

Visual Semantic Segmentation Based on Few/Zero-Shot Learning: An Overview

Wenqi Ren, Yang Tang*, *Senior Member, IEEE*, Qiyu Sun, Chaoqiang Zhao, Qing-Long Han*, *Fellow, IEEE*

Abstract—Visual semantic segmentation aims at separating a visual sample into diverse blocks with specific semantic attributes and identifying the category for each block, and it plays a crucial role in environmental perception. Conventional learning-based visual semantic segmentation approaches count heavily on large-scale training data with dense annotations and consistently fail to estimate accurate semantic labels for unseen categories. This obstruction spurs a craze for studying visual semantic segmentation with the assistance of few/zero-shot learning. The emergence and rapid progress of few/zero-shot visual semantic segmentation make it possible to learn unseen-category from a few labeled or zero-labeled samples, which advances the extension to practical applications. Therefore, this paper focuses on the recently published few/zero-shot visual semantic segmentation methods varying from 2D to 3D space and explores the commonalities and discrepancies of technical settlements under different segmentation circumstances. Specifically, the preliminaries on few/zero-shot visual semantic segmentation, including the problem definitions, typical datasets, and technical remedies, are briefly reviewed and discussed. Moreover, three typical instantiations are involved to uncover the interactions of few/zero-shot learning with visual semantic segmentation, including image semantic segmentation, video object segmentation, and 3D segmentation. Finally, the future challenges of few/zero-shot visual semantic segmentation are discussed.

Index Terms—Few-shot learning, zero-shot learning, low-shot learning, semantic segmentation, computer vision, deep learning.

I. INTRODUCTION

VISUAL semantic segmentation targets at fine-grained classification for collected samples, such as images, videos, and 3D meshes. It has gradually drawn research interests in virtue of the extensive applications to self-driving [1], [2], medical diagnosis [3], [4], remote sensing [5], and so on. As three typical representatives in visual semantic segmentation, image semantic segmentation (ISS) [6], [7], video object segmentation (VOS) [8], [9] and 3D segmentation

This work was supported by National Key Research and Development Program of China (2021YFB1714300), the National Natural Science Foundation of China (62233005), and in part by the CNPC Innovation Fund under Grant 2021D002-0902, Fundamental Research Funds for the Central Universities and Shanghai AI Lab. Qiyu Sun is sponsored by Shanghai Gaofeng and Gaoyuan Project for University Academic Program Development. (*Corresponding author: Yang Tang and Qing-Long Han.)

Wenqi Ren, Yang Tang, Qiyu Sun and Chaoqiang Zhao are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China. Qiyu Sun is with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, China (e-mail: wenqiren9801@163.com; yangtang@ecust.edu.cn; qysun291@163.com; zhaocqilc@gmail.com).

Qing-Long Han is with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia (e-mail: qhan@swin.edu.au).

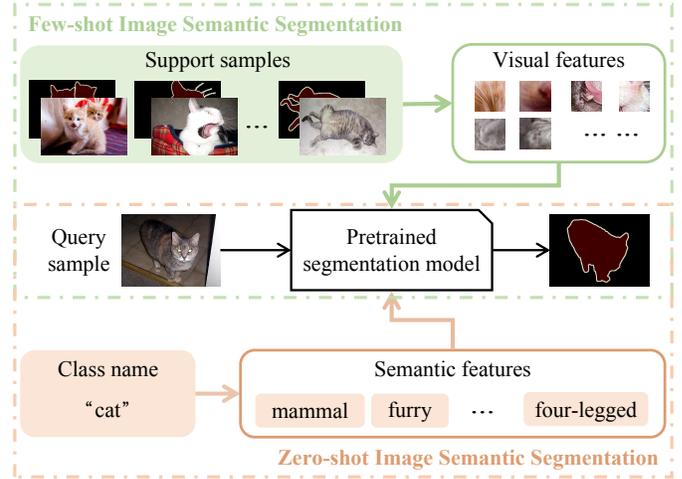


Fig. 1. An example of few/zero-shot visual semantic segmentation, where the segmentation model is pre-trained on the categories except “cat”. The visual samples are selected from the PASCAL VOC [13].

(3DS) [10], [11] have been well-developed. Among them, ISS and VOS attempt to automatically assign a label for each pixel located in 2D images, while 3DS endeavors to allocate annotations to target shapes or objects for given 3D samples, such as point clouds [10], [11] and 3D meshes [12].

Thanks to the advance in deep learning, the conventional supervised learning-based visual semantic segmentation approaches have made considerable breakthroughs in both real-time performance and prediction accuracy [118]–[121]. They learn to fit the distribution of specific classes from a great deal of training samples and strive to segment them on test samples. Whereas, these visual semantic segmentation algorithms still confront various challenges. Firstly, a great deal of training data with fine-grained annotations is desperately required for these segmentation models to learn a category of interest, where the annotations are expensive to obtain. Particularly, in 3D point cloud segmentation, the order of millions is commonly reached by the number of captured point clouds [122], leading to laborious labeling processes. Moreover, the structure of point clouds is irregular, which further attaches enormous challenges to manual annotations. Secondly, these algorithms rely heavily on an implicit assumption that the test categories must be the same as the training ones [123]–[125]. Consequently, when dealing with unseen categories after training, these segmentation models always suffer from severe performance degradation, which hinders their flexible application in dynamic scenarios. Although the methods based

TABLE I
A SUMMARY OF TECHNICAL SOLUTIONS FOR FEW/ZERO-SHOT VISUAL SEMANTIC SEGMENTATION TASKS

Space	Samples	Scenarios	Profiles	Solutions	Typical datasets	Literatures
2D	Images	Few-shot ISS	Predict pixel-wise labels for unseen categories with a few annotated images	Metric Parameter Fine-tune Memory	PASCAL-5 ⁱ [14], COCO-20 ² [15]	[16]–[37] [14], [38], [39] [40]–[42] [43], [44]
		Zero-shot ISS	Predict pixel-wise labels for unseen categories with zero annotated images	Metric Generative	PASCAL VOC [13], PASCAL Context [45]	[46]–[51] [52]–[55]
	Videos	Few-shot VOS	Segment the object specified in the first frame in the remaining frames	Metric Fine-tune Memory	Youtube-VOS [56], DAVIS 2016 [57]	[58]–[67] [68]–[72] [73]–[84]
		Zero-shot VOS	Segment the primary object in the video sequence without annotations	Metric Fine-tune Memory	Youtube-VOS [56], DAVIS 2016 [57]	[85]–[98] [99] [100]–[103]
3D	Point clouds /3D meshes	Few-shot 3DS	Assign labels for unseen categories with a few annotated 3D samples	Metric Parameter Fine-tune	ShapeNet Part [104], ScanNet [105]	[106]–[109] [110] [111]–[114]
		Zero-shot 3DS	Assign labels for unseen categories with zero annotated 3D samples	Generative	ScanNet [105], S3DIS [115]	[116], [117]

¹ Each technical solution is represented by its first word.

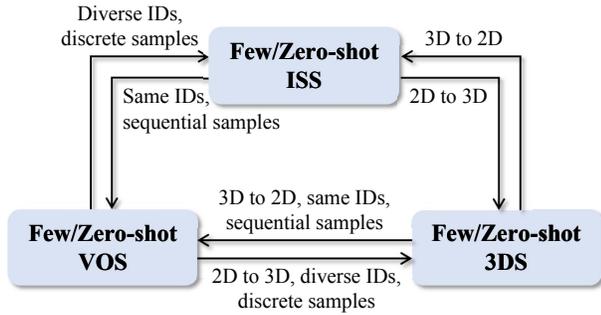


Fig. 2. The relationship among ISS, VOS, and 3DS under both few-shot and zero-shot cases.

on weakly-supervised learning [126]–[128] alleviate the data hunger significantly, they still require that the categories encountered at test-time must have been seen in training, leading to the problem of cross-class adaptation failing to be addressed. These limitations encourage the employment of few-shot learning and zero-shot learning [129]–[131] on visual semantic segmentation, which promotes the skills of segmentation models in dealing with unseen categories, even though they have only been exposed to a few labeled, or even unlabeled, samples, as shown in Fig. 1.

In recent years, a significant number of meaningful and valuable few/zero-shot visual semantic segmentation studies [132]–[134] have been spawned, which eliminates the barriers on the cross-class adaptation with a limited number of annotated samples. To summarize these approaches, some related and well-written surveys were published [132]–[134]. Compared with these previous efforts [132]–[134], this paper creatively summarizes relevant studies from the perspective of problem settings and technical solutions of few/zero-shot learning and strengthens the relationship between different segmentation methods, including ISS, VOS, and 3DS, in both few-shot and zero-shot scenarios. In addition, this paper provides some discussions on open challenges that few/zero-

shot learning brought to visual semantic segmentation, such as cross-domain few/zero-shot segmentation and generalized few/zero-shot segmentation. In summary, the main contributions of this paper are as follows:

- The comparison among the problem settings of different few/zero-shot visual segmentation tasks and a summary of technical solutions are provided.
- The advancements of few/zero-shot visual semantic segmentation are reviewed and the dissimilarities of technical solutions in different segmentation tasks are specified.
- The open challenges that involve data, algorithms, and applications for few/zero-shot visual semantic segmentation are discussed to provide the enlightenment to follow-up researchers.

The rest of this paper is organized as follows. Section II provides the comparison on the problem settings and typical datasets of different few/zero-shot visual segmentation tasks and gives a summary of technical solutions. Section III, Section IV and Section V concentrate on the advancements of ISS, VOS and 3DS in both few-shot and zero-shot circumstances, respectively, and illustrate the technical settlements in diverse segmentation scenarios. Section VI analyzes the open challenges and applications for few/zero-shot visual semantic segmentation.

II. PRELIMINARIES

Few-shot learning expects to learn unseen categories from a few labeled training samples [129], [131], [137]. When the annotated samples are unavailable, the few-shot learning problem turns into a zero-shot learning problem [129], [137]. Therefore, zero-shot learning is the boundary case of few-shot learning [129], [138], [139]. This section attempts to provide detailed definitions, typical datasets, and technical remedies for few/zero-shot visual semantic segmentation.

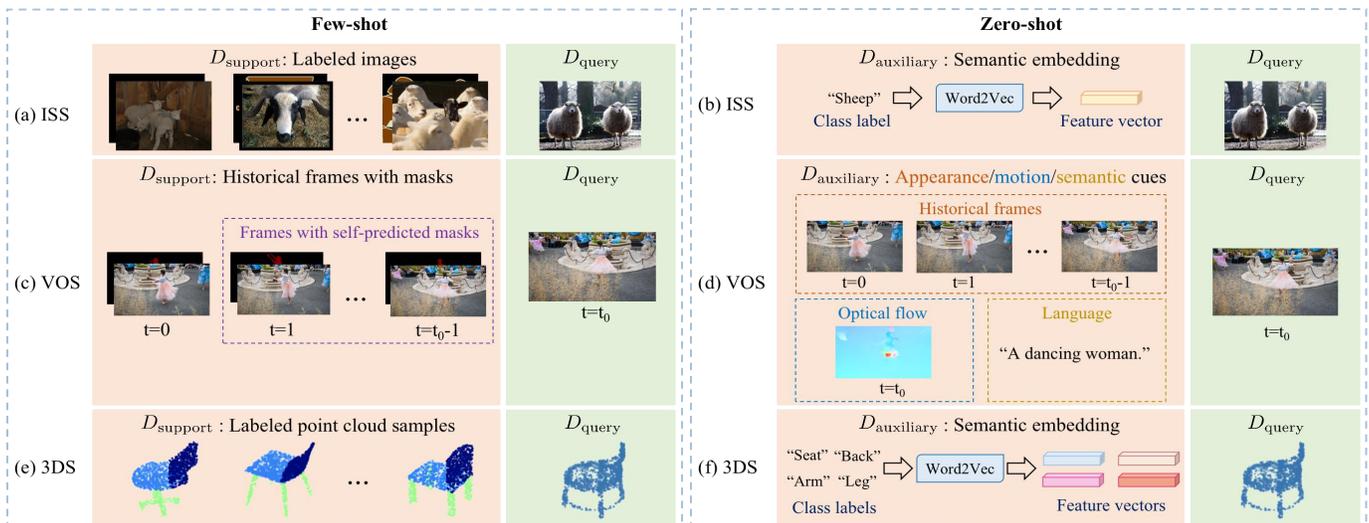


Fig. 3. The conventional settings of different few/zero-shot visual segmentation tasks. The ISS samples are selected from the PASCAL VOC [13], the VOS samples are picked from DAVIS 2016 [57], and the 3DS exemplars are sampled from ShapeNet Part [104]. The Word2Vec [135], [136] is the operation to map a class label as a feature vector, which is called word embedding [130] and represents the semantic attributes of the corresponding category.

