Offline-to-Online Knowledge Distillation for Video Instance Segmentation

Hojin Kim[1]*, Seunghun Lee[2], Hyeon Kang[2] and Sunghoon Im[2]†
[1]SI Analytics, Daejeon, Korea
[2]Department of Electrical Engineering & Computer Science, DGIST, Daegu, Korea
[1]hojin.kim@si-analytics.ai
[2]{lsh5688, hyeonkang, sunghoonim}@dgist.ac.kr

Abstract

In this paper, we present offline-to-online knowledge distillation (OOKD) for video instance segmentation (VIS), which transfers a wealth of video knowledge from an offline model to an online model for consistent prediction. Unlike previous methods that have adopted either an online or offline model, our single online model takes advantage of both models by distilling offline knowledge. To transfer knowledge correctly, we propose query filtering and association (QFA), which filters irrelevant queries to exact instances. Our KD with QFA increases the robustness of feature matching by encoding object-centric features from a single frame supplemented by long-range global information. We also propose a simple data augmentation scheme for knowledge distillation in the VIS task that fairly transfers the knowledge of all classes into the online model. Extensive experiments show that our method significantly improves the performance in video instance segmentation, especially for challenging datasets, including long, dynamic sequences. Our method also achieves state-of-the-art performance on YTVIS-21, YTVIS-22, and OVIS datasets, with mAP scores of 46.1%, 43.6%, and 31.1%, respectively.

1. Introduction

Video instance segmentation (VIS) is the task of detecting, segmenting, and tracking object instances simultaneously in a given video [40]. It can be categorized into two groups: online and offline approaches. Offline methods [1,2,5,12,16,18,22,34,36] input a whole video clip and segment the instances of the entire video sequence in a single step. These models encode global video knowledge by leveraging detected objects in a video sequence. This per-clip pipeline generally shows superior performance over per-frame online methods by associating richer information across the entire video sequence.

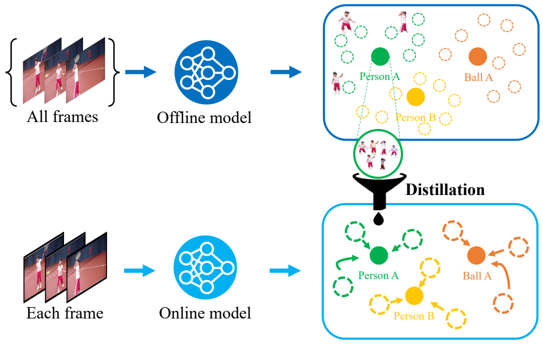


Figure 1. **Basic concept of our method.** An offline model aggregates the instance-specific features extracted from all frames. We distill the instance feature knowledge encoding global video information into the online model for better instance feature matching.

Despite the robustness of offline methods, very recent research trends are leaning toward online approaches [4, 6, 9, 15, 17, 21, 40, 41], which segment objects per frame and keep track of instances, using every single video frame as input. This is useful because many real-time applications (*e.g.*, autonomous driving and surveillance systems) require on-the-fly instance segmentation. However, they still suffer from inaccurate predictions due to poor matching performance. This problem occurs because the online method produces inconsistent features for the same instance. Moreover, the object-centric feature extraction method makes the matching algorithm to be confused when multiple instances with the same class appear in a frame. This is the fundamental limitation of online VIS methods and is easily observed in the YouTubeVIS2022 (YTVIS-22) long video benchmark [39].

To tackle these issues while maintaining online applicability, we present Offline-to-Online Knowledge Distillation (OOKD), as shown in Fig. 1. The basic idea behind OOKD is to train an online model using per-instance features from an offline model as proxy features. The existing online models [15, 37] rely entirely on pair-based loss, which measures the pairwise distance between data in the

*Works processed at DGIST
†Corresponding Author

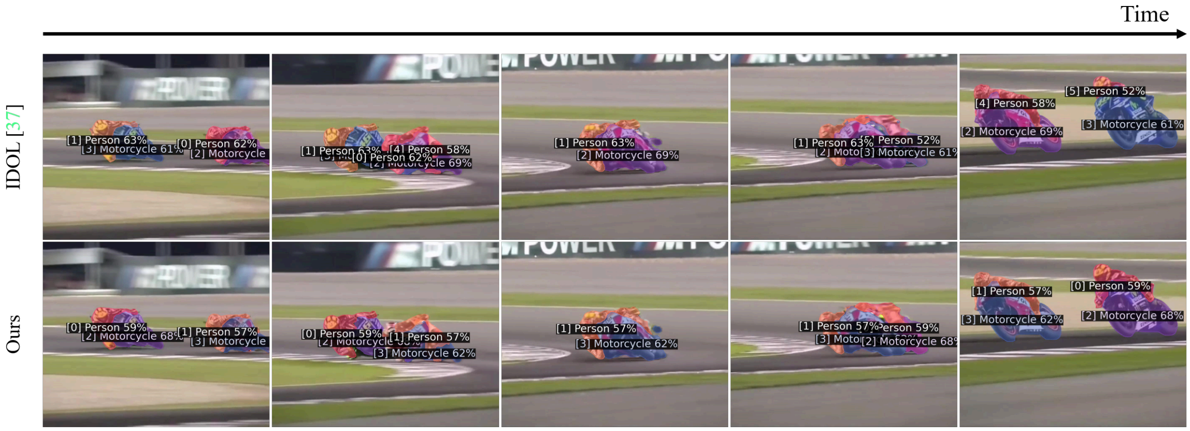


Figure 2. **Qualitative results for the OVIS dataset.** We compare ours (bottom) with the state-of-the-art online method, IDOL [37] (up).

