MDQE: Mining Discriminative Query Embeddings to Segment Occluded Instances on Challenging Videos

Minghan Li Shuai Li Wangmeng Xiang Lei Zhang*
Department of Computing, Hong Kong Polytechnic University
liminghan0330@gmail.com, {csshuaili, cswxiang, cslzhang}@comp.polyu.edu.hk

Abstract

While impressive progress has been achieved, video instance segmentation (VIS) methods with per-clip input often fail on challenging videos with occluded objects and crowded scenes. This is mainly because instance queries in these methods cannot encode well the discriminative embeddings of instances, making the query-based segmenter difficult to distinguish those 'hard' instances. To address these issues, we propose to mine discriminative query embeddings (MDQE) to segment occluded instances on challenging videos. First, we initialize the positional embeddings and content features of object queries by considering their spatial contextual information and the inter-frame object motion. Second, we propose an inter-instance mask repulsion loss to distance each instance from its nearby non-target instances. The proposed MDQE is the first VIS method with per-clip input that achieves state-of-the-art results on challenging videos and competitive performance on simple videos. In specific, MDQE with ResNet50 achieves 33.0% and 44.5% mask AP on OVIS and YouTube-VIS 2021, respectively. Code of MDQE can be found at https://github.com/MinghanLi/MDQE_CVPR2023.

1. Introduction

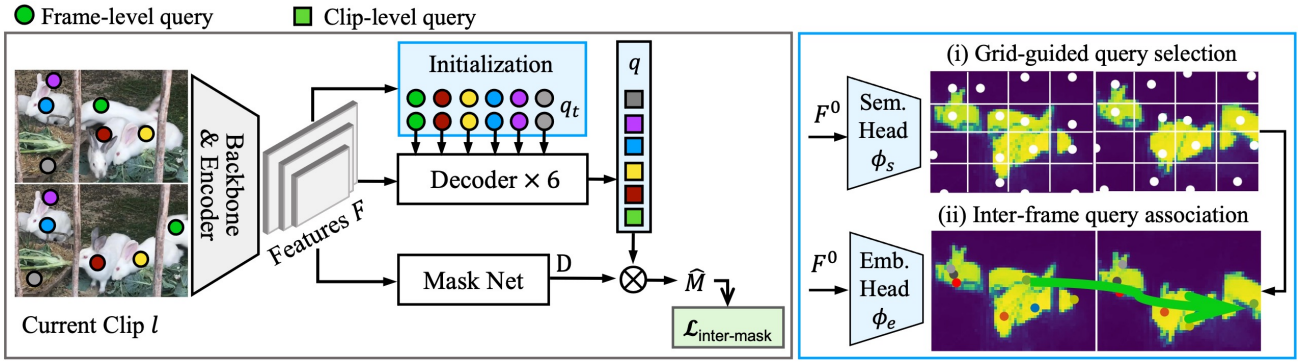
Video instance segmentation (VIS) [51] aims to obtain pixel-level segmentation masks for instances of different classes over the entire video. The current VIS methods can be roughly divided into two paradigms: per-frame input based methods [3, 18, 20, 26, 47, 51, 52] and per-clip input based methods [1, 19, 27, 45, 46, 48, 53]. The former paradigm first partitions the whole video into individual frames to segment objects frame by frame, and then associate the predicted instance masks across frames, while the latter takes per-clip spatio-temporal features as input to predict multi-frame instance masks with the help of embedding learning [1], graph neural networks [41] and transformer networks [17, 19, 45, 46].

The recently proposed per-clip VIS methods [19, 45, 46, 48] have set new records on the YouTube-VIS datasets [51], achieving significant performance improvement over the per-frame VIS methods [3, 5, 12, 14, 26, 52, 54]. SeqFormer [46] and VITA [17] locate an instance in each frame and aggregate temporal information to learn powerful representations of video-level instances via a naive weighted manner and a video-level decoder, respectively. However, on the challenging OVIS dataset [37], which includes occluded or similar-looking instances in crowded scenes, the per-clip VIS methods lag behind the per-frame ones. Actually, the recently developed per-frame method IDOL [47] records state-of-the-art performance on OVIS by introducing contrastive learning [10, 35, 44] to learn inter-frame instance embeddings. We argue that the per-clip VIS methods should be able to exploit richer spatial-temporal features and achieve better performance than their per-frame counterparts. However, there are two main issues that limit the existing per-clip methods to achieve this goal.

First, existing query-based VIS methods adopt zero or random input as the positional embeddings and content features of object queries in decoder layers, which cannot encode spatio-temporal prior of objects, resulting in poor results on challenging videos. Second, during training, the existing mask prediction loss mainly forces each query to match the pixels of its target instance [21, 42] and mismatch the pixels of other instances and the background. No further inter-instance clue has been exploited to teach the segmenter to distinguish mutually occluded instances.

To address the above issues, we propose to mine discriminative query embeddings (MDQE) to better segment hard instances on challenging videos for per-clip VIS methods. First, we propose to improve the initialization of object queries to specify discriminative spatial-temporal priors. We divide the activation map of each frame into several patches via a grid and select the peak point in each patch as the initial positions of frame-level queries, and then associate them across frames by embedding similarity to ensure that frame-level queries in the same grid of the video clip

*Corresponding author.



(a) Overview of MDQE architecture (b) Frame-level query initialization

Figure 1. (a) The proposed MDQE architecture consists of a backbone and encoder that extract multi-scale features F from a video clip, a query initialization module that produces temporally-aligned frame-level queries q_t, a decoder that decodes discriminative clip-level queries q, and a Mask Net that generates mask features D. The mask features D and clip-level queries q are combined via a linear combination to obtain the clip-level instance mask $\hat{M}$, which is supervised by our proposed inter-instance mask repulsion loss in Sec. 3.3. (b) The frame-level query initialization consists of two steps: grid-guided query selection and inter-frame query association, resulting in temporally-aligned frame-level queries. Please refer to Sec. 3.2 for more details.

can correspond to the same object. Second, to teach the query-based segmenter to distinguish occluded instances, we replace the original mask prediction loss with an inter-instance mask repulsion loss, which forces each query to activate the pixels of its target instance and suppress the pixels of its surrounding non-target instances.

