International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02195-4

Video Instance Segmentation in an Open-World

Omkar Thawakar[1] · Sanath Narayan[2] · Hisham Cholakkal[1] · Rao Muhammad Anwer[1,3] · Salman Khan[1] · Jorma Laaksonen[3] · Mubarak Shah[4] · Fahad Shahbaz Khan[1,5]

Received: 15 December 2023 / Accepted: 4 July 2024

Abstract
Existing video instance segmentation (VIS) approaches generally follow a closed-world assumption, where only seen category instances are identified and spatio-temporally segmented at inference. Open-world formulation relaxes the close-world static-learning assumption as follows: (a) first, it distinguishes a set of known categories as well as labels an unknown object as 'unknown' and then (b) it incrementally learns the class of an unknown as and when the corresponding semantic labels become available. We propose the first open-world VIS approach, named OW-VISFormer, that introduces a novel feature enrichment mechanism and a spatio-temporal objectness (STO) module. The feature enrichment mechanism based on a light-weight auxiliary network aims at accurate pixel-level (unknown) object delineation from the background as well as distinguishing category-specific known semantic classes. The STO module strives to generate instance-level pseudo-labels by enhancing the foreground activations through a contrastive loss. Moreover, we also introduce an extensive experimental protocol to measure the characteristics of OW-VIS. Our OW-VISFormer performs favorably against a solid baseline in OW-VIS setting. Further, we evaluate our contributions in the standard fully-supervised VIS setting by integrating them into the recent SeqFormer, achieving an absolute gain of 1.6% AP on Youtube-VIS 2019 val. set. Lastly, we show the generalizability of our contributions for the open-world detection (OWOD) setting, outperforming the best existing OWOD method in the literature. Code, models along with OW-VIS splits are available at https://github.com/OmkarThawakar/OWVISFormer.

Keywords Open-world segmentation · Video instance segmentation · Object-detection · Video object detection · Transformers

1 Introduction

Video instance segmentation (VIS) strives to simultaneously classify, segment and track all object instances from a set of semantic classes in a given video. The problem is challenging since a diverse set of objects are desired to be accurately tracked and segmented despite real-world issues such as, fast motion, large intra-class variation and background clutter. Most existing VIS approaches (Wu et al., 2022; Wang et al., 2021; Yang et al., 2021; Ke et al., 2021) typically follow a close-world assumption, *i.e.*, all object categories to be detected are provided during the training and only seen object classes are spatio-temporally segmented at inference. *E.g.*, existing VIS methods evaluated on the popular Youtube-VIS benchmark (Yang et al., 2019; Xu et al., 2021) assume that annotated (known) instances of all 40 semantic categories to be segmented and tracked are available during training. Here, such a training scheme treats unannotated (unknown) objects as background. Therefore, the closed-world assumption poses issues to existing VIS methods when recognizing novel (unknown) object class instances.

The open-world problem formulation (Joseph et al., 2021; Gupta et al., 2022) relaxes the closed-world assumption by enabling the VIS model at each training episode to identify unknown object category instances as belonging to the 'unknown' class while simultaneously learning to spatio-temporally segment a given set of 'known' objects. Afterwards, these identified unknowns can be passed to an oracle, which annotates a set of object categories of interest. Then, the VIS model takes into account these new knowns

Communicated by Hong Liu.

✉ Omkar Thawakar
omkar.thawakar@mbzuai.ac.ae

1 Mohamed bin Zayed University of AI, Abu Dhabi, UAE

2 Technology Innovation Institute, Abu Dhabi, UAE

3 Aalto University, Espoo, Finland

4 University of Central Florida, Orlando, USA

5 Linköping University, Linköping, Sweden

Published online: 30 July 2024

to incrementally update its knowledge without requiring to be retrained from scratch using the previous known object categories. However, such an open-world formulation poses additional issues over the standard VIS problem challenges by requiring the model to also (i) distinguish unknown objects *and* (ii) recognize them later with the arrival of progressive training data, in a unified manner. Although the open-world setting has been explored recently for detection (Joseph et al., 2021; Gupta et al., 2022) and image segmentation (Kuniaki et al., 2022; Wang et al., 2022), to the best of our knowledge, we are the first to investigate the problem and introduce a novel approach for *open-world video instance segmentation* (OW-VIS).

When designing an OW-VIS framework, one plausible way is to extend a fully-supervised VIS approach by introducing a pseudo-labeling scheme to identify potential unknown objects. These potential pseudo-unknowns along with the ground-truth known instances can then be utilized to learn a foreground-background class-agnostic separation as well as performing class-specific known vs.unknown instance classification. Existing fully-supervised VIS approaches typically employ an ImageNet (Russakovsky et al., 2015) pre-trained classification backbone for multi-scale feature extraction to be used in the encoder. The same features can also be utilized in a bottom-up pseudo-labeling scheme in OW-VIS. However, such a pre-trained classification-based framework is likely to struggle in the OW-VIS paradigm (see Fig. 1), where the aim is to accurately distinguish a class-agnostic unknown object from the background as well as class-specific known categories at the *pixel-level*. To achieve such an accurate pixel-level (unknown) object delineation from the background, we argue that dedicated shallow features are especially desired to complement the high-level semantic pre-trained features. Moreover, since the selection of pseudo-unknowns relies on the activation's in the selected feature map, it is further desired to enhance their strengths in the foreground regions (known and unknown) for learning better objectness priors.

Expanding on the motivation for extending OWOD (Joseph et al., 2021; Gupta et al., 2022) to OW-VIS, our novel approach addresses several key challenges and directions inherent in the open-world video instance segmentation domain. First, we aim to develop algorithm capable of effectively handling the dynamic nature of video data, including object motion, occlusions, and varying perspectives over time, while simultaneously accommodating unknown object categories. Secondly, by introducing OW-VIS as a new benchmark, we seek to foster research and development efforts focused on advancing the state-of-the-art in video instance segmentation under open-world conditions. Moreover, we anticipate that OW-VIS will serve as a testbed for evaluating the scalability, generalization, and adaptability of existing video instance segmentation models across diverse datasets and real-world scenarios. By highlighting these potential problems and directions, our work aims to stimulate further innovation and collaboration within the research community, ultimately leading to the development of more robust and versatile video instance segmentation solutions.

1.1 Contributions

We propose an OW-VIS approach, named OW-VISFormer, that introduces (*i*) a feature enrichment mechanism, which aims to better differentiate class-agnostic foreground vs. background as well as aid in class-specific known vs.unknown instance classification and (*ii*) a spatio-temporal objectness (STO) module that strives to identify candidate pseudo-unknowns. The feature enrichment mechanism is based on a light-weight auxiliary network that is trained from *scratch* and generates dedicated shallow features to complement the high-level semantic standard *pre-trained* features. The resulting extended features are enriched by the encoder and then use in the STO module. Our STO module employs a contrastive loss that distinguishes candidate pseudo-unknowns from the background by enhancing the foreground activations. As a result, improved video instance mask predictions are obtained for both the known and unknown classes (see Fig. 1). Furthermore, we introduce carefully curated open-world splits of Youtube-VIS dataset for a rigorous evaluation of OW-VIS problem.