A. Problem Definition

As shown in Table I, few/zero-shot ISS methods [16], [17], [55] strive to learn a powerful segmentation model that can estimate pixel-wise labels for unseen objects, even though they have only been exposed to a small number of labeled, or even unlabeled, images. When learning a specific unseen category with a few or zero annotated images, there is no constraint on the identities of the objects in training and reasoning samples. Few/zero-shot VOS [58], [102], [103], which aims at segmenting foreground objects under the guidance of a limited number of labeled frames, can be regarded as a particular case of ISS [16], [17], [55], as the images segmented here are continuous in time steps, and the target objects generally have the same identities but different appearances in sequential frames. Few/zero-shot 3DS [106], [107], [117] conducts segmentation in 3D space rather than 2D space, where the inputs are disordered 3D samples, such as point clouds and meshes, rather than regular 2D images. The settings and relationship of them are illustrated in Fig. 2 and Fig. 3, respectively.

In this section, we provide a problem definition of visual semantic segmentation in both few-shot and zero-shot scenarios. In each circumstance, we take ISS as an instance and deliver a detailed definition. In addition, we discuss the differences among the settings of few-shot ISS, few-shot VOS, and few-shot 3DS, and extend these problem settings into zero-shot scenarios. The nomenclature of the symbolism involved in this paper is listed in Table II.

1) *Few-shot Visual Segmentation*: The problem definition of few-shot ISS is firstly discussed, which will be further extended to few-shot VOS and few-shot 3DS. Few-shot ISS aims to learn a segmentation model from a few labeled images for target categories, as illustrated in Table I. A few-shot ISS task T attempts to learn a powerful segmentation model that is capable of predicting the pixel-wise semantic category from D_{support} for images in D_{query} . The $D_{\text{support}} = \{(x_i, m_i)\}_{i=1}^{I_{\text{support}}}$ called the support set denotes a

TABLE II
THE NOMENCLATURE OF THE SYMBOLISM INVOLVED IN THIS PAPER

Symbolism	Meaning
T	Few-shot/zero-shot task
x	Visual samples
m	Annotations
w	Category description
e	Word embeddings
n	Noise
ω	Weights of the prediction layer
\hat{m}	Prediction of x
\hat{e}	Prediction of e
\mathcal{X}_T	Input space of unseen classes
\mathcal{M}_T	Label space of unseen classes
\mathcal{X}_S	Input space of base classes
\mathcal{M}_S	Label space of base classes
D_{support}	Support set
D_{query}	Query set
D_{source}	Dataset of base classes
$D_{\text{auxiliary}}$	Auxiliary dataset
I_{support}	Number of samples in D_{support}
I_{query}	Number of samples in D_{query}
I_{source}	Number of samples in D_{source}

dataset containing I_{support} images x with their corresponding labels m , and $D_{\text{query}} = \{(x_k)\}_{k=1}^{I_{\text{query}}}$ termed the query set represents a dataset composed of I_{query} unlabeled images. The images from both D_{support} and D_{query} have the same marginal distribution $P_{\mathcal{X}_T}$ of the input space \mathcal{X}_T and the same label space \mathcal{M}_T . Typically, the I_{support} is set to CK , where the I_{support} samples consists of C classes, and each class provides K samples (generally K is set to 1 or 5 [43]). In these circumstances, the task T is allowed to be termed as a C -way K -shot task. Fig. 3(a) indicates an example of a few-shot ISS task, where the category of both D_{support} and D_{query} is “sheep”.

Whereas, directly adopting the conventional supervised

learning to reap a satisfactory segmentation model from insufficient data D_{support} is an exceptionally ambitious and challenging attempt. As a consequence, most of few-shot ISS approaches resort to the assistance of prior knowledge, which is distilled from an accessible dataset $D_{\text{source}} = \{(x_j, m_j)\}_{j=1}^{I_{\text{source}}}$ (if any) and is conducive to dealing with the target T . The source dataset D_{source} contains I_{source} labeled samples for base classes ($I_{\text{source}} \gg I_{\text{support}}$) and shares the same input space $\mathcal{X}_S = \mathcal{X}_T$ but distinct label space \mathcal{M}_S with D_{support} and D_{query} . By combining the available supervision information of T with the learned prior knowledge, obtaining a high-quality model becomes more feasible and reasonable [129].

The settings of few-shot VOS and few-shot 3DS follow that of few-shot ISS, but also preserve their own distinct identities. The samples to be processed in few-shot VOS are sequential frames. Given only the first frame of a video with annotations, few-shot VOS strives to predict pixel-wise labels for the specific objects in the remaining frames. Few-shot VOS is a sequential learning problem that segments moving objects with shared identities frame by frame. Since the appearance of the objects changes dramatically between different frames, the D_{support} in few-shot VOS are allowed to be comprised of the first frame with given labels and/or the historical frames with self-predicted masks to capture reliable object characteristics. For few-shot 3DS, it can be observed in Fig. 3(c) that the type of inputs is generally constructed by 3D exemplars. Take few-shot point cloud part segmentation for example. The D_{support} is made of point cloud samples whose subparts are annotated by unseen categories, and the D_{query} is composed of the ones with subparts of the same categories as the D_{support} .

2) *Zero-shot Visual Segmentation*: When ISS encounters zero-shot scenarios, the task T will become more complex to address. In this circumstance, the D_{support} fails to access any images with supervision signals, which hinders the learning on the target T . To address this issue, zero-shot ISS ordinarily attempts to transfer some supervision information from other modalities, denoting as the auxiliary dataset $D_{\text{auxiliary}}$, to enable the learning practicable, as depicted in Fig. 1. Specifically, auxiliary information [130] from semantic embeddings is applied to tackle the zero-shot ISS task T , as shown in Fig. 3(d). The dataset $D_{\text{auxiliary}} = \{(w_j, e_j)\}_{j=1}^N$ contains N categories, where w_j and e_j represent the category description and the word embedding corresponding to the j -th specific class, respectively. Thanks to $D_{\text{auxiliary}}$, it is hopeful for zero-label unseen objects to be separated effectively.

As a special case of few-shot VOS, the primary target object in the video sequence is demanded to be segmented without accessible annotations in zero-shot VOS. Since there is abundant temporal information that is hidden in the sequential frames or optical flows, zero-shot VOS often exploits appearance and motion cues to provide adequate supervision signals for task T . Moreover, language expressions can also supervise the model to segment objects of interest, as illustrated in Fig. 3(d). It can be seen in Fig. 3(f) that zero-shot 3DS follows the scheme of zero-shot ISS, which also uses word embeddings as crucial sources of auxiliary supervision signals. However, the target samples in zero-shot 3DS are 3D visual data rather

than 2D images, which is distinguished from zero-shot ISS.

B. Typical Datasets

In this section, two typical datasets of each visual semantic segmentation scenario are selected to be discussed. The picked datasets are displayed in Table I.

PASCAL-5ⁱ [14] is a popular benchmark designed for few-shot ISS, which is created by PASCAL VOC 2012 [13] with additional annotations from SDS dataset [140]. There are 20 categories, split into 4 subsets with 5 classes per subset.

COCO-20ⁱ [15] is the largest dataset for few-shot ISS, which is built from MS COCO benchmark [141]. It covers 80 common categories, which are divided into 4 folds with 20 categories per fold.

PASCAL VOC [13] can be adopted for image classification, object detection, and object semantic segmentation. For segmentation tasks, there are 1464 available samples for training and 1449 samples for validation with 20 categories in total.

PASCAL Context [45] is proposed for scene parsing, which covers both indoor and outdoor scenarios. It contains 4998 training and 5105 validation samples with 59 classes.

Youtube-VOS [56] is presented for VOS tasks, which has 94 classes in total. It consists of 3471 videos with 65 seen categories for training, 507 videos with 65 seen and 26 unseen categories for validation, and 541 videos with 65 seen and 29 unseen categories for testing.

DAVIS 2016 [57] is designed for segmenting fore- and background objects in videos, where only binary annotations are provided. There are 30 videos for training and 20 videos for validation.

ShapeNet Part [104] is applied for 3D fine-grained point cloud segmentation. It comprises 16881 3D shape instances of 16 classes. Each instance is further labeled by 2-5 part annotations, leading to 50 part categories.

ScanNet [105] is a 3D point cloud dataset captured by RGB-D cameras, which contains 1513 diverse indoor scenarios (1201 for training and 312 for testing) with 20 classes.

S3DIS [115] is a dataset for 3D semantic segmentation, which includes RGB-D data and 3D point clouds. The 3D point clouds are scanned from 271 indoor environments with 13 categories.

C. Technical Solutions

Given a query image x_k , the existing few/zero-shot visual segmentation approaches strive to estimate the posterior probability of classes the pixels in x_k belong to. For convenience, we describe this as follows:

$$\hat{m}_k = \arg \max P(m_k | x_k). \quad (1)$$

We further consider the Bayesian rule, then the above equation can be expressed as:

$$\hat{m}_k = \arg \max P(x_k | m_k) P(m_k), \quad (2)$$

where $P(x_k | m_k)$ is the conditional distribution of x_k given the mask m_k , and $P(m_k)$ is the prior distribution of m_k .

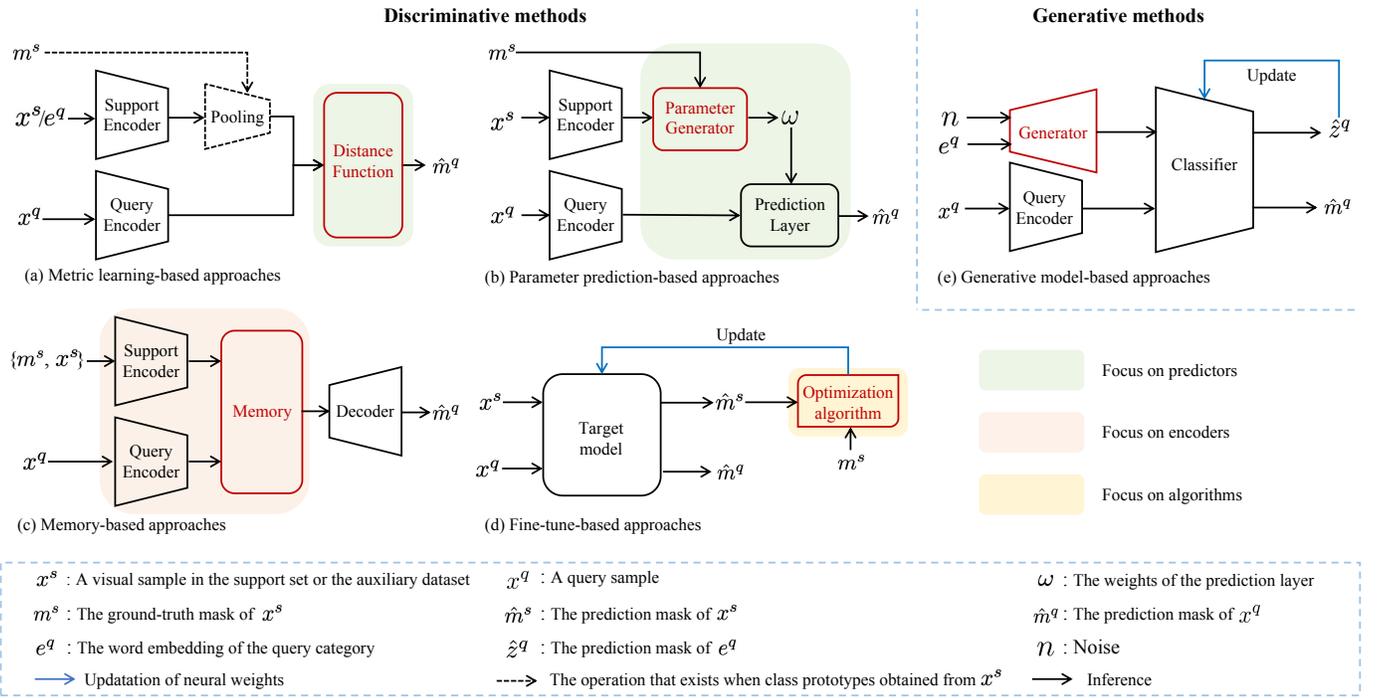


Fig. 4. Conventional frameworks of few/zero-shot visual segmentation. The principal differences of these frameworks are highlighted in red.

Based on Eq. 1 and Eq. 2, we can summarize the few/zero-shot visual segmentation approaches into discriminative approaches and generative approaches, where the discriminative approaches attempt to build their frameworks to maximize $P(m|x)$ and the generative approaches strive to model $P(x|m)P(m)$. Thus, discriminative methods expect to construct a powerful segmentation model so that it can maximize the posterior probability of classes on query samples. Discriminative methods tend to capture a more effective mask predictor, grasp an optimization algorithm with fast convergence, or enhance the representations of objects, leading to four sub-categories: metric learning-based, parameter prediction-based, fine-tune-based, and memory-based methods. The conventional frameworks of the above four technical solutions are exhibited in Fig. 4(a)-(d). As for the generative approaches, a generator is frequently built to fit the distribution of x or its features conditioned on classes in m , where $P(m)$ is generally supposed to be a uniform distribution, as shown in Fig. 4(e). In the rest of this section, we will describe these five types of technical solutions in detail.