The proposed VIS method with per-clip input, namely MDQE, is the first to achieve contrastive learning of instance embeddings via query initialization and the inter-instance mask repulsion supervision, which can effectively segment hard instances on challenging videos. Our experiments on both OVIS and YouTube-VIS datasets validate that MDQE with per-clip input achieve competitive performance with its per-frame competitors.

2. Related work

Our work is related to the many per-frame VIS methods, the recently proposed per-clip VIS methods, as well as the methods for learning query embeddings.

Per-frame input based VIS methods. A popular VIS pipeline [5, 12, 18, 20, 26, 28, 31, 47, 51, 52] is to extend the representative image instance segmentation methods [4,6,8,15,39] by adapting a frame-to-frame instance tracker. For example, in [12, 18, 51], the clues such as category score, box/mask IoU and instance embedding similarity are integrated into the tracker. However, these trackers may struggle in distinguishing instances with similar appearance. Inspired by contrastive learning [7, 10, 23, 35, 44], IDOL [47] learns discriminative instance embeddings for multiple object tracking frame by frame, achieving state-of-the-art results on OVIS [37]. Besides, clip-to-clip trackers [3,28,37] propagates the predicted instance masks from a key frame to other frames using deformable convolution [3,9], non-local block [28], correlation [26,37], graph neural network [41], *etc*. By exploiting the temporal redundancy among overlapped frames, clip-to-clip trackers improve much the performance of per-frame methods.

Per-clip input based VIS methods. A clip-in clip-out VIS pipeline was firstly proposed in [1] to model a video clip as a single 3D spatio-temporal pixel embedding. In recent years, transformer based per-clip methods [19,45,46,48,53] have achieved impressive progress on the YouTube-VIS datasets [51]. VisTR [45] views the VIS task as a direct end-to-end parallel sequence prediction problem, but it needs a large memory to store spatio-temporal features. To solve the issue, IFC [19] transfers inter-frame information via efficient memory tokens, and SeqFormer [46] locates an instance in each frame and aggregates temporal information to predict video-level instances. To keep object temporal consistency, EfficientVIS [48] transfers inter-clip query embeddings via temporal dynamic convolution.

However, per-clip VIS methods do not perform well on the challenging OVIS videos [37] with occluded objects in crowded scenes. Actually, occlusion-aware models have been developed for related tasks [21, 33, 43, 50]. For instance, a bilayer convolutional network is developed in [22] to infer the occluder and occluded instances in image segmentation. A repulsion detection loss is designed in [43] to distance the bounding box of an object from the surrounding non-target objects for detecting individual pedestrian in a crowd. Inspired by these works, we propose an inter-instance mask repulsion loss to distinguish the pixels of each instance from its nearby non-target instances.

Query initialization. Existing query-based VIS methods adopt zero-initialized (*e.g.*, DETR [6]) or randomly-initialized (*e.g.*, Deformable DETR [55]) inputs as initial queries. The initial queries cannot encode well the spatio-

temporal priors of objects, making the query-based segmenter difficult to distinguish occluded instances with similar appearance. Actually, query initialization with contextual and positional information has been used in many computer vision tasks [2, 13, 25, 36] for higher performance or faster convergence. However, it has not been well explored in VIS. In this paper, we thus propose a query initialization method to obtain temporally-aligned frame-level queries.

3. Methodology

We outline the proposed MDQE from the perspective of query-based mask prediction in Sec. 3.1, then introduce the two major parts of MDQE: object query initialization in Sec. 3.2 and the inter-instance mask repulsion loss in Sec. 3.3. Finally, we present the training loss and near-online inference process in Sec. 3.4.

3.1. Framework Overview

An input video is partitioned into a few video clips, each with T frames. As illustrated in the left of Fig. 1, during training, a video clip first passes through the backbone and encoder to obtain its multi-scale features F, where the one at the largest scale is represented as F^0. The feature F^0 is then used as the input to our proposed query initialization module in Sec. 3.2 to produce temporally-aligned frame-level queries $\{q_t\}_{t=1}^{T}$ (denoted by circles), which will be fed into the decoder to obtain discriminative clip-level queries q (denoted by squares). On the other hand, the multi-scale features F are input into the Mask Net to generate mask features D, which are then combined with the clip-level queries q to produce clip-level instance masks $\hat{M}$ using linear combination [4, 39]. Finally, the predicted masks $\hat{M}$ are supervised by our proposed inter-instance mask repulsion loss in Sec. 3.3 to distance each instance from its nearby non-target instances, implementing contrastive learning of instance masks.

3.2. Frame-level Query Initialization

In this subsection, we initialize frame-level queries with instance-specific information and improve the decoder architecture to mine discriminative instance embeddings.

Existing query-based VIS methods typically adopt zero or random input as the initial positional embeddings and content features of object queries, following DETR [6] and deformable DETR [55]. However, a recent method [36] called *grid-guided query selection* was proposed to endow object queries with positional and contextual information of the input image, as shown in the top of Fig. 1(b). This method inputs the features F^0 into a semantic segmentation head (ϕ_s with three linear layers) to predict the class-aware activation map: $S = \phi_s(F^0) \in R^{c \times T \times H^0 \times W^0}$, where c is the number of categories and H^0, W^0 are the height and