Our extensive quantitative and qualitative evaluations demonstrate the effectiveness of the proposed OW-VISFormer leading to consistent improvement in performance, compared to the baseline. In addition, we also validate our proposed contributions in the standard fully-supervised VIS problem setting by introducing them into the recent SeqFormer (Wu et al., 2022), achieving an absolute gain of 1.6% in overall AP on the Youtube-VIS 2019 val. set. Lastly, we demonstrate the generalizability of our two contributions for the open-world detection (OWOD) problem setting by integrating them into the recent OW-DETR (Gupta et al., 2022). On the challenging MS COCO OWOD split, our approach outperforms the recent OW-DETR on all the tasks for both the 'known' and 'unknown'.

2 Open-World Video Instance Segmentation

2.1 Problem Formulation

Let $\mathcal{D}^t = \{\mathcal{V}^t, \mathcal{Y}^t\}$ be a progressive dataset at time t containing N_t videos $\mathcal{V}^t = \{V_1, \cdots, V_{N_t}\}$ with corresponding labels $\mathcal{Y}^t = \{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_{N_t}\}$. Here, $V_i \in \mathcal{R}^{L_i \times 3 \times H \times W}$ denotes a video of length L_i frames with spatial resolution $H \times W$, while $\boldsymbol{Y}_i = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ denotes the ground-truth mask

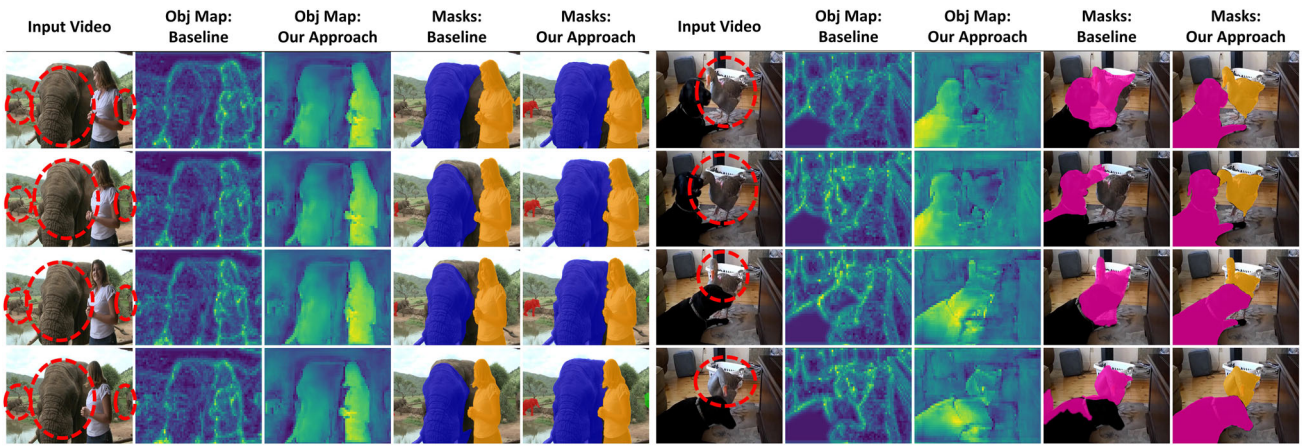

Fig. 1 Video instance segmentation in an open-world (OW-VIS) setting illustrating the first step of identifying 'known' and 'unknown' objects. For each example video frame, we show the corresponding objectness map obtained from backbone features in the case of the baseline (col 2 and 7) and the spatio-temporal objectness (STO) module output map for our OW-VISFormer (col 3 and 8). Moreover, we show the respective output segmentation mask for each frame in case of the baseline (col 4 and 9) and OW-VISFormer (col 5 and 10). In these example videos, the unknown (in red dashed line) objects are three *elephants* (on the left), and the *duck* (on the right). The known objects in these videos are *person* (on the left) and *dog* (on the right). Compared to the baseline, OW-VISFormer accurately segments all the 'unknown' and 'known' object instances. Best viewed zoomed in. Additional results are presented in the supplementary

annotations of a set of K object instances present in the video. Here, $\boldsymbol{y}_j \in \mathcal{R}^{L_i \times h \times w}$ denotes the set of masks predicted for an object instance j in L_i frames of a video V_i. Let $\mathcal{K}^t = \{1, 2, \cdots, C\}$ denote the known object categories at time t and $\mathcal{U}^t = \{C+1, \cdots\}$ be a set of unknown classes that are likely to be encountered at test time.

As discussed earlier, our OW-VIS first learns a model $\mathcal{M}^t$ that can spatio-temporally segment an unseen class instance at time t as belonging to the unknown class (denoted by label 0) in addition to segmenting the instances of previously encountered known classes $\mathcal{K}^t$. Next, a set of these unknown instances $\boldsymbol{U}^t$ identified by $\mathcal{M}^t$ are then taken as input to an oracle, which labels n novel classes of interest and provides new training examples for the corresponding n classes. Then, these n classes are considered as known and added to the previously known C classes, such that $\mathcal{K}^{t+1} = \mathcal{K}^t + \{C+1, \cdots, C+n\}$. Then, $\mathcal{M}^t$ is incrementally trained to obtain an updated model $\mathcal{M}^{t+1}$, which can spatio-temporally segment all object instances belonging to classes in $\mathcal{K}^{t+1}$ without forgetting the previously learned classes in $\mathcal{K}^t$. This cycle of spatio-temporally segmenting unknown instances and incremental learning of new knowledge continues over the model's life-time.

2.2 Baseline OW-VIS Framework

We base our approach on the recent fully-supervised (FS) SeqFormer (Wu et al., 2022). It utilizes a standard pre-trained backbone network for multi-scale feature extraction followed by a deformable transformer (Zhu et al., 2021) and a segmentation block for video instance mask prediction. Here, an M-frame video clip $\mathbf{v} \subset V_i$ is input to a pre-trained backbone network. Then, the resulting multi-scale features of each frame are input to an encoder that outputs feature maps of the same size as the input. The encoder output features together with q learnable instance query embeddings $\boldsymbol{Q}^I$ are input to the decoder. Consequently, the decoder outputs q instance features $\boldsymbol{F}^I$ that are then used for video mask prediction.

A straightforward way to extend the above FS SeqFormer to OW-VIS (Sect. 2.1) is to introduce a pseudo-labeling scheme for selecting potential unknown objects followed by learning to categorize these identified pseudo-unknowns into a single unknown class. One way to design such a pseudo-labeling scheme is to utilize object proposals[1] having high activations in their corresponding regions of the backbone feature maps as candidates for unknown class, as in Gupta et al. (2022). These pseudo-unknowns can be then used along with ground-truth known instances to learn a foreground-background *class-agnostic* separation as well as perform *class-specific* known vs.unknown instance classification. We refer to this as our baseline OW-VIS framework.

3 Proposed OW-VIS Framework

Overall Architecture: Fig. 2a shows the overall OW-VISFormer framework with a standard pre-trained backbone. To circumvent using a fully-supervised ImageNet

[1] Proposals are obtained from the instance features $\boldsymbol{Q}^I$ and only those remaining after selecting the ground-truth class instances through Hungarian matching are considered.

