國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

基於影像與文字共同分解之文本監督式語意分割

Image-Text Co-Decomposition for Text-Supervised Semantic Segmentation

吳季嘉

Ji-Jia Wu

指導教授: 莊永裕 博士

Advisor: Yung-Yu Chuang Ph.D.

共同指導教授: 林彥宇 博士

Co-Advisor: Yen-Yu Lin Ph.D.

中華民國 113 年 5 月

May, 2024

doi:10.6342/NTU202400974

Acknowledgements

能夠完成這一篇碩士論文，我要由衷感謝我的指導教授莊永裕老師和林彥宇老師在我碩士生涯中給予的指導與支持。謝謝林老師在每週的會議中不僅幫助我確立研究方向，更指出了我未曾察覺的問題。謝謝莊老師在實驗室的團體會議中幫我確認研究的方向是否有所偏差，並點出實驗中未注意到的盲點。非常榮幸能在兩位老師的指導下做學術研究，讓我不只完成了高品質的碩士論文，還讓論文有榮幸投稿上國際頂尖會議 CVPR 2024 並進行發表。

我亦要感謝與我進行同一個研究專案的學長及學弟妹們的幫助和支持。感謝劉育綸教授、陳敏弘博士與胡厚寧博士在研究方向和學術討論上的許多建議。同時，也感謝學弟妹張家豪、莊婕妤和陳存佩在實驗設計和數據分析上的協助，這使我的研究更加完整。

另外，我也要感謝實驗室同學們，千祐、則偉、秉茂、湧致、王安、力仁、菩提、啟維一直以來的相互協助與扶持，儘管大家做的研究主題不同，研究生的兩年生活因為有你們而變得更加豐富。

最後要感謝我的家人，沒有你們的支持，我無法克服在研究生活中遇到的各種困難與挑戰。你們的鼓勵以及經濟上的支持讓我不用去擔心生活上的瑣事，讓我可以全力去完成研究，並完成這篇論文。

doi:10.6342/NTU202400974

摘要

本篇論文旨在解決文本監督式語義分割問題。在這個任務中，我們希望能僅透過影像-文字配對而無需密集標註，訓練出一個語義分割模型，在圖像中對任意視覺概念進行分割。現有方法顯示，透過圖像-文字配對進行對比學習，可以有效地將影像局部與文字含義對齊。我們注意到此學習方式存在問題：一段文字通常包含多個語義概念，而語義分割則傾向於針對單一物件進行分割。為解決此問題，我們提出了一個新框架，名為 Image-Text **Co-De**composition （**CoDe**），在此框架中，配對的圖像與文字被共同分解為一組影像區域和文字片段的配對，並透過對比學習來強化影像區域與文字片段之間的對齊。此外，我們提出了一種提示學習機制，目的是強調影像和文字中分割出的影像區段或文字片段，從而使視覺語言模型能夠對這些影像區域和文字片段提取出更有效的特徵。實驗結果顯示，我們的方法在六個數據集上相較於現有的文本監督式語義分割方法較為有效。我們將程式碼公開在 https://github.com/072jiajia/image-text-co-decomposition。

關鍵字：文本監督式學習、語意分割、多模態學習、提示學習、視覺-語言模型

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

Abstract

This paper addresses text-supervised semantic segmentation, aiming to learn a model capable of segmenting arbitrary visual concepts within images by using only image-text pairs without dense annotations. Existing methods have demonstrated that contrastive learning on image-text pairs effectively aligns visual segments with the meanings of texts. We notice that there is a discrepancy between text alignment and semantic segmentation: A text often consists of multiple semantic concepts, whereas semantic segmentation strives to create semantically homogeneous segments. To address this issue, we propose a novel framework, Image-Text **Co-De**composition (**CoDe**), where the paired image and text are jointly decomposed into a set of image regions and a set of word segments, respectively, and contrastive learning is developed to enforce region-word alignment. To work with a vision-language model, we present a prompt learning mechanism that derives an extra representation to highlight an image segment or a word segment of interest, with which more effective features can be extracted from that segment. Comprehensive experimental results demonstrate that our method performs favorably against existing text-supervised semantic segmentation methods on six benchmark datasets. The code is available at `https://github.com/072jiajia/image-text-co-decomposition`.

Keywords: Text-supervised learning, Semantic segmentation, Multi-modal learning, Prompt learning, Vision-language model

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

Contents

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

List of Figures

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

List of Tables

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

Chapter 1 Introduction

1.1 Background and Motivation

Semantic segmentation is essential to various applications [11, 15, 51] in computer vision but is hindered by several critical challenges. First, the expensive cost of acquiring pixel-level annotations limits the applicability of fully supervised semantic segmentation methods. Second, most existing methods [40, 44, 53] are developed to work on predefined categories and leave themselves inapplicable to rare or unseen concepts described by free-form text. To address these obstacles, a new research direction has emerged in vision-language models, referred to as text-supervised semantic segmentation [5, 28, 45–47, 50]. This task develops segmentation models capable of assigning labels across large vocabularies of concepts and supporting semantic segmentation model training without pixel-wise annotations.

fig. 1.1 compares existing methods for text-supervised semantic segmentation by grouping their cross-domain alignment mechanisms into three categories, including image-text, region-text, and region-word alignment. Despite the differences, most of these methods compensate for the lack of pixel-wise annotations on broad semantic concepts by exploring abundant image-text pairs on the internet. The textual descriptions bring extensive knowledge across diverse categories. Thus, existing methods typically apply a

doi:10.6342/NTU202400974

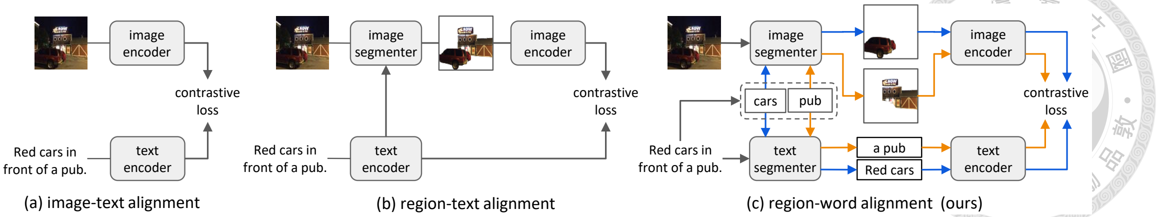

Figure 1.1: Existing methods perform text-supervised semantic segmentation by learning either (a) image-text alignment or (b) region-text alignment. This paper presents (c) region-word alignment via image-text co-decomposition, where the image and the text are decomposed into object regions and word segments, respectively, while contrastive learning is used to establish cross-modal correspondences between these image and word segments.

vision-language model such as CLIP [34] to textual descriptions to acquire the semantic context of the corresponding images for segmentation model learning.

The image-text alignment is widely adopted in the literature e.g. [28, 45, 46]. As depicted in fig. 1.1a, methods of this group derive an image encoder and a text encoder by aligning them in a joint embedding space. They then use their proposed zero-shot transfer techniques to enable the two encoders to predict segmentation output. Despite the simplicity, they introduce unfavorable discrepancies between the training and testing phases since we aim to match the semantic features from the text to the corresponding image segments rather than the whole image during testing.

To mitigate this issue, the region-text alignment is explored. As shown in fig. 1.1b, methods of this group such as [5] utilize a pre-trained visual-language model to derive an additional image segmenter that discovers concepts described by the text. They enforce the consistency between the segmented region and the text but suffer from the discrepancy between the region-text alignment and semantic segmentation: A text may consist of multiple concepts, such as pub and car in fig. 1.1b, while semantic segmentation aims to identify regions of the same concept.

