Transformer-Based Visual Segmentation: A Survey

Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy, *Senior Member, IEEE*

(Survey Paper)

***Abstract*—Visual segmentation seeks to partition images, video frames, or point clouds into multiple segments or groups. This technique has numerous real-world applications, such as autonomous driving, image editing, robot sensing, and medical analysis. Over the past decade, deep learning-based methods have made remarkable strides in this area. Recently, transformers, a type of neural network based on self-attention originally designed for natural language processing, have considerably surpassed previous convolutional or recurrent approaches in various vision processing tasks. Specifically, vision transformers offer robust, unified, and even simpler solutions for various segmentation tasks. This survey provides a thorough overview of transformer-based visual segmentation, summarizing recent advancements. We first review the background, encompassing problem definitions, datasets, and prior convolutional methods. Next, we summarize a meta-architecture that unifies all recent transformer-based approaches. Based on this meta-architecture, we examine various method designs, including modifications to the meta-architecture and associated applications. We also present several specific subfields, including 3D point cloud segmentation, foundation model tuning, domain-aware segmentation, efficient segmentation, and medical segmentation. Additionally, we compile and re-evaluate the reviewed methods on several well-established datasets. Finally, we identify open challenges in this field and propose directions for future research.**

***Index Terms*—Vision transformer review, dense prediction, image segmentation, video segmentation, scene understanding.**

Manuscript received 2 June 2023; revised 17 April 2024; accepted 22 July 2024. Date of publication 29 July 2024; date of current version 5 November 2024. This work was supported in part by The Alan Turing Institute (U.K.) through the project 'Turing-DSO Labs Singapore Collaboration' under Grant SDCfP2\100009. This study was also supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative and Singapore MOE AcRF Tier 1 under Grant RG16/21, as well as cash and in-kind contributions from the industry partner(s). Recommended for acceptance by N. Sebe. *(Corresponding author: Guangliang Cheng.)*

Xiangtai Li, Haobo Yuan, Wenwei Zhang, Ziwei Liu, and Chen Change Loy are with S-Lab, Nanyang Technological University, Singapore 639798 (e-mail: xiangtai94@gmail.com; wenwei.zhang@ntu.edu.sg; ccloy@ntu.edu.sg).

Henghui Ding is with the Institute of Big Data, Fudan University, Shanghai 200437, China (e-mail: henghui.ding@gmail.com).

Jiangmiao Pang and Kai Chen are with the Shanghai AI Laboratory, Shanghai 200240, China (e-mail: pangjiangmiao@gmail.com; chenkai@pjlab.org.cn).

Guangliang Cheng is with the University of Liverpool, L69 7ZX Liverpool, U.K. (e-mail: Guangliang.Cheng@liverpool.ac.uk).

The project page can be found at https://github.com/lxtGH/Awesome-Segmentation-With-Transformer.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3434373, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3434373

I. Introduction

VISUAL segmentation aims to group pixels of the given image or video into a set of semantic regions. It is a fundamental problem in computer vision and involves numerous real-world applications, such as robotics, automated surveillance, image/video editing, social media, autonomous driving, *etc*. Starting from the hand-crafted features [1], [2] and classical machine learning models [3], [4], [5], segmentation problems have been involved with a lot of research efforts. During the last ten years, deep neural networks, Convolution Neural Networks (CNNs) [6], [7], [8], such as Fully Convolutional Networks (FCNs) [9], [10], [11], [12] have achieved remarkable successes for different segmentation tasks and led to much better results. Compared to traditional segmentation approaches, CNNs based approaches have better generalization ability. Because of their exceptional performance, CNNs and FCN architecture have been the basic components in the segmentation research works.

Recently, with the success of natural language processing (NLP), transformer [13] is introduced as a replacement for recurrent neural networks [14]. Transformer contains a novel self-attention design and can process various tokens in parallel. Then, based on transformer design, BERT [15] and GPT-3 [16] scale the model parameters up and pre-train with huge unlabeled text information. They achieve strong performance on many NLP tasks, accelerating the development of transformers into the vision community. Recently, researchers applied transformers to computer vision (CV) tasks. Early methods [17], [18] combine the self-attention layers to augment CNNs. Meanwhile, several works [19], [20] used pure self-attention layers to replace convolution layers. After that, two remarkable methods boost the CV tasks. One is *vision transformer (ViT)* [21], which is a pure transformer that directly takes the sequences of image patches to classify the full image. It achieves state-of-the-art performance on multiple image recognition datasets. Another is *detection transformer (DETR)* [22], which introduces the concept of object query. Each object query represents one instance. The object query replaces the complex anchor design in the previous detection framework, which simplifies the pipeline of detection and segmentation. Then, the following works adopt improved designs on various vision tasks, including representation learning [23], [24], object detection [25], segmentation [26], low-level image processing [27], video understanding [28], 3D scene understanding [29], and image/video generation [30].

As for visual segmentation, recent state-of-the-art methods are all based on transformer architecture. Compared with CNN

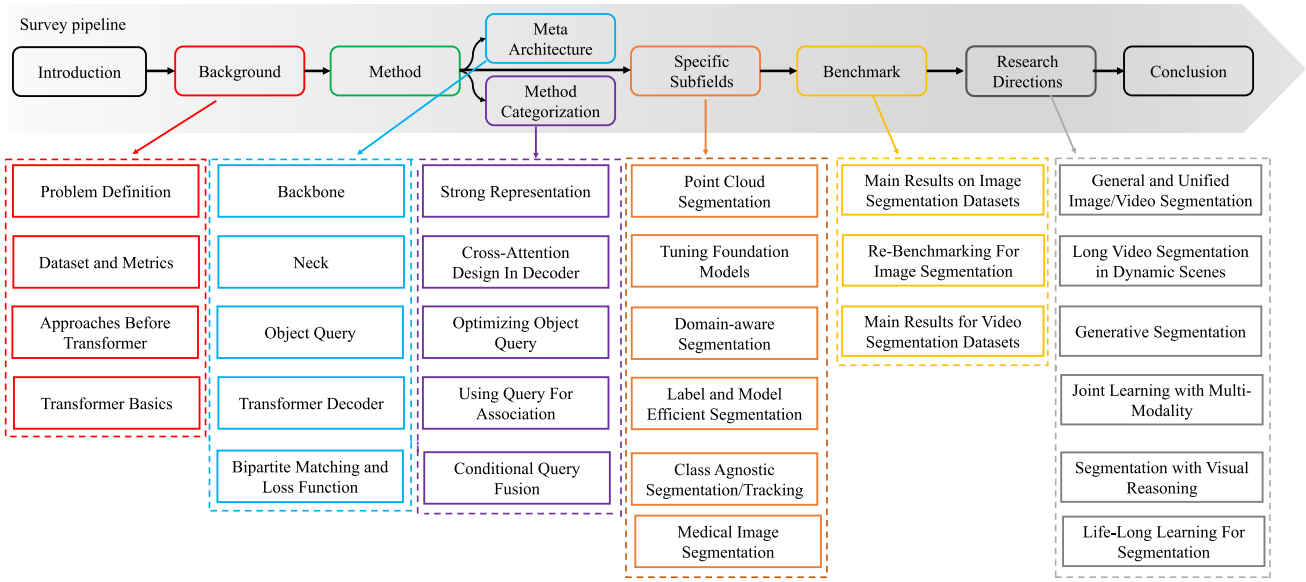

Fig. 1. A diagram that summarizes this survey. Different colors represent specific sections. Best viewed in color.

approaches, most transformer-based approaches have simpler pipelines but stronger performance. Because of a rapid upsurge in transformer-based vision models, there are several surveys on vision transformer [31], [32], [33]. However, most of them mainly focus on general transformer design and its application on several specific vision tasks [34], [35], [36]. Meanwhile, there are previous surveys on the deep-learning-based segmentation [37], [38], [39]. However, to the best of our knowledge, there are *no surveys* focusing on using vision transformers for visual segmentation or query-based object detection. We believe it would be beneficial for the community to summarize these works and keep tracking this evolving field.

• *Contribution:* In this survey, we systematically introduce recent advances in transformer-based visual segmentation methods. We start by defining the task, datasets, and CNN-based approaches and then move on to transformer-based approaches, covering existing methods and future work directions. Our survey groups existing representative works from a more technical perspective of the method details. In particular, for the main review part, we first summarize the core framework of existing approaches into a meta-architecture in Section III-A, which is an extension of DETR [22]. By changing the components of the meta-architecture, we divide existing approaches into six categories in Section III-B, including Representation Learning, Interaction Design in Decoder, Optimizing Object Query, Using Query For Association, and Conditional Query Generation.

Moreover, we also survey closely related specific subfields, including point cloud segmentation, tuning foundation models, domain-aware segmentation, data/model efficient segmentation, class agnostic segmentation and tracking, and medical segmentation. We also evaluate the performance of influential works published in top-tier conferences and journals on several widely used segmentation benchmarks. Additionally, we provide an overview of previous CNN-based models and relevant literature in other areas, such as object detection, object tracking, and referring segmentation in the background section.

• *Scope:* This survey will cover several mainstream segmentation tasks, including semantic segmentation, instance segmentation, panoptic segmentation, and their variants, such as video and point cloud segmentation. Additionally, we cover related subfields in Section IV. We focus on transformer-based approaches and only review a few closely related CNN-based approaches for reference. Although there are many preprints or published works, we only include the most representative works.

• *Organization:* The rest of the survey is organized as follows. Overall, Fig. 1 shows the pipeline of our survey. We first introduce the background knowledge on problem definition, datasets, and CNN-based approaches in Section II. Then, we review representative papers on transformer-based segmentation methods in Sections III and IV. We compare the experiment results in Section V. Finally, we raise the future directions in Section VI and conclude the survey in Section VII. We provide more benchmarks and details in the appendix, available online.

II. Background

In this section, we first present a unified problem definition of different segmentation tasks. Then, we detail the common datasets and evaluation metrics. Next, we present a summary of previous approaches before the transformer. Finally, we present a review of basic concepts in transformers. To facilitate understanding of this survey, we list the brief notations in Table I for reference.

A. Problem Definition

• *Image Segmentation:* Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the goal of image segmentation is to output a group of masks $\{y_i\}_{i=1}^{G} = \{(m_i, c_i)\}_{i=1}^{G}$ where c_i denotes the ground truth class label of the binary mask m_i and G is the number of masks, $H \times W$ are the spatial size. According to the scope

TABLE I
NOTATION AND ABBREVIATIONS USED IN THIS SURVEY

Notations	Descriptions
SS / IS /PS	Semantic Segmentation / Instance Segmentation / Panoptic Segmentation
VSS / VIS / VPS	Video Semantic / Instance / Panoptic Segmentation
DVPS	Depth-aware Panoptic Segmentation
PPS	Part-aware Panoptic Segmentation
PCSS / PCIS / PCPS	Point Cloud Semantic /Instance / Panoptic Segmentation
RIS / RVOS	Referring Image Segmentation / Referring Video Object Segmentation
VLM	Vision Language Model
VOS / (V)OD	Vido Object Segmentation / (Video) Object Detection
MOTS	Multi-Object Tracking, and Segmentation
CNN / ViTs	Convolution Neural Network / Vision Transformer
SA / MHSA / MLP	Self-Attention / Multi-Head Self Attention / Multi-Layer Perceptron
(Deformable) DETR	(Deformable) DEtection TRansformer
mIoU	mean Intersection over Union (SS, VSS)
PQ/VPQ	Panoptic Quality / Video Panoptic Quality (PS, VPS)
mAP	mean Average Precision (IS, VIS)
STQ	Segmentation and Tracking Quality (VPS)

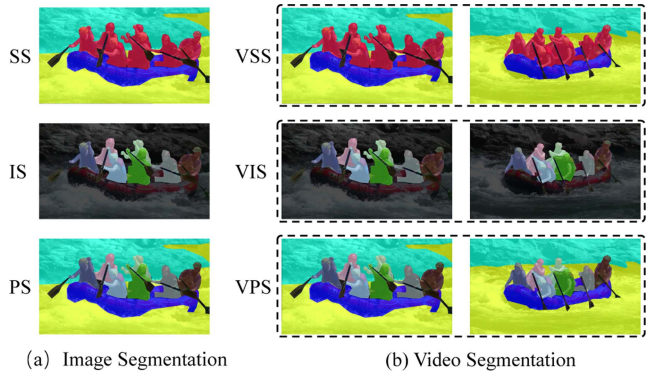

Fig. 2. Illustration of different segmentation tasks. The examples are sampled from the VIP-Seg dataset [40]. For (V)SS, the same color indicates the same class. For (V)IS and (V)PS, different instances are represented by different colors.

of class labels and masks, image segmentation can be divided into three different tasks, including semantic segmentation (SS), instance segmentation (IS), and panoptic segmentation (PS), as shown in Fig. 2(a). For SS, the classes may be foreground objects (thing) or background (stuff), and each class only has one binary mask that indicates the pixels belonging to this class. Each SS mask does not overlap with other masks. For IS, each class may have more than one binary mask, and all the classes are foreground objects. Some IS masks may overlap with others. For PS, depending on the class definition, each class may have a different number of masks. For the countable thing class, each class may have multiple masks for different instances. For the uncountable stuff class, each class only has one mask. Each PS mask does not overlap with other masks. One can understand image segmentation from the pixel view. Given an input $I \in \mathbb{R}^{H\times W\times 3}$, the output of image segmentation is a two-channel dense segmentation map $S = \{k_j, c_j\}_{j=1}^{H\times W}$. In particular, k indicates the identity of the pixel j, and c means the class label of pixel j. For SS, the identities of all pixels are zero. For IS, each instance has a unique identity. For PS, the pixels belonging to the thing classes have a unique identity. The pixel identities of the stuff class are zero. From both two perspectives, the PS unifies both SS and IS. We present the visual examples in Fig. 2.

• *Video Segmentation:* Given a video clip input as $V \in \mathbb{R}^{T\times H\times W\times 3}$, where T represents the frame number, the goal of video segmentation is to obtain a mask tube $\{y_i\}_{i=1}^{N} = \{(m_i, c_i)\}_{i=1}^{N}$, where N is the number of the tube masks $m_i \in \{0,1\}^{T\times H\times W}$, and c_i denotes the class label of the tube m_i. Video panoptic segmentation (VPS) requires temporally consistent segmentation and tracking results for each pixel. Each tube mask can be classified into countable thing classes and countless stuff classes. Each thing tube mask also has a unique ID for evaluating tracking performance. For stuff masks, the tracking is zero by default. When $N = C$ and the task only contains stuff classes, and all thing classes have no IDs, VPS turns into video semantic segmentation (VSS). If $\{y_i\}_{i=1}^{N}$ overlap and C only contains the thing classes and all stuff classes are ignored, VPS turns into video instance segmentation (VIS). We present the visual examples that summarize the difference among VPS, VIS, and VSS with $T = 2$ in Fig. 2(b).

• *Related Problems:* Object detection and instance-wise segmentation (IS/VIS/VPS) are closely related tasks. Object detection involves predicting object bounding boxes, which can be considered a coarse form of IS. After introducing the DETR model, many works have treated object detection and IS as the same task, as IS can be achieved by adding a simple mask prediction head to object detection. Similarly, video object detection (VOD) aims to detect objects in every video frame. In our survey, we also examine query-based object detectors for both object detection and VOD. Point cloud segmentation is another segmentation task, where the goal is to segment each point in a point cloud into pre-defined categories. We can apply the same definitions of semantic segmentation, instance segmentation, and panoptic segmentation to this task, resulting in point cloud semantic segmentation (PCSS), point cloud instance segmentation (PCIS), and point cloud panoptic segmentation (PCPS). Referring segmentation is a task that aims to segment objects described in natural language text input. There are two subtasks in referring segmentation: referring image segmentation (RIS), which performs language-driven segmentation, and referring video object segmentation (RVOS), which segments and tracks a specific object in a video based on required text inputs. Finally, video object segmentation (VOS) involves tracking an object in a video by predicting pixel-wise masks in every frame, given a mask of the object in the first frame.

