.

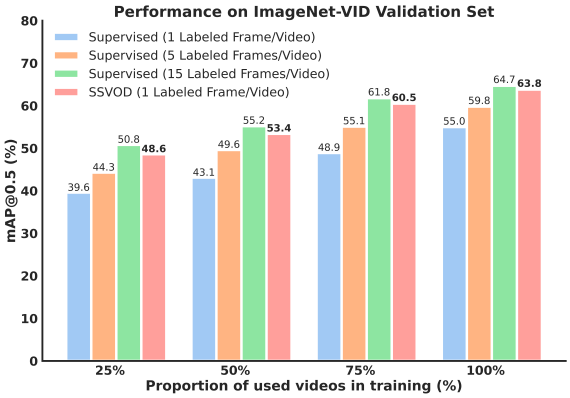

Figure 1. Supervised performance greatly depends on the availability of annotated frames per video. Our proposed SSVOD largely reduces the performance gap between sparse annotations (1 frame/video) and dense annotations (15 frames/video).

Video comes with additional challenges compared to static images, particularly arising from motion deblurring, pose variations, and camera defocus under fast motion [30, 38]. Numerous approaches have been studied to exploit the rich temporal information presented in videos to improve video object detection [14, 15]. However, the performance of supervised video object detectors greatly depends on the availability of annotated frames per video. In case of extremely sparse annotations of one labeled frame per video, an average of 22% supervised performance reduction is observed compared to the dense supervision of 15 labeled

*Work done while interned at Bosch Research North America

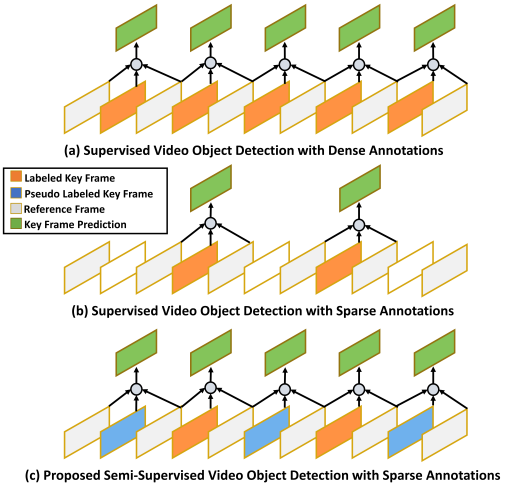

Figure 2. (a) Supervised video object detection requires *dense annotations* to explore different time-steps of videos. (b) With *sparse annotations*, the supervised method generates sub-optimal performance due to insufficient temporal exploration. (c) SSVOD leverages semi-supervised learning to *estimate robust pseudo-labels* across different time-steps by exploiting available *sparse annotations*, thereby reducing annotation burden.

frames per video (Fig. 1). For sparse annotations, standard supervised training strategies cannot explore different time-steps over the video, which results in sub-optimal performance (Fig. 2b). Despite having large inter-frame correlations in a video, collecting annotations for a large number of frames per video is expensive, time-consuming, and often redundant. Our primary motivation is to *reduce the performance gap between sparse and dense annotations by fully exploiting the inherent temporal dynamics of videos*.

The advancement of semi-supervised image object detection has introduced several key challenges in pseudo-label estimation for unlabeled images [17, 18, 31]. Wrong estimation of pseudo-labels for object classes and bounding boxes results in confirmation bias that diminishes the advantage of additional unlabeled data [27]. Image-based techniques offer various pseudo-label selection and filtering techniques to systematically filter robust pseudo-labels for unlabeled images [18, 31]. However, such techniques customized for static images cannot utilize the rich temporal information presented in videos for searching reliable pseudo-labels across groups of frames. Therefore, naive integration of image-based pseudo-label selection techniques into state-of-the-art (SOTA) video object detectors yield sub-optimal performance, thereby demanding more specialized solutions for videos (Table 1).

Moreover, existing semi-supervised learning approaches on videos [12, 19] primarily focus on the post-processing of detected bounding boxes to estimate pseudo-labels, which are referred as *box-level methods* [38]. However, SOTA video object detectors mainly operate on the feature space of groups of frames to aggregate motion cues from surrounding *reference frames* into the target *key frame* prediction [17, 18, 31]. Hence, instead of operating on final image-level predictions of video detectors, it is necessary to exploit the temporal feature space of groups of frames to estimate most consistent pseudo-labels in the target *key frame*.

In this paper, we introduce SSVOD, a semi-supervised learning framework that exploits the motion cues present in videos to greatly reduce annotation burden of SOTA video object detectors (Fig. 2c). Inspired by semi-supervised image detectors, we propose a teacher-student framework in video object detection to train on sparsely labeled data. Instead of only operating on final image-level predictions of target *key frames*, we estimate optical *flow-warped predictions* from each surrounding *reference frame* to leverage temporal consistency estimation in pseudo-label selection (Fig. 3). Using these predictions, we selectively identify the most reliable pseudo-labels for object class and bounding boxes in the target *key frame*. In particular, we introduce cross-IoU and cross-divergence based object pair matching across a group of predictions to search the most consistent bounding boxes and class labels, respectively. Moreover, to strike a balance between confirmation bias and uncertainty noise in pseudo-labels, we combine hard-class training with soft-label distillation based on their consistency. Our proposed SSVOD largely closes the performance gap between sparse and dense annotations by achieving around 98% of supervised mAP with over 95% sparsity in annotations (Fig. 1). Moreover, SSVOD provides average 7.5% higher mAP than naive integration of video object detectors and SOTA semi-supervised image techniques (Table 1). Our contributions are summarized as follows.