To address the aforementioned issues in the image-text and region-text alignments, we propose a novel framework, Image-Text **Co-De**composition (**CoDe**), to achieve region-

doi:10.6342/NTU202400974

word alignment. As illustrated in fig. 1.1c, we utilize a visual-language model to construct an image segmenter and a text segmenter: The former decomposes an image into image segments, while the latter decomposes a text into word segments. In addition, there exist one-to-one correspondence between image and word segments. This way, the discrepancy between training and testing is alleviated since each image segment is derived from a single concept given by the corresponding word segment.

The proposed CoDe framework comprises four components: an image segmenter, a text segmenter, a region-word alignment module, and a prompt learning module. We randomly select nouns in the text. For each selected noun e.g., “car”, the image segmenter identifies the image segment that matches the noun, i.e., the region of the car, while the text segmenter discovers the corresponding word segment, i.e., “red cars.” The region-word alignment is developed to enforce the consensus between the image and word segments. To better work with a vision-language model, we present a prompt learning module to derive an extra representation, enabling more effective feature extraction.

1.2 Contributions

The main contributions of this work are as follows:

- We propose a new framework, Image-Text Co-Decomposition (CoDe), to learn the region-word alignment for eliminating train-test and image-text discrepancies, facilitating text-supervised semantic segmentation.

- We propose a prompt learning method to address domain shift issues arising from blank areas during the highlighting process and enhance the alignment between

doi:10.6342/NTU202400974

highlighted regions and highlighted words.

- Our method effectively carries out zero-shot semantic segmentation and performs favorably against the state-of-the-art methods on six benchmark datasets.

1.3 Publication

The core of this thesis builds upon the following peer-reviewed publication:

- Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. **"Image-Text Co-Decomposition for Text-Supervised Semantic Segmentation."** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. [42]

doi:10.6342/NTU202400974

Chapter 2 Related Works

2.1 Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation focuses on segmenting any concepts within images, even those unseen during training, based solely on textual descriptions. Its three important branches are discussed as follows:

Semi-supervised setting with mask-annotations. Methods of this group such as [16, 17, 24, 26, 31, 48] learn from dense annotations to produce high-quality segmentation masks, and then utilize image-text pairs and pre-trained vision-language models to extend the segmentation capability to a larger target vocabulary. Despite the remarkable results, these methods are hindered by their reliance on costly dense annotations, posing a challenge in cases where such annotations are difficult to obtain.

Training-free methods. Another line of research e.g. [38, 41, 55] makes the most of large pre-trained models for open-vocabulary segmentation without training. MaskCLIP [55] introduces a modification to the final layer of the CLIP image encoder, yielding dense feature maps that could be employed as initial segmentation maps for further refinement. ReCo [38] constructs an image archive and makes use of retrieval and co-segmentation to identify co-occurrence regions among a specific category. Although these methods eliminate the process of training, the results exhibit significant room for improvement, which shows the need for additional supervision to accomplish this task.

Text-supervised semantic segmentation. It strikes a balance between the two aforementioned branches. Methods of this group are discussed in detail in the following because our method belongs to this group.

2.2 Text-Supervised Semantic Segmentation

Text-supervised semantic segmentation [4, 5, 28, 33, 36, 43, 45–47, 50] decomposes an image into semantic regions according to text descriptions. Unlike semi-supervised methods relying on a few images with mask annotations during training, methods of this group aim to learn semantic masks solely from text-based guidance. We roughly divide existing methods into two categories based on their cross-modal alignment between the image and text domains.

Image-text alignment. These methods train an image encoder alongside a text encoder to align pairs of image and text in a joint embedding space. They use zero-shot transfer to enable the encoders to produce segmentation results. GroupViT [46] introduces a bottom-up approach within Transformers, grouping image patches into regions and utilizing object semantics derived from texts to guide training. SimSeg [50] further introduces a pretraining method that densely aligns visual and language representations, enabling the trained image encoder to generate segmentation masks in a zero-shot manner.

Region-text alignment. Another line of research targets at aligning the embedding of a region, instead of the whole image, with text descriptions. For instance, TCL [5] learns

doi:10.6342/NTU202400974

to segment specific regions within an image while ensuring consistency between the segmented region and the original text. It enables the model to segment the relevant region described in the text.

These methods for text-supervised semantic segmentation have shown that employing vision-language models and contrastive learning on image-text pairs enables aligning visual concepts with the meaning of the whole text. We notice that a text is usually a mix of multiple semantic concepts, but semantic segmentation aims to discover semantically homogeneous segments. To address this issue, inspired by the region-word matching techniques [8, 21, 22, 39] for cross-modal retrieval, we introduce image-text co-decomposition, where the image and the text are decomposed into image and word segments, respectively, and contrastive learning is adopted to enforce cross-modal consensus between these image and word segments. It turns out that image-text co-decomposition results in consistent performance gains on multiple benchmarks.

2.3 Prompt Tuning for Vision-Language Models

Emerged from natural language processing [23, 25, 27], prompt tuning focuses on parameter-efficient adaptation of large pre-trained models to new tasks. In computer vision [20, 56–58], pioneering work such as CoOp [56, 57] incorporates learnable tokens into the CLIP text encoder, enhancing the classification task performance. Recent studies e.g. [12, 14, 35] leverage prompt tuning in the text modality for extending CLIP's capabilities to various applications such as detection and segmentation tasks. Notably, prompt learning methods are also applicable to the visual domain. VPT [19] employs prompt tuning in the visual modality by inserting learnable vectors into Vision Transformers. Further studies [18, 26] explore tuning methods that directly incorporate learnable prompts into the input image within the RGB domain to address downstream tasks.

Drawing inspiration from the success of these methods, our method leverages the capabilities of prompt tuning on segment feature extraction in both the visual and text domains. Prompt learning is beneficial in this work when applying contrastive learning to the visual and textual features extracted by a vision-language model.

doi:10.6342/NTU202400974

Chapter 3 Methodology

In this section, we first provide an overview of our method for image-text co-decomposition and define the notations in section 3.1. Then, we specify the three major modules of our method, including 1) the image-text co-segmentation module in section 3.2, 2) the region-word highlighting module in section 3.3, and 3) the region-word alignment module in section 3.4. These modules work harmoniously to address the region-word alignment for text-supervised semantic segmentation and enhance model performances. Finally, implementation details are given in section 3.5.

