Prompting Multi-Modal Image Segmentation with Semantic Grouping

Qibin He

University of Chinese Academy of Sciences, Beijing, China
qibin.he@outlook.com

Abstract

Multi-modal image segmentation is one of the core issues in computer vision. The main challenge lies in integrating common information between modalities while retaining specific patterns for each modality. Existing methods typically perform full fine-tuning on RGB-based pre-trained parameters to inherit the powerful representation of the foundation model. Although effective, such paradigm is not optimal due to weak transferability and scarce downstream data. Inspired by the recent success of prompt learning in language models, we propose the **Gr**o**u**ping **P**rompt **T**uning Framework (GoPT), which introduces explicit semantic grouping to learn modal-related prompts, adapting the frozen pre-trained foundation model to various downstream multi-modal segmentation tasks. Specifically, a class-aware uni-modal prompter is designed to balance intra- and inter-modal semantic propagation by grouping modality-specific class tokens, thereby improving the adaptability of spatial information. Furthermore, an alignment-induced cross-modal prompter is introduced to aggregate class-aware representations and share prompt parameters among different modalities to assist in modeling common statistics. Extensive experiments show the superiority of our GoPT, which achieves SOTA performance on various downstream multi-modal image segmentation tasks by training only $< 1\%$ model parameters.

Introduction

Semantic segmentation aims to assign a semantic category to each pixel in the observed scene, which plays an important role in various applications (Zhao et al. 2017; Liu et al. 2022a; Zheng et al. 2021). With the development of sensor technology, multi-modal fusion using multiple data sources for segmentation has become one of the core issues in image interpretation (Hazirbas et al. 2017; Zhou, Ruan, and Canu 2019; Zhang et al. 2021).

Thanks to the success of deep learning, multi-modal fusion has recently been specifically referred to as deep multi-modal fusion (Hazirbas et al. 2017; Wang et al. 2022b) through end-to-end neural integration of multiple image modalities, and has shown significant advantages over uni-modal segmentation (Park et al. 2017; Li et al. 2023). Deep learning pipelines for multi-modal segmentation are usually expected to capture strong semantics and rich spatial details, which benefit greatly from joint training on multi-modal data sources. According to the type of fusion, existing methods are generally classified into *alignment-based* fusion (Valada et al. 2020; Wang et al. 2020) and *aggregation-based* fusion (Wang et al. 2022a; Zhou et al. 2022a). Despite the fruitful progress, these methods still face great challenges: *integrating common information between modalities while retaining the specific patterns of each modality remains dilemmatic*.

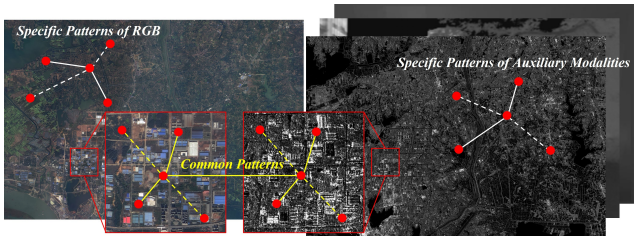

Figure 1: Multi-modal images with complex structural relations. Here, red dots indicate objects with solid lines indicating similarity to each other and dashed lines indicating dissimilarity.

More specifically, (i) *Cross-modal Gaps*. Modalities with different imaging mechanisms may have heterogeneous gaps, *i.e.*, the same object has different descriptions in different data sources and requires inter-modal alignment, as shown in fig. 1. However, alignment-based fusion mostly provides invalid inter-modal fusion due to weak information exchange through only training the alignment loss (Zhou et al. 2022b; Shivakumar et al. 2020). (ii) *Information Imbalances*. The effective information that different modalities can provide differs due to various features across domains, *e.g.*, acquisition frequency, visual disparity, *etc*. However, aggregation-based fusion tends to ignore intra-modal propagation, resulting in insufficient balance between inter-modal knowledge sharing and intra-modal information processing (Xu et al. 2017; Gao et al. 2021; Ordun 2023).

Given the above challenges, we attempt to learn discriminative and compact modality-specific and shared patterns via multi-modal grouping to alleviate cross-modal gaps and information imbalances in multi-modal image segmentation. Grouping aims to reorganize pixels into candidate groups

to provide explicit semantic priors and facilitate recognition (Roelfsema and Houtkamp 2011; Watt and Phillips 2000; Hosseinaee et al. 2021). Considering that multi-modal methods usually employ RGB-based pre-trained segmenters, we propose to introduce the semantic grouping mechanism into the fine-tuning on task-oriented training sets. While the full fine-tuning method is effective, it is inefficient and causes a large burden of parameter storage (Lee and Kwon 2017; Lin et al. 2020). In addition, due to limited sample labeling, full fine-tuning cannot take advantage of the pre-training knowledge of the foundation model trained on large-scale datasets to obtain generalized representations (Sun, Zuo, and Liu 2019). Inspired by recent prompt tuning in Natural Language Processing (NLP) (Lester, Al-Rfou, and Constant 2021; Liu et al. 2022b), we try to improve the adaptation efficiency of downstream multi-modal segmentation by freezing the foundation model and fine-tuning only the visual prompt parameters.

