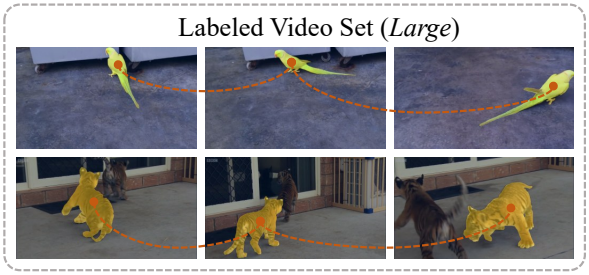
Minimizing Labeled, Maximizing Unlabeled: An Image-Driven Approach for Video Instance Segmentation

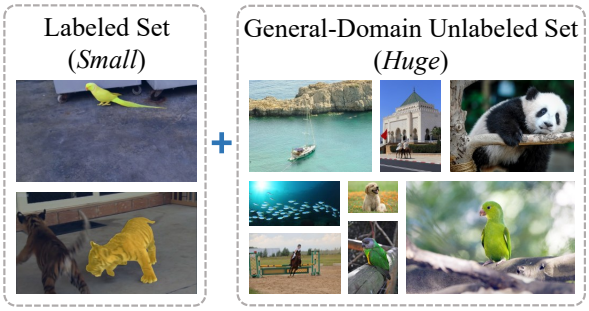
Fangyun Wei[1*] Jinjing Zhao[1*] Kun Yan[2†] Chang Xu[1†]
[1]The University of Sydney [2]Peking University
{fwei8714,jzha0100}@uni.sydney.edu.au kyan2018@pku.edu.cn c.xu@sydney.edu.au

Abstract

Traditional video instance segmentation (VIS) models rely on extensive per-frame video annotations, which are both time-consuming and costly. In this paper, we present MinMaxVIS, a novel VIS framework that reduces the dependency on fully labeled video datasets by utilizing a small set of labeled images from the target domain along with a large volume of general-domain, unlabeled images. MinMaxVIS operates in three stages: first, a preliminary segmentation model is trained on the small labeled set from the target domain; this model then retrieves relevant instances from the unlabeled dataset to build a high-quality pseudo-labeled set, ensuring a rich content alignment with the target domain while avoiding the inefficiencies of large-scale semi-supervised learning across the entire unlabeled set. Finally, we train MinMaxVIS on a combination of labeled and pseudo-labeled data, addressing challenges such as noise in pseudo-labels and instance association across frames. To simulate object continuity, we augment static images to create paired frames, allowing MinMaxVIS to capture instance associations effectively. MinMaxVIS outperforms the prior image-driven approach, MinVIS, achieving superior mAP scores with significantly reduced labeled data. For instance, MinMaxVIS with a Swin-L backbone attains 62.2 mAP on YouTube-VIS 2019 using only 2% labeled data and additional unlabeled images from SA-1B. This surpasses MinVIS, which uses the same backbone trained on fully labeled YouTube-VIS 2019, by 0.6 mAP.



(a) Traditional video-based VIS training.



(b) MinMaxVIS image-based training.

Figure 1. (a) Traditional VIS models rely on fully labeled video frames with instance association across frames, demanding extensive manual annotations. (b) MinMaxVIS enables effective video instance segmentation using only a small set of labeled target-domain images and a vast amount of unlabeled general-domain images, significantly reducing annotation costs while maximizing data efficiency.

1. Introduction

The objective of video instance segmentation (VIS) is to identify, segment, and consistently track each unique object instance across frames in a video sequence. Traditional VIS methods [18, 20, 21, 25, 40] primarily rely on fully labeled video datasets, where each frame requires segmentation labels, class labels, and association labels that link instances across frames. This approach, while effective, demands extensive labeling, which is both time-consuming and costly, especially for high-quality datasets with detailed annotations.

In this paper, we introduce a novel VIS framework that reduces dependency on fully labeled video data. As shown in Figure 1, instead of relying on labeled video sequences, our approach, MinMaxVIS, requires only a minimal amount of labeled images from the target domain and

*Equal contribution.
†Corresponding author.

leverages a large quantity of general-domain, unlabeled internet images. For instance, we utilize merely around 1,200 labeled images from the YouTube-VIS 2019 benchmark [45]—representing only 2% of all frames—alongside 2.8 million unlabeled images from the SA-1B dataset [23]. This setup eliminates the need for fully labeled video data and addresses the high cost of annotations. The name "MinMaxVIS" reflects our goal: to minimize the use of labeled images and maximize the utility of unlabeled data.

The training process of MinMaxVIS consists of three stages. First, we train a preliminary segmentation model on the small labeled set from the target domain. This model then serves as a retrieval tool to identify relevant instances from the vast, unlabeled set of general-domain images, focusing on high-precision retrieval to build a pseudo-labeled dataset, with each pseudo-label generated by the preliminary model. This retrieval stage is crucial in our framework, as the unlabeled set contains millions of images, most of which do not include instances relevant to the target domain. Performing semi-supervised learning [16, 34, 43, 44] directly on this vast dataset would be computationally costly and inefficient. Instead, by using the preliminary model to selectively retrieve only the most relevant images, we minimize data redundancy and ensure that the pseudo-labeled set is rich in domain-specific content. Finally, we train MinMaxVIS on a combination of the small labeled image set and the expansive pseudo-labeled set, achieving a robust model with minimal manual annotation.