3.1 Method Overview

Image-text co-decomposition enables text-supervised segmenters to learn region-word consensus when segmenting an image X^v with a paired text X^t. Our method aims to jointly learn an image segmenter F^v and a text segmenter F^t with solely the supervision from a set of K image-text pairs, $D = \{X^v_k, X^t_k\}_{k=1}^K$, where no annotations are given. In addition, we optimize two learnable prompts, including a region prompt P^v and a word prompt P^t, to alleviate the unfavorable effect of blank embeddings caused by applying a vision-language model to masked images or texts for feature extraction.

fig. 3.1 illustrates the pipeline of our method, consisting of three modules, includ-

doi:10.6342/NTU202400974

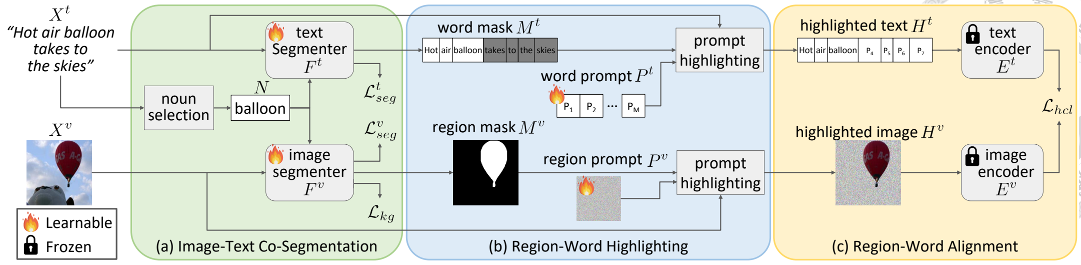

Figure 3.1: **Training pipeline of our method for image-text co-decomposition.** Our method consists of three major modules, including (a) the image-text co-segmentation module where the image and text segmenters estimate the region and word masks according to a selected noun, respectively, (b) the region-word highlighting module where the estimated masks together with two learnable prompts produce the highlighted image and text, and (c) the region-word alignment module where contrastive learning is applied to the embedded object regions and word segments to accomplish region-word alignment.

ing the image-text co-segmentation, region-word highlighting, and region-word alignment modules. For an input image-text pair (X^v, X^t), we initiate the process by randomly selecting a noun N, e.g., balloon in the figure, from the text X^t using the noun selector [2]. This selected noun serves as a query. We take the query N along with the image X^v as input to the image segmenter F^v to generate the region mask M^v showing the estimated object region specified by the query. Similarly, a text segmenter F^t takes the query N and the text X^t as input and estimates the word mask M^t indicating the associated word segment.

Subsequently, we apply the region mask M^v to the image X^v to crop the estimated object region. For the estimated background, i.e., the region outside the mask M^v, we crop the corresponding region from the learned region prompt P^v. The highlighted image H^v is yielded by combining the cropped object and background regions. Similarly, the highlighted text H^t is generated by combining the text X^t inside the mask M^t and the word prompt P^t outside the mask M^t. We extract features from the highlighted image and text by using the image encoder E^v and the text encoder E^t of CLIP [34], respectively. The procedure is repeated for each image-text pair and each selected noun. It follows that the region-word alignment is accomplished by contrastive learning [7]. Four loss

doi:10.6342/NTU202400974

functions, including $\mathcal{L}_{\text{kg}}$, $\mathcal{L}^{v}_{\text{seg}}$, $\mathcal{L}^{t}_{\text{seg}}$, and $\mathcal{L}_{\text{hcl}}$, are used for network optimization, and will be elaborated in the following.

3.2 Image-Text Co-Segmentation

The image-text co-segmentation module comprises a noun selector, an image segmenter, and a text segmenter, as shown in fig. 3.1a. Taking an image-text pair (X^v, X^t) as input, this module aims at jointly identifying an object region in image X^v and its accompanying word segment in text X^t according to a randomly selected noun.

To begin with, we employ the noun selector [2], which takes the text X^t as input and extracts a set of J nouns, $\{N_j\}_{j=1}^{J}$, in X^t. For each noun N_j, we carry out region mask generation, where the image segmenter F^v predicts a region mask M^v specifying the area in image X^v relevant to noun N_j. A similar task word mask generation is performed by the text segmenter F^t, which seeks a word mask M^t matching noun N_j. The tasks of region and word mask generation are depicted as follows.

Region mask generation. The image segmenter F^v takes image X^v and noun N_j as input. It encodes the image into a pixel-wise embedding $\mathbf{x}^v \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ is the image resolution and C is the channel dimension. We also compute the noun embedding $\mathbf{n}_j \in \mathbb{R}^C$ for noun N_j. The image segmenter generates a region mask $M^v \in \mathbb{R}^{H \times W}$ by performing the dot product between the noun embedding $\mathbf{n}_j$ and every location of the image embedding $\mathbf{x}^v$.

In this work, we use the image segmentation model in [5] to serve as the image segmenter F^v, and employ its corresponding loss, denoted by $\mathcal{L}^{v}_{\text{seg}}$ here, to help derive the

doi:10.6342/NTU202400974

image segmenter. This loss considers segment regularization and contrastive learning that can be directly applied to the segmentation results along with the noun embedding. We use the KgCoOp method [49] to obtain the noun embedding $\mathbf{n}_j$, as it avoids the pitfalls of improper prompt selection. It appends learnable context tokens to the noun, forming pseudo-sentences for optimal prompt tuning. The noun embedding loss $\mathcal{L}_{\text{kg}}$ [49] is included to improve the accuracy of these embeddings, i.e.,

$$\mathcal{L}_{\text{kg}} = ||\mathbf{n}_j - \mathbf{n}'_j||_2^2, \tag{3.1}$$

where the $\mathbf{n}'_j \in \mathbb{R}^C$ represents the knowledge-guided noun embedding generated from hand-crafted prompts such as “a photo of a N_j” using the text encoder.

Word mask generation. The text segmenter F^t takes the text X^t and the noun N_j as input. For text feature extraction, we consider the CLIP text encoder appended with two learnable multi-head attention layers. With the resultant feature extractor $\tilde{E}^t$, the word-wise features of text X^t are obtained via $\mathbf{x}^t = \tilde{E}^t(X^t) \in \mathbb{R}^{L\times C}$, where L is the text length, i.e., the number of word tokens. The word-specific logits $\boldsymbol{\ell}_j = [\ell_{j,i}]_{i=1}^L \in \mathbb{R}^L$ for noun N_j are computed via

$$\boldsymbol{\ell}_j = w \cdot \mathbf{x}^t \mathbf{n}_j + b, \tag{3.2}$$

where w and b are two learnable parameters, and $\mathbf{n}_j \in \mathbb{R}^C$ is the noun embedding.

Since each word in text X^t belongs to either one of the J word segments associated with nouns $\{N_j\}_{j=1}^J$ or none of them, the word mask $M^t = [m_i^t]_{i=1}^L \in \mathbb{R}^L$ for noun N_j is obtained by applying the softmax function over all the J noun-associated segments, i.e.,

$$m_i^t = \frac{\exp(\ell_{j,i})}{1 + \sum_{j'=1}^J \exp(\ell_{j',i})}, \text{ for } 1 \leq i \leq L, \tag{3.3}$$

where the additional 1 in the denominator is included for the case where word i does not belong to any noun-associated segments. The word mask M^t for noun N_j is produced.

According to the softmax function defined in Eq. 3.3, we get the probabilities of word i over $J+1$ cases, namely belonging to one of the J noun-associated segments or none of them. We compile a pseudo label vector $\mathbf{p} = \{p_i\} \in \{0,1\}^L$, where p_i takes value 1 if word i belonging to the jth noun-associated segment gets the highest probability, and 0 otherwise. We develop the text segmentation loss $\mathcal{L}^t_{\text{seg}}$, which is the cross-entropy loss on the word mask M^t with respect to the pseudo label vector $\mathbf{p}$, and can help learn the text segmenter F^t.

3.3 Region-Word Highlighting

We present a prompt learning method to reliably extract features from an image region or a word segment using a vision-language model. Specifically, we propose a region-highlighting prompt learning method and a word-highlighting prompt learning method, as shown in fig. 3.1b.