embedding space. OOKD encourages our online model to extract instance-specific features embedding global feature knowledge from a single frame. This allows the online model to learn consistent features, even for dynamic, deformable, and occluded objects, and achieve robust instance matching as shown in Fig. 2.

We also propose Query Filtering and Association (QFA) to build high-purity instance features encoded by whole video sequences. Incorrect predictions of offline models can construct offline knowledge that is less instance discriminative, which leads to performance degradation after the distillation. The QFA module filters out bad queries and associates instances of the same instance when building offline knowledge from the entire video sequence. This QFA module is also utilized to link and precisely align the instance features from offline and online models.

Lastly, we find that the VIS task suffers from a class imbalance problem, which makes a model weak at predicting labels from minorities. For example, the YTVIS-19 training set contains 1654 'human' class instances out of the total 3774 instances. Therefore, we propose a simple yet effective data augmentation for VIS, called Minor-Paste (Minor class copy-Paste). We adopt the copy-paste scheme [7, 18], but selectively sample and paste the instance masks of minor classes. This module is designed to fairly transfer the knowledge of all classes from the teacher network to the student network.

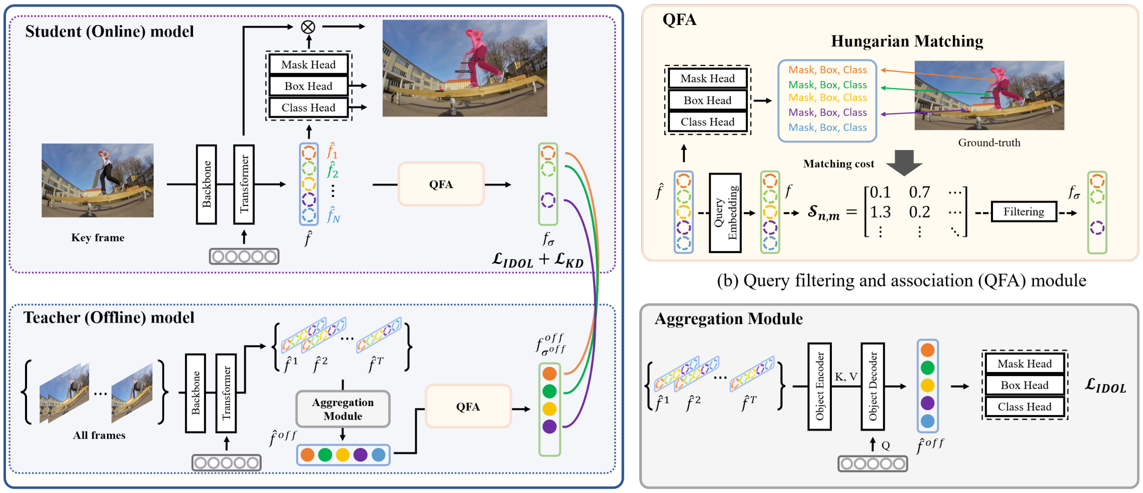
Extensive experiments show that our method outperforms state-of-the-art methods on all benchmark datasets including YTVIS21 [38], YTVIS22 [39], and OVIS [25]. We also observe that our method noticeably improves performance, especially on long video datasets, which validates the effectiveness of the proposed knowledge distillation method and augmentation schemes for VIS. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce an Offline-to-Online Knowledge Distillation (OOKD) method for VIS.
- We design Query Filtering and Association (QFA) for better offline feature extraction and distillation.
- We introduce a data augmentation scheme for VIS (Minor-Paste) that allows our student network to learn fairly for all class representations.
- The proposed method effectively handles the limitation of online VIS on long videos and outperforms both state-of-the-art online and offline VIS models.

2. Related Work

Offline Video Instance Segmentation Offline methods [1, 2, 5, 16, 18, 22, 34, 36] input all scenes at once and predict the instance segmentation labels for all sequence. VisTR [34] is an early work applying the Transformer [31] to offline VIS. IFC [16] presents inter-frame communication Transformers to share inter-frame knowledge with other frames and reduce memory usage. Some works [16, 42] introduce 'near-online' methods that divide the whole video into several clips and use each clip to predict labels. VITA [12] uses frame-level object tokens and associates the collection of the features for global video understanding.