B. Datasets and Metrics

• *Commonly Used Datasets:* For image segmentation, the most commonly used datasets are COCO [43], ADE20k [44] and Cityscapes [45]. For video segmentation, the most used datasets are VSPW [49] and Youtube-VIS [50]. We will compare several dataset results in Section V. More datasets are listed in the Table II.

• *Common Metric:* For SS and VSS, the commonly used metric is mean intersection over union (mIoU), which calculates the pixel-wised Union of Interest between output image and video masks and ground truth masks. For IS, the metric is mask mean average precision (mAP), which is extended from the object detection via replacing box IoU with mask IoU. For VIS,

TABLE II
COMMONLY USED DATASETS AND METRIC FOR TRANSFORMER-BASED SEGMENTATION

Dataset	Samples (train/val)	Task	Evaluation Metrics	Characterization
Pascal VOC [41]	1,464 / 1,449	SS	mIoU	PASCAL Visual Object Classes (VOC) 2012 dataset contains 20 object categories.
Pascal Context [42]	4,998 / 5,105	SS	mIoU	PASCAL Context dataset is an extension of PASCAL VOC containing 400+ classes (usually 59 most frequently).
COCO [43]	118k / 5k	SS / IS / PS	mIoU / mAP / PQ	MS COCO dataset is a large-scale dataset with 80 thing categories and 91 stuff categories.
ADE20k [44]	20,210 / 2,000	SS / IS / PS	mIoU / mAP / PQ	ADE20k dataset is a large-scale dataset exhaustively annotated with pixel-level objects and object part labels.
Cityscapes [45]	2,975 / 500	SS / IS / PS	mIoU / mAP / PQ	Cityscapes dataset focuses on semantic understanding of urban street scenes, captured in 50 cities.
Mapillary [46]	18k / 2k	SS / PS	mIoU / PQ	Mapillary dataset is a large-scale dataset with accurate high-resolution annotations.
RefCOCO [47]	42k / 4k	RIS	mIoU	A large-scale dataset for classic reference segmentation based on the COCO.
gRefCOCO [48]	79k / 8k	RIS	mIoU	A large-scale dataset for generalized referring segmentation based on the COCO.
VSPW [49]	2,906 / 343	VSS	mIoU	VPSW is a large-scale high-resolution dataset with long videos focusing on VSS.
Youtube-VIS-2019 [50]	2,238 / 302	VIS	AP	Extending from Youtube-VOS, Youtube-VIS is with exhaustive instance labels.
OVIS [51]	607 / 140	VIS	AP	A large-scale occluded video instance segmentation benchmark.
VIP-Seg [40]	2,806 / 343	VPS	VPQ & STQ	Extending from VSPW, VIP-Seg adds extra instance labels for VPS task.
Cityscape-VPS [52]	2,400 / 300	VPS	VPQ	Cityscapes-VPS dataset extracts from the val split of Cityscapes dataset, adding temporal annotations.
KITTI-STEP [53]	5,027 / 2,981	VPS	STQ	KITTI-STEP focuses on the long videos in the urban scenes.
DAVIS-2017 [54]	4,219 / 2,023	VOS	J / F / J&F	DAVIS focuses on video object segmentation.
Youtube-VOS [55]	3,471 / 474	VOS	J / F / J&F	A large-scale video object segmentation benchmark.
MOSE [56]	1,507 / 311	VOS	J / F / J&F	Tracking and segmenting objects in complex environments.
MeViS [57]	1,712 / 140	RVOS	J / F / J&F	Tracking and segmenting target objects referred by motion expressions.

the metric is 3D mAP, which extends mask mAP in a spatial-temporal manner. For PS, the metric is the panoptic quality (PQ), which unifies both thing and stuff prediction by setting a fixed threshold 0.5. For VPS, the commonly used metrics are video panoptic quality (VPQ) and segmentation tracking quality (STQ). The former extends PQ into temporal window calculation, while the latter decouples the segmentation and tracking in a per-pixel-wised manner. Note that there are other metrics, including pixel accuracy and temporal consistency. For simplicity, we only report the primary metrics used in the literature. We present the detailed formulation of these metrics in the supplementary material.

C. Segmentation Approaches Before Transformer

• *Semantic Segmentation:* Prior to the emergence of ViT and DETR. SS was typically approached as a dense pixel classification problem, as initially proposed by FCN. Then, the following works are all based on the FCN framework. These methods can be divided into the following aspects, including better encoder-decoder frameworks [58], [59], larger kernels [60], [61], multiscale pooling [11], [62], multiscale feature fusion [12], [63], [64], [65], non-local modeling [18], [66], [67], efficient modeling [68], [69], [70], and better boundary delineation [71], [72], [73], [74]. After the transformer was proposed, with the goal of global context modeling, several works design variants of self-attention operators to replace the CNN prediction heads [66], [75].

• *Instance Segmentation:* IS aims to detect and segment each object, which goes beyond object detection. Most IS approaches focus on how to represent instance masks beyond object detection, which can be divided into two categories: top-down approaches [76], [77] and bottom-up approaches [78], [79]. The former extends the object detector with an extra mask head. The designs of mask heads are various, including FCN heads [76], [80], diverse mask encodings [81], and dynamic kernels [77], [82]. The latter performs instance clustering from semantic segmentation maps to form instance masks. The performance of top-down approaches is closely related to the choice of detector [83], while bottom-up approaches depend on both semantic segmentation results and clustering methods [84]. Besides, there are also several approaches [85], [86] using gird representation to learn instance masks directly. The ideas using kernels and different mask encodings are also extended into several transformer-based approaches, which will be detailed in Section III.

• *Panoptic Segmentation:* Previous works for PS mainly focus on how to fuse the results of both SS and IS, which treats PS as two independent tasks. Based on IS subtask, the previous works can also be divided into two categories: top-down approaches [87], [88] and bottom-up approaches [84], [89], according to the way to generate instance masks. Several works use a shared backbone with multitask heads to jointly learn IS and SS, focusing on mutual task association. Meanwhile, several bottom-up approaches [84], [89] use the sequential pipeline by performing instance clustering from semantic segmentation results and then fusing both. In summary, most PS methods include complex pipelines and are highly engineered.

• *Video Segmentation:* The research for VSS mainly focuses on better spatial-temporal fusion [90] or acceleration using extra cues [91], [92] in the video. VIS requires segmenting and tracking each instance. Most VIS approaches [52], [93], [94], [95] focus on learning instance-wised spatial, temporal relation, and feature fusion. Several works learn the 3D temporal embeddings. Like PS, VPS [52] can also be top-down [52] and bottom-up approaches [96]. The top-down approaches learn to link the temporal features and then perform instance association online. In contrast, the bottom-up approaches predict the center map of the near frame and perform instance association in a separate stage. Most of these approaches are highly engineering. For example, MaskPro [93] adopts state-of-the-art IS segmentation models [80], deformable CNN [97], and offline mask propagation in one system. There are also several video segmentation tasks, including video object segmentation (VOS) [56], [98], referring video segmentation [57], multi-Object tracking, and segmentation (MOTS) [99].

• *Point Cloud Segmentation:* This task aims to group point clouds into semantic or instance categories, similar to image and video segmentation. Depending on the input scene, it is typically categorized as either indoor or outdoor scenes. Indoor scene segmentation mainly includes

point cloud semantic segmentation (PSS) and point cloud instance segmentation (PIS). PSS is commonly achieved using the Point-Net [100], [101], while PIS can be achieved through two approaches: top-down approaches [102], [103] and bottom-up approaches [104], [105]. The former extracts 3D bounding boxes and uses a mask learning branch to predict masks, while the latter predicts semantic labels and utilizes point embedding to group points into different instances. For outdoor scenes, point cloud segmentation can be divided into point-based [100], [106] and voxel-based [107], [108] approaches. Point-based methods focus on processing individual points, while voxel-based methods divide the point cloud into 3D grids and apply 3D convolution. Like panoptic segmentation, most 3D panoptic segmentation methods [109], [110], [111], [112], [113] first predict semantic segmentation results, separate instances based on these predictions and fuse the two results to obtain the final results.

D. Transformer Basics

• *Vanilla Transformer* [13] is a seminal model in the transformer-based research field. It is an encoder-decoder structure that takes tokenized inputs and consists of stacked transformer blocks. Each block has two sub-layers: a multi-head self-attention (MHSA) layer and a position-wise fully-connected feed-forward network (FFN). The MHSA layer allows the model to attend to different parts of the input sequence while the FFN processes the output of the MHSA layer. Both sub-layers use residual connections and layer normalization for better optimization.

In the vanilla transformer, the encoder and decoder both use the same architecture. However, the decoder is modified to include a mask that prevents it from attending to future tokens during training. Additionally, the decoder uses sine and cosine functions to produce positional embeddings, which allow the model to understand the order of the input sequence. Subsequent models such as BERT and GPT-2 have built upon its architecture and achieved state-of-the-art results on a wide range of natural language processing tasks.

• *Self-Attention:* The core operator of the vanilla transformer is the self-attention (SA) operation. Suppose the input data is a set of tokens $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{N\times c}$. N is the token number and c is token dimension. The positional encoding P may be added into $I = X + P$. The input embedding I goes through three linear projection layers ($W^q \in \mathbb{R}^{c\times d}, W^k \in \mathbb{R}^{c\times d}, W^v \in \mathbb{R}^{c\times d}$) to generate Query (Q), Key (K), and Value (V):

$$Q = IW^q, K = IW^k, V = IW^v, \tag{1}$$

where d is the hidden dimension. The Query and Key are usually used to generate the attention map in SA. Then the SA is performed as follows:

$$O = \text{SA}(Q, K, V) = \text{Softmax}(QK^\intercal)V. \tag{2}$$

According to (2), given an input X, self-attention allows each token x_i to attend to all the other tokens. Thus, it has the ability of global perception compared with local CNN operator. Motivated by this, several works [18], [114] treat it as a fully-connected graph or a non-local module for visual recognition task.

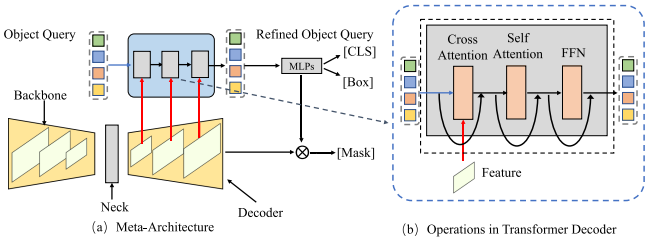

Fig. 3. Illustration of (a) meta-architecture and (b) common operations in the decoder.

• *Multi-Head Self-Attention:* In practice, multi-head self-attention (MHSA) is more commonly used. The idea of MHSA is to stack multiple SA sub-layer in parallel, and the concatenated outputs are fused by a projection matrix $W^{fuse} \in \mathbb{R}^{d\times c}$

$$O = \text{MHSA}(Q, K, V) = \text{concat}([\text{SA}_\text{i}, ..\text{SA}_\text{H}])W^{fuse}, \tag{3}$$

where $\text{SA}_\text{i} = \text{SA}(Q_i, K_i, V_i)$ and H is the number of the head. Different heads have individual parameters. Thus, MHSA can be viewed as an ensemble of SA.

• *Feed-Forward Network:* The goal of feed-forward network (FFN) is to enhance the non-linearity of attention layer outputs. It is also called multi-layer perceptron (MLP) since it consists of two successive linear layers with non-linear activation layers.

III. Methods: A Survey

In this section, based on DETR-like meta-architecture, we review the key techniques of transformer-based segmentation. As shown in Fig. 3, the meta-architecture contains a feature extractor, object query, and a transformer decoder. Then, according to the meta-architecture, we survey existing methods by considering the modification or improvements to each component of the meta-architecture in Sections III-B1, III-B2 and III-B3. Finally, based on such meta-architecture, we present several detailed applications in Sections III-B4 and III-B5.

A. Meta-Architecture

• *Backbone:* Before ViTs, CNNs were the standard approach for feature extraction in computer vision tasks. To ensure a fair comparison, many research works [22], [76], [115] used the same CNN models, such as ResNet50 [7]. Some researchers [18], [89] also explored the combination of CNNs with self-attention layers to model long-range dependencies. ViT, on the other hand, utilizes a standard transformer encoder for feature extraction. It has a specific input pipeline for images, where the input image is split into fixed-size patches, such as 16 $\times$ 16 patches. These patches are then processed through a linear embedding layer. Then, the positional embeddings are added to each patch. Afterward, a standard transformer encoder encodes all patches. It contains multiple multi-head self-attention and feed-forward layers. For instance, given an image $I \in \mathbb{R}^{H\times W\times 3}$, ViT first reshapes it into a sequence of flattened 2D patches: $I_p \in \mathbb{R}^{N\times P^2\times 3}$, where N is the number of patches and P is

the patch size. With patch embedding operations, the final input is $I_{in} \in \mathbb{R}^{N \times P^2 \times C}$, where C is the embedding channel. To perform classification, an extra learnable embedding "classification token" (CLS) is added to the sequence of embedded patches. After the standard transformer for all patches, $I_{out} \in \mathbb{R}^{N \times P^2 \times C}$ is obtained. For segmentation tasks, ViT is used as a feature extractor, meaning that I_{out} is resized back to a dense map $F \in \mathbb{R}^{H \times W \times C}$.

• *Neck:* Feature pyramid network (FPN) has been shown effective in object detection and instance segmentation [116], [117], [118] for scale variation modeling. FPN maps the features from different stages into the same channel dimension C for the decoder. Several works [83], [119] design stronger FPNs via cross-scale modeling using dilation or deformable convolution. For example, Deformable DETR [25] proposes a deformable FPN to model cross-scale fusion using deformable attention. Lite-DETR [120] further refines the deformable cross-scale attention design by efficiently sampling high-level features and low-level features in an interleaved manner. The output features are used for decoding the boxes and masks. The role of FPN is the same as previous detection-based or FCN-based segmentation methods. The FPN generates multi-scale features to handle and balance both small and large objects in the scene. For the transformer-based method, FPN architecture is often used to refine object queries from different scales, which can lead to stronger results than single-scale refinement.

• *Object Query:* Object query is first introduced in DETR [22]. It plays as the dynamic anchors that are used in detectors [76], [115]. In practice, it is a learnable embedding $Q_{obj} \in \mathbb{R}^{N_{ins} \times d}$. N_{ins} represents the maximum instance number. The query dimension d is usually the same as feature channel c. Object query is refined by the cross-attention layers. Each object query represents one instance of the image. During the training, each ground truth is assigned with one corresponding query for learning. During the inference, the queries with high scores are selected as output. Thus, object query simplifies the design of detection and segmentation models by eliminating the need for hand-crafted components such as non-maximum suppression (NMS). The flexible design of object query has led to many research works exploring its usage in different contexts, which will be discussed in more detail in Section III-B.

• *Transformer Decoder:* Transformer decoder is a crucial architecture component in transformer-based segmentation and detection models. Its main operation is cross-attention, which takes in the object query Q_{obj} and the image/video feature F. It outputs a refined object query, denoted as Q_{out}. The cross-attention operation is derived from the vanilla transformer architecture, where Q_{obj} serves as the query, and F is used as the key and value in the self-attention mechanism. After obtaining the refined object query Q_{out}, it is passed through a prediction FFN, which typically consists of a 3-layer perceptron with a ReLU activation layer and a linear projection layer. The FFN outputs the final prediction, which depends on the specific task. For example, for classification, the refined query is mapped directly to class prediction via a linear layer. For detection, the FFN predicts the normalized center coordinates, height, and width of the object bounding box. For segmentation, the output embedding is used to perform dot product with feature F, which results in the binary mask logits. The transformer decoder iteratively repeats cross-attention and FFN operations to refine the object query and obtain the final prediction. The intermediate predictions are used for auxiliary losses during training and discarded during inference. The outputs from the last stage of the decoder are taken as the final detection or segmentation results. We show the detailed process in Fig. 3(b).

• *Mask Prediction Representation:* Transformer-based segmentation approaches adopt two formats to represent the mask prediction: pixel-wise prediction as FCNs and per-mask-wise prediction as DETR. The former is used in semantic-aware segmentation tasks, including SS, VSS, VOS, and *etc*. The latter is used in instance-aware segmentation tasks, including IS, VIS, and VPS, where each query represents each instance.

