Preprints.org

Article

Not peer-reviewed version

Text-to-Image Segmentation with Open-Vocabulary and Multitasking

Lihu Pan *, Yunting Yang *, Zhengkui Wang, Rui Zhang

Posted Date: 9 April 2024

doi: 10.20944/preprints202404.0631.v1

Keywords: image segmentation; open vocabulary; multitask; multi-modal interaction

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Text-to-Image Segmentation with Open-Vocabulary and Multitasking

Lihu Pan [1,*], Yunting Yang [1], Zhengkui Wang [2] and Rui Zhang [1]

[1] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan, 030024, China; yangyunting606@163.com(Y.Y.); zhangrui@tyust.edu.cn(R.Z.)

[2] InfoComm Technology Cluster, Singapore Institute of Technology, 138683, Singapore; zhengkui.wang@singaporetech.edu.sg

* Correspondence: panlh@tyust.edu.cn

Abstract: Open-vocabulary learning has recently gained prominence as a means to enable image segmentation for arbitrary categories based on textual descriptions. This advancement has extended the applicability of segmentation systems to a broader range of generally purpose scenarios. However, current methods often revolve around specialized architectures and parameters tailored to specific segmentation tasks, resulting in a fragmented landscape of segmentation models. In response to these challenges, we introduce OVAMTSeg, a versatile framework designed for Open-Vocabulary and Multitask Image Segmentation. OVAMTSeg harnesses adaptive prompt learning to empower the model to capture category-sensitive concepts, enhancing its robustness across diverse multi-task and scenario contexts. Text prompts are employed to effectively capture semantic and contextual features of the text, while cross-attention and cross-modal interactions enable the fusion of image and text features. Furthermore, a transformer-based decoder is incorporated for dense prediction. Extensive experimental results underscore the effectiveness of OVAMTSeg, showcasing its state-of-the-art performance and superior generalization capabilities across three segmentation tasks. Notable achievements include a 47.5 mIoU in referring expression segmentation, 51.6 mIoU on Pascal-VOC with four unseen classes, 46.6 mIoU on Pascal-Context in zero-shot segmentation, 65.9 mIoU on Pascal-5i, and 35.7 mIoU on COCO-20i datasets for one-shot segmentation.

Keywords: image segmentation; open vocabulary; multitask; multi-modal interaction

1. Introduction

Image segmentation represents a deeply explored and pivotal domain within the field of computer vision. Its primary objective is the simultaneous categorization and grouping of pixels belonging to distinct objects within an image. Recent strides in image segmentation owe their success largely to the availability of expansive dataset [1–3], meticulously annotated to include pixel-level masks and object category labels. However, these annotations, although invaluable, come at a significant cost in terms of time and labor. Consequently, the predefined categories within current segmentation tasks remain restricted in scope, far removed from the vast and diverse lexicon that humans employ to describe the complexities of the real world. Such limitations in the learning objectives of existing segmentation systems impose a substantial impediment to scalability, particularly when attempting to accommodate richer and more encompassing semantic nuances.

To address the inherent constraints of predefined categories and unlock the potential for handling custom-defined classes beyond the confines of training data, the paradigm of open-vocabulary learning has gained prominence. Open-vocabulary learning leverages the power of large-scale visual-language pre-training models, exemplified by prominent models like CLIP[4] and ALIGN[5], to compute semantic similarity between visual concepts and textual descriptions. Notably, a burgeoning body of research in segmentation based open-vocabulary studies[6,7] has emerged with the goal of devising task-specific architectures and parameters tailored to single segmentation tasks. For instance, ZSSeg[6] harnesses the capabilities of off-the-shelf pre-trained CLIP

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

models and demonstrates competitive performance in open-vocabulary semantic segmentation. However, when extending these methods to a broader spectrum of segmentation scenarios, they exhibit significant limitations. Firstly, a unified model cannot be seamlessly applied to address multiple segmentation tasks, necessitating retraining and the deployment of numerous custom models for diverse tasks. Additionally, while CLIPSeg [8] successfully handles multiple segmentation tasks within a compact framework, its reliance on fixed-format text prompts (e.g., "photos of...") may impose restrictions on the generalization of human language understanding in practical applications. Moreover, it lacks the adaptability to dynamically adjust the modality correlation degree for different tasks and data, and it is not inherently designed for open-vocabulary tasks.

In response to the challenges outlined above, we propose OVAMTSeg, a framework tailored for open-vocabulary and multitask image segmentation. OVAMTSeg, designed with precision and versatility in mind, is driven by two primary objectives:

(1) Multitasking: OVAMTSeg seamlessly adapts to a spectrum of tasks, encompassing referring expression segmentation, zero-shot, and one-shot image segmentation.

(2) Open-Vocabulary: OVAMTSeg exhibits the capacity to generalize across a wide array of segmentation categories, embracing the flexibility to accommodate arbitrary categories.

OVAMTSeg unfolds as a two-stage segmentation paradigm. The initial stage entails the extraction of universal mask proposals, while the subsequent stage is dedicated to the precise segmentation of these masks. Crucially, OVAMTSeg operates as a unified framework, cultivating a profound understanding of both textual and visual features for segmentation tasks, driven by text and image prompts. An adaptive prompt learning mechanism is introduced to encode category-specific concepts into the textual abstraction, endowing OVAMTSeg with the versatility to tackle diverse segmentation tasks spanning arbitrary categories, all within a single, unified model.

To further elevate its performance, OVAMTSeg augments the text encoder and integrates a multimodal interaction module. This strategic enhancement facilitates the dynamic adjustment of modality correlations, enabling a more adaptable fusion of text and image features across distinct tasks and datasets.