Our approach addresses two primary challenges inherent to this VIS framework: noise in the pseudo-labeled data and the need of instance association when training on static image data, in contrast to classical video datasets [29, 45, 46], which typically provide instance association labels across frames. Pseudo-labeled images may contain false negatives—instances of target categories that are not recalled—resulting in noisy labels. To mitigate this, we introduce a selective gradient backpropagation technique. This technique ensures that only foreground queries and high-confidence background queries contribute to network training, while gradients from low-confidence background queries, which can be either true negatives (i.e., backgrounds) or false negatives (i.e., missed foreground instances), are detached, thereby reducing the impact of label noise. The second challenge, instance association, is critical as VIS requires linking the same object instance across frames. Since our training is conducted on static images rather than videos, we simulate video data by creating paired images through augmentations. This pairing allows us to associate the same object across frames by maximizing feature similarity for corresponding instances and minimizing it for non-matching instances.

Previous work by MinVIS [18] has shown that VIS models can be trained effectively on static images. We build upon MinVIS [18], enhancing segmentation robustness on image datasets while minimizing reliance on labeled data and efficiently utilizing extensive unlabeled data. Our approach, MinMaxVIS, outperforms MinVIS [18], which relies on 100% labeled data, by leveraging only 2%-10% labeled data alongside a large pool of unlabeled data, across benchmarks such as YouTube-VIS 2019 [45], YouTube-VIS 2021 [46], and OVIS [29]. For instance, on YouTube-VIS 2019, MinMaxVIS achieves an mAP of 62.2 with a Swin-L backbone, using just 2% labeled training data and additional unlabeled data from SA-1B [23], surpassing MinVIS with the same backbone, trained on the fully labeled dataset, by 0.6 mAP.

2. Related Works

Instance Segmentation. Image instance segmentation, which involves predicting the classes and segmentation masks of individual objects within an image, has advanced significantly in parallel with object detection improvements [5, 30, 31, 35]. Mask R-CNN [13] was a pioneering method in this field, extending Faster R-CNN [31] by adding a mask head to generate object masks. Subsequent approaches have refined this framework, focusing on enhancing bounding box quality [3, 6, 28] or improving mask precision [9, 19, 22, 33, 49]. To bypass the intermediate step of region of interest (RoI) detection, some methods [2, 10, 42] leverage one-stage detectors [30, 35]. For instance, YOLACT [2] builds on RetinaNet [32], generating prototype masks followed by per-instance mask coefficients. Additionally, SOLO [37] and SOLOv2 [38] eliminate the need for bounding box prediction by using object center features in a one-stage pipeline.

More recently, query-based methods [8, 11, 12, 17, 24, 53, 54] have gained popularity, inspired by DETR [5], the first Transformer-based end-to-end detector. For example, Mask2Former [8] introduces masked attention to focus cross-attention within predicted mask regions, enhancing feature localization. MaskDINO [24] extends DINO [50] by adding a mask prediction branch, thereby unifying object detection and segmentation. However, these methods face limitations in video instance segmentation as they lack effective mechanisms for associating objects across frames.

Video Instance Segmentation. Video instance segmentation (VIS) is first introduced by [45]. The goal of VIS is to segment and track object instances across frames in a video sequence. Current VIS methods are generally categorized into two types: offline methods [1, 7, 14, 20, 26, 39, 40, 48] and online methods [4, 15, 18, 21, 25, 27, 41, 45, 47].

Offline VIS methods process an entire video (or a video clip) in one pass, predicting instance segmentation results for all frames simultaneously. For example, VisTR [39] was the first to apply Transformers [36] to VIS, capturing spatial and temporal features across the entire video. Seq-

Former [40] dynamically predicts mask sequences for each frame by aggregating temporal information, learning robust video-level representations for each instance. VITA [14] achieves video-level understanding by associating frame-level object tokens without relying on spatio-temporal backbones. However, this video-in and video-out approach is not well-suited for streaming applications.

Online VIS methods, in contrast, process videos frame by frame, making them more suitable for streaming scenarios. MaskTrack R-CNN [45] extends Mask R-CNN [13] by adding a tracking branch to handle segmentation and tracking simultaneously. Following this approach, other methods [4, 27, 47] generate frame-by-frame outputs and associate instances across frames using various matching techniques. IDOL [41] enhances instance association by learning discriminative embeddings via contrastive learning, based on Deformable-DETR [56]. More recently, TCOVIS [25] improves temporal feature consistency through global instance assignment and spatio-temporal enhancement, while VISAGE [21] enhances object distinction by incorporating appearance cues during tracking. Additionally, the DVIS series [51, 52, 55] offers flexible segmentation capabilities, supporting both online and offline processing modes. The work most relevant to ours is MinVIS [18], which is based on Mask2Former [8] and allows for training VIS models using static images. However, MinVIS relies on fully labeled video datasets from the target domain. In contrast, our approach enables training with only a small set of labeled images from the target domain while effectively leveraging the knowledge from large, general-domain unlabeled datasets.

3. Methodology

Problem Formulation. Let $\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{I}_i, \boldsymbol{L}_i)\}_{i=1}^{N}$ represent a small, labeled image set from the target domain, where $\boldsymbol{I}_i$ is an image and $\boldsymbol{L}_i$ contains the labeled instances within $\boldsymbol{I}_i$, including class labels and segmentation masks for each instance. Here, N denotes the total number of labeled images. Additionally, we have a large, unlabeled image set $\mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{I}_j\}_{j=1}^{M}$, sourced from a general domain, where M represents the total number of unlabeled images. These images feature diverse content that, while not specifically annotated for the target domain, provides valuable general features that can enrich the final MinMaxVIS model. Notably, $M \gg N$, emphasizing the disparity in size between the unlabeled and labeled datasets.

