General and Task-Oriented Video Segmentation

Mu Chen[1], Liulei Li[1], Wenguan Wang[2], Ruijie Quan[2], and Yi Yang[2⋆]

[1] ReLER Lab, AAII, University of Technology Sydney, Australia
[2] ReLER Lab, CCAI, Zhejiang University, China
https://github.com/kagawa588/GvSeg

Abstract. We present GvSeg, a **g**eneral **v**ideo **seg**mentation framework for addressing four different video segmentation tasks (*i.e.*, instance, semantic, panoptic, and exemplar-guided) while maintaining an identical architectural design. Currently, there is a trend towards developing general video segmentation solutions that can be applied across multiple tasks. This streamlines research endeavors and simplifies deployment. However, such a highly homogenized framework in current design, where each element maintains uniformity, could overlook the inherent diversity among different tasks and lead to suboptimal performance. To tackle this, GvSeg: **i)** provides a holistic disentanglement and modeling for segment targets, thoroughly examining them from the perspective of appearance, position, and shape, and on this basis, **ii)** reformulates the query initialization, matching and sampling strategies in alignment with the task-specific requirement. These architecture-agnostic innovations empower GvSeg to effectively address each unique task by accommodating the specific properties that characterize them. Extensive experiments on seven gold-standard benchmark datasets demonstrate that GvSeg surpasses all existing specialized/general solutions by a significant margin on four different video segmentation tasks.

Keywords: Video segmentation · General solution · Task-orientation

1 Introduction

Identifying target objects and then inferring their spatial locations over time in a pixel observation constitute fundamental challenges in computer vision [1]. Depending on discriminating unique instances or semantics associated with targets, exemplary tasks include: *exemplar-guided* video segmentation (EVS) that tracks objects with given annotations at the first frame, video *instance* segmentation (VIS), video *semantic* segmentation (VSS), and video *panoptic* segmentation (VPS) which entails the delineation of foreground instance tracklets, while simultaneously assigning semantic labels to each video pixel. Prevalent work primarily adheres to discrete technical protocols customized for each task, showcasing promising results [2–21]. Nevertheless, these approaches necessitate meticulous architectural designs for each unique task, thereby posing challenges

⋆ Corresponding author: Yi Yang (yangyics@zju.edu.cn)

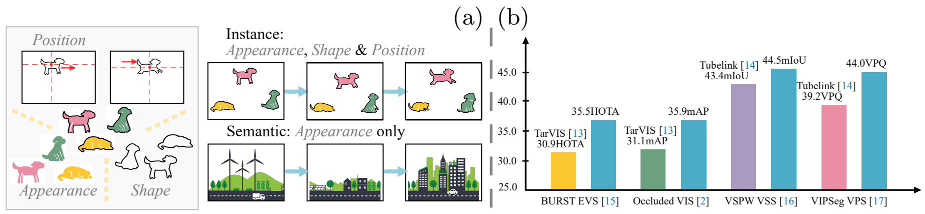

Fig. 1: (a) We render holistic modeling on segment targets by disentangling them into appearance, shape and position. (b) By adjusting the involvement of the above three factors into tracking and segmentation according to task requirement, GvSeg achieves remarkable improvement compared to prior top-leading general solutions.

in facilitating research endeavors devoting on one task to another. Recently, there have been efforts in shifting the above *task-specific* paradigm to a *general* solution that can be applied across multiple distinct tasks [22–26]. However, one concern naturally arises that such a highly homogenized framework would overlook the diversity between tasks, potentially leading to suboptimal performance. For instance, the segmenting and tracking of objects like *human* prioritize *instance discrimination* in VIS but lean towards *semantic recognition* in VSS. However, prior general approaches adopt exactly same query initialization, matching and space-time learning strategies [22, 23, 26], lacking tailored differentiation within the algorithm design that caters to the specific properties of individual tasks.

In this work, we present GvSeg, a **g**eneral **v**ideo **seg**mentation framework to address EVS, VIS, VSS, and VPS that can seamlessly accommodate ***task-oriented*** properties into the learning and inference process, while maintaining an ***identical*** architectural design. To achieve this, we rethink video segmentation in two aspects: ❶ what are the key factors that constitute segment targets (*i.e.*, *instance*, *thing*, and *stuff*), and ❷ how to leverage these key factors to build a unique sequential observation for each specific task within a general model. To address ❶, we delve deeply into the mechanism of how individuals can effectively discriminate moving instances or background stuff. The most intuitive answer in this regard is appearance, aligning with current video solutions where binary masks are classified solely based on visual representations (*i.e.*, **appearance**) [4, 27–29]. However, human perception extends beyond mere appearance [30–33]. For instance, we can also recognize moving entities such as cats in low-light conditions by referring to sketches (*i.e.*, **shape**), and distinguish distinct instances on the basis of respective spatial locations (*i.e.*, **position**), even in fast motion. Therefore, it is noteworthy that the instances to be segmented usually carry rich cues encompassing not only appearance but also position and shape characteristics. In light of the analysis above, we could assert three significant observations that contribute to the resolution of ❷: **First**, it becomes evident that current solutions downplay the importance of position and consistently ignore shape, in favor of solely appearance-based discrimination. To tackle this, we derive a *shape-position descriptor* for each object, followed by encoding them into the cross-frame query matching process to enable the participation of three key factors in discriminating corresponding instances across the entire video. **Second**, it is crucial to acknowledge that the engagement of appearance, position, and shape cues should be adjusted in accordance with the task requirements. In current general solutions, all queries are roughly initialized as empty and matched in the same manner. However, for semantic classes VSS and background *stuff* in VPS, there is no instance discrimination and overly emphasize shape/location cues would harm the generalization of the model to various targets with the same semantics. Concerning this, we advocate for a tailored query initialization and object association strategies for each task by adjusting the relative contribution of three key elements. **Third**, owing to the absence of disentanglement on segment targets, the widely used temporal contrastive learning [4, 22, 26, 34] strategy for object association in current solutions is deemed suboptimal. Concretely, prior work empirically chooses objects in nearby frames as positive samples, remaining unaware of why excluding the same instance in distant frames. In fact, entities moving in long temporal range may display similar **appearance**, but undergo strong **shape** distortion, rendering them unsuitable as positive samples for instance discrimination. Therefore, we devise a task-oriented sampling strategy that caters to *thing* and *stuff*, where instance examples are selectively sampled from the entire video by referring to shape similarity and location distance. This not only makes full use of the pre-defined *shape-position descriptors*, but also recollects valuable samples that were arbitrarily discarded in prior work. In a similar spirit, the *stuff* examples are gathered from the whole dataset which renders rich semantic description for each semantic class. Through an in-depth analysis of the essential elements that compose segmentation targets and subsequently derive task-oriented insights, our work exhibits several compelling facets: **First**, it not only recognizes but also effectively harnesses the unique nature of each task, enabling seamless accommodation of task-specific properties into segmentation models. **Second**, all of our designs are architecture-agnostic, preserving a uniform structural to efficiently address task diversity. **Third**, GvSeg substantially attains remarkable performance on each task. Notably, it surpasses existing general solutions by **4.6%** HOTA on BURST [35], **1.3%** AP on YouTube-VIS 2021 [2], **4.8%** AP on Occluded-VIS [36], **1.1%** mIoU on VSPW [37], **4.8%** VPQ on VIPSeg [38], establishing new SOTA.

2 Related Work

Exemplar-guided Video Segmentation (EVS). Given the hint which can be mask, bounding box, or point at one video frame, EVS aims to propagate the mask-level predictions to subsequent frames [25,35]. Therefore, the standard video object segmentation (VOS) task can be viewed as a specific instance of EVS – mask-guided video segmentation. Recent promising solutions for the mask-guided task mainly implemented in a *matching-based* manner which classifies pixels in current frame according to the feature similarities of target objects in reference frames [18, 20, 21, 39–42, 42–54]. To solve the bounding box and point-guided tasks, current solutions typically have to regress a pseudo ground-truth mask via pre-processing [25, 35]. In contrast, GvSeg simply adapts various kinds of hints by initializing object queries from features within regions delineated by hints.