Online Video Instance Segmentation Online methods [4, 6, 9, 15, 17, 21, 40, 41] segment instance labels; every single video frame is given as input. An early online VIS model, MaskTrack R-CNN [40], follows the Mask R-CNN [10] framework with a modified tracking head to match instances between frames. It becomes a generalized framework, and subsequent works follow the pipeline. Multi-Object Tracking and Segmentation (MOTS) [24, 32] is a similar task to online VIS; it predicts segmentation and tracking of all objects except class labels. MinVIS [15] trains queries to be discriminative between intra-frame object instances and uses them for instance tracking. IDOL [37] introduces



(a) Offline-to-online knowledge distillation pipeline

(b) Query filtering and association (QFA) module

(c) Object token aggregation module

Figure 3. **Overview of proposed method.** (a) Our pipeline consists of a student and teacher model with the same architecture of backbone network, Transformer (DeformableDETR [44]) decoder, and query embedding. The student model extracts the features $\{f_n^t\}_{n=\{1,\ldots N\}}$ for each instance n from a single frame v^t, which is used as the input of task heads. The feature f_n^t also passes the query filtering and association (QFA) module to remove the queries associated with the wrong prediction, as illustrated in (b). The teacher model produces the offline feature f_n^{off} for each instance by aggregating the features $\{\hat{f}_n^t\}_{n=\{1,\ldots,N\}}^{t=\{1,\ldots,T\}}$ for all instances from all the frames, as illustrated in (c). Then, we transfer the offline knowledge in the embedded space into online models by imposing a cosine similarity loss $\mathcal{L}_{KD}$ between each instance feature, as well as a task loss $\mathcal{L}_{IDOL}$. (b) The QFA module conducts Hungarian Matching [20] between the prediction of each query and the ground-truth label. The queries with low matching costs are filtered out. (c) The aggregation module unifies the queries of identical objects in the whole video through the object encoder and decoder, which is trained using $\mathcal{L}_{IDOL}$.

a memory-based association strategy, which applies contrastive learning to obtain more discriminative instance embeddings. We adopt the IDOL model as our baseline.

Asymmetric Knowledge Distillation Knowledge distillation in neural networks has been widely studied [8, 13]. This technique mainly targets model compression, distilling knowledge from a bigger teacher model to a smaller student model performing the same task. However, recent studies [14, 26, 35, 43] present new concepts of knowledge distillation; this involves the transfer of asymmetrical knowledge learned from a teacher model to a student model. SVT [26] introduces a self-distillation method that transfers features from global views to features for local views. One study on action detection [43] proposes to transfer the knowledge from the offline action detection model to an online model. LiDAR Distillation [35] distills rich knowledge from the LiDAR data with higher beams to lower beams. This serves as a data compression leading to higher performance with sparse data. The monocular 3D object detection method [14] is trained by distilling feature-based knowledge from 3D LiDAR points.

Inspired by the above papers, We aim to expand the asymmetrical concept to online VIS by utilizing offline video knowledge. To achieve this, we studied an association between the offline model and the online model, targeting the online model to learn the advantages of the offline model.

Data Augmentation It is generally known that data augmentation techniques are widely applied in computer vision tasks and are driving performance improvement [28]. The augmentation scheme is applied not only to the classifications [28] but also to the detection [33] and segmentation [7]. Recently, Tubeformer [18] extends the copy-paste method [7] to clip-paste for video-level recognition. We employ this technique, but more frequently copy and paste the instance mask of the minorities to effectively distill teacher knowledge on all instance representations into the student network.

3. Method

First, we briefly describe the overview of IDOL [37], which is the baseline for our online model in Sec. 3.1. Then, we introduce the offline knowledge extraction method in Sec. 3.2 and the knowledge distillation method in Sec. 3.3. Lastly, we describe the Minor-Paste data augmentation scheme in Sec. 3.4.

3.1. Online VIS model

In this paper, we adopt IDOL [37] for a baseline of our online VIS model, which consists of an image encoder, a Transformer decoder, and prediction heads. Given an input frame $v^t \in \mathbb{R}^{H \times W \times 3}$ of a video $V = \{v^1, ..., v^T\}$, either CNN or Transformer backbone extracts feature maps. The extracted features and N learnable object queries pass through DeformableDETR [44] decoder to transform the queries into instance features $\{\hat{f}_n^t\}_{n=\{1,...,N\}}$ with C hidden dimension ($\hat{f}_n^t \in \mathbb{R}^C$). Lastly, dynamic mask head [30] decodes instance features into segmentation mask, bounding box, and a class of instances. The model is optimized with a classification loss $\mathcal{L}_{cls}$, a bounding box loss $\mathcal{L}_{box}$, a segmentation mask loss $\mathcal{L}_{mask}$, and a contrastive loss $\mathcal{L}_{embed}$ as follows:

$$\mathcal{L}_{\text{IDOL}} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{mask} + \lambda_3 \mathcal{L}_{embed}, \tag{1}$$

where $\lambda_{\{1,2,3\}}$ are the balancing terms among the losses. Each query $\hat{f}_n^t$ from a frame v^t is the input of the task heads and applied to contrastive embedding to obtain instance embeddings f_n^t. During inference, instance embeddings from previous frames are summed with embeddings in memory banks with specific weights. Embeddings in memory banks are utilized to match feature similarities between current and memory instances and to track instance IDs.