- We introduce a novel semi-supervised video object detection framework to tackle practical challenges of video object detection with sparse annotations.
- We propose a motion-aware robust pseudo class label and bounding box filtering approach by exploiting the temporal feature space of group of frames.
- We present an extensive experimental study that shows significant performance improvements on ImageNet-VID, Epic-KITCHENS, and YouTube-VIS datasets.

2. Related Work

2.1. Video Object Detection

Several supervised video object detectors have been explored over the years [9, 22, 30, 38]. Initial work has focused on post-processing of sequential video predictions with image-based object detectors [14, 15]. FGFA [38] first introduced flow-guided reference feature aggregation

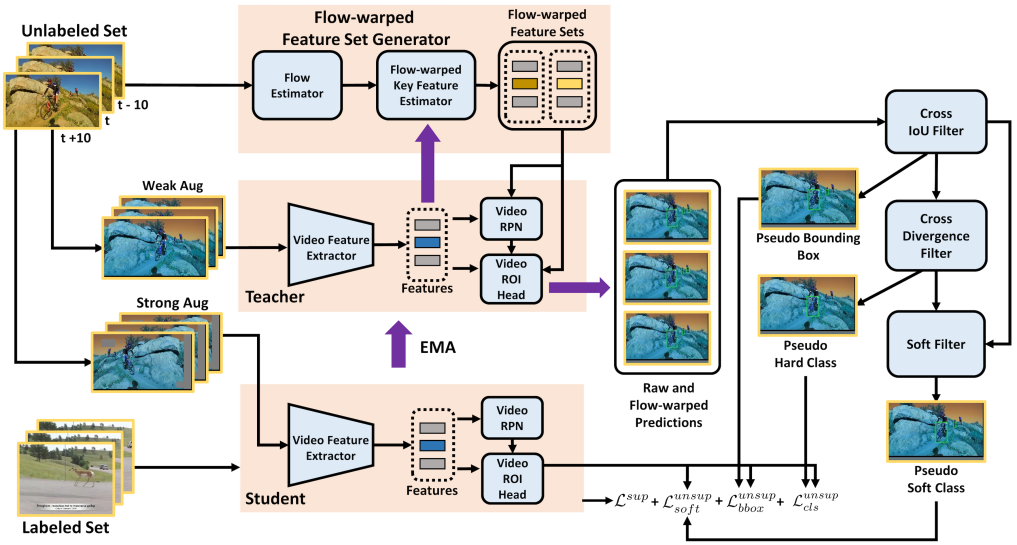

Figure 3. The overview of the semi-supervised video object detection (SSVOD). Each labeled/unlabeled set consists of one *key* and several *reference frames*. The teacher video detector operates on the unlabeled set to generate pseudo-labels. A flow-warped feature generator warps reference features using motion flow estimates to generate flow-warped feature sets. The *key frame* predictions generated from the raw and flow-warped feature sets are processed with a three-stage filtering scheme to separate pseudo bounding boxes, hard class labels, and soft-labels. The student video detector is trained on unlabeled sets with pseudo-labels, as well as on labeled sets.

from nearby frames by utilizing motion maps. SELSA [30] replaced the motion-based aggregation by full sequence-level semantics with object cross-attention weights. Later, TROI [8] improved upon SELSA by integrating temporal alignment of region-of-interest (ROI) features using the surrounding frames. TF-blender [4] proposed dense blending of the reference frame aggregation by considering each pair of surrounding frames. Relation distillation networks [6] distilled object relations from nearby reference frames to augment object features. Recently, an end-to-end transformer-based approach was introduced with sequential spatial and temporal processing [11]. However, the performance of all these approaches has been limited by the availability of the annotated frames from different temporal horizons of videos.

2.2. Semi-Supervised Object Detection

Recently, several semi-supervised image object detection techniques have been proposed. First, STAC [24] introduced a teacher-student framework for semi-supervised object detection that estimates the pseudo-labels with a pre-trained teacher. The unbiased teacher [17] approach introduced an end-to-end approach to update the teacher as a moving average of the weights from the student. However, it only considered the pseudo class labels. Soft-Teacher [31] introduced box-jittering and modified confidence thresholding of the pseudo-labels. Instant-teaching [37] proposed a co-rectify scheme by maintaining two models for generating the pseudo-labels. A single-stage detector based scheme was introduced in [18]. However, these static image-based detection approaches cannot exploit the temporal dynamics [30, 38] for overcoming video-specific challenges, such as deblurring, pose variations, and camera defocus.

Some of the prior work has focused on semi-supervised approaches for video. Misra *et al.* [19] proposed a dynamic approach to gradually learn and accumulate the unknown objects from videos. Hu *et al.* [12] introduced a pseudo-label propagation scheme on the detected objects in the unlabeled frames. However, these non end-to-end video object detection schemes are primarily built upon image-based detectors which cannot utilize the motion guided feature enhancement techniques developed in video object detectors.