The objective is to train an effective video instance segmentation model, MinMaxVIS, by leveraging the combination of the small labeled image set $\mathcal{D}_{\text{labeled}}$ from the target domain, and the large unlabeled image set $\mathcal{D}_{\text{unlabeled}}$ from the general domain, allowing MinMaxVIS to generalize well on the target domain.

Overview. The MinMaxVIS architecture, illustrated in Figure 2, consists of three primary stages: preliminary segmentation model training on the small labeled image set, high-precision retrieval on the unlabeled set, and MinMaxVIS training. First, a segmentation model is trained on $\mathcal{D}_{\text{labeled}}$, establishing a foundation for subsequent steps (Section 3.1). This trained model is then used to perform high-precision retrieval on the large unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$, constructing a pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$ that is rich in target-domain content (Section 3.2). In the final MinMaxVIS training stage, $\mathcal{D}_{\text{labeled}}$ and $\mathcal{D}_{\text{pseudo}}$ are combined, integrating selective gradient backpropagation to mitigate noise in pseudo-labels, augmented paired images to simulate continuity across frames, and instance association to link instances across frames (Section 3.3). The inference process for MinMaxVIS is detailed in Section 3.4.

3.1. Preliminary Segmentation Model Training

The first stage involves training a preliminary segmentation model $\boldsymbol{S}_{\boldsymbol{\theta}}$ on the small labeled dataset $\mathcal{D}_{\text{labeled}}$. This model is designed to facilitate high-precision retrieval of relevant instances from the unlabeled image set $\mathcal{D}_{\text{unlabeled}}$, ensuring a rich content alignment with the target domain while avoiding the inefficiencies of large-scale semi-supervised learning across the entire unlabeled set.

In our implementation, we employ Mask2Former [8] with a Swin-L backbone as the preliminary segmentation model, which is trained using standard classification and segmentation loss functions. Given that $\mathcal{D}_{\text{labeled}}$ is relatively small, this training process is efficient and fast, allowing us to establish a foundational model for subsequent retrieval and pseudo-labeling stages.

3.2. High-Precision Retrieval

In the second stage, the preliminary segmentation model $\boldsymbol{S}_{\boldsymbol{\theta}}$ trained on $\mathcal{D}_{\text{labeled}}$ is used to retrieve relevant instances from the large, unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$, as shown in Figure 2(b). This retrieval step aims to construct a pseudo-labeled set, $\mathcal{D}_{\text{pseudo}}$, that contains content highly aligned with the target domain while minimizing irrelevant data.

The high-precision retrieval process begins by passing each image $\boldsymbol{I}_j$ from $\mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{I}_j\}_{j=1}^{M}$ through the segmentation model $\boldsymbol{S}_{\boldsymbol{\theta}}$, which generates a set of predictions $\{(c_u, \boldsymbol{M}_u)\}_{u=1}^{U}$ for $\boldsymbol{I}_j$. Here, c_u represents the confidence score of the u-th prediction, $\boldsymbol{M}_u$ is the corresponding mask, and U denotes the total number of predictions for the image. To ensure high precision, only predictions with confidence scores exceeding a predefined threshold τ are retained as pseudo-labels. Formally, for each image $\boldsymbol{I}_j$, we keep predictions $\widehat{\boldsymbol{L}}_j = \{(c_u, \boldsymbol{M}_u) | c_u \geq \tau\}_{u=1}^{U}$, where τ is set to a high value to minimize false positives, potentially allowing some false negatives (missed instances) while ensuring that retained instances are highly reliable. The issue of false negatives is mitigated in the MinMaxVIS training