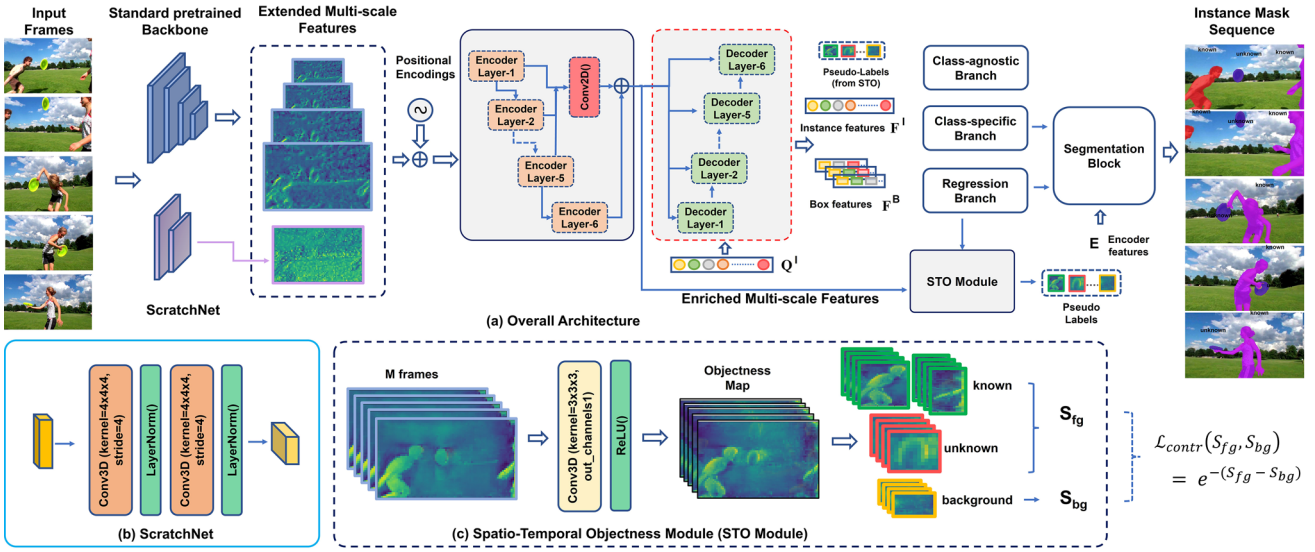

Fig. 2 **a** Overall architecture of the proposed OW-VISFormer framework. It comprises a standard pre-trained backbone, a light-weight Scratch-Net, an encoder-decoder followed by class-specific, class-agnostic and regression branches along with a spatio-temporal objectness (STO) module. Different from the pre-trained backbone, **b** the Scratch-Net is trained from scratch with random weight initialization during the OW-VIS training. Its light-weight architecture comprises two 3D convolution and normalization layers. The resulting feature maps are then integrated with the standard backbone features to construct extended multi-scale features which are then input to the encoder that outputs enriched features. Within the decoder, the instance queries cross-attend to the enriched features. The instance and box features output by the decoder are input to the class-specific (multi-class), class-agnostic and regression branches. To effectively learn the class-agnostic and class-specific branches with the unknown instances in the OW-VIS setting, we introduce a **c** spatio-temporal objectness (STO) module trained with a contrastive loss ($\mathcal{L}_{contr}$) for generating instance-level pseudo-labels. Consequently, the instance features from the decoder along with the encoder features are used within the segmentation block for the video instance mask prediction of known and unknown classes in the proposed OW-VISFormer framework

pre-trained network, we use the popular self-supervised ResNet50 DINO (Caron et al., 2021) ImageNet-1K backbone. For accurate pixel-level object delineation from the background, we introduce a feature enrichment mechanism that utilizes a novel light-weight ScratchNet Fig. 2b which is trained from scratch through random weight initialization. Our ScratchNet, comprising two 3D convolution and normalization layers, takes the same input frames as the standard pre-trained backbone stream. The resulting shallow features from ScratchNet are integrated with high-level semantic features from the standard pre-trained backbone to produce *extended multi-scale features*. These extended features are then input to the encoder to obtain *enriched multi-scale features*. The decoder within our OW-VISFormer framework aggregates the enriched features from all encoder layers, which are then cross-attended with the instance queries $\boldsymbol{Q}^I$. The decoder then outputs the instance and box features ($\boldsymbol{F}^I$ and $\boldsymbol{F}^B$), which are input to the three branches: class-specific, class-agnostic and regression. As discussed earlier, in the OW-VIS setting both the class-specific as well as the class-agnostic branches are required to be learned with the unknown instances. To this end, we propose a spatio-temporal objectness (STO) module, shown in Fig. 2c, comprising a 3D convolution layer for spatio-temporally aggregating the objectness information of video instances across multiple frames M. The STO module is trained with a contrastive loss ($\mathcal{L}_{contr}$) and enables improved foreground-background separability resulting in a better unknown instance mask prediction. Finally, the instance features from the decoder and enriched multi-scale features from the encoder are utilized within the segmentation block to produce video instance mask predictions for both the known and unknown classes. Unlike OW-DETR (Gupta et al., 2022), which is tailored for static image processing, our proposed OW-VISFormer is specifically designed for the videos based on Wu et al. (2022), addressing both spatial and temporal aspects of video instance segmentation. In contrast to OW-DETR (Gupta et al., 2022), our OW-VISFormer contains the ScratchNet, a lightweight auxiliary network, generates dedicated shallow features designed to enhance the delineation of moving objects across frames. Additionally, our approach employs a spatio-temporal objectness (STO) module that integrates 3D convolution to aggregate objectness information over time, further distinguishing it from OW-DETR's focus on spatial processing. These enhancements make OW-VISFormer particularly adept at handling the complexities of video instance segmentation in open-world settings, offering significant improvements over OW-DETR in tracking

and segmenting objects with high temporal fidelity. Next, we describe our feature enrichment mechanism.

3.1 Feature Enrichment Mechanism

Light-weight ScratchNet: the OW-VIS setting requires accurate *pixel-level* (unknown) object delineation from the background as well as distinguishing category-specific known classes. Therefore, shallow features capturing distinct edge and boundary information are desired to complement the high-level semantic features. A straightforward strategy to integrate the shallow feature information is to re-use the initial layer features from the standard pre-trained backbone. However, we empirically observe this to achieve inferior performance compared to the features generated using the proposed ScratchNet. We conjecture that the low-/mid-level shallow features from the initial layers of the standard pre-trained backbone are better adapted for the task of image classification, which prefers translation invariance. In contrast, the proposed light-weight ScratchNet produces dedicated shallow features that aid in accurate video instance mask prediction for both unknown as well as known categories. The light-weight ScratchNet is trained from *scratch* through random weights initialization. It comprises two layers of 3D convolution, each followed by a layer normalization, as shown in Fig. 2b. Both convolutional layers perform a non-overlapping convolution operation on their inputs with a kernel size of $4 \times 4 \times 4$ and a stride of 4. ScratchNet produces complementary shallow features, which are then integrated as an additional feature scale along with multi-scale pre-trained backbone features, resulting in extended multi-scale features.