1) *Metric Learning-based Methods*: Metric learning-based approaches attempt to learn a general distance function so that it can provide higher affinity scores to more similar features and lower scores to more distinct features for any category. The distance functions can be any algorithms or networks so long as they can calculate the distance between two representations. Since metric learning-based methods achieve segmentation through calculating pairwise similarity, the structures of few/zero-shot tasks, such as the number of categories, would not be strictly required [142], making them more flexible. Nevertheless, the high correlation between the source and target tasks is a potential condition for metric

learning-based methods to be well-implemented [129].

2) *Parameter Prediction-based Methods*: Different from metric learning-based approaches focusing on learning a powerful predictor transferable cross tasks, parameter prediction-based methods target designing a unique predictor for each task. To this end, a parameter generator is devised to predict the neural weights of the prediction layer. In this instance, the parameters of the predictor are updated through a simple forward propagation so that fast cross-class adaptation can be achieved. However, it is difficult for the generator to estimate large-scale model parameters [43].

3) *Fine-tune-based Methods*: Fine-tune-based methods intend to develop an optimization algorithm, so that the segmentation model can fast converge to model the posterior probability for a given few-shot (or zero-shot) task. Fine-tune-based methods have a powerful adaptation ability even if the unseen classes obey the different distribution from seen ones [143]. However, fine-tune-based methods involve some hyperparameters (i.e., the iteration steps), which are significant for model performance on target tasks. Moreover, due to the gradient calculation in backpropagation, a longer adaptation period is demanded to update parameters.

4) *Memory-based Methods*: Memory-based approaches aim to cooperate with the encoder to enhance the representations of objects. Specifically, memory-based approaches take advantage of some tools, such as RNNs and memory networks, to store previously seen information for assisting the representations of target classes. Thanks to the ability of these tools in capturing temporal information, memory-based approaches play a crucial role in few/zero-shot sequential learning problems. Nevertheless, memory-based methods generally contribute to high computational costs.

TABLE III
A SUMMARY ON COMMON ADVANTAGES AND DISADVANTAGES OF THE FIVE FEW-SHOT SOLUTIONS.

Type of Methods		Characteristics	Method performance					
			Training complexity		Adaptation period		Computational costs	
			Simple	Difficult	Short	Long	Low	High
Discriminative	Metric	learn a similarity metric applicable cross tasks	✓		✓		✓	
	Parameter	Learn a task-specific predictor by a parameter generator	✓		✓		✓	
	Fine-tune	Learn a task-specific model by an optimization algorithm	✓			✓		✓
	Memory	Cache previously seen information to enhance the features with encoders	✓		✓			✓
Generative		Synthesize labeled instances to enable discriminative solutions		✓		✓		✓

¹ Each technical solution is represented by its first word.

5) *Generative Model-based Methods*: Generative model-based methods tend to build a generator, which is learned from base classes to obtain an ability to fit $P(x|m)$ for a given task. The generator aims to synthesize labeled instances of unseen classes (i.e., visual features), which assists in modeling the class-specific distribution and enabling the parameter adjustment on segmentation models. Moreover, synthesized instances of both base and unseen classes are sometimes allowed to train the classifier together to alleviate the prediction bias towards base classes [144]. Whereas, the distribution of generated instances fails to obey the real distribution in most cases, and the training difficulties and inference costs are intractable in few/zero-shot problems.

6) *Summary*: Apart from the distinctive pros and cons of the five technical solutions mentioned above, the comparison of training complexity, adaptation period, and computational costs are further illustrated in Table III. In terms of training complexity, since the generator in generative model-based methods needs to be trained alone, the training step is more difficult than other methods. In respect of the adaptation period, owing to multiple forward and backward propagations, the time costs of fine-tune-based and generative model-based methods are higher than other solutions. Referring to computational costs, thanks to the single forward propagation and similarity calculation between a few features, metric learning-based and parameter prediction-based approaches have lower computational overhead. Besides, as for applicable scenarios, due to the access to the distribution of unseen classes via D_{support} , few-shot circumstances incline to adopt discriminative solutions, while zero-shot cases have a preference for generative model-based methods as their unavailability on the distribution of target unseen classes. Furthermore, thanks to the high flexibility of metric learning-based methods and the excellent ability of memory-based ones in storing valuable information, metric learning-based and memory-based approaches have made outstanding contributions in both few-shot and zero-shot scenarios.

III. IMAGE SEMANTIC SEGMENTATION

In this section, we review recently proposed ISS studies that are exposed to only a few or zero annotated training

samples and group them into few-shot ISS methods and zero-shot ISS methods, according to whether the annotated samples are accessible or not.

A. Few-shot Image Semantic Segmentation

Learning a pixel-wise classifier for unseen categories from a small number of labeled samples has attracted more and more attention. In recent years, several few-shot ISS methods [16]–[18] have been proposed. Since the methods based on generative models require higher training difficulties and inference costs to generate pseudo-labeled samples, few-shot ISS generally tends to discriminative solutions, which will be demonstrated as follows.

1) *Metric Learning-based Methods*: Metric learning plays a vital role in tackling few-shot ISS, where the approaches [16]–[22] based on prototype networks [145] are in a dominant position. Different from the conventional learning-based approaches [146]–[150], where the learned prototype of a class is an approximate estimate of the optimal prototype, these few-shot approaches [16]–[22] aim to obtain a class-specific prototype, which may not be an approximation of the optimal prototype, as long as it can provide the information of objects and enable higher similarity scores for the query features that have the same semantic classes as the prototype. However, it is insufficient to describe a category only by a vectorial prototype, which inspires some methods [27], [28], [30] aiming at generating multiple prototypes for each class. In addition, directly performing element-level dense matching between support and query features [32]–[34] is also a feasible way to break through this obstacle. The relevant approaches are depicted as follows.

Early approaches [16]–[18] attempt to conduct feature matching with a single class descriptor. Prototype networks were first exploited for segmentation on unseen objects with the assistance of a few labeled samples in [16]. Specifically, a two-branch network was designed to estimate segmentation maps for query images. The first branch took the support set as the input and output a global class descriptor, while the second branch leveraged the generated prototype as the guidance to tune the segmentation results of the query set. The work in [16] inspired the follow-up research in [17],

where binary segmentation was directly completed based on the cosine distance calculated between the feature vectors located on query features and the class-specific prototype [17]. To make full use of the support information, the query samples and their predicted masks were further regarded as a new support set to guide the segmentation of the original support samples. Unlike the work mentioned above [17] which applies a fixed distance function, other approaches [18]–[22] utilize a learnable neural network to predict the segmentation maps, which can be regarded as a learnable distance metric to implicitly measure the similarity between support and query features. These approaches fuse support cues of target classes with query features and decode the fused features to output final segmentation results. A common fusion strategy is to concatenate query features with the tiled prototypes [18], [19] or support feature maps [20] along the channel dimension, where multi-scale features [21] and multi-class label information [22] are considered to enhance the representations of query samples. Apart from concatenating directly along the channel dimension [18]–[22], other methods, such as element-level addition [23], re-weighted by attention map [24], and similarity guidance [25], [26], are also some feasible ways to conduct the integration between support and query features.

Though these methods mentioned above have made undeniable contributions to few-shot ISS, some of them [19], [21], [22] apply the masked average pooling operation to generate a holistic descriptor for each semantic category, giving rise to some issues. Firstly, the insufficiency of annotated category-specific samples makes the prototype learner fail to output a robust class representation. Secondly, due to the appearance variations between support and query samples, it is challenging to capture rich and fine-grained semantic information only by a global feature vector. To deal with this dilemma, some follow-up approaches attempt to generate multiple prototypes for each semantic category [27], [28], [30] or conduct dense matching between support and query images [32]–[34].

From the perspective of generating multiple prototypes for each class, the similarity measurement is conducted between each generated prototype and query features. A common way is to divide the object into different parts according to a particular mechanism and generate a corresponding prototype for each part. Object semantics were decomposed to assist in the generation of multiple prototypes [27]. To change the number of prototypes adaptively, similar support feature vectors with different spatial positions were grouped to generate a specific prototype in [28]. In order to obtain more fine-grained feature representations, multiple part-aware prototypes were further refined with the help of unlabeled samples in [29]. In addition, three different descriptors were designed from multiple aspects for a specific object, which would be employed to conduct feature matching with query features [30]. Instead of obtaining deterministic prototypes, the distribution of generated prototypes was estimated to simulate the uncertainty caused by limited training images and object variations, which improved the robustness of the segmentation model [31].

From the perspective of dense matching, a pyramid graph network was presented to capture the dense correspondences between support and query features at different scales [32].

A democratic attention network was proposed to focus more on pixels where the object was located, building a robust correspondence between support and query images [33]. A harmonic feature activation strategy was proposed, which jointly exploited exclusive support features for pixel-level semantic matching [34]. A novel cross-attention mechanism was proposed for aggregating more relevant pixel-wise features in support images into query ones [35]. A bipartite graph was built and a graph attention mechanism as well as weight adjustment strategy were applied to promote more target-object pixels to participate in the segmentation on query images [36]. The dense correlations of foreground and background were explored, which alleviated the information loss caused by prototype learning and dense matching of a foreground feature pair [37].

2) *Parameter Prediction-based Methods*: In few-shot ISS, parameter prediction-based methods are frequently employed to modify the weights of the classifier for cross-class adaptation, as demonstrated in Fig. 4(b). By employing this, the segmentation network trained on base classes can quickly enhance the segmentation ability on unseen classes.

A two-branch network, consisted of a conditional branch and a segmentation branch, was proposed to tackle cross-category segmentation in [14]. The conditional branch input a class-specific image with its mask and predicted the weights of the logistic regression layer for adapting to a target object. Unlike the conditional branch, the predominant duty of the segmentation branch was to extract high-level semantic features from query images. Through the logistic regression layer with replaced parameters, the pixel-wise semantic labels could be generated from the extracted query features. Instead of leveraging support samples merely, query images were also employed to the generation of classifier weights [38]. In contrast to replacing the classifier parameters directly, the weights of the classifier were added dynamically so that the model can master both base and unseen categories in [39].

3) *Fine-tune-based Methods*: The fine-tune-based few-shot ISS aims to adopt an optimization algorithm to refine the parameters of the pre-trained segmentation network for learning unseen categories. The segmentation network was refined iteratively by minimizing the error calculated from support predictions and their corresponding masks in [40]. With the help of parameter refinement, the performance degradation resulting from the inter-class gap between the offline and online stages was alleviated. An embedding network and a differentiable linear classification model were proposed so that the parameters of the linear classification model could be updated more efficiently while the embedding network generalizing among diverse classes in [41]. Different from the approaches [40], [41] adopting episode training in the offline stage, a transductive inference strategy based on standard supervised learning was resorted to obtain a feature extractor on base classes [42]. In the inference phase, a linear classifier was refined by means of minimizing a loss function based on labeled support images and the statistical characteristics of unlabeled query ones. Apart from the cross-category adaptation, the shift caused by distribution diversity between training and inference data was also concerned, making it more desirable

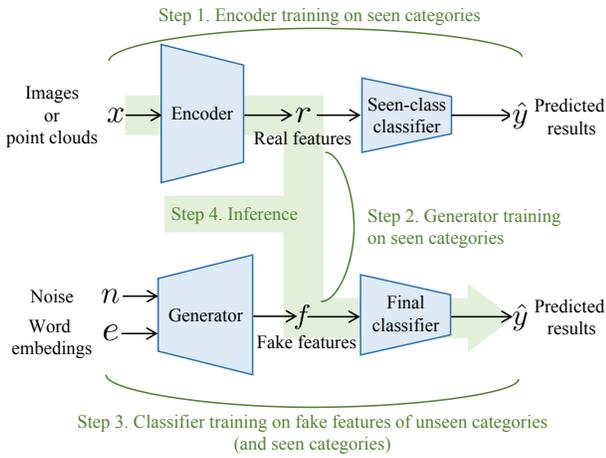


Fig. 5. Implementation process of zero-shot ISS and zero-shot 3DS approaches based on generative model (adapted from [52], [116])

for real applications.

4) *Memory-based Methods:* In the memory-based few-shot ISS, the previously seen information is reserved to help the segmentation on query samples. Common attributes and prior information of diverse categories with noticeable visual differences were stored into an external memory, which was manipulated to transfer labels for unseen classes [43]. However, the computational loads of the model increased with the growth of memory size [43]. The features of target classes with different resolutions were memorized, which reduced the computational consumption [44]. The stored features were then extracted to obtain more cross-resolution information and accurate segmentation results.

B. Zero-shot Image Semantic Segmentation

Zero-shot ISS focuses on performing segmentation for zero-label categories unavailable in the training process. Attributable to the absence of supervision signals on class-specific visual samples, it is tricky for zero-shot ISS to modify the model weights directly. Consequently, flexible metric learning-based settlements and generative model-based methods capable of generating pseudo labeled data are adopted by zero-shot ISS for learning new categories.

1) *Metric Learning-based Methods:* In zero-shot ISS, the features encoded from query samples are in a latent visual space while the word embeddings are in a semantic space. Due to the inconsistency between the two embedding spaces, it is unreasonable to carry out feature matching directly. Therefore, the metric learning-based approaches endeavor to map the visual and/or semantic features into a shared space, which can be the semantic space [46]–[48], the visual space [50] and the shared latent space [51], so that the distance measurement can be conducted.