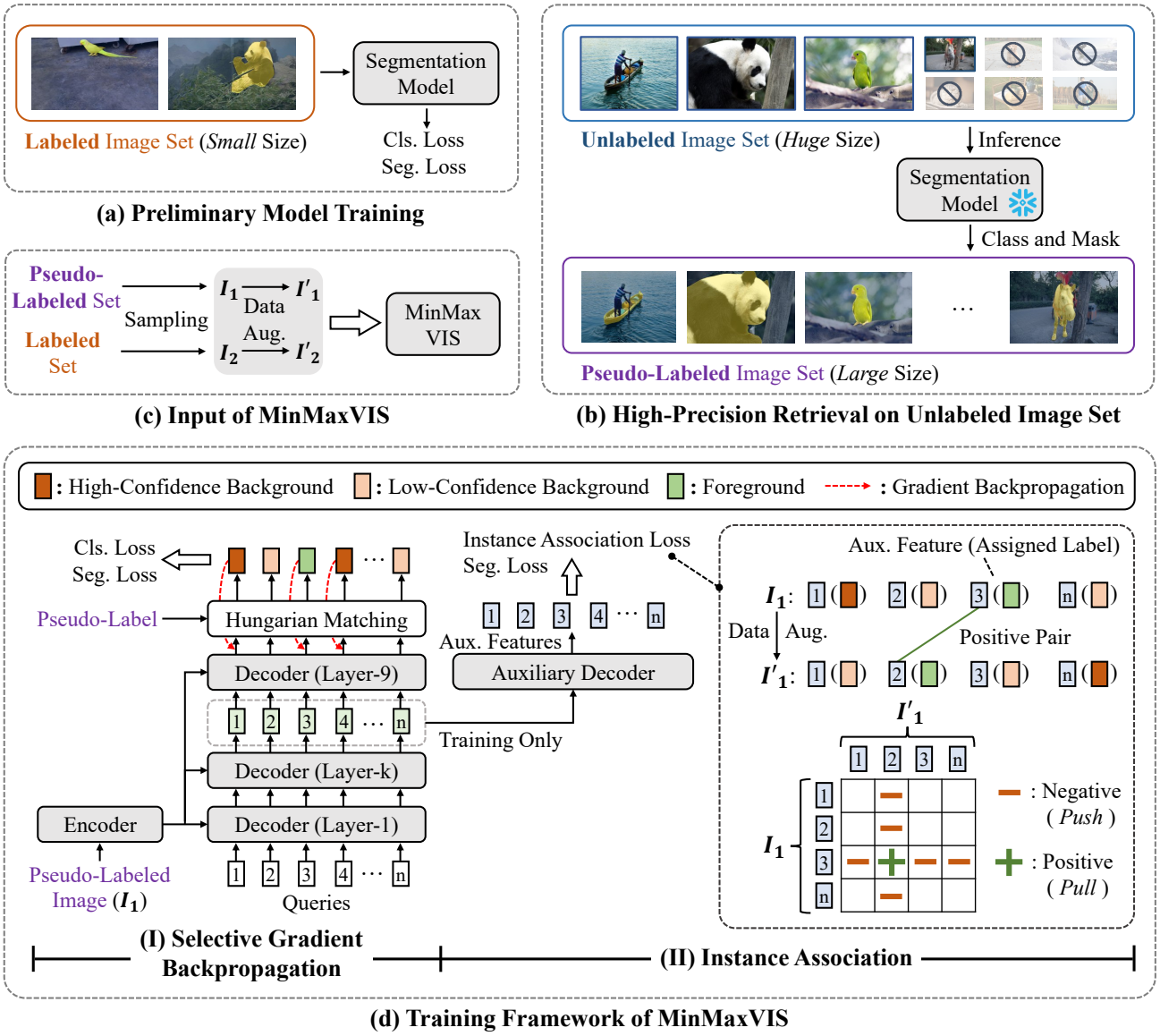


Figure 2. Overview of the MinMaxVIS framework, consisting of three main stages: (a) Preliminary segmentation model training on a small labeled set; (b) High-precision retrieval from a large unlabeled image dataset to create a pseudo-labeled set containing only high-confidence samples; (c) Input preparation for MinMaxVIS, incorporating both labeled and pseudo-labeled sets; (d) MinMaxVIS employs an encoder-decoder architecture with (I) selective Gradient Backpropagation to mitigate noisy pseudo-labels and (II) an auxiliary decoder with an instance association loss applied on augmented image pairs. *The process for generating auxiliary features for $\boldsymbol{I}'$, the augmented version of the original image $\boldsymbol{I}$, is identical to that of $\boldsymbol{I}$. For simplicity, we omit the illustration of processing $\boldsymbol{I}'$.*

stage through selective gradient backpropagation, as discussed in Section 3.3.

Images without any predictions meeting the confidence threshold τ are discarded. This process yields a filtered pseudo-labeled set, $\mathcal{D}_{\text{pseudo}} = \{(\boldsymbol{I}_j, \widehat{\boldsymbol{L}}_j)\}_{j=1}^{\widehat{M}}$, where $\widehat{M}$ represents the number of images in the pseudo-labeled set after filtering. This curated set $\mathcal{D}_{\text{pseudo}}$, along with the small labeled dataset $\mathcal{D}_{\text{labeled}}$, serves as the foundation for the final MinMaxVIS training stage, providing high-confidence, target-domain content. Notably, $\widehat{M} \ll M$, indicating that $\mathcal{D}_{\text{pseudo}}$ is significantly smaller than $\mathcal{D}_{\text{unlabeled}}$, which enhances the training efficiency of MinMaxVIS.

3.3. MinMaxVIS Training

MinMaxVIS is trained on the combination of a small labeled dataset $\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{I}_i, \boldsymbol{L}_i)\}_{i=1}^{N}$, and a large pseudo-labeled dataset $\mathcal{D}_{\text{pseudo}} = \{(\boldsymbol{I}_j, \widehat{\boldsymbol{L}}_j)\}_{j=1}^{\widehat{M}}$, with $N \ll \widehat{M}$. As illustrated in Figure 2(d), MinMaxVIS adopts an encoder-

decoder architecture with an additional auxiliary decoder designed specifically to enhance instance association.

Encoder-Decoder. As illustrated in Figure 2(d), the encoder (ResNet-50 or Swin-L) processes each input image—either from the labeled set $\mathcal{D}_{\text{labeled}}$ or the pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$—to extract multi-resolution image features. Following the approach of MinVIS [18] and Mask2Former [8], these extracted features serve as inputs to the decoder, which consists of K layers ($K = 9$ by default). The decoder begins with n learnable object queries, which are progressively refined across each decoder layer. Each layer consists of three components: 1) a self-attention layer; 2) a cross-attention layer, where image features serve as keys and values, and object queries (or query features) act as queries; and 3) an FFN layer. The output of each decoder layer undergoes Hungarian Matching to classify queries as either foreground or background. Foreground queries are supervised with both classification and segmentation losses, while background queries are supervised only with classification loss. The loss functions used are consistent with those in MinVIS [18].

To facilitate learning VIS on pseudo-labeled images, we apply a selective gradient backpropagation strategy to each layer of the main decoder, helping to mitigate the impact of false negatives in the pseudo-labels. Furthermore, we incorporate an auxiliary branch parallel to the main decoder to improve instance association learning throughout MinMaxVIS training.