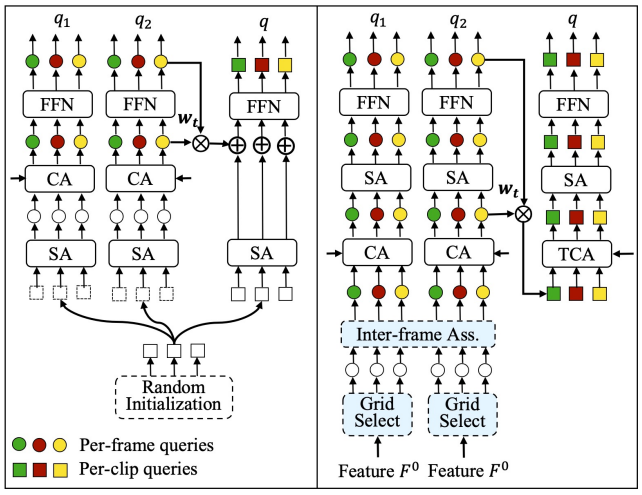


Figure 2. Architecture comparison of the first decoder layer between (a) Seqformer [46] and (b) MDQE, where 'SA', 'CA' and 'TCA' refer to self-attention, cross-attention and temporal cross-attention layer, respectively. Please refer to Sec. 3.2 for details.

width of features. The activation map is evenly divided into several patches through a grid, and the peak point with the highest response (the white dots) in each patch is selected. The coordinates $p \in R^2$ and features $F_p^0 \in R^d$ of the peak point are then assigned as the initial positions and content features of the query, where d is the dimension of features.

However, the grid-guided query selection may not be able to cover an object with the same grid across multiple frames because object motion or camera jitters will cause position shifts over time. To improve temporal consistency, we extend the grid-guided query selection by incorporating the *inter-frame query association*, as illustrated in the bottom of Fig. 1(b). For a video clip, we first perform the above grid-guided query selection frame by frame to obtain the initial frame-level queries, and then input the content features of these queries into an embedding head (ϕ_e with three linear layers) and output their embeddings: $e_p = \phi_e(F_p^0) \in R^{d_e}$, where p represents the query position and d_e is the dimension of embeddings. To obtain temporally-aligned queries, we calculate the embedding similarity between each query in the central frame and neighboring queries within a large window in the non-central frame. The query in the non-central frame with the highest similarity is assigned as the matched query. Note that the size of the window increases if two frames are far apart. After applying the above inter-frame query association, frame-level queries within the same grid are roughly aligned across multiple frames.

In training, we employ the commonly used focal loss [29] to supervise the class-aware activation map, denoted as $\mathcal{L}_{\text{init-sem}}$. Besides, we employ contrastive learning [35, 47] to pull the query embeddings with the same instance ID over

multiple frames closer, and push them away from each other otherwise. For an object query at position p, its contrastive loss of embeddings is defined as:

$$\mathcal{L}_{\text{init-reid}} = -\log \frac{\exp(e_p \cdot e_{p'}^{r+})}{\exp(e_p \cdot e_{p'}^{r+}) + \sum_{r-} \exp(e_p \cdot e_{p'}^{r-})},$$

where e_p represents its query embeddings, $e_{p'}^{r+}$ and $e_{p'}^{r-}$ denote the embeddings vectors of its neighbouring queries at position p' with the same instance ID and with different instance IDs, respectively.

Since our query initialization can provide instance-specific spatial-temporal information for frame-level queries, we further adjust the decoder architecture to take full advantage of it. We compare the architectures of the first decoder layer of SeqFormer [46] and our MDQE in Fig. 2. As shown in Fig. 2(a), SeqFormer employs random input as clip-level queries, calculates cross-attention between clip-level queries and per-frame encoded features to specify frame-level queries in each frame, and finally updates clip-level queries as the weighted combination of frame-level queries. We refer to this architecture of coping embeddings from clip-level instance queries to frame-level object queries as 'I2O'.

Our MDQE utilizes a different decoder architecture, as shown in Fig. 2(b). it first computes cross-attention and then self-attention, and integrates frame-level object queries into clip-level instance queries ('O2I' for short). The clip-level queries $q \in R^{N\times d}$ are calculated similarly to SeqFormer [46]: $q = \sum_{t=1}^{T} w_t \cdot q_t$, where $q_t \in R^{N\times d}$ indicates frame-level queries, and $w_t = \text{FFN}(q_t) \in R^{N\times 1}$ is the time weight of each frame in the clip. Additionally, we add an extra temporal cross-attention (TCA) layer to mine and integrate discriminative features of instances in a larger spatio-temporal receptive field. The attention module in TCA extracts deformable points from multi-frame single-scale feature maps [55]. Our proposed query initialization can early associate instances across frames, providing a good warm-up to the decoder with the 'O2I' architecture. This helps to reduce confusing masks among crowded instances.

3.3. Inter-instance Mask Repulsion Loss

The clip-level instance masks $\hat{M} \in R^{N\times T\times H\times W}$ can be produced by the linear combination of mask features $D \in R^{d\times T\times H\times W}$ and clip-level queries $q \in R^{N\times d}$: $\hat{M} = qD$, as shown in Fig. 1(a). During training, the predicted instance masks $\hat{M}$ are supervised by the ground-truth instance masks M via the binary cross-entropy (BCE) loss and the Dice loss [11, 34]. The typical instance mask prediction loss thus can be formulated as:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{BCE}}(\hat{M},\ M) + \mathcal{L}_{\text{Dice}}(\hat{M},\ M), \qquad (1)$$

where $M \in R^{K\times T\times H\times W}$ is the ground-truth (GT) instance mask, and K is the number of matched GT instances. The

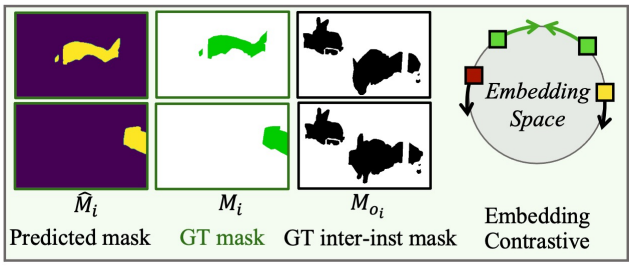


Figure 3. Schematic diagram of the inter-inst mask repulsion loss on the instance with ID i, corresponding to the instance with green circles in Fig. 1.

formula of Dice loss is as follows:

$$\mathcal{L}_{\text{Dice}}(\hat{M},\ M) = \frac{1}{K}\sum_{i=1}^{K} 1 - \frac{2|\hat{M}_i \odot M_i|}{|\hat{M}_i| + |M_i|} \qquad (2)$$

where $\odot$ denotes the point-wise matrix multiplication and $|\cdot|$ sums the values in the matrix.