3. Methodology

3.1. Overview of Supervised Video Object Detection

Supervised video object detection primarily focuses on aggregating surrounding context from nearby *reference frames* to enhance detection of a target *key frame*. In particular, supervised training requires annotations on *key frames*, whereas *reference frames* are the nearby frames primarily used for feature enhancement. With access to densely annotated *key frames* over the video, supervised methods can learn temporal dynamics across different time-steps (Fig. 2a). In contrast, having access to sparsely annotated frames in a video limits the temporal learning of supervised methods (Fig. 2b).

Consider a dataset $\mathcal{D}$ consisting M videos where $\mathcal{D} =$

$\{V^1, V^2, \ldots, V^M\}$. Each training video contains N frames with n_k annotated key frames and n_r reference frames, such that $n_k + n_r = N$. The supervised training objective for a video object detection network Z_θ parameterized by θ can be expressed as

$$\underset{\theta}{argmin} \sum_{m=1}^{M} \sum_{t=1}^{n_k} \mathcal{L}(Z_\theta(R_m^{t-i}, K_m^t, R_m^{t+i}); y_m^t), \quad (1)$$

where $\mathcal{L}$ is a pre-defined loss function, K_m^t, y_m^t denote the t^{th} *key frame* and corresponding annotation in the m^{th} video, respectively, and R_m^{t-i}, R_m^{t+i} represent the *reference frame* at time $t-i$ and $t+i$, respectively. We note that here, and in the rest of the paper, boundary cases are taken care of by ensuring $(t-i)$ is always positive.

The supervised training objective is limited by the availability of annotated *key frames*. Though annotations for *reference frames* are not required, it is necessary to use the nearby *reference frames* for proper feature enhancement. With a limited number of labeled *key frames* in a particular video, the majority of the frames from other time-steps will remain unused in the supervised setting.

3.2. Problem Formulation for SSVOD

Instead of only relying on labeled *key frames*, the semi-supervised approach targets robust pseudo-label generation for unlabeled *key frames* throughout the video (Fig. 2c). Therefore, training can be continued across different time-steps of the video utilizing the labeled and pseudo-labeled *key frames*. Given that each video contains n_k^u unlabeled *key frames* and n_k^l labeled *key frames* such that $n_k^u + n_k^l = n_k$, the semi-supervised objective can be defined as follows

$$\underset{\theta}{\text{argmin}} \left(\sum_{m=1}^{M} \sum_{t=1}^{n_k^l} \mathcal{L}(Z_\theta(R_m^{t-i}, K_m^t, R_m^{t+i}); y_m^t) + \sum_{m=1}^{M} \sum_{t=1}^{n_k^u} \mathcal{L}(Z_\theta(R_m^{t-i}, K_m^t, R_m^{t+i}); p_m^t) \right), \quad (2)$$

where p_m^t denotes the pseudo-label generated for the t^{th} unlabeled *key frame* from the m^{th} video. Therefore, the proposed semi-supervised formulation ensures utilization of sufficient key-reference pairs across the video without being entirely limited by human annotations.

3.3. SSVOD Overview

Our proposed semi-supervised video object detection (SSVOD) framework facilitates training of SOTA video object detectors with sparse annotations (Fig. 3). To utilize the temporal information with video detectors, SSVOD operates on a group of frames to aggregate features of nearby *reference frames* for *key frame* prediction. For having sparsely annotated *key frames*, SSVOD inherits the

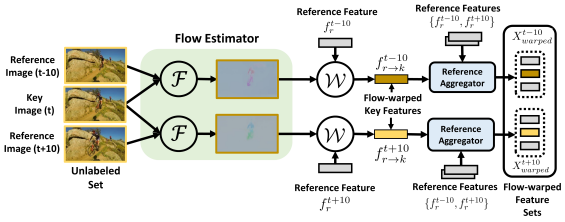

Figure 4. Overview of flow-warped feature set generation scheme. The flow estimator extracts the motion map between each key-reference pair. A feature warping is carried out on each *reference feature*. The key feature is replaced with the flow-warped feature to generate *flow-warped feature sets*.

teacher-student framework that is extensively used in semi-supervised learning [17, 31] for generating robust pseudo-labels on unlabeled *key frames*. The teacher network is derived by the exponential moving average (EMA) weights of the student network which operates on weakly-augmented unlabeled sets to estimate pseudo-labels on unlabeled *key frames*. The student network is trained on strongly augmented unlabeled sets using estimated pseudo-labels along with labeled sets. The performance of semi-supervised learning mostly depends on the correct estimation and filtering of pseudo-labels. SSVOD searches for the most consistent and robust pseudo-labels on unlabeled *key frames* by exploiting the temporal feature space of nearby frames.