Enriched Multi-scale Encoder Features: The extended multi-scale features are combined with positional encodings and input to a deformable encoder (Zhu et al., 2021) consisting of six layers of multi-scale deformable attention. The standard deformable encoder in our OW-VIS baseline (Sect. 2.2) outputs the attended multi-scale features from its final layer alone, which are tailored for known category mask prediction. This can likely lead to deterioration in some of the relevant features for accurate mask prediction of unknown instances in our OW-VIS setting. To alleviate this issue, we combine features from all encoder layers in order to learn enriched features for predicting accurate video masks for both known and unknown instances. To this end, multi-scale features from all but the final encoder layer are scale-wise fused through a convolution operation and added with the features of the final encoder layer, resulting in enriched multi-scale features. Such a feature fusion enables improving unknown video instance mask prediction while preserving the known category predictions. Consequently, the encoder outputs enriched multi-scale features that are better suited for the OW-VIS task. These enriched multi-scale features are then input to the decoder as well as our spatio-temporal objectness (STO) module described next.

3.2 Spatio-Temporal Objectness Module

In the OW-VIS setting, both category-specific as well as class-agnostic branches require learning with the unknown instances. To this end, we introduce a spatio-temporal objectness (STO) module (Fig. 2c) that generates instance-level pseudo-labels and is trained using a contrastive loss. The STO module $G_{STO}(\cdot)$ consists of a 3D convolutional layer with output channels equal to one. It takes the d-dimensional enriched encoder features $\boldsymbol{E}_k$ corresponding to the M frames at spatial scale $k = 1/16$ as input and outputs an objectness map $O_{map} \in \mathcal{R}^{M \times H/16 \times W/16}$.

Pseudo-labeling Unknown Instances: Given the q instance features $\boldsymbol{F}^I$ output by the decoder, we employ the Hungarian matching loss (Kuhn, 1955) that identifies the best matching queries for the K known instances in the input video. For each of the remaining $q - K$ instance predictions, their objectness scores s_i (where $i \in \{K+1, \cdots, q\}$) are computed by spatio-temporally aggregating the activation strengths of the objectness maps $O_{map} = G_{STO}(\boldsymbol{E}_k)$ within the corresponding predicted box regions across $m = [1, \cdots, M]$ frames, given by

$$s_i = \sum_{m=1}^{M} \frac{1}{h_i^m \cdot w_i^m} \sum_{x_i^m - 0.5w_i^m}^{x_i^m + 0.5w_i^m} \sum_{y_i^m - 0.5h_i^m}^{y_i^m + 0.5h_i^m} G_{STO}(\boldsymbol{E}_k), \tag{1}$$

where $\boldsymbol{b}_i^m = [x_i^m, y_i^m, w_i^m, h_i^m]$ denotes the box proposal predicted in the regression branch for the i^{th} instance in m^{th} frame. Here, (x_i^m, y_i^m), w_i^m and h_i^m denote the center, width and height, respectively. The resulting $q - K$ scores (s_i) are sorted in decreasing order and the top-p_u instances are employed as pseudo-unknown during training. Furthermore, the remaining $q - (K + p_u)$ instances are considered as background instances.

Given that the selection of pseudo-unknowns depends on the activations in the objectness map O_{map}, we introduce a contrastive loss that aims to better separate the class-agnostic foreground regions from the background in the objectness map. The foreground score S_{fg} is computed by aggregating the objectness scores s_i of the foreground video instances, *i.e.*, $i \in [1, \cdots, K + p_u]$ (both known and pseudo-unknown). Similarly, the background score S_{bg} is obtained by aggregating the objectness scores s_i of the background video instances, *i.e.*, $i \in [K + p_u + 1, \cdots, q]$. The constrastive loss is then given by

$$\mathcal{L}_{contr}\left(S_{fg}, S_{bg}\right) = e^{-(S_{fg} - S_{bg})}, \tag{2}$$

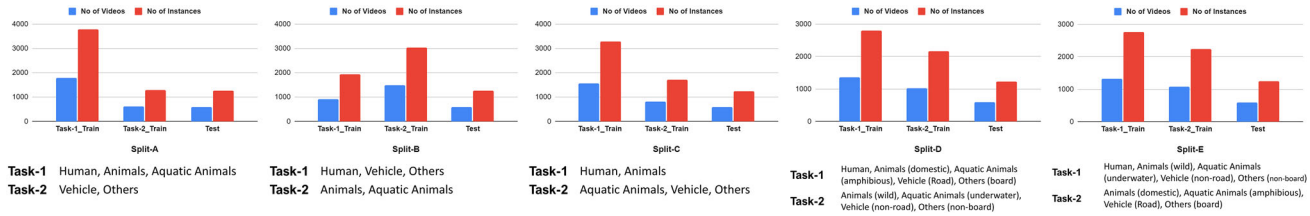

Fig. 3 The proposed task composition in our OW-VIS evaluation setting based on the Youtube-VIS dataset. Here, for each task in the corresponding split we show the number of videos and instances (objects). We construct five splits **A**–**E** each having two tasks (Task-1 and Task-2). The super-categories for each task within the splits are shows

$$\text{where} \quad S_{fg} = \sum_{i=1}^{K+p_u} s_i \quad \text{and} \quad S_{bg} = \sum_{i=K+p_u+1}^{q} s_i \,. \tag{3}$$

The resulting pseudo-unknown and background instances along with the ground-truth known instances are employed for training the class-specific and class-agnostic branches. Furthermore, as in Wu et al. (2022), the instance and box features ($\boldsymbol{F}^I$ and $\boldsymbol{F}^B$) output by the decoder, along with the enriched multi-scale features $\boldsymbol{E}$ are utilized as input to the segmentation block for predicting the video masks of both known and unknown instances.

3.3 Training and Inference

Training: Our proposed OW-VISFormer framework is trained with the loss formulation given by

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_r + \alpha \mathcal{L}_f + \mathcal{L}_{contr}, \tag{4}$$

where $\mathcal{L}_c$, $\mathcal{L}_f$, $\mathcal{L}_r$ and $\mathcal{L}_{contr}$, respectively denote the classification (class-specific branch), foreground objectness (class-agnostic branch), regression (box and mask) and constrastive (STO module) loss terms. While $\mathcal{L}_c$ and $\mathcal{L}_f$ are computed using the focal loss (Lin et al., 2017), $\mathcal{L}_r$ is the standard l_1 loss. Furthermore, a balanced set of exemplars is utilized to finetune the model after the incremental step in each task for alleviating catastrophic forgetting, as in Gupta et al. (2022); Joseph et al. (2021).

Inference: At test time, the class-specific branch predictions of the C known classes are utilized and top-k instances are selected as known instances. Furthermore, among the remaining $q - k$ instances, top-k with high unknown class probability are selected as unknown instances. Finally, the box and instance features from the decoder for the corresponding known and unknown instances predicted, along with the enriched multi-scale encoder features are input to the segmentation block for predicting the video masks.