• *Bipartite Matching and Loss Function:* Object query is usually combined with bipartite matching [121] during training, uniquely assigning predictions with ground truth. This means each object query builds the one-to-one matching during training. Such matching is based on the matching cost between ground truth and predictions. The matching cost is defined as the distance between prediction and ground truth, including labels, boxes, and masks. By minimizing the cost with the Hungarian algorithm [121], each object query is assigned by its corresponding ground truth. For object detection, each object query is trained with classification and box regression loss [115]. For instance-aware segmentation, each object query is trained via both mask classification loss and segmentation loss. The output masks are obtained via the inner product between object query and decoder features. The segmentation loss usually contains binary cross-entropy loss and dice loss [122].

• *Discussion on Scope of Meta-Architecture:* We admit our meta-architecture may **not** cover all transformer-based segmentation methods. In semantic segmentation, methods such as Segformer [123] and SETR [124] employ a fully connected layer and predict each pixel as previous FCN-based methods [9], [62], [125]. These methods concentrate on enhanced feature representation. We argue that this represents a basic form of our meta-architecture, wherein each query corresponds to a class category. The cascaded cross-attention layers are omitted, and bipartite matching is removed. Thus, the object query plays the same role as a fully connected layer. In addition, meta-architecture represents the latest design philosophy. Nearly all recent state-of-the-art methods [126], [127], [128], [129] adopt this meta-architecture. In particular, different methods may add more components to adapt their tasks and requirements. Thus, we review recent works by modifying each component based on this meta-architecture.

B. Method Categorization

In this section, we review five aspects of transformer-based segmentation methods. Rather than classifying the literature by the task settings, our goal is to extract the essential and common techniques used in the literature. We summarize the methods, techniques, related tasks, and corresponding references in Table III. Most approaches are based on the meta-architecture

TABLE III
TRANSFORMER-BASED SEGMENTATION METHOD CATEGORIZATION

Method Categorization	Tasks	Reference
Representation Learning (Sec. 3.2.1)		
•Better ViTs Design	SS / IS	[21], [130], [131], [132], [133], [134], [135], [136], [137], [138]
•Hybrid CNNs / transformers / MLPs	SS / IS	[23], [123], [139], [140], [141], [142], [143], [144], [145], [146], [147]
•Self-Supervised Learning	SS / IS	[24], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157]
Cross-Attention Design in Decoder (Sec. 3.2.2)		
•Improved Cross-Attention Design	SS / IS / PS	[25], [158], [159], [160], [161], [162], [163], [164], [165]
•Spatial-Temporal Cross-Attention Design	VSS / VIS / VPS	[166], [167], [168], [169], [170], [171], [172], [173]
Optimizing Object Query (Sec. 3.2.3)		
•Adding Position Information into Query	IS / PS	[174], [175], [176], [177]
•Adding Extra Supervision into Query.	IS / PS	[178], [179], [180], [181], [182], [183], [184]
Using Query For Association (Sec. 3.2.4)		
•Query for Instance Association	VIS / VPS	[172], [185], [186], [187], [188], [189]
•Query for Linking Multi-Tasks	VPS / DVPS / PS / PPS / IS	[128], [190], [191], [192], [193], [194]
Conditional Query Generation (Sec. 3.2.5)		
•Conditional Query Fusion on Language Features	RIS / RVOS	[48], [57], [195], [196], [197], [198], [199], [200], [201], [202]
•Conditional Query Fusion on Cross Image Features	SS/ VOS / SS / Few Shot SS	[203], [204], [205], [206], [207], [208], [209]

We select the representative works for reference.

described in Section III-A. We list the comparison of representative works in Table IV.

1) Strong Representations: Learning a strong feature representation always leads to better segmentation results. Taking the SS task as an example, SETR [124] is the first to replace CNN backbone with the ViT backbone. It achieves state-of-the-art results on the ADE20 k dataset without bells and whistles. After ViTs, researchers start to design better vision transformers. We categorize the related works into three aspects: better vision transformer design, hybrid CNNs/transformers/MLPs, and self-supervised learning.

• *Better ViTs Design:* Rather than introducing local bias, these works follow the original ViTs design and process feature using the original MHSA for token mixing. DeiT [130] proposes knowledge distillation and provides strong data augmentation to train ViT efficiently. Starting from DeiT, nearly all ViTs adopt the stronger training procedure. MViT-V1 [131] introduces the multiscale feature representation and pooling strategies to reduce the computation cost in MHSA. MViT-V2 [132] further incorporates decomposed relative positional embeddings and residual pooling design in MViT-V1, which leads to better representation. Motivated by MViT, from the architecture level, MPViT [133] introduces multiscale patch embedding and multi-path structure to explore tokens of different scales jointly. Meanwhile, from the operator level, XCiT [134] operates across feature channels rather than token inputs and proposes cross-covariance attention, which has linear complexity in the number of tokens. This design makes it easy to adapt to segmentation tasks, which always have high-resolution inputs. Pyramid ViT [135] is the first work to build multiscale features for detection and segmentation tasks. There are also several works [136], [137], [138] exploring cross-scale modeling via MHSA, which exchange long-range information on different feature pyramids.

• *Hybrid CNNs/Transformers/MLPs:* Rather than modifying the ViTs, many works focus on introducing local bias into ViT or using CNNs with large kernels directly. To build a multi-stage pipeline, Swin [23], [210] adopts shift-window attention in a CNN style. They also scale up the models to large sizes and achieve significant improvements on many vision tasks. From an efficient perspective, Segformer [123] designs a light-weight transformer encoder. It contains a sequence reduction during MHSA and a light-weight MLP decoder. Segformer achieves better speed and accuracy trade-off for SS. Meanwhile, several works [139], [140], [141], [142] directly add CNN layers to a transformer to explore the local context. Several works [211], [212] explore the pure MLPs design to replace the transformer. With specific designs such as shifting and fusion [211], MLP models can also achieve comparable results with ViTs. Later, several works [143], [144] point out that CNNs can achieve stronger results than ViTs if using the same data augmentation pipeline. In particular, DWNet [144] re-visits the training pipeline of ViTs and proposes dynamic depth-wise convolution. Then, ConvNeXt [143] uses the larger kernel depth-wise convolution and a stronger data training pipeline. It achieves stronger results than Swin [23]. Motivated by ConvNeXt, SegNext [145] designs a CNN-like backbone with linear self-attention and performs strongly on multiple SS benchmarks. Meanwhile, Meta-Former [146] shows that the meta-architecture of ViT is the key to achieving stronger results. Such meta-architecture contains a token mixer, a MLP, and residual connections. The token mixer is a simple MHSA layer. Meta-Former shows that the token mixer is not as important as meta-architecture. Using simple pooling as a token mixer can achieve stronger results. Following the Meta-Former, recent work [147] re-benchmarks several previous works using a unified architecture to eliminate unfair engineering techniques. However, under stronger settings, the authors find the spatial token mixer design still matters. Meanwhile, several works [214] explore the MLP-like architecture for dense prediction.

• *Self-Supervised Learning (SSL):* SSL has achieved huge progress in recent years [148], [149], [215]. Compared with supervised learning, SSL exploits unlabeled data via specially designed pseudo tasks and can be easily scaled up. MoCo-v3 [150] is the first study that trains ViTs in SSL. It freezes the patch projection layer to stabilize the training process. Motivated by BERT, BEiT [151] proposes the BERT-like per-training (Mask Image Modeling, MIM) of vision transformers. After BEiT, MAE [24] shows that ViTs can be trained with the simplest MIM style. By masking a portion of input tokens and reconstructing the RGB images, MAE achieves better results than supervised training. As a concurrent work, MaskFeat [152] mainly

TABLE IV
REPRESENTATIVE WORKS SUMMARIZATION AND COMPARISON IN SECTION III

Method	Task	Input/Output	Transformer Architecture	Highlight
Strong Representations (Sec. 3.2.1)				
SETR [124]	SS	Image/Semantic Masks	Pure transformer + CNN decoder	the first vision transformer to replace CNN backbone in SS.
Segformer [123]	SS	Image/Semantic Masks	Pure transformer + MLP head	a light-weight transformer backbone with simple MLP prediction head.
MAE [24]	SS/IS	Image/Semantic Masks	Pure transformer + CNN decoder	a MIM pretraining framework for plain ViTs, which achieves better results than supervised training.
SegNext [145]	SS	Image/Semantic Masks	Transformer + CNN	a large kernel CNN backbone with linear self-attention layer.
Cross-Attention Design in Decoder (Sec. 3.2.2)				
Deformable DETR [25]	OD	Image/Box	CNN + query decoder	a new multi-scale deformable attention and a new encoder-decoder framework.
OCRNet [66]	SS	Image/Semantic Masks	CNN + query decoder	introduces category queries and uses one cross-attention layer to model global context efficiently.
Segmenter [225]	SS	Image/Semantic Masks	ViT + query decoder	uses ViT backbone and category queries to directly output each class mask.
Sparse-RCNN [158]	OD	Image/Box	CNN + query decoder	a new dynamic convolution layer and combine object query with RoI-based detector.
AdaMixer [233]	OD	Image/Box	CNN + query decoder	a new multiscale query-based decoder and refine query with multiscale features.
Max-Deeplab [26]	PS	Image/Panoptic Masks	CNN + attention + query decoder	the first pure mask supervised panoptic segmentation method and a two path framework (query and CNN features).
K-Net [163]	SS/IS/PS	Image/Panoptic Masks	CNN + query decoder	the first work using kernels to unify image segmentation tasks and a new mask-based dynamic kernel update module.
Mask2Former [226]	SS/IS/PS	Image/Panoptic Masks	CNN + query decoder	design masked cross-attention and fully utilize the multiscale features in the decoder.
kMax-Deeplab [229]	SS/PS	Image/Panoptic Masks	CNN + query decoder	proposes a new k-mean style cross-attention by replacing softmax with argmax operation.
VisTR [166]	VIS	Video/Instance Masks	CNN + query decoder	the first end-to-end VIS method and each query represent a tracked object in a clip.
VITA [237]	VIS	Video/Instance Masks	CNN + query decoder	use the fixed object detector and process all frame queries with extra encoder-decoder and global queries.
TubeFormer [173]	VSS/VIS/VPS	Video/Panoptic Masks	CNN + query decoder	a tube-like decoder with a token exchanging mechanism within the tube, which unifies three video segmentation tasks in one framework.
Video K-Net [172]	VSS/VIS/VPS	Video/Panoptic Masks	CNN + query decoder	unified online video segmentation and adopt object query for association and linking.
Optimizing Object Query (Sec. 3.2.3)				
Conditional DETR [174]	OD	Image/Box	CNN + query decoder	add a spatial query to explore the extremity regions to speed up the DETR training.
DN-DETR [178]	OD	Image/Box	CNN + query decoder	add noisy boxes and de-noisy loss to stable query matching and improve the coverage of DETR.
Group-DETR [183]	OD	Image/Box	CNN + query decoder	introduce one-to-many assignment by extending more queries into groups.
Mask-DINO [180]	IS/PS	Image/Panoptic Masks	CNN + query decoder	boost instance/panoptic segmentation with object detection datasets.
Using Query For Association (Sec. 3.2.4)				
MOTR [187]	MOT	Video/Box	CNN + query decoder	design an extra tracking query for object association.
MiniVIS [188]	VIS	Video/Instance Masks	CNN/transformer + query decoder	perform video instance segmentation with image level pretraining and image object query for tracking.
Polyphonicformer [128]	D-VPS	Video/(Depth+Panoptic Masks)	CNN/transformer + query decoder	use object query and depth query to model instance-wise mask and depth prediction jointly.
X-Decoder [238]	SS/PS	Image /Panoptic Masks	CNN/transformer + query decoder	jointly pre-train image segmentation and language model and perform zero-shot inference on multiple segmentation tasks.
LMPM [57]	RVOS	(Video+Text)/Instance Masks	CNN/transformer + query decoder	capture motion by associating frame-level object tokens from an off-the-shelf instance segmentation model.
Conditional Query Generation (Sec. 3.2.5)				
VLT [195], [223]	RIS	(Image+Text)/Instance Masks	CNN + transformer decoder	design a query generation module to produce language conditional queries for transformer decoder.
LAVT [196]	RIS	(Image+Text)/Instance Masks	Transformer + CNN decoder	design gated cross-attention between pyramid features and language features.
MTTR [199]	RVOS	(Video+Text)/Instance Masks	Transformer + query decoder	perform spatial-temporal cross-attention between language features and object query.
X-DETR [239]	OD	(Image+Text)/Box	Transformer + query decoder	perform directly alignment between language features and object query.
CyCTR [203]	Few-Shot SS	(Image+Masks)/Instance Masks	Transformer + query decoder	design a cycle cross-attention between features in support images and query images.

studies reconstructing targets of the MIM framework, such as the histogram of oriented gradient (HOG) features. The following works focus on improving the MIM framework [153], [154] or replacing the backbone of ViTs with CNN architecture [155], [216]. **********DINO series [216] find the self-supervised learned feature itself has grouping effects, which is always used in unsupervised learning contexts. (Section IV-D) Recently, several works [156], [217] on VLM also adopt SSL by utilizing easily obtained text-image pairs. Recent work [157] demonstrates the effectiveness of VLM in downstream tasks, including IS and SS. Moreover, several recent works [218] adopt multi-modal SSL pre-training and design a unified model for many vision tasks. For video representation learning, most current works [219], [220], [221] verify such representation learning on action or motion learning, such as action recognition. Several works [202], [222] adopt a video backbone for video segmentation. However, for video segmentation, from the method design perspective, most works focus on matching and association of entities or pixels, which is discussed in Sections III-B2 and III-B4.

2) Cross-Attention Design in Decoder: In this section, we review the new transformer decoder designs. We categorize the decoder design into two groups: one for improved cross-attention design in image segmentation and the other for spatial-temporal cross-attention design in video segmentation. The former focuses on designing a better decoder to refine the

original decoder in the original DETR. The latter extends the query-based object detector and segmenter into the video domain for VOD, VIS, and VPS, focusing on modeling temporal consistency and association.

• *Improved Cross-Attention Design:* Cross-attention is the core operation of meta-architecture for segmentation and detection. Current solutions for improved cross-attention mainly focus on designing new or enhanced cross-attention operators and improved decoder architectures. Following DETR, Deformable DETR [25] proposes deformable attention to efficiently sample point features and perform cross-attention with object query jointly. Meanwhile, several works bring object queries into previous RCNN frameworks. Sparse-RCNN [158] uses RoI pooled features to refine the object query for object detection. They also propose a new dynamic convolution and self-attention to enhance object query without extra cross-attention. In particular, the pooled query features reweight the object query, and then self-attention is applied to the object query to obtain the global view. After that, several works [159], [160] add the extra mask heads for IS. QueryInst [159] adds mask heads and refines mask query with dynamic convolution. Meanwhile, several works [161], [223] extend Deformable DETR by directly applying MLP on the shared query. Inspired by MEInst [81], SOLQ [161] utilizes mask encodings on object query via MLP. By applying the strong Deformable DETR detector and Swin transformer [23] backbone, it achieves remarkable results on IS. However, these works still need extra box supervision, which makes the system complex. Moreover, most RoI-based approaches for IS have low mask quality issues since the mask resolution is limited within the boxes [71].