Generally, in two-stage SOTA video detectors, the video feature extractor separately generates features from each *key* and *reference frame*. Later, the video region proposal network (RPN) and region-of-interest (ROI) head aggregates the reference features into the key feature to generate *raw predictions* on *key frames*. To leverage temporal consistency estimation on these *raw predictions*, we introduce estimation of *flow-warped predictions* from each *reference frame* utilizing motion cues with a flow estimator. In this regard, we propose generation of *flow-warped feature sets* from each key-reference pair based on the motion flow (Fig. 4, Sec 3.4). These *flow-warped feature sets* are processed simultaneously with the video RPN and ROI head to estimate *flow-warped predictions* on the target *key frame*. Therefore, instead of generating single *key frame* predictions, SSVOD exploits the inherent temporal feature accumulation over a group of frames in the SOTA two-stage video detectors to estimate sets of predictions based on motion flow.

As a next step, we estimate temporal consistency across generated *raw* and *flow-warped predictions* to selectively filter robust pseudo bounding box and pseudo class labels on unlabeled *key frames* (Sec 3.5). For pseudo bounding box, we estimate cross-IoU (intersection over union) across each object pair in *raw* and *flow-warped predictions* to filter objects maintaining high regional overlapping. For pseudo

class labels, we further estimate the cross-divergence in class predictions across object pairs with high regional overlapping to filter class labels with high consistency over different predictions. Soft-labels facilitate learning on inter-class relationships with more label-noise for higher uncertainty [23, 26], whereas wrong estimation of hard-labels leads to confirmation bias [27]. To strike a balance between these two, we combine hard-label training with soft-label distillation based on measured class-label consistency.

3.4. Flow-warped Feature Set Generation

Inspired by FGFA [38] which has introduced flow-guided feature aggregation in video object detection, we propose flow-warped feature set generation to estimate temporal prediction consistency. The flow-warped feature sets contain all the reference image features and the flow-warped key image features generated from motion warping of each of the reference image features. The details are presented in Figure 4. The t^{th} raw feature set X^t_{raw} having the key and reference image features of a particular video within range $[t-i, t+i]$ is denoted by

$$X^t_{raw} = \{f^{t-i}_r, \ldots, f^t_k, \ldots, f^{t+i}_r\}, \tag{3}$$

where $f^{t-i}_r, f^t_k, f^{t+i}_r$ denote features generated from R^{t-i}, K^t, and R^{t+i} image frames, respectively.

A flow network $\mathcal{F}(.)$ [7] is used to generate the flow-map of each key-reference pair, and the feature warping function $\mathcal{W}(.)$ [38] is carried out on each reference image features using the estimated flow map to generate flow-warped key feature $f_{r\to k}$. Thus, the estimated flow-warped feature set $X_{flow-warped}$ from given *key-reference features* is given by

$$f^{t+j}_{r\to k} = \mathcal{W}(f^{t+j}_r, \mathcal{F}(K^t, R^{t+j})), \tag{4}$$

$$X^{t+j}_{flow-warped} = \{f^{t-i}_r, \ldots, f^{t+j}_{r\to k}, \ldots, f^{t+i}_r\}, \tag{5}$$

$$\forall j \in \{-i, \ldots, i\}, j \neq 0.$$

In general, only two *reference frames* are used with one *key frame* in each feature set for training. Hence, the estimation of these sets is computationally efficient.

3.5. Robust Pseudo-Label Selection

On the target *key frame*, we obtain raw P_{raw} and flow-warped predictions $P_{flow-warped}$ after processing raw X_{raw} and flow-warped feature sets $X_{flow-warped}$, respectively. To estimate the temporal consistency across these predictions for filtering robust pseudo bounding boxes, hard and soft class labels, we introduce three stages of selection following initial confidence thresholding.

Cross-IoU based Pseudo Bounding Box Selection. As a first step, we estimate object pair matching between the raw and each flow-warped *key frame* prediction based on maximum overlap. For the k^{th} object in the t^{th} frame o^t_k, the matched object in the $(t+j)^{th}$ flow-warped frame o^{t+j}_k is estimated as

$$o^{t+j}_k = \underset{w\in\{1,\ldots,n^{t+j}\}}{\text{argmax}} \text{IoU}(o^t_k, o^{t+j}_w), \tag{6}$$

$$\forall k \in \{1, \ldots, n^t\}, j \in \{-i, \ldots, i\}, j \neq 0,$$

where n^t, n^{t+j} denotes number of objects at the t^{th} *key frame* and the $(t+j)^{th}$ *reference frames*, respectively. In the following, we estimate the mean cross-IoU (xIoU) for each object of the t^{th} raw prediction P^t_{raw} which represents the consistency in bounding boxes where

$$\text{xIoU}(k) = \tfrac{1}{2i}\textstyle\sum_{j=-i}^{i} \text{IoU}(o^t_k, o^{t+j}_k)\ \forall k \in \{1, \ldots, n^t\}. \tag{7}$$

Finally, we filter out the objects with high xIoU scores to obtain the pseudo bounding boxes P^t_{bbox}, which is given by

$$P^t_{bbox} = \mathbb{I}(\text{xIoU} > \zeta_{IoU}) P^t_{raw,bbox}, \tag{8}$$

where $\mathbb{I}(.)$ is the indicator function, and ζ_{IoU} denotes the threshold IoU.