3.2. Offline Knowledge Extration

We aim to distill instance-distinctive feature knowledge from an offline model into our online model. We extract the representative features for each instance from the entire video frame and use them as offline knowledge. For effective distillation, we design the offline model for learning more representative features on the feature space shared with the online model. Thus, we use the pre-trained online model whose structure is the same as our target online model to extract frame-level instance-centric knowledge. We associate the collection of knowledge across an entire video sequence as illustrated in Fig. 3.

We first pass every single frame v^t into the baseline online VIS model, defined in Sec. 3.1, to extract the per-frame instance query $\{\hat{f}_n^t\}_{n=\{1,...,N\}}$. Then, we aggregate every instance query $\hat{F}_n = \{\hat{f}_n^t\}_{n=\{1,...,N\}}^{t=\{1,...,T\}}$ from the whole set of video frames using object token association [12] to obtain offline knowledge $\{\hat{f}_n^{off}\}_{n=\{1,...,N\}}$. Each instance feature $\hat{f}_n^{off} \in \mathbb{R}^C$ embeds video-level information for each instance. The aggregation module consists of an object encoder and an object decoder as shown in Fig. 3-(c). The encoder builds intercommunication of queries along the temporal axis, employing self-attention modules. The augmented instance features and N learnable object queries are passed through the object decoder to embed the offline instance information into the queries. The feature aggregation module is trained by passing the offline instance query $\hat{f}_n^{off}$ through the dynamic mask head and minimizing the loss to instance masks, bounding boxes, and classes in (1). We only train the object encoder and decoder with the frozen baseline online model. We obtain offline instance embedding f_n^{off} by passing the learned query $\hat{f}_n^{off}$ through contrastive embedding similar to the method in the previous section.

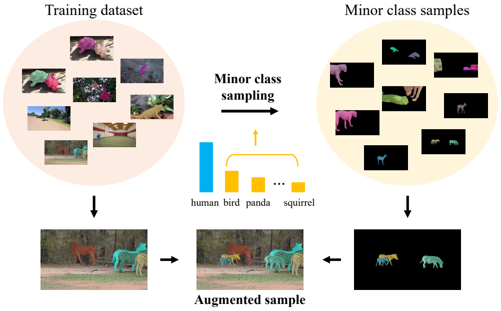


Figure 4. Illustration of our augmentation scheme, Minor-paste.

3.3. Offline-to-Online Knowledge Distillation

Distilling incorrect or irrelevant knowledge into student models can degrade the model's performance. This situation is commonly caused by the transfer of knowledge from false predictions in the teacher model and in pairs of mismatched instances between the teacher and student models. To address the issue, we propose Query Filtering and Association (QFA) in Fig. 3-(b), which maps predicted instance embeddings to ground-truth instances one-to-one while removing wrong predictions. Suppose we have M ground-truth bounding boxes $B \in \mathbb{R}^{M \times 4}$ and classes $C \in \mathbb{R}^{M \times N_c}$ with N_c class labels in a single frame appearing in a sampled training video. We define a matching cost matrix $\mathcal{S} \in \mathbb{R}^{N \times M}$ by measuring the localization errors of the bounding box predictions $\hat{B} \in \mathbb{R}^{N \times 4}$ and the confidence errors of the class prediction $\hat{C} \in \mathbb{R}^{N \times N_c}$ as follows:

$$\mathcal{S}_{n,m} = \mathcal{L}_c(\hat{C}_n, C_m) + \lambda_b \mathcal{L}_b(\hat{B}_n, B_m), \tag{2}$$

where $\mathcal{L}_c$ is cross entropy and $\mathcal{L}_b$ is generalized IoU [27], by following [29]. The indices n and m are for the prediction and ground truth instances. We find the optimal index σ_m for ground truth instance m, which has the lowest cost among all N predictions, as follows:

$$\sigma_m =_{n \in \{1,...,N\}} \mathcal{S}_{n,m}. \tag{3}$$

We conduct the same process for an offline model to find the optimal index σ_m^{off} for offline predictions as well. These optimal indices σ_m and σ_m^{off} are utilized to match the instances between online and offline models. Given pairs of the matched features f_{σ_m} and $f_{\sigma_m^{off}}^{off}$, we compute distillation loss, which maximizes the cosine similarity between them, as follows:

$$\mathcal{L}_{KD} = \frac{1}{M} \sum_{m=1}^{M} \left(1 - \frac{f_{\sigma_m^{off}}^{off} \cdot f_{\sigma_m}}{\| f_{\sigma_m^{off}}^{off} \| \| f_{\sigma_m} \|} \right). \tag{4}$$