Selective Gradient Backpropagation. This strategy is applied exclusively to images from the pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$. As outlined in Section 3.2, our high-precision retrieval process only retains predictions with exceptionally high confidence scores. While this approach ensures the precision of pseudo-labeled instances, it also introduces training noise from false negatives (missed instances). The core idea of selective gradient backpropagation is to utilize the model's background confidence scores to identify queries that may correspond to false positives and detach their gradients during training, as shown in Figure 2(d.I).

Concretely, given a pseudo-labeled image $\boldsymbol{I}_j \in \mathcal{D}_{\text{pseudo}}$ with its corresponding pseudo-label $\widehat{\boldsymbol{L}}_j$, MinMaxVIS encodes $\boldsymbol{I}_j$ along with n object queries. Each of the K decoder layers then produces n query features. Let $\{\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_n^k\}$ represent the n query features generated by the k-th decoder layer ($1 \leq k \leq K$). Hungarian matching is applied to assign a label (foreground instance or background) to each query in $\{\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_n^k\}$, based on the alignment of their confidence scores, segmentation masks, and the pseudo-label $\widehat{\boldsymbol{L}}_j$.

For queries assigned as background, their confidence scores are used to identify potential false negatives—queries with lower background confidence scores are more likely to be false negatives. For convenience, let $\{\overline{\boldsymbol{q}}_1^k, \ldots, \overline{\boldsymbol{q}}_{n'}^k\}$ denote the set of queries assigned to a background label, where n' is the size of this set, and $\{\overline{b}_1, \ldots, \overline{b}_{n'}\}$ denote their corresponding background confidence scores. To reduce the impact of noise in pseudo-labels, uncertain background queries are assigned lower gradient weights. Selective gradient backpropagation (SGB) is applied to $\{\overline{\boldsymbol{q}}_1^k, \ldots, \overline{\boldsymbol{q}}_{n'}^k\}$. Formally, the SGB loss function is defined as:

$$\mathcal{L}_{SGB} = \frac{1}{n'} \sum_{i=1}^{n'} w_i * \mathcal{L}_{cls}(\overline{b}_i), \tag{1}$$

where $\mathcal{L}_{cls}$ is the classification loss, and w_i represents the weight of each background query. We explore several variants for w_i:

- Confidence-weight: $w_i = \overline{b}_i$.
- Squared-confidence-weight: $w_i = (\overline{b}_i)^2$.
- Truncation-weight: $w_i = \mathbb{1}_{[b_i \geq \beta]}$, where β is a threshold ensuring that only queries with confidence scores above β contribute to the background classification. By default, we use this configuration.

For foreground queries, the loss functions remain unchanged. The SGB loss function is applied at each of the K decoder layers.

Instance Association. VIS requires consistent tracking of the same object instance across frames. Unlike conventional VIS datasets that provide association labels for training videos, our model is trained on static images without such labels. To enable instance association learning, we simulate video-like data by generating paired images through augmentations. As depicted in Figure 2(d.II), an auxiliary decoder, running in parallel with the main decoder, processes these image pairs (an original image and its augmented version) to aid in instance association. The auxiliary decoder consists of six layers, each mirroring the structure of a main decoder layer. The output from the $\overline{K}$-th layer of the main decoder is used as input to the auxiliary decoder.

Given an image $\boldsymbol{I}$ from either the labeled set $\mathcal{D}_{\text{labeled}}$ or the pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$, with its corresponding label $\boldsymbol{L}$ or pseudo-label $\widehat{\boldsymbol{L}}$, we pass it, along with the n object queries, through the encoder, main decoder, and auxiliary decoder. The auxiliary decoder produces n auxiliary features, denoted as $\{\boldsymbol{a}_1^{\boldsymbol{I}}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}}\}$. We utilize the Hungarian matching results from the final layer of the main decoder to assign a foreground or background label to each of the n auxiliary features (see Figure 2(d.II). Similarly, for the augmented image $\boldsymbol{I}'$, we use the same process to generate its auxiliary features $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ and the assignment results. Since $\boldsymbol{I}'$ is the augmented version of $\boldsymbol{I}$, we have established associations for the same instances across the two images. For each foreground query in $\{\boldsymbol{a}_1^{\boldsymbol{I}}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}}\}$,

we locate the matching foreground query in $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ that corresponds to the same instance label. We then maximize the feature similarity for these matched pairs, while minimizing similarity for non-matching pairs to reinforce instance association.

Formally, the instance association loss function for image $\boldsymbol{I}$ is defined as:

$$\begin{aligned}\mathcal{L}_{IA}^{\boldsymbol{I}} = &-\frac{1}{G}\sum_{g=1}^{G}\log\sigma(\boldsymbol{a}_g^{\boldsymbol{I}}\cdot\mathcal{M}(\boldsymbol{a}_g^{\boldsymbol{I}})) \\ &-\frac{1}{G(n-1)}\sum_{g=1}^{G}\sum_{h=1}^{n-1}\log\sigma(-\boldsymbol{a}_g^{\boldsymbol{I}}\cdot\mathcal{N}(\boldsymbol{a}_g^{\boldsymbol{I}})_h),\end{aligned} \tag{2}$$

where G represents the total number of foreground instances in image $\boldsymbol{I}$, n is the total number of queries, and $\sigma(\cdot)$ denotes the sigmoid function. Here, $\boldsymbol{a}_g^{\boldsymbol{I}}$ represents the auxiliary feature of the g-th foreground instance in $\boldsymbol{I}$, $\mathcal{M}(\boldsymbol{a}_g^{\boldsymbol{I}})$ identifies the matching auxiliary feature in $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ that corresponds to the same instance label as $\boldsymbol{a}_g^{\boldsymbol{I}}$, and $\mathcal{N}(\boldsymbol{a}_g^{\boldsymbol{I}})$ identifies the $n-1$ non-matching auxiliary features in $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ for $\boldsymbol{a}_g^{\boldsymbol{I}}$. The first term in Eq.2 maximizes feature similarity for matching pairs, while the second term minimizes similarity for non-matching pairs.