In summary, OVAMTSeg emerges as a task-flexible, category agnostic, and performance-driven framework. The following concisely lists our contributions:

- We introduce OVAMTSeg, a universal open-vocabulary framework renowned for its capacity to efficiently segment images based on arbitrary text or image prompts. OVAMTSeg effectively addresses the intricate challenges posed by zero-shot, one-shot, and referring expression segmentation tasks.
- •Adaptive prompt learning empowers OVAMTSeg to explicitly encode category-specific information into a compact textual abstraction, facilitating the model's adeptness in generalizing to diverse textual descriptions. Additionally, we enhance the text encoder and introduce a multimodal interaction module to optimize cross-model fusion.
- Our model's efficiency and effectiveness are meticulously demonstrated through comprehensive evaluations across various benchmark datasets. Extensive experimental results conclusively establish that our proposed model surpasses current standards by a substantial margin, rendering it a highly viable choice for multitask deployment.

2. Related Work

2.1. Open Vocabulary Segmentation

In recent years, deep learning techniques [9–14] have advanced image segmentation [15–20], especially in open-vocabulary segmentation, addressing unseen categories. Research divides into two main areas: mapping visual features to semantics and cross-modal alignment with pre-trained models. SPNet [7] uses a unique mapping strategy to project visual features onto a fixed semantic word coding matrix, facilitating the prediction of category probability distributions. ZS3Net [3] extends SPNet by mapping semantic space to visual space, generating pixel-level features for previously unseen categories and supervising the visual segmentation model. On the other hand,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

cross-modal alignment utilizes the strong zero-shot abilities of pre-trained cross-modal models like CLIP[4] for executing open vocabulary segmentation tasks. LSeg [21] calculates pixel-wise image features using a convolutional neural network and aligns them with text embeddings from a pre-trained text model. These methods leverage cross-modal alignment and pre-trained models to innovate open lexical segmentation.

2.2. Multitask Image Segmentation Architecture

Multitask image segmentation architecture aims to unify various segmentation tasks, eliminating the need for separate training modes. Leading universal segmentation methods like MaskFormer [22] treat segmentation as a mask classification problem, excelling in semantic and panoptic tasks. CLIPSeg [8] offers adaptability to new tasks using text or image prompts during inference, sparing the cost of retraining. This hybrid approach accommodates referring expression, zero-shot, and one-shot segmentation, informing our multitask segmentation model.

2.3. Prompt Learning

Prompt learning, originally a prominent concept in natural language processing [22–24], has gained widespread recognition in vision and visual language models [6,25]. CoOp [25] presents ongoing immediate enhancement of subsequent data to synchronize with pre-trained visual language models. DenseCLIP [6] fine-tunes the pre-trained text encoder by employing distinct prompt templates for functions like detection and segmentation, ensuring precise alignment of text and visual features. MAPLE [26] argues that prompts should be applied in all modes or branches to achieve optimal performance, deviating from the conventional practice of applying prompts in only one mode.

In the context of open-vocabulary segmentation tasks, prompt templates are meticulously crafted based on provided category labels and subsequently transformed into text embeddings. These embeddings are then deployed for alignment with the representations of unseen classes, facilitating the achievement of remarkable results.

3. Methods

Our proposed framework is centered on a unified open-vocabulary segmentation pardigm aimed at optimizing an all-encompassing model to excel in referring expression, one-shot, and zero-shot segmentation tasks spanning arbitrary categories. At its core, this framework is founded upon the CLIP model, serving as the foundational backbone as depicted in Figure 1. To enhance its capabilities, we have made significant improvements to both the text prompt and text encoder. Moreover, we have incorporated an efficient multi-modal interaction module for seamless fusion of image and text features. Additionally, we employ a compact and parametrically efficient transformer decoder to extend the model's functionality. We establish a connection between the decoder and the CLIP encoder, inspired by the U-Net network structure[27]. Activations at specific layers S are extracted from the visual encoder and then mapped onto the token embedding size D our decoder. To guide the decoder in segmenting the target, we employ the FiLM module[28] to modulate the input activation. The decoder generates binary segmentation by linearly projecting the output of its last transformer block. In our experiments, we use a projection dimension of 64. To leverage the semantic information of deep features, we extract the activation information of CLIP in the transformer module using S=[3,7,9], resulting in a three-layer decoder. While keeping the improved CLIP encoder frozen, we train the decoder to perform the image segmentation task.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

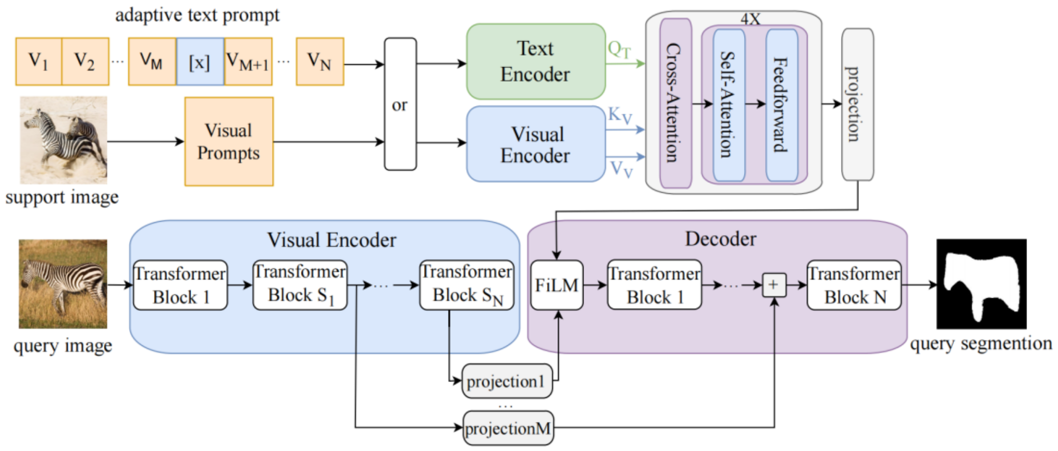

Figure 1. Architecture of OVAMTSeg: Our enhancement of the static CLIP model(green and blue) involves adding a transformer that divides the query image according to either a support image or a text prompt. Fusing text and visual features through a multi-modal interaction module(gray).