Video Instance Segmentation (VIS). Extending beyond detecting and segmenting instances within images, VIS further engages in the active tracking of individual objects across video frames. According to the process of video sequences, existing solutions for VIS fall into three categories [28]: *online*, *semi-online*, and *offline*. The *online* methods take each frame as inputs and associate instances through hand-designed rules [2, 55–57], integrating learnable matching algorithms [58–63], or deploying query matching frameworks [4, 27, 34, 64–66]. The *semi-online* solutions typically divide long videos into clips and model the representations of instances by leveraging rich spatio-temporal information [23, 67–69]. Conversely, *offline* methods predict the instance sequence for an entire video in a single step [3, 5, 6, 63, 70, 71] which require a growing amount of GPU memory as the video length extends, limiting their application in real-world scenarios.

Video Semantic Segmentation (VSS). Building upon the principle of semantic segmentation [72–79], VSS extends this concept to video sequences, so as to capture the evolution of scenes and objects over time. Existing solutions can generally be classified into two main paradigms. The *motion-based* approaches [80–84] employ optical flow to model dynamic scenes. Though workable in certain scenarios, they rely heavily on the accuracy of flow maps and are prone to error accumulation [1]. On the other hand, the *attention-based* methods take advantage of the attention mechanism [8–10] or Transformer [85, 86] to aggregate temporal cues. This contributes to improved coherence among predictions of individual frames.

Video Panoptic Segmentation (VPS). With the emergence of seminal work [11], there has been a research trend [13, 87–92] dedicated to unifying video instance and semantic segmentation. Though showing the promise of general video segmentation, the early work [13, 87, 88] utilizes task-specific heads to handle instance and semantic segmentation separately, and assembles the panoptic predictions through post-processing. Recent algorithms typically leverage unified queries for the detection and tracking of both *thing* and *stuff* objects [89–92]. However, they demonstrate sub-optimal performance compared to task-specified solutions, emphasizing the urgency for the development of more powerful solutions.

General Video Segmentation (GVS). In order to address the limitations of task-specific models that lack the flexibility to generalize across different tasks and result in redundant research efforts, GVS aims at an all-inclusive solution for multiple video segmentation tasks. A limited number of studies [22–26, 93, 94] have ventured in this direction. However, [22, 23, 93] exhibits inferior performance compared to dedicated, task-specific methods. [25] achieves remarkable results but requires extensive pre-training on various large-scale, pixel-level annotated datasets. Inspired by these pioneers, GvSeg **i)** delves deeper into the segment targets across tasks, offering a disentanglement and modeling for them, **ii)** harnesses insights gained from **i)** to adapt task-oriented property without any modification to network architecture or training objectives, and **iii)** contributes to a robust solution that outperforms all existing specialized/general models.

Query-Based Segmentation. Image segmentation has witnessed substantial progress with top-performing approaches primarily falling into the *query-based* paradigm. Such paradigm directly models targets by introducing a set of learnable embeddings as queries to search for objects of interest and subsequently decode masks from image features. Inspired by DETR [95], the latest research [93, 96–99] takes this paradigm a step further by harnessing the Transformer architecture. This trend also spills over into video segmentation with recent solutions [22, 23, 25, 26, 93] all building upon their image segmentation counterparts. In contrast to prior work that focused solely on object appearance, GvSeg provides a holistic modeling of targets by encoding the relative position and shape cues into queries. This is particularly valuable for the tracking of instance objects. As a result, the query matching process can harness appearance, shape, and position information, enhancing object association across frames.

3 Methodology

Problem Statement. Video segmentation seeks to partition a video clip $V \in \mathbb{R}^{THW\times 3}$ containing T frames of size $H\times W$ into K non-overlap tubes linked along the time axis:

$$\{Y_k\}_{k=1}^{K} = \{(M_k, c_k)\}_{k=1}^{K}, \tag{1}$$

where each tube mask $M_k \in \{0,1\}^{T\times H\times W}$ is labeled with a category $c_k \in \{1, \cdots, C\}$. The value of K varies across tasks: in VSS, it is consistent with the number of predefined semantic categories; in EVS and VIS, it is adjusted in response to the instance count; and in VPS, it is the sum of *stuff* categories and *thing* entities.
Tracking by Query Matching. Inspired by the success of *query-based* object detectors, [4, 22, 34] propose to associate instances based on the query embeddings. Specifically, given a set of N randomly initialized queries $\{\boldsymbol{q}_n^t\}_{n=1}^{N}$, we can derive the object-centric representation $\{\hat{\boldsymbol{q}}_n^t\}_{n=1}^{N}$ for frame V^t by:

$$\{\hat{\boldsymbol{q}}_n^t\}_{n=1}^{N} = \mathcal{D}(\mathcal{E}(V^t), \{\boldsymbol{q}_n^t\}_{n=1}^{N}), \tag{2}$$

where $\mathcal{E}$ and $\mathcal{D}$ are the Transformer encoder and decoder. Here $\hat{\boldsymbol{q}}_n^t$ refines rich appearance representation for a specific object. The tracking is done by applying Hungarian Matching on the affinity matrix $\mathcal{S}_{ij} = \texttt{cosine}(\hat{\boldsymbol{q}}_i^t, \hat{\boldsymbol{q}}_j^{t+1})$ computed between $\hat{\boldsymbol{q}}_i^t$ and $\hat{\boldsymbol{q}}_j^{t+1}$ of two successive frame V^t and V^{t+1}. As such, instances exhibiting identical attributes across the video sequence are linked automatically.

3.1 GvSeg: Task-Oriented Property Accommodation Framework

GvSeg seeks to advance general video segmentation through controllable emphasis on instance discrimination and semantic comprehension according to task requirements. Concretely, we first devise a new shape-position descriptor to accurately reveal the shape and location of targets. Then, by adjusting the engagement of above shape-position descriptor during cross-frame query matching, we could realize controllable association for instance and background stuff, respectively.

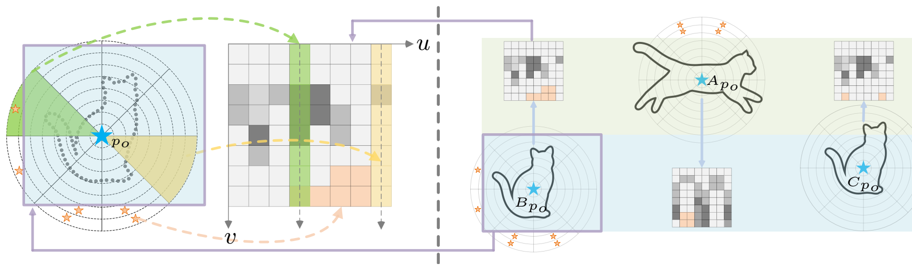

Fig. 2: Illustration of **shape-position descriptor** (§3.1).

Finally, we give an analysis on the limitation of current temporal contrastive learning and devise a task-oriented sampling strategy to tackle encountered issues.