The loss in Eq.2 is applied to the original image $\boldsymbol{I}$. Similarly, a symmetric loss $\mathcal{L}_{IA}^{\boldsymbol{I}'}$ is applied to the augmented image $\boldsymbol{I}'$. The final instance association loss is defined as $\mathcal{L}_{IA} = \mathcal{L}_{IA}^{\boldsymbol{I}} + \mathcal{L}_{IA}^{\boldsymbol{I}'}$, and it is applicable to images from both the labeled and pseudo-labeled sets. Additionally, we apply the same segmentation loss used in the main decoder to the auxiliary decoder to enhance training stability.

3.4. Inference

The auxiliary decoder, equipped with the instance association loss, is used to enhance the discriminative ability of query features generated by the main decoder, allowing them to effectively associate the same instances across frames. During inference, the auxiliary decoder is removed, and only the encoder-decoder architecture is retained.

MinMaxVIS employs streaming inference (a.k.a online inference). For a continuous video stream, each frame is processed by MinMaxVIS (excluding the auxiliary decoder) to produce n predictions, each corresponding to an object query. Each prediction includes a classification score and a mask prediction. Finally, Hungarian matching is applied to associate predictions across consecutive frames, as illustrated in Figure 3.

4. Experiment

Datasets. We evaluate our approach on three datasets: YouTube-VIS 2019 [45], YouTube-VIS 2021 [46], and OVIS [29], using two data settings for each. For YouTube-VIS 2019, which has 61,845 labeled frames in its training set, we experiment with 1% and 2% of the data, resulting in 618 and 1,236 images, respectively. For YouTube-VIS 2021, with a training set of 90,160 frames, we also use 1% and 2% data settings, yielding 901 and 1,803 images, respectively. Given the complexity of OVIS, we experiment with 5% and 10% of its training set, which consists of 42,149 images, resulting in 2,107 and 4,214 images, respectively. For the general-domain unlabeled dataset, we randomly sample a subset from SA-1B [23], consisting of 2.8 million images.

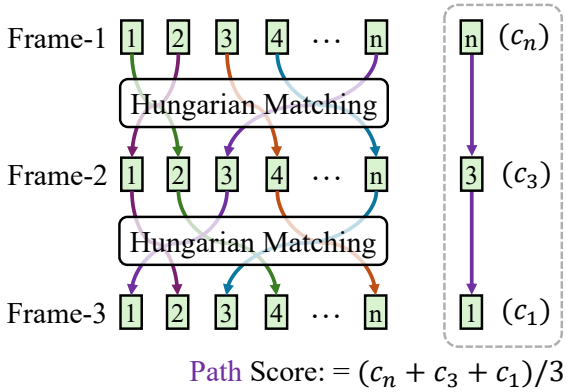


Figure 3. Illustration of the inference process (example with three frames). Each frame is independently processed by MinMaxVIS to produce n query features (indicated by green rectangles) from the main decoder. Each query feature generates a classification score c and a mask prediction. Hungarian matching is then applied between pairs of consecutive frames to associate predictions based on the similarity of query features, resulting in n paths across the frames. The path score is computed by averaging the classification scores along the path. This path score is then used as the final classification score for all predictions along the path.

Implementation Details. For YouTube-VIS 2019 (1%) and YouTube-VIS 2019 (2%), we retrieve 25,669 and 28,033 pseudo-labeled images from SA-1B using the high-precision retrieval method described in Section 3.2. For YouTube-VIS 2021 (1%) and YouTube-VIS 2021 (2%), the retrieved pseudo-labeled images are 24,088 and 25,807, respectively. For OVIS (5%) and OVIS (10%), we retrieve 15,966 and 17,006 pseudo-labeled images. In high-precision retrieval, we set $\tau = 0.99$. For categories with fewer than 1,000 retrieved pseudo-labeled images, we reduce τ to 0.9. In MinMaxVIS, the number of object queries n is set to 200 for the Swin-L backbone and 100 for the ResNet-50 backbone. When using the "truncation-weight" in Eq.1, we set $\beta = 0.5$. In each training batch, the ratio of labeled to pseudo-labeled images is set to 1:4. The total batch size is 64, including both original and augmented images. MinMaxVIS is trained for 6,000 iterations with an initial learning rate of 1e-4, which is reduced by a factor of 0.1 after 4,000 iterations.