To this end, we propose **Gr**ouping **P**rompt **T**uning (GoPT), which provides unified visual prompts for multi-modal image segmentation by explicitly grouping context semantics. As shown in Figure fig. 2, unlike full fine-tuning the RGB segmenters combined with other auxiliary modal branches, GoPT only needs to learn modality-specific visual prompts to maximize the inheritance of prior knowledge from large-scale RGB pretraining. Specifically, GoPT first inserts the class-aware uni-modal prompter (CUP) into the frozen foundation model, which employs uni-modal grouping to extract modality-complementary features. Then, by introducing the alignment-induced cross-modal prompter (ACP) to group cross-modal spatial contexts, GoPT mines object patterns from updated embeddings. Notably, GoPT is a general framework for various multi-modal image segmentation, including RGB-Depth (RGB-D), RGB-Thermal (RGB-T), and RGB-Synthetic Aperture Radar (RGB-SAR) segmentation. We summarize the main contributions of this paper as follows:

- A grouping prompt tuning framework for task-oriented multi-modal image segmentation is proposed, which can be generalized to various tasks, *i.e.*, RGB-D, RGB-T and RGB-SAR segmentation. By simplifying auxiliary modalities to a few prompts instead of designing additional network branches, GoPT effectively adapts the off-the-shelf RGB-based pre-trained foundation model to downstream multi-modal segmentation.
- A class-aware uni-modal prompter is designed to balance intra- and inter-modal semantic propagation by grouping modality-specific class tokens, thereby improving the adaptability of spatial information.
- An alignment-induced cross-modal prompter is introduced to aggregate class-aware representations and share prompt parameters among different modalities to assist in modeling common statistics.
- Extensive experiments show the superiority of our GoPT, which achieves SOTA performance on multiple downstream multi-modal image segmentation tasks by training only $< 1\%$ parameters.

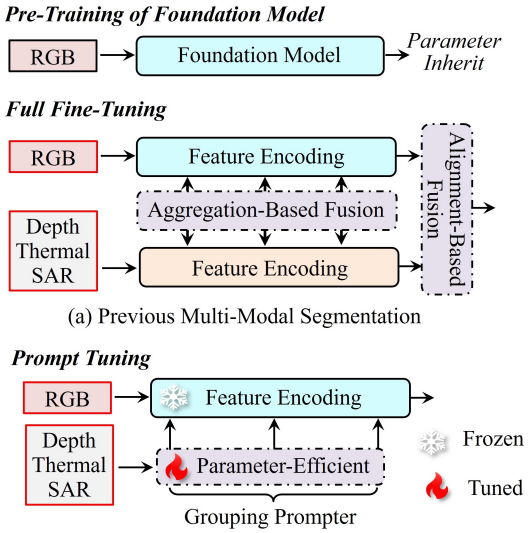

Figure 2: Motivation of GoPT. It introduces parameter-efficient prompt learning into multi-modal segmentation and combines semantic grouping to adapt the foundation model to downstream tasks with a more concise network structure.

Related Work

Multi-Modal Image Segmentation

As a mainstream method for multi-modal image segmentation, deep multi-modal fusion aims to enhance fine-grained details and pixel-level semantics using multiple data sources to combat uni-modal defects (Zhou, Ruan, and Canu 2019; Ha et al. 2017; Hazirbas et al. 2017). In fact, relevant methods of multi-modal segmentation are basically divided into alignment- and aggregation-based fusion. Alignment-based fusion methods (Valada et al. 2020; Wang et al. 2020) employ conditional loss to align subnetwork embedding while maintaining the full propagation of each subnetwork. These methods align multi-modal features by applying similarity rules, with maximizing mean difference being the most commonly used (Park et al. 2017). However, focusing solely on the entire uniform distribution may miss specific patterns in each mode/domain. Thus, Wang *et al.* (Wang et al. 2022c) offer a possible mitigation of this problem, which preserves modality-specific information while associating modality-common features. Whereas aggregation-based fusion methods apply specific operators to combine multi-modal subnetworks into a single network and fuse features, *e.g.*, concatenating (Shivakumar et al. 2020), averaging (Sun, Zuo, and Liu 2019), and attention-based modules (Wang et al. 2022b). Considering the inadequacy of intra-modal propagation, recent aggregation-based methods perform feature fusion while maintaining subnetworks of all modalities (Wang et al. 2022a; Zhou et al. 2022a).