3.1. Adaptive Prompt Learning

Traditional approaches to prompt template design often rely on human linguistic expertise to craft smooth and effective semantic templates. While such hand-crafted templates are intuitive and easy to comprehend, they come with inherent limitations, demanding substantial manual effort and expertise, thus incurring high costs. In response to this challenge, we introduce an innovative adaptive prompt learning module. This module transforms diverse classes of text into a set of learnable vectors, which are subsequently amalgamated into text embeddings, simplifying the training of the model.

The introduction of adaptive prompts has not only broadened the applicability of our OVAMTSeg framework to a wider spectrum of unseen categories but has also significantly enhanced performance in open-domain scenarios. Specifically, the adaptive prompt P is generated based on the following template, where $[x]$ represents semantic categories or phrases.

$$P = [V]_1[V]_2 \cdots [V]_M[x][V]_{M+1} \cdots [V]_N, \tag{1}$$

Where each $[V]_n(n \in \{1,2,\ldots,N\})$ has the same dimension (512 dimensions) as the word embedding$[x]$, is a learnable vector, and N is a hyperparameter that specifies the number of context tokens. During the training, we are given the semantic categories or phrases involved. These adaptive prompts are then embedded in the pre-trained text encoderΨ:

$$E = \Psi\big(P([x])\big). \tag{2}$$

E represents the obtained text embedding vector. Given the high flexibility of the input categories or phrases, it can be seamlessly adapted to unseen categories in the Open-Vocabulary segmentation task.

3.2. Feature Extraction Dual-Encoder

Text Encoder. Text encoders are employed to convert text information into a high dimensional vector representation, allowing text and images to be compared in the singular embedded area. The text encoder adopts the ResNet-50 structure. Our modifications involve replacing the replacement of the global average pooling layer with an attention pooling mechanism, utilizing multi-head QKV attention. This enhancement enables the text encoder to effectively capture semantic and contextual features of the text, including cross-word associations, which improves its semantic representation capacity. The text encoder embeds N possible labels into an unbroken vector space $\mathrm{T} \in \mathbb{R}^{\mathrm{n_t} \times \mathrm{d_t}}$, resulting in N vectors$\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_\mathrm{N} \in \mathbb{R}^{\mathrm{n_t} \times \mathrm{d_t}}$, with their arrangement unaffected by the input label sequence. The number N is flexible and can vary freely.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

Image Encoder. Our image encoder utilizes the Visual Transformer It begins by converting the three-dimensional image $x \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$into a two-dimensional sequence $x_1 \in \mathbb{R}^{N\times(p^2\times C)}$, where$(H \times W)$represents the original image size, C is the channel count, and the image is divided into evenly sized (P, P) blocks. This results in $N = HW/P^2$ blocks, which also function as the actual length of input sequence for the Visual Transformer. Next, the patch sequence undergoes transformation into $1D$ tokens $\{f_i^v\}_{i=1}^n$ through a trainable linear projection. To capture positional information, positional embeddings and an additional $[CLS]$token are introduced. These tokens $f_{cls}^v, f_1^v, \cdots, f_N^v$ are then fed into S_N -layer transformer blocks to model correlations among individual patches, where $N = 11$. Finally, a linear projection maps f_{cls}^v to the combined space of image and text embedding, acting as the overarching representation of the image.

3.3. Multimodal Interaction Module

In order to achieve complete interaction between image and text modes, we designed an proficient multi-modal for merging image and text embedding. In contract to other well-known multi-modal interaction modules, our design of the model enhances the accuracy of image and text correspondence more accurately and improves the semantic understanding ability of the model. The multi-modal interaction module consists of a multi-head cross-attention layer and four transformer modules. Given text and supporting images. The text hidden representation $T = \{t_1, t_2, \cdots, t_s\}$, $T \in \mathbb{R}^{n_t \times d_t}$ and image hidden representation $I^* \in \mathbb{R}^{n_i \times d_i}$ are obtained by the text and image encoder respectively. $I^*$is the image feature mapping output at the last layer of the image encoder. Since the dimension of $I^*$is different from the dimension of T, the hidden representation dimension needs to be converted to the same dimension as T. It can be converted by the following formula:

$$I = I^* W_I + b_I, \tag{3}$$

Where $I = \{i_1, i_2, \cdots i_{n_i}\}$, $I \in \mathbb{R}^{n_i \times d_i}$. Then, the text T and image I are fed into the multi-modal interaction module. To fuse image and text representations more effectively, the text representation T is used as the query (Q), and the image representation I is used as the key (K) and value (V). This allows the model to adjust the correlation between patterns, thereby achieving more flexible text-image fusion for different tasks and data. Full interaction between image and text representations can be achieved by:

$$F = Transformer\big(MAC(Q, K, V)\big), \tag{4}$$

Where $MAC(\cdot)$ is multi-head cross-attention, which can be achieved in the following ways:

$$MAC(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{5}$$

where d is the embedding dimension of tokens.