Evaluation. We evaluate our approach on the validation

Method	Backbone	Setting	YouTube-VIS 2019			YouTube-VIS 2021			OVIS		
			1%	2%	100%	1%	2%	100%	5%	10%	100%
VITA [14]	ResNet-50	Offline	32.3	37.0	49.8	28.2	33.5	45.7	8.2	10.4	19.6
DVIS [51]	ResNet-50	Online	23.9	28.3	51.2	27.0	32.3	46.4	10.9	20.6	30.4
DVIS++ [52]	ResNet-50	Online	29.8	36.1	55.5	31.5	33.9	50.0	10.7	14.4	37.2
GenVIS [15]	ResNet-50	Online	31.9	32.7	50.0	26.4	31.4	47.1	11.3	14.7	35.8
DVIS-DAQ [55]	ResNet-50	Online	33.4	38.9	55.2	32.3	33.0	50.4	10.8	16.5	38.7
Image-Driven Approach											
MinVIS [18]	ResNet-50	Online	33.7	37.8	47.4	29.6	31.1	44.2	22.2	22.5	25.0
MinMaxVIS (Ours)	ResNet-50	Online	**46.6**	**50.0**	-	**44.6**	**45.7**	-	**24.7**	**26.8**	-
VITA [14]	Swin-L	Offline	52.9	56.0	63.0	43.9	47.6	57.5	13.8	18.8	27.7
DVIS [51]	Swin-L	Online	36.8	36.2	63.9	43.9	49.6	58.7	23.6	29.5	46.0
DVIS++ [52]	VIT-L	Online	49.2	52.4	67.7	44.5	47.1	62.3	18.3	26.1	49.6
GenVIS [15]	Swin-L	Online	53.8	56.6	64.0	45.0	46.1	59.6	14.7	21.5	45.2
DVIS-DAQ [55]	VIT-L	Online	51.8	57.7	68.3	46.6	50.4	62.4	15.9	24.0	53.7
Image-Driven Approach											
MinVIS [18]	Swin-L	Online	56.2	58.5	61.6	49.8	52.4	55.3	36.1	37.0	39.4
MinMaxVIS (Ours)	Swin-L	Online	**60.9**	**62.2**	-	**54.1**	**55.6**	-	**37.5**	**39.2**	-

Table 1. Performance comparison (mAP in %) of various methods on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS datasets across different labeled data settings. MinMaxVIS, built upon MinVIS, effectively leverages unlabeled data to achieve superior performance in low-data regimes, significantly outperforming MinVIS. Both MinMaxVIS and MinVIS are image-driven approaches. MinMaxVIS even achieves results comparable to or exceeding full-set MinVIS (100% labeled data) across multiple settings.

sets of YouTube-VIS 2019, YouTube-VIS 2021, and OVIS. Consistent with standard practice, mAP is used as the evaluation metric.

4.1. Main Results

Table 1 presents the performance comparison of various methods on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS datasets across different labeled data settings: 1% and 2% for YouTube-VIS datasets, and 5% and 10% for OVIS. We evaluate MinMaxVIS with both ResNet-50 and Swin-L backbones and compare it to recent approaches using the same backbones. For each method, we reproduce its performance in each low-data regime. For approaches that require video input, we simulate video sequences by applying data augmentations to individual images. Adapting these approaches to leverage unlabeled image data directly is challenging, as they differ in network architectures.

MinMaxVIS, built on MinVIS [18], falls within the category of image-driven approaches. Our comparison with MinVIS reveals two key findings:

- *Effective Use of Unlabeled Data.* MinMaxVIS effectively leverages information from unlabeled data, significantly outperforming low-data MinVIS with both Swin-L and ResNet-50 backbones across all three datasets. For example, MinMaxVIS with a ResNet-50 backbone achieves 12.9% and 12.2% higher mAP than MinVIS under the YouTube-VIS 2019 (1%) and YouTube-VIS 2019 (2%) settings, respectively.

Component	mAP
Baseline (MinVIS [18])	58.5
+High-Precision Retrieval	60.3
+Selective Gradient Backpropagation	61.4
+Instance Association	62.2

Table 2. Main components of MinMaxVIS.

- *Competitive Performance with Full-Set MinVIS.* MinMaxVIS also outperforms or achieves comparable results to full-set MinVIS (using 100% labeled data) across all settings. For instance, with only 2% labeled data from YouTube-VIS 2019 and 2021, MinMaxVIS with a ResNet-50 backbone surpasses full-set MinVIS with the same backbone by 2.6% and 1.5% mAP, respectively.

4.2. Ablation Studies

For all ablation studies, we use MinMaxVIS with a Swin-L backbone on YouTube-VIS 2019 using 2% labeled data.

Main Components. Table 2 presents the ablation study results showing the impact of each main component in MinMaxVIS. Starting from the baseline (MinVIS [18]), high-precision retrieval gathers a pseudo-labeled set with strong alignment to the target domain. Leveraging this set alongside the small labeled set boosts the mAP from 58.5% to 60.3%. The inclusion of selective gradient backpropagation further improves performance, raising the mAP to 61.4% by effectively handling noise in the pseudo-labels. Finally,

Strategy	mAP
Confidence-Weight	60.7
Squared-Confidence-Weight	61.7
Truncation-Weight	62.2

Table 3. Ablation study on selective gradient backpropagation strategies proposed in Section 3.3.

Threshold (β)	0.0	0.3	0.5	0.7	0.9
mAP	61.4	61.2	62.2	61.6	60.1

Table 4. Impact of threshold β in truncation-weight strategy for selective gradient backpropagation.

incorporating instance association increases the mAP to 62.2%, emphasizing the importance of enhancing instance association across frames.

Selective Gradient Backpropagation. Table 3 compares different weighting strategies for selective gradient backpropagation in MinMaxVIS, as introduced in Section 3.3. The truncation-weight strategy, with a threshold β set to 0.5, achieves the highest mAP, indicating that selectively filtering out low-confidence background queries enhances performance. Table 4 examines the impact of varying the threshold β in the truncation-weight strategy for selective gradient backpropagation on mAP performance. Lower β values introduce more training noise, while higher β values may result in insufficient background classification training. A threshold of 0.5 achieves the highest mAP, balancing noise reduction with effective training.