To fix the issues of extra box heads, several works remove the box prediction and adopt pure mask-based approaches. Earlier work, OCRNet [66] characterizes a pixel by exploiting the representation of the corresponding object class that forms a category query. Then, Segmenter [224] adopts a strong ViT backbone with the class query to directly decode class-wise masks. Pure mask-based approaches directly generate segmentation masks from high-resolution features and naturally have better mask quality. Max-Deeplab [26] is the first to remove the box head and design a pure-mask-based segmenter for PS. It also achieves stronger performance than box-based PS method [83]. It combines a CNN-transformer hybrid encoder [89] and a transformer decoder as an extra path. Max-Deeplab still needs extra auxiliary loss functions, such as semantic segmentation loss, and instance discriminative loss. K-Net [163] uses mask pooling to group the mask features and designs a gated dynamic convolution to update the corresponding query. By viewing the segmentation tasks as convolution with different kernels, K-Net is the first to unify all three image segmentation tasks, including SS, IS, and PS. Meanwhile, MaskFormer [164] extends the original DETR by removing the box head and transferring the object query into the mask query via MLPs. It proves simple mask classification can work well enough for all three segmentation tasks. Compared to MaskFormer, K-Net is good at training data efficiency. This is because K-Net adopts mask pooling to localize object features and then update object queries accordingly. Motivated by this, Mask2Former [225] proposes masked cross-attention and replaces the cross-attention in MaskFormer. Masked cross-attention makes object query only attend to the object area, guided by the mask outputs from previous stages. Mask2Former also adopts a stronger Deformable FPN backbone [25], stronger data augmentation [226], and multiscale mask decoding. The above works only consider updating object queries. To handle this, CMT-Deeplab [227] proposes an alternating procedure for object query and decoder features. It jointly updates object queries and pixel features. After that, inspired by the k-means clustering algorithm, kMaX-DeepLab [228] proposes k-means cross-attention by introducing cluster-wise argmax operation in the cross-attention operation. Meanwhile, PanopticSegformer [165] proposes a decoupling query strategy and deeply supervised mask decoder to speed up the training process. For real-time segmentation setting, SparseInst [229] proposes a sparse set of instance activation maps highlighting informative regions for each foreground object.

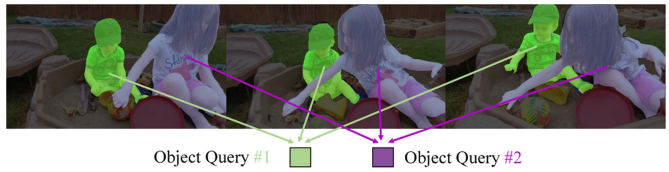

Fig. 4. Illustration of object query in video segmentation.

Besides segmentation tasks, several works speed up the convergence of DETR by introducing new decoder designs, and most approaches can be extended into IS. Several works bring such semantic priors in the DETR decoder. SAM-DETR [230] projects object queries into semantic space and searches salient points with the most discriminative features. SMAC [231] conducts location-aware co-attention by sampling features of high near estimated bounding box locations. Several works adopt dynamic feature re-weights. From the multiscale feature perspective, AdaMixer [232] samples feature over space and scales using the estimated offsets. It dynamically decodes sampled features with an MLP, which builds a fast-converging query-based detector. ACT-DETR [233] clusters the query features adaptively using a locality-sensitive hashing and replaces the query-key interaction with the prototype-key interaction to reduce cross-attention cost. From the feature re-weighting view, Dynamic-DETR [234] introduces dynamic attention to both the encoder and decoder parts of DETR using RoI-wise dynamic convolution. Motivated by the sparsity of the decoder feature, Sparse-DETR [235] selectively updates the referenced tokens from the decoder and proposes an auxiliary detection loss on the selected tokens in the encoder to keep the sparsity. In summary, dynamically assigning features into query learning speeds up the convergence of DETR.

• *Spatial-Temporal Cross-Attention Design:* After extending the object query in the video domain, each object query represents a tracked object across different frames, which is shown in Fig. 4. The simplest extension is proposed by VisTR [166] for VIS. VisTR extends the cross-attention in DETR into multiple frames by stacking all clip features into flattened spatial-temporal features. The spatial-temporal features also

involve temporal embeddings. During inference, one object query can directly output spatial-temporal masks without extra tracking. Meanwhile, TransVOD [167] proposes to link object query and corresponding features across the temporal domain. It splits the clips into sub-clips and performs clip-wise object detection. TransVOD utilizes the local temporal information and achieves better speed and accuracy trade-off. IFC [170] adopts message tokens to exchange temporal context among different frames. The message tokens are similar to learnable queries, which perform cross-attention with features in each frame and self-attention among the tokens. After that, TeViT [168] proposes a novel messenger shift mechanism for temporal fusion and a shared spatial-temporal query interaction mechanism to utilize both frame-level and instance-level temporal context information. Seqformer [171] combines Deformable-DETR and VisTR in one framework. It also proposes to use image datasets to augment video segmentation training. Mask2Former-VIS [169] extends masked cross-attention in Mask2Former [225] into temporal masked cross-attention. Following VisTR, it also directly outputs spatial-temporal masks.

In addition to VIS, several works [163], [165], [225] have shown that query-based methods can naturally unify different segmentation tasks. Following this pipeline, there are also several works [172], [173] solving multiple video segmentation tasks in one framework. In particular, based on K-Net [163], Video K-Net [172] proposes to unify VPS/VIS/VSS via tracking and linking kernels and works in an online manner. Meanwhile, TubeFormer [173] extends Max-Deeplab [26] into the temporal domain by obtaining the mask tubes. Cross-attention is carried out in a clip-wise manner. During inference, the instance association is performed by mask-based matching. Moreover, several works [236] propose the local temporal window to refine the global spatial-temporal cross-attention. For example, VITA [236] aggregates the local temporal query on top of an off-the-shelf transformer-based image instance segmentation model [225]. Recently, several works [239], [240] have explored the cross-clip association for video segmentation. In particular, Tube-Link [240] designs a universal video segmentation framework via learning cross-tube relations. It performs better than task-specific methods in VSS, VIS, and VPS.

3) Optimizing Object Query: Compared with Faster-RCNN [115], DETR [22] needs a much longer schedule for convergence. Due to the critical role of object query, several approaches have launched studies on speeding up training schedules and improving performance. According to the methods for the object query, we divide the following literature into two aspects: adding position information and adopting extra supervision. The position information provides the cues to sample the query feature for faster training. The extra supervision focuses on designing specific loss functions in addition to default ones in DETR.

• *Adding Position Information into Query:* Conditional DETR [174] finds cross-attention in DETR relies highly on the content embeddings for localizing the four extremities. The authors introduce conditional spatial query to explore the extremity regions explicitly. Conditional DETR V2 [175] introduces the box queries from the image content to improve detection results. The box queries are directly learned from image content, which is dynamic with various image inputs. The image-dependent box query helps locate the object and improve the performance. Motivated by previous anchor designs in object detectors, several works bring anchor priors in DETR. The Efficient DETR [241] adopts hybrid designs by including query-based and dense anchor-based predictions in one framework. Anchor DETR [176] proposes to use anchor points to replace the learnable query and also designs an efficient self-attention head for faster training. Each object query predicts multiple objects at one position. DAB-DETR [177] finds the localization issues of the learnable query and proposes dynamic anchor boxes to replace the learnable query. Dynamic anchor boxes make the query learning more explainable and explicitly decouple the localization and content part, further improving the detection performance.

• *Adding Extra Supervision into Query:* DN-DETR [178] finds that the instability of bipartite graph matching causes the slow convergence of DETR and proposes a denoising loss to stabilize query learning. In particular, the authors feed GT bounding boxes with noises into the transformer decoder and train the model to reconstruct the original boxes. Motivated by DN-DETR, based on Mask2Former, MP-Former [242] finds inconsistent predictions between consecutive layers. It further adds class embeddings of both ground truth class labels and masks to reconstruct the masks and labels. Meanwhile, DINO [179] improves DN-DETR via a contrastive way of denoising training and a mixed query selection for better query initialization. Mask DINO [180] extends DINO by adding an extra query decoding head for mask prediction. Mask DINO [180] proposes a unified architecture and joint training process for both object detection and instance segmentation. By sharing the training data, Mask DINO can scale up and fully utilize the detection annotations to improve IS results. Meanwhile, motivated by contrastive learning, IUQ [181] introduces two extra supervisions, including cross-image contrastive query loss via extra memory blocks and equivalent loss against geometric transformations. Both losses can be naturally adapted into query-based detectors. Meanwhile, there are also several works [182], [183], [184], [243] exploring query supervision from the target assignment perspective. In particular, since DETR lacks the capability of exploiting multiple positive object queries, DE-DETR [243] first introduces one-to-many label assignment in query-based instance perception framework, to provide richer supervision for model training. Group DETR [183] proposes group-wise one-to-many assignments during training. H-DETR [182] adds auxiliary queries that use one-to-many matching loss during training. Rather than adding more queries, Co-DETR [184] proposes a collaborative hybrid training scheme using parallel auxiliary heads supervised by one-to-many label assignments. All these approaches drop the extra supervision heads during inference. These extra supervision designs can be easily extended to query-based segmentation methods [163], [225].

4) Using Query for Association: Benefiting from the simplicity of query representation, several recent works have adopted it as an association tool to solve downstream tasks.

There are mainly two usages: one for instance-level association and the other for task-level association. The former adopts the idea of instance discrimination, for instance-wise matching problems in video, such as joint segmentation and tracking. The latter adopts queries to link features for multitask learning.

• *Using Query for Instance Association:* The research in this area can be divided into two aspects: one for designing extra tracking queries and the other for using object queries directly. TrackFormer [185] is the first to treat multi-object tracking as a set prediction problem by performing joint detection and tracking-by-attention. TransTrack [186] uses the object query from the last frame as a new track query and outputs tracking boxes from the shared decoder. MOTR [187] introduces the extra track query to model the tracked instances of the entire video. In particular, MOTR proposes a new tracklet-awared label assignment to train track queries and a temporal aggregation module to fuse temporal features. There are also several works [57], [172], [188], [189], [240] adopting object query solely for tracking. In particular, MiniVIS [188] directly uses object query for matching without extra tracking head modeling for VIS, where it adopts image instance segmentation training. Both Video K-Net [172] and IDOL [189] learn the association embeddings directly from the object query using a temporal contrastive loss. During inference, the learned association embeddings are used to match instances across frames. These methods are usually verified in VIS and VPS tasks. All methods pre-train their image baseline on image datasets, including COCO and Cityscapes, and fine-tune their video architecture in the video datasets.

• *Using Query for Linking Multi-Tasks:* Several works [128], [190], [244], [245] use object query to link features across different tasks to achieve mutual benefits. Rather than directly fusing multitask features, using object query fusion not only selects the most discriminative parts to fuse but also is more efficient than dense feature fusion. In particular, Panoptic-PartFormer [190] links part and panoptic features via different object queries into an end-to-end framework, where joint learning leads to better part segmentation results. Several works [128], [191] combine segmentation features, and depth features using the MHSA layer on corresponding depth query and segmentation query, which unify the depth prediction and panoptic segmentation prediction via shared masks. Both works find the mutual effect for both segmentation and depth prediction. Recently, several works [193], [194] have adopted the vision transformers with multiple task-aware queries for multi-task dense prediction tasks. In particular, they treat object queries as task-specific hidden features for fusion and perform cross-task reasoning using MSHA on task queries. Moreover, in addition to dense prediction tasks, FashionFormer [192] unifies fashion attribute prediction and instance part segmentation in one framework. It also finds the mutual effect on instance segmentation and attribute prediction via query sharing. Recently, X-Decoder [237] uses two different queries for segmentation and language generation tasks. The authors jointly pre-train two different queries using large-scale vision language datasets, where they find both queries can benefit corresponding tasks, including visual segmentation and caption generation.

5) Conditional Query Fusion: In addition to using object query for multitask prediction, several works adopt conditional query design for cross-modal and cross-image tasks. The query is conditional on the task inputs, and the decoder head uses such a conditional query to obtain the corresponding segmentation masks. Based on the source of different inputs, we split these works into two aspects: language features and image features.

• *Conditional Query Fusion From Language Feature:* Several works [48], [48], [57], [195], [196], [197], [199], [200], [201], [222], [245], [246], [247] adopt conditional query fusion according to input language feature for both referring image segmentation (RIS) [48] and referring video object segmentation (RVOS) [57] tasks. In particular, VLT [195], [222] first adopts the vision transformer for the RIS task and proposes a query generation module to produce multiple sets of language-conditional queries, which enhances the diversified comprehensions of the language. Then, it adaptively selects the output features of these queries via the proposed query balance module. Following the same idea, LAVT [196] designs a new gated cross-attention fusion where the image features are the query inputs of a self-attention layer in the encoder part. Compared with previous CNN approaches [248], [249], using a vision transformer significantly improves the language-driven segmentation quality. With the help of CLIP's knowledge, CRIS [198] proposes vision-language decoding and contrastive learning for achieving text-to-pixel alignment. Meanwhile, several works [57], [199], [202], [250] adopt video detection transformer in Section III-B2 for the RVOS task. MTTR [199] models the RVOS task as a sequence prediction problem and proposes both language and video features jointly. Recently, several works [57], [245] explore referring VOS under fast motion condition settings. Each object query in each frame combines the language features before sending it into the decoder. To speed up the query learning, ReferFormer [202] designs a small set of object queries conditioned on the language as the input to the transformer. The conditional queries are transformed into dynamic kernels to generate tracked object masks in the decoder. With the same design as VisTR, ReferFormer can segment and track object masks with given language inputs. In this way, each object tracklet is controlled by a given language input. In addition to referring segmentation tasks, MDETR [251] presents an end-to-end modulated detector that detects objects in an image conditioned on a raw text query. In particular, they fuse the text embedding directly into visual features and jointly train the fused feature and object query. X-DETR [238] proposes an effective architecture for instance-wise vision-language tasks via using dot-product to align vision and language. In summary, these works fully utilize the interaction of language features and query features.

• *Condition Query Fusion From Image Feature:* Several tasks take multiple images as references and refine corresponding object masks of the main image. The multiple images can be support images in few shot segmentation [203], [209], [252] or the same input image in matting [205], [253] and semantic segmentation [206], [207]. These works aim to model the correspondences between the main image and other images via condition query fusion. For SS, StructToken [207] presents a

new framework by doing interactions between a set of learnable structure tokens and the image features, where the image features are the spatial priors. In the video, BATMAN [208] fuses optical flow features and previous frame features into mixed features and uses such features as a query to decode the current frame outputs. For few-shot segmentation, CyCTR [203] aggregates pixel-wise support features into query features. In particular, CyCTR performs cross-attention between features from different images in a cycle manner, where support image features and query image features are the query inputs of the transformer jointly. Meanwhile, MM-Former [209] adopts a class-agnostic method [225] to decompose the query image into multiple segment proposals. Then, the support and query image features are used to select the correct masks via a transformer module. Then, for few-shot instance segmentation, RefTwice [254] proposes an object query enhanced framework to weight query image features via object queries from support queries. In image matting, MatteFormer [205] designs a new attention layer called prior-attentive window self-attention based on Swin [23]. The prior token represents the global context feature of each trimap region, which is the query input of window self-attention. The prior token introduces spatial cues and achieves thinner matting results. In summary, according to the different tasks, the image features play as the decoder features in previous Section III-B2, which enhance the features in the main images.

IV. Specific Subfields

In this section, we revisit several related subfields that adopt vision transformers for segmentation tasks. The subfields include point cloud segmentation, domain-aware segmentation, label and model efficient segmentation, class agnostic segmentation, tracking, and medical segmentation.

A. Segmentation

• *Semantic Level Point Cloud Segmentation:* Like image segmentation and video semantic segmentation, adopting transformers for semantic level processing mainly focuses on learning a strong representation (Section III-B1). The works [29], [255] focus on transferring the success in image/video representation learning into the point cloud. Early works [29] directly use modified self-attention as backbone networks and design U-Net-like architectures for segmentation. In particular, Point-Transformer [29] proposes vector self-attention and subtraction relation to aggregate local features progressively. The concurrent work PCT [255] also adopts a self-attention operation and enhances input embedding with the support of farthest point sampling and nearest neighbor searching. However, the ability to model long-range context and cross-scale interaction is still limited. Stratified-Transformer [256] extends the idea of Swin Transformer [23] into the point cloud and dived 3D inputs into cubes. It proposes a mixed key sampling method for attention input and enlarges the effective receptive field via merging different cube outputs. Meanwhile, several works also focus on better pre-training or distilling the knowledge of 2D pre-trained models. PointBert [257] designs the first Masked Point Modeling (MPM) task to pre-train point cloud transformers. It divides a point cloud into several local point patches as the input of a standard transformer. Moreover, it also pre-trains a point cloud Tokenizer with a discrete variational autoEncoder to encode the semantic contents and train an extra decoder using the reconstruction loss. Following MAE [24], several works [258], [259] simply the MIM pretraining process. Point-MAE [258] divides the input point cloud into irregular point patches and randomly masks them at a high ratio. Then, it uses a standard transformer-based autoencoder to reconstruct the masked points. Point-M2AE [259] designs a multiscale MIM pretraining by making the encoder and decoder into pyramid architectures to model spatial geometries and multilevel semantics progressively. Meanwhile, benefiting from the same transformer architecture for point cloud and image, several works adopt image pre-trained standard transformer by distilling the knowledge from large-scale image dataset pre-trained models.