4 Experiments

4.1 Experiment 1: OW-VIS Setting

Datasets: We adapt the popular Youtube-VIS (Xu et al., 2021) dataset to construct open-world VIS (OW-VIS) data splits. we For each split, we group the 40 categories into two mutually exclusive sets, thereby constructing tasks with non-overlapping classes $\{\mathcal{T}_1, \mathcal{T}_2\}$ such that, the $\mathcal{T}_2$ categories are not known during task $\mathcal{T}_1$. Then, during the learning of task $\mathcal{T}_t$, all the category labels belonging in $\{\mathcal{T}_\alpha : \alpha \leq t\}$ are considered as *known*. Similarly, labels belonging to $\{\mathcal{T}_\alpha : \alpha > t\}$ are considered as *unknown* during evaluation. To construct a test set at time t, we consider 20% of videos in tasks $\{\mathcal{T}_\alpha : \alpha \leq t\}$ for *known* evaluation and 20% of videos in tasks $\{\mathcal{T}_\alpha : \alpha > t\}$ for *unknown* evaluation. By grouping super-categories of all 40 Youtube-VIS categories in different ways, we created 5 such splits as shown in Fig. 3. As Youtube-VIS validation annotations are not available, we split training set further into train and test set for ll the splits explained. Additional details are provided in the supplementary.

Evaluation Metric: We adapt the standard VIS evaluation metrics for evaluating the OW-VIS setting. For the known classes as well as the unknown class, the standard overall average precision (AP) and average recall (AR) are used.

Implementation Details: We employ the self-supervised DINO (Caron et al., 2021) ResNet50 Imagenet-1k pre-trained backbone to extract multi-scale features using the $conv_3$, $conv_4$ and $conv_5$ stages. The resulting multi-scale features are mapped to the same feature dimension of 256 through convolution, as in Wu et al. (2022); Zhu et al. (2021). The two convolution layers in ScratchNet employ kernel size 4×4 and stride 4 along with the output channels set to 256. The encoder and decoder are six layers each with latent dimension being 256. The number of instance queries q is set to 300. Our OW-VISFormer is implemented in PyTorch−1.8 (Paszke et al., 2019) and learned using the Adam optimizer with learning rate set to 10^{-4}. The model is trained on Task-1 data for 18 epochs. In our training strategy, we prioritize the seamless transition from Task-1 to Task-2 by employing an incremental learning approach enhanced

Table 1 Comparison between the baseline and our OW-VISFormer on the five splits (A-E) introduced for OW-VIS setting

		Task-1				Task-2					
		Known		Unknown		Previously Known		Current Known		Both	
		AP	AR-1	AP	AR-1	AP	AR-1	AP	AR-1	AP	AR-1
Split A	Baseline	35.3	34.8	6.7	9.5	30.6	31.2	30.3	30.7	30.4	31.0
	Ours	36.7	37.3	10.0	11.9	32.5	33.9	34.1	34.9	33.3	34.4
Split B	Baseline	31.0	31.4	2.7	5.1	28.9	30.5	31.7	32.6	30.3	31.6
	Ours	32.2	33.1	6.5	8.9	30.4	32.2	35.1	35.7	32.7	33.9
Split C	Baseline	33.9	33.3	4	7.2	28.7	32.6	32.1	32.6	30.4	28.7
	Ours	36.4	35.2	7.1	9.6	30.6	33.2	35.0	35.1	32.8	34.1
Split D	Baseline	31.4	34.5	3.3	6.4	29.7	30.9	30.2	32.2	30.0	31.6
	Ours	33.6	35.0	6.9	9.7	31.7	32.2	33.5	34.2	32.6	33.2
Split E	Baseline	32.0	35.1	3.5	6.5	29.7	31.2	30.4	31.7	30.1	29.7
	Ours	35.1	36.3	5.6	8.9	31.3	32.1	33.9	36.0	32.6	34.0

For the Task-1 evaluation, the results are reported in terms of overall AP and recall (AR-1) for both 'Known' classes and the 'Unknown' category. For the Task-2, which involves the incremental learning step, we report the OW-VIS setting results for 'Previously Known', 'Current Known' along with 'Both'. Note that the 'Unknown' class performance is not reported for Task-2 since all 40 classes are 'Known'. Our proposed OW-VISFormer achieves consistent gains over the baseline on all splits across all tasks for both 'Known' and 'Unknown' classes

with memory replay techniques. Following the initial training on Task-1, we proceed to incrementally train our model on Task-2, devoting 12 epochs to this phase using newly introduced Task-2 data samples. To mitigate the risk of catastrophic forgetting, we incorporate a memory replay training step, where a subset comprising 20% of previously encountered Task-1 data samples is randomly selected and utilized for an additional 2 epochs. This careful methodology ensures that our model retains knowledge from Task-1 while effectively adapting to the demands of Task-2, thus bolstering its overall performance and robustness.

Quantitative Comparison: We compare the performance of the baseline (2.2) and our OW-VISFormer (3) on the OW-VIS splits (see Fig. 3). Table 1 shows the comparison on all five splits and the corresponding two tasks in each split. We report the performance in terms of AP and recall for both known and unknown. For the Task-1 in split A, the baseline achieves a known class AP of 35.3% and AR-1 score of 34.8%. Our OW-VISFormer achieves consistent improvement in performance in terms of both AP and AR-1 by achieving 36.7% and 37.3%, respectively. Notably, OW-VISFormer obtains a considerable gain in performance in the case of unknown class, in terms of both AP and AR-1, owing to the proposed feature enrichment mechanism and the STO module. When the unknown class labels are progressively labeled in Task-2, we observe OW-VISFormer to better maintain both objectives of (i) spatio-temporally segmenting the new known categories and (ii) not forgetting the previously known classes. Moreover, we observe a consistent improvement in performance from OW-VISFormer over the baseline for other splits. We further analyzed the computational complexity of OW-VISFormer in comparison with the baseline where baseline shows the 292 GFLOPS whereas OW-VISFormer has 296 GFLOPS. Similarly, we have computed FPS for bothe models on A100 GPU where baseline shows 10.1 FPS whereas our model shows 9.7 FPS. This minor trade-off is outweighed by the significant gains in segmentation accuracy delivered by OW-VISFormer.

Ablation Study: We also conduct an experiment by progressively integrating our contributions into the baseline on one of the difficult splits (split B) in Table 3. For the known classes, the baseline achieves a AP score of 31.0%. The performance on the known classes is improved by the introduction of our proposed feature enrichment mechanism (Sect. 3.1) with an AP score of 31.5%. Furthermore, the integration of our STO module (Sect. 3.2) that generates instance-level pseudo-labels improves the AP score to 32.2%, leading to an absolute overall gain of 1.2% over the baseline. For the unknown class, the baseline achieves AP and AR-1 scores of 2.7% and 5.1%. Integrating the proposed feature enrichment mechanism significantly improves the performance to 4.5% and 6.1%, in terms of AP and AR-1. The introduction of the STO module leads to a consistent improvement in both AP and AR-1, achieving absolute final gains of 3.8% and 3.6% over the baseline. We further perform an experiment to validate the impact of the dedicated shallow features generated from the proposed ScratchNet. To this end, we use the shallow features $conv_2$ from the standard pre-trained backbone instead in our OW-VISFormer. The reduces the performance in terms of both known and unknown AP from 32.2% and 6.5% to 31.2% and 3.8%. This inferior performance likely suggests that the shallow features from standard pre-trained backbone are more suited for image classification task. In contrast, the shallow features obtained through our ScratchNet are trained