The majority of these approaches [46]–[48] take the semantic space as the common embedding space. Thus, the function projecting visual features into the semantic space should be devised in advance. Zero-annotation ISS on novel categories was conducted by two steps [46]. Firstly, a visual-semantic

embedding module was devised to project the class-specific visual information into the semantic space, where every mapped embedding vector could be regarded as the representation of a particular pixel of the query sample. Secondly, a mask was predicted based on the similarity between the semantic and pixel embeddings. Despite that achievements have been made in [46], the prediction bias towards observed classes arises when transferring the knowledge from the observed classes to unseen categories. To cope with this issue, the studies in [47], [48] endeavor to train the model on observed categories together with the information from unseen ones. Image captions rather than word embeddings were adopted to dig out supervision signals for unseen classes, where the position cues of unseen classes could be provided by base ones [47]. Target-class semantic embeddings were utilized in training to reduce the prediction bias, which aimed to learn the shareable information concealed in the source and target semantic embeddings [48]. Specifically, a saliency detection strategy was proposed to effectively distinguish the area where the source objects were located so that the pixel-level label assignment for both source and target classes could be conducted in their respective areas. Nevertheless, these methods [46], [48] do not pay attention to the unfavorable impact resulting from noisy and outlying samples, which contributes to biased estimation of target classes [49]. Based on this, Bayesian uncertainty estimation [151] was introduced to aggregate more discriminative samples on seen classes to carry out pixel-wise label prediction [49]. However, extra parameters were involved in estimating the uncertainty of input images.

There are also some approaches [50], [51] conducting zero-shot ISS in other embedding spaces. Distinct from classifying pixels in the semantic space [46]–[48], the segmentation of unseen classes was also allowed to be conducted in the visual space [50]. The potential distribution of semantic embeddings was constructed, where the embeddings were then decoded with visual features to measure the pair-wise similarity. A joint embedding space was opted to train a visual encoder and a semantic encoder in [51]. Moreover, two complementary loss functions were presented to learn more representative embeddings and a decision boundary modification scheme was proposed for the prediction bias.

2) *Generative Model-based Methods:* It can be seen in Fig. 4(e) that the generative model-based methods strive to estimate labeled instances to enable the weight modification on the classifier. As one of the representative methods, the encoder was trained firstly on seen categories to estimate real visual features in [52]. The encoded features and semantic embeddings of seen classes were then exploited to learn a generator, which could take semantic embeddings as input and output the fake visual features for unseen categories. The generated features of never-seen categories (and real features of seen categories) were finally applied to refine the classifier weights at inference time, as illustrated in Fig. 5. The research in [52] spawns some variants [53]–[55], where contextual information [53], [54] and inter-class structural relationship [55] were considered to synthesize higher-quality visual features.

TABLE IV
A SUMMARY OF FEW/ZERO-SHOT ISS METHODS INVOLVED IN THIS SURVEY.

Methods	Years	Scenarios		Technical Solutions					Main contributions
		Few-shot	Zero-shot	Discriminative				Generative	
				Metric	Parameter	Fine-tune	Memory		
Shaban <i>et al.</i> [14]	2017	✓			✓				Two-branched architecture
Dong <i>et al.</i> [16]	2018	✓		✓					Prototype learning
Rakelly <i>et al.</i> [20]	2018	✓		✓					FCNs
Wang <i>et al.</i> [17]	2019	✓		✓					Parametric classification
Zhang <i>et al.</i> [18]	2019	✓		✓					Mask refinement
Hu <i>et al.</i> [23]	2019	✓		✓					Multi-scale feature fusion
Zhang <i>et al.</i> [32]	2019	✓		✓					GNN, attention mechanism
Tian <i>et al.</i> [21]	2020	✓		✓					Prior information
Yang <i>et al.</i> [24]	2020	✓		✓					Transformation module
Zhang <i>et al.</i> [25]	2020	✓		✓					Feature representation
Liu <i>et al.</i> [29]	2020	✓		✓					Multi-prototype representation
Wang <i>et al.</i> [33]	2020	✓		✓					Attention mechanism
Liu <i>et al.</i> [39]	2020	✓			✓				Attention mechanism
Yang <i>et al.</i> [40]	2020	✓				✓			Online refinement strategy
Tian <i>et al.</i> [41]	2020	✓				✓			Linear classifier
Liu <i>et al.</i> [34]	2021	✓		✓					Harmonic feature activation
Li <i>et al.</i> [28]	2021	✓		✓					Multi-prototype representation
Yang <i>et al.</i> [27]	2021	✓		✓					Multi-prototype representation
Zhang <i>et al.</i> [30]	2021	✓		✓					Multi-prototype representation
Wang <i>et al.</i> [31]	2021	✓		✓					Probabilistic framework
Zhang <i>et al.</i> [35]	2021	✓		✓					Transformer
Zhuge <i>et al.</i> [38]	2021	✓			✓				Feature fusion
Boudiaf <i>et al.</i> [42]	2021	✓				✓			Transductive inference
Wu <i>et al.</i> [43]	2021	✓					✓		Memory network
Xie <i>et al.</i> [44]	2021	✓					✓		Memory network
Chen <i>et al.</i> [22]	2022	✓		✓					Multi-class guidance
Lang <i>et al.</i> [19]	2022	✓		✓					Base-class prediction
Shi <i>et al.</i> [37]	2022	✓		✓					Feature fusion
Liu <i>et al.</i> [26]	2022	✓		✓					Weight-sparsification
Gao <i>et al.</i> [36]	2022	✓		✓					Attention mechanism
Xian <i>et al.</i> [46]	2019		✓	✓					Projection function
Kato <i>et al.</i> [50]	2019		✓	✓					Projection function
Bucher <i>et al.</i> [52]	2019		✓					✓	Zero-shot framework
Tian <i>et al.</i> [47]	2020		✓	✓					Image captions
Gu <i>et al.</i> [53]	2020		✓					✓	Feature generator
Li <i>et al.</i> [55]	2020		✓					✓	Feature generator
Hu <i>et al.</i> [49]	2020		✓	✓					Uncertainty estimation
Lu <i>et al.</i> [48]	2021		✓	✓					Feature enhancement
Baek <i>et al.</i> [51]	2021		✓	✓					Loss function
Gu <i>et al.</i> [54]	2022		✓					✓	Feature generator

¹ Each technical solution is represented by its first word.

C. Summary

To elaborate this section intelligibly, we further describe the involved few/zero-shot ISS approaches in Table IV. It can be seen that the methods based on metric learning occupy a crucial position in few-shot ISS, where the diversity of feature representation, distance functions and feature matching strategies give researchers inexhaustible motivation for innovations. Due to the characteristics of zero annotation in zero-shot ISS, it is challenging to directly leverage semantic embeddings to optimize visual segmentation models. Therefore, easy-to-implement metric learning-based approaches are preferred by zero-shot ISS. Moreover, adopting generative model-based approaches to convert zero-shot ISS into few-shot problems is also popular in breakthrough these limitations.

IV. VIDEO OBJECT SEGMENTATION

Few/zero-shot VOS is the extension of few/zero-shot ISS in temporal dimension. When the objects in a task have the same identities and support and query images are continuous on time stamps, a few/zero-shot ISS problem becomes a few/zero-shot VOS challenge. Compared with few/zero-shot ISS, few/zero-shot VOS has no requirement on manual construction of few/zero-shot tasks, where a video of an unseen class can be naturally regarded as a target task. The common VOS settings of the support set are demonstrated in Fig. 6. The previous frame with the predicted mask and/or the first frame with given annotations can be adopted to guide the segmentation of the current frame, and the intermediate frames spanning from the first to previous frames are also allowed to be leveraged. Moreover, compared with ISS, few/zero-shot VOS burdens fewer

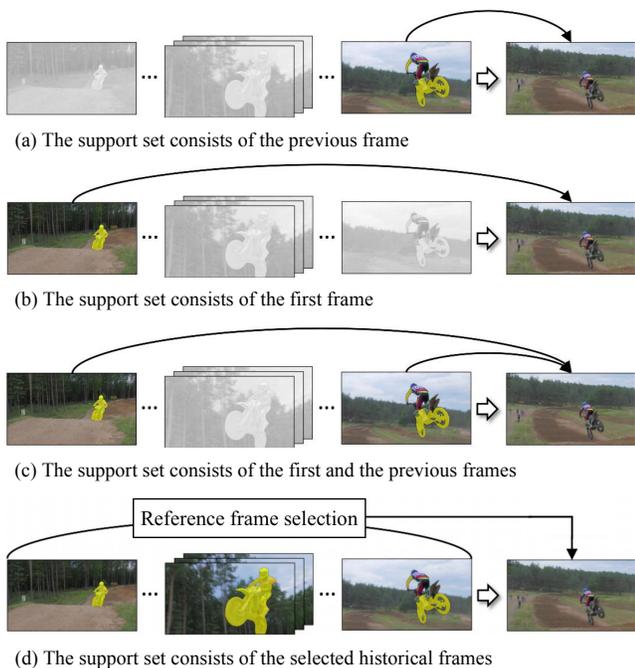


Fig. 6. The settings of the support set in few-shot VOS (adapted from [73]), where the mask of the first frame is the ground truth and the masks of intermediate and previous ones are self-estimated masks. These constructions are also commonly-used in zero-shot VOS, where the mask of the first frame is unavailable.

appearance variations and enjoys richer temporal information, where appearance and temporal cues in support samples can be applied to enhance the representation of query samples, leading to memory-based methods preferred in this case.

A. Few-shot Video Object Segmentation

Due to the inter-frame similarity of the object appearance in a video, it is reasonable to adopt metric learning-based approaches to separate the area similar to the given object from the current frame. In addition, for a given video, the previously seen historical frames play a considerable role in enhancing the representations of objects in the current frame, leading to the prosperity of memory-based approaches. Different from few-shot ISS, few-shot VOS has higher requirements on the speed of model adaptation, which is desired to be larger than the frame rate. However, owing to the problem of time-consuming, fine-tune-based approaches are facing a decline. In this section, we will describe these solutions separately.

1) *Metric Learning-based Methods*: Compared with few-shot ISS, only the first frame is labeled in few-shot VOS. It is challenging to construct a reliable class descriptor merely based on a single frame. Thus, few-shot VOS usually carries out dense matching between the features generated by current and reference frames. In this section, we categorize and review each method according to the different selection of reference frames.

The first frame [58], [59] or the previous frame [60] (or both [61]–[63]) is to be given preference by earlier studies for comprising support sets, as depicted in Fig. 6(a)–(c). Object tracking was combined with VOS to estimate masks as well

as rotated bounding boxes simultaneously, with the assistance of the pairwise similarity between the features mapped from the current frame and the initial template [58]. Taking the low efficiency caused by pixel-wise dense matching and the curse of dimension in the Euclidean space [152] into account, a hypersphere embedding space was built, where the cosine distance was replaced with a convolution layer to accelerate the similarity measurement [59]. As an alternative to the initial frame, the previous frame adjacent to the current frame was employed to provide appearance information of the target object in [60]. By measuring the distance between the features generated from the previous and current frames, a coarse segmentation map of the current frame could be obtained, which would be further refined for more accurate prediction. For obtaining global and local correlations, a novel pixel-level matching mechanism was presented, where the current frame was employed to calculate correlations with both first and historical frames in [61]. To alleviate mismatching on similar objects in the background, foreground and background areas were addressed identically in [62]. Taking the appearance and structure information on the target objects into account, a structure modeling branch was constructed to encode the information of the complete object and its components in [63].

Nevertheless, these approaches [58]–[63] only leverage the initial frame and/or the adjacent frame, making it tricky to deal with the challenge of object deformation, occlusion, and model shift. Therefore, other metric learning-based approaches [64]–[66] aim to explore a reasonable intermediate frame sampling mechanism, as illustrated in Fig. 6(d), to capture richer temporal information and tackle the shift caused by sparse matching. To explain the appearance variations of the target objects, more recent and less long-distance frames were selected in [64]. An observed video was clipped into different snippets, from each of which a frame was selected to build a support set together with the labeled first frame, and the final segmentation map was obtained by averaging the predictions conducted on the target frame and each support frame in [65]. Different from the manually designed sampling strategy [64], [65], an adaptive selection mechanism was presented based on the similarity in object appearance and the accuracy on predicted masks in [66]. Despite the utilization of intermediate frames can bring about performance gains, redundant computational consumption exists when the appearance variations of target objects are slight cross different frames. On that account, a new segmentation network was proposed to dynamically adjust the processing strategy according to the appearance changes across frames, which reduced the unnecessary computational occupation [67].