• *Instance Level Point Cloud Segmentation:* As shown in Section II, previous PCIS / PCPS approaches are based on manually-tuned components, including a voting mechanism that predicts hand-selected geometric features for top-down approaches and heuristics for clustering the votes for bottom-up approaches. Both approaches involve many hand-crafted components and post-processing, The usage of transformers in instance-level point cloud segmentation is similar to the image or video domain, and most works use bipartite matching for instance-level masks for indoor and outdoor scenes. For example, Mask3D [260] proposes the first Transformer-based approach for 3D semantic instance segmentation. It models each object instance as an instance query and uses the transformer decoder to refine each instance query by attending to point cloud features at different scales. Meanwhile, SPFormer [261] learns to group the potential features from point clouds into super-points [262], and directly predicts instances through instance query with a masked-based transformer decoder. The super-points utilize geometric regularities to represent homogeneous neighboring points, which is more efficient than all point features. The transformer decoder works similarly to Mask2Former, where the cross-attention between instance query and super-point features is guided by the attention mask from the previous stage. PUPS [263] proposes a unified PPS system for outdoor scenes. It presents two types of learnable queries named semantic score and grouping score. The former predicts the class label for each point, while the latter indicates the probability of grouping ID for each point. Then, both queries are refined via grouped point features, which share the same ideas from previous Sparse-RCNN [158] and K-Net [163]. Moreover, PUPS also presents a context-aware mixing to balance the training instance samples, which achieves the new state-of-the-art results [264].

B. Tuning Foundation Models

We divide this section into two aspects: vision adapter design and open vocabulary learning. The former introduces new ways to adapt the pre-trained large-scale foundation models for downstream tasks. The latter tries to detect and segment

unknown objects with the help of the pre-trained vision language model and zero-shot knowledge transfer on unseen segmentation datasets. The core idea for vision adapter design is to extract the knowledge of foundation models and design better ways to fit the downstream settings. For open vocabulary learning, the core idea is to align pre-trained VLM features into current detectors to achieve novel class classification.

• *Vision Adapter and Prompting Modeling:* Following the idea of prompt tuning in NLP, early works [265], [266] adopt learnable parameters with the frozen foundation models to better transfer the downstream datasets. These works use small image classification datasets for verification and achieve better results than original zero-shot results [217]. Meanwhile, there are several works [267] designing adapter and frozen foundation models for video recognition tasks. In particular, the pre-trained parameters are frozen, and only a few learnable parameters or layers are tuned. Following the idea of learnable tuning, recent works [268], [269] extend the vision adapter into dense prediction tasks, including segmentation and detection. In particular, ViT-Adapter [268] proposes a spatial prior module to solve the issue of the location prior assumptions in ViTs. The authors design a two-stream adaption framework using deformable attention and achieve comparable results in downstream tasks. From the CLIP knowledge usage view, DenseCLIP [269] converts the original image-text in CLIP to a pixel-text matching problem and uses the pixel-text score maps to guide the learning of dense prediction models. From the task prompt view, CLIPSeg [270] builds a system to generate image segmentations based on arbitrary prompts at test time. A prompt can be a text or an image where the CLIP visual model is frozen during training. In this way, the segmentation model can be turned into a different task driven by the task prompt. Previous works only focus on a single task. OneFormer [271] extends the Mask2Former with multiple target training setting and perform segmentation driven by the task prompt. Moreover, using a vision adapter and text prompt can easily reduce the taxonomy problems of each dataset and learn a more general representation for different segmentation tasks. Recently, SAM [272] proposes more generalized prompting methods, including mask, points, box, and text. The authors build a larger dataset with 1 billion masks. SAM achieves good zero-shot performance in various segmentation datasets.

• *Open Vocabulary Learning:* Recent studies [246], [273], [274], [275], [276], [277], [278] focus on the open vocabulary and open world setting, where their goal is to detect and segment novel classes, which are not seen during the training. Different from zero-shot learning, an open vocabulary setting assumes that large vocabulary data or knowledge can provide cues for final classification. Most models are trained by leveraging pre-trained language-text pairs, including captions and text prompts, or with the help of VLM. Then, trained models can detect and segment the novel classes with the help of weakly annotated captions or existing publicly available VLM. In particular, VilD [274] distills the knowledge from a trained open vocabulary image classification model CLIP into a two-stage detector. However, VilD still needs an extra visual CLIP encoder for visual distillation. To handle this, Forzen-VLM [279] adopts the frozen visual clip model and combines the scores of both learned visual embedding and CLIP embedding for novel class detection. From the data augmentation view, MViT [280] combines the Deformable DETR and CLIP text encoder for the open world class-agnostic detection, where the authors build a large dataset by mixing existing detection datasets. Motivated by the more balanced samples from image classification datasets, Detic [275] improves the performance of the novel classes with existing image classification datasets by supervising the max-size proposal with all image labels. OV-DETR [276] designs the first query-based open vocabulary framework by learning conditional matching between class text embedding and query features. Besides these open vocabulary detection settings, recent works [281], [282] perform open vocabulary segmentation. In particular, L-Seg [282] presents a new setting for language-driven semantic image segmentation and proposes a transformer-based image encoder that computes dense per-pixel embeddings according to the language inputs. OpenSeg [281] learns to generate segmentation masks for possible candidates using a DETR-like transformer. Then it performs visual-semantic alignments by aligning each word in a caption to one or a few predicted masks. BetrayedCaption [283] presents a unified transformer framework by joint segmentation and caption learning, where the caption part contains both caption generation and caption grounding. The novel class information is encoded into the network during training. With the goal of unifying different segmentation with text prompts, FreeSeg [284] adopts a similar pipeline as OpenSeg to crop frozen CLIP features for novel class classification. Meanwhile, open set segmentation [285] requires the model to output class agnostic masks and enhance the generality of segmentation models. Recently, ODISE [286] uses a frozen diffusion model as the feature extractor, a Mask2Former head, and joint training with caption data to perform open vocabulary panoptic segmentation. There are also several works [287] focusing on open-world object detection, where the task detects a known set of object categories while simultaneously identifying unknown objects. In particular, OW-DETR [287] adopts the DETR as the base detector and proposes several improvements, including attention-driven pseudo-labeling, novelty classification, and objectness scoring. In summary, most approaches [284], [288] adopt the idea of region proposal network [115] to generate class-agnostic mask proposals via different approaches, including anchor-based and query-based decoders in Section III-A. Then, the open vocabulary problem turns into a region-level matching problem to match the visual region features with pre-trained VLM language embedding.

C. Domain-Aware Segmentation

• *Domain Adaption:* Unsupervised Domain Adaptation (UDA) aims at adapting the network trained with source (synthetic) domain into target (real) domain [45], [289] without access to target labels. UDA has two different settings, including semantic segmentation and object detection. Before ViTs, the previous works [290] mainly design domain-invariant representation learning strategies. DAFormer [291] replaces the outdated backbone with the advanced transformer backbone [123] and

proposes three training strategies, including rare class sampling, thing-class ImageNet feature loss, and a learning rate warm-up method. It achieves new state-of-the-art results and is a strong baseline for UDA segmentation. Then, HRDA [292] improves DAFormer via a multi-resolution training approach and uses various crops to preserve fine segmentation details and long-range contexts. Motivated by MIM [24], MIC [293] proposes a masked image consistency to learn spatial context relations of the target domain as additional clues. MIC enforces the consistency between predictions of masked target images and pseudo-labels via a teacher-student framework. It is a plug-in module that is verified among various UDA settings. For detection transformers on UDA, SFA [294] finds feature distribution alignment on CNN brings limited improvements. Instead, it proposes a domain query-based feature alignment and a token-wise feature alignment module to enhance. In particular, the alignment is achieved by introducing a domain query and performing the domain classification on the decoder. DA-DETR [295] proposes a hybrid attention module (HAM), which contains a coordinate attention module and a level attention module along with the transformer encoder. A single domain-aware discriminator supervises the output of HAM. MTTrans [296] presents a teacher-student framework and a shared object query strategy. Meanwhile, SePiCo [297] introduces a new framework that extracts the semantic meaning of individual pixels to learn class-discriminative and class-balanced pixel representations. It supports both CNN and Transformer architecture. The image and object features between source and target domains are aligned at local, global, and instance levels.

• *Multi-Dataset Segmentation:* The goal of multi-dataset segmentation is to learn a universal segmentation model on various domains. MSeg [298] re-defines the taxonomies and aligns the pixel-level annotations by relabeling several existing semantic segmentation benchmarks. Then, the following works try to avoid taxonomy conflicts via various approaches. For example, Sentence-Seg [299] replaces each class label with a vector-valued embedding. The embedding is generated by a language model [15]. To further handle inflexible one-hot common taxonomy, LMSeg [300] extends such embedding with learnable tokens [265] and proposes a dataset-specific augmentation for each dataset. It dynamically aligns the segment queries in MaskFormer [164] with the category embeddings for both SS and PS tasks. Meanwhile, there are several works on multi-dataset object detection [301], [302]. In particular, Detection-Hub [302] proposes to adapt object queries on language embedding of categories per dataset. Rather than previously shared embedding for all datasets, it learns semantic bias for each dataset based on the common language embedding to avoid the domain gap. Meanwhile, several works [303], [304] focus on segmentation domain generation, which directly transfers learned knowledge from one domain to the remaining domains. TarVIS [127] jointly pre-trains one video segmentation model for different tasks spanning multiple benchmarks, where it extends Mask2Former into the video domain and adopts the unified image datasets pretraining and video fine-tuning. Recently, OMG-Seg [126] has unified multi-dataset segmentation, image/video segmentation, and open-vocabulary segmentation in one shared model and achieved using one model to segment all entities.

D. Label and Model Efficient Segmentation

• *Weakly Supervised Segmentation:* Weakly supervised segmentation methods learn segmentation with weaker annotations, such as image labels and object boxes. For weakly supervised semantic segmentation, previous works [305], [306] improve the typical CNN pipeline with class activation maps (CAM) and use refined CAM as training labels, which requires an extra model for training. ViT-PCM [307] shows the self-supervised transformers [150] with a global max pooling can leverage patch features to negotiate pixel-label probability and achieve end-to-end training and test with one model. MCTformer [305] adopts the idea that the attended regions of the one-class token in the vision transformer can be leveraged to form a class-agnostic localization map. It extends to multiple classes by using multiple class tokens to learn interactions between the class tokens and the patch tokens to generate the segmentation labels. For weakly supervised instance segmentation, previous works [308], [309], [310] mainly leverage the box priors to supervise mask heads. Recently, MAL [308] shows that vision transformers are good mask auto-labelers. It takes the box-cropped images as inputs and adopts a teacher-student framework, where the two vision transformers are trained with multiple instances loss [308]. MAL proves the zero-shot segmentation ability and achieves nearly mask-supervised performance on various baselines. Meanwhile, several works [311], [312] explore the text-only supervision for semantic segmentation. One representative work, GroupViT [311] adopts ViT to group image regions into progressively larger shaped segments.

• *Unsupervised Segmentation:* Unsupervised segmentation performs segmentation without any labels [313]. Before ViTs, recent progress [314] leverages the ideas from self-supervised learning. DINO [216] finds that the self-supervised ViT features naturally contain explicit information on the segmentation of input image. It finds that the attention maps between the CLS token and feature to describe the segmentation of objects. Instead of using the CLS token, LOST [315] solves unsupervised object discovery by using the key component of the last attention layer for computing the similarities between the different patches. Several works are aiming at finding the semantic correspondence of multiple images. Then, by utilizing the correspondence maps as guidance, they achieve better performance than DINO. Given a pair of images, SETGO [316] finds the self-supervised learned features of DINO have semantically consistent correlations. It proposes to distill unsupervised features into high-quality discrete semantic labels. Motivated by the success of VLM, ReCo [317] adopts the language-image pre-trained model, CLIP, to retrieve large unlabeled images by leveraging the correspondences in deep representation. Then, it performs co-segmentation among both input and retrieved images. There are also several works adopting sequential pipelines. MaskDistill [318] first identifies groups of pixels that likely belong to the same object with a bottom-up model. Then, it clusters the

object masks and uses the result as pseudo ground truths to train an extra model. Finally, the output masks are selected from the offline model according to the object score. FreeSOLO [319] first adopts an extra self-supervised trained model to obtain the coarse masks. Then, it trains a SOLO-based instance segmentation model via weak supervision. CutLER [320] proposes a new framework for multiple object mask generation. It first designs the MaskCut to discover multiple coarse masks based on the self-supervised features (DINO). Then, it adopts a detector to recall the missing masks via a loss-dropping strategy. Finally, it further refines mask quality via self-training

• *Mobile Segmentation:* Most transformer-based segmentation methods have huge computational costs and memory requirements, which make these methods unsuitable for mobile devices. Different from previous real-time segmentation methods [68], [70], [321], the mobile segmentation methods need to be deployed on mobile devices with considering both power cost and latency. Several earlier works [322], [323], [324], [325], [326], [327] focus on a more efficient transformer backbone. In particular, Mobile-ViT [323] introduces the first transformer backbone for mobile devices. It reduces image patches via MLPs before performing MHSA and shows better task-level generalization properties. There have also been several works on designing mobile semantic segmentation using transformers. TopFormer [328] proposes a token pyramid module that takes the tokens from various scales as input to produce the scale-aware semantic feature. SeaFormer [329] proposes a squeeze-enhanced axial transformer that contains a generic attention block. The block mainly contains two branches: a squeeze axial attention layer to model efficient global context and a detail enhancement module to preserve the details. RAP-SAM [327] proposes a new unified setting to put real-time interactive segmentation, panoptic segmentation, and video segmentation into one framework.

E. Class Agnostic Segmentation and Tracking

• *Fine-grained Object Segmentation:* Several applications, such as image and video editing, often need fine-grained details of object mask boundaries. Earlier CNN-based works focus on refining the object masks with extra convolution modules [71], or extra networks [330]. Most transformer-based approaches [331], [332], [333], [334], [335] adopt vision transformers due to their fine-grained multiscale features and long-range context modeling. Transfiner [331] refines the region of the coarse mask via a quad-tree transformer. By considering multiscale point features, it produces more natural boundaries while revealing details for the objects. Then, Video-Transfiner [332] refines the spatial-temporal mask boundaries by applying Transfiner [331] to the video segmentation method [166]. It can refine the existing video instance segmentation datasets [50]. PatchDCT [336] adopts the idea of ViT by making object masks into patches. Then, each mask is encoded into a DCT vector [337], and PatchDCT designs a classifier and a regressor to refine each encoded patch. Entity segmentation [338] aims to segment all visual entities without predicting their semantic labels. Its goal is to obtain high-quality and generalized segmentation results.

• *Video Object Segmentation:* Recent approaches for VOS mainly focus on designing better memory-based matching methods [339]. Inspired by the Non-local network [18] in image recognition tasks, the representative work STM [339] is the first to adopt cross-frame attention, where previous features are seen as memory. Then, the following works [204] design a better memory-matching process. associating objects with transformers (AOT) [204] matches and decodes multiple objects jointly. The authors propose a novel hierarchical matching and propagation, named long short-term transformer, where they joint persevere an identity bank and long-short term attention. XMem [340] proposes a mixed memory design to handle the long video inputs. The mixed memory design is also based on the self-attention architecture. Meanwhile, Clip-VOS [341] introduces per-clip memory matching for inference efficiency. Recently, to enhance instance-level context, Wang et al. [342] adds an extra query from Mask2Former into memory matching for VOS.