4. Experimental Results

4.1. Experimental Settings

The experiments were conducted using the PyTorch [29] deep learning framework in Python 3.8, within the Anaconda3 environment, and implemented using the PyCharm compiler. The training of our model consisted of 20,000 iterations, utilizing the PhraseCut+ dataset. We employed a batch size of 64 and set the image size to 352×352 pixels. During the model's training process, we employed the AdamW optimization algorithm to iteratively optimize the network. The initial learning rate was set to 0.001. As the model approaches the global minimum of the loss function, the corresponding learning rate gradually decreases towards a minimum point. To achieve this, we utilized the Cosine Annealing method for dynamic learning rate adjustment without warm-up. The minimum learning rate was set to 0.0001. Simultaneously, we adopted the technique of automatic mixed precision to enhance the training process. Our model's loss function exclusively employed the binary cross-entropy function.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

4.2. Datasets

PhraseCut+ is employed as the primary dataset for our training process. It is an extension of the PhraseCut dataset [30], encompassing 345,486 phrase regions meticulously labeled with category names, attributes, and object relationships within the images. Each phrase corresponds to a binary segmentation mask in the image. PhraseCut+ enhances the original dataset by introducing visual support samples and negative samples. This augmentation makes it particularly well-suited for models that jointly process text and visual inputs. The dataset extends the phrase regions using automatically learned templates and ensures partial object visibility by considering object position information during random cropping.

Pascal-VOC2012 [31]serves as the basis for our zero-shot image segmentation experiments. Pascal-VOC2012 is a comprehensive dataset featuring images for training, validation, and testing. The validation subset consists of 1,449 images, each accompanied by pixel-level semantic segmentation masks used for image segmentation tasks. It includes annotations for 20 object categories along with a background category. We follow previous studies' practice [1,28] by partitioning these 20 categories into 16 seen classes and 4 unseen classes ("cow," "motorbike," "airplane," and "sofa") to evaluate open word segmentation performance. Additionally, we employ the **Pascal-5i dataset** [21], a one-shot segmentation dataset derived from Pascal-VOC2012, which contains 20 object classes evenly distributed across four sections.

For our zero-shot segmentation experiment, we employ the **Pascal-Context dataset** [32], which consists of 5,105 validation samples, spanning 59 item classes in addition to a background class. Similarly to Pascal-VOC, we select four unseen categories ("cow," "motorbike," "sofa," and "cat") from the foreground category for evaluation.

The **COCO-20i dataset** [33] is commonly used for one-shot segmentation tasks and is derived from the COCO dataset. It includes annotations for 80 object classes evenly divided into four sections, each containing 20 categories.

The test results involve selecting the best-performing group from among these four sections. These datasets provide a diverse and comprehensive foundation for evaluating the performance of our proposed open-vocabulary and multitask image segmentation framework.

4.3. Evaluation Metrics.

In our experimental evaluation, we employ several key metrics to assess the performance of our model: Mean Intersection over Union ($mIoU$), Intersection over Union for Foreground (IoU_{FG}), Intersection over Union for Binary Segmentation (IoU_{BIN}), and Average Precision (AP).

Mean Intersection over Union ($mIoU$): $mIoU$ is calculated as the average of the Intersection over Union (IoU) values across different foreground object categories. It is defined as:

$$mIoU = \frac{1}{C}\sum_{C=1}^{C} IoU_C. \tag{6}$$

Where C represents the number of object categories, and IoU_C signifies the Intersection over Union for category C. This metric quantifies the degree of overlap between the predicted segmentation and the ground truth labels for each category, providing an overall measure of segmentation quality.

Intersection over Union for Binary Segmentation (IoU_{BIN}): IoU_{BIN} assesses segmentation performance without considering specific object classes. It computes the average of the IoU values for both foreground and background regions across all test images:

$$IoU_{BIN} = \frac{1}{2}(IoU_{FG} + IoU_{BG}). \tag{7}$$

Where IoU_{FG} represents the IoU for the foreground, and IoU_{BG} represents the IoU for the background. This metric provides insights into the segmentation quality when class information is disregarded.

Average Precision (AP): AP is determined by the area beneath the recall-precision curve (ROC). The assessment focuses on the model's capacity to differentiate between accurate and inaccurate matches, shedding light on its balance between precision and recall.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

4.4. Comparison to State-of-the-Art Methods

Referring Expression Segmentation Comparative Experiment. Referring expression segmentation associates natural language reference expressions with corresponding objects in the image to perform semantic segmentation on the image. We conducted a comprehensive comparison of OVAMTSeg with contemporary state-of-the-art open-vocabulary referring expression segmentation methods, including MDETR [34], HulaNet [35], Mask-RCNN [35], RMI [35], and CLIPSeg [8]. The results are summarized in Table 1. Our approach outperforms the two-stage HulaNet approach. However, the $mIoU$ of OVAMTSeg is worse than MDETR, which operates at full image resolution and received two rounds of fine-tuning on PhraseCut. Notably, OVAMTSeg outperforms CLIPSeg with the same training method in all aspects, indicating that our model is effective.

Table 1. Referring Expression Segmentation performance on PhraseCut+.