Backbone	Type	Method	mAP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
MsgShifT	Offline	TeViT [42]	37.9	61.2	42.1	35.1	44.6
ResNet-50	Offline	VisTR [34]	31.8	51.7	34.5	29.7	36.9
		IFC [16]	36.6	57.9	39.3	-	-
		SeqFormer [36]	40.5	62.5	43.6	36.2	48.0
		Mask2Former [5]	40.6	60.9	41.8	-	-
		VITA [12]	45.7	67.4	49.5	**40.9**	53.6
	Online	M-RCNN [40]	28.6	48.9	29.6	26.5	33.8
		STMask [21]	30.6	49.4	32.0	26.4	36.0
		SipMask [4]	31.7	52.5	34.0	30.8	37.8
		Cross-VIS [41]	34.2	54.4	37.9	30.4	38.2
		VISOLO [9]	36.9	54.7	40.2	30.6	40.9
		InstanceFormer [19]	40.8	62.4	43.7	36.1	48.1
		DeVIS [3]	43.1	66.8	46.6	38.0	50.1
		MinVIS [15]	44.2	66.0	48.1	39.2	51.7
		IDOL [37]	43.9	68.0	**49.6**	38.0	50.9
		OOKD (Ours)	**46.1**	**69.6**	49.2	40.8	**55.5**
Swin-L	Offline	VITA [12]	57.5	80.6	61.0	**47.7**	62.6
	Online	InstanceFormer [19]	51.0	73.7	56.9	42.8	56.0
		DeVIS [3]	54.4	77.7	59.8	43.8	57.8
		MinVIS [15]	55.3	76.6	62.0	45.9	60.8
		IDOL [37]	56.1	80.8	63.5	45.0	60.1
		OOKD (Ours)	**59.2**	**82.6**	**65.0**	47.2	**64.3**

Table 1. Quantative comparison of our method to state-of-the-art methods on the YTVIS-21 dataset. Best scores are highlighted with **bold**.

We additionally impose the distillation loss on the loss with a balancing term λ_4 defined in (1), as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{IDOL}} + \lambda_4 \mathcal{L}_{KD}. \tag{5}$$

3.4. Minor-paste

We also propose a simple yet effective augmentation scheme, called Minor-paste (minor-class copy-paste) in Fig. 4, for knowledge distillation in video instance segmentation. We aim to transfer the knowledge of the teacher model into the student model equally, regardless of class labels. To do so, we compute the sampling probability p_c^s for each instance c as follows:

$$p_c^s = k\frac{\max(p_c) - p_c}{\max(p_c) - \min(p_c)}, \text{ where } c \in \{1, ..., N_c\}, \tag{6}$$

where p_c is the proportion of the number of classes c to the total number of classes in the entire training data set. We set the scale parameter k at 0.7, which controls the probability of data augmentation. We sample video clips containing at least one minor class whose p_c is less than 10%. Then, we randomly crop the instance regions and paste them onto other target video clips based on the sampling probability p_c^s. We observe that the YTVIS-21 training datasets contains about 35.5% of the 'Human' class and 0.3% of the 'Squirrel' class. Based on these probabilities, for example, the most major class 'Human' has never been augmented and the most minor class 'Squirrel' has a 70% chance to be augmented. We use GT instance segmentation masks and class IDs to determine the area to crop and copy the instance.

4. Experimental Results

4.1. Experimental Setup

Datasets: We examine our method on YTVIS-21 [38], YTVIS-22 [39], and OVIS [25]. These datasets are more recent, and challenging datasets than YTVIS-19 [40].

YTVIS-19 is the first video instance segmentation dataset originated from Video Object Segmentation (VOS) datasets. It includes 2,238 training, 302 validation, and 343 test data of high-resolution YouTube video clips. They have 40 different object categories, and the video frame interval is 5. YTVIS-19 has a small number of instances and a small amount of class variety for a single video (an average of 1.3 classes and 1.7 instances per video for the training set).

YTVIS-21 is an upgraded version of YTVIS-19, which has additional 747 training data and 119 validation data. There are also 40 different object categories, but with some minor changes to the object class. It includes a total of 2,985 training, 421 validation, and 453 test videos with an average of 1.5 classes and 3.4 instances per video for the training set.