2) *Fine-tune-based Methods*: The fine-tune-based methods endeavor to optimize the model for each video sequence based on the first labeled frame (and the subsequent frames with their self-predicted masks), where over-fitting risks, time-consuming problems and handcrafted hyperparameters hinder their development. To alleviate over-fitting and generalize to new video objects more flexibly, a novel architecture consisting of a segmentation sub-network and a lightweight appearance sub-network was proposed, where only the appearance one was updated online [69]. In response to the chal-

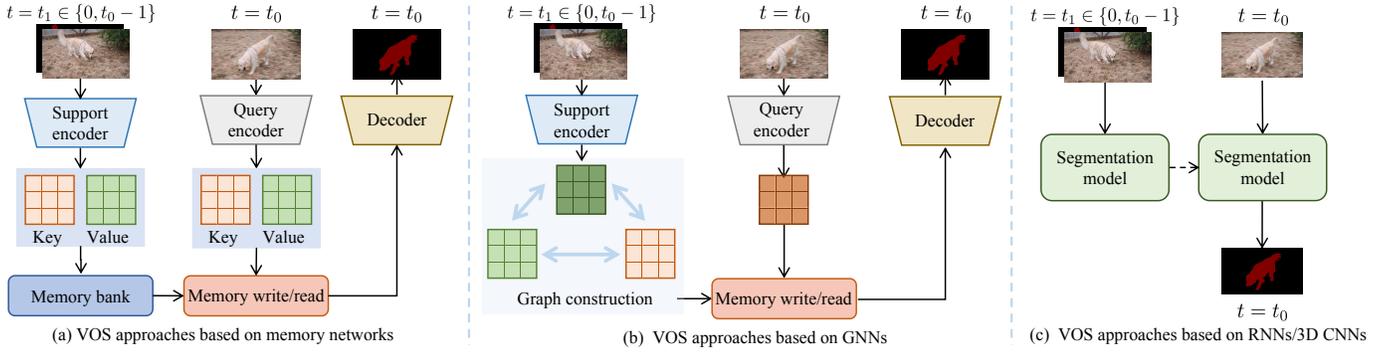


Fig. 7. Data flows in memory-based VOS methods, in which the masks of historical frames are only leveraged in few-shot scenarios. Memory networks and GNNs are adopted to cache the valuable information explicitly, while RNNs and 3D CNNs, which are leveraged as some subparts of the segmentation model, are employed to store the knowledge implicitly. The squares with different colors represent diverse features.

lengths of time-consuming and handcrafted hyperparameters, optimization-based meta-learning [153] was integrated into online adaptation for efficient VOS [70]. In the offline stage, the common knowledge between different objects was mined by training on multiple similar segmentation tasks, and the optimal model initialization as well as parameter-level learning rates were provided. In the online adaptation, the first and previous frames were leveraged to fine-tune the initialized segmentation model with provided learning rates, reaping the characteristics of unseen objects and adapting to annoying appearance variations. Whereas, providing learning rates for each parameter places restrictions on large-scale segmentation networks. To address this issue, neuron-level learning rates were explored, which considerably eased the requirements on the number of learning rates in [71]. Moreover, VOS could be decomposed into object detection and local-mask prediction, so that the fine-tuning operation could be carried out with the assistance of bounding box propagations. Despite higher efficiency can be reached, these methods [70], [71] conducted optimization with fixed learning rates, making them challenging to adaptively deal with target objects with diverse appearances [72]. To handle this, an evaluation criterion was proposed for the meta learner and the learning rates were automatically modified for the upcoming frames [72].

3) *Memory-based Methods*: To cache previously seen information, memory networks [73]–[75], GNNs [83] and RNNs [83], [84] are leveraged in memory-based methods, as shown in Fig. 7.

As one of representative approaches, a space-time memory network was devised, which could store helpful information from observed frames [73], [74]. Peculiarly, when the estimation was performed on the current frame, a pixel-wise matching on key maps was conducted to read out the beneficial value maps for enhancing the representations of the current frame. Furthermore, to attenuate the adverse effects resulting from the appearance deformation, the features stored in the memory were updated dynamically with the segmentation of video frames, as demonstrated in Fig. 7(a). However, these methods [73], [74] conduct global-to-global dense matching strategies between query and memory features, leading to erroneous segmentation on background objects that are similar to the target ones. To cope with the mismatching, some studies

[75]–[81] attempt to build non-global relationships between features. A memory network based on the Gaussian kernel was developed to compel features focusing on a single object, which reduced the occurrence of mismatching [75], [76]. A local-to-local matching between the regions of memory and query frames was performed [77], which mitigated the mismatching on similar objects and brought about lower computational consumption than global-to-global methods [73], [74]. Whereas, these methods [73], [74], [77] only perform matching at coarse scales, which makes them challenging to capture fine-grained information [78]. Considering this, a novel reading mechanism was proposed to set up multi-scale correspondences, where coarse correlations of dense matching were employed to guide sparse matching at finer scales [78].

Nevertheless, these methods [73], [74], [78] only conduct memorization on the latest information without abandoning the antiquated or useless one, resulting in the increasing computational costs and memory occupation with the progress of the segmentation. Consequently, a considerable number of approaches [79]–[81] have been presented to manage the memory features efficiently. A global representation with a fixed size was learned from all support frames, which decoupled the computational consumption with the length of the given video sequence [79]. A feature bank was embedded into the memory network to aggregate useful features and forget worthless ones dynamically [80]. Without storing the key and value maps for each instance individually, the correlation between query and memory frames was delved to retrieve the value map encoded from multiple instances, where a voting mechanism was employed for feature aggregation [81]. Aside from CNN-based memory networks utilized by these mentioned approaches [73]–[81], graph-based memory networks are also employed [82]. In [82], a novel graph memory network was designed to store the information of the support frames with their masks. To quickly adapt to the visual variations on different timestamps, a learnable controller was added to update features and store more abundant information dynamically on the premise of maintaining a fixed memory size.

Furthermore, GNNs and RNNs also play a crucial role in few-shot VOS, as shown in Fig. 7(b) and Fig. 7(c). GNNs and GRUs were jointly adopted to model short-term and long-

term temporal information, respectively, where the short-term information was cached in nodes of GNNs while the long-term one is stored in the GRU modules [83]. The GRU modules were also embedded into the segmentation model to propagate features for proposal generation [84].

B. Zero-shot Video Object Segmentation

Zero-shot VOS targets separating primary moving objects without labeled samples available. As seen in Fig. 3(d), distinct from zero-shot ISS and zero-shot 3DS extracting supervision signals from semantic embeddings, the auxiliary information in zero-shot VOS is allowed to be appearance cues (i.e., unlabeled historical frames), motion cues (i.e., optical flow), or semantic cues (i.e., language descriptions). The identity of the object in zero-shot VOS, whose appearance has strong correlations between adjacent frames, is the same among diverse frames. Therefore, it is reasonable and intuitive to adopt memory-based methods instead of more complex generative model-based approaches for this task, leading to some principal solutions in zero-shot VOS: metric learning-based, fine-tune-based and memory-based settlements.

1) *Metric Learning-based Methods*: Due to the differences in auxiliary samples, the pairwise similarity can be calculated between appearance features of historical and current frames, motion and appearance features of the current frame, and semantic features of language expressions and appearance features of frames.

From the standpoint of the similarity captured from inter-frame appearance embeddings, the dominant challenges lie in reference frame selection and feature matching mechanisms. A feasible way is to conduct pixel-level non-local matching between initial and current frames, where inter-frame information [85] and intra-frame information [86] are leveraged. However, non-local matching may lead to unexpected computational expenditure. Targeting at this issue, local matching was carried out only on object locations [87]. To make full utilization of the global information of a given video, the correlation between two order-independent frames was grasped in [88], which helped to determine the area where required to be segmented.

When it comes to the similarity calculated by the current frame and the optical flow, a well-designed fusion strategy is contrived to implicitly align motion features to appearance ones. Given a current frame, a pixel-level interaction between the prominent motion map and object proposals was performed, which eliminated the obstacles derived from moving background and unchanging objects [89]. To encourage the appearance features generated by distinct convolution stages, a multi-level interaction operation was proposed to interplay with the motion ones, so that the representations of objects could be decoded into more accurate segmentation masks [90]. To suppress the misleading knowledge caused by noisy optical flows, the methods of completing the discontinuous edges of optical flows [91], dynamically adjusting the effects of motion and appearance cues on spatio-temporal representations [92], and promoting consistent features and suppressing incompatible ones [93] were presented. Moreover, depth maps and

static saliency were integrated, where the representations of foreground objects were enlarged and purified in [94].

The similarity also can be calculated between semantic features of language expressions and appearance features of frames. Two attention modules were carefully designed to encourage temporal consistency and prevent model shift in [95]. Trivial and non-trivial linguistic phrases were encoded into language features to identify referents more accurately in [96]. To encode semantic and spatial information of objects at multiple levels, visual features of different scales interacted with linguistic features in the encoder [97]. A multimodal transformer was presented in [98] to conduct cross-modal interaction between semantic and appearance features.

2) *Fine-tune-based Methods*: Since language features generally carry unrelated information of objects, fusing language features with visual ones may lead to inferior segmentation accuracy. To address this issue, a learning-to-learn pattern was proposed, where target-specific cues can be obtained by fine-tuning the parameters of the transfer function [99].

3) *Memory-based Methods*: The memory tools shown in Fig. 7 are also applied to zero-shot scenarios. RNNs were employed to perpetrate implicit memory in both spatial and temporal dimensions [100]. In the spatial dimension, RNNs were exploited to explore other objects at different locations within the given frame. In the temporal dimension, the role of RNNs was to maintain the relationships across different frames. Assisted by these settings, the representations of multiple instances could be enhanced simultaneously by only a single forward propagation. Moreover, GRUs [101] and 3D CNNs [102] were embedded into the architecture to extract spatio-temporal information from video sequences. A fully connected graph was established to mine the interrelationship between any two frames in [103], where video frames were encoded as nodes and their correlations were modeled as edges. When separating the object of interest frame by frame, the information stored in the graph was dynamically updated to obtain comprehensive knowledge and precise segmentation masks.

C. Summary

In the discussions of this section, it can be found that compared with ISS, few/zero-shot solutions face a great deal of new challenges and opportunities when addressing VOS problems. The extreme inadequacy of labeled data is the first challenge, which leads to prototype networks that thrive in ISS are not suitable for VOS. Secondly, the utilization of intermediate frames may cause cumulative errors on the current frame, where a selection mechanism to pick up high-quality reference frames needs to cooperate with few/zero-shot solutions. Thirdly, the requirements on the real-time performance of VOS are much higher than that of ISS, which requires additional consideration when designing frameworks. Concerning opportunities, it is worth mentioning that VOS can access richer information, such as the correlations among support frames and query ones, which leads to generous variants of technical solutions. In addition, due to the similarity among ISS and VOS tasks, some ideas proposed in ISS can

TABLE V
A SUMMARY OF FEW/ZERO-SHOT VOS METHODS.

Methods	Years	Scenarios		Technical Solutions			Flow	Main contributions
		Few-shot	Zero-shot	Metric	Fine-tune	Memory		
Wang <i>et al.</i> [58]	2019	✓		✓				Multi-task
Voigtlaender <i>et al.</i> [61]	2019	✓		✓				Feature embedding, matching mechanism
Khoreva <i>et al.</i> [68]	2019	✓			✓		✓	Data augmentation
Xiao <i>et al.</i> [70]	2019	✓			✓			Meta learner
Oh <i>et al.</i> [73]	2019	✓				✓		Memory network, matching mechanism
Lyu <i>et al.</i> [84]	2019	✓				✓		Conv-GRU module, multi-task
Hu <i>et al.</i> [60]	2020	✓		✓			✓	Refinement network, attention mechanism
Yang <i>et al.</i> [62]	2020	✓		✓				Feature embedding
Zhang <i>et al.</i> [64]	2020	✓		✓				Transductive inference
Liu <i>et al.</i> [65]	2020	✓		✓				Reference frame selection
Robinson <i>et al.</i> [69]	2020	✓			✓			Optimization technique
Meinhardt <i>et al.</i> [71]	2020	✓			✓			Optimization technique
Seong <i>et al.</i> [75]	2020	✓				✓		Memory network
Li <i>et al.</i> [79]	2020	✓				✓		Feature embedding
Liang <i>et al.</i> [80]	2020	✓				✓		Memory network, loss function
Lu <i>et al.</i> [82]	2020	✓				✓		Memory network
Zhang <i>et al.</i> [83]	2020	✓				✓		Memory network
Yin <i>et al.</i> [59]	2021	✓		✓				Matching mechanism, feature embedding
Park <i>et al.</i> [67]	2021	✓		✓				Dynamic network
Xu <i>et al.</i> [72]	2021	✓			✓		✓	Optimization technique, loss function
Xie <i>et al.</i> [77]	2021	✓				✓	✓	Memory network
Seong <i>et al.</i> [78]	2021	✓				✓		Memory network, matching mechanism
Cheng <i>et al.</i> [81]	2021	✓				✓		Memory matching mechanism
Zhu <i>et al.</i> [63]	2022	✓		✓				Structure modeling
Hong <i>et al.</i> [66]	2022	✓		✓				Reference frame selection
Oh <i>et al.</i> [74]	2022	✓				✓		Memory network
Seong <i>et al.</i> [76]	2022	✓				✓		Memory network
Yang <i>et al.</i> [85]	2019		✓	✓				Aggregation technique
Zhuo <i>et al.</i> [89]	2019		✓	✓			✓	Feature fusion
Wang <i>et al.</i> [103]	2019		✓			✓		GNN
Ventura <i>et al.</i> [100]	2019		✓			✓		ConvLSTM decoder
Tokmakov <i>et al.</i> [101]	2019		✓			✓	✓	ConvGRU module
Gu <i>et al.</i> [87]	2020		✓	✓				Self-attention mechanism
Zhou <i>et al.</i> [90]	2020		✓	✓			✓	Attention mechanism
Seo <i>et al.</i> [95]	2020		✓	✓				Attention mechanism
Bellver <i>et al.</i> [96]	2020		✓	✓				Non-trivial referring expressions
Mahadevan <i>et al.</i> [102]	2020		✓			✓		3D CNN decoder
Liu <i>et al.</i> [86]	2021		✓	✓				Feature fusion, matching mechanism
Zhou <i>et al.</i> [91]	2021		✓	✓			✓	Feature refinement
Yang <i>et al.</i> [92]	2021		✓	✓			✓	Attention mechanism, feature fusion
Ji <i>et al.</i> [93]	2021		✓	✓			✓	Attention mechanism, feature refinement
Zhao <i>et al.</i> [94]	2021		✓	✓			✓	Multi-source fusion
Yang <i>et al.</i> [97]	2021		✓	✓			✓	Feature fusion
Lu <i>et al.</i> [88]	2022		✓	✓				Siamese network, attention mechanism
Botach <i>et al.</i> [98]	2022		✓	✓				Transformer framework
Li <i>et al.</i> [99]	2022		✓		✓			Feature fusion, optimization technique

¹ Each method is represented by its first word.