Model	mIoU	IoU$_{FG}$	AP
MDETR[34]	53.7	-	-
HulaNet[35]	41.3	50.8	-
Mask-RCNN top[35]	39.4	47.4	-
RMI[35]	21.1	42.5	-
CLIPSeg[8]	43.4	54.7	76.7
Ours	**47.5**	**57.1**	**80.4**

Zero-Shot Segmentation Comparative Experiment. As shown in Tables 2 and 3, we compare the open vocabulary zero-shot segmentation performance on Pascal-VOC and Pascal-Context datasets, including SPNet[7], ZS3Net[3], CSRL[6], CaGNet[36], OSR[37], JoEm[38], CLIPSeg[8] . Tab.2 and Tab.3 can be condened as the following observations: i) OVAMTSeg achieves 51.6% and 46.6% mIoU towards unseen classes on Pascal-VOC and Pascal-Context, which goes beyond the best method CLIPSeg and OSR by +4.3% and +3.5%, respectively. It suggests that OVAMTSeg can be adapted to a wider range of scenarios. ii) Among the baselines, OSR demonstrates promising results for the seen classes but encounters challenges with respect to the unseen classes. In comparison, our model performs well in unseen classes compared to models trained on Pascal-VOC and Pascal-Context datasets. This variation in performance can be attributed to the fact that other models are trained on datasets with fixed classes, whereas OVAMTSeg is capable of discerning a wider range of classes.iii) OVAMTSeg and CLIPSeg perform better on unseen classes compared to seen classes. This could be because seen classes pose inherent difficulties in segmentation, while unseen classes tend to be relatively larger and easier to segment.

Table 2. Performance in zero-shot segmentation using Pascal-VOC with 4 unseen classes. $mIoUs$ and $mIoUu$ denote the mIoU(%) of classes that are observed and those that are not. The training of our model occurs on PhraseCut+ with the Pascal classes removed. The term IN-seen refers to the pre-training phase of ImageNet where previously unseen classes are eliminated.

Model	pre-train	mIoUs	mIoUu
SPNet[7]	IN	67.3	21.8
ZS3Net[3]	IN-seen	66.4	23.2
CSRL[6]	IN-seen	69.8	31.7
CaGNet[36]	IN	69.5	40.2
OSR[37]	IN-seen	75.0	44.1
JoEm[38]	IN-seen	67.0	33.4
CLIPSeg[8]	CLIP	20.8	47.3
Ours	CLIP	**28.3**	**51.6**

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

Table 3. Zero-shot segmentation performance on Pascal-Context with 4 unseen classes. Our model is trained on PhraseCut+ with the Pascal classes removed.

Model	pre-train	mIoUs	mIoUu
SPNet[7]	IN	36.3	18.1
ZS3Net[3]	IN-seen	37.2	24.9
CSRL[6]	IN-seen	39.8	23.9
CaGNet[36]	IN	24.8	18.5
OSR[37]	IN-seen	41.1	43.1
JoEm[38]	IN-seen	36.9	30.7
CLIPSeg[8]	CLIP	16.8	40.2
Ours	CLIP	**25.3**	**46.6**

One-Shot Segmentation Comparative Experiment. Differing from zero-shot segmentation, one-shot segmentation requires the understanding of both text prompts and annotated support images. To aid in this comprehension, we used the same visual prompts methods as CLIPSeg, including object cropping, background blur, and darkening. We evaluate the proposed model on PASCAL-5i, COCO-20i, and compare the results with recent methods. To ensure fairness during training, classes that overlap with these datasets are removed. As shown in Table 4, OVAMTSeg achieves 85.6% AP, outperforming CLIPSeg by 4.3%, respectively. Table 5 presents a performance comparison when the model is trained on the COCO-20i dataset. Similarly, we note consistent findings where our model performs effectively. OVAMTSeg also attains the highest results with an AP of 86.6%. The COCO-20i results show that OVAMTSeg also performs well when trained on other datasets than PhraseCut+. Particularly, HSNet and PFENet exhibit superior performance in the performance metric of mIoU, which can be attributed to their explicit design for one-shot segmentation.

Table 4. One-shot performance on Pascal-5i.

Model	vis.backb.	mIoU	IoU$_{BIN}$	AP
PPNet[39]	RN50	52.8	69.2	-
RePRI[40]	RN50	59.1	-	-
PFENet[41]	RN50	60.8	73.3	-
HSNet[42]	RN50	64.0	76.7	-
MGNet[43]	RN50	52.1	68.2	-
HCNet[44]	RN50	62.1	71.7	-
DRNet[45]	RN50	53.3	72.8	-
SRPNet[46]	RN50	61.5	-	-
CLIPSeg[8]	ViT(CLIP)	59.5	75.0	82.3
Ours	ViT(CLIP)	**65.9**	**77.1**	**86.8**

Tian et al. [41] replace visual samples in the network with textual label word vectors and employ PFENet for zero-shot segmentation, adhering to the protocol of zero-shot segmentation. and utilize PFENet for zero-shot segmentation following the one-shot segmentation protocol. In this context, OVAMTSeg significantly surpasses their performance scores, as evidenced in Table 6. This indicates the challenge in applying one-time oriented techniques such as PFENet to different tasks, in contrast to our OVAMTSeg, which exhibits significant generalization potential.

Table 5. One-shot performance on COCO-20i(OVAMTSeg trained on COCO-20i).

Model	vis.backb.	mIoU	IoU$_{BIN}$	AP
PPNet[39]	RN50	29.0	-	-
RePRI[40]	RN50	34.0	-	-
PFENet[41]	RN50	35.8	-	-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

HSNet[42]	RN50	39.2	68.2	-
MGNet[43]	RN50	34.9	63.9	-
HCNet[44]	RN50	40.7	63.4	-
DRNet[45]	RN50	36.5	60.9	-
CLIPSeg[8]	ViT(CLIP)	33.2	58.4	40.5
Ours	ViT(CLIP)	**35.7**	**62.5**	**46.3**

Table 6. Zero-shot performance on Pascal-5i. Scores were derived by adhering to the one- shot segmentation assessment method.