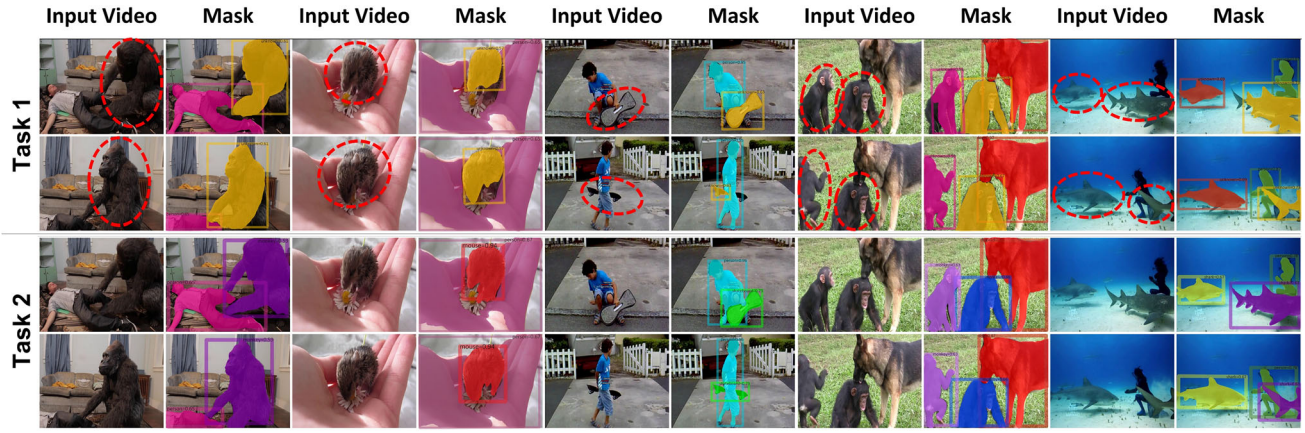

Fig. 4 Qualitative OW-VIS results on example video frames from the the test sets of different splits. Here, for each example video, we show the segmentation masks obtained from our OW-VISFormer when trained only on Task-1 categories (row 1 and 2). The instance mask predictions for the same video frames are shown after *incrementally* learning with Task-2 categories (row 3 and 4). From left to right: the unknown objects are encircled by red dashed lines in the input video frames during the Task-1 evaluation. The unknown objects are accurately segmented first as 'unknown' in Task-1 and then later correctly classified and spatio-temporally segmented into their respective 'known' classes during Task-2 evaluation. Best viewed zoomed in. Additional results are in the supplementary

Table 2 State-of-the-art comparison on YouTube-VIS 2019 val set

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN	30.3	51.1	32.6	31.0	35.5
SipMask-VISCao et al. (2020)	32.5	53.0	33.3	33.5	38.9
VisTRWang et al. (2021)	36.2	59.8	36.9	37.2	42.4
CrossVISYang et al. (2021)	36.6	57.3	39.7	36.0	42.0
IFCHwang et al. (2021)	42.8	65.8	46.8	43.8	51.2
DeVISCaelles et al. (2022)	44.4	66.7	48.6	42.4	51.6
SeqFormerWu et al. (2022)	45.1	66.9	50.5	45.6	54.1
Our Approach	**46.7**	**69.1**	**51.7**	**46.1**	**54.9**

All results are reported using the same ResNet-50 backbone. Our approach outperforms the recent SeqFormer (Wu et al., 2022) with an absolute gain of 1.6% in terms of overall AP. Best results are in bold

from scratch on OW-VIS data and therefore dedicated to aid in accurate video mask prediction for both unknown as well as known classes.

Qualitative Analysis: Fig. 4 shows the qualitative results from our OW-VISFormer on example test video frames of different OW-VIS splits. The first two rows shows the segmentation results obtained on the corresponding video frames in the Task-1 evaluation ('known' and 'unknown' objects). The last two rows show the results from Task-2 evaluation which involves incrementally training with Task-2 categories. In the Task-1, OW-VISFormer is able to first accurately segment the different 'known' and 'unknown' instances. Then, in the Task-2 evaluation, OW-VISFormer successfully identifies the correct categories for the same unknown instances that were introduced in the Task-2 learning while still accurately predicting the instance masks for the previously known categories from Task-1.

Table 3 Performance comparison of baseline and progressively integrated Ow-VISFormer components on split B

Method	Known AP	Unknown AP	Unknown AR_1
Baseline	31.0	2.7	5.1
+ ScratchNet	31.5	4.5	6.1
+ STO Module	32.2	6.5	8.9

The table presents the Average Precision (AP) scores for known and unknown classes, along with the Average Recall@1 score for unknown classes. The improvements achieved by integrating ScratchNet, STO module are demonstrated relative to the baseline

Results with additional baseline: Table 5 shows the effectiveness of our proposed method OW-VISFormer when integrated with different baselines. We consider recently introduced VITA (Heo et al., 2022) and DVIS (Zhang et al., 2023) and adopted it to open-world instance segmentation task by adding DINO (Caron et al., 2021) backbone

Table 4 State-of-the-art comparison for the open-world object detection (OWOD) problem on MS COCO split of Gupta et al. (2022)

Task IDs	Task-1		Task-2				Task-3				Task-4		
	U-Recall	mAP	U-Recall	mAP			U-Recall	mAP			mAP		
		Current Known		Previously Known	Current Known	Both		Previously Known	Current Known	Both	Previously Known	Current Known	Both
ORE-EBUI Joseph et al. (2021)	1.5	61.4	3.9	56.5	26.1	40.6	3.6	38.7	23.7	33.7	33.6	26.3	31.8
OW-DETR Gupta et al. (2022)	5.7	71.5	6.2	62.8	27.5	43.8	6.9	45.2	24.9	38.5	38.2	28.1	33.1
Our Approach	**8.8**	**72.1**	**8.4**	**63.3**	**28.1**	**44.5**	**9.1**	**45.4**	**25.5**	**38.9**	**38.7**	**29.3**	**34.0**

The comparison is presented in terms of unknown category recall (U-Recall) and the known class mAP for the Task-1. For the remaining tasks (2-4) involving incremental learning steps, we report the mAP scores for 'Previously Known', 'Current Known' along with 'Both'. Furthermore, the U-Recall is only reported for tasks 1-3 since all classes are known in the final task 4. When using the same backbone, our approach outperforms the recent OW-DETR on all tasks for both the 'Known' classes and the 'Unknown' class. Best results are in bold

and pseudo-labelling scheme as described in Sect. 2.2. OW-VISFormer enhances these baselines by integrating Scratch-Net and the STO module. Notably, our approach consistently outperforms the baselines across all tasks, particularly demonstrating significant improvements in the AP and AR metrics for unknown classes on both splits. This underscores the effectiveness of OW-VISFormer in addressing the challenges of open-world video instance segmentation.

4.2 Experiment 2: Fully-Supervised VIS Setting

We further evaluate the effectiveness of our proposed contributions (feature enrichment mechanism and STO module) in the standard full-supervised (FS) VIS problem setting. Our main intuition is that the proposed contributions can likely aid in reducing the confusion of an object region (both known and unknown) being called as a background. Hence, this can serve as an additional learning step for better object identification. To this end, we integrate the feature enrichment mechanism (Sect. 3.1) and our STO module (Sect. 3.2) into the recent SeqFormer (Wu et al., 2022). Table 2 shows the results in the standard FS setting on the Youtube-VIS 2019 val. set. Our approach obtained by integrating the feature enrichment and the STO module within the standard SeqFormer obtains overall AP of 46.7%, leading to an absolute gain of 1.6% in AP over recent best (Wu et al., 2022).