However, due to the lack of large-scale multi-modal training sets, data-driven deep learning models are facing potential risks of increasing overfitting (Lee and Kwon 2017). Therefore, the above two types of methods usually load

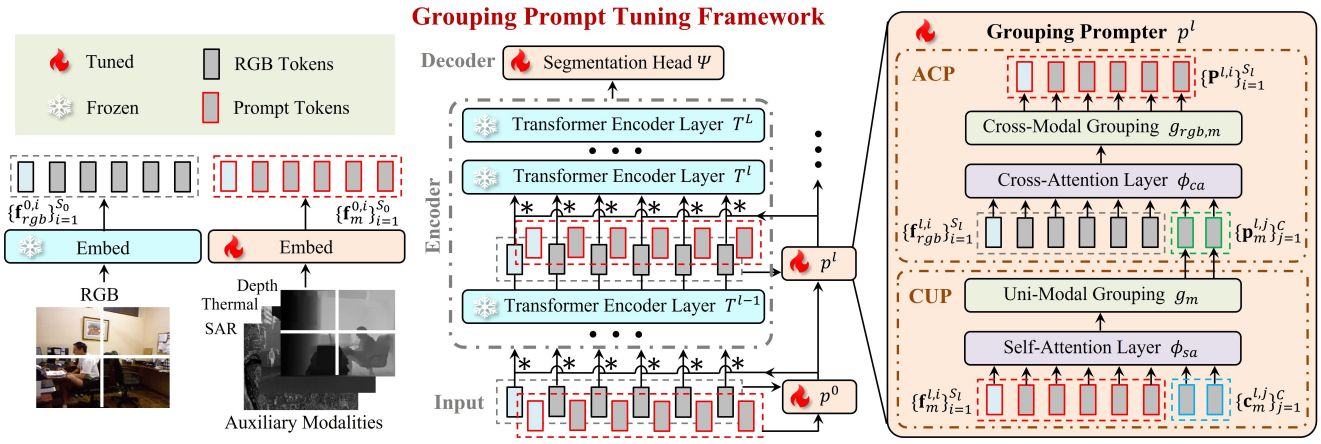

Figure 3: Overview of GoPT. RGB and auxiliary modalities are first to generate corresponding tokens with patch embedding, and then fed to the L-layer stacked visual transformer for feature encoding. Grouping prompters $\{p^l\}_{l=0}^{L-1}$ are inserted into the foundation model to learn visual prompts, where CUP improves intra-modal semantic propagation by learning modality-specific class tokens, while ACP aggregates class-aware representations and assists in modeling modality-common statistics.

RGB-based pre-trained model parameters first, and then fine tune on specific downstream task-oriented datasets (Gao et al. 2021). In this work, we propose a fusion method based on semantic grouping prompts to explore the adaptation from RGB-based pre-trained models to downstream multi-modal segmentation tasks, which only requires tuning a few parameters to achieve intra-modal learning and inter-modal interaction.

Visual Prompt Learning

Large pre-trained models have long relied on fine-tuning to perform specific downstream tasks, where all model parameters typically need to be updated during downstream data training (Erhan et al. 2010; Devlin et al. 2018; He et al. 2022). This approach requires storing multiple task-oriented copies of the entire pre-trained model. Therefore, prompt tuning has been proposed as a new paradigm to overcome such parameter-inefficient dilemma (Lester, Al-Rfou, and Constant 2021; Liu et al. 2022b). It greatly improves the performance of many downstream NLP tasks and shows a strong generalization ability on transfer learning. Recently, prompts have been applied to visual tasks. Sandler *et al.* (Sandler et al. 2022) introduce memory tokens to embed a set of learnable vectors in each transformer layer. Jia *et al.* (Jia et al. 2022) propose a similar idea and study the universality and feasibility of visual prompts through extensive experiments on several recognition tasks across backbone architectures and multiple domains. Unlike most existing methods that focus on uni-modal recognition tasks, our work aims to explore the core ideas of visual prompt learning and design a prompt tuning framework for multi-modal image segmentation to replace the fully fine-tuning paradigm.

Methodology

In this work, we propose GoPT to adapt the RGB-based pre-trained foundation model to downstream multi-modal segmentation tasks in a simple but efficient manner. Compared with full fine-tuning on the foundation model, GoPT only needs to tune a few prompt parameters to achieve satisfactory cross-modal transfer learning capabilities. The overall architecture is shown in fig. 3.