Shape-Position Descriptor. Inspired by shape context [100], a shape-position descriptor is constructed to represent the spatial distribution and shape of target objects. First, it describes shape cues by encoding the relative geometric relationships of points in object contours relative to the object center. As shown in Fig. 2, given the contour $G \in \{0,1\}^{H \times W}$ of a target object which can be easily derived from masks, a set P with M anchor points (*i.e.*,) are evenly sampled:

$$\mathcal{P} = \{p_m = (x,y) \,|\, G(x,y) = 1,\ 1 \leq m \leq M\}. \tag{3}$$

Above anchor points are transformed into polar coordinates with the central point p_o of targets (*i.e.*, ★) as the reference point. The polar coordinate is a histogram divided into a grid of $u \times v$ bins with u angle divisions and v radius divisions. Next we calculate the number of anchor points falling within each bin:

$$\boldsymbol{H}_{i,j} = \sum_{m=1}^{M} \left\{ \begin{array}{ll} \frac{1}{\sqrt{d_{\mathrm{model}}}} & \text{if} \quad |\theta_m - \hat{\theta}_i| \leq \frac{\Delta\theta}{2} \quad \text{and} \quad |r_m - \hat{r}_j| \leq \frac{\Delta r}{2} \\ 0 & \text{otherwise} \end{array} \right\}, \tag{4}$$

where $\Delta\theta$, Δr, and $(\hat{\theta}_i, \hat{r}_j)$ are the angle span, radius span, and center point of each bin, (θ_m, r_m) is the polar coordinate of anchor point p_m, d_{model} is the embedding dimension of model. As such, $\boldsymbol{H}$ expresses the spatial configuration of contour G relative to center point (*i.e.*, p_o) in a compact and robust way. As depicted in Fig. 2, instances with different shapes (*i.e.*, target A and B) present varying distributions of $\boldsymbol{H}$ which demonstrates the capability to encode the shape cues of target objects. Moreover, we equip $\boldsymbol{H}$ with the ability to account for the relative spatial location of target objects by setting $\boldsymbol{H}_{i,j} = -1/\sqrt{d_{\mathrm{model}}}$ if the center point of a bin (*i.e.*, ☆) falls outside of masks. Therefore, instances with similar shapes but different locations (*i.e.*, target B and C) would yield similar distribution of positive values, but distinct distribution of negative values, effectively evolving above shape descriptor into a **shape-position** descriptor.

Shape- and Position-Aware (SPA) Query Matching. Given the above analysis, a set of shape-position descriptors $\{\boldsymbol{H}_k\}_{k=1}^{K}$ could be derived from each object k within the mask. We then aim to facilitate the awareness of shape-position

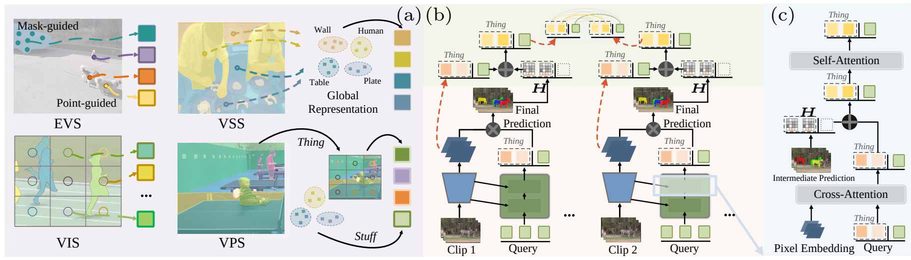

Fig. 3: (a) Task-oriented queries initialization. (b) Task-oriented object association tailored w.r.t. *thing* and *stuff* objects. (c) Shape- and position-aware query matching.

cues for object association between frames, by integrating such descriptors into the query matching process. To achieve this, as shown in Fig. 3 (c), we draw inspiration from the absolute position encoding (APE) which is widely adopted in Transformer [101]. Specifically, during mask decoding, N query embeddings $\{\boldsymbol{q}_n\}_{n=1}^{N}$ is interacting with the backbone feature $\boldsymbol{F}$ to retrieve object-centric feature in each decoder layer by:

$$\boldsymbol{q}^l = \texttt{CrossAttn}(\boldsymbol{q}^{l-1}, \boldsymbol{F}), \quad \boldsymbol{q}^l = \texttt{SelfAttn}(\boldsymbol{q}^l, \boldsymbol{q}^l) \tag{5}$$

Where l is the layer index. Typically, a Hungarian Matching matrix $\mathbb{1}^l \in \{0,1\}^{N\times K}$ between N predictions generated from query embeddings and K ground truth objects can be derived from each decoding layer. Following the principle of APE, where the position encodings $\mathbf{P}$ is integrated into $\boldsymbol{q}$: $\boldsymbol{q} \leftarrow \boldsymbol{q} + \mathbf{P}$, we assign $\{\boldsymbol{H}_k\}_{k=1}^{K}$ to K elements in $\boldsymbol{q}$ that corresponds to the object described in ground truth by referring to $\mathbb{1}^{l-1}$ produced from prior decoding layer: $\boldsymbol{q}^l \leftarrow \boldsymbol{q}^l + \mathbb{1}^{l-1} \cdot \boldsymbol{H}$ before conducting `SelfAttn`. Note the K elements in $\{\boldsymbol{H}_k\}_{k=1}^{K}$ are flattened and bilinearly interpolated to size d_{model}, and then stacked together to get $\boldsymbol{H} \in \mathbb{R}^{K\times d_{\text{model}}}$. In this way, the query embeddings can **i)** well attend to and discriminate corresponding objects by injecting the descriptors into `SelfAttn`, and **ii)** be aware to shape-position cues after mask decoding (*i.e.*, $\hat{\boldsymbol{q}}$ in Eq. 2). To further reinforce the consideration to shape and position of targets in $\hat{\boldsymbol{q}}$, we compile $\boldsymbol{H}$ into the affinity-based query matching between two adjacent frames:

$$\mathcal{S}_{ij} = \texttt{cosine}(\hat{\boldsymbol{q}}_i^t + \boldsymbol{H}_i^t, \hat{\boldsymbol{q}}_j^{t+1} + \boldsymbol{H}_j^{t+1}). \tag{6}$$

As such, each query embedding is seamlessly incorporated with the unique attributes of corresponding objects, thereby endowing them with a heightened sensitivity to specific targets when matching with other frames afterward.

Task-Oriented Query Initialization & Object Association. To orient the model towards specific tasks, existing work usually employs dedicated queries (*i.e.*, *stuff*/*thing* query) for semantic/instance segmentation [90, 102], and process them parallel by modifying the model into a two-path architecture. In contrast, GvSeg smartly addresses this challenge by dynamically adjusting the involvement

of three key constitutes, *i.e.*, **appearance**, **shape**, and **position** within the query initialization (*i.e.*, Fig. 3 (a)) and object association (*i.e.*, Fig. 3 (b)).
• **EVS** underscores the utilization of given hints to guide the segmentation of subsequent frames. To flexibly unleash the potential of different kinds of hints under the *track by query matching* paradigm, we propose to initialize the query embeddings from backbone features sampled within hinted regions. Specifically, for the point-guided task which provides a single point $p_k = (x, y)$ to indicate the target object, the backbone feature at corresponding location can be sampled by:

$$\boldsymbol{f}_k = \texttt{sample}(\boldsymbol{F},\ p_k), \tag{7}$$

where the implementation of `sample` follows PointRent [103]. Then, the query embedding is initialized with f_k: $\bar{\boldsymbol{q}}_k = \texttt{FFN}(\boldsymbol{f}_k)$ to fulfill the guidance ability of given exemplars where `FFN` is a feed-forward network. For the mask and box guided tasks, we sample multiple f_k and average them to get the feature that comprehensively describes target objects. Finally, SPA query matching is applied to enhance instance discrimination during the object association between frames.
• **VIS** emphasizes the tracking of instances which usually exhibits unique attributes for discrimination. To encode these instance-specific properties (*e.g.*, location, appearance) into query embeddings, we follow [104] to initialize $\boldsymbol{q} \in \mathbb{R}^{N \times D}$ from the backbone features. Concretely, we partition the backbone features into $S \times S$ grids and flatten them, resulting in $\{\boldsymbol{F}_i\}_{i=1}^{S \times S}$. We then randomly select N elements from this set for the initialization of queries and obtain $\{\bar{\boldsymbol{q}}_i\}_{i=1}^{N}$:

$$[\bar{\boldsymbol{q}}_0; \cdots; \bar{\boldsymbol{q}}_N] = \texttt{FFN}(\boldsymbol{F}). \tag{8}$$