Cross-Divergence based Pseudo Hard Class Selection. In some cases, the bounding box estimation can be accurate, even though the object class is wrong. To address this issue, we estimate the mean cross KL-divergence (xDiv) for each object o^t_k in the raw prediction P^t_{raw} with its corresponding flow-warped object o^{t+j}_k, such that

$$\text{xDiv}(k) = \tfrac{1}{2i}\textstyle\sum_{j=-i}^{i} D_{KL}(o^t_k, o^{t+j}_k),\ \forall k \in \{1, \ldots, n^t\}. \tag{9}$$

The bounding boxes with lower KL-divergence are filtered for hard class labeling with threshold η_{div}, such that

$$P^t_{cls} = \mathbb{I}(\text{xDiV} < \eta_{div}) P^t_{raw,cls}. \tag{10}$$

Confidence-Aware Pseudo Soft Class Selection. To enhance the exploration of the generated pseudo-labels, class predictions from remaining bounding boxes in $P^t_{raw,cls}$ are accumulated after confidence thresholding with threshold γ_c to generate the soft-class distribution P^t_{soft} , *i.e.*,

$$P^t_{soft} = \mathbb{I}(c > \gamma_c; \text{xDiV} > \eta_{div}; \text{xIoU} < \zeta_{IoU}) P^t_{raw,cls}. \tag{11}$$

A soft-distillation is carried out between the class prediction of student network $P^t_{std,cls}$ and the filtered soft-class distribution from teacher P^t_{soft} at time t, such that

$$\mathcal{L}^t_{soft} = D_{KL}(P^t_{std,cls}, P^t_{soft}). \tag{12}$$

Finally, the objective loss to train the student is defined as

$$\mathcal{L} = \mathcal{L}^{sup}_{cls} + \mathcal{L}^{sup}_{bbox} + \mathcal{L}^{unsup}_{cls} + \mathcal{L}^{unsup}_{bbox} + \mathcal{L}^{unsup}_{soft}. \tag{13}$$

Table 1. Performance comparison on ImageNet-VID validation set under **single labeled key frame per video** setting with different sampling rates and under **multiple labeled key frames per video** setting from 25% training videos. * denotes our improved implementation for the VOD task. All the results are average of three independent runs. No ImageNet-DET pre-training is considered.

Method	Single Labeled Key Frame (Percentage)				Multiple Labeled Key Frames (Number)		
	25%	50%	75%	100%	5	10	15
(a) Supervised Image Baseline [20]	30.1 ± 0.25	37.6 ± 0.22	42.5 ± 0.18	47.8 ± 0.29	35.2 ± 0.26	39.9 ± 0.26	43.1 ± 0.18
(b) STAC [25]	37.2 ± 0.31	44.2 ± 0.18	46.6 ± 0.24	50.9 ± 0.19	39.5 ± 0.29	41.3 ± 0.19	-
(c) Soft-Teacher [31]	40.2 ± 0.26	46.7 ± 0.16	49.7 ± 0.18	54.7 ± 0.21	40.8 ± 0.32	42.2 ± 0.27	-
(d) Unbiased Teacher [17]	39.1 ± 0.16	44.6 ± 0.26	48.3 ± 0.21	53.4 ± 0.28	40.1 ± 0.23	41.8 ± 0.18	-
(e) Supervised Video Baseline [30]	39.6 ± 0.23	43.1 ± 0.17	48.9 ± 0.19	55.0 ± 0.28	44.3 ± 0.32	47.4 ± 0.24	50.8 ± 0.28
(f) Baseline [30] + STAC [25]	43.4 ± 0.26	47.9 ± 0.21	53.5 ± 0.18	57.4 ± 0.12	45.6 ± 0.21	48.3 ± 0.25	-
(g) Baseline [30] + Soft-Teacher [31]	45.2 ± 0.21	50.2 ± 0.22	55.8 ± 0.20	59.3 ± 0.19	46.4 ± 0.23	48.8 ± 0.19	-
(h) Baseline [30] + Unbiased Teacher [17]	44.7 ± 0.28	51.0 ± 0.25	55.1 ± 0.23	58.8 ± 0.17	46.2 ± 0.24	48.7 ± 0.18	-
(i) PseudoProp [12] + Soft-Teacher [31]	41.1 ± 0.26	47.6 ± 0.26	51.6 ± 0.21	56.5 ± 0.17	45.3 ± 0.27	48.0 ± 0.16	-
(j) Misra *et al.** [19]	42.9 ± 0.23	48.5 ± 0.25	53.1 ± 0.19	57.7 ± 0.21	47.1 ± 0.22	49.2 ± 0.21	-
(k) Yan *et al.* [32]*	40.5 ± 0.23	44.2 ± 0.18	50.3 ± 0.21	52.5 ± 0.26	46.2 ± 0.23	47.0 ± 0.27	-
(l) Ours (SSVOD)	**48.6** ± 0.23	**53.4** ± 0.18	**60.5** ± 0.24	**63.8** ± 0.19	**49.7** ± 0.24	**50.4** ± 0.22	-

4. Experiments

4.1. Dataset and Evaluation Setup

Following prior work [8, 30, 38], we primarily use the large-scale ImageNet-VID dataset [21] for experiments. The training set contains 3,862 videos collected at a frame rate of 25 frames per second (fps) to 30 fps, and the validation set contains 555 videos. A total of 30 object categories are included, which is a subset of ImageNet-DET dataset containing 200 categories of static images [21]. Existing supervised schemes used the ImageNet-DET dataset for pre-training which is a large-scale static-image dataset with similar objects. Since such pre-training on ImageNet-DET disrupts the sparse annotation constraints on the ImageNet-VID performance, we primarily focus on the ImageNet-VID dataset with sparse annotations. However, some experiments are done to analyze the effect of ImageNet-DET inclusion. For training, we considered the same 15 uniformly sampled labeled *key frames* per video as in prior work [4, 30, 38] with the following two settings.