Preliminaries

Problem Setup. Multi-modal image segmentation improves the performance of the original segmenter F pre-trained on RGB data $\mathbf{x}_{rgb}$ by introducing an additional spatially aligned input $\mathbf{x}_m$, where the subscript m indicates other auxiliary modalities (*e.g.*, depth, thermal infrared, or synthetic aperture radar). Therefore, multi-modal image segmentation can be expressed as $F : \{\mathbf{x}_{rgb}, \mathbf{x}_m\} \rightarrow \mathbf{S}$, where $\mathbf{S}$ is the segmentation mask.

Foundation Model. Generally, the segmenter F can be decomposed into $\Psi \circ B$, where $B : \mathbf{x}_{rgb} \rightarrow \mathbf{f}_{rgb}$ indicates the feature extraction function, and the segmentation head $\Psi : \mathbf{f}_{rgb} \rightarrow \mathbf{S}$ predicts the final result. In our scheme, B is a powerful L-layer vision transformer backbone, *i.e.*, MAE (He et al. 2022) pre-trained ViT (Dosovitskiy et al. 2020). Specifically, the input sample $\mathbf{x}_{rgb}$ is first divided into non-overlapping patches, and the corresponding initialization tokens are generated by linear projection. Formally, the tokens of the l-th layer transformer T^l are indicated as $\{\mathbf{f}_{rgb}^{l,i}\}_{i=1}^{S_l}$, where S_l represents the number of tokens in the l-th layer. The forward propagation process can be defined as:

$$\{\mathbf{f}_{rgb}^{l,i}\}_{i=1}^{S_l} = T^l(\{\mathbf{f}_{rgb}^{l-1,i}\}_{i=1}^{S_{l-1}}), \quad l = 1, 2, \ldots, L \tag{1}$$

$$\mathbf{S} = \Psi(\{\mathbf{f}_{rgb}^{L,i}\}_{i=1}^{S_L}), \tag{2}$$

where $\{\mathbf{f}_{rgb}^{L,i}\}_{i=1}^{S_L}$ is the feature output of the last encoding layer. For more information on the RGB foundation model, please see MAE (He et al. 2022) and SETR (Zheng et al. 2021).

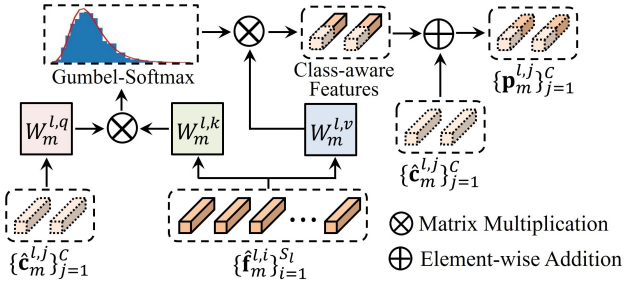

Figure 4: Structure of uni-modal grouping module g_m. It captures the similarity between auxiliary modal tokens $\{\hat{\mathbf{f}}_m^{l,i}\}_{i=1}^{S_l}$ and learned class tokens $\{\hat{\mathbf{c}}_m^{l,j}\}_{j=1}^{C}$. Through computing one-hot assignments using Gumbel-Softmax on class tokens, auxiliary modal tokens assigned to the same group are merged together to update prompts $\{\mathbf{p}_m^{l,j}\}_{j=1}^{C}$.

Overview

Multi-modal image segmentation provides an additional auxiliary modality stream, which is spatially aligned and time-synchronized with the RGB stream. As shown in fig. 3, GoPT first feeds two inputs $\{\mathbf{x}_{rgb}, \mathbf{x}_m\}$ to the patch embedding layer to obtain the corresponding RGB tokens $\{\mathbf{f}_{rgb}^{0,i}\}_{i=1}^{S_0}$ and auxiliary modality tokens $\{\mathbf{f}_m^{0,i}\}_{i=1}^{S_0}$. Then, $\{\mathbf{f}_{rgb}^{0,i}\}_{i=1}^{S_0}$ is fed to the foundation model, $\{\mathbf{f}_{rgb}^{0,i}\}_{i=1}^{S_0}$ and $\{\mathbf{f}_m^{0,i}\}_{i=1}^{S_0}$ are sent to the grouping prompter to generate modality-specific prompts. The learned prompts are added to the original RGB stream as residuals:

$$\{\mathbf{f}^{l,i}\}_{i=1}^{S_l} = \{\mathbf{f}_{rgb}^{l,i} * \mathbf{p}^{l,i}\}_{i=1}^{S_l}, \quad l = 0, 1, \ldots, L-1 \tag{3}$$

where $\{\mathbf{f}^{l,i}\}_{i=1}^{S_l}$ indicates the token that will be fed to the next layer of the foundation model, $\{\mathbf{p}^{l,i}\}_{i=1}^{S_l}$ is the prompt token from the l-th grouping prompter, and "$*$" is element-wise multiplication. Adding prompts directly to the intermediate features of the foundation model enables our GoPT to be easily deployed on existing pre trained segmenters. The introduction of hierarchical prompters can also make full use of the semantic analysis of different modalities and different levels. Notably, the network parameters related to the RGB modality are frozen during training, including patch embedding and feature extraction.