Model	vis.backb.	mIoU	IoU$_{BIN}$	AP
PFENet[41]	VGG16	54.2	-	-
LSeg[21]	ViT(CLIP)	52.3	67.0	-
CLIPSeg[8]	ViT(CLIP)	72.4	83.1	93.5
Ours	ViT(CLIP)	**78.5**	**87.3**	**93.8**

4.5. Ablation Study

Adaptive Prompt Analysis. We evaluate various prompt configurations to assess the importance of the adaptive prompt in the context of open vocabulary segmentation, as presented in Table 7. The static template prompt utilizes the sentence structure "A photo of class," where "class" is replaced with specific class names. These class-specific prompts are then encoded into the text features. As shown in Table 7, the adaptive prompt yields a noteworthy enhancement of 2.2% mIoU and 2.3% AP performance for referring expression segmentation, respectively, when compared to the fixed prompt. Similarly, the adaptive prompt leads to 2.6% mIoU performance improvement tover the fixed prompt for unseen classes in zero-shot segmentation, and achieves a lead of 2.6% mIoU and 1.2% AP for one-shot segmentation. This underscores the role of the adaptive prompt in facilitating the capture of category-sensitive concepts through learnable parameters.

Multi-Task Analysis. To assess the benefits of the multitask approach in OVAMTSeg, we performed a comparative examination by comparing the model's performance with single-task training for individual tasks. As depicted in Table 7, the outcomes under the multitask category are derived from a single unified model, while the single-task results are obtained from three separate individual models. Our model achieves remarkable results, with a 47.5% $mIoU$ and an 80.4% AP on referring expression segmentation, surpassing the performance of the one-shot counterparts. Furthermore, open-vocabulary zero-shot and one-shot segmentation exhibit consistent and impressive results, particularly in terms of performance on unseen classes. OVAMTSeg consistently improves metrics across almost all tasks, highlighting the effectiveness of the multi-task training approach in enhancing network generalization and surpassing the performance of one-shot models.

Text Encoders. OVAMTSeg inherently accommodates various text encoders. We illustrate the impact of employing different text encoders in Table 7. It is important to highlight that all text encoders feature the same transformer-based architecture that purely operates on text prompts. The main difference between the encoders is the image encoder that was paired during CLIP pre-training (for example, the text encoder denoted by "ViT-B/32" was trained in conjunction with a ViT-B/16 image encoder) and the size of the embedding dimension. We observed that employing RN50×16 achieves the highest performance among all text encoders. We hypothesize that this is due to the larger embedding dimension provided by this encoder.

Table 7. Comparison of various task paradigms, prompt strategies, and text encoders. Evaluation includes Referring Expression Segmentation performance on PhraseCut+, zero-shot segmentation performance on Pascal-VOC, and one-shot performance on Pascal-5i.

Method	Referring Expression		Zero-Shot		One-Shot	
	mIoU	AP	mIoUs	mIoUu	mIoU	AP

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

Prompt	Fixed	45.4	78.1	57.8	49.0	63.3	85.4
	Adaptive	47.5	80.4	28.3	51.6	65.9	86.6
Task Paradigm	Single-Task	41.3	-	75.0	44.1	64.0	-
	Multi-Task	47.5	80.4	28.3	51.6	65.9	86.6
Text Encoder	ViT-B/32(512)	43.6	79.0	25.5	49.2	63.5	85.2
	ViT-B/16 (512)	44.4	79.3	26.0	49.8	64.1	85.4
	RN50 × 4(640)	44.8	79.4	26.2	50.1	64.5	85.6
	RN50 × 16(768)	46.2	79.8	27.4	51.0	65.6	85.9

Component Analysis. We conducted ablation studies on OVAMTSeg, and the results are presented in Table 8. When the model operates without any adaptive prompt, text encoder extension, and replaces the attention-based multi-modal interaction with a simple concatenation sum, its performance across various metrics is notably subpar. The introduction of adaptive prompt learning significantly enhances the model's performance, particularly in the context of unseen classes during one-shot segmentation. Subsequently, the text encoder extension and multi-modal interaction components are gradually incorporated into the framework. This incremental integration yields performance improvements across various tasks during inference. These results underscore the critical contributions of each component. Adaptive prompt learning facilitates OVAMTSeg in capturing category-sensitive characteristics, while the text encoder extension aids in effectively capturing textual features. Additionally, the multi-modal interaction component plays a pivotal role in enhancing the cross-modal alignment of visual and text features.

Table 8. Ablation studies were conducted to analyze the effectiveness of the proposed modules. Referring Expression Segmentation performance on PhraseCut+. Zero-shot segmentation performance on Pascal-VOC, and one-shot performance on Pascal-5i.

Adaptive Prompt	**Text Encoder Extension**	**Multimodal Interaction**	**Referring Expression**		**Zero-Shot**		**One-Shot**	
			mIoU	**AP**	**mIoU*s***	**mIoU*u***	**mIoU**	**AP**
✗	✗	✗	43.6	72.8	23.1	25.3	60.2	78.1
✓	✗	✗	45.9	77.6	25.9	48.8	63.7	83.5
✓	✓	✗	46.6	78.3	26.8	49.9	64.6	84.3
✓	✓	✓	47.5	80.4	28.3	51.6	65.9	86.6