However, the mask prediction supervision gives priority to match the pixels of its target instance, and then mismatch the pixels of other instances and the background. This may make the query-based segmenter converge to a shortcut of object spatial locations, resulting in imprecise segmentation prediction and confusing masks on occluded instances.

In fact, the relative relationship between each instance and its surrounding instances can provide contrastive features to the query-based segmenter. Inspired by the success of contrastive learning [7, 21, 44], we design an inter-instance mask repulsion loss to suppress the pixels belonging to the nearby non-target instances. For the i-th instance, we define its nearby non-target instances via the intersection of union of bounding boxes (BIoU):

$$o_i = \{j \mid \max_{t\in[1,T]} \text{IoU}(B_{ti}, B_{tj}) > \epsilon, \forall j \in [1, K], j \neq i\}, \qquad (3)$$

where ϵ is a threshold to control the number of nearby non-target samples, which is set to 0.1 by default. Let the union of nearby GT instance masks be the complementary GT inter-instance mask, *i.e.*, $M_{o_i} = \cup_{j\in o_i} M_j$, which contains all pixels of its nearby non-target instances, as illustrated in Fig. 3. Since most annotations used in instance segmentation are not mutually exclusive, we further set $M_{o_i} = M_{o_i} \cap (1 - M_i)$ to remove pixels contained in the GT mask and the GT inter-instance mask at the same time.

Typically, the supervision for predicting instance masks includes the BCE and Dice losses in Eq. (1). In order to train the segmenter to perceive the relative relationships between each instance and its surroundings, we enhance the original BCE and Dice losses by incorporating inter-instance repulsion supervision, named inter-instance BCE loss and inter-instance Dice loss, respectively. Specifically, we adopt a weighted BCE loss to assign a larger weight for the pixels belonging to the target instance and its nearby instances.

The formula of the inter-instance BCE loss is:

$$\mathcal{L}_{\text{BCE-inter}} = \frac{1}{|W_i|} \sum_{p=1}^{N} W_{ip} \, \text{BCE}(\hat{M}_{ip}, M_{ip}), \quad (4)$$

where p and N indicate the pixel position index and the total number of points in the mask, respectively. The corresponding inter-instance weight W_{ip} is set to α (2 by default), if $M_{ip} = 1$ or $M_{o_i p} = 1$, otherwise 1.

On the other hand, we introduce an inter-instance mask repulsion loss, which involves M_{o_i} into the Dice loss to explicitly suppress the pixels of nearby non-target instances. The formula of inter-instance Dice loss is:

$$\mathcal{L}_{\text{Dice-inter}} = 1 - \frac{2|\hat{M}_i \odot M_i| \; + \; |(1 - \hat{M}_i) \odot M_{o_i}|}{|\hat{M}_i| \; + \; |M_i| \; + \; |M_{o_i}|}. \quad (5)$$

If an instance is isolated to other instances, *i.e.*, $|M_{o_i}| = 0$, the inter-instance Dice loss degrades to the original Dice loss in Eq. (2). In terms of gradient back-propagation, the pixels that belong to the target instance and its nearby instances will have a larger gradient value compared to other pixels in the image.

Finally, our inter-instance mask repulsion loss is

$$\mathcal{L}_{\text{inter-mask}} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}_{\text{BCE-inter}}(\hat{M}_i, M_i, M_{o_i}) + \mathcal{L}_{\text{Dice-inter}}(\hat{M}_i, M_i, M_{o_i}). \quad (6)$$

The above loss considers both the matching of pixels belonging to the target instance and the mismatching of pixels belonging to the nearby non-target instances, therefore providing instance discriminative information to the segmenter for producing more accurate instance masks.

3.4. Training and Inference Details

We employ Deformable DETR [55] as the transformer framework, and SeqFormer [46] as the clip-level VIS baseline. The training losses of our proposed MDQE is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{box}} + \lambda_3 \mathcal{L}_{\text{inter-mask}} + \lambda_4 \mathcal{L}_{\text{init-sem}} + \lambda_5 \mathcal{L}_{\text{init-reid}},$$

where we adopt the focal loss for classification, and the smooth L_1 loss and the generalized IoU loss [38] for bounding box regression. During training, we empirically set $\lambda_1 = 2$, $\lambda_2 = 2$, $\lambda_3 = 4$, $\lambda_4 = 2$, and $\lambda_5 = 0.5$.

During inference, MDQE processes the testing video clip by clip in a near-online fashion. Multiple frames (more than T frames) are loaded into the backbone and encoder to obtain encoded features, which are then extracted clip by clip to the decoder to output clip-level queries. Overlapping frames between clips are only used in the decoder, making MDQE fast. In each clip, instances with classification confidence below a threshold are removed, and their masks and embeddings are added to the memory pool to remember objects from previous clips. For a new clip, denoted as the l-th video clip for clarity, MDQE generates instance masks $\hat{M}$ and clip-level embeddings q with high classification confidence, and extracts memory-based instance masks $\hat{M}^m$ and embeddings q^m from the previous T_{mem} clips. The Hungarian matching algorithm is applied on the score matrix S to associate instances across clips:

$$S = \beta_1 \, \text{mIoU}(\hat{M}^m, \hat{M}) + \beta_2 \, \text{sim}(q^m, \, q), \quad (7)$$

where 'mIoU' and 'sim' measure the mask IoU of instance masks and embedding similarity of instance queries between the memory pool and the current clip, respectively. β_1 and β_2 balance the proportions of the two losses, which are set to 1 by default. This process is illustrated in Fig. 4.