Grouping Prompt Tuning

The proposed grouping prompter is inserted into multiple stages of the foundation model to achieve rapid learning by fine-tuning only a few parameters. It mainly includes class-aware uni-modal prompter (CUP) and alignment-induced cross-modal prompter (ACP). CUP balances intra- and inter-modal information processing by learning modality-specific class tokens, while ACP aggregates class-aware representations to model modality-public statistics.

Class-Aware Uni-Modal Prompter (CUP). To explicitly group class-aware matching semantics from the auxiliary modality $\mathbf{x}_m$, we design CUP that performs hierarchical progressive grouping of visual concepts. CUP introduces modality-specific class tokens $\{\mathbf{c}_m^{l,j}\}_{j=1}^{C}$ to help prompt the original patch features $\{\mathbf{f}_m^{l,i}\}_{i=1}^{S_l}$, where C indicates the number of semantic categories. We apply a self-attention layer to temporally aggregate uni-modal features and align the features with class token embedding:

$$\{\hat{\mathbf{f}}_m^{l,i}\}_{i=1}^{S_l}, \{\hat{\mathbf{c}}_m^{l,j}\}_{j=1}^{C} = \{\phi_{sa}(\mathbf{z}_m^{l,h}, \mathbf{Z}_m^l, \mathbf{Z}_m^l)\}_{h=1}^{S_l+C}, \tag{4}$$

$$\mathbf{Z}_m^l = \{\mathbf{z}_m^{l,h}\}_{h=1}^{S_l+C} = [\{\mathbf{f}_m^{l,i}\}_{i=1}^{S_l}; \{\mathbf{c}_m^{l,j}\}_{j=1}^{C}], \tag{5}$$

where $[;]$ indicates the concatenation operator, and ϕ_{sa} indicates the self-attention function:

$$\phi_{sa}(\mathbf{z}_m^{l,h}, \mathbf{Z}_m^l, \mathbf{Z}_m^l) = \mathrm{Softmax}(\frac{\mathbf{z}_m^{l,h}\mathbf{Z}_m^{l\top}}{\sqrt{d}})\mathbf{Z}_m^l, \tag{6}$$

where d indicates the feature dimension of each token. Then, the uni-modal grouping module $g_m(\cdot)$ takes the aggregated features and learned semantic class tokens as input to generate class-aware prompt embedding:

$$\{\mathbf{p}_m^{l,j}\}_{j=1}^{C} = g_m(\{\hat{\mathbf{f}}_m^{l,i}\}_{i=1}^{S_l}, \{\hat{\mathbf{c}}_m^{j}\}_{j=1}^{C}), \tag{7}$$

as shown in fig. 4. During grouping, all uni-modal features belonging to the same class token are merged into the new uni-modal class-aware features. Similarity matrix $\mathbf{A}_m^{l,(i,j)}$ between class tokens and uni-modal features is computed via $\mathrm{Gumbel\text{-}Softmax}$ operator (Jang, Gu, and Poole 2017):

$$\begin{aligned}\mathbf{A}_m^{l,(i,j)} &= \mathrm{Gumbel\text{-}Softmax}(W_m^{l,q}\hat{\mathbf{f}}_m^{l,i} \cdot W_m^{l,k}\hat{\mathbf{c}}_m^{l,j}) \\ &= \frac{\exp(W_m^{l,q}\hat{\mathbf{f}}_m^{l,i} \cdot W_m^{l,k}\hat{\mathbf{c}}_m^{l,j} + \gamma_m^{l,j})}{\sum_{h=1}^{C}\exp(W_m^{l,q}\hat{\mathbf{f}}_m^{l,i} \cdot W_m^{l,k}\hat{\mathbf{c}}_m^{l,h}) + \gamma_m^{l,h}},\end{aligned} \tag{8}$$

where $W_m^{l,q}$ and $W_m^{l,k}$ are the learned weights for linear projection of the input modality features and class tokens, respectively. $\{\gamma_m^{l,j}\}$ are independent samples drawn randomly from the $\mathrm{Gumbel}(0, 1)$ distribution. Based on such similarity, we perform $\mathrm{one\text{-}hot}$ with argmax on all prompt embedding to compute the groups to assign image tokens to. Considering that the $\mathrm{one\text{-}hot}$ assignment of argmax is non-differentiable, the straight through technique in (Van Den Oord, Vinyals et al. 2017) is introduced:

$$\hat{\mathbf{A}}_m^l = \mathrm{one\text{-}hot}(\mathbf{A}_{m,\mathrm{argmax}}^l) + \mathbf{A}_m^l - \mathrm{sg}(\mathbf{A}_m^l), \tag{9}$$

where sg indicates the gradient stop operator. $\hat{\mathbf{A}}_m^l$ obtains one-hot values assigned to individual groups, and has the same gradient as $\mathbf{A}_m^l$. This $\mathrm{one\text{-}hot}$ assignment is called a hard assignment. After assigning image tokens to different learnable prompts, we merge the embedding of all tokens belonging to the same group to update the class-aware features:

$$\begin{aligned}\mathbf{p}_m^{l,j} &= g_m(\{\hat{\mathbf{f}}_m^{l,i}\}_{i=1}^{S_l}, \hat{\mathbf{c}}_m^{j}) \\ &= \hat{\mathbf{c}}_m^{j} + W_m^{l,o}\frac{\sum_{i=1}^{S_l}\hat{\mathbf{A}}_m^{l,(i,j)}W_m^{l,v}\hat{\mathbf{f}}_m^{l,i}}{\sum_{i=1}^{S_l}\hat{\mathbf{A}}_m^{l,(i,j)}},\end{aligned} \tag{10}$$

where $W_m^{l,o}$ and $W_m^{l,v}$ indicate the learning weights for merging projection features. Soft assignments (*i.e.*, $\mathbf{A}_m^l$ instead of $\hat{\mathbf{A}}_m^l$) can also be chosen to calculate eq. (10), but it has been empirically found that hard assignments allow for more efficient grouping.

Alignment-Induced Cross-Modal Prompter (ACP). The inter-modal gap caused by different imaging mechanisms requires fine-grained cross-modal interactions to facilitate information fusion. Since the alignment relationship between RGB and auxiliary modalities is explicitly available, we design ACP to aggregate class-aware representations from the auxiliary modality. According to the semantic similarity of explicit grouping, key patterns from other data sources are combined into RGB stream to generate new cross-modal alignment-induced prompts $\{\mathbf{p}^{l,i}\}_{i=1}^{S_l}$:

$$\{\mathbf{p}^{l,i}\}_{i=1}^{S_l} = g_{rgb,m}(\{\mathbf{p}_m^{l,j}\}_{j=1}^{C}, \{\mathbf{p}_{rgb,m}^{l,i}\}_{i=1}^{S_l}), \tag{11}$$

$$\{\mathbf{p}_{rgb,m}^{l,i}\}_{i=1}^{S_l} = \{\phi_{ca}(\mathbf{f}_{rgb}^{l,i}, \mathbf{P}_m^l, \mathbf{P}_m^l)\}_{i=1}^{S_l}, \tag{12}$$

$$\mathbf{P}_m^l = \{\mathbf{p}_m^{l,h}\}_{h=1}^{C}, \tag{13}$$

where $g_{rgb,m}$ is a grouping module similar to g_m in eq. (7), and ϕ_{ca} indicates the cross-attention function:

$$\phi_{ca}(\mathbf{f}_{rgb}^{l,i}, \mathbf{P}_m^l, \mathbf{G}_m^l) = \mathrm{Softmax}(\frac{\mathbf{f}_{rgb}^{l,i}\mathbf{P}_m^{l\top}}{\sqrt{d}})\mathbf{P}_m^l, \tag{14}$$

where d indicates the feature dimension. In the actual implementation, we insert grouping prompters into different levels of the foundation model, and use the inheritance mechanism to construct a progressive hierarchy. Specifically, modality-specific class tokens in the CUP module inherit from the cross-modal alignment-induced feature of the previous layer:

$$\{\mathbf{c}_m^{l+1,j}\}_{j=1}^{C} = \{\mathbf{p}_m^{l,j}\}_{j=1}^{C}. \tag{15}$$

The class tokens (*i.e.*, $\{\mathbf{c}_m^{0,j}\}_{j=1}^{C}$) input for the initial stage are only a set of learnable parameters with random initialization.

Optimization The multi-modal segmentation model is initialized by the parameters of the RGB-based pre-trained foundation model. While the data stream propagates throughout the model during prompt tuning, we only update gradient values for a few specific parameters, *i.e.* grouping prompter and segmentation head Ψ. Due to the differences in the category settings in different downstream multi-modal datasets, the parameters of Ψ also need to be fine-tuned. Even so, the default version of our GoPT still contains only 0.97M trainable parameters and beats the full fine-tuning paradigm. Tuning with a small number of prompt parameters can promote the model convergence in a short time and effectively inherit the prior knowledge of the pre-trained foundation model, so as to achieve rapid learning.