Region highlighting prompt. When the region mask M^v is directly applied to the image X^v via $M^v * X^v$, where $*$ denotes the element-wise multiplication operation, it makes specific regions of the image being zeroed out, resulting in what we refer to as blank areas. When a pre-trained vision-language model like CLIP is applied to these areas, the domain distribution may shift due to the introduction of zero tokens, which are unseen in natural images. To mitigate this issue, we introduce a region highlighting prompt, which is a learnable, universal image representation, denoted by P^v. This representation is used alongside the original image in the process of feature extraction. The highlighted image

doi:10.6342/NTU202400974

is then obtained via

$$H^v = X^v * M^v + P^v * (1 - M^v). \tag{3.4}$$

In this way, the blank areas are filled with the corresponding areas of the region prompt P^v alleviating the unfavorable effect of domain shift.

Word highlighting prompt. A similar challenge arises in the text domain when applying the word mask M^t to text X^t. The resultant zero tokens in the masked part unintentionally carry meanings of specific words, leading to potential inaccuracies. To mitigate this issue, we introduce a word highlighting prompt, represented as a learnable, universal text representation P^t. The highlighted text H^t is obtained by

$$H^t = X^t * M^t + P^t * (1 - M^t). \tag{3.5}$$

Since the masked part is filled with content from P^t, the risk of including unexpected text meanings is reduced.

3.4 Region-Word Alignment

In the following, we describe how our method achieves region-word alignment. Our objective is to optimize mutual evidence between the highlighted object regions and the highlighted word segments, as illustrated in fig. 3.1c.

Contrastive loss on highlighted region-word pairs. To achieve region-word alignment, we compute the highlighted region embedding $\mathbf{e}^v$ and highlighted word segment embedding $\mathbf{e}^t$ from the highlighted region-word pair by using the image and text encoders

doi:10.6342/NTU202400974

of CLIP by

$$\mathbf{e}^v = E^v(H^v) \quad \text{and} \quad \mathbf{e}^t = E^t(H^t), \tag{3.6}$$

where E^v and E^t are the CLIP image and text encoders, respectively.

We adopt batch optimization for model training. Each batch has several triplets, each of which is composed of an image, its paired text, and a randomly selected noun from the text. Each triplet yields a region embedding and a word embedding via Eq. 3.6. Suppose that there are B triplets in this batch. We create a similarity matrix $S = [S_{i,j}] \in \mathbb{R}^{B \times B}$, where $S_{i,j}$ stores the cosine similarity between the ith region embedding $\mathbf{e}^v_i$ and the jth word segment embedding $\mathbf{e}^t_j$. We adopt the symmetric version of InfoNCE loss to develop the highlighted region-word pair contrastive loss, which enhances the similarity of related region-word pairs while reducing it for unrelated pairs:

$$\begin{aligned} \mathcal{L}_{\text{hcl}} = &- \frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^{B} \exp(S_{i,j}/\tau)} \\ &- \frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^{B} \exp(S_{j,i}/\tau)}, \end{aligned} \tag{3.7}$$

where τ is a learnable temperature. Notably, even though nouns may be selected multiple times across image-caption pairs, the corresponding highlighted regions H^v and highlighted texts H^t vary, ensuring the effectiveness of the InfoNCE loss in precise region-word alignment.

Loss functions and optimization. In sum, the proposed network for image-text co-decomposition is optimized using a composite loss that combines the knowledge-guided, image segmentation, text segmentation, and highlighted region-word pair contrastive losses, defined as follows:

$$\mathcal{L} = \lambda_{\text{kg}}\mathcal{L}_{\text{kg}} + \lambda^{v}_{\text{seg}}\mathcal{L}^{v}_{\text{seg}} + \lambda^{t}_{\text{seg}}\mathcal{L}^{t}_{\text{seg}} + \lambda_{\text{hcl}}\mathcal{L}_{\text{hcl}}. \tag{3.8}$$

3.5 Implementation Details

We utilize NLTK's [2] part-of-speech tagging algorithm for noun selection. For image segmentation, we utilize TCL's image segmenter [5] to generate image masks, and we adopt the training loss in TCL, which relies solely on the image-caption pairs to yield $\mathcal{L}^{v}_{\text{seg}}$. For text segmentation, we use a CLIP text encoder appended with two multi-head attention layers as the text segmenter $\tilde{E}^{t}$. Our model is trained on the CC3M and CC12M datasets. The resolution of input images is set to 224×224. For each forward pass of an image-text pair, we randomly select 2 nouns from the text. The loss weights are set as follows: $\lambda_{\text{kg}} = 8.0$, $\lambda^{v}_{\text{seg}} = 1.0$, $\lambda^{t}_{\text{seg}} = 1.0$, and $\lambda_{\text{hcl}} = 0.5$ in the experiments. We train the model with a batch size of 64 on four NVIDIA 2080Ti GPUs and with a learning rate of $5 \times *10^{-6}$ for a total of $50,000$ iterations with $15,000$ warmup steps and a cosine schedule. AdamW optimizer [29] is used with a weight decay of 0.05. To improve the quality of the predicted mask during the evaluation phase, we adopt the post-processing approach described in TCL [5], which uses pixel-adaptive mask refinement (PAMR) [1] for mask refinement.

doi:10.6342/NTU202400974

Chapter 4 Experiments

4.1 Datasets and Evaluation Settings

We utilize image-text datasets to train our proposed model and perform extensive experiments on six commonly used semantic segmentation benchmarks to validate our method.

Training datasets. We trained our model on two image-text datasets, Conceptual Captions 3M (CC3M) [37] and Conceptual 12M (CC12M) [6] containing 3M and 12M image-text pairs respectively. They have been widely adopted for training text-supervised semantic segmentation methods.

Evaluation datasets. We used six zero-shot semantic segmentation benchmarks to validate the zero-shot transfer capability of our model on categories that were not specifically trained. As in previous work [5], the benchmarks can be categorized into two groups, with and without background classes. Benchmarks with a background generally label areas that do not belong to any predefined categories as "background," which is usually removed by considering a probability threshold in text-supervised semantic segmentation. For this category, we use the validation split of the following datasets: PASCAL VOC 2012 [13], PASCAL Context [32], and COCO-Object [3]. They each contain 20, 59, and 80 foreground classes, respectively, with an additional background class. For the "without

doi:10.6342/NTU202400974

Methods	Publication	Training Dataset	VOC	Context	Object	Stuff	City	ADE	Avg.
GroupViT [46]	CVPR 2022	CC3M+CC12M+YFCC14M	49.5	19.0	24.3	12.6	6.9	8.7	20.2
ViL-Seg [28]	ECCV 2022	CC12M	37.3	18.9	18.1				
ViewCo [36]	ICLR 2023	CC12M+YFCC14M	52.4	23.0	23.5				
OVSegmentor [47]	CVPR 2023	CC12M	53.8	20.4	25.1				
SimSeg [50]	CVPR 2023	CC3M+CC12M	57.4	26.2	29.7				
TCL [5]	CVPR 2023	CC3M+CC12M	55.0	30.4	31.6	22.4	24.0	17.1	30.1
SegCLIP [30]	ICML 2023	CC3M + COCO	52.6	24.7	26.5				
CoCu [45]	NeurIPS 2023	CC3M+CC12M+YFCC14M	51.4	23.6	22.7	15.2	22.1	12.3	24.6
PGSeg [52]	NeurIPS 2023	CC12M+RedCaps12M	53.2	23.8	28.7				
CoDe (Ours)	CVPR 2024	CC3M+CC12M	**57.7**	**30.5**	**32.3**	**23.9**	**28.9**	**17.7**	**31.8**

Table 4.1: **Text-supervised semantic segmentation performance comparison in terms of mIoU.** The proposed method is compared with nine SOTA methods on six popular semantic segmentation datasets: PASCAL VOC (VOC), PASCAL Context (Context), COCO-Object (Object), COCO-Stuff (Stuff), Cityscapes (City) and ADE20K (ADE). For each compared method, the training dataset column lists its training datasets. Several methods used datasets in addition to CC3M and CC12M, such as YFCC14M, COCO and RedCaps12M. When applicable, we also provide an average mIoU across all six datasets. For each dataset, the best method is indicated by bold fonts, whereas the second best method is underlined.

background category," we evaluated our model with the validation split of COCO-Stuff [3], Cityscapes [10], and ADE20K [54] datasets. Each of them contains 171, 19, and 150 classes, respectively. In this category, all images are fully annotated, which is exceptionally challenging. Using datasets in this category, our model can be tested for its ability to recognize a variety of concepts. We employ mean intersection-over-union (mIoU) as our evaluation metric.