Single Labeled Key Frame per Video. Only one out of these 15 key frames is used as a labeled *key frame* per video while the remaining 14 are used as unlabeled *key frames* during training. We sample 25%, 50%, 75%, and 100% videos with a single labeled *key frame*. The nearby surrounding frames of these *key frames* are considered reference frames where annotations are not required.

Multiple Labeled Key Frames per Video. We sample 5, 10, and 15 labeled *key frames* from a total of 15 per video to represent different degrees of annotation sparsity during training. Moreover, we sample 1,000 videos (≈ 25%) from the whole training dataset. Following standard practice [30, 38], nearby surrounding frames of labeled/unlabeled *key frames* are used as *reference frames*.

Additional Datasets. Since ImageNet-VID lacks diversity of objects, we also study the performance on challenging Epic-KITCHENS [5] dataset that contains 290 classes from 272 videos taken from 28 kitchen environments. For additional comparisons on video object detection, we use YouTube-VIS [33] dataset having 40 object categories with 2238, 302, and 343 training, validation, and test videos.

Following prior works [30, 38], we follow the same convention to report the results with the standard mAP evaluation metric. The categorization of different size objects and different motion objects follows the design in [38]. All experiments are conducted with three different sets generated with independent sampling.

4.2. Implementation Details

We use the implementation and hyper-parameters based on MMTracking [3]. For each *key frame*, we use two *reference frames* in training and 30 *reference frames* in evaluation following prior work [8, 30, 38]. For the video object detector, we mostly focus on SELSA network [30] unless otherwise specified. We use the COCO-pretrained Faster-RCNN network [20] as the base object detection network with FPN [16] and ResNet-50 [10] backbone. More details can be found in supplementary materials.

4.3. Main Results

In this section, we present the key results obtained with the proposed SSVOD framework. We also compare SSVOD with the image and video baselines. We report the mAP with IoU set to 50%, denoted as mAP@IoU=0.5, on the ImageNet-VID validation set. We note that each method is associated with a letter for clarity in reporting the results (*e.g.*, SSVOD (l), see Table 1).

Single Labeled Key Frame per Video Performance. In this setup, supervised video baseline provides an average of +7 mAP performance improvement over image baseline, as shown in Table 1. SSVOD (l) significantly outperforms the supervised baselines by achieving +46.4%

Table 2. Ablation with various video object detector. Our SSVOD scheme provides consistent improvements over supervised baseline across various detectors.

detector	method	mAP	mAP@0.5	mAP@0.75
FGFA [38]	Supervised	23.4	51.3	28.8
	Ours	**31.9** (+8.5)	**58.7** (+7.4)	**36.4** (+7.6)
SELSA [30]	Supervised	32.1	55.0	34.7
	Ours	**39.2** (+9.2)	**63.8** (+8.8)	**43.1** (+8.4)
TROI [8]	Supervised	33.9	56.2	37.1
	Ours	**42.6** (+8.7)	**64.8** (+8.6)	**44.8** (+7.7)

Table 3. Ablation study on the effect of different loss components on pseudo-labels.

Hard class	Bounding box	Soft class	mAP	mAP@0.5	mAP@0.75
			32.1	55.0	34.7
✓			35.6	58.8	37.9
✓	✓		37.9	61.1	41.7
✓	✓	✓	**39.2**	**63.8**	**43.1**

and $+22.8\%$ higher mAP than supervised image (a), and video (e) baselines, respectively. We study the effect of SOTA semi-supervised image techniques [17, 25, 31] for pseudo-label filtering on image and video baselines. We notice considerable improvements over the supervised baseline with these techniques, *e.g.*, Soft-Teacher [31] improves the image baseline by average $+7.7$ mAP (c), and video baseline by $+5.8$ mAP (g). However, since these image techniques cannot exploit temporal dynamics of a group of frames in videos, they provide sub-optimal improvement over video baseline. By overcoming such limitations with motion-guided pseudo-label filtering, SSVOD achieves average $+4$ mAP higher than the best performing naive integration of video baseline and Soft-Teacher (g). Though PseudoProp [12] leverages motion propagation in pseudo-label estimation, its non end-to-end image-based feature processing primarily focuses on the post-processing of generated outputs rather than exploration of temporal feature space of groups of frames. Hence, it provides sub-optimal performance even when integrated with SOTA image techniques (i). The iterative supervised training and object tracking approach introduced in (i) for generating pseudo-labels are time-consuming (requiring 30 iterations) and fail to produce reliable pseudo-labels on distant frames due to its reliance on tracking performance. The method in (k) generates pseudo-labels by relying on annotated frames at regular intervals, leading to incorrect pseudo-label estimation on distant frames under high sparsity of annotations.