Experiments

Downstream Tasks

GoPT achieves the unification of several downstream multi-modal image segmentation tasks, among which three challenging tasks are selected to verify the effectiveness and superiority of the proposed method. (i) For RGB-D segmentation, we provide the comparison results of NYUDv2 (Silberman et al. 2012) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015). (ii) For RGB-T segmentation, we evaluate our segmenter on MFNet (Ha et al. 2017) and PST900 (Shivakumar et al. 2020). (iii) For RGB-SAR segmentation, we report experimental results on WHU-OS (Li et al. 2022). All experimental settings are kept the same, *i.e.*, prompt tuning on these downstream tasks without any specific modulation.

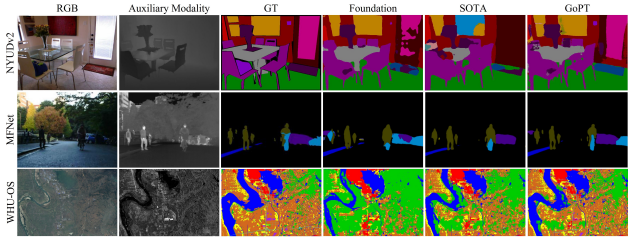

Figure 5: Qualitative visual comparison with foundation and SOTA segmenters on various multi-modal datasets.

Experimental Setup

To fully verify the superiority of the propsed method, when fine-tuning GoPT on downstream tasks, we only use the training sets corresponding to the above datasets without introducing other algorithms for data expansion. GoPT is trained on 1 NVIDIA Tesla A100 GPU with a batch size of 64 and fine-tuning epochs of 60. AdamW (Loshchilov and Hutter 2019) is employed as training optimizer, where the initial learning rate is 4×10^{-5} and scheduled following the polynomial annealing policy. The parameters of the pre-trained foundation model remain fixed, while the learnable prompt parameters are initialized with the xavier uniform scheme (Glorot and Bengio 2010). Concretely, our models include GoPT (B) and GoPT (L), corresponding to MAE (He et al. 2022) pre-trained ViT-B and ViT-L (Dosovitskiy et al. 2020) as the foundation model.

Comparison With State-of-the-Arts

NYUDv2. NYUDv2 (Silberman et al. 2012) is an indoor understanding dataset containing 1,449 RGB-Depth (RGB-D) pairs of size 640×480, split into 795/654 for `train`/`test` with 40 classes. In table 1, we compare our method with various SOTA methods. The results show that our GoPT achieves new records better than previous methods in all metrics, and even the basic version improves by 4.8% mIoU compared to our RGB-based foundation segmenter (Zheng et al. 2021). This also verifies the efficiency of GoPT to utilize depth information to assist segmentation.

SUN RGB-D. SUN RGB-D (Song, Lichtenberg, and Xiao 2015) is one of the most challenging benchmarks for indoor semantic segmentation, containing 10,335 RGB-D images of 37 semantic classes. We employ the standard `train`/`test` split. As shown in table 1, our GoPT (L) outperforms previous competing methods, achieving 52.3% mIoU. GoPT outperforms the foundation segmenter by 3.4% mIoU, while another prompt-based segmenter EVP (Liu et al. 2023) only boosts 0.9% mIoU over the baseline.

MFNet. MFNet (Ha et al. 2017) is an urban street dataset containing 1,569 RGB-T pairs of size 640×480 with 8 classes. 820 pairs are collected during the day and the other

Method	Modal	NYUDv2			SUN RGB-D		
		PAcc.	mAcc.	mIoU	PAcc.	mAcc.	mIoU
(Long et al. 2015)	RGB	60.0	42.2	29.2	68.4	41.1	29.0
(Lin et al. 2020)	RGB	74.4	59.6	47.6	81.1	57.7	47.0
(Zheng et al. 2021)	RGB	76.9	61.1	49.3	82.7	59.3	48.9
(Liu et al. 2023)	RGB	77.5	64.2	50.7	84.1	61.5	49.8
(Hazirbas et al. 2017)	RGB-D	68.1	50.4	37.9	76.3	48.3	37.3
(Valada et al. 2020)	RGB-D	75.2	60.5	48.7	81.0	58.1	45.7
(Park et al. 2017)	RGB-D	76.0	62.8	50.1	81.5	60.1	47.7
(Wang et al. 2020)	RGB-D	77.0	64.0	51.2	–	–	–
(Wang et al. 2022c)	RGB-D	77.7	65.0	52.5	83.5	63.2	51.1
(Wang et al. 2022b)	RGB-D	78.6	66.2	53.3	84.0	63.3	51.4
GoPT (B)	RGB-D	79.8	**67.5**	54.1	**85.7**	64.2	52.1
GoPT (L)	RGB-D	**80.1**	67.4	**54.3**	85.5	**64.6**	**52.3**

Table 1: Comparison results on the NYUDv2 and SUN RGB-D datasets with SOTAs for RGB-D segmentation. Evaluation metrics include pixel accuracy (PAcc.) (%), mean accuracy (mAcc.) (%), and mean IoU (mIoU) (%).