As such, queries could involve appearance and location cues for diverse instances present in the frame. Similarly to EVS, we apply SPA query matching for object association to enable more precise instance discrimination across the entire video.
• **VSS** prioritizes semantic understanding of each class. Therefore, to enhance the thorough grasp of semantics, we continuously collect the query embeddings corresponding to each semantic class during training. More precisely, given N queries $\boldsymbol{q} \in \mathbb{R}^{N \times D}$, we gather K entities from them based on the bipartite matching results $\mathbb{1} \in \{0, 1\}^{K \times N}$ between predictions generated from $\boldsymbol{q}$ and ground truth:

$$\bar{\boldsymbol{q}} = \mathbb{1} \odot \boldsymbol{q} \in \mathbb{R}^{K \times D}. \tag{9}$$

Here $\bar{\boldsymbol{q}}$ encodes the semantic-specific properties for each class, and we momentously update it in each training step to approximate the global representation of semantic classes over the entire dataset. During inference, we initialize object queries for each frame from $\bar{\boldsymbol{q}}$. Note we do not apply SPA query matching for VSS, as shape and location cues would harm semantic-level tracking.
• **VPS** integrates both instance-discrimination for foreground *thing* classes and semantic interpretation for background *stuff* categories. We thus combine the query initialization and association strategies used in VIS and VSS, to facilitate the effective recognition and tracking for *thing* and *stuff* classes, respectively.
Task-Oriented Temporal Contrastive Learning. The performance of current *track by query matching-based* solutions depends significantly on the temporal

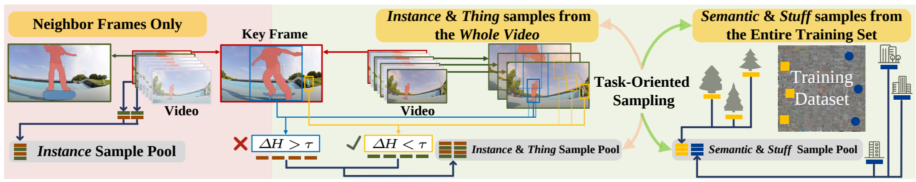

Fig. 4: Illustration of **task-oriented temporal contrastive learning** (§3.1). Prior work considers solely *instance* objects, and samples are restricted within neighbor frames. In UvSeg, *instance* & *thing* samples are collected from the whole video according to shape and location similarity, while *semantic* & *stuff* samples are gathered from the entire training set to capture diver shapes and appearances of each semantic class.

contrastive learning (TCL) between frames. Given a key frame, prior methods [22, 26, 34] typically select reference frames from the temporal neighborhood, while ignoring all other frames. This leads to limited positive/negative samples for effective contrastive learning which relies on a substantial quantity of samples to achieve optimal performance. To maximize the usage of these discarded samples, we devise a smart sampling strategy that caters to individual tasks and addresses the challenge of accurately distinguishing the positive ones from them (*i.e.*, Fig. 4). Specifically, for tasks leaning towards instance discrimination (*i.e.*, VIS, EVS and *thing* in VPS), it is essential to note that not all identical instances in the same video are suitable as positive samples. This is due to the strong variations in shape and spatial location among instances, which can disrupt the local consistency between the same instance at nearby frames that usually manifest similar shape and position. To tackle this, in contrast to existing work arbitrarily discards samples in distant frames, we sample examples across the whole video by measuring the shape and location similarity. The variation of shape-position descriptors (*i.e.*, ΔH) belonging to the same instance but at frame V^t and V^{t+n} is computed via:

$$\Delta H = \frac{\|\boldsymbol{H}^{t+n} - \boldsymbol{H}^{t}\|_2}{\|\boldsymbol{H}^{t}\|_2}. \tag{10}$$

We set a threshold $\tau = 0.2$ and consider the query embedding associated with $\boldsymbol{H}^{t+n}$ as a positive example if ΔH is smaller than τ; otherwise, it is deemed negative. As such, we involve distant frames into the reference set which enriches the diversity of samples and bolsters the robustness of TCL. On the other hand, for VSS and background *stuff* classes in VPS, samples are relaxed to select from the whole training set, as larger mount of entities with diverse appearance, shape, and location will improve the grasp of semantics. To implement this, we maintain a first-in-first-out queue $\mathcal{Q}$ that contains $N_{\mathcal{Q}}$ queries for each pre-defined semantic class. Elements in $\mathcal{Q}$ will engage in TCL and be updated with new samples at each training step. We set $N_{\mathcal{Q}}$ to a relatively small number (*e.g.*, 100), which incurs negotiable cost in training time but considerable improvement in performance.

Table 1: Quantitative results for VPS on VIPSeg [38] and KITTI-STEP [12] (§4.1), and VSS on VSPW [37] (§4.2).

Method	Backbone	General Solution	VIPSeg `val`				KITTI-STEP `val`				VSPW `val`		
			VPQ	VPQ^{Th}	VPQ^{St}	STQ	VPQ	STQ	AQ	SQ	mIoU	mVC_8	mVC_{16}
VPSNet [11]	R-50	✗	14.0	14.0	14.2	20.8	0.43	0.56	0.52	0.61	-	-	-
Mask-Prop [11]	R-50	✗	-	-	-	-	-	0.67	0.63	0.71	-	-	-
MotionLab [12]	R-50	✗	-	-	-	-	0.40	0.58	0.51	0.67	-	-	-
SiamTrack [13]	R-50	✗	17.2	17.3	17.3	21.1	-	-	-	-	-	-	-
TCB [37]	R-101	✗	-	-	-	-	-	-	-	-	37.5	86.9	82.1
DVIS [94]	R-50	✗	43.2	43.6	42.8	42.8	-	-	-	-	-	-	-
Mask2Former [93]	R-50	✓	-	-	-	-	-	-	-	-	38.4	87.5	82.5
TubeFormer [23]	R-50	✓	26.9	-	-	38.6	0.51	0.70	0.64	0.76	-	-	-
Video K-Net [22]	R-50	✓	26.1	-	-	31.5	0.46	0.71	0.70	0.71	37.9	87.0	82.1
TarVIS [25]	R-50	✓	33.5	39.2	28.5	43.1	-	0.70	0.70	0.69	-	-	-
DEVA [105]	R-50	✓	38.3	-	-	41.5	-	-	-	-	-	-	-
Tube-Link [26]	R-50	✓	39.2	-	-	39.5	0.51	0.68	0.67	0.69	43.4	89.2	85.4
GvSeg	R-50	✓	**44.0**	**44.4**	**42.4**	**44.9**	**0.53**	**0.71**	**0.69**	**0.71**	**44.5**	**90.5**	**86.4**
CFFM [10]	MiT-B5	✗	-	-	-	-	-	-	-	-	49.3	90.8	87.1
MRCFA [86]	MiT-B2	✗	-	-	-	-	-	-	-	-	49.9	90.9	87.4
DVIS [94]	Swin-L	✗	57.6	59.9	55.5	55.3	-	-	-	-	-	-	-
Video K-Net [22]	Swin-B	✓	-	-	-	-	-	-	-	-	57.2	90.1	87.8
TarVIS† [25]	Swin-L	✓	48.0	58.2	39.0	52.9	-	-	-	-	-	-	-
DEVA [105]	Swin-L	✓	52.2	-	-	52.2	-	-	-	-	-	-	-
Tube-Link [26]	Swin-B	✓	50.4	-	-	49.4	0.56	0.72	0.69	0.74	62.3	91.4	89.3
GvSeg	Swin-B	✓	**55.3**	**57.2**	**52.3**	**52.4**	**0.58**	**0.74**	**0.73**	**0.74**	**63.2**	**91.8**	**89.4**
GvSeg	Swin-L	✓	**57.9**	**59.7**	**56.1**	**55.6**	-	-	-	-	**65.5**	**93.8**	**91.6**

3.2 Implementation Details

Network Configuration. GvSeg is a semi-online video segmentation framework built upon the *tracking by query matching* paradigm [4]. It comprises an *image-level segmenter* to extract frame-level queries, and an object associator to match query embeddings across frames. The *image-level segmenter* is implemented as Mask2Former [93] with both ResNet-50 [106] and Swin-L [107] as the backbone. Given the most recent work typically adopts clip-level inputs for richer temporal cues [5, 26, 28], in alignment with this trend, GvSeg takes a clip containing three frames as input each time. The size of points set $\mathcal{P}$ derived from object contour is fixed to 200 to make the shape-position descriptor effectively characterize objects of varying scales. We employ $u = 36$ angle divisions and $v = 12$ radius divisions to capture point distribution in finer granularity.