F. Medical Image Segmentation

CNNs have achieved milestones in medical image analysis. In particular, the U-shaped architecture and skip-connections [343], [344] have been widely applied in various medical image segmentation tasks. With the success of ViTs, recent representative works [345], [346] adopt vision transformers into the U-Net architecture and achieve better results. TransUNet [345] merges transformer and U-Net, where the transformer encodes tokenized image patches to build the global context. Then decoder upsamples the encoded features, which are then combined with the high-resolution CNN feature maps to enable precise localization. Swin-Unet [346] designs a symmetric Swin-like [23] decoder to recover fine details. TransFuse [347] combines transformers and CNNs in a parallel style, where global dependency and low-level spatial details can be efficiently captured jointly. UNETR [348] focuses on 3D input medical images and designs a similar U-Net-like architecture. The encoded representations of different layers in the transformer are extracted and merged with a decoder via skip connections to get the final 3D mask outputs.

V. Benchmark Results

In this section, we report recent transformer-based visual segmentation and tabulate the performance of previously discussed algorithms. For each reviewed field, the most widely used datasets are selected for performance benchmark in Sections V-A and V-C. We further re-benchmark several representative works in Section V-B using the same data augmentations and feature extractor. Note that we only list *published works* for reference. For simplicity, we have excluded several works on representation learning and only present specific segmentation methods. For a comprehensive method comparison, please refer to the supplementary material that provides a more detailed analysis. In addition, several works [349], [350], [351] achieve better results. However, due to the extra datasets [352] they used, we do not list them here.

TABLE V
BENCHMARK RESULTS ON SEMANTIC SEGMENTATION VALIDATION DATASETS

Method	backbone	COCO-Stuff	Cityscapes	ADE20K	Pascal-Context
SETR [124]	ViT-Large	-	82.2	50.3	55.8
Segmenter [225]	ViT-Large	-	81.3	53.6	59.0
SegFormer [123]	MiT-B5	46.7	84.0	51.8	-
SegNext [145]	MSCAN-L	47.2	83.9	52.1	60.9
ConvNext [143]	ConvNeXt-XL	-	-	54.0	-
MAE [24]	ViT-L	-	-	53.6	-
K-Net [163]	Swin-L	-	-	54.3	-
MaskFormer [164]	Swin-L	-	-	55.6	-
Mask2Former [226]	Swin-L	-	84.3	57.3	-
CLUSTSEG [354]	Swin-B	-	-	57.4	-
OneFormer [272]	ConvNeXt-XL	-	84.6	58.8	-

The results are with mIoU metric.

TABLE VI
BENCHMARK RESULTS ON INSTANCE SEGMENTATION OF COCO VALIDATION DATASETS

Method	backbone	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
SOLQ [161]	ResNet50	39.7	-	-	21.5	42.5	53.1
K-Net [163]	ResNet50	38.6	60.9	41.0	19.1	42.0	57.7
Mask2Former [226]	ResNet50	43.7	-	-	23.4	47.2	64.8
Mask DINO [180]	ResNet50	46.3	69.0	50.7	26.1	49.3	66.1
Mask2Former [226]	Swin-L	50.1	-	-	29.9	53.9	72.1
Mask DINO [180]	Swin-L	52.3	76.6	57.8	33.1	55.4	72.6
OneFormer [272]	Swin-L	49.0	-	-	-	-	-

The results are with the mAP metric.

TABLE VII
BENCHMARK RESULTS ON PANOPTIC SEGMENTATION VALIDATION DATASETS

Method	backbone	COCO	Cityscapes	ADE20K	Mapillary
PanopticSegFormer [165]	ResNet50	49.6	-	36.4	-
K-Net [163]	ResNet50	47.1	-	-	-
Mask2Former [226]	ResNet50	51.9	62.1	39.7	36.3
K-Max Deeplab [229]	ResNet50	53.0	64.3	42.3	-
Mask DINO [180]	ResNet50	53.0	-	-	-
Max-Deeplab [26]	MaX-L	51.1	-	-	-
PanopticSegFormer [165]	Swin-L	55.8	-	-	-
Mask2Former [226]	Swin-L	57.8	66.6	48.1	45.5
CMT-Deeplab [228]	Axial-R104-RFN	55.3	-	-	-
K-Max Deeplab [229]	ConvNeXt-L	58.1	68.4	50.9	-
OneFormer [272]	Swin-L	57.9	67.2	51.4	-
Mask DINO [180]	Swin-L	58.3	-	-	-
CLUSTSEG [354]	Swin-B	59.0	-	-	-

The results are with PQ metric.

A. Main Results on Image Segmentation Datasets

• *Results On Semantic Segmentation Datasets:* In Table V, Mask2Former [225] and OneFormer [271] perform the best on Cityscapes and ADE20 K dataset, while SegNext [145] achieves the best results on COCO-Stuff and Pascal-Context datasets.

• *Results on COCO Instance Segmentation:* In Table VI, Mask DINO [180] achieves the best results on the COCO instance segmentation with both ResNet and Swin-L backbones.

• *Results on Panoptic Segmentation:* In Table VII, for panoptic segmentation, Mask DINO [180] and K-Max Deeplab [228] achieve the best results on the COCO dataset. K-Max Deeplab also achieves the best results on Cityscapes. OneFormer [271] performs the best on ADE20 K.

B. Re-Benchmarking for Image Segmentation

• *Motivation:* We perform re-benchmarking on two segmentation tasks: semantic segmentation and panoptic segmentation on four public datasets, including ADE20 K, COCO, Cityscapes, and COCO-Stuff datasets. In particular, we want to explore the effect of the transformer decoder. Thus, we use the same encoder [7] and neck architecture [25] for a fair comparison.

TABLE VIII
BENCHMARK RESULTS ON VIDEO SEMANTIC SEGMENTATION OF VPSW VALIDATION DATASETS

Method	backbone	mIoU	mVC_8	mVC_{16}
TCB [49]	ResNet101	37.5	87.0	82.1
Video K-Net [172]	ResNet101	38.0	87.2	82.3
CFFM [355]	MiT-B5	49.3	90.8	87.1
MRCFA [356]	MiT-B5	49.9	90.9	87.4
TubeFormer [173]	Axial-ResNet	63.2	92.1	87.9

The results are with mIoU and mVC (mean Video Consistency) metrics.

TABLE IX
EXPERIMENT RESULTS ON SEMANTIC SEGMENTATION DATASETS

Method	backbone	COCO-Stuff	Cityscapes	ADE20K	Param	FPS
Segformer+ [123]	MiT-B2	41.6	81.6	47.5	30.5	25.5
K-Net+ [163]	MiT-B2	36.3	81.4	45.9	36.2	23.3
K-Net+ [163]	ResNet50	35.2	81.3	43.9	40.2	20.3
MaskFormer+ [164]	ResNet50	37.1	80.1	44.9	45.0	23.7
Mask2Former [226]	ResNet50	38.8	80.4	48.0	44.0	19.4

The results are with mIoU metric.

TABLE X
EXPERIMENT RESULTS ON INSTANCE SEGMENTATION DATASETS

Method	mAP	AP@50	AP@75	APs	APm	APl	FPS
QueryInst [159]	40.7	62.7	44.4	20.6	43.9	60.6	20.9
SOLQ [161]	39.6	61.2	43.2	19.2	44.2	60.2	18.2
K-Net+ [163]	41.5	64.2	44.4	20.3	45.2	62.4	21.2
MaskFormer+ [226]	36.9	58.4	38.9	16.3	39.2	58.1	15.6
Mask2Former [226]	43.1	65.6	46.5	22.7	46.7	64.7	13.4

We report results on the validation set using the ResNet50 backbone. The results are with mAP metric.

TABLE XI
EXPERIMENT RESULTS ON PANOPTIC SEGMENTATION DATASETS

Method	COCO	Cityscapes	ADE20K	Param	FPS
PanopticSegFormer [165]	50.1	60.2	36.4	51.0	10.3
K-Net+ [163]	49.2	59.7	35.1	40.2	21.2
MaskFormer+ [226]	47.2	53.1	36.3	45.1	13.4
Mask2Former [226]	52.1	62.3	39.2	44.0	15.6
YOSO [357]	49.2	59.3	38.2	42.2	35.2

We report results on the validation set using the ResNet50 backbone. The results are with PQ metric.

• *Results on Semantic Segmentation:* As shown in Table IX, we carry out re-benchmark experiments for SS. In particular, using the same neck architecture, Segformer+ [123] achieves the best results on COCO-Stuff and Cityscapes. Mask2Former achieves the best result on the ADE-20 k dataset.

• *Results on Instance Segmentation:* In Table X, we also explore the instance segmentation methods on COCO datasets. Under the same neck architecture, we observe gains on both K-Net and MaskFormer, compared with origin results in Table VI. Mask2Former achieve the best results.

• *Results on Panoptic Segmentation:* In Table XI, we present the re-benchmark results for PS. In particular, Mask2Former achieves the best results on all three datasets. Compared with K-Net and MaskFormer, both K-Net+ and MaskFormer+ achieve over 3-4% improvements due to the usage of stronger neck [25], which close the gaps between their original results and Mask2Former.

TABLE XII
BENCHMARK RESULTS ON VIDEO INSTANCE SEGMENTATION VALIDATION DATASET

Method	backbone	YT-VIS-2019	YT-VIS-2021	OVIS
VISTR [166]	ResNet50	36.2	-	-
IFC [170]	ResNet50	42.8	36.6	-
Seqformer [171]	ResNet50	47.4	40.5	-
Mask2Former-VIS [169]	ResNet50	46.4	40.6	-
IDOL [189]	ResNet50	49.5	43.9	30.2
VITA [237]	ResNet50	49.8	45.7	19.6
Min-VIS [188]	ResNet50	47.4	44.2	25.0
GenVIS [359]	ResNet50	51.3	46.3	35.8
Tube-Link [241]	ResNet50	52.8	47.9	29.5
CTVIS [358]	ResNet50	55.1	50.1	35.5
SeqFormer [171]	Swin-L	59.3	51.8	-
Mask2Former-VIS [169]	Swin-L	60.4	52.6	-
IDOL [189]	Swin-L	64.3	56.1	42.6
VITA [237]	Swin-L	63.0	57.5	27.7
Min-VIS [188]	Swin-L	61.6	55.3	39.4
GenVIS [359]	Swin-L	63.8	60.1	45.4
Tube-Link [241]	Swin-L	64.6	58.4	-
CTVIS [358]	Swin-L	65.6	61.2	46.9

The results are with the mAP metric.

TABLE XIII
BENCHMARK RESULTS ON VIDEO PANOPTIC SEGMENTATION VALIDATION DATASETS

Method	backbone	Cityscapes-VPS (VPQ)	KITTI-STEP (STQ)	VIP-Seg (STQ)
VIP-Deeplab [96]	ResNet50	60.6	-	-
PolyphonicFormer [128]	ResNet50	65.4	-	-
Tube-PanopticFCN [40]	ResNet50	-	-	31.5
TubeFormer [173]	Axial-ResNet	-	70.0	-
Video K-Net [172]	Swin-B	62.2	73.0	46.3
TubeLink [241]	Swin-B	-	72.0	49.4
SLOT-VPS [360]	Swin-L	63.7	-	-

The results are with VPQ and STQ metrics.

C. Main Results for Video Segmentation Datasets

• *Results On Video Semantic Segmentation:* In Table VIII, we report VSS results on VPSW. Among the methods, TubeFormer [173] achieves the best results.

• *Results on Video Instance Segmentation:* In Table XII, for VIS, CTVIS [357] achieves the best result on YT-VIS-2019 and YT-VIS-2021 using ResNet50 backbone. GenVIS [358] achieves better results on OVIS using ResNet50 backbone. When adopting Swin-L backbone, CTVIS [357] achieves the best results.

• *Results on Video Panoptic Segmentation:* In Table XIII, for VPS, SLOT-VPS [360] achieves the best results on Cityscapes-VPS. TubeLink [241] achieves the best results on the VIP-Seg dataset. Video K-Net [172] achieves the best results on the KITTI-STEP dataset.

VI. FUTURE DIRECTIONS

• *General and Unified Image/Video Segmentation:* The trend of using transformers to unify diverse segmentation tasks is gaining traction. Recent studies [26], [126], [129], [163], [172], [173], [271] have employed query-based transformers for various segmentation tasks within a unified architecture. A promising research avenue is the integration of image and video segmentation tasks in a universal model across different datasets. Such models may achieve general, robust segmentation capabilities in multiple scenarios, like detecting rare classes for improved robotic decision-making. This approach holds significant practical value, particularly in applications like robot navigation and autonomous vehicles.

• *Joint Learning with Multi-Modality:* Transformers' inherent flexibility in handling various modalities positions them as ideal for unifying vision and language tasks. Segmentation tasks, which offer pixel-level information, can enhance associated vision-language tasks such as text-image retrieval and caption generation [360]. Recent studies [237], [361], [362], [363] demonstrate the potential of a universal transformer architecture that concurrently learns segmentation alongside visual language tasks, paving the way for integrated multi-modal segmentation learning.

• *Life-Long Learning for Segmentation:* Existing segmentation methods are usually benchmarked on closed-world datasets with a set of predefined categories, i.e., assuming that the training and testing samples have the same categories and feature spaces that are known beforehand. However, realistic scenarios are usually open-world and non-stationary, where novel classes may occur continuously [364]. For example, unseen situations can occur unexpectedly in self-driving vehicles and medical diagnoses. There is a distinct gap between the performance and capabilities of existing methods in realistic and open-world settings. Thus, it is desired to gradually and continuously incorporate novel concepts into the existing knowledge base of segmentation models, making the model capable of lifelong learning.

• *Long Video Segmentation in Dynamic Scenes:* Long videos introduce several challenges [56], [57], [365]. First, existing video segmentation methods are designed to work with short video inputs and may struggle to associate instances over longer periods. Thus, new methods must incorporate long-term memory design and consider the association of instances over a more extended period. Second, maintaining segmentation mask consistency over long periods can be difficult, especially when instances move in and out of the scene. This requires new methods to incorporate temporal consistency constraints and update the segmentation masks over time. Third, heavy occlusion can occur in long videos, making it challenging to segment all instances accurately. New methods should incorporate occlusion reasoning and detection to improve segmentation accuracy. Finally, long video inputs often involve various scene inputs, which can bring domain robustness challenges for video segmentation models. New methods must incorporate domain adaptation techniques to ensure the model can handle diverse scene inputs. In short, addressing these challenges requires the development of new long video segmentation models that incorporate advanced memory design, temporal consistency constraints, occlusion reasoning, and detection techniques.

• *Generative Segmentation:* With the rise of stronger generative models, recent works [366], [367], [368] solve image segmentation problems via generative modeling, inspired by a stronger transformer decoder and high-resolution representation in the diffusion model [369]. Adopting a generative design avoids the transformer decoder and object query design, which makes the entire framework simpler. However, these generative models typically introduce a complicated training pipeline. A simpler training pipeline is needed for further research.

• *Segmentation with Visual Reasoning:* Visual reasoning [370], [371], [372], [373], [374] requires the robot to understand the connections between objects in the scene, and this understanding plays a crucial role in motion planning. Previous research has explored using segmentation results as input to visual reasoning models for various applications, such as object tracking and scene understanding. Joint segmentation and visual reasoning can be a promising direction, with the potential for mutual benefits for both segmentation and relation classification. By incorporating visual reasoning into the segmentation process, researchers can leverage the power of reasoning to improve the segmentation accuracy, while segmentation can provide better input for visual reasoning.

VII. Conclusion

This survey provides a comprehensive review of recent advancements in transformer-based visual segmentation, which, to our knowledge, is the first of its kind. The paper covers essential background knowledge and an overview of previous works before transformers and summarizes more than 120 deep-learning models for various segmentation tasks. The recent works are grouped into six categories based on the meta-architecture of the segmenter. Additionally, the paper reviews five specific subfields and reports the results of several representative segmentation methods on widely-used datasets. To ensure fair comparisons, we also re-benchmark several representative works under the same settings. Finally, we conclude by pointing out future research directions for transformer-based visual segmentation.

References

[1] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *Int. J. Comput. Vis.*, vol. 43, pp. 7–27, 2001.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[3] X. Y. Stella and J. Shi, "Multiclass spectral clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 313–319.

[4] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 54.1–54.10.

[5] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, pp. 321–331, 1988.