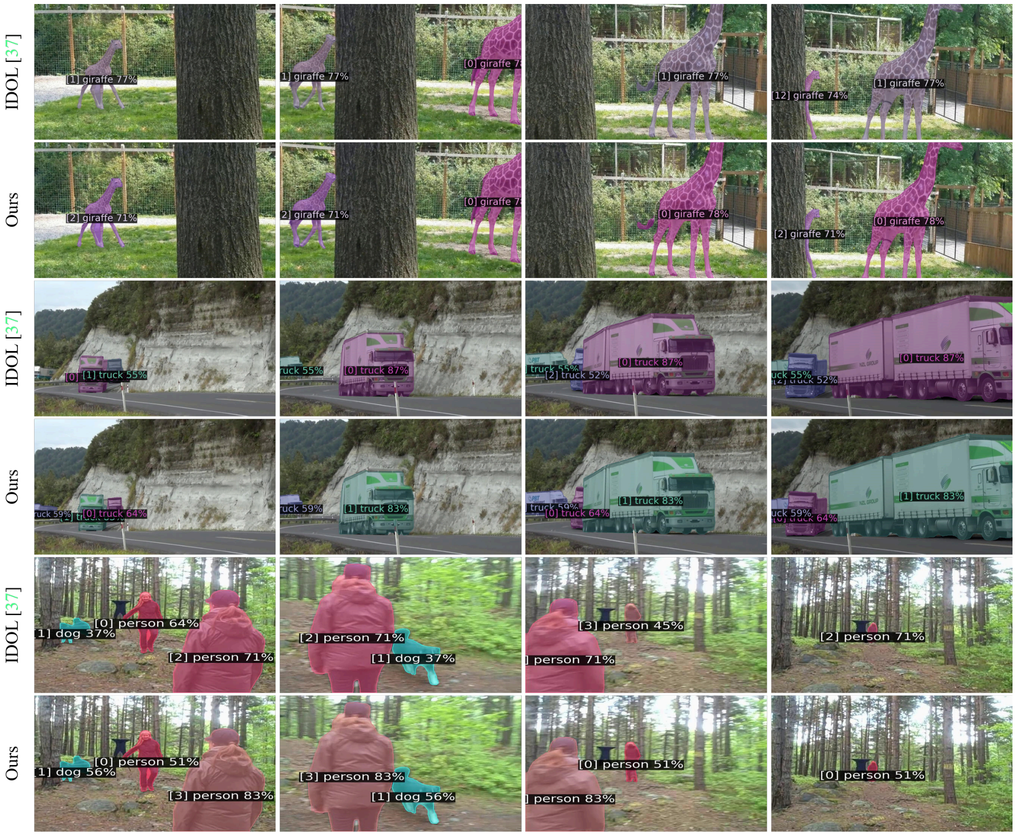


Figure 5. Qualitative comparison of our method to IDOL [37] for the YTVIS-22 dataset.

YTVIS-22 has the same training set as YTVIS-21, but 71 videos are additionally included in the YTVIS-21 validation set. These additional videos, named 'long videos', have longer frame intervals of 20, from longer video sequences For clarification, the existing validation datasets are named 'short video'; these have a frame interval of 5. We conduct the experiments by training the models [5, 12, 15, 37] provided by the authors and reported in Tab. 3. For a fair comparison, we train all the competitive methods and ours with the same training environments.

OVIS data is a very challenging dataset that contains long video sequences with a large number of objects and more frequent occlusion. It consists of 607 training, 140 validation, and 154 test videos. This dataset has a large number of instances despite its average class diversity (average 1.4 class and 5.9 instances per video for the training set). This property makes VIS models more difficult to distinguish from each instance because each instance has a similar appearance.

Evaluation Metric: We use standard metrics for VIS, the average precision (AP), and average recall (AR) with the video intersection over Union (IoU) of the mask sequences as the threshold.

Baselines: We use ResNet-50 [11] and Swin-L [23] backbones. ResNet-50 is the most standard and widely used backbone for VIS. Swin-L is a recent backbone that provides the best performance in VIS.

Implementation Details: Unless otherwise noted, we follow the hyper-parameter setting of IDOL [37] for our online model and VITA [12] for offline knowledge aggregation. We set loss balencing term as $\lambda_{\{1,2,3,4\}} = 2, 2, 1, 1$. We sample 4 frames to train the offline model in Sec. 3.2. We train the ResNet-50-based model in eight RTX3090 GPUs and the Swin-L-based model in eight A6000 GPUs. Our method using the Resnet-50 and Swin-L backbones runs at 30.6 fps and 17.6 fps for per-frame inference of the YTVIS-21 dataset, respectively.

4.2. Comparison to State-of-the-art Methods

YTVIS-21: We conduct the performance comparison of our model to the recent competitive methods for the YTVIS-21 dataset in Tab. 1. Our method achieves the highest mAP for both ResNet-50 and Swin-L backbone by reaching mAP performances of 46.1% and 59.2%, respectively. Compared to the state-of-the-art online model IDOL [37], the proposed method shows approximately 2% and 3% performance improvement on both backbones respectively. Interestingly, OOKD outperforms the state-of-

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
IFC [16]	13.1	27.8	11.6	9.4	23.9
SeqFormer [36]	15.1	31.9	13.8	10.4	27.1
VITA [12]	19.6	41.2	17.4	11.7	26.0
M-RCNN [40]	10.8	25.3	8.5	7.9	14.9
SipMask [4]	10.2	24.7	7.8	7.9	15.8
CrossVIS [41]	14.9	32.7	12.1	10.3	19.8
InstanceFormer [19]	20	40.7	18.1	12.0	27.1
DeVIS [3]	23.7	47.6	20.8	12.0	28.9
MinVIS [15]	25.0	45.5	24.0	13.9	29.7
IDOL [37]	30.2	51.3	30.0	**15.0**	37.5
OOKD (Ours)	**31.1**	**52.8**	**32.7**	**15.0**	**39.6**

Table 2. Quantative comparison of our method to state-of-the-art methods on the OVIS validation set. Offline models [12, 16, 36] are sorted by placing them on the top of the table. All results are conducted with ResNet-50 backbone. The best results are highlighted with **bold**.

type	Method	AP	mAP_S	mAP_L
Offline	Mask2former [5]	36.3	40.2	32.3
	VITA [12]	38.8	45.7	31.9
Online	MinVIS [15]	34.5	43.5	25.6
	IDOL [37]	39.3	44.7	33.9
	OOKD (Ours)	**43.6**	**46.1**	**41.2**

Table 3. Quantative comparison of our method to state-of-the-art methods on the YTVIS-22 dataset. All experiments are conducted with ResNet-50 backbone.

the-art offline model, VITA [12].