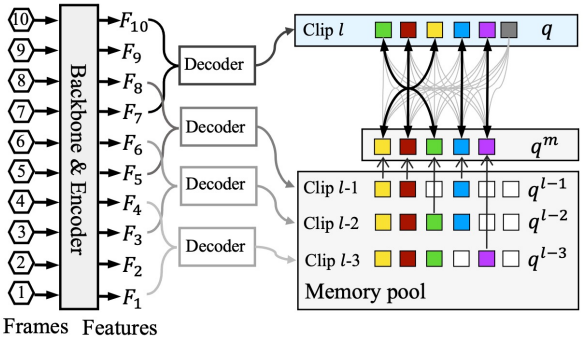


Figure 4. Near-online inference with a clip-by-clip tracker.

4. Experimental Results

4.1. Experiments setup

Datasets. YouTube-VIS [51] 2019/2021 respectively contain 2,283/2,985 training, 302/421 validation, and 343/453 test videos over 40 categories. All the videos are annotated for every 5 frames. The number of frames per video is between 19 and 36. OVIS [37] includes 607 training, 140 validation and 154 test videos, scoping 25 object categories. Different from YouTube-VIS series, OVIS dataset includes longer videos, up to 292 frames, with more objects per frame and varying occlusion levels. The proportions of objects with no occlusion, slight occlusion, and severe occlusion are 18.2%, 55.5%, and 26.3%, respectively. Note that 80.2% of the instances are severely occluded in at least one frame, and only 2% of the instances are not occluded in any frame.

Evaluation metrics. The commonly used metrics, including average precision (AP) at different IoU thresholds, average recall (AR) and the mean value of AP (mAP), are adopted for VIS model evaluation. OVIS divides all instances into three groups called slightly occluded (AP_{so}), moderately occluded (AP_{mo}) and heavily occluded (AP_{ho}),

Init.	Arch.	TCA	mAP	AP_{50}	AP_{75}	AP_{so}	AP_{mo}	AP_{ho}
	I2O		15.4	31.3	14.3	31.8	17.3	3.2
✓	I2O		19.8	40.6	18.2	36.3	22.6	6.5
✓	O2I		24.2	47.5	22.9	40.9	27.3	8.4
✓	O2I	✓	25.6	49.1	24.9	41.9	29.0	11.2

(a) Initialization for frame-level queries.

w	Assoc.	mAP	AP_{50}	AP_{75}	AP_{so}	AP_{mo}	AP_{ho}
0		28.5	53.0	26.9	47.6	32.5	11.9
3	✓	29.7	55.6	27.1	48.9	34.5	12.2
5	✓	30.6	57.2	28.2	49.3	35.1	13.6
7	✓	30.5	57.1	28.6	49.1	33.7	13.7

(b) Inter-frame query association, where w controls the window size.

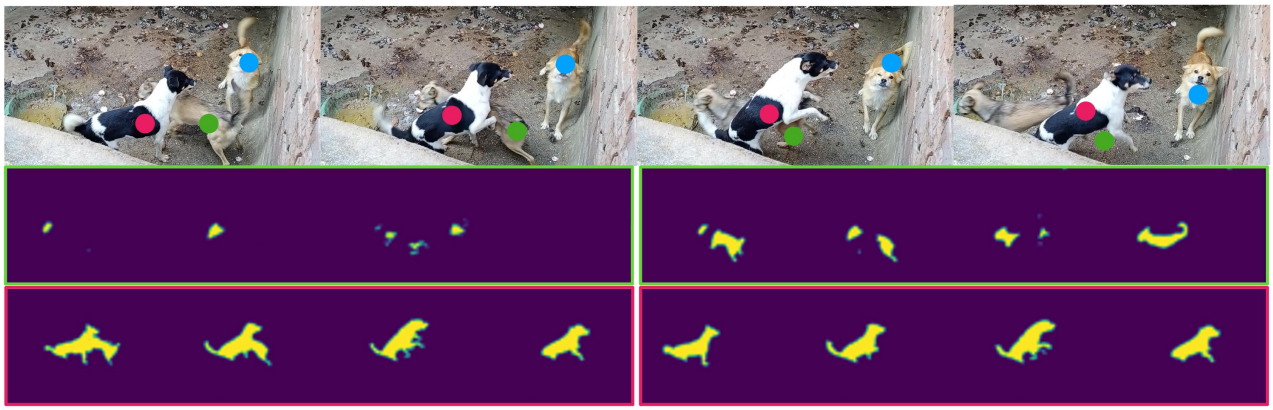
$\mathcal{L}_{\text{BCE-inter}}$	$\mathcal{L}_{\text{Dice-inter}}$	ϵ	mAP	AP_{50}	AP_{75}	AP_{so}	AP_{mo}	AP_{ho}
			29.0	51.6	29.5	44.7	31.3	11.8
2		0.1	30.5	55.6	29.5	46.7	33.1	12.9
2	✓	0.1	31.2	56.8	30.4	48.6	34.5	13.5
2	✓	0.5	30.9	56.4	30.5	47.2	34.2	13.3

(c) Inter-instance mask repulsion loss.

β_1	β_2	T_{mem}	mAP	AP_{50}	AP_{75}	AP_{so}	AP_{mo}	AP_{ho}
1		-	29.1	54.1	27.7	46.5	32.8	12.9
	1	10	28.3	53.4	27.1	47.1	31.3	11.6
1	1	10	30.6	57.2	28.2	49.3	35.1	13.6
1	1	5	30.4	56.4	28.7	49.4	35.2	13.2

(d) Tracking. β_1 and β_2 control the proportions of mIoU and similarity.

Table 1. Ablation studies of MDQE on OVIS valid set, where the input video is 480p, the clip length is $T = 4$, and the number of overlapped frames across clips is $T - 1$, respectively. ϵ controls the number of nearby non-target samples.