Multiple Labeled Key Frames per Video Performance. In this setting, we observe consistent performance improvement with the increase in labeled key frames per video for all baselines, as shown in Table 1. For the supervised video baseline (e), we notice $+4.7$, $+7.8$, $+11.2$ increase in mAP with the increase of labeled *key frames* from 1 to 5, 10, and 15, respectively. This shows that supervised performance greatly depends upon the availability of labeled *key frames* per video. In general, the supervised video baselines (e) maintain superior performance over image baselines (a) and inclusion of semi-supervised image techniques considerably improve the supervised performance as before. However, our SSVOD (l) outperforms all the baselines by a considerable margin and largely closes the gap between sparse training and full-supervised training, achieved with all 15 labeled *key frames* per video.

Table 4. Ablation study on the effect of different number of unlabeled *key frames*.

# of unlabeled key frames	mAP	mAP@0.50	mAP@0.75
1	34.7	54.5	38.8
5	37.1	59.4	41.6
10	38.6	62.1	42.7
14	**39.2**	**63.8**	**43.1**

Table 5. The effect of ImageNet-DET dataset integrated as the unlabeled set using single key frame from ImageNet-VID dataset in the labeled set.

Unlabeled Dataset	mAP
ImageNet-VID (14 Key)	39.2
ImageNet-VID (14 Key) + ImageNet-DET	**41.1** (+1.9)

Table 6. Additional study with SOTA VOD and flow estimators on SSVOD framework in the single-frame (100% videos) setting.

VOD	**Flow Estimator**	**ImageNet-VID**
SELSA [30]	FlowNet [7]	63.8 ± 0.19
SELSA [30]	FlowFormer [13]	68.4 ± 0.18
TransVOD [36]	FlowNet [7]	67.2 ± 0.23
TransVOD [36]	FlowFormer [13]	**70.5** ± 0.26

Table 7. Additional comparisons on Epic-KITCHENS and YouTube-VIS Datasets on single-frame (100% videos) and multi-frame (5 frames per video on 25% videos) settings.

Method/Dataset	EPIC-KITCHENS [5]		YouTube-VIS [33]	
	Single	**Multi**	**Single**	**Multi**
Supervised [30]	30.4 ± 0.18	25.8 ± 0.26	45.3 ± 0.29	38.8 ± 0.15
PseudoProp [12]	32.4 ± 0.29	29.2 ± 0.24	47.6 ± 0.17	41.8 ± 0.23
Misra et al. [19]*	31.6 ± 0.23	28.1 ± 0.22	48.9 ± 0.19	42.2 ± 0.17
Yan et al. [32]*	29.2 ± 0.20	26.7 ± 0.24	43.7 ± 0.25	41.4 ± 0.23
SSVOD (ours)	$\mathbf{36.7 \pm 0.22}$	$\mathbf{30.9 \pm 0.23}$	$\mathbf{53.6 \pm 0.21}$	$\mathbf{45.1 \pm 0.21}$

4.4. Ablation Studies

In this section, we validate the effect of each design choice and parameter setting. All the ablation studies are conducted on the full training set with a single labeled key frame per video, unless specified otherwise.

Effect of Different Video Object Detectors. Though we used SELSA [30] as a proof-of-concept for most experiments, we ablate the effect of different video object detectors on the SSVOD framework as shown in Table 2. Among

supervised schemes, SELSA detector [30] provides superior performance to FGFA [38], while TROI [8] exhibits the best performance. The SSVOD performance improvements are consistent with its supervised counterpart across different detectors. TROI detector achieves the best performance with 56.2 mAP in supervised setting and its performance reaches 64.8 mAP for SSVOD (+8.6 points higher).

Effect of Different Loss Components on Pseudo-Labels. We study the effect of different loss components on pseudo-labels (see Table 3). Integrating hard-label cross-entropy loss improves performance by +3 points. Further applying pseudo bounding box regression loss, we get +2.3 mAP improvements. Finally, by incorporating soft-label distillation, we achieve the best performance of 63.8 mAP that shows the effectiveness of the three loss components.

Number of Unlabeled *Key Frames*. We evaluate the performance for different numbers of unlabeled *key frames*. The performance consistently improves with the increasing number of unlabeled *key frames*, as shown in Table 4. When a higher number of *key frames* is used, SSVOD framework can explore different temporal regions of the video thereby producing better performance. We argue that SSVOD makes the annotations of multiple *key frames* per video redundant. With one labeled and 14 unlabeled *key frames*, the SSVOD achieves 63.8 mAP while the supervised baseline with 15 labeled *key frames* is at 64.7 mAP.

Correctness of Generated Pseudo-Labels. We study pseudo-label mAP on single labeled key frame per video setting with 25% training videos: (See Table 1 for naming) Method (f) is 57.1, (g) is 59.8, (h) is 58.5, (i) is 55.2, and ours (l) is 64.9. The higher mAP achieved by SSVOD demonstrates better quality of pseudo-label generation.

Effect of ImageNet-DET as Unlabeled Set. As described earlier, ImageNet-VID contains 30 objects which are a subset of ImageNet-DET with 200 object classes. Existing supervised schemes pre-train the video detector on the ImageNet-DET dataset that provides large performance gains [8, 30, 38]. Since ImageNet-DET is a static image dataset, video pre-training is conducted by considering the same image as both *key* and *reference frames*. However, such pre-training violates the constraints introduced by sparse annotations. Instead of such pre-training, we study the effects of ImageNet-DET integration as an unlabeled set. The results are given in Table 5. Despite having several unseen classes, we notice considerable performance improvement with the integration of ImageNet-DET.