YTVIS-22: We compare our method to the most recent competitive methods [5,12,15,37] with the YTVIS-22 dataset. We report quantitative results with mAP for short video (mAP_S) and long video (mAP_L) in Tab. 3. Because the results for long videos are not reported in the papers, we use the source codes provided by the authors to measure the average precision in this experiment. We achieve the best performance among all competitive methods in both short and long video datasets. We observe that the performance improvement for long videos is significant as ours outperforms the state-of-the-art offline (VITA) method with 9.3% and the online (IDOL) method with 7.3%. These results show that object-centric features associated with global video information extracted by OOKD enable robust feature matching between images with long time intervals. Accurate feature matching is the key to increasing mAP scores.

We also evaluate the methods qualitatively in Fig. 5. All the competitive methods produce high-quality segmentation results while predicting wrong instance labels. This problem is significant, especially in the existing online-based methods. It is because the inconsistent instance features per frame are extracted, and it makes the matching difficult. On the other hand, the proposed method correctly predicts the instance IDs although it is processed in an online manner. The matching performance is improved by the proposed distillation method, which is the key to accurate video instance segmentation.

OVIS: The quantitative comparisons for the OVIS dataset are shown in Tab. 2. The results also demonstrate that the proposed method outperforms all the competitive methods even with challenging datasets, OVIS. It is widely known that offline VIS methods are struggling for video instance segmentation with long and dynamic videos [12,15]. As is known, the experiments show that the recent online methods [3, 15, 37] are generally better than the recent offline methods [12, 16, 36]. Despite the limitations of offline methods, our online method distilled by offline knowledge outperforms the state-of-the-art online method. This demonstrates the effectiveness of offline knowledge distillation. We believe that the instance features aggregated by global video information act as the proxies of each instance and the proxies guide all the features for the same instance to be consistent. This guideline helps the feature matching to be more robust even for the instances with appearance changes, and this is analyzed in detail in Sec. 4.3.

4.3. Ablation Study

Knowledge Distillation with/without QFA: To demonstrate the effectiveness of the query filtering and association (QFA), we conduct the ablation study in Tab. 4. The results in the first and last rows are the results of the baseline model and our method, respectively. The results in the second row are from a baseline model with the same knowledge distillation method without QFA-based query matching. None of the results in Tab. 4 adopts the data augmentation. The results show that KD without QFA degrades the performance of the online model by about 0.4% mAP_S and 2.8% mAP_L. It is because the online model is trained by distilling mismatched offline knowledge into online features. The order of instance IDs is not always consistent, and the instance labels should be matched to distill knowledge correctly. Our model with KD and QFA improves the performance of the pure baseline model by approximately 1.0% mAP_S and 4.5% mAP_L thanks to the QFA finding corresponding features. This shows that knowledge distillation is effective only with the proposed QFA. One interesting observation here is that the performance gains on long video datasets are significant. This demonstrates that the offline-to-online knowledge distillation helps the online model to extract consistent features, even with the long video frames containing large appearance changes.

Synergy of Minor-Paste and KD We conduct the ablation study on Minor-Paste and KD in Tab. 5. Minor-Paste improves the performance of the baseline model by 1.2% and 0.9% for the short video and the long video, respec-

KD	QFA	mAP$_S$	mAP$_L$
X	-	44.6	33.8
O	X	44.2	31.0
O	O	45.6	38.3

Table 4. Ablation study on knowledge distillation (KD) and KD with query filtering and association (QFA). We measure mAP$_S$ for short videos and mAP$_L$ for long videos in the YTVIS-22 dataset.

Minor-Paste	KD+QFA	mAP$_S$	mAP$_L$
X	X	44.6	33.8
O	X	45.8	34.7
X	O	45.6	38.3
O	O	46.1	41.2

Table 5. Ablation study on Minor-paste and our KD (KD+QFA).

tively. Moreover, our method with both augmentation and KD+QFA significantly improves the baseline model, especially for the long video datasets from 33.8% to 41.2%. Our KD+QFA brings 4.5% out of a total of 7.4% improvements and can be regarded as the core component that derives the success of our model. Another interesting point here is the proposed augmentation scheme is more effective on our model than on the baseline model for the long video sequences (ours: 2.9% and baseline: 0.9% improvements). This shows that the richer knowledge of the minor class helps the knowledge distillation. We also compare our Minor-Paste to the conventional copy-paste method [18] in Tab. 7. As the proposed method shows better performance than the conventional method, mitigating the data imbalance problem helps better knowledge distillation.