² “Flow” represents the methods involving optical flows.

be transferred into few/zero-shot VOS scenarios. Table V summarizes the few/zero-shot VOS methods introduced in this paper.

V. 3D SEGMENTATION

Few/zero-shot 3DS can alleviate the requirements for expensive annotations of unseen classes and address cross-class adaptation as ISS and VOS problems. However, since the data is disordered and nonstructural 3D samples rather than regular RGB images, as can be seen from Fig. 3(e) and Fig.

3(f), few/zero-shot 3DS is more challenging than 2D cases, where architectures designed for 2D samples generally fail to be employed into 3D circumstances. Consequently, the existing few/zero-shot 3DS methods draw lessons from the 2D approaches and develop applicable protocols for 3D data structures. In this section, we laconically review the existing few/zero-shot 3DS approaches.

TABLE VI
A SUMMARY OF FEW/ZERO-SHOT 3DS METHODS.

Methods	Years	Scenarios		Technical Solutions			Point	Main contributions
		Few-shot	Zero-shot	Discriminative		Generative		
				Metric	Parameter			
Sharma <i>et al.</i> [112]	2019	✓				✓	Embedding network	
Chen <i>et al.</i> [114]	2019	✓				✓	Autoencoder	
Sharma <i>et al.</i> [113]	2020	✓				✓	Self-supervised pre-training	
Chen <i>et al.</i> [106]	2020	✓		✓		✓	Prototype network	
Yuan <i>et al.</i> [108]	2020	✓		✓			Descriptor generator	
Wang <i>et al.</i> [109]	2020	✓		✓		✓	Shape morphing	
Zhao <i>et al.</i> [107]	2021	✓		✓		✓	Multi-prototype representation	
Hao <i>et al.</i> [110]	2021	✓			✓	✓	Meta learner	
Huang <i>et al.</i> [111]	2021	✓				✓	Meta learner	
Michele <i>et al.</i> [116]	2019		✓			✓	Zero-shot framework, dataset	
Liu <i>et al.</i> [117]	2022		✓			✓	Dataset, embedding network	

¹ Each method is represented by its first word.

² “Point” indicates that the sample type is point clouds.

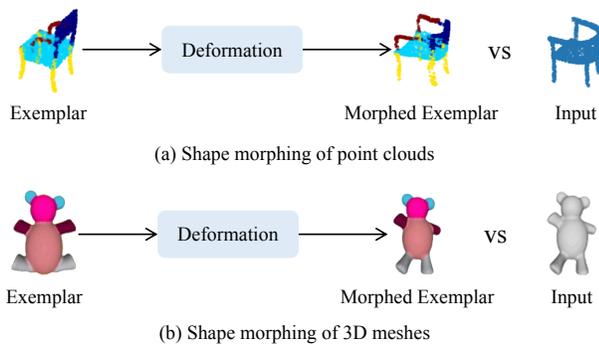


Fig. 8. The operation of shape morphing in few-shot 3DS based on metric learning (adapted from [109]). The samples are selected from [108], [109].

A. Few-shot 3D Segmentation

The majority of available few-shot 3DS methods counting on the metric learning paradigm have been relatively deeply researched, while a small proportion of them resorting to other tools, such as fine-tune-based or parameter prediction-based strategies, have been studied rarely. This section focuses on the existing few-shot 3DS methods and discusses them as follows.

1) *Metric Learning-based methods*: The methods based on metric learning generally transfer the labels from the given 3D exemplars to the input ones by calculating the similarity or proximity in a shared space. Dissimilar to few-shot 2D segmentation allocating labels in an embedding space regularly, the approaches related to 3DS also have a propensity to conduct labels assignment in the input space.

From the perspective of taking the embedding space as the public metric space, prototype networks are also integrated into 3DS for grappling with limited training samples. The complicated point clouds in the support set were mapped as multiple prototypes of diverse shapes, which were then applied to measure the similarity with query embeddings in [106]. Multiple prototypes were estimated to propagate labels based on correlations between prototypes and query points [107].

Regarding 3D shape space as the common metric space, a conventional completion paradigm is to alter the given 3D exemplar into the same shape as the input and then to

transfer labels from the deformed exemplar to the input for segmentation, as indicated in Fig. 8. In [108], labeled meshes were first morphed to reap shapes that were the same as inputs. Then, a label migration was conducted according to the distance between points sampled from the deformed and input 3D meshes. The 3D template was transformed towards the input shape, which facilitated the subsequent label alignment operation [109]. Moreover, instead of calculating the spatial distance between two points directly, a probability distribution function was learned to assist the predictions on part-specific labels.

2) *Parameter Prediction-based Methods*: Parameter prediction-based methods were also applied to tackle few-shot 3DS. A potential space associated with diverse 3DS functions was learned from a significant number of tasks, where the limited training data about unseen shapes was leveraged to dynamically generate a task-specific function for rapid generalization on target shapes in [110].

3) *Fine-tune-based Methods*: There are also some approaches that have recourse to fine-tune policies to cope with few-shot 3DS. In addition to meta-learning [111], pre-training paradigms constructed on self-supervised [112], [113] or unsupervised learning [114] are also integrated into few-shot 3DS.

Under the support of contrastive learning and meta-learning, pre-training offline was conducted to represent more discriminative features on shapes and estimate the optimal model initialization in [111]. Based on the initialized parameters, the model could effectively adjust itself and converge quickly on new tasks with a handful of training samples. A self-supervised loss function was developed in [112] to assist in the estimation on dimension-fixed feature vectors of 3D points. When a small amount of training data was provided, the segmentation model initialized by the pre-trained embedding network could be further fine-tuned for performance gains. A self-supervised strategy was adopted in [113] to heuristically represent a point cloud exemplar in a metric space ahead of schedule. The learned point embedding network was then employed to learn the segmentation network under a few samples. In [114], a novel autoencoder was designed, where the encoder aimed

to represent 3D shape in a feature space, and the branched decoder endeavored to learn a compact representation for each frequently occurring shape. Under this setting, the model pre-trained in an unsupervised way could quickly acclimate to the segmentation of new 3D shapes by an uncomplicated adjustment with the guidance of only one or a few labeled training data.

B. Zero-shot 3D Segmentation

Existing zero-shot 3DS approaches generally adopt generative models to cope with unseen 3D objects with zero labels, which also follows the learning paradigm illustrated in Fig. 5. Similar to generative model-based ISS methods [53], [55], a generator was trained on the seen training data to generate pseudo but semantically consistent features for the unseen point cloud samples [116]. The generated pseudo representations were employed to fine-tune a classifier for overcoming the challenges arising from the absence of unseen point cloud categories and disequilibrium size between seen and unseen samples. Analogous ideas were also adopted to tackle the zero-shot 3D scene segmentation in [117], where a regularizer was designed to assist the generation on semantically consistent features of both seen and unseen classes. These synthesized features were then leveraged to adjust the classifier for accurate prediction on unseen objects.

C. Summary

Table VI sums up the few/zero-shot 3DS methods reviewed in this paper. It can be discovered that compared with ISS and VOS, the cooperation between few-shot learning (or zero-shot learning) and 3DS is still in its infancy. Due to the disorder and complexity of 3D samples, these technical solutions that show great effectiveness in 2D space are tricky to be applied in 3D space directly, which limits the development of few/zero-shot 3DS. How to deal with this challenge effectively with a particular solution is still a promising topic in the future.

VI. DISCUSSION

In spite of the fact that few/zero-shot visual semantic segmentation has achieved a considerable breakthrough and effectively settled the issues caused by a few or even zero annotated samples in both 2D and 3D space, there are still a variety of obstacles to their applications in real scenes. This section aims to discuss some open challenges and list some applications of few/zero-shot visual semantic segmentation.

Cross-domain transferability: Transferability plays a crucial role for the computer vision community [154]. For visual semantic segmentation, the transferability is principally reflected in two aspects: crossing different domains and crossing diverse categories, where the cross-category transferability has received extensive attention in few/zero-shot visual semantic segmentation. However, the existing methods are required to follow a solid premise: the samples of unseen classes must have the same distribution as base classes, which places restrictions on the serviceability in real scenarios [42]. Although visual semantic segmentation methods based on domain adaptation can bridge the inter-domain gap, they fail to generalize

to unseen categories effectively. Therefore, integrating domain adaptation with cross-category adaptation seamlessly (i.e., adopting adversarial training with few-shot learning [155]) can be a promising way to significantly boost the performance of few/zero-shot segmentation models in different domains.

Generalized few/zero-shot segmentation: Generalized few/zero-shot visual segmentation requires the model to realize the segmentation on both base and unseen classes. Most of few/zero-shot visual segmentation algorithms tend to deal with the segmentation of unseen categories while ignoring the performance on base classes in inference. For example, the methods based on parameter prediction adjust the classifier weights only for a specified category, which may lead to performance degradation on base classes. Furthermore, practical applications also demand that the segmentation model should be able to segment multiple categories of objects, including seen and unseen categories. Therefore, pursuing generalized few/zero-shot segmentation algorithms is one of the valuable future topics.

Weakly-supervised few/zero-shot segmentation: Even though the existing few/zero-shot visual segmentation approaches have greatly alleviated the requirements for annotations of target objects, a large number of base-class samples with labels are still indispensable [26], [77], [78] to learn to separate unseen categories of interest. However, it is difficult and challenging to collect such a great deal of labeled base-class samples in practice. The strictness of practical applications entails that few/zero-shot visual semantic segmentation methods are competent to precisely separate unseen categories in a more efficient way. Consequently, combining weakly-supervised learning, or even unsupervised learning [156], with few/zero-shot visual semantic segmentation will be a constructive subject in further development.

Multi-modal supervision: Most of few-shot visual semantic segmentation approaches are inclined to rely on insufficient supervision signals from a single-modal support set, which makes them tricky and laborious to deal with the gap between support and query samples effectively. For example, in few-shot ISS, the type of support samples is all 2D images. However, other modalities, such as language expressions, also can be used as auxiliary supervision signals in addition to images. The zero-shot visual semantic segmentation utilizes the supervisory information from other modalities, such as word embeddings and optical flows, which effectively fills the vacancy in supervision information and conducts reliable predictions for target categories. Therefore, it is promising for few-shot visual semantic segmentation to turn to other modalities to augment the supervision information. In addition, the utilization of more modalities is also of great significance to further promote the richness of supervision information for zero-shot visual semantic segmentation.

Lightweight network: Nowadays, embedded devices play a significant role in both civil and military fields, which puts forward higher requirements on the computational overhead and runtime costs. Whereas, the majority of few/zero-shot visual semantic segmentation approaches resort to different backbones, such as VGG16 [157], or more complicated structure for cross-class prediction, which requires higher com-

putational costs and memory usage and is disadvantageous for deployments. Moreover, although real-time performance can be achieved by decreasing the model size, undesirable performance degradation may be inevitably incurred [158]. Therefore, how to design a lightweight network with both real-time performance and high-quality prediction is a direction worthy of efforts for few/zero-shot visual semantic segmentation.

Cross-task collaboration: It is natural that the role of a few/zero-shot visual semantic segmentation model is insufficient to contend with diverse challenges in real scenarios. Therefore, combining visual semantic segmentation with other computer vision tasks will be a promising direction to expand the abilities of the visual semantic segmentation model in some aspects. For instance, it is advantageous to leverage few-shot object detection to provide more efficient annotations for weakly-supervised few-shot visual semantic segmentation [159]. Moreover, collaborating across diverse tasks will also promote cooperation and information sharing between various tasks, promising to maximize the value of limited samples.