For zero-shot semantic segmentation evaluation, we rely solely on the image segmenter. The image segmenter processes the input image in conjunction with class names from each dataset to produce segmentation predictions. In accordance with the settings of prior work [5], we adopt the class names provided by the default version of MMSegmentation [9] and adhere to its post-processing methodology.

4.2 Quantitative Comparisons

We compare the proposed method with nine text-supervised semantic segmentation methods on the six datasets. **??** reports the mIoU values. The numbers have been taken directly from the original papers. All methods were tested on the three datasets of the "with background class," but only three methods (GroupViT [46], CoCu [45] and TCL [5]) were tested on the dataset of the "without background class." For those three methods, we also report their average mIoU values across all six datasets. It is also worth noting that these methods use different combinations of training datasets, as indicated in the dataset column of **??**.

Our method achieves the best performance in all six datasets, while TCL [5] and SimSeg [50] are the runners-up. In terms of average mIoU, our method (CoDe) achieves

doi:10.6342/NTU202400974

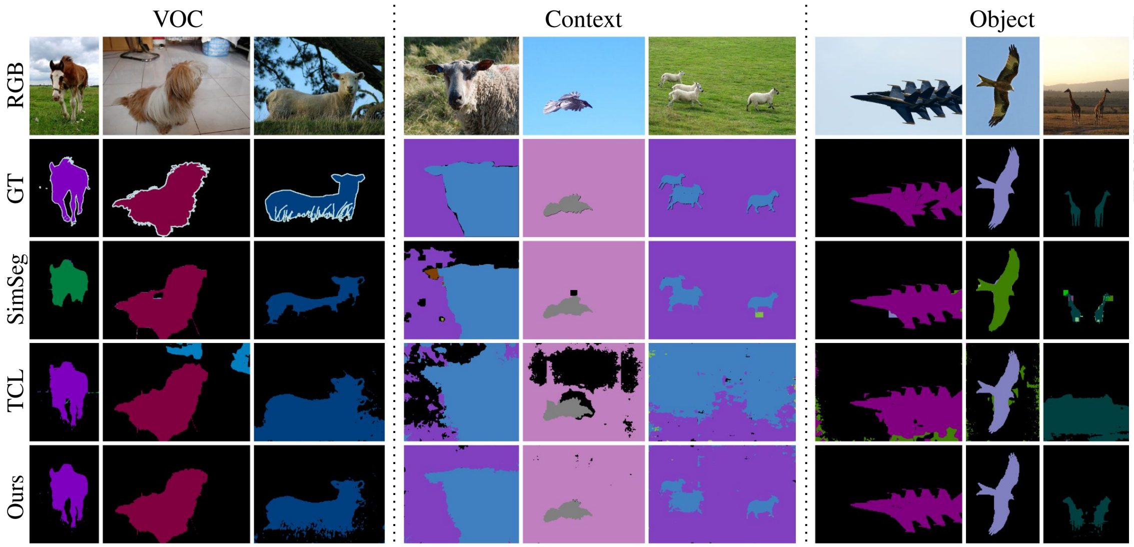

Figure 4.1: **Qualitative comparisons.** The proposed method is compared with the two most competitive methods, TCL [5] and SimSeg [50], on PASCAL VOC, PASCAL Context, and COCO Object datasets. Our method provides more precise object boundaries and effectively localizes objects within images without misclassification, leading to more accurate segmentation.

31.8 whereas TCL achieves 30.1, resulting in a 5.65% improvement. The result demonstrates the effectiveness of our image-text co-decomposition method in addressing the alignment-level train-test discrepancy that exists in previous methods by directly learning the region-word alignment.

4.3 Qualitative Results

Visual comparison with existing methods. fig. 4.1 visually compares the segmentation results of our methods and two runners-up, TCL [5] and SimSeg [50], on the PASCAL VOC, PASCAL Context, and COCO Object datasets.

This figure illustrates the fundamental benefit of our approach, which involves the direct learning of region-word alignments. Our model effectively establishes a strong connection between object regions and word segments, allowing a better understanding of how objects are represented in images. Through this enhanced understanding, both seg-

doi:10.6342/NTU202400974

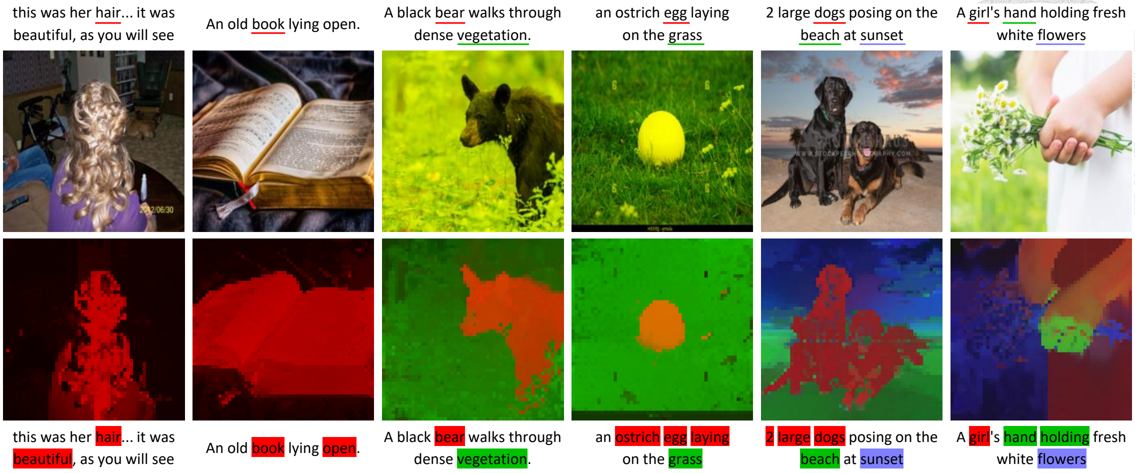
this was her hair... it was beautiful, as you will see
An old book lying open.
A black bear walks through dense vegetation.
an ostrich egg laying on the grass
2 large dogs posing on the beach at sunset
A girl's hand holding fresh white flowers
this was her hair... it was beautiful, as you will see
An old book lying open.
A black bear walks through dense vegetation.
an ostrich egg laying on the grass
2 large dogs posing on the beach at sunset
A girl's hand holding fresh white flowers

Figure 4.2: **Visualization of the results of our image-text co-decomposition method.** The first two rows display text and images, representing input image-text pairs. In each text, nouns are underlined with different colors. Our method uses these nouns as queries for performing image-text co-decomposition. Using our image-text co-decomposition method, the last two rows depict the method's output, where regions and word segments associated with different nouns appear in corresponding colors.

mentation quality and localization capabilities can be improved. As a result, our method provides more accurate classification and more precise masks than other methods.