[6] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–16.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[12] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2393–2402.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[16] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 159.

[17] H. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 270–286.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[19] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10073–10082.

[20] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3463–3472.

[21] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[23] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[24] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16.

[26] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "MaX-DeepLab: End-to-end panoptic segmentation with mask transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5463–5474.

[27] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.

[28] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 813–824.

[29] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 16239–16248.

[30] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3D object detection with pointformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7463–7472.

[31] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[32] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, 2022, Art. no. 200.

[33] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[34] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," 2022, *arXiv:2201.05991*.

[35] J. Lahoud et al., "3D vision with transformers: A survey," 2022, *arXiv:2208.04309*.

[36] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," 2022, *arXiv:2206.06488*.

[37] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[38] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[39] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7099–7122, Jun. 2023.

[40] J. Miao et al., "Large-scale video panoptic segmentation in the wild: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21001–21011.

[41] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.

[42] R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 891–898.

[43] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[44] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, pp. 302–321, 2019.

[45] M. Cordts et al., "The CityScapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[46] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5000–5009.

[47] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–85.

[48] C. Liu, H. Ding, and X. Jiang, "GRES: Generalized referring expression segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23592–23601.

[49] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "VSPW: A large-scale dataset for video scene parsing in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4133–4143.

[50] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5187–5196.

[51] J. Qi et al., "Occluded video instance segmentation: A benchmark," *Int. J. Comput. Vis.*, vol. 130, no. 8, pp. 2022–2039, 2022.

[52] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9856–9865.

[53] M. Weber et al., "STEP: Segmenting and tracking every pixel," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021.

[54] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 Davis challenge on video object segmentation," 2017, *arXiv: 1704.00675*.

[55] N. Xu et al., "YouTube-VOS: A large-scale video object segmentation benchmark," 2018, *arXiv: 1809.03327*. [Online]. Available: http://arxiv.org/abs/1809.03327

[56] H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai, "MOSE: A new dataset for video object segmentation in complex scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 20167–20177.

[57] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy, "MeViS: A large-scale benchmark for video segmentation with motion expressions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2694–2703.

[58] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.

[59] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Trans. Image Process.*, vol. 29, pp. 3520–3533, 2020.

[60] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1743–1751.

[61] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8885–8894.

[62] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv: 1706.05587*.

[63] B. Shuai, H. Ding, T. Liu, G. Wang, and X. Jiang, "Toward achieving robust low-level and high-level scene parsing," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1378–1390, Mar. 2019.

[64] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang, "Gated fully fusion for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11418–11425.

[65] X. Li et al., "Global aggregation then local distribution for scene parsing," *IEEE Trans. Image Process.*, vol. 30, pp. 6829–6842, 2021.

[66] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.

[67] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr, "Dual graph convolutional network for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 254–264.

[68] X. Li et al., "Semantic flow for fast and accurate scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 775–793.

[69] X. Li et al., "SFNet: Faster, accurate, and domain agnostic semantic segmentation via semantic flow," 2022, *arXiv:2207.04415*.

[70] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–349.

[71] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9796–9805.

[72] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6818–6828.

[73] X. Li et al., "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 435–452.

[74] H. He et al., "Enhanced boundary learning for glass-like object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15839–15848.

[75] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[76] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[77] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 282–298.

[78] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool, "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8837–8845.

[79] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," 2017, *arXiv: 1708.02551*.

[80] K. Chen et al., "Hybrid task cascade for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4974–4983.

[81] R. Zhang, Z. Tian, C. Shen, M. You, and Y. Yan, "Mask encoding for single shot instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10223–10232.

[82] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9156–9165.

[83] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10213–10224.

[84] B. Cheng et al., "Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12472–12482.

[85] X. Chen, R. Girshick, K. He, and P. Dollár, "TensorMask: A foundation for dense object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2061–2069.

[86] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1487.

[87] Y. Xiong et al., "UPSNet: A unified panoptic segmentation network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8818–8826.

[88] Y. Li et al., "Fully convolutional networks for panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 214–223.

[89] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.

[90] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video CNNs through representation warping," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4463–4472.

[91] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–868.

[92] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4141–4150.

[93] G. Bertasius and L. Torresani, "Classifying, segmenting, and tracking object instances in video with mask propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9736–9745.

[94] Y. Fu, L. Yang, D. Liu, T. S. Huang, and H. Shi, "CompFeat: Comprehensive feature aggregation for video instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1361–1369.
[95] X. Li et al., "Improving video instance segmentation via temporal pyramid routing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6594–6601, May 2023.
[96] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "VIP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3997–4008.
[97] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9308–9316.
[98] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.
[99] P. Voigtlaender et al., "MOTS: Multi-object tracking and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7942–7951.
[100] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.
[101] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.
[102] B. Yang et al., "Learning object bounding boxes for 3D instance segmentation on point clouds," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 6737–6746.
[103] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, "GSPN: Generative shape proposal network for 3D instance segmentation in point cloud," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3947–3956.
[104] W. Wang, R. Yu, Q. Huang, and U. Neumann, "SGPN: Similarity group proposal network for 3D point cloud instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2569–2578.
[105] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "PointGroup: Dual-set point grouping for 3D instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4866–4875.
[106] J. Mao, X. Wang, and H. Li, "Interpolated convolutional networks for 3D point cloud understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1578–1587.
[107] Q. Hu et al., "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11105–11114.
[108] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "(AF)2-S3Net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12547–12556.
[109] Z. Zhou, Y. Zhang, and H. Foroosh, "Panoptic-PolarNet: Proposal-free LiDAR point cloud panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13194–13203.
[110] S. Xu, R. Wan, M. Ye, X. Zou, and T. Cao, "Sparse cross-scale attention network for efficient LiDAR panoptic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2920–2928.
[111] F. Hong, H. Zhou, X. Zhu, H. Li, and Z. Liu, "LiDAR-based panoptic segmentation via dynamic shifting network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13090–13099.
[112] M. Aygun et al., "4D panoptic LiDAR segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5527–5537.
[113] X. Zhu et al., "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9939–9948.
[114] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A2-Nets: Double attention networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 350–359.
[115] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[116] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.
[117] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[118] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.
[119] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.
[120] F. Li et al., "Lite DETR: An interleaved multi-scale encoder for efficient DETR," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18558–18567.
[121] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, pp. 83–97, 1955.
[122] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[123] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[124] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[125] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[126] X. Li et al., "OMG-seg: Is one model good enough for all segmentation?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27948–27959.
[127] A. Athar, A. Hermans, J. Luiten, D. Ramanan, and B. Leibe, "TarViS: A unified approach for target-based video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18738–18748.
[128] H. Yuan et al., "PolyphonicFormer: Unified query learning for depth-aware video panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 582–599.
[129] J. Wu, Y. Jiang, B. Yan, H. Lu, Z. Yuan, and P. Luo, "UniRef: Segment every reference object in spatial and temporal spaces," 2023, *arXiv:2312.15715*.
[130] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[131] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6804–6815.
[132] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4794–4804.
[133] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7277–7286.
[134] A. Ali et al., "XCiT: Cross-covariance image transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 20014–20027.
[135] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 548–558.
[136] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 347–356.
[137] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 323–339.
[138] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9961–9970.
[139] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12165–12175.
[140] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9355–9366.
[141] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 22–31.
[142] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive Bias," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 28522–28535.
[143] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.

[144] Q. Han et al., "On the connection between local attention and dynamic depth-wise convolution," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–25.

[145] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 1140–1156.

[146] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10809–10819.

[147] J. Dai et al., "Demystify transformers & convolutions in modern image deep networks," 2022, *arXiv:2211.05781*.

[148] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[149] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[150] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9620–9629.

[151] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 14668–14678.

[152] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14648–14658.

[153] Y. Gandelsman, Y. Sun, X. Chen, and A. A. Efros, "Test-time training with masked autoencoders," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 2130.

[154] R. Hu, S. Debnath, S. Xie, and X. Chen, "Exploring long-sequence masked autoencoders," 2022, *arXiv:2210.07224*.

[155] P. Gao, T. Ma, H. Li, J. Dai, and Y. Qiao, "ConvMAE: Masked convolution meets masked autoencoders," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 35632–35644.

[156] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[157] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," 2022, *arXiv:2212.00794*.

[158] P. Sun et al., "SparseR-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14449–14458.

[159] Y. Fang et al., "Instances as queries," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6890–6899.

[160] J. Hu et al., "ISTR: End-to-end instance segmentation via transformers," 2021, *arXiv:2105.00637*.

[161] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, "SOLQ: Segmenting objects by learning queries," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 21898–21909.

[162] H. He et al., "BoundarySqueeze: Image segmentation as boundary squeezing," 2021, *arXiv:2105.11668*.

[163] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: Towards unified image segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 10326–10338.

[164] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 17864–17875.

[165] Z. Li et al., "Panoptic SegFormer: Delving deeper into panoptic segmentation with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1270–1279.

[166] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8741–8750.

[167] Q. Zhou et al., "TransVOD: End-to-end video object detection with spatial-temporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7853–7869, Jun. 2023.

[168] S. Yang et al., "Temporally efficient vision transformer for video instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2875–2885.

[169] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2Former for video instance segmentation," 2021, *arXiv:2112.10764*.

[170] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Video instance segmentation using inter-frame communication transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 13352–13363.

[171] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "SeqFormer: Sequential transformer for video instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 553–569.

[172] X. Li et al., "Video K-Net: A simple, strong, and unified baseline for video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18825–18835.

[173] D. Kim et al., "TubeFormer-DeepLab: Video mask transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13904–13914.

[174] D. Meng et al., "Conditional DETR for fast training convergence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3631–3640.

[175] X. Chen, F. Wei, G. Zeng, and J. Wang, "Conditional DETR v2: Efficient detection transformer with box queries," 2022, *arXiv:2207.08914*.

[176] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor DETR: Query design for transformer-based detector," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2567–2575.

[177] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–20.

[178] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13609–13617.

[179] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–19.

[180] F. Li et al., "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3041–3050.

[181] W. Wang, J. Liang, and D. Liu, "Learning equivariant segmentation with instance-unique querying," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 932.

[182] D. Jia et al., "DETRs with hybrid matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19702–19712.

[183] Q. Chen et al., "Group DETR: Fast DETR training with group-wise one-to-many assignment," 2022, *arXiv:2207.13085*.

[184] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," 2022, *arXiv:2211.12860*.

[185] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8834–8844.

[186] P. Sun et al., "TransTrack: Multiple-object tracking with transformer," 2020, *arXiv: 2012.15460*.

[187] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 659–675.

[188] D.-A. Huang, Z. Yu, and A. Anandkumar, "MinVIS: A minimal video instance segmentation framework without video-based training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 31265–31277.

[189] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 588–605.

[190] X. Li, S. Xu, Y. Yang, G. Cheng, Y. Tong, and D. Tao, "Panoptic-PartFormer: Learning a unified model for panoptic part segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 729–747.

[191] N. Gao et al., "PanopticDepth: A unified framework for depth-aware panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1622–1632.

[192] S. Xu, X. Li, J. Wang, G. Cheng, Y. Tong, and D. Tao, "Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 545–563.

[193] Y. Xu et al., "Multi-task learning with multi-query transformer for dense prediction," 2022, *arXiv:2205.14354*.

[194] H. Ye and D. Xu, "Inverted pyramid multi-task transformer for dense scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 514–530.

[195] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 16301–16310.

[196] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "LAVT: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18134–18144.

[197] N. Kim, D. Kim, C. Lan, W. Zeng, and S. Kwak, "ReSTR: Convolution-free referring image segmentation using transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18124–18133.

[198] Z. Wang et al., "CRIS: CLIP-driven referring image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11676–11685.
[199] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4975–4985.
[200] Z. Ding, T. Hui, J. Huang, X. Wei, J. Han, and S. Liu, "Language-bridged spatial-temporal interaction for referring video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4954–4963.
[201] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, "Towards robust referring image segmentation," 2022, *arXiv:2209.09554*.
[202] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4964–4974.
[203] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 21984–21996.
[204] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 2491–2502.
[205] G. Park, S. Son, J. Yoo, S. Kim, and N. Kwak, "MatteFormer: Transformer-based image matting via prior-tokens," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11686–11696.
[206] B. Shi et al., "A transformer-based decoder for semantic segmentation with multi-level context mining," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 624–639.
[207] F. Lin, Z. Liang, J. He, M. Zheng, S. Tian, and K. Chen, "StructToken: Rethinking semantic segmentation with structural prior," 2022, *arXiv:2203.12612*.
[208] Y. Yu, J. Yuan, G. Mittal, L. Fuxin, and M. Chen, "BATMAN: Bilateral attention transformer in motion-appearance neighboring space for video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 612–629.
[209] S. Jiao et al., "Mask matching transformer for few-shot segmentation," 2022, *arXiv:2301.01208*.
[210] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11999–12009.
[211] S. Chen, E. Xie, C. Ge, R. Chen, D. Liang, and P. Luo, "CycleMLP: A MLP-like architecture for dense prediction," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–21.
[212] I. O. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.
[213] J. Guo et al., "Hire-MLP: Vision MLP via hierarchical rearrangement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 816–826.
[214] X. Chen et al., "Context autoencoder for self-supervised representation learning," *Int. J. Comput. Vis.*, vol. 132, pp. 208–223, 2024.
[215] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, and Z. Yuan, "Designing BERT for convolutional networks: Sparse and hierarchical masked modeling," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–16.
[216] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640.
[217] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[218] H. Li et al., "Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks," 2022, *arXiv:2211.09808*.
[219] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022, *arXiv:2203.12602*.
[220] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," 2022, *arXiv:2205.09113*.
[221] Z. Liu et al., "Video swin transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3192–3201.
[222] H. Ding, C. Liu, S. Wang, and X. Jiang, "VLT: Vision-language transformer and query generation for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7900–7916, Jun. 2023.
[223] X. Yu, D. Shi, X. Wei, Y. Ren, T. Ye, and W. Tan, "SOIT: Segmenting objects with instance-aware transformers," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3188–3196.
[224] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 7242–7252.
[225] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1280–1289.
[226] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[227] Q. Yu et al., "CMT-DeepLab: Clustering mask transformers for panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2550–2560.
[228] Q. Yu et al., "k-means mask transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 288–307.
[229] T. Cheng et al., "Sparse instance activation for real-time instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4423–4432.
[230] G. Zhang, Z. Luo, Y. Yu, K. Cui, and S. Lu, "Accelerating DETR convergence via semantic-aligned matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 939–948.
[231] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3601–3610.
[232] Z. Gao, L. Wang, B. Han, and S. Guo, "AdaMixer: A fast-converging query-based object detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5354–5363.
[233] M. Zheng et al., "End-to-end object detection with adaptive clustering transformer," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 226–236.
[234] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic DETR: End-to-end object detection with dynamic attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2968–2977.
[235] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse DETR: Efficient end-to-end object detection with learnable sparsity," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–23.
[236] M. Heo, S. Hwang, S. W. Oh, J.-Y. Lee, and S. J. Kim, "VITA: Video instance segmentation via object token association," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23109–23120.
[237] X. Zou et al., "Generalized decoding for pixel, image and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15116–15127.
[238] Z. Cai et al., "X-DETR: A versatile architecture for instance-wise vision-language tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 290–308.
[239] I. Shin et al., "Video-kMaX: A simple unified approach for online and near-online video panoptic segmentation," 2023, *arXiv:2304.04694*.
[240] X. Li, H. Yuan, W. Zhang, J. Pang, G. Cheng, and C. C. Loy, "Tube-link: A flexible cross tube baseline for universal video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 13877–13887.
[241] Z. Yao, J. Ai, B. Li, and C. Zhang, "Efficient DETR: Improving end-to-end object detector with dense prior," 2021, *arXiv:2104.01318*.
[242] H. Zhang et al., "MP-Former: Mask-piloted transformer for image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18074–18083.
[243] W. Wang, J. Zhang, Y. Cao, Y. Shen, and D. Tao, "Towards data-efficient detection transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 88–105.
[244] X. Li et al., "PanopticPartFormer: A unified and decoupled view for panoptic part segmentation," 2023, *arXiv:2301.00954*.
[245] S. He and H. Ding, "Decoupling static and hierarchical motion perception for referring video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13332–13341.
[246] S. He, H. Ding, and W. Jiang, "Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19498–19507.
[247] C. Liu, X. Li, and H. Ding, "Referring image editing: Object-level image editing via referring expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13128–13138.
[248] C. Liu, X. Jiang, and H. Ding, "Instance-specific feature propagation for referring segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 3657–3667, 2022.
[249] G. Luo et al., "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10031–10040.