4.6. Qualitative Results

Figure 2 presents qualitative results, highlighting the notable differences in prediction accuracy between OVAMTSeg and CLIPSeg for identical text prompts. For example, in the second image where the text prompt is "glass," OVAMTSeg's predictions exhibit precise delineation of the two glasses, accurately capturing their outlines. In contrast, CLIPSeg's predictions result in a less refined separation of the glasses' shapes. Similarly, when the text prompt is "bottle," OVAMTSeg correctly identifies and predicts two small bottles, demonstrating its ability to discern subtle distinctions. However, CLIPSeg's predictions in this scenario are less accurate, as it erroneously identifies both the wine glass and the bottle. Despite these successes, certain challenges persist. In the second image, where the text indicates "fork," both OVAMTSeg and CLIPSeg incorrectly predict the knife to be a fork, in addition to accurately predicting the strengths and limitations of both OVAMTSeg and CLIPSeg in their responses to specific text prompts.the presence of an actual fork. These visual comparisons underscore the strengths and limitations of both OVAMTSeg and CLIPSeg in their responses to specific text prompt.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

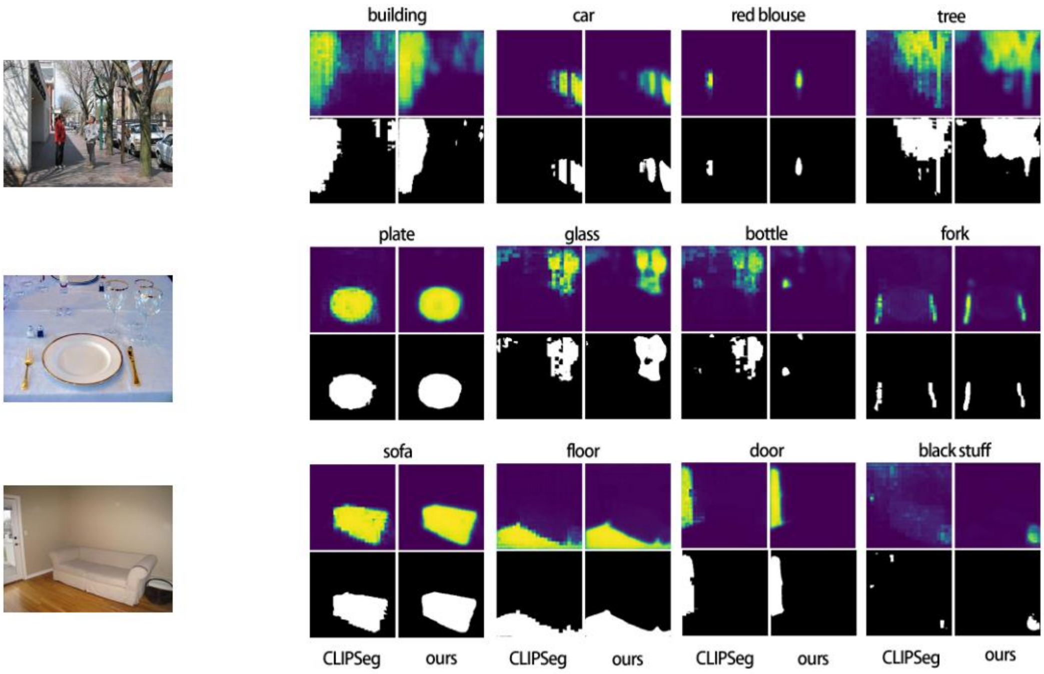

Figure 2. Qualitative results of the multi-task open vocabulary segmentation. We compare the segmentation results of the proposed OVAMTSeg and CLIPSeg.

5. Conclusions

In this paper, we have presented OVAMTSeg, a universal framework designed to excel in multitask open-vocabulary segmentation. Our investigation delved into a novel text prompt method, showcasing competitive performance across a spectrum of tasks, including referring expression, zero-shot, and one-shot image segmentation. Furthermore, we have enhanced the text encoder to bolster its semantic representation capabilities. Additionally, we introduced a multi-modal interaction module that dynamically adjusts the correlation between modalities. This adaptation enables more flexible fusion of textual and image features across diverse tasks and datasets. Our contributions stand as a testament to the potential of open-vocabulary and multitask segmentation. We are convinced that our work not only provides valuable insights but also outlines a promising direction for future research in this dynamic field.

The research results indicate that the image size should be around 350×350 pixels, as going significantly larger or smaller can impact experimental accuracy negatively. In addition, our experiments are confined to a few benchmarks. In future studies, additional modalities like sound and touch could be integrated. Second, we will attempt to improve the model to accommodate inputs of any image size.

Author Contributions: Funding acquisition, Lihu Pan; Methodology, Lihu Pan and Yunting Yang; Supervision, Lihu Pan, Zhengkui Wang and Rui Zhang; Visualization, Yunting Yang; Writing – original draft, Yunting Yang; Writing – review & editing, Lihu Pan.

Funding: This work was supported by the National Science Foundation of Shanxi Province(No.202203021221145) and Shanxi Province Graduate Joint Training Demonstration Base Project(No. 2022JD11). This study was funded by the Shanxi Provincial Committee for the Management of Overseas Chinese.