Training. Following the standard protocols [5,23,26,48,108] in video segmentation, the maximum training iteration is set to 10K for OVIS/VSPW/VIPSeg/KITTI and 15K for YouTube-VOS_{18}/YouTube-VIS_{21} with a mini-batch size of 16. The AdamW optimizer with initial learning rate 0.001 is adopted. More details regarding training and testing can be found in *Supplementary Material*.

4 Experiment

4.1 Results for Video Panoptic Segmentation

Dataset. VIPSeg [38] provides 2,806/323 videos in `train`/`test` splits which covers 232 real-world scenarios and 58/66 thing/stuff classes. KITTI-STEP [12]

Table 2: Quantitative results for VIS on OVIS [36] and YouTube-VIS$_{21}$ [2] (§4.3).

Method	Backbone	General Solution	Occluded-VIS `val`					Youtube-VIS$_{21}$ `val`				
			AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$	AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
SipMask [55]	R-50	✗	10.2	24.7	7.8	7.9	15.8	31.7	52.5	34.0	30.8	37.8
InsPro [27]	R-50	✗	-	-	-	-	-	37.6	58.7	0.9	32.7	41.4
SeqFormer [6]	R-50	✗	-	-	-	-	-	40.5	62.4	43.7	36.1	48.1
VITA [5]	R-50	✗	19.6	41.2	17.4	11.7	26.0	45.7	67.4	49.5	40.9	53.6
MinVIS [4]	R-50	✗	25.0	45.5	24.0	13.9	29.7	44.2	66.0	48.1	39.2	51.7
IDOL [34]	R-50	✗	30.2	51.3	30.0	15.0	37.5	43.9	68.0	49.6	38.0	50.9
MDQE [66]	R-50	✗	33.0	57.4	32.2	15.4	38.4	44.5	67.1	48.7	37.9	49.8
DVIS [94]	R-50	✗	34.1	59.8	32.3	15.9	41.1	-	-	-	-	-
GenVIS [28]	R-50	✗	34.5	59.4	35.0	16.6	38.3	47.1	67.5	51.5	41.6	54.7
TCOVIS [109]	R-50	✗	35.3	60.7	36.6	15.7	39.5	49.5	71.2	53.8	41.3	55.9
CTVIS [110]	R-50	✗	35.5	60.8	34.9	16.1	41.9	50.1	73.7	54.7	41.8	59.5
TubeFormer [23]	R-50	✓	-	-	-	-	-	41.2	60.4	44.7	40.4	54.0
CAROQ [24]	R-50	✓	25.8	47.9	25.4	14.2	33.9	43.3	64.9	47.1	39.3	52.7
TarVIS [25]	R-50	✓	31.1	52.5	30.4	15.9	39.9	48.3	69.6	53.2	40.5	55.9
Tube-Link [26]	R-50	✓	29.5	51.5	30.2	15.5	34.5	47.9	70.0	50.2	42.3	55.2
GvSeg	R-50	✓	**35.9**	**50.7**	**38.0**	**16.6**	**40.1**	**49.6**	**72.0**	**53.1**	**42.7**	**56.7**
GenVIS [28]	Swin-L	✗	45.4	69.2	47.8	18.9	49.0	59.6	80.9	65.8	48.7	65.0
TCOVIS [109]	Swin-L	✗	46.7	70.9	49.5	19.1	50.8	61.3	82.9	68.0	48.6	65.1
CTVIS [110]	Swin-L	✗	46.9	71.5	47.5	19.1	52.1	61.2	84.0	68.8	48.0	65.8
CAROQ [24]	Swin-L	✓	-	-	-	-	-	54.5	75.4	60.5	45.5	61.4
TarVIS [25]	Swin-L	✓	43.2	67.8	44.6	18.0	50.4	60.2	81.4	67.6	47.6	64.8
Tube-Link [26]	Swin-L	✓	-	-	-	-	-	58.4	79.4	64.3	47.5	63.6
GvSeg	Swin-L	✓	**49.7**	**74.9**	**52.0**	**18.9**	**54.5**	**60.7**	**82.9**	**69.7**	**47.5**	**65.7**

is an urban street-view dataset with 12/9 videos for `train`/`val`. It includes 19 semantic classes, with two of them (*pedestrians* and *cars*) having tracking IDs.

Performance. As illustrated by Table 1, GvSeg achieves dominant results on VIPSeg [38], presenting an improvement up to **4.8%/5.4%** in terms of VPQ/STQ over the SOTA [26] with ResNet-50 as backbone. This reinforces our belief that accommodating task-oriented property into general video segmentation is imperative. Such an assertion gets further support on KITTI-STEP [12] that GvSeg outperforms all existing solutions by significant margins in STQ and AQ, which focus more on the coherent association of identical objects.

4.2 Results for Video Semantic Segmentation

Dataset. VSPW [37] has 2,806/343 in-the-wild videos with 198,224/24,502 frames for `train`/`val`, and provides pixel-level annotations for 124 semantic categories.

Performance. As shown in Table 1, based on ResNet-50, GvSeg outperforms all competitors and achieves **44.5%** mIoU. In particular, the **90.5%/86.4%** scores in terms of mVC$_8$/mVC$_{16}$ are comparable to MRCFA [86] which utilizes Swin-B as the backbone and yields much higher mIoU. This suggests that, benefited by task-oriented temporal contrast learning, GvSeg can produce more consistent prediction across frames. When integrated with Swin-B, GvSeg demonstrates **0.9%** gains over Tube-Link [26], confirming the superiority of our approach.

Table 3: Quantitative results for EVS on YouTube-VOS$_{18}$ [111], and BURST [35] (§4.4).