Cross-task learning: As summarized above, the technical solutions of ISS, VOS, and 3DS in both few-shot and zero-shot scenarios have some commonalities and universalities, which makes the solutions proposed in one segmentation field can be referenced by another one. For example, on the one hand, some methods [33], [34], which are proposed to solve the appearance gap between support and query samples in few-shot ISS, can be used to deal with long-time sequences in few-shot VOS. On the other hand, some strategies [62], [75], which are proposed to address mismatching of similar objects in few-shot VOS, can also be considered in few-shot ISS.

Applications: Compared with conventional visual semantic segmentation, few/zero-shot semantic segmentation algorithms have more significant application advantages. Take the petrochemical industry for example. Firstly, few/zero-shot semantic segmentation methods can be applied in the situations where labeled training data is scarce, such as oil and gas pipeline leakage detection [160]. Secondly, few/zero-shot visual-semantic segmentation provides a flexible solution for fast cross-class adaptation, which plays an important role in addressing diverse types of pipeline defects, such as deformation, corrosion, scaling and cracks [161]. Thirdly, combining the limited number of cross-modal samples (i.e., infrared images) and breaking through the barriers of multi-modal few/zero-shot semantic segmentation methods can advance the applications in all-weather scenarios, such as safety production monitoring of petroleum and petrochemical products [162], [163]. In a word, employing few/zero-shot semantic segmentation in real scenarios, such as petrochemical industry, plays a positive role in promoting the production and life to become more intelligent and autonomous.

VII. CONCLUSION

This paper focuses on the applications of few-shot learning on visual semantic segmentation. To this end, we perform an exhaustive and systematic survey on related works of three typical few/zero-shot segmentation tasks, including few/zero-shot

ISS, few/zero-shot VOS, and few/zero-shot 3DS. Moreover, we explore the commonalities and discrepancies of few/zero-shot segmentations under different segmentation circumstances. Besides, we analyze these existing few/zero-shot segmentation methods and list some challenges and valuable directions for follow-up researchers.

REFERENCES

- [1] L. Chen, X. Hu, W. Tian, H. Wang, D. Cao, and F. Wang, "Parallel planning: A new motion planning framework for autonomous driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 236–246, 2018.
- [2] C. Sun, J. M. U. Vianney, Y. Li, L. Chen, L. Li, F. Wang, A. Khajepour, and D. Cao, "Proximity based automatic data annotation for autonomous driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 2, pp. 395–404, 2020.
- [3] W. He, M. Liu, Y. Tang, Q. Liu, and Y. Wang, "Differentiable automatic data augmentation by proximal update for medical image segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1315–1318, 2022.
- [4] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.
- [5] Q. Cheng, Y. Zhou, H. Huang, and Z. Wang, "Multi-attention fusion and fine-grained alignment for bidirectional image-sentence retrieval in remote sensing," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1532–1535, 2022.
- [6] C. Zhang, Y. Tang, C. Zhao, Q. Sun, Z. Ye, and J. Kurths, "Multitask GANs for semantic segmentation and depth completion with cycle consistency," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5404–5415, 2021.
- [7] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [8] P. Patil, A. Dudhane, A. Kulkarni, S. Murala, A. Gonde, and S. Gupta, "An unified recurrent video object segmentation framework for various surveillance environments," *IEEE Transactions on Image Processing*, vol. 30, pp. 7889–7902, 2021.
- [9] S. Garg and V. Goel, "Mask selection and propagation for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1680–1690.
- [10] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Learning semantic segmentation of large-scale point clouds with random sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [11] G. Te, W. Hu, A. Zheng, and Z. Guo, "Rgcnn: Regularized graph cnn for point cloud segmentation," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 746–754.
- [12] X. Chen, A. Golovinskiy, and T. Funkhouser, "A benchmark for 3D mesh segmentation," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–12, 2009.
- [13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [14] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proceedings of the British Machine Vision Conference*, 2017.
- [15] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 622–631.
- [16] N. Dong and E. Xing, "Few-shot semantic segmentation with prototype learning," in *Proceedings of the British Machine Vision Conference*, vol. 3, no. 4, 2018.
- [17] K. Wang, J. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9197–9206.
- [18] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

- [19] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8057–8067.
- [20] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," 2018.
- [21] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, 2020.
- [22] T. Chen, G. Xie, Y. Yao, Q. Wang, F. Shen, Z. Tang, and J. Zhang, "Semantically meaningful class prototype learning for one-shot image segmentation," *IEEE Transactions on Multimedia*, vol. 24, pp. 968–980, 2022.
- [23] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 8441–8448.
- [24] Y. Yang, F. Meng, H. Li, Q. Wu, X. Xu, and S. Chen, "A new local transformation module for few-shot segmentation," in *Proceedings of the International Conference on Multimedia Modeling*, vol. 11962, 2020, pp. 76–87.
- [25] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-One: Similarity guidance network for one-shot semantic segmentation," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.
- [26] Y. Liu, B. Jiang, and J. Xu, "Axial assembled correspondence network for few-shot semantic segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, pp. 1–11, 2022.
- [27] B. Yang, F. Wan, C. Liu, B. Li, X. Ji, and Q. Ye, "Part-based semantic transform for few-shot semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [28] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8334–8343.
- [29] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, vol. 12354, 2020, pp. 142–158.
- [30] X. Zhang, Y. Wei, Z. Li, C. Yan, and Y. Yang, "Rich embedding features for one-shot semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [31] H. Wang, Y. Yang, X. Cao, X. Zhen, C. Snoek, and L. Shao, "Variational prototype inference for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 525–534.
- [32] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9587–9595.
- [33] H. Wang, X. Zhang, Y. Hu, Y. Yang, X. Cao, and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 730–746.
- [34] B. Liu, J. Jiao, and Q. Ye, "Harmonic feature activation for few-shot semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3142–3153, 2021.
- [35] G. Zhang, G. Kang, Y. Yang, and Y. Wei, "Few-shot segmentation via cycle-consistent transformer," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [36] H. Gao, J. Xiao, Y. Yin, T. Liu, and J. Shi, "A mutually supervised graph attention network for few-shot segmentation: The perspective of fully utilizing limited samples," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [37] X. Shi, D. Wei, Y. Zhang, D. Lu, M. Ning, J. Chen, K. Ma, and Y. Zheng, "Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation," *arXiv preprint arXiv:2207.08549*, 2022.
- [38] Y. Zhuge and C. Shen, "Deep reasoning network for few-shot semantic segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5344–5352.
- [39] L. Liu, J. Cao, M. Liu, Y. Guo, Q. Chen, and M. Tan, "Dynamic extension nets for few-shot semantic segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1441–1449.
- [40] X. Yang, B. Wang, K. Chen, X. Zhou, S. Yi, W. Ouyang, and L. Zhou, "BriNet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation," in *Proceedings of the British Machine Vision Conference*, 2020.
- [41] P. Tian, Z. Wu, L. Qi, L. Wang, Y. Shi, and Y. Gao, "Differentiable meta-learning model for few-shot semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 087–12 094.
- [42] M. Boudiaf, H. Kervadec, Z. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 979–13 988.
- [43] Z. Wu, X. Shi, G. Lin, and J. Cai, "Learning meta-class memory for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 517–526.
- [44] G. Xie, H. Xiong, J. Liu, Y. Yao, and L. Shao, "Few-shot semantic segmentation with cyclic memory network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7293–7302.
- [45] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtaşun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891–898.
- [46] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero- and few-label semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8256–8265.
- [47] G. Tian, S. Wang, J. Feng, L. Zhou, and Y. Mu, "Cap2seg: Inferring semantic and spatial context from captions for zero-shot image segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4125–4134.
- [48] H. Lu, L. Fang, M. Lin, and Z. Deng, "Feature enhanced projection network for zero-shot semantic segmentation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2021, pp. 14 011–14 017.
- [49] P. Hu, S. Sclaroff, and K. Saenko, "Uncertainty-aware learning for zero-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 713–21 724, 2020.
- [50] N. Kato, T. Yamasaki, and K. Aizawa, "Zero-shot semantic segmentation via variational mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [51] D. Baek, Y. Oh, and B. Ham, "Exploiting a joint embedding space for generalized zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9536–9545.
- [52] M. Bucher, T. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [53] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1921–1929.
- [54] —, "From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [55] P. Li, Y. Wei, and Y. Yang, "Consistent structural relation learning for zero-shot segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 317–10 327, 2020.
- [56] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-VOS: Sequence-to-sequence video object segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 585–601.
- [57] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.
- [58] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.
- [59] Y. Yin, D. Xu, X. Wang, and L. Zhang, "Directional deep embedding and appearance learning for fast video object segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [60] P. Hu, G. Wang, X. Kong, J. Kuen, and Y. Tan, "Motion-guided cascaded refinement network for video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 08, pp. 1957–1967, 2020.

- [61] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L. Chen, "FEELVOS: Fast end-to-end embedding learning for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9481–9490.
- [62] Z. Yang, Y. Wei, and Y. Yang, "Collaborative video object segmentation by foreground-background integration," in *Proceedings of the European Conference on Computer Vision*, vol. 12350, 2020, pp. 332–348.
- [63] W. Zhu, J. Li, J. Lu, and J. Zhou, "Separable structure modeling for semi-supervised video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 330–344, 2022.
- [64] Y. Zhang, Z. Wu, H. Peng, and S. Lin, "A transductive approach for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6949–6958.
- [65] W. Liu, G. Lin, T. Zhang, and Z. Liu, "Guided co-segmentation network for fast video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1607–1617, 2020.
- [66] L. Hong, W. Zhang, L. Chen, W. Zhang, and J. Fan, "Adaptive selection of reference frames for video object segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 1057–1071, 2022.
- [67] H. Park, J. Yoo, S. Jeong, G. Venkatesh, and N. Kwak, "Learning dynamic network using a reuse gate function in semi-supervised video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8405–8414.
- [68] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for video object segmentation," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1175–1197, 2019.
- [69] A. Robinson, F. Lawin, M. Danelljan, F. Khan, and M. Felsberg, "Learning fast and robust target models for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7406–7415.
- [70] H. Xiao, B. Kang, Y. Liu, M. Zhang, and J. Feng, "Online meta adaptation for fast video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1205–1217, 2019.
- [71] T. Meinhardt and L. Leal-Taixé, "Make one-shot video object segmentation efficient again," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [72] C. Xu, L. Wei, Z. Cui, T. Zhang, and J. Yang, "Meta-VOS: Learning to adapt online target-specific segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4760–4772, 2021.
- [73] S. Oh, J. Lee, N. Xu, and S. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9226–9235.
- [74] —, "Space-time memory networks for video object segmentation with user guidance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 442–455, 2022.
- [75] H. Seong, J. Hyun, and E. Kim, "Kernelized memory network for video object segmentation," in *Proceedings of European Conference on Computer Vision*, vol. 12367, 2020, pp. 629–645.
- [76] —, "Video object segmentation using kernelized memory network with multiple kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [77] H. Xie, H. Yao, S. Zhou, S. Zhang, and W. Sun, "Efficient regional memory network for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1286–1295.
- [78] H. Seong, S. Oh, J. Lee, S. Lee, S. Lee, and E. Kim, "Hierarchical memory matching network for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12889–12898.
- [79] Y. Li, Z. Shen, and Y. Shan, "Fast video object segmentation using the global context module," in *Proceedings of the European Conference on Computer Vision*, vol. 12355, 2020, pp. 735–750.
- [80] Y. Liang, X. Li, N. Jafari, and J. Chen, "Video object segmentation with adaptive feature bank and uncertain-region refinement," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3430–3441, 2020.
- [81] H. Cheng, Y. Tai, and C. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [82] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Gool, "Video object segmentation with episodic graph memory networks," in *Proceedings of European Conference on Computer Vision*, vol. 12348, Springer, 2020, pp. 661–679.
- [83] K. Zhang, L. Wang, D. Liu, B. Liu, Q. Liu, and Z. Li, "Dual temporal memory network for efficient video object segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1515–1523.
- [84] Y. Lyu, G. Vosselman, G. Xia, and M. Ying Yang, "LIP: Learning instance propagation for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [85] Z. Yang, Q. Wang, L. Bertinetto, W. Hu, S. Bai, and P. Torr, "Anchor diffusion for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 931–940.
- [86] D. Liu, D. Yu, C. Wang, and P. Zhou, "F2Net: Learning to focus on the foreground for unsupervised video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2109–2117.
- [87] Y. Gu, L. Wang, Z. Wang, Y. Liu, M. Cheng, and S. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10869–10876.
- [88] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo, "Zero-shot video object segmentation with co-attention siamese networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2228–2242, 2022.
- [89] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Unsupervised online video object segmentation with motion property understanding," *IEEE Transactions on Image Processing*, vol. 29, pp. 237–249, 2019.
- [90] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, and L. Shao, "Motion-attentive transition for zero-shot video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 13066–13073.
- [91] Y. Zhou, X. Xu, F. Shen, X. Zhu, and H. Shen, "Flow-edge guided unsupervised video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
- [92] S. Yang, L. Zhang, J. Qi, H. Lu, S. Wang, and X. Zhang, "Learning motion-appearance co-attention for zero-shot video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1564–1573.
- [93] G. Ji, K. Fu, Z. Wu, D. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4922–4933.
- [94] X. Zhao, Y. Pang, J. Yang, L. Zhang, and H. Lu, "Multi-source fusion and automatic predictor selection for zero-shot video object segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2645–2653.
- [95] S. Seo, J. Lee, and B. Han, "URVOS: Unified referring video object segmentation network with a large-scale benchmark," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 208–223.
- [96] M. Bellver, C. Ventura, C. Silberer, I. Kazakos, J. Torres, and X. Giro-i Nieto, "RefVOS: A closer look at referring expressions for video object segmentation," *arXiv preprint arXiv:2010.00263*, 2020.
- [97] Z. Yang, Y. Tang, L. Bertinetto, H. Zhao, and P. H. Torr, "Hierarchical interaction network for video object segmentation from referring expressions," in *Proceedings of the British Machine Vision Conference*, 2021.
- [98] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4985–4995.
- [99] D. Li, R. Li, L. Wang, Y. Wang, J. Qi, L. Zhang, T. Liu, Q. Xu, and H. Lu, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- [100] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i Nieto, "RVOS: End-to-end recurrent network for video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5277–5286.
- [101] P. Tokmakov, C. Schmid, and K. Alahari, "Learning to segment moving objects," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 282–301, 2019.
- [102] S. Mahadevan, A. Athar, A. Ošep, S. Hennen, L. Leal-Taixé, and B. Leibe, "Making a case for 3D convolutions for object segmentation in videos," in *Proceedings of the British Machine Vision Conference*, 2020.