Data Availability Statement: The data presented in this study are derived from the following resources available in the public domain: https://github.com/ChenyunWu/PhraseCutDataset, https://github.com/dvlab-research/PFENet.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

1. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR), 2019; pp. 5693-5703.
2. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*. **2017**, 40, 834-848.
3. Bucher, M.; Vu, T.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*. **2019**, 32.
4. Radford, A.; Kim, J.; Hallacy, C.; Ramesh, Aditya.; Goh, Gabriel.; Agarwal, Sandhini.; Sastry, Girish.; Askell, Amanda.; Mishkin, Pamela.; Clark, Jack.; Krueger, Gretchen.; Sutskever, Ilya. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (PMLR), July 2021; pp. 8748-8763.
5. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. Proceedings of the 38th International Conference on Machine Learning (PMLR), July 2021; pp.4904-4916.
6. Li, P.; Wei, Y.; Yang, Y. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*. **2020**, 33, 10317-10327.
7. Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; Akata, Z. Semantic projection network for zero-and few-label semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019; pp.8256-8265.
8. Lüddecke, T.; Ecker, A. Image segmentation using text and image prompts. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022; pp.7086-7096.
9. Li, J.; Qi, Q.; Wang, J.; Ce, Ge.; Li, Y.; Yue, Z.; Sun, H. OICSR: Out-in-channel sparsity regularization for compact deep neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019; pp.7046-7055.
10. Wu, J.; Li, G.; Liu, S.; Lin, L. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. Proceedings of the AAAI Conference on Artificial Intelligence, April 2020; pp. 12386-12393.
11. Wu, J.; Chen, T.; Wu, H.; Zhi, Yang.;Luo, G.; Lin, L. Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia*. **2020**, 23, 2413-2427.
12. Xia, X.; Li, J.; Wu, J.; Wang, X .; Xiao, X .; Zheng, M .; Wang, R. TRT-ViT: TensorRT-oriented vision transformer. *arXiv preprint arXiv:2205.09579*, **2022**.
13. Xiao, X.; Yang, Y.; Ahmad, T.; Jin, L.; Chang, T. Design of a very compact cnn classifier for online handwritten chinese character recognition using dropweight and global pooling. 2017 14th IAPR international conference on document analysis and Recognition (ICDAR), November 2017; pp. 891-895.
14. Ren, Y.; Wu, J.; Xiao, X.; Yang, J. Online multi-granularity distillation for gan compression. Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp.6793-6803.
15. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*. **2021**, 34, 17864-17875.
16. Qin, J.; Wu, J.; Li, M.; Zheng, M .; Wang, X . Multi-granularity distillation scheme towards lightweight semi-supervised semantic segmentation. European Conference on Computer Vision. Cham: Springer Nature Switzerland, October 2022; pp.481-498.
17. Qin, J.; Wu, J.; Xiao, X.; Li, L.; Wang, X. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. Proceedings of the AAAI conference on artificial intelligence, June 2022; pp. 2117-2125.
18. Sun, K.; Xiao, B.; Liu, D, Wang, J. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019; pp.5693-5703.
19. Li, P.; Wei, Y.; Yang, Y. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*. **2020**, 33,10317-10327.
20. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp.3684-3692.
21. Li, B.; Weinberger, K.; Belongie, S.; Koltun,V.; Ranftl, R. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*. **2022**.
22. Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. **2021**.
23. Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.; Lester, B.; Du, N.; Dai, A.; Le, Q. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*. **2021**.
24. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*. **2021**.
25. Zhou, K.; Yang, J.; Loy, C.; LiuZ. Learning to prompt for vision-language models. *International Journal of Computer Vision*. **2022**, 130,2337-2348.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0631.v1

26. Khattak, M.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F. Maple: Multi-modal prompt learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023; pp.19113-19122.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9 2015; pp.234-241.
28. Dumoulin, V.; Perez, E.; Schucher, N.; Strub, F.; Vries, H.; Courville, A.; Bengio, Y. Feature-wise transformations. *Distill*. **2018**, 3, e11.
29. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. **2017**.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. European conference on computer vision. Cham: Springer International Publishing, 2020; pp.213-229.
31. Everingham, M.; Winn, J. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*. **2012**, 2007(1-45): 5.
32. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014; pp.891-898.
33. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C . Microsoft coco: Common objects in context. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12 2014, Proceedings, Part V 13. Springer International Publishing, pp.740-755.
34. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp.1780-1790.
35. Wu, C.; Lin, Z.; Cohen, S.; Bui, T.; Maji, S. Phrasecut: Language-based image segmentation in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp.10216-10225.
36. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. Proceedings of the 28th ACM International Conference on Multimedia, 2020; pp.1921-1929.
37. Zhang, H.; Ding, H. Prototypical matching and open set rejection for zero-shot semantic segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp.6974-6983.
38. Baek, D.; Oh, Y.; Ham, B. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp.9536-9545.
39. Liu, Y.; Zhang, X.; Zhang, S.; He, X. Part-aware prototype network for few-shot semantic segmentation. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Proceedings, Part IX 16. Springer International Publishing, August 23–28, 2020; pp.142-158.
40. Boudiaf, M.; Kervadec, H.; Masud, Z.; Piantanida, P.; Ayed, I.; Dolz, J. Few-shot segmentation without meta-learning: A good transductive inference is all you need? Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021; pp.13979-13988.
41. Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; Jia, J. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*. **2020**, 44, 1050-1065.
42. Min, J.; Kang, D.; Cho, M. Hypercorrelation squeeze for few-shot segmentation. Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp.6941-6952.
43. Chang, Z.; Lu, Y.; Wang, X.; Ran, X. Mgnet: Mutual-guidance network for few-shot semantic segmentation. *Engineering Applications of Artificial Intelligence*. **2022**, 116,105431.
44. Fang, Z.; Gao, G.; Zhang, Z.; Zhang, A. Hierarchical context-agnostic network with contrastive feature diversity for one-shot semantic segmentation. *Journal of Visual Communication and Image Representation*. 2023, 90, 103754.
45. Tang, M.; Zhu, L.; Xu, Y.; Zhao, M. Dual-stream reinforcement network for few-shot image segmentation. *Digital Signal Processing*. **2023**, 134, 103911.
46. Ding, H.; Zhang, H.; Jiang X. Self-regularized prototypical network for few-shot semantic segmentation. *Pattern Recognition*, **2023**, 133,109018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.