Method	Backbone	General Solution	YouTube-VOS$_{18}$ `val` (Mask-guide)					BURST `val` (Point-guide)		
			$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	H_{all}	H_{com}	H_{unc}
Box Tracker [112]	R-50	✗	-	-	-	-	-	12.7	31.7	7.9
STCN [48]	R-50	✗	83.0	81.9	86.5	77.9	85.7	24.4	44.0	19.5
XMem [49]	R-50	✗	85.7	84.6	89.3	80.2	88.7	32.3	47.5	28.6
UNINEXT [113]	R-50	✓	77.0	76.8	81.0	70.8	79.4	-	-	-
TarVIS [25]	R-50	✓	79.2	79.7	84.2	72.9	79.9	30.9	43.2	27.8
GvSeg	R-50	✓	**81.5**	**80.9**	**86.0**	**75.4**	**83.7**	**35.9**	**49.6**	**32.7**
UNINEXT [113]	ConvNeXt-L	✓	78.1	79.1	83.5	71.0	78.9	-	-	-
TarVIS [25]	Swin-L	✓	82.1	82.3	86.5	76.1	83.5	37.5	51.7	34.0
GvSeg	Swin-L	✓	**84.3**	**82.7**	**87.9**	**78.5**	**87.1**	**40.9**	**55.5**	**36.3**

4.3 Results for Video Instance Segmentation

Dataset. Occluded VIS [36] is specifically designed to tackle the challenging scenario of object occlusions. It consists of 607/140 long videos with up to 292 frames for `train`/`val` and spans 25 object categories with a high density of instances. YouTube-VIS$_{21}$ [2] comprises 2,985/421 high resolution videos for `train`/`val`. It extensively covers 40 object classes with 8,171 unique instances.
Performance. From Table 2 we can observe that GvSeg provides a considerable performance gain over existing methods on Occluded-VIS [36]. Notably, it outperforms the prior specalized/general solution SOTA CTVIS [110]/TarVIS [25] by **0.4%/4.8%** in terms of mAP with ResNet-50 as the backbone. When adopting Swin-L, GvSeg showcases far better performance, achieving up to **49.7%** mAP which earns an impressive **2.8%** improvement against CTVIS. Moreover, we report performance on YouTube-VIS$_{21}$ [2]. As seen, GvSeg surpasses the main rival (*i.e.*, TarVIS), by **1.3%/0.5%** with ResNet-50/Swin-L as backbone.

4.4 Results for Exemplar-guided Video Segmentation

Dataset. YouTube-VOS$_{18}$ [111] includes 3, 471/474 videos for `train`/`val`. The videos are sampled at 30 FPS and annotated per 5 frames with multiple objects. BURST [35] contains 500/993/1, 421 videos for `train`/`val`/`test`. It provides mask/point/bounding box as exemplars and averages over 1000 frames per video.
Performance. To make a fair comparison with existing work which usually tests on BURST without training, we train GvSeg on YouTube-VOS$_{18}$ and randomly adopt mask or point exemplars as the guidance. Then the performance is evaluated with mask exemplar on YouTube-VOS$_{18}$ and point exemplar on BURST. As shown in Table 3, GvSeg yields satisfactory performance on YouTube-VOS$_{18}$, *i.e.*, surpassing the general counterpart (*i.e.*, TarVIS [25]) by **2.3%/2.2%** in terms of $\mathcal{G}$ score with ResNet-50/Swin-L as the backbone. We also provide the point-guided segmentation results on BURST. As seen, GvSeg surpasses current solutions by a large margin across all metrics. For instance, When compared with task-specialized approaches (*e.g.*, XMem [49]), our approach still earns **3.6%** improvement. Note existing work has to adopt an additional offline model for mask prediction with given points, while our method natively supports points as the exemplar, contributing to the superiority in both efficiency and effectiveness.

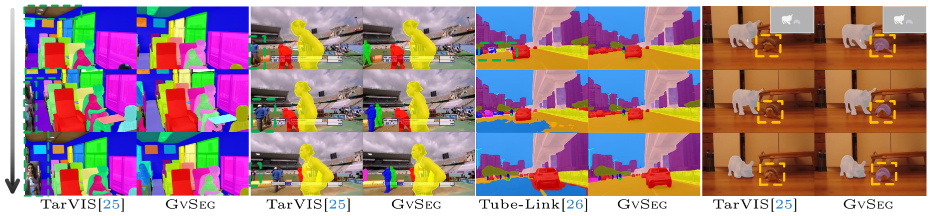

Fig. 5: Visual comparison results on VIPSeg-VPS [38], YouTube-VIS$_{21}$ [2], VSPW-VSS [37] and YouTube-VOS$_{18}$ [111] (§4.5).

4.5 Qualitative Results

In Fig. 5, we visualize the comparisons of GvSeg against the top-leading methods on four different tasks (*i.e.*, VPS, VIS, VSS, and EVS). As seen, GvSeg gives more precise and consistent predictions in challenging scenarios.

4.6 Diagnostic Experiment

For more detailed analysis, we conduct a set of ablative studies on VIPSeg-VPS [38] with ResNet-50 as the backbone.

Key Component Analysis. We investigate the improvements brought by each component of GvSeg in Table 4a where 'SPA' indicates 'shape-position aware'. First, it can be observed that SPA query matching brings a considerable improvement over the Baseline, *i.e.*, **1.8%**/**1.2%** concerning VPQ and STQ. This verifies our modeling of segment targets by disentangling them into appearance, shape, and position. Moreover, the adoption of task-oriented strategies for query initialization, object association, and temporal contrastive learning (TCL) elevates the results to a new level. Finally, we combine all these designs together which results in GvSeg and obtains the optimal performance. This confirms the compatibility of each component and the effectiveness of our whole algorithm.

Matching Threshold & Queue Length. The results with different threshold τ and queue length N_Q utilized in task-oriented TCL are reported in Table 4b. Though larger size of samples in the queue contributes to higher scores, we remain N_Q to 100 which gives nearly no impact in training speed and memory usage.

Histogram Size. In Table 4c, we investigate the impact of the number of bins within the polar-style histogram for building position-shape descriptor. As seen, there is minor change in performance if $u \times v$ is large enough (*e.g.*, > 200) to capture the fine-grained variation in shape and location.

Task-Oriented Object Association. We probe the impact of integrating distinct cues into object association in Table 4d. By comparing *Row* #2 to #1 we can observe that considering shape and position can boost the performance for *thing* objects. In stark contrast, the inclusion of these cues causes negative impacts and yields less favorable results for *stuff* objects (*i.e.*, *Row* #3 *vs.* #1).

Table 4: A set of ablative studies on VIPSeg-VPS [38] `val` with ResNet-50 [106] as the backbone (§4.6). The adopted settings are marked in red.

Component	VPQ ↑	STQ ↑
Baseline	36.0	37.3
+ SPA query matching	37.8	38.5
+ Task-oriented init.&asso.	40.1	40.7
+ Task-oriented TCL	41.2	42.0
GvSeg	**44.0**	**44.9**

(a) Component analysis

τ	$N_{\mathcal{Q}}$	VPQ ↑	STQ ↑
0.1	100	43.3	43.9
0.2	**100**	**44.0**	**44.9**
0.2	200	44.1	45.1
0.3	100	43.6	44.4
0.3	200	43.7	44.6

(b) Task-oriented TCL

Angle u	Radius v	VPQ ↑	STQ ↑
12	6	43.1	43.8
24	12	43.6	44.3
36	**12**	**44.0**	**44.9**
36	18	43.9	45.0
48	12	44.0	44.8

(c) Shape-position descriptor

#	*Thing*		*Stuff*		VPQ ↑	STQ ↑
	Appear.	Shape & Pos.	Appear.	Shape & Pos.		
1	✓		✓		42.1	43.1
2	✓	✓	✓		**44.0**	**44.9**
3	✓		✓	✓	41.7	42.8
4	✓	✓	✓	✓	42.9	43.4

(d) Task-oriented query association

#	*Thing*		*Stuff*		VPQ ↑	STQ ↑
	Frame	Video	Frame	Dataset		
1	✓		✓		40.1	40.7
2	✓			✓	42.4	43.3
3		✓	✓		43.0	43.9
4		✓		✓	**44.0**	**44.9**

(e) Task-oriented example sampling

This proves the necessity and urgency to cater to the task-oriented property which emphasizes more on *instance discrimination* or *semantic understanding*.

Task-Oriented Example Sampling. To determine the contribution of our devised example sampling strategy utilized in TCL, we examine the performance w.r.t. *thing* and *stuff* categories in Table 4e where 'Frame' refers to selecting samples from nearby frames, 'Video' indicates gathering samples across the entire video based on shape-position descriptor for instance discrimination, and 'Dataset' means storing samples in a queue to enhance the comprehension of semantics. As seen, both 'Video' and 'Dataset' level sampling for *thing* and *stuff* classes boost the scores significantly. This verifies our core insight that current sampling strategy in TCL is sub-optimal, and we can improve it by rendering a more holistic modeling on segment targets to select richer and more suitable samples.