The SimSeg[50] model, which learns from image-text alignments, occasionally assigns objects to the wrong classes. On the other hand, TCL [5], which is based on region-text alignment, produces coarser semantic masks. Accordingly, these observed limitations are most likely a result of the alignment-level discrepancy between the train and test, which may lead to suboptimal performance.

Visualization of image-text co-segmentation results. fig. 4.2 presents a visualization of the results obtained by our model. We denote regions and word segments associated with the different nouns in the corresponding colors. It demonstrates that our method effectively segments object regions within images based on various input nouns. It simultaneously segments corresponding word segments within the associated text, creating a harmonious alignment between the object region and the word segment.

C.	W.	R.	VOC	Context	Object	Stuff	City	ADE	Avg.
			54.4	27.6	32.7	22.5	25.0	16.6	29.8
✓			56.2	29.2	**32.9**	23.3	27.5	17.0	31.0
✓	✓		56.1	29.3	32.6	23.6	**29.0**	17.3	31.3
✓	✓	✓	**57.7**	**30.5**	32.3	**23.9**	28.9	**17.7**	**31.8**

Table 4.2: **Ablation study.** The baseline model is augmented with the image-text co-decomposition method (C.), the word highlighting prompt (W.), and the region highlighting prompt (R.), one at a time. We report the mIoU values of the resultant models on the six datasets and their averages.

The region-word alignment plays a pivotal role in our approach, serving as a supervisory signal for the model. By taking advantage of this alignment, our model not only performs visual localization but also captures correlations within the language domain. It indicates that our trained model possesses a more comprehensive understanding of the segmentation task.

4.4 Ablation Study

Contributions of individual components. The ablation study in table 4.2 assesses the contribution of the proposed components, including the image-text co-decomposition method, the word highlighting prompt, and the region highlighting prompt. Without the co-decomposition method, our baseline model only trains the image segmenter, resulting in an average mIoU of 29.8. Afterward, each proposed component is added to the baseline model one at a time to verify its contribution. As a result of adding the image-text co-decomposition module alone, the average mIoU has been increased to 31.0. It suggests that the image-text co-decomposition method can achieve region-word alignment and enhance localization capability. The model is further enhanced with the addition of word highlighting prompts and image highlighting prompts, resulting in further performance improvement.

λ_{hcl}	0.05	0.1	0.25	0.5	0.75	1.0
Avg.	30.6	31.2	31.7	**31.8**	31.5	30.8

Table 4.3: **Sensitivity analysis on the hyperparameter λ_{hcl}.** By varying λ_{hcl}, we examine the corresponding average mIoU values of all six datasets.

It demonstrates that the highlighting prompt learning method enhances feature extraction and strengthens alignment between regions and words.

Hyperparameter sensitivity analysis. table 4.3 investigates the impact of the loss weight for the highlighted region-word pair contrastive loss, denoted as λ_{hcl} in eq. (3.8). We observe that, when we apply the highlighted region-word pair contrastive loss in our training phase, the performance consistently outperforms our baseline model. The method is robust to the parameter to some degree as it achieves reasonable performance for a wide range of values. When λ_{hcl} is set to 0.5, our model achieves a peak score of 31.8. It is evident from these results that the image-text co-decomposition method is superior to the image-text decomposition method for achieving region-word alignment.

Effectiveness of jointly decomposing text. We validate the effectiveness of decomposing text by assessing the performance enhancement achieved by generating word masks, as opposed to simply using extracted nouns. This experiment is conducted by modifying the calculation of $\mathcal{L}_{\text{hcl}}$. Instead of using word segment embeddings as mentioned in section 3.4, we opt to compute the similarity matrix S using region embeddings with the embeddings of *individual nouns*. The average mIoU across all benchmarks is 30.2%, which is below our method's 31.8%. This indicates the benefits of using word segments encompassing extra words associated with each noun. The contextual information encoded in these additional words can serve as valuable supervisory signals, thereby improving

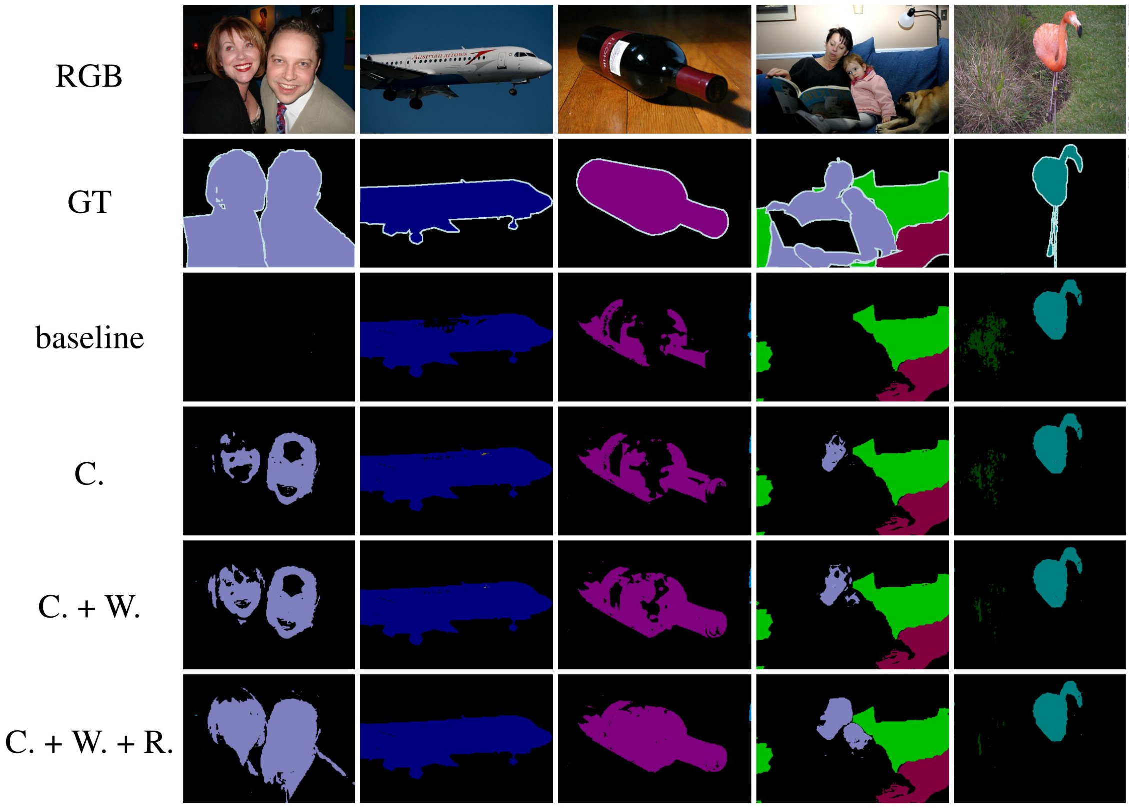

Figure 4.3: **Ablation studies.** We improve the baseline model by incrementally including (C.) the image-text co-decomposition module, (W.) the word highlighting prompt, and (R.) the region highlighting prompt. We present the segmentation results of the resulting models on the images of the PASCAL VOC [13] dataset.

performance.

4.5 Ablation Study Visualization

In the following, we conduct ablation studies by visualizing the effects of the proposed components in our method, including the image-text co-decomposition method, the word highlighting prompt, and the region highlighting prompt. To this end, fig. 4.3 offers the visual comparison of segmentation results produced by the variants of our method on five images of the PASCAL VOC [13] dataset.