4.3 Experiment 3: Open-World Detection Setting

Lastly, we also validate the generalizability of our feature enrichment and the STO module for open-world object detection (OWOD) in images. The OWOD problem has recently gained popularity with evaluations being performed on the challenging MS COCO dataset (Lin et al., 2014). To this end, we integrate our feature enrichment and STO module (replacing 3D convolutions with 2D) into the recent transformers-based OWOD framework, named OW-DETR (Gupta et al., 2022). Table 4 shows the results on the challenging MS COCO split introduced in Gupta et al. (2022). For a fair comparison, we employ the same self-supervised ResNet50 backbone as in the standard OW-DETR (Gupta et al., 2022). We report the results of ORE-EBUI (Joseph et al., 2021) and our baseline OW-DETR from Gupta et al. (2022). In the case of Task-1, our approach achieves an impressive performance particularly on the 'Unknown' object class by increasing the U-Recall from 5.7% to 8.8%, while also improving the detection performance on the 'Current Known' classes. Further, the performance on both 'Unknown' and 'Known' are consistently improved in the subsequent tasks as well. For the final task (task-4) involving all 80 classes from MS COCO as 'Known', our approach improves the performance over the recent OW-DETR in the case of 'Both' previously known and unknown classes from 33.1% to 34.0% mAP.

5 Relation to Prior Art

Video Instance Segmentation: Existing video instance segmentation (VIS) methods can be categorized based on the underlined detection architecture such as, two-stage (Yang et al., 2019; Lin et al., 2020; Bertasius & Torresani, 2020), single-stage (Ke et al., 2021; Cao et al., 2020; Athar et al., 2020; Fu et al., 2021) and transformer-based approaches (Wang et al., 2021; Wu et al., 2022; Yang et al., 2021; Hwang et al., 2021; Caelles et al., 2022; Thawakar et al., 2022). Most two-stage VIS approaches extend a two-stage detector such as, Mask R-CNN He et al. (2017) by integrating a tracking branch. Most single-stage VIS methods adapt the one-stage pipeline, where the final mask prediction is obtained through a linear combination of mask bases. Recently, several works explored transformers-based detection architecture (Zhu et al., 2021; Carion et al., 2020) to formulate VIS as a direct end-to-end sequence prediction. These approaches predominantly solve the problem following a closed-world assumption, where the annotated instances of all (known) semantic classes to be spatio-temporally segmented are available during instance. This assumption posses issues to most existing VIS approaches when classi-

Table 5 Comparison between the baselines and our OW-VISFormer on the two splits (A-B) introduced for OW-VIS setting

		Task-1				Task-2					
		Known		Unknown		Previously Known		Current Known		Both	
		AP	AR-1	AP	AR-1	AP	AR-1	AP	AR-1	AP	AR-1
Split A	`Baseline(VITA` Heo et al. (2022))	38.7	37.5	7.8	11.3	32.8	33.3	33.1	32.4	33.1	33.5
	`Ours(VITA`6 Heo et al. (2022))	39.2	39.1	12.2	14.6	34.9	35.3	35.9	36.7	35.3	36.2
Split B	`Baseline(VITA` Heo et al. (2022))	33.1	33.5	4.3	7.2	31.2	32.7	33.9	34.2	32.6	33.1
	`Ours(VITA` Heo et al. (2022))	35.1	35.3	8.3	10.2	32.7	34.2	37.4	37.8	34.6	36.2
Split A	`Baseline(DVIS` Zhang et al. (2023))	39.2	38.6	8.4	12.7	33.6	34.5	34.2	33.8	34.6	34.4
	`Ours(DVIS` Zhang et al. (2023))	40.3	30.8	13.6	15.9	36.1	36.5	37.4	37.9	6.8	37.5
Split B	`Baseline(DVIS` Zhang et al. (2023))	34.6	34.8	5.9	8.5	32.4	33.9	35.2	35.8	33.9	34.7
	`Ours(DVIS` Zhang et al. (2023))	36.3	36.7	9.5	11.3	34.1	35.7	38.8	39.3	35.1	37.7

Here, the baselines (VITA Heo et al. (2022) and DVIS Zhang et al. (2023)) were adopted according to the our proposed open-world setting as discussed in Sect. 2.2. Our proposed OW-VISFormer built on top of two respective baselines (VITA Heo et al. (2022) and DVIS Zhang et al. (2023)) achieves consistent gains on splits (A-B) across all tasks for both 'Known' and 'Unknown' classes

fying a novel (unknown) object class instance. Furthermore, recent progress in video instance segmentation (VIS) has demonstrated notable enhancements in both model precision and computational efficiency. Mask2former (Cheng et al., 2022) adeptly captures temporal coherence and spatial intricacies through the utilization of a mask transformer. SG-Net (Liu et al., 2021) introduces a pioneering one-stage Spatial Granularity Network, dynamically adjusting instance mask resolutions to optimize both speed and accuracy. Wu et al. (2022) underscore the benefits of online models for real-time VIS applications. Additionally, Heo et al. (2023) present a versatile approach aimed at addressing a wide array of video segmentation tasks. Other noteworthy contributions include DVIS (Zhang et al., 2023) and Tube-link (Li et al., 2023), each contributing unique insights into video instance segmentation methodologies. Furthermore, recent works (Li et al., 2023; Naseer et al., 2021; Ranasinghe et al., 2022; Awais et al., 2023; Dudhane et al., 2023; Thawakar et al., 2023; Dudhane et al., 2024) offer valuable perspectives on the evolving landscape of transformer-based techniques in visual segmentation tasks.

Video Object Segmentation: In the field of video object detection, recent research has focused on integrating spatial-temporal transformers to enhance detection performance. Zhou et al. (2022) introduces a comprehensive framework TransVOD for end-to-end video object detection, while PTSEFormer (Han et al., 2022) proposes a progressive approach to improve temporal and spatial modeling in video object detection. Similarly, Geng et al. (2022) introduced the RSTT for real-time video super-resolution, utilizing a transformer architecture to process both spatial and temporal dimensions efficiently. These developments underscore the integration of advanced spatial-temporal processing in VIS, pointing towards sophisticated solutions for handling complex video scenarios in real-time.

Open Vocabulary Detection and Segmentation: Several recent works have explored the open-world problem setting, where a model learns to distinguish an unknown object instance as 'unknown' while also identifying the given set of 'known' semantic object classes. Such an open-world setting has been investigated in detection (Joseph et al., 2021; Gupta et al., 2022), image instance segmentation (Kuniaki et al., 2022; Wang et al., 2022), and object tracking and segmentation (Liu et al., 2021; Wang et al., 2021). The emergence of open vocabulary segmentation and detection techniques presents new opportunities for handling diverse and evolving object classes (Rasheed et al., 2023). Gu et al. (2021) explores knowledge distillation techniques to enable open-vocabulary object detection. Additionally, Wu et al. (2024) provides a comprehensive overview of methodologies and challenges in open-vocabulary learning. Notably, Wu et al. (2024) and OpenVIS (Guo et al., 2023) present pioneering efforts in extending the capabilities of video instance segmentation models to accommodate an open vocabulary of object classes.