- [103] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9236–9245.
- [104] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 605–613.
- [105] A. Dai, A. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839.
- [106] X. Chen, C. Zhang, G. Lin, and J. Han, "Compositional prototype network with multi-view comparison for few-shot point cloud semantic segmentation," *arXiv preprint arXiv:2012.14255*, 2020.
- [107] N. Zhao, T. Chua, and G. Lee, "Few-shot 3D point cloud semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8873–8882.
- [108] S. Yuan and Y. Fang, "ROSS: Robust learning of one-shot 3D shape segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1961–1969.
- [109] L. Wang, X. Li, and Y. Fang, "Few-shot learning of part-specific probability space for 3D shape segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4504–4513.
- [110] Y. Hao and Y. Fang, "3D meta-segmentation neural network," *arXiv preprint arXiv:2110.04297*, 2021.
- [111] H. Huang, X. Li, L. Wang, and Y. Fang, "3D-MetaConNet: Meta-learning for 3D shape classification and segmentation," in *Proceedings of the International Conference on 3D Vision*, 2021, pp. 982–991.
- [112] G. Sharma, E. Kalogerakis, and S. Maji, "Learning point embeddings from shape repositories for few-shot segmentation," in *Proceedings of the International Conference on 3D Vision*, 2019, pp. 67–75.
- [113] C. Sharma and M. Kaul, "Self-supervised few-shot learning on point clouds," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7212–7221, 2020.
- [114] Z. Chen, K. Yin, M. Fisher, S. Chaudhuri, and H. Zhang, "BAE-NET: Branched autoencoder for shape co-segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8490–8499.
- [115] I. Armeni, O. Sener, A. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1534–1543.
- [116] B. Michele, A. Boulch, G. Puy, M. Bucher, and R. Marlet, "Generative zero-shot learning for semantic segmentation of 3D point clouds," in *Proceedings of the International Conference on 3D Vision*, 2021, pp. 992–1002.
- [117] B. Liu, S. Deng, Q. Dong, and Z. Hu, "Segmenting 3D hybrid scenes via zero-shot learning," *arXiv preprint arXiv:2107.00430*, 2021.
- [118] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *arXiv preprint arXiv:2101.06085*, 2021.
- [119] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.
- [120] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.
- [121] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F. Wang, "FISS GAN: A generative adversarial network for foggy image semantic segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 8, pp. 1428–1439, 2021.
- [122] M. Cheng, L. Hui, J. Xie, and J. Yang, "SSPC-Net: Semi-supervised semantic 3d point cloud segmentation network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1140–1147.
- [123] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [124] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [125] Q. Sun, C. Zhao, Y. Tang, and F. Qian, "A survey on unsupervised domain adaptation in computer vision tasks," *SCIENTIA SINICA Technologica*, vol. 52, no. 1, pp. 26–54, 2022.
- [126] Y. Zhang, Z. Li, Y. Xie, Y. Qu, C. Li, and T. Mei, "Weakly supervised semantic segmentation for large-scale point cloud," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3421–3429.
- [127] B. Kim, S. Han, and J. Kim, "Discriminative region suppression for weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1754–1761.
- [128] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27408–27421, 2021.
- [129] Y. Wang, Q. Yao, J. Kwok, and L. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.
- [130] W. Wang, V. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–37, 2019.
- [131] Y. Hu, A. Chapman, G. Wen, and D. Hall, "What can knowledge bring to machine learning?—a survey of low-shot learning for structured data," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 3, pp. 1–45, 2022.
- [132] S. Luo, Y. Li, P. Gao, Y. Wang, and S. Serikawa, "Meta-Seg: A survey of meta-learning for image segmentation," *Pattern Recognition*, p. 108586, 2022.
- [133] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 4, pp. 1–47, 2020.
- [134] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. Van Gool, "A survey on deep learning technique for video segmentation," *arXiv preprint arXiv:2107.01153*, 2021.
- [135] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [136] D. Wang, Y. Li, Y. Lin, and Y. Zhuang, "Relational knowledge transfer for zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- [137] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022.
- [138] S. Antonelli, D. Avola, L. Cinque, D. Crisostomi, G. Foresti, F. Galasso, M. Marini, A. Mecca, and D. Pannone, "Few-shot object detection: A survey," *ACM Computing Surveys*, 2021.
- [139] Y. Song, T. Wang, S. Mondal, and J. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *arXiv preprint arXiv:2205.06743*, 2022.
- [140] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 297–312.
- [141] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [142] J. Lu, P. Gong, J. Ye, and C. Zhang, "Learning from very few samples: A survey," *arXiv preprint arXiv:2009.02653*, 2020.
- [143] Y. Guo, N. Codella, L. Karlinsky, J. Codella, J. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 124–141.
- [144] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. Lim, and X. Wang, "A review of generalized zero-shot learning methods," *arXiv preprint arXiv:2011.08641*, 2020.
- [145] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.
- [146] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2582–2593.
- [147] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7303–7313.
- [148] U. Michieli and M. Ozay, "Prototype guided federated learning of visual feature representations," *arXiv preprint arXiv:2105.08982*, 2021.
- [149] J. Hwang, S. Yu, J. Shi, M. Collins, T. Yang, X. Zhang, and L. Chen, "Segsort: Segmentation by discriminative sorting of segments," in

Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7334–7344.

- [150] L. Ke, X. Li, M. Danelljan, Y. Tai, C. Tang, and F. Yu, “Prototypical cross-attention networks for multiple object tracking and segmentation,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 1192–1203, 2021.
- [151] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [152] A. Jain, R. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.
- [153] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 1126–1135.
- [154] C. Zhang, J. Wang, G. Yen, C. Zhao, Q. Sun, Y. Tang, F. Qian, and J. Kurths, “When autonomous systems meet accuracy and transferability through AI: A survey,” *Patterns*, vol. 1, no. 4, p. 100050, 2020.
- [155] A. Zhao, M. Ding, Z. Lu, T. Xiang, Y. Niu, J. Guan, and J. Wen, “Domain-adaptive few-shot learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1390–1399.
- [156] M. S. Amac, A. Sencan, B. Baran, N. Ikizler-Cinbis, and R. G. Cinbis, “MaskSplit: Self-supervised meta-learning for few-shot semantic segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1067–1077.
- [157] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the International Conference on Learning Representations*, 2015.
- [158] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, “Monocular depth estimation based on deep learning: An overview,” *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612–1627, 2020.
- [159] S. Bonechi, P. Andreini, M. Bianchini, and F. Scarselli, “Generating bounding box supervision for semantic segmentation with deep learning,” in *Artificial Neural Networks in Pattern Recognition*, 2018, pp. 190–200.
- [160] Q. Lu, Q. Li, L. Hu, and L. Huang, “An effective low-contrast sf6 gas leakage detection method for infrared imaging,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.
- [161] P. Ravishankar, S. Hwang, J. Zhang, I. Khalilullah, and B. Eren-Tokgoz, “DARTS—drone and artificial intelligence reconsolidated technological solution for increasing the oil and gas pipeline resilience,” *International Journal of Disaster Risk Science*, pp. 1–12, 2022.
- [162] W. Zhou, J. Liu, J. Lei, L. Yu, and J. Hwang, “GMNet: Graded-feature multilabel-learning network for RGB-Thermal urban scene semantic segmentation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.
- [163] P. Patil, A. Dudhane, S. Chaudhary, and S. Murala, “Multi-frame based adversarial learning approach for video surveillance,” *Pattern Recognition*, vol. 122, p. 108350, 2022.



Wenqi Ren received the B.S. degree in electrical engineering and automation, from the East China University of Science and Technology, Shanghai, China, in 2020, where she is currently working toward the Ph.D. degree in control science and engineering. Her research interests include meta-learning, few-shot learning, domain adaptation, and scene understanding.



Yang Tang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Donghua University in 2006 and 2010, respectively. From 2008 to 2010, he was a Research Associate with The Hong Kong Polytechnic University, China. From 2011 to 2015, he was a Post-Doctoral Researcher with the Humboldt University of Berlin, Germany, and with the Potsdam Institute for Climate Impact Research, Germany. He is now a Professor with the East China University of Science and Technology. His current research interests

include distributed estimation/control/optimization, cyber-physical systems, hybrid dynamical systems, computer vision, reinforcement learning and their applications.

He was a recipient of the Alexander von Humboldt Fellowship and has been the ISI Highly Cited Researchers Award by Clarivate Analytics from 2017. He is a Senior Board Member of Scientific Reports, an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Circuits and Systems-I: Regular Papers, IEEE Transactions on Emerging Topics in Computational Intelligence, IEEE Systems Journal and Engineering Applications of Artificial Intelligence (IFAC Journal), etc. He is a Leading Guest Editor for special issues in IEEE Transactions on Emerging Topics in Computational Intelligence and IEEE Transactions on Cognitive and Developmental Systems.



Qiyu Sun received the B.S. degree in automation, from the East China University of Science and Technology, Shanghai, China, in 2019, where she is currently working toward the Ph.D. degree in control science and engineering. Her research interests include 3D scene understanding, domain adaptation, and deep learning.



Chaoqiang Zhao received the B.S. degree in automation from the East China University of Science and Technology, Shanghai, China, in 2018, where he is currently working toward the Ph.D. degree in control science and engineering. His research interests include depth perception, visual odometry, and stereo vision.



Qing-Long Han (Fellow, IEEE) received the B.Sc. degree in Mathematics from Shandong Normal University, Jinan, China, in 1983, and the M.Sc. and Ph.D. degrees in Control Engineering from East China University of Science and Technology, Shanghai, China, in 1992 and 1997, respectively.

Professor Han is Pro Vice-Chancellor (Research Quality) and a Distinguished Professor at Swinburne University of Technology, Melbourne, Australia. He held various academic and management positions at Griffith University and Central Queensland University, Australia. His research interests include networked control systems, multi-agent systems, time-delay systems, smart grids, unmanned surface vehicles, and neural networks.

Professor Han was awarded The 2021 Norbert Wiener Award (the Highest Award in systems science and engineering, and cybernetics) and The 2021 M. A. Sargent Medal (the Highest Award of the Electrical College Board of Engineers Australia). He was the recipient of The 2022 IEEE SMC Society Andrew P. Sage Best Transactions Paper Award, The 2021 IEEE/CAA Journal of Automatica Sinica Norbert Wiener Review Award, The 2020 IEEE Systems, Man, and Cybernetics Society Andrew P. Sage Best Transactions Paper Award, The 2020 IEEE Transactions on Industrial Informatics Outstanding Paper Award, and The 2019 IEEE Systems, Man, and Cybernetics Society Society Andrew P. Sage Best Transactions Paper Award.

Professor Han is a Member of the Academia Europaea (The Academy of Europe). He is a Fellow of The International Federation of Automatic Control (IFAC) and a Fellow of The Institution of Engineers Australia (IEAust). He is a Highly Cited Researcher in both Engineering and Computer Science (Clarivate Analytics). He has served as an AdCom Member of IEEE Industrial Electronics Society (IES), a Member of IEEE IES Fellows Committee, and Chair of IEEE IES Technical Committee on Networked Control Systems. Currently, he is Co-Editor-in-Chief of IEEE Transactions on Industrial Informatics, Deputy Editor-in-Chief of IEEE/CAA JOURNAL OF AUTOMATICA SINICA, Co-Editor of Australian Journal of Electrical and Electronic Engineering, an Associate Editor for 12 international journals, including the IEEE TRANSACTIONS ON CYBERNETICS, IEEE INDUSTRIAL ELECTRONICS MAGAZINE, Control Engineering Practice, and Information Sciences, and a Guest Editor for 14 Special Issues.