(a) Typical mask prediction loss in Eq. (1)

(b) Our inter-instance mask repulsion loss in Eq. (6)

Figure 5. Segmentation on heavily occluded objects by MDQE with typical mask prediction loss and our inter-instance mask repulsion loss. The red, green and blue dots highlight the initial query positions selected by our proposed query initialization method. The last two rows display the predicted masks on the occluded instance and the occluder instance.

where the occlusion scores of instances are in the range of [0, 0.25], [0.25, 0.5] and [0.5, 0.75], respectively, whose proportions are 23%, 44% and 49% accordingly.

Implementation details. Our method is implemented on top of detectron2 [49]. Following [46], we set six layers with multi-scale deformable attention module in encoder and decoder, and employ 200 object queries with dynamic one-to-many label assignment as in [47]. We first pre-train our model on COCO [30], and then finetune it on VIS datasets [37, 51]. In pre-training, we sample an image as the key frame, and resize it in a relative range [0.6, 1] as the reference frame. All models with ResNet50 backbone [16] are trained on eight GeForce RTX 3090 GPUs with 24G RAM, and models with Swin Large backbone [32] are trained on four NVIDIA RTX A6000 with 48G RAM. The initial learning rate is 0.0001, and it decays by 0.1 at 20k and 24k iterations. During training and inference, unless specified, all frames are resized to a shorter edge size of 360 on YouTube-VIS datasets or 480 on OVIS dataset, and the clip length is taken as $T = 4$ on ResNet50 backbone and $T = 3$ on Swin Large backbone, respectively.

4.2. Ablation study

This section performs ablation studies on MDQE and discusses the impact of query embeddings on the segmentation performance. The OVIS valid set is used.

Query initialization. In Table 1a, we explore the performance improvements brought by query initialization. The baseline Seqformer [46] with 'I2O' decoder architecture achieves 15.4% mask AP. By using grid-guided query selection to initialize frame-level queries, the performance is improved to 19.8% mask AP. By employing both query initialization and the 'O2I' decoder architecture, the mask AP and AP_{50} increase to 24.2% and 47.5%, respectively. Additionally, the temporal cross-attention layer (TCA) further improves mask AP by about 1%. The proposed query initialization significantly enhances instance detection and segmentation accuracy in challenging videos.

Inter-frame query association. In Table 1b, we investigate the impact of window size w on inter-frame query association for query initialization. Without inter-frame query

Type	Methods	YouTube-VIS 2021					OVIS					FPS	Params
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}		
Per-frame (360p)	MaskTrack [51]	28.6	48.9	29.6	-	-	10.8	25.3	8.5	7.9	14.9	20.0	58.1M
	STMask [26]	31.1	50.4	33.5	26.9	35.6	15.4	33.9	12.5	8.9	21.4	28.0	-
	CrossVIS [52]	33.3	53.8	37.0	30.1	37.6	14.9	32.7	12.1	10.3	19.8	39.8	37.5M
	InstFormer [24]	40.8	62.4	43.7	36.1	48.1	20.0	40.7	18.1	12.0	27.1	-	44.3M
	IDOL [47]	43.9	**68.0**	**49.6**	38.0	50.9	24.3	45.1	23.3	14.1	33.2	30.6	43.1M
	MinVIS [18]	44.2	66.0	48.1	39.2	51.7	26.3	47.9	25.1	**14.6**	30.0	52.4	44.0M
Per-clip (360p)	VisTR* [45]	31.8	51.7	34.5	29.7	36.9	10.2	25.7	7.7	7.0	17.4	30.0	57.2M
	IFC* [19]	36.6	57.9	39.3	-	-	13.1	27.8	11.6	9.4	23.9	46.5	39.3M
	TeViT [53]	37.9	61.2	42.1	35.1	44.6	17.4	34.9	15.0	11.2	21.8	68.9	161.8M
	SeqFormer* [46]	40.5	62.4	43.7	36.1	48.1	15.1	31.9	13.8	10.4	27.1	72.3	49.3M
	VITA [17]	**45.7**	67.4	49.5	**40.9**	**53.6**	19.6	41.2	17.4	11.7	26.0	33.7	57.2M
	MDQE (our)	44.5	67.1	48.7	37.9	49.8	**29.2**	**55.2**	**27.1**	14.5	**34.2**	37.8	51.4M
720p	IDOL [47]	-	-	-	-	-	30.2	51.3	30.0	15.0	37.5	-	43.1M
	MDQE (ours)	-	-	-	-	-	**33.0**	**57.4**	**32.2**	**15.4**	**38.4**	13.5	51.4M

Table 2. Quantitative performance comparison of VIS methods with ResNet50 backbone on benchmark YouTube-VIS 2021 and OVIS datasets. Note that MinVIS and VITA adopt stronger masked-attention decoder layers proposed in Mask2Former [8]. FPS is computed on YouTube-VIS 2021 valid set, and symbol "-" means the results are not available or applicable. Best in **bold**, second with underline.

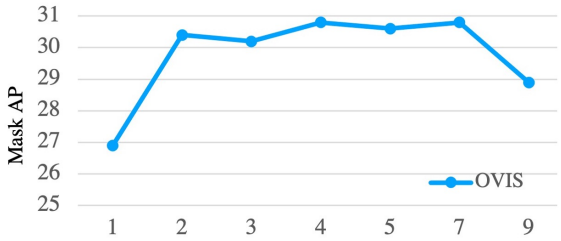


Figure 6. Ablation study on clip length with near-online inference.