Method	Modal	MFNet		PST900
		mAcc.	mIoU	mIoU
(Zhao et al. 2017)	RGB	51.9	44.7	68.7
(Liu et al. 2022a)	RGB	57.6	51.4	70.6
(Ha et al. 2017)	RGB-T	45.1	39.7	57.0
(Sun, Zuo, and Liu 2019)	RGB-T	63.1	53.2	74.6
(Shivakumar et al. 2020)	RGB-T	55.5	48.6	68.4
(Zhou et al. 2021)	RGB-T	72.3	55.5	77.1
(Zhou et al. 2022b)	RGB-T	72.7	54.8	78.5
(Zhou et al. 2022a)	RGB-T	75.2	56.1	78.6
(Li et al. 2023)	RGB-T	75.9	56.2	80.5
GoPT (B)	RGB-T	**77.1**	57.4	81.3
GoPT (L)	RGB-T	77.0	**57.7**	**81.5**

Table 2: Comparison results on the MFNet and PST900 datasets with SOTAs for RGB-T segmentation.

Method	Modal	OA	AA	Kappa
(Melgani and Bruzzone 2004)	RGB-SAR	60.5	57.4	45.4
(Li et al. 2015)	RGB-SAR	48.3	49.3	34.7
(Xu et al. 2017)	RGB-SAR	65.8	67.8	41.8
(Lee and Kwon 2017)	RGB-SAR	66.3	68.9	54.0
(Gao et al. 2021)	RGB-SAR	67.9	67.3	55.2
(Wang et al. 2022a)	RGB-SAR	72.5	66.5	59.8
GoPT (B)	RGB-SAR	**72.9**	69.3	60.7
GoPT (L)	RGB-SAR	72.7	**70.4**	**60.9**

Table 3: Comparison results on the WHU-OS dataset with SOTAs for RGB-SAR segmentation. Evaluation metrics include overall accuracy (OA) (%), average accuracy (AA) (%), and Kappa coefficient (%).

749 pairs were captured at night. The `train` set consists of 50% daytime images and 50% nighttime images, while the `val` and `test` sets contain 25% daytime images and 25% nighttime images. We compare the proposed GoPT with recent RGB-T segmenter in table 2. The results show that our GoPT (B) achieves astonishing mAcc. and mIoU of 77.1% and 57.4%, beating the well-designed RGB-T segmenter RSFNet (Li et al. 2023).

PST900. PST900 (Shivakumar et al. 2020) contains 894 synchronized and calibrated RGB-T image pairs with 5 semantic classes. The ratio of `train`/`test` set is 2/1. As shown in table 2, GoPT (L) achieves the best performance compared to other SOTA methods. We argue that prompt tuning exploits thermal information to enable GoPT to adapt well to diverse segmentation scenarios.

WHU-OS. WHU-OS (Li et al. 2022) is a remote sensing dataset containing 100 RGB-SAR images of 5556×3704 pixels with 6 land-cover classes, splitting into 60%/20%/20% for `train`/`val`/`test`. In the experiment, the images are cropped to 512×512 patches without overlapping. table 3 shows that GoPT (B) is better than the runner-up MIFNet (Wang et al. 2022a) by 2.8% AA and 0.9% Kappa, and is better than other classic RGB-SAR segmenters. Although the high signal-to-noise ratio of remote sensing images is more challenging for model representation capabilities, our GoPT still shows convincing adaptability by explicitly grouping multi-modal contexts.

Ablation Studies and Analyses

We further explore the properties of our method on various downstream tasks, including experimental results of GoPT (B) on several representative benchmarks, *i.e.*, NYUDv2, MFNet and WHU-OS.

Visualization. We visualize the segmentation results on various tasks in fig. 5. It can be seen that GoPT has a more accurate semantic discrimination for some complex situations under the prompt of the auxiliary modality. For example, GoPT successfully captures sparse details of foreground objects and connects them to form an overall representation when encountering illumination changes and background interference. This shows that GoPT can perceive and parse diverse scenes in a more robust and generalized manner.

Variant Analysis. To verify the effectiveness of the grouping prompt, we study different variants of GoPT. For fair comparison, we introduce VPT (Jia et al. 2022) style variants and GoPT-shallow, where the input prompts of VPT-shallow and VPT-deep are replaced by embeddings of auxiliary modalities to adapt to multi-modal segmentation. Notably, in order to align RGB and auxiliary modalities, we

Method	Params	NYUDv2			MFNet		WHU-OS		
		PAcc.	mAcc.	