5 Conclusion

We present GvSeg, the first general video segmentation solution that accommodates task-oriented properties into model learning. To achieve this, we first render a holistic investigation on segment targets by disentangling them into three essential constitutes: appearance, shape, and position. Then, by adjusting the involvement of these three key elements in query initialization and object association, we realize customizable prioritization of *instance discrimination* or *semantic understanding* to address different tasks. Moreover, task-oriented temporal contrastive learning is proposed to accumulate a diverse range of informative samples that considers both local consistency and semantic understanding properties for tracking instances and semantic/background classes, respectively. In this manner, GvSeg offers tailored consideration for each individual task and consistently obtains top-leading results in four video segmentation tasks.

References

1. Zhou, T., Porikli, F., Crandall, D.J., Van Gool, L., Wang, W.: A survey on deep learning technique for video segmentation. IEEE TPAMI **45**(6), 7099–7122 (2022) 1, 4
2. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 1, 3, 4, 11, 12, 13
3. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR (2021) 1, 4
4. Huang, D.A., Yu, Z., Anandkumar, A.: Minvis: A minimal video instance segmentation framework without video-based training. In: NeurIPS (2022) 1, 2, 3, 4, 5, 10, 11
5. Heo, M., Hwang, S., Oh, S.W., Lee, J.Y., Kim, S.J.: Vita: Video instance segmentation via object token association. In: NeurIPS (2022) 1, 4, 10, 11
6. Wu, J., Jiang, Y., Bai, S., Zhang, W., Bai, X.: Seqformer: Sequential transformer for video instance segmentation. In: ECCV (2022) 1, 4, 11
7. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020) 1
8. Paul, M., Danelljan, M., Van Gool, L., Timofte, R.: Local memory attention for fast video semantic segmentation. In: IROS (2021) 1, 4
9. Ji, W., Li, J., Bian, C., Zhou, Z., Zhao, J., Yuille, A.L., Cheng, L.: Multispectral video semantic segmentation: A benchmark dataset and baseline. In: CVPR (2023) 1, 4
10. Sun, G., Liu, Y., Ding, H., Probst, T., Van Gool, L.: Coarse-to-fine feature mining for video semantic segmentation. In: CVPR (2022) 1, 4, 10
11. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video panoptic segmentation. In: CVPR (2020) 1, 4, 10
12. Weber, M., Xie, J., Collins, M., Zhu, Y., Voigtlaender, P., Adam, H., Green, B., Geiger, A., Leibe, B., Cremers, D., et al.: Step: Segmenting and tracking every pixel. In: NeurIPS (2021) 1, 10, 11
13. Woo, S., Kim, D., Lee, J.Y., Kweon, I.S.: Learning to associate every segment for video panoptic segmentation. In: CVPR (2021) 1, 4, 10
14. Liang, C., Wang, W., Zhou, T., Miao, J., Luo, Y., Yang, Y.: Local-global context aware transformer for language-guided video segmentation. IEEE TPAMI **45**(8), 10055–10069 (2023) 1
15. Hui, T., Liu, S., Ding, Z., Huang, S., Li, G., Wang, W., Liu, L., Han, J.: Language-aware spatial-temporal collaboration for referring video segmentation. IEEE TPAMI **45**(7), 8646–8659 (2023) 1
16. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023) 1
17. Wang, W., Shen, J., Li, X., Porikli, F.: Robust video object cosegmentation. IEEE TIP **24**(10), 3137–3148 (2015) 1
18. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR (2015) 1, 3
19. Wang, W., Shen, J., Xie, J., Porikli, F.: Super-trajectory for video segmentation. In: ICCV (2017) 1
20. Lu, X., Wang, W., Shen, J., Tai, Y.W., Crandall, D.J., Hoi, S.C.: Learning video object segmentation from unlabeled videos. In: CVPR (2020) 1, 3
21. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Gool, L.V.: Video object segmentation with episodic graph memory networks. In: ECCV (2020) 1, 3

22. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Video k-net: A simple, strong, and unified baseline for video segmentation. In: CVPR (2022) 2, 3, 4, 5, 9, 10
23. Kim, D., Xie, J., Wang, H., Qiao, S., Yu, Q., Kim, H.S., Adam, H., Kweon, I.S., Chen, L.C.: Tubeformer-deeplab: Video mask transformer. In: CVPR (2022) 2, 4, 5, 10, 11
24. Choudhuri, A., Chowdhary, G., Schwing, A.G.: Context-aware relative object queries to unify video instance and panoptic segmentation. In: CVPR (2023) 2, 4, 11
25. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: CVPR (2023) 2, 3, 4, 5, 10, 11, 12, 13
26. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Tube-link: A flexible cross tube baseline for universal video segmentation. In: ICCV (2023) 2, 3, 4, 5, 9, 10, 11, 13
27. He, F., Zhang, H., Gao, N., Jia, J., Shan, Y., Zhao, X., Huang, K.: Inspro: Propagating instance query and proposal for online video instance segmentation. In: NeurIPS (2022) 2, 4, 11
28. Heo, M., Hwang, S., Hyun, J., Kim, H., Oh, S.W., Lee, J.Y., Kim, S.J.: A generalized framework for video instance segmentation. In: CVPR (2023) 2, 4, 10, 11
29. Qin, Z., Lu, X., Nie, X., Liu, D., Yin, Y., Wang, W.: Coarse-to-fine video instance segmentation with factorized conditional appearance flows. IEEE/CAA Journal of Automatica Sinica **10**(5), 1192–1208 (2023) 2
30. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: Human vision and electronic imaging VI (2001) 2
31. Loomis, J.M., Philbeck, J.W., Zahorik, P.: Dissociation between location and shape in visual space. Journal of Experimental Psychology: Human Perception and Performance **28**(5), 1202 (2002) 2
32. Wang, W., Yang, Y., Pan, Y.: Visual knowledge in the big model era: Retrospect and prospect. arXiv preprint arXiv:2404.04308 (2024) 2
33. Yang, Z., Chen, G., Li, X., Wang, W., Yang, Y.: Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In: ICML (2024) 2
34. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: ECCV (2022) 3, 4, 5, 9, 11
35. Athar, A., Luiten, J., Voigtlaender, P., Khurana, T., Dave, A., Leibe, B., Ramanan, D.: Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In: WACV (2023) 3, 4, 12
36. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: A benchmark. IJCV **130**(8), 2022–2039 (2022) 3, 11, 12
37. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: CVPR (2021) 3, 10, 11, 13
38. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: CVPR (2022) 3, 10, 11, 13, 14
39. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by multi-scale foreground-background integration. IEEE TPAMI **44**, 4701–4712 (2021) 3
40. Miao, J., Wei, Y., Yang, Y.: Memory aggregation networks for efficient interactive video object segmentation. In: CVPR (2020) 3