Type	Methods	AP	AP_{so}	AP_{mo}	AP_{ho}
Per-frame	IDOL [47]	24.3	34.8	29.0	8.1
	MinVIS [18]	26.3	45.8	31.6	9.9
Per-clip	SeqFormer [46]	15.1	31.8	17.3	3.2
	VITA [17]	19.6	32.3	20.4	8.2
	MDQE (ours)	**29.2**	**46.6**	**33.9**	**13.1**

Table 3. Occlusion metrics of SOTA methods on OVIS valid set.

association ($w = 0$), the performance is only 28.5% mask AP. When the window size is set to 3, 5 and 7, the mask AP increases by 1.2%, 2.1% and 2.0%, receptively. So we set the default size as 5. The top row of Fig. 5 visualizes the initial query positions selected by our proposed query initialization method. We see that it keeps temporal consistency and avoids confusion between occluded objects. Visualization of the initial query positions on more videos can be found in the **supplementary materials**.

Inter-instance mask repulsion loss. In Table 1c, we compare the performance of typical mask prediction loss and our inter-instance mask repulsion loss. MDQE with the typical mask prediction loss achieves only 29.0% mask AP, and performs poorly in all occluded objects. When the weight α of the inter-instance BCE loss is set to 2, the performance increases to 30.5% mask AP, and introducing inter-instance Dice loss further increases mask AP to 31.2%. If we set a higher threshold ϵ in Eq. (3) to only consider heavily occluded objects, the performance drops by 1.4% AP_{so} on slightly occluded objects. Overall, our inter-instance mask repulsion loss brings significant improvements on all occlusion metrics, 3.9% AP_{so}, 3.2% AP_{mo} and 1.7% AP_{ho}, respectively, resulting in better instance segmentation results on challenging videos. The visual comparison of instance masks on a video with heavily occluded objects is displayed in Fig. 5.

Tracking. In Table 1d, we compare the performance of clip-by-clip tracker using different weights in the score matrix in Eq. (7). Using only the mIou term ($\beta_1 = 1$) or the embedding similarity term ($\beta_2 = 1$) respectively achieves 29.1% and 28.3% mask AP, while the mIou term performs better on moderately and heavily occluded objects. By enabling both the two terms, the mask AP increases to 30.6% and all occlusion metrics improve significantly. If the number of frames T_{mem} in memory pool is reduced from 10 to 5 frames, the performance will drop slightly.

Effect of the clip length. In Fig. 6, we explore the VIS results by varying the clip length. MDQE with $T = 1$ achieves 27% mask AP, slightly higher than the state-of-the-art per-frame input method MinVIS [18] in Table 2. The mask AP fluctuates between 30.0% and 31% with the increase of the frame number of input clip, peaks at around 31.0% mask AP with 7-frame per clip, and then falls to 29% mask AP with 9-frame per clip. When $T = 9$, the performance drops because of the complex trajectories of objects in long clips, which can be alleviated by increasing the training samples of long videos.

Data	Methods	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
YT21	IDOL [47]	56.1	80.8	63.5	45.0	60.1
	MinVIS [18]	55.3	76.6	62.0	45.9	60.8
	SeqFromer [46]	51.8	74.6	58.2	42.8	58.1
	VITA [17]	57.5	80.6	61.0	47.7	62.6
	MDQE (ours)	56.2	80.0	61.1	44.9	59.1
OVIS	MinVIS [18]	41.6	65.4	43.4	18.6	44.9
	VITA [17]	27.7	51.9	24.9	14.9	33.0
	MDQE (ours)	41.0	67.9	42.7	18.3	45.2
	IDOL † [47]	42.6	65.7	45.2	17.9	49.6
	MDQE † (ours)	42.6	67.8	44.3	18.3	46.5

Table 4. Quantitative performance comparison of VIS methods with Swin Large (SwinL) backbone [40] on benchmark VIS datasets. Symbol '†' means that the input video is of 720p.

4.3. Main Results

With ResNet50 backbone, we compare in Table 2 the proposed MDQE with state-of-the-art methods on OVIS and YouTube-VIS 2021 datasets. For those methods marked by '*', we employ video-in video-out offline inference on YouTube-VIS valid sets (less than 84 frames), and clip-in clip-out inference with overlapping frames between clips on OVIS (at most 292 frames). The occlusion metrics of SOTA methods on OVIS is provided in Table 3 as well. In addition, we report the performance comparison between recently proposed VIS methods with Swin Large backbone in Table 4. Due to the limit of space, the experiments on ResNet101 backbone and YouTube-VIS 2019 valid set can be found in the **supplementary materials**.

YouTube-VIS [51] 2021 valid set. From Table 2, we can see that early VIS methods [5, 26, 51, 52] using dense anchors only obtain around 30% mask AP, while recently proposed VIS methods [17, 18, 47] using sparse object queries can reach more than 44% mask AP. Since the videos in YouTube-VIS 2021 are short and simple, the performance gap between per-frame input and per-clip input based methods is not significant. Based on the strong decoder layers proposed in Mask2Former, the frame-level method MinVIS [18] and the clip-level method VITA [17] respectively reset new state-of-the-arts with 44.2% and 45.7% mask AP. Without using the masked-attention in Mask2Former, our proposed MDQE achieves 44.5% mask AP, which is only slightly lower than VITA [17].

OVIS [37] valid set. OVIS is much more difficult than YouTube-VIS. Its videos have much longer duration with occluded and similar-looking objects. The early per-frame methods MaskTrack R-CNN [51] and CrossVIS [52] achieve only 10.8% and 14.9% mask AP, respectively. The recent per-frame methods with query-based transformer decoder layers, IDOL [47] and MinVIS [18], bring impressive improvements, achieving 24.3% and 26.3% mask AP, respectively. However, the query-based transformer methods with per-clip input show unexpectedly low performance, such as 15.1% by Seqformer [46]. By introducing object token association between frames, VITA [17] achieves 19.6% mask AP. In comparison, our MDQE can reach 29.2% mask AP, bringing 9.6% performance improvement. Besides, by using videos of 720p, our MDQE can further improve the mask AP from 29.2% to 33.0%, which is the best result using ResNet50 backbone by far.