The image-text co-decomposition module equips the model with the region-word alignment ability to localize objects in the images accurately. This module aligns words with corresponding regions in the image, leading to more precise segmentation results. Furthermore, both the word and region highlighting prompts contribute to feature extraction, improving the model's ability to capture the details of the objects. Hence, the resultant model is more effective in segmenting the whole objects of interest.

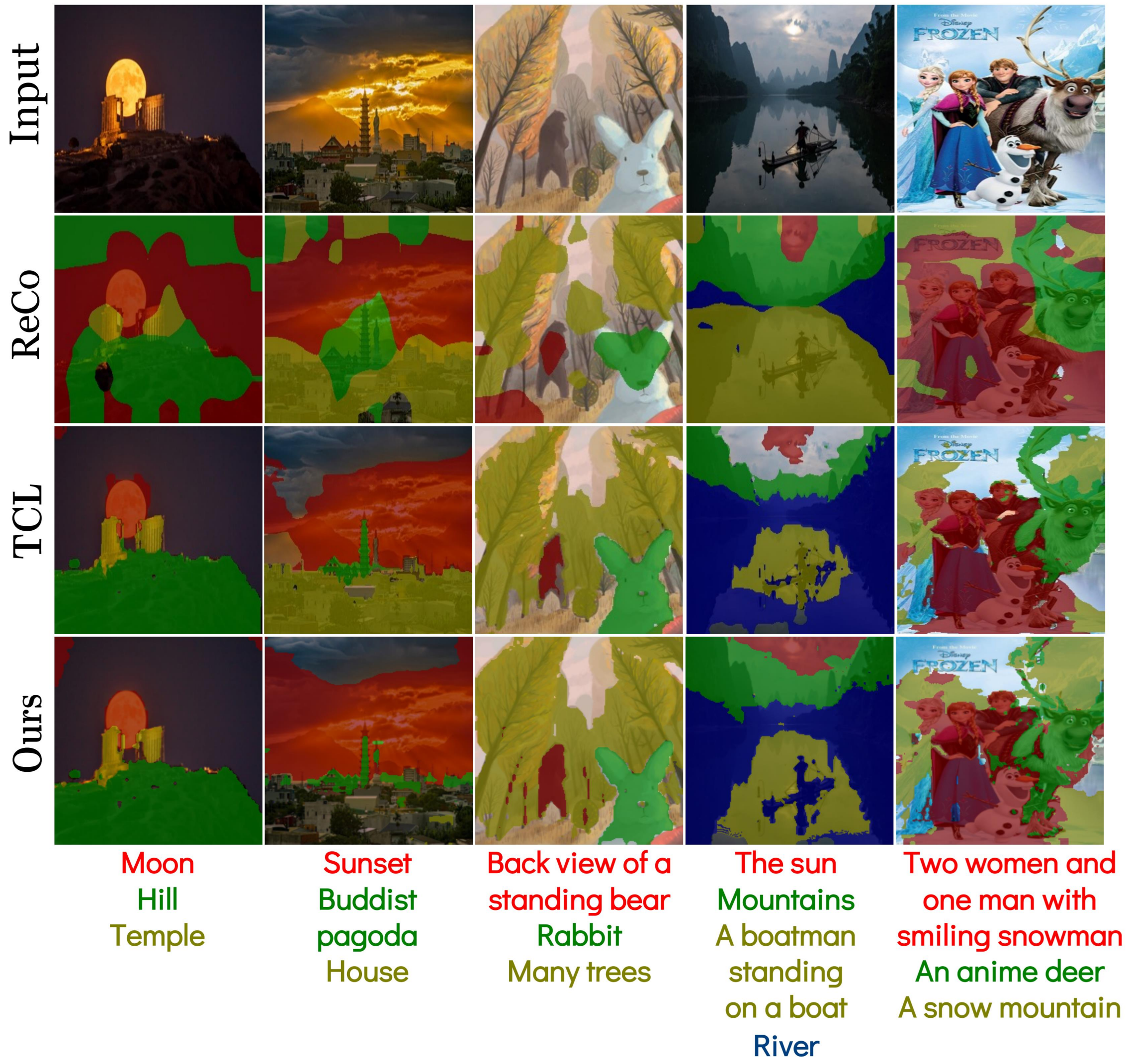

Figure 4.4: **Examples in the wild.** We show predictions on wild images with free-form text queries. Texts used as target classes are shown at the bottom of the images.

4.6 Multi-Noun Queries

Fig. 4.4 shows predictions on wild web images with various text queries using the same images and queries selected from Fig. 5 of TCL [5]. Although our method is primarily designed and trained for single-noun queries, the figure demonstrates its effectiveness in processing more complex queries.

4.7 Failure case visualization

In fig. 4.5, we show several failure cases of our method and two competing methods, TCL [5] and SimSeg [50], on the images of the PASCAL VOC [13] dataset.

The first example in fig. 4.5a shows a common limitation of existing methods: When segmenting the “person” class, most methods focus on the most distinctive areas, namely the face in this example, and suffer from the variations in the clothes, resulting in the segment that does not cover the entire person. The second example in fig. 4.5b depicts a scenario, where unexpected variations are present, i.e., people showing in a television monitor. All three methods segment the outer borders of the monitor. Compared to TCL and SimSeg, our method can further segment the individuals within the monitor. Although the ground truth covers the entire TV monitor, this example validates the effectiveness of our model in localizing the individuals present on the screen.

fig. 4.5c, fig. 4.5d, and fig. 4.5e showcase instances where co-occurrent objects, such as trains and tracks, airplanes and contrails, and boats and water, tend to be segmented together even though they are of different semantic categories. This is a challenge for our method and the two competing methods TCL [5] and SimSeg [50]. These visualization examples emphasize the difficulties of accurate segmentation and the challenges in aligning model predictions with ground truth annotations. They provide insights into the limitations of current segmentation approaches and suggest future research directions.