Effect of SOTA VOD and flow estimators in SSVOD. We use FlowNet [7], and baseline VODs [30] as proof-of-concept. We ablate SOTA FlowFormer [13] and TransVOD [36] in SSVOD (Table 6). Integration of superior baseline models in SSVOD significantly improves performance, which is consistent with our prior observations.

Performance on additional datasets. We present results on two additional benchmark datasets (EPIC-KITCHENS [5] and Youtube-VIS [33]) for VOD tasks (See Table 7). SSVOD consistently achieves superior performance on these benchmark datasets over existing methods.

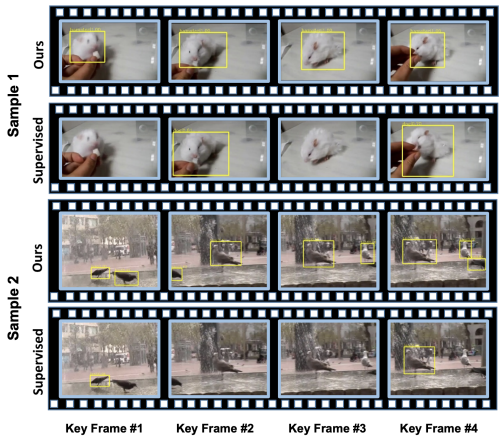

Figure 5. Qualitative visualization of supervised and SSVOD performance: SSVOD demonstrates better temporal consistency in predictions compared to its supervised counterpart over both challenging single and multi-object scenarios.

Qualitative Visualization of Performance: We study the qualitative performance of SSVOD and its corresponding supervised video baseline on challenging single and multi-object videos, as shown in Fig. 5 (see appendix for more visualizations). SSVOD demonstrates visibly better temporal consistency than its supervised counterparts.

5. Conclusion

In this paper, we introduce a novel semi-supervised learning framework (SSVOD) to overcome the limitations of existing supervised video object detection approaches for sparse annotations. Instead of the naive integration of existing semi-supervised image techniques on SOTA video detectors, SSVOD exploits temporal feature space of groups of frames to search robust pseudo-labels based on motion consistency. Moreover, SSVOD is found to be detector invariant which can scale-up performance with improved supervised baselines. Our proposed three-stage pseudo-label selection for bounding-box regression, hard-label classification, and soft-label distillation largely contributes to the final performance gain. Through effective utilization of unlabeled frames in videos, SSVOD achieves around 98% of densely supervised performance by using over 95% sparser annotations significantly outperforming other baselines.

Acknowledgements

This work was supported in part by ONR Minerva program, iMAGiNE - the Intelligent Machine Engineering Consortium at UT Austin, and a UT Cockrell School of Engineering Doctoral Fellowship.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 1

[2] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 695–714, Cham, 2020. Springer International Publishing. 1

[3] M Contributors. Mmtracking: Openmmlab video perception toolbox and benchmark, 2020. 6

[4] Yiming Cui, Liqi Yan, Zhiwen Cao, and Dongfang Liu. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8138–8147, 2021. 3, 6

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 6, 7, 8

[6] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019. 3

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 5, 7, 8

[8] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, and Huamin Feng. Temporal roi align for video object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1442–1450, 2021. 3, 6, 7, 8

[9] Fei He, Naiyu Gao, Jian Jia, Xin Zhao, and Kaiqi Huang. Queryprop: Object query propagation for high-performance video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 834–842, 2022. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[11] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021. 3

[12] Shu Hu, Chun-Hao Liu, Jayanta Dutta, Ming-Ching Chang, Siwei Lyu, and Naveen Ramakrishnan. Pseudoprop: Robust pseudo-label generation for semi-supervised object detection in autonomous driving systems. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4389–4397, 2022. 1, 2, 3, 6, 7

[13] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European Conference on Computer Vision*, pages 668–685. Springer, 2022. 7, 8

[14] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017. 1, 2

[15] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 817–825, 2016. 1, 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[17] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 4, 6, 7

[18] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 2, 3

[19] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3602, 2015. 2, 3, 6, 7

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[21] 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. 6

[22] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. Yolov: Making still image object detectors great at video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2254–2262, 2023. 2

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk,

Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 5

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

[25] 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. 6, 7

[26] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3132–3141, 2021. 5

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 5

[28] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 1

[29] Jianfeng Wang, Thomas Lukasiewicz, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, and Alexandros Neophytou. NP-match: When neural processes meet semi-supervised learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22919–22934. PMLR, 17–23 Jul 2022. 1

[30] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. *ICCV 2019*, 2019. 1, 2, 3, 6, 7, 8

[31] 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. 2, 3, 4, 6, 7

[32] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7284–7293, 2019. 6, 7

[33] 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. 6, 7, 8

[34] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 1

[35] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021. 1

[36] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7, 8

[37] 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. 1, 3

[38] 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. 1, 2, 3, 5, 6, 7, 8

[39] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8848–8857, 2019. 1