41. Wu, R., Lin, H., Qi, X., Jia, J.: Memory selection network for video propagation. In: ECCV (2020) 3
42. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: ICCV (2019) 3
43. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: CVPR (2019) 3
44. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019) 3
45. Wang, W., Shen, J., Porikli, F., Yang, R.: Semi-supervised video object segmentation with super-trajectories. IEEE TPAMI **41**(4), 985–998 (2018) 3
46. Seong, H., Oh, S.W., Lee, J.Y., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: CVPR (2021) 3
47. Mao, Y., Wang, N., Zhou, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: ICCV (2021) 3
48. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In: NeurIPS (2021) 3, 10, 12
49. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022) 3, 12
50. Li, L., Zhou, T., Wang, W., Yang, L., Li, J., Yang, Y.: Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In: CVPR (2022) 3
51. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Per-clip video object segmentation. In: CVPR (2022) 3
52. Yu, Y., Yuan, J., Mittal, G., Fuxin, L., Chen, M.: Batman: Bilateral attention transformer in motion-appearance neighboring space for video object segmentation. In: ECCV (2022) 3
53. Zhang, Y., Li, L., Wang, W., Xie, R., Song, L., Zhang, W.: Boosting video object segmentation via space-time correspondence learning. In: CVPR (2023) 3
54. Li, L., Wang, W., Zhou, T., Li, J., Yang, Y.: Unified mask embedding and correspondence learning for self-supervised video segmentation. In: CVPR (2023) 3
55. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV (2020) 4, 11
56. Liu, D., Cui, Y., Tan, W., Chen, Y.: Sg-net: Spatial granularity network for one-stage video instance segmentation. In: CVPR (2021) 4
57. Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Crossover learning for fast online video instance segmentation. In: ICCV (2021) 4
58. Han, S.H., Hwang, S., Oh, S.W., Park, Y., Kim, H., Kim, M.J., Kim, S.J.: Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In: CVPR (2022) 4
59. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: ICCV (2021) 4
60. Zhu, F., Yang, Z., Yu, X., Yang, Y., Wei, Y.: Instance as identity: A generic online paradigm for video instance segmentation. In: ECCV (2022) 4
61. Li, M., Li, S., Li, L., Zhang, L.: Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In: CVPR (2021) 4

62. Ke, L., Li, X., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Prototypical cross-attention networks for multiple object tracking and segmentation. In: NeurIPS (2021) 4
63. Lin, H., Wu, R., Liu, S., Lu, J., Jia, J.: Video instance segmentation with a propose-reduce paradigm. In: ICCV (2021) 4
64. Koner, R., Hannan, T., Shit, S., Sharifzadeh, S., Schubert, M., Seidl, T., Tresp, V.: Instanceformer: An online video instance segmentation framework. In: AAAI (2023) 4
65. Liu, Q., Wu, J., Jiang, Y., Bai, X., Yuille, A.L., Bai, S.: Instmove: Instance motion for object-centric video segmentation. In: CVPR (2023) 4
66. Li, M., Li, S., Xiang, W., Zhang, L.: Mdqe: Mining discriminative query embeddings to segment occluded instances on challenging videos. In: CVPR (2023) 4, 11
67. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV (2020) 4
68. Wu, J., Yarram, S., Liang, H., Lan, T., Yuan, J., Eledath, J., Medioni, G.: Efficient video instance segmentation via tracklet query and proposal. In: CVPR (2022) 4
69. Yang, S., Wang, X., Li, Y., Fang, Y., Fang, J., Liu, W., Zhao, X., Shan, Y.: Temporally efficient vision transformer for video instance segmentation. In: CVPR (2022) 4
70. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR (2020) 4
71. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. In: NeurIPS (2021) 4
72. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: ICCV (2021) 4
73. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: CVPR (2022) 4
74. Li, L., Wang, W., Yang, Y.: Logicseg: Parsing visual semantics with neural logic learning and reasoning. In: ICCV (2023) 4
75. Chen, M., Zheng, Z., Yang, Y., Chua, T.S.: Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In: ACM MM (2023) 4
76. Li, L., Zhou, T., Wang, W., Li, J., Yang, Y.: Deep hierarchical semantic segmentation. In: CVPR (2022) 4
77. Li, L., Wang, W., Zhou, T., Quan, R., Yang, Y.: Semantic hierarchy-aware segmentation. IEEE TPAMI (2023) 4
78. Chen, M., Zheng, Z., Yang, Y.: Transferring to real-world layouts: A depth-aware framework for scene adaptation. arXiv preprint arXiv:2311.12682 (2023) 4
79. Zhou, T., Wang, W.: Cross-image pixel contrasting for semantic segmentation. IEEE TPAMI (2024) 4
80. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: CVPR (2018) 4
81. Mahasseni, B., Todorovic, S., Fern, A.: Budget-aware deep semantic video segmentation. In: CVPR (2017) 4
82. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: CVPR (2018) 4
83. Jain, S., Wang, X., Gonzalez, J.E.: Accel: A corrective fusion network for efficient semantic segmentation on video. In: CVPR (2019) 4
84. Liu, Y., Shen, C., Yu, C., Wang, J.: Efficient semantic video segmentation with per-frame inference. In: ECCV (2020) 4

85. Li, J., Wang, W., Chen, J., Niu, L., Si, J., Qian, C., Zhang, L.: Video semantic segmentation via sparse temporal transformer. In: ACM MM (2021) 4
86. Sun, G., Liu, Y., Tang, H., Chhatkuli, A., Zhang, L., Van Gool, L.: Mining relations among cross-frame affinities for video semantic segmentation. In: ECCV (2022) 4, 10, 11
87. Qiao, S., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In: CVPR (2021) 4
88. Kreuzberg, L., Zulfikar, I.E., Mahadevan, S., Engelmann, F., Leibe, B.: 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In: ECCV (2022) 4
89. Zhou, Y., Zhang, H., Lee, H., Sun, S., Li, P., Zhu, Y., Yoo, B., Qi, X., Han, J.J.: Slot-vps: Object-centric representation learning for video panoptic segmentation. In: CVPR (2022) 4
90. Yuan, H., Li, X., Yang, Y., Cheng, G., Zhang, J., Tong, Y., Zhang, L., Tao, D.: Polyphonicformer: unified query learning for depth-aware video panoptic segmentation. In: ECCV (2022) 4, 7
91. He, J., Wang, Y., Wang, L., Lu, H., Luo, B., He, J.Y., Lan, J.P., Geng, Y., Xie, X.: Towards deeply unified depth-aware panoptic segmentation with bi-directional guidance learning. In: ICCV (2023) 4
92. Shin, I., Kim, D., Yu, Q., Xie, J., Kim, H.S., Green, B., Kweon, I.S., Yoon, K.J., Chen, L.C.: Video-kmax: A simple unified approach for online and near-online video panoptic segmentation. arXiv preprint arXiv:2304.04694 (2023) 4
93. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022) 4, 5, 10
94. Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., Wan, P.: Dvis: Decoupled video instance segmentation framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1282–1291 (2023) 4, 10, 11
95. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 5
96. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS (2021) 5
97. Wang, W., Liang, J.C., Liu, D.: Learning equivariant segmentation with instance-unique querying. In: NeurIPS (2022) 5
98. Ding, Y., Li, L., Wang, W., Yang, Y.: Clustering propagation for universal medical image segmentation. In: CVPR (2024) 5
99. Liang, J.C., Zhou, T., Liu, D., Wang, W.: Clustseg: Clustering for universal segmentation. In: ICML (2023) 5
100. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE TPAMI **24**(4), 509–522 (2002) 6
101. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 7
102. Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., Lu, T.: Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In: CVPR (2022) 7
103. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: CVPR (2020) 8
104. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: NeurIPS (2020) 8
105. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: ICCV (2023) 10

106. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10, 14
107. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 10
108. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021) 10
109. Li, J., Yu, B., Rao, Y., Zhou, J., Lu, J.: Tcovis: Temporally consistent online video instance segmentation. In: ICCV (2023) 11
110. Ying, K., Zhong, Q., Mao, W., Wang, Z., Chen, H., Wu, L.Y., Liu, Y., Fan, C., Zhuge, Y., Shen, C.: Ctvis: Consistent training for online video instance segmentation. In: ICCV (2023) 11, 12
111. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 12, 13
112. Hoffhues, A., Luiten, J.: Trackeval (2020), https://github.com/JonathonLuiten/TrackEval 12
113. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: CVPR (2023) 12