Token Aggregation Module: We report the quantitative comparison to the token average in Tab. 6 to demonstrate the effectiveness of the token aggregation module. The token average is a method that averages instance feature tokens representing the same instance as a representative token of instance for the entire video. The results show that both methods improve the performance, but the token aggregation module improves more than the averaging method.

Instance Feature Similarity: We perform further analysis of feature similarity to figure out why the proposed KD induces a performance improvement. We report a histogram of the similarity between two features of the same instances that appeared at different frames in Fig. 6. We randomly sample a pair of instances from all frames in 100 videos on the YTVIS-21 training dataset and measure cosine similarity. The results show that higher feature similarity is obtained after OOKD is applied. The feature extraction with higher similarity among the same instances induces better performance. It is because online VIS models find instance IDs by selecting the instance with the highest feature similarity among all queries. Thus, our model achieves better performance than the conventional method.

Method	mAP$_S$	mAP$_L$
IDOL	44.7	33.9
Token Average	45.5	38.3
Token Aggregation Module	46.1	41.2

Table 6. Comparison on token aggregation methods.

Method	mAP$_S$	mAP$_L$
KD without augmentation	45.6	38.3
KD with Clip-Paste [18]	44.4	40.2
KD with Minor-Paste (Ours)	46.1	41.2

Table 7. Ablation study on augmentation methods.

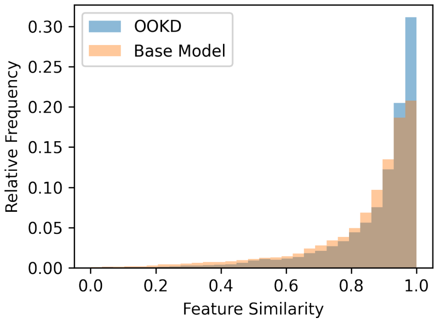


Figure 6. Histogram of feature similarity between two features from the same instance. The similarities of the features from our method and base model, IDOL [37] are indicated by blue and orange bars, respectively.

5. Conclusion

In this paper, we propose OOKD, offline-to-online knowledge distillation for video instance segmentation. Our method transfers the richer instance representation from an offline model into an online model. To teach the student model correctly, we present a query filtering and association (QFA) that filters out irrelevant queries and finds the correct matching pairs between student and teacher queries. This enables a single online model to take both advantages of online and offline models, which boosts the robustness while maintaining the ability for on-the-fly inference. Robustness is further enhanced by our method of minor-paste augmentation that alleviates the class imbalance issues. Extensive experiments have shown that our method improves the performance of the VIS, even in long and dynamic videos. We also achieve state-of-the-art performance on YTVIS-21, YTVIS-22, and OVIS datasets by reaching mAP up to 46.1%, 43.6% and 31.1%, respectively.

Acknowledgement This work was supported by Institute of Information Communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210908).

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 *European Conference on Computer Vision*, pages 158–177. Springer, 2020.

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020.

[3] Adrià Caelles, Tim Meinhardt, Guillem Brasó, and Laura Leal-Taixé. Devis: Making deformable transformers work for video instance segmentation. *arXiv preprint arXiv:2207.11103*, 2022.

[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 *European Conference on Computer Vision*, pages 1–18. Springer, 2020.

[5] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[6] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1361–1369, 2021.

[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021.

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[9] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2896–2905, 2022.

[10] 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.

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

[12] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022.

[13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[14] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. 2022.

[15] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*, 2022.

[16] 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.

[17] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *Advances in Neural Information Processing Systems*, 34:1192–1203, 2021.

[18] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13914–13924, 2022.

[19] Rajat Koner, Tanveer Hannan, Suprosanna Shit, Sahand Sharifzadeh, Matthias Schubert, Thomas Seidl, and Volker Tresp. Instanceformer: An online video instance segmentation framework. *arXiv preprint arXiv:2208.10547*, 2022.

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[21] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11215–11224, 2021.

[22] 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.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[24] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022.

[25] 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.

[26] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised

video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022.

[27] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[28] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[29] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[30] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019.

[33] Hao Wang, Qilong Wang, Fan Yang, Weiqi Zhang, and Wangmeng Zuo. Data augmentation for object detection via progressive and selective instance-switching. *arXiv preprint arXiv:1906.00358*, 2019.

[34] 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.

[35] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. *arXiv preprint arXiv:2203.14956*, 2022.

[36] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. *Computer Vision ECCV*, 2022.

[37] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *arXiv preprint arXiv:2207.10661*, 2022.

[38] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S Huang. Youtube-vis dataset 2021 version, 2021.

[39] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S Huang. Youtube-vis dataset 2022 version, 2022.

[40] 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.

[41] 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.

[42] 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.

[43] Peisen Zhao, Lingxi Xie, Jiajie Wang, Ya Zhang, and Qi Tian. Progressive privileged knowledge distillation for online action detection. *Pattern Recognition*, 129:108741, 2022.

[44] 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.