doi:10.6342/NTU202400974

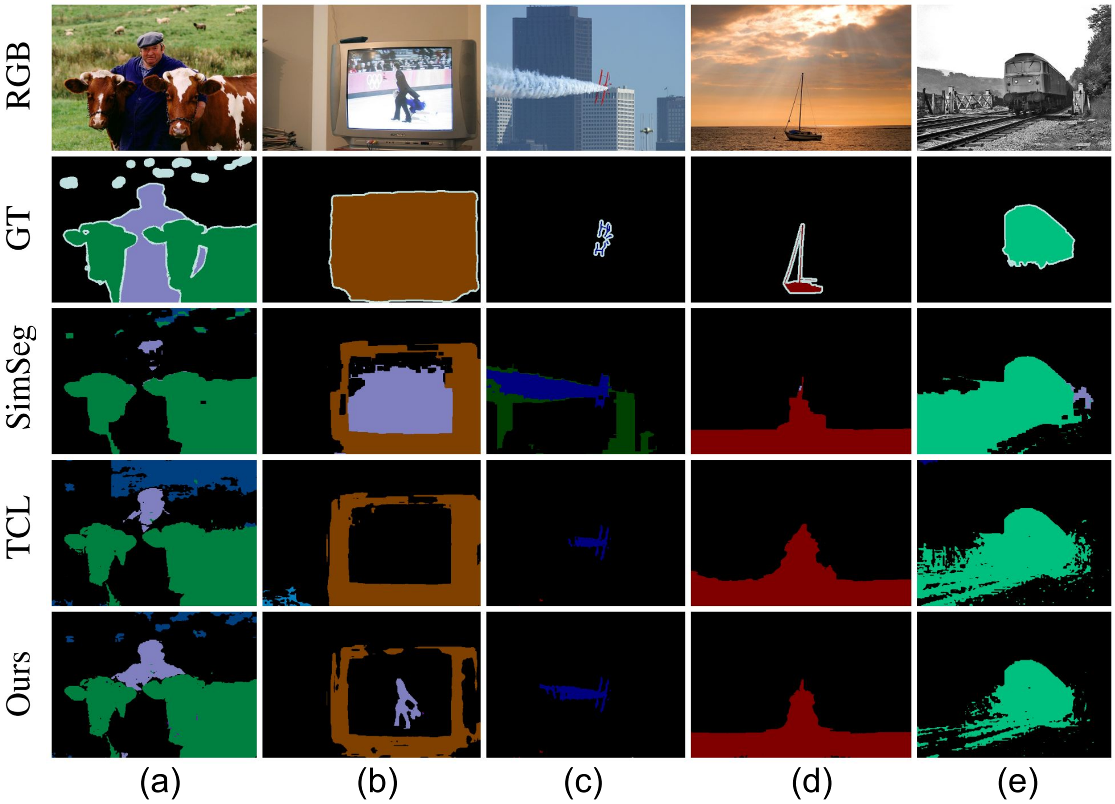

Figure 4.5: **Failure cases.** The proposed method is compared with the two most competitive methods, TCL [5] and SimSeg [50], on the images of the PASCAL VOC [13] dataset.

Training Time. On four NVIDIA 2080Ti GPUs, it takes eight hours to train the baseline model with only the image segmenter. On the same devices, it takes twelve hours to train our image-text co-decomposition method, which requires training an additional text segmenter. In light of the improved performance as described above, the longer training period can be justified.

doi:10.6342/NTU202400974

doi:10.6342/NTU202400974

Chapter 5 Conclusions

We propose Image-Text Co-Decomposition (CoDe) to address cross-domain alignment discrepancies in the existing methods for text-supervised semantic segmentation. First, our method decomposes image-text pairs into corresponding regions and word segments to enforce the region-word alignment. CoDe, underpinned by contrastive learning, alleviates the train-test discrepancy by unifying image-text and region-text alignments to region-word alignment. Then, we introduce a region-highlighting prompt learning method to enhance feature extraction on masked images or texts for precise region-word alignment. Moreover, CoDe surpasses state-of-the-art methods in zero-shot semantic segmentation across six benchmark datasets. This novel approach opens new possibilities for research in vision-language models and their broader applications in computer vision.

doi:10.6342/NTU202400974

References

[1] N. Araslanov and S. Roth. Single-stage semantic segmentation from image labels. In CVPR, 2020.

[2] S. Bird, E. Klein, and E. Loper. Natural language processing with Python: analyzing text with the natural language toolkit. 2009.

[3] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. In CVPR, 2018.

[4] K. Cai, P. Ren, Y. Zhu, H. Xu, J. Liu, C. Li, G. Wang, and X. Liang. Mixreorg: Cross-modal mixed patch reorganization is a good mask learner for open-world semantic segmentation. In ICCV, 2023.

[5] J. Cha, J. Mun, and B. Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In CVPR, 2023.

[6] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In CVPR, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.

doi:10.6342/NTU202400974

[8] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus. Probabilistic embeddings for cross-modal retrieval. In CVPR, 2021.

[9] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.

[11] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS, 2020.

[12] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li. Learning to prompt for open-vocabulary object detection with vision-language model. In CVPR, 2022.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.

[14] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In ECCV, 2022.

[15] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst., 2020.

[16] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels. In ECCV, 2022.

[17] C. Han, Y. Zhong, D. Li, K. Han, and L. Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In ICCV, 2023.

[18] Q. Huang, X. Dong, D. Chen, W. Zhang, F. Wang, G. Hua, and N. Yu. Diversity-aware meta visual prompting. In CVPR, 2023.

[19] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In ECCV, 2022.

[20] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. Maple: Multi-modal prompt learning. In CVPR, 2023.

[21] D. Kim, N. Kim, and S. Kwak. Improving cross-modal retrieval with set of diverse embeddings. In CVPR, 2023.

[22] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In ECCV, 2018.

[23] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691, 2021.

[24] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation. In ICLR, 2022.

[25] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.

[26] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In CVPR, 2023.

doi:10.6342/NTU202400974

[27] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023.

[28] Q. Liu, Y. Wen, J. Han, C. Xu, H. Xu, and X. Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In ECCV, 2022.

[29] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[30] H. Luo, J. Bao, Y. Wu, X. He, and T. Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. ICML, 2023.

[31] C. Ma, Y. Yang, Y. Wang, Y. Zhang, and W. Xie. Open-vocabulary semantic segmentationwith frozen vision-language models. In BMVC, 2022.

[32] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. CVPR, 2014.

[33] P. Pandey, M. Chasmai, M. Natarajan, and B. Lall. A language-guided benchmark for weakly supervised open vocabulary semantic segmentation. arXiv preprint arXiv:2302.14163, 2023.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

[35] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In CVPR, 2022.

doi:10.6342/NTU202400974

[36] P. Ren, C. Li, H. Xu, Y. Zhu, G. Wang, J. Liu, X. Chang, and X. Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. ICLR, 2023.

[37] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL, 2018.

[38] G. Shin, W. Xie, and S. Albanie. Reco: Retrieve and co-segment for zero-shot transfer. NIPS, 2022.

[39] Y. Song and M. Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In CVPR, 2019.

[40] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In ICCV, 2021.

[41] J. Wang, X. Li, J. Zhang, Q. Xu, Q. Zhou, Q. Yu, L. Sheng, and D. Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773, 2023.

[42] J.-J. Wu, A. C.-H. Chang, C.-Y. Chuang, C.-P. Chen, Y.-L. Liu, M.-H. Chen, H.-N. Hu, Y.-Y. Chuang, and Y.-Y. Lin. Image-text co-decomposition for text-supervised semantic segmentation. arXiv preprint arXiv:2404.04231, 2024.

[43] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. ICCV, 2023.

[44] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer:

Simple and efficient design for semantic segmentation with transformers. NIPS, 2021.

[45] Y. Xing, J. Kang, A. Xiao, J. Nie, S. Ling, and S. Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In NIPS, 2023.

[46] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang. Groupvit: Semantic segmentation emerges from text supervision. In CVPR, 2022.

[47] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In CVPR, 2023.

[48] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai. Side adapter network for open-vocabulary semantic segmentation. In CVPR, 2023.

[49] H. Yao, R. Zhang, and C. Xu. Visual-language prompt tuning with knowledge-guided context optimization. In CVPR, 2023.

[50] M. Yi, Q. Cui, H. Wu, C. Yang, O. Yoshie, and H. Lu. A simple framework for text-supervised semantic segmentation. In CVPR, 2023.

[51] X. Yuan, J. Shi, and L. Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications, 2021.

[52] F. Zhang, T. Zhou, B. Li, H. He, C. Ma, T. Zhang, J. Yao, Y. Zhang, and Y. Wang. Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. NIPS, 2023.

[53] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In CVPR, 2021.

[54] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. IJCV, 2019.

[55] C. Zhou, C. C. Loy, and B. Dai. Extract free dense labels from clip. In ECCV, 2022.

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In CVPR, 2022.

[57] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. IJCV, 2022.

[58] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang. Prompt-aligned gradient for prompt tuning. In ICCV, 2023.