Instance Association. In Section 3.3, we introduce an auxiliary decoder with the proposed instance association loss applied to it. Each decoder layer generates $2 \times n$ auxiliary query features for an image pair, which includes an original image and its augmented version. First, we identify foreground query features that correspond to the same instances across both images, forming several positive pairs. The remaining queries are treated as background, and we construct a set of negative pairs, where each pair consists of a foreground query feature from one image and a background query feature from the other. By default, as shown in Figure 2(d.II), we apply a "Pull(F-F) + Push(F-B)" strategy, where we pull the similarities of positive pairs (foreground-foreground) closer and push the similarities of negative pairs (foreground-background) further apart. Here, we explore an alternative approach, termed "Pull(F-F) + Push(F-B) + Push(B-B)". In this variant, we additionally push similarities between background query features from both images (background-background). However, as shown in Table 5, this approach underperforms compared to our default setting, as it likely introduces unnecessary separation between background queries, which does not contribute effectively to instance association.

Strategy	mAP
Pull(F-F) + Push(F-B)	62.2
Pull(F-F) + Push(F-B) + Push(B-B)	60.2

Table 5. Study on the instance association strategies.

Layer ID	3	6	9
mAP	62.2	61.5	61.4

Table 6. Impact of main decoder layer selection on instance association performance.

Ratio	1:2	1:4	1:8	1:16
mAP	60.6	62.2	61.6	61.3

Table 7. Impact of different ratios of labeled to pseudo-labeled images within a training batch.

The input to the auxiliary decoder is derived from the output of the $\overline{K}$-th layer of the main decoder. Table 6 explores the impact of different choices for $\overline{K}$ on performance. Using features from a shallower layer of the main decoder yields the best results, as these less semantic features allow the auxiliary decoder to better capture intra-class variability, which is beneficial for instance association. In contrast, features from deeper layers of the main decoder are more influenced by the classification loss, leading to greater inter-class discrimination but reduced intra-class variability. Since instance association relies on preserving intra-class similarities, shallower layer features are more suitable for this purpose.

Ratio of Labeled to Pseudo-Labeled Images. Table 7 examines the impact of different ratios of labeled to pseudo-labeled images within a training batch. A ratio of $1:4$ produces the best results.

5. Conclusion

This paper presents MinMaxVIS, an efficient framework for VIS that reduces reliance on fully labeled video data by combining a small set of labeled images from the target domain with a large, unlabeled general-domain dataset. MinMaxVIS operates through three stages: (1) preliminary segmentation model training on the small labeled set, (2) high-precision retrieval to create a pseudo-labeled set from the unlabeled data, and (3) training on both labeled and pseudo-labeled sets. To address the challenges of training VIS on static images, MinMaxVIS incorporates selective gradient backpropagation to mitigate noise in pseudo-labels and an auxiliary decoder to improve instance association via augmented image pairs. Experimental results show that MinMaxVIS outperforms the full-data MinVIS, an image-driven method, across multiple settings, achieving higher mAP with substantially less labeled data.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 158–177, 2020. 2

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[3] 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. 2

[4] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 1–18, 2020. 2, 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019. 2

[7] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *CoRR*, abs/2112.10764, 2021. 2

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 3, 5

[9] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 660–676. Springer, 2020. 2

[10] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022. 2

[11] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34: 21898–21909, 2021. 2

[12] Junjie He, Pengyu Li, Yifeng Geng, and Xuansong Xie. Fastinst: A simple query-based model for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23663–23672, 2023. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3

[14] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 2, 3, 7

[15] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023. 2, 7

[16] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 2

[17] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021. 2

[18] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 1, 2, 3, 5, 7

[19] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019. 2

[20] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 1, 2

[21] Hanjung Kim, Jaehyun Kang, Miran Heo, Sukjun Hwang, Seoung Wug Oh, and Seon Joo Kim. VISAGE: video instance segmentation with appearance-guided enhancement. In *European Conference on Computer Vision*, pages 93–109, 2024. 1, 2, 3

[22] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 2

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 6

[24] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2

[25] Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Tcovis: Temporally consistent online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1097–1107, 2023. 1, 2, 3

[26] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1739–1748, 2021. 2

[27] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9816–9825, 2021. 2, 3

[28] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 2

[29] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2, 6

[30] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788, 2016. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2

[32] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 2

[33] Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13926–13935, 2021. 2

[34] 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

[35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9626–9635, 2019. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[37] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer, 2020. 2

[38] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 2

[39] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 2

[40] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2(3): 4, 2021. 1, 2, 3

[41] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022. 2, 3

[42] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020. 2

[43] 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

[44] Kun Yan, Xiao Li, Fangyun Wei, Jinglu Wang, Chenbin Zhang, Ping Wang, and Yan Lu. Two-shot video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2257–2267, 2023. 2

[45] 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. 2, 3, 6

[46] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 2, 6

[47] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 8043–8052, 2021. 2, 3

[48] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2885–2895, 2022. 2

[49] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6861–6869, 2021. 2

[50] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[51] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 3, 7

[52] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 3, 7

[53] Zicheng Zhang, Tong Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, QiXiang Ye, and Wei Ke. Language-driven visual consensus for zero-shot semantic segmentation. *arXiv preprint arXiv:2403.08426*, 2024. 2

[54] Jinjing Zhao, Fangyun Wei, and Chang Xu. Hybrid proposal refiner: Revisiting detr series from the faster r-cnn perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17416–17426, 2024. 2

[55] Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Dvis-daq: Improving video segmentation via dynamic anchor queries. *arXiv preprint arXiv:2404.00086*, 2024. 3, 7

[56] 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. 3