[250] D. Wu, X. Dong, L. Shao, and J. Shen, "Multi-level representation learning with semantic alignment for referring video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4986–4995.

[251] A. Kamath, M. Singh, Y. LeCun, I. Misra, G. Synnaeve, and N. Carion, "MDETR–modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1760–1770.

[252] L. Cao, Y. Guo, Y. Yuan, and Q. Jin, "Prototype as query for few shot semantic segmentation," 2022, *arXiv:2211.14764*.

[253] H. Ding, H. Zhang, C. Liu, and X. Jiang, "Deep interactive image matting with feature propagation," *IEEE Trans. Image Process.*, vol. 31, pp. 2421–2432, 2022.

[254] Y. Han et al., "Reference twice: A simple and unified baseline for few-shot instance segmentation," 2023, *arXiv:2301.01156*.

[255] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, pp. 187–199, 2021.

[256] X. Lai et al., "Stratified transformer for 3D point cloud segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8490–8499.

[257] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19291–19300.

[258] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 604–621.

[259] R. Zhang et al., "Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training," 2022, *arXiv:2205.14401*.

[260] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3D for 3D semantic instance segmentation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 8216–8223.

[261] J. Sun, C. Qing, J. Tan, and X. Xu, "Superpoint transformer for 3D scene instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2393–2401.

[262] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4558–4567.

[263] S. Su et al., "PUPS: Point cloud unified panoptic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2339–2347.

[264] J. Behley et al., "A dataset for semantic segmentation of point cloud sequences," 2019, *arXiv: 1904.01416*.

[265] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16795–16804.

[266] R. Zhang et al., "Tip-Adapter: Training-free CLIP-adapter for better vision-language modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 493–510.

[267] Z. Lin et al., "Frozen CLIP models are efficient video learners," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 388–404.

[268] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–20.

[269] Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18061–18070.

[270] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7076–7086.

[271] J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi, "OneFormer: One transformer to rule universal image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2989–2998.

[272] A. Kirillov et al., "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[273] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14393–14402.

[274] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–20.

[275] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 350–368.

[276] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary DETR with conditional matching," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 106–122.

[277] S. He, H. Ding, and W. Jiang, "Primitive generation and semantic-related alignment for universal zero-shot segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11238–11247.

[278] H. Zhou et al., "Rethinking evaluation metrics of open-vocabulary segmentaion," 2023, *arXiv:2311.03352*.

[279] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "F-VLM: Open-vocabulary object detection upon frozen vision and language models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–20.

[280] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang, "Class-agnostic object detection with multi-modal transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 512–531.

[281] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 540–557.

[282] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–13.

[283] J. Wu et al., "Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 21881–21891.

[284] J. Qin et al., "FreeSeg: Unified, universal and open-vocabulary image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19446–19455.

[285] W. Wang, M. Feiszli, H. Wang, and D. Tran, "Unidentified video objects: A benchmark for dense, open-world segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10756–10765.

[286] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2955–2966.

[287] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: Open-world detection transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9225–9234.

[288] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2945–2954.

[289] Z. Liu et al., "Open compound domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12403–12412.

[290] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4084–4094.

[291] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9914–9925.

[292] L. Hoyer, D. Dai, and L. Van Gool, "HRDA: Context-aware high-resolution domain-adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 372–391.

[293] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "MIC: Masked image consistency for context-enhanced domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11721–11732.

[294] W. Wang et al., "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1730–1738.

[295] J. Zhang, J. Huang, Z. Luo, G. Zhang, and S. Lu, "DA-DETR: Domain adaptive detection transformer by hybrid attention," 2021, *arXiv:2103.17084*.

[296] J. Yu et al., "MTTrans: Cross-domain object detection with mean teacher transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 629–645.

[297] B. Xie, S. Li, M. Li, C. H. Liu, G. Huang, and G. Wang, "SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9004–9021, Jul. 2023.

[298] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "MSeg: A composite dataset for multi-domain semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2876–2885.

[299] W. Yin, Y. Liu, C. Shen, A. v. d. Hengel, and B. Sun, "The devil is in the labels: Semantic segmentation from sentences," 2022, *arXiv:2202.02002*.

[300] Z. Qiang et al., "LMSeg: Language-guided multi-dataset segmentation," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–12.

[301] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7561–7570.

[302] L. Meng et al., "Detection hub: Unifying object detection datasets via query adaptation on language embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11402–11411.
[303] L. Hoyer, D. Dai, and L. Van Gool, "Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation," 2023, *arXiv:2304.13615*.
[304] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. H. Lee, "Style-hallucinated dual consistency learning: A unified framework for visual domain generalization," *Int. J. Comput. Vis.*, vol. 132, pp. 837–853, 2024.
[305] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4300–4309.
[306] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12272–12281.
[307] S. Rossetti, D. Zappia, M. Sanzari, M. Schaerf, and F. Pirri, "Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 446–463.
[308] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 6582–6593.
[309] S. Lan et al., "DiscoBox: Weakly supervised instance segmentation and semantic correspondence from box supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3386–3396.
[310] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5443–5452.
[311] J. Xu et al., "GroupViT: Semantic segmentation emerges from text supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18113–18123.
[312] M. Yi, Q. Cui, H. Wu, C. Yang, O. Yoshie, and H. Lu, "A simple framework for text-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7071–7080.
[313] L. Ke, M. Danelljan, H. Ding, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask-free video instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22857–22866.
[314] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10032–10042.
[315] O. Siméoni et al., "Localizing objects with self-supervised transformers and no labels," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 310–320.
[316] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–26.
[317] G. Shin, W. Xie, and S. Albanie, "ReCo: Retrieve and co-segment for zero-shot transfer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 33754–33767.
[318] W. Van Gansbeke, S. Vandenhende, and L. Van Gool, "Discovering object masks with transformers for unsupervised semantic segmentation," 2022, *arXiv:2206.06363*.
[319] X. Wang et al., "FreeSOLO: Learning to segment objects without annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14156–14166.
[320] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3124–3134.
[321] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.
[322] M. Maaz et al., "EdgeNeXt: Efficiently amalgamated CNN-transformer architecture for mobile vision applications," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2022, pp. 3–20.
[323] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–26.
[324] J. Zhang et al., "Rethinking mobile block for efficient neural models," 2023, *arXiv:2301.01146*.
[325] W. Liang et al., "Expediting large-scale vision transformer for dense prediction without fine-tuning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 35462–35477.
[326] C. Zhou, X. Li, C. C. Loy, and B. Dai, "EdgeSAM: Prompt-in-the-loop distillation for on-device deployment of SAM," 2023, *arXiv:2312.06660*.
[327] S. Xu et al., "RAP-SAM: Towards real-time all-purpose segment anything," 2024, *arXiv:2401.10228*.
[328] W. Zhang et al., "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12073–12083.
[329] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "SeaFormer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–19.
[330] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8887–8896.
[331] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4402–4411.
[332] L. Ke, H. Ding, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, "Video mask transfiner for high-quality video instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 731–747.
[333] Q. Liu, Z. Xu, G. Bertasius, and M. Niethammer, "SimpleClick: Interactive image segmentation with simple vision transformers," 2022, *arXiv:2210.11006*.
[334] M. Wang, H. Ding, J. H. Liew, J. Liu, Y. Zhao, and Y. Wei, "SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 3492.
[335] Y. Song, Q. Zhou, X. Li, D.-P. Fan, X. Lu, and L. Ma, "BA-SAM: Scalable bias-mode attention mask for segment anything model," 2024, *arXiv:2401.02317*.
[336] Q. Wen, J. Yang, X. Yang, and K. Liang, "PatchDCT: Patch refinement for high quality instance segmentation," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–15.
[337] X. Shen et al., "DCT-mask: Discrete cosine transform mask representation for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8720–8729.
[338] L. Qi et al., "Open world entity segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8743–8756, Jul. 2023.
[339] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9225–9234.
[340] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 640–658.
[341] K. Park, S. Woo, S. W. Oh, I. S. Kweon, and J.-Y. Lee, "Per-clip video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1342–1351.
[342] J. Wang et al., "Look before you match: Instance understanding matters in video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2268–2278.
[343] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2015, pp. 234–241.
[344] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.
[345] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[346] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2022, pp. 205–218.
[347] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, Springer, 2021, pp. 14–24.
[348] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1748–1758.
[349] H. Zhang et al., "A simple framework for open-vocabulary segmentation and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 1020–1031.
[350] X. Zou et al., "Segment everything everywhere all at once," 2023, *arXiv:2304.06718*.
[351] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, and T. Darrell, "Hierarchical open-vocabulary universal image segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 936.

[352] S. Shao et al., "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8429–8438.
[353] J. Liang, T. Zhou, D. Liu, and W. Wang, "CLUSTSEG: Clustering for universal segmentation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 20787–20809.
[354] G. Sun, Y. Liu, H. Ding, T. Probst, and L. Van Gool, "Coarse-to-fine feature mining for video semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3116–3127.
[355] G. Sun, Y. Liu, H. Tang, A. Chhatkuli, L. Zhang, and L. Van Gool, "Mining relations among cross-frame affinities for video semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 522–539.
[356] J. Hu, L. Huang, T. Ren, S. Zhang, R. Ji, and L. Cao, "You only segment once: Towards real-time panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17819–17829.
[357] K. Ying et al., "CTVIS: Consistent training for online video instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 899–908.
[358] M. Heo et al., "A generalized framework for video instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14623–14632.
[359] Y. Zhou et al., "Slot-VPS: Object-centric representation learning for video panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3083–3093.
[360] H. Ding, S. Cohen, B. Price, and X. Jiang, "PhraseClick: Toward achieving flexible interactive segmentation by phrase and click," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 417–435.
[361] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "UNIFIED-IO: A unified model for vision, language, and multi-modal tasks," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–34.
[362] L. Qi et al., "Generalizable entity grounding via assistance of large language model," 2024, *arXiv:2402.02555*.
[363] H. Yuan, X. Li, C. Zhou, Y. Li, K. Chen, and C. C. Loy, "Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively," 2024, *arXiv:2401.02955*.
[364] H. Zhang and H. Ding, "Prototypical matching and open set rejection for zero-shot semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6954–6963.
[365] Y. Zhou, T. Zhang, S. Ji, S. Yan, and X. Li, "DVIS-DAQ: Improving video segmentation via dynamic anchor queries," 2024, *arXiv:2404.00086*.
[366] T. Chen, L. Li, S. Saxena, G. Hinton, and D. J. Fleet, "A generalist framework for panoptic segmentation of images and videos," 2022, *arXiv:2210.06366*.
[367] C. Wang et al., "Explore in-context segmentation via latent diffusion models," 2024, *arXiv:2403.09616*.
[368] J. Xie, W. Li, X. Li, Z. Liu, Y. S. Ong, and C. C. Loy, "MosaicFusion: Diffusion models as data augmenters for large vocabulary instance segmentation," 2023, *arXiv:2309.13042*.
[369] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.
[370] J. Johnson et al., "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3668–3678.
[371] J. Yang et al., "Panoptic video scene graph generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18675–18685.
[372] J. Yang, Y. Z. Ang, Z. Guo, K. Zhou, W. Zhang, and Z. Liu, "Panoptic scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 178–196.
[373] J. Yang et al., "4D panoptic scene graph generation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 3053.
[374] J. Wang, Z. Wen, X. Li, Z. Guo, J. Yang, and Z. Liu, "Pair then relation: Pair-net for panoptic scene graph generation," 2023, *arXiv:2307.08699*.

Xiangtai Li received the PhD degree from Peking University, in 2022. He currently works as a research fellow with S-Lab, a member of the Multimedia Laboratory of NTU (MMLab@NTU) with Nanyang Technological University. His research interests include computer vision and machine learning with a focus on scene understanding, segmentation, and video understanding. Several of his works have been published in top-tier conferences and journals. He serves as the reviewer for top-tier conferences and journals, including CVPR, ICML, ECCV, ICCV, ICLR, NeurIPS, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *International Journal of Computer Vision*.

Henghui Ding received the BE degree from Xi'an Jiaotong University, in 2016, and the PhD degree from Nanyang Technological University (NTU), Singapore, in 2020. He was research scientist with ByteDance, postdoctoral researcher with ETH and NTU. He is currently a tenure-track professor with Fudan University. He serves as associate editors for *IET Computer Vision* and *Visual Intelligence*. He serves/served as area chairs for CVPR'24, NeurIPS'24, ACM MM'24, BMVC'24, Senior Program Committee members for AAAI'(22-25) and IJCAI'(23-24). His research interests include computer vision and machine learning.

Haobo Yuan received the master's degree from Wuhan University, in 2023. He is a research associate with S-Lab of Nanyang Technological University. His research interests include computer vision and machine learning. He has published several works in top-tier conferences and journals.

Wenwei Zhang received the BEng degree from Computer Science School, Wuhan University, in 2019. He is currently working toward the final-year PhD degree with Nanyang Technological University. He is a member of the Multimedia Laboratory of NTU (MMLab@NTU), affiliated with the NTU S-Lab, supervised by Prof. Chen Change Loy. His research interests focus on 3D/2D object detection and segmentation. He also devotes himself to OpenMMLab projects, an open source projects for academic research and industrial applications, covering a wide range of computer vision tasks. He is a core maintainer of MMDetection, MMDetection3D, and MMCV in the OpenMMLab projects.

Jiangmiao Pang is a research scientist at Shanghai AI Laboratory. He received his PhD degree from Zhejiang University in 2021. His research interests cover robotics and multimodal learning. He has published over 30 papers with 1000+ citations in top-tier conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, TPAMI, etc. He leads the development of OpenRobotLab in Shanghai AI Laboratory.

Guangliang Cheng received the PhD degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing. He is currently a reader with the Department of Computer Science, University of Liverpool. Previously, He was a vice research director with SenseTime. Before that, he was a postdoc researcher with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China. His research interests include computer vision, autonomous driving and robotic perception, scene understanding, domain adaptation, and remote sensing image processing.

Kai Chen received the BEng degree from Tsinghua University, in 2015, and the PhD degree from the Chinese University of Hong Kong, in 2019. He is a research scientist with Shanghai AI Laboratory. His research interests cover computer vision and large language models. He has published more than 30 papers with more than 9000 citations in top-tier conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, etc. He also leads the development of OpenMMLab, an open-source computer vision algorithm platform.

Ziwei Liu is currently a Nanyang assistant professor with Nanyang Technological University, Singapore. His research revolves around computer vision, machine learning, and computer graphics. He has published extensively in top-tier conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *ACM Transactions on Graphics*, and *Nature Machine Intelligence*. He is the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, CVPR Best Paper Award Candidate, and MIT Technology Review Innovators under 35 Asia Pacific. He serves as an area chair of CVPR, ICCV, NeurIPS, and ICLR, as well as an associate editor of *International Journal of Computer Vision*.

Chen Change Loy (Senior Member, IEEE) received the PhD degree in computer science from the Queen Mary University of London, in 2010. He is a professor of computer vision with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is also an adjunct associate professor with The Chinese University of Hong Kong. He is the Lab director of MMLab@NTU and co-associate director of S-Lab. Prior to joining NTU, he served as a research assistant professor with the MMLab of The Chinese University of Hong Kong, from 2013 to 2018. He serves as an associate editor of the *International Journal of Computer Vision (IJCV)*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* and *Computer Vision and Image Understanding (CVIU)*. He also serves/served as an area chair of top conferences, such as ICCV, CVPR, ECCV, ICLR, and NeurIPS.