6 Conclusions

We proposed an approach, named OW-VISFormer, to address open-world VIS (OW-VIS). OW-VISFormer introduces a feature enrichment mechanism to produce enriched features and a spatio-temporal objectness module that generates instance-level pseudo-labels. Based on a light-weight auxiliary network that generates shallow features, which are combined with pre-trained high-level features. The resulting features are then input to encoder to obtain enriched multi-scale features. We further propose a spatio-temporal objectness (STO) module that produces instance-level pseudo-labels and is trained using a contrastive loss. Moreover, we pro-

pose OW-VIS splits to identify unknown, segment known and unknown along with progressively segmenting new semantic classes. OW-VISFormer achieves competitive performance in three settings: OW-VIS, FS-VIS and OWOD.

Acknowledgements The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725, the LUMI supercomputer hosted by CSC (Finland) and the LUMI consortium, and by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., & Leibe, B. (2020). Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*.

Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., & Khan, F.S. (2023). Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:2307.13721.

Bertasius, G., & Torresani, L. (2020). Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*.

Caelles, A., Meinhardt, T., Brasó, G., & Leal-Taixé, L. (2022). DeVIS: Making deformable transformers work for video instance segmentation. arXiv:2207.11103.

Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., & Shao, L. (2020). Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *ICCV*.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1290–1299.

Dudhane, A., Thawakar, O., Zamir, S.W., Khan, S., Khan, F.S., & Yang, M.-H. (2024). Dynamic pre-training: Towards efficient and scalable all-in-one image restoration. arXiv preprint arXiv:2404.02154.

Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., & Yang, M.-H. (2023). Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, pp. 5703–5712. IEEE.

Fu, Y., Yang, L., Liu, D., Huang, T.S., & Shi, H. (2021). Compfeat: Comprehensive feature aggregation for video instance segmentation. *AAAI*.

Geng, Z., Liang, L., Ding, T., & Zharkov, I. (2022). Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *CVPR*, pp. 17441–17451.

Gu, X., Lin, T.Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921

Guo, P., Huang, T., He, P., Liu, X., Xiao, T., Chen, Z., & Zhang, W. (2023). Openvis: Open-vocabulary video instance segmentation. arXiv preprint arXiv:2305.16835.

Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., & Shah, M. (2022). Ow-detr: Open-world detection transformer. In *CVPR*.

Han, W., Jun, T., Xiaodong, L., Shanyan, G., Rong, X., & Li, S. (2022). Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. *ECCV*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R.B. (2017). Mask r-cnn. In *ICCV*.

Heo, M., Hwang, S., Hyun, J., Kim, H., Oh, S.W., Lee, J.-Y., & Kim, S.J. (2023). A generalized framework for video instance segmentation. In *CVPR*, pp. 14623–14632.

Heo, M., Hwang, S., Oh, S. W., Lee, J. Y., & Kim, S. J. (2022). Vita: Video instance segmentation via object token association. *NeurIPS, 35*, 23109–23120.

Hwang, S., Heo, M., Oh, S. W., & Kim, S. J. (2021). Video instance segmentation using inter-frame communication transformers. *NeurIPS, 34*, 13352–13363.

Joseph, K., Khan, S., Khan, F.S., & Balasubramanian, V.N. (2021). Towards open world object detection. In *CVPR*.

Ke, L., Li, X., Danelljan, M., Tai, Y. W., Tang, C.K., & Yu, F. (2021). Prototypical cross-attention networks for multiple object tracking and segmentation. In *NeurIPS*.

Kuhn., H.W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly2*(1-2), 83–97.

Kuniaki, S., Ping, H., Trevor, D., & Saenko, K. (2022). Learning to detect every thing in an open world. ECCV.

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z., & Loy, C.C. (2023). Transformer-based visual segmentation: A survey. arXiv preprint arXiv:2304.09854

Li, X., Yuan, H., Zhang, W., Cheng, G., Pang, J., & Loy, C.C. (2023). Tube-link: A flexible cross tube baseline for universal video segmentation. arXiv preprint arXiv:2303.12782.

Lin, T., Goyal, P., Girshick, R.B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *ICCV*.

Lin, C., Hung, Y., Feris, R., & He, L. (2020). Video instance segmentation tracking with a modified vae architecture. In *CVPR*.

Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In: ECCV.

Liu, D., Cui, Y., Tan, W., & Chen, Y. (2021). Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*.

Liu, Y., Zulfikar, I.E., Luiten, J., Dave, A., Ramanan, D., Leibe, B., Ošep, A., & Leal-Taixé, L. (2021). Opening up open-world tracking. In *CVPR*.

Naseer, M., Ranasinghe, K., Khan, S., Khan, F.S., & Porikli, F. (2021). On improving adversarial transferability of vision transformers. arXiv preprint arXiv:2106.04169.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (2019). (eds.) *NeurIPS*, pp. 8024–8035. Curran Associates, Inc.,. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ranasinghe, K., Naseer, M., Khan, S., Khan, F.S., & Ryoo, M.S. (2022). Self-supervised video transformer. In *CVPR*, pp. 2874–2884.

Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., & Khan, F.S. (2023). Fine-tuned clip models are efficient video learners. In *CVPR*, pp. 6545–6554.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., & Li, F. (2015). Imagenet large scale visual recognition challenge. In *IJCV*.

Thawakar, O., Anwer, R.M., Laaksonen, J., Reiner, O., Shah, M., & Khan, F.S. (2023). 3d mitochondria instance segmentation with spatio-temporal transformers. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention, pp. 613–623. Springer.

Thawakar, O., Narayan, S., Cao, J., Cholakkal, H., Anwer, R.M., Khan, M.H., Khan, S., Felsberg, M., & Khan, F.S. (2022). Video instance segmentation via multi-scale spatio-temporal split attention transformer. In *ECCV*, pp. 666–681. Springer

Wang, W., Feiszli, M., Wang, H., & Tran, D. (2021). Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, pp. 10776–10785.

Wang, W., Feiszli, M., Wang, H., Malik, J., & Tran, D. (2022). Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *CVPR*, pp. 4422–4432.

Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., & Xia, H. (2021). End-to-end video instance segmentation with transformers. CVPR.

Wu, J., Jiang, Y., Zhang, W., Bai, X., & Bai, S. (2022). Seqformer: a frustratingly simple model for video instance segmentation. ECCV.

Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., & Jiang, X., *et al.* (2024). Towards open vocabulary learning: A survey. *TPAMI*.

Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., & Bai, X. (2022). In defense of online models for video instance segmentation. In *ECCV*, pp. 588–605. Springer.

Xu, N., Yang, L., Yang, J., Yue, D., Fan, Y., Liang, Y., & Huang, T.S. (2021). YouTube-VIS Dataset 2021 Version. https://youtube-vos.org/dataset/vis.

Yang, L., Fan, Y., & Xu, N. (2019). Video instance segmentation. In *ICCV*.

Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., & Liu, W. (2021). Crossover learning for fast online video instance segmentation. In *ICCV*.

Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., & Wan, P. (2023). Dvis: Decoupled video instance segmentation framework. arXiv preprint arXiv:2306.03413.

Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., & Wan, P. (2023). Dvis: Decoupled video instance segmentation framework. In: ICCV, pp. 1282–1291.

Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., & Tao, D. (2022). Transvod: end-to-end video object detection with spatial-temporal transformers. *TPAMI*.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.