We compare the occlusion metrics of competing VIS methods on OVIS valid set in Table 3. One can see that MDQE achieves impressive improvements on AP_{mo} and AP_{ho} metrics, validating that MDQE can handle the moderately and heavily occluded objects very well.

Swin Large backbone. VIS models with the Swin Large backbone can have higher detection and segmentation abilities on challenging videos. Due to limited space, only the recently developed transformer-based methods are compared in Table 4. IDOL [47], SeqFormer [46] and our MDQE adopt the deformable DETR transformer architecture, while VITA [17] and MinVIS [18] employ the stronger masked-attention transformer architecture. On both YouTube-VIS 2021 and OVIS valid sets, MDQE obtains competitive performance with VITA [17] and MinVIS [18]. With 720p video input, IDOL [47] with inter-frame object re-association obtains 42.6% mask AP on OVIS. Due to our limited computational memory, we only take 2-frame video clips as inputs to train our MDQE; however, it can still reach 42.6% mask AP. We believe MDQE can achieve higher performance if more frames are used in the clip.

Parameters and Speed. We follow Detectron2 [49] to calculate the parameters and FPS. As shown in Table 2, compared with the latest per-clip method VITA [17], our MDQE has 51.4M parameters and runs at 37.8 FPS, saving about 9.4% parameters and speeding up 12% the run-time.

The visualization of example segmentation results on challenging videos by the competing VIS methods can be found in the **supplementary materials**.

5. Conclusions

We proposed to mine discriminative query embeddings to segment occluded instances on challenging videos. We first initialized the positional and content embeddings of frame-level queries by considering instance spatio-temporal features. We then performed contrastive learning on instance embeddings by proposing a new inter-instance mask repulsion loss. The proposed per-clip VIS method, termed as MDQE, was validated on OVIS and YouTube-VIS datasets. The experimental results showed that the mined discriminative embeddings of instance queries can teach the query-based segmenter to better distinguish occluded instances in crowded scenes, improving significantly performance on the challenging OVIS dataset.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1090–1099, 2022. 3

[3] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9739–9748, 2020. 1, 2

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Int. Conf. Comput. Vis.*, pages 9157–9166, 2019. 2, 3

[5] 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. *arXiv preprint arXiv:2007.14772*, 2020. 1, 2, 8

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 2, 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607. PMLR, 2020. 2, 4

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 7

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Int. Conf. Comput. Vis.*, pages 764–773, 2017. 2

[10] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 1, 2

[11] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 4

[12] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. *arXiv preprint arXiv:2105.01928*, 2021. 1, 2

[13] Fei He, Naiyu Gao, Jian Jia, Xin Zhao, and Kaiqi Huang. Queryprop: Object query propagation for high-performance video object detection. In *AAAI*, volume 36, pages 834–842, 2022. 3

[14] Fei He, Haoyang Zhang, Naiyu Gao, Jian Jia, Yanhu Shan, Xin Zhao, and Kaiqi Huang. Inspro: Propagating instance query and proposal for online video instance segmentation. *arXiv preprint arXiv:2301.01882*, 2023. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 6

[17] 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. 1, 7, 8

[18] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Adv. Neural Inform. Process. Syst.*, 2022. 1, 2, 7, 8

[19] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Adv. Neural Inform. Process. Syst.*, 34:13352–13363, 2021. 1, 2, 7

[20] 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. *Adv. Neural Inform. Process. Syst.*, 34:1192–1203, 2021. 1, 2

[21] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4019–4028, 2021. 1, 2, 4

[22] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Adv. Neural Inform. Process. Syst.*, 33:18661–18673, 2020. 2

[24] 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. 7

[25] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 3

[26] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11215–11224, 2021. 1, 2, 7, 8

[27] Minghan Li and Lei Zhang. One-stage video instance segmentation: From frame-in frame-out to clip-in clip-out. *arXiv preprint arXiv:2203.06421*, 2022. 1

[28] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. *arXiv preprint arXiv:2103.13746*, 2021. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 6

[31] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9816–9825, 2021. 2

[32] 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 *Int. Conf. Comput. Vis.*, pages 10012–10022, October 2021. 6

[33] Jiaxu Miao, Yu Wu, and Yi Yang. Identifying visible parts via pose estimation for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[35] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 164–173, 2021. 1, 2, 3

[36] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. *arXiv preprint arXiv:2207.02255*, 2022. 3

[37] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: Dataset and challenge. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 2, 5, 6, 8

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 658–666, 2019. 5

[39] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. *arXiv preprint arXiv:2003.05664*, 2020. 2, 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 8

[41] Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *Int. Conf. Comput. Vis.*, pages 10797–10806, 2021. 1, 2

[42] Wenguan Wang, James Liang, and Dongfang Liu. Learning equivariant segmentation with instance-unique querying. *arXiv preprint arXiv:2210.00911*, 2022. 1

[43] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7774–7783, 2018. 2

[44] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3024–3033, 2021. 1, 2, 4

[45] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 7

[46] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[47] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 1, 2, 3, 6, 7, 8

[48] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 959–968, 2022. 1, 2

[49] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6, 8

[50] Jin Xie, Yanwei Pang, Hisham Cholakkal, Rao Anwer, Fahad Khan, and Ling Shao. Psc-net: learning part spatial co-occurrence for occluded pedestrian detection. *Science China Information Sciences*, 64(2):1–13, 2021. 2

[51] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Int. Conf. Comput. Vis.*, pages 5188–5197, 2019. 1, 2, 5, 6, 7, 8

[52] 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 *Int. Conf. Comput. Vis.*, pages 8043–8052, 2021. 1, 2, 7, 8

[53] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2885–2895, 2022. 1, 2, 7

[54] Feng Zhu, Zongxin Yang, Xin Yu, Yi Yang, and Yunchao Wei. Instance as identity: A generic online paradigm for video instance segmentation. In *Eur. Conf. Comput. Vis.* Springer, 2022. 1

[55] 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. 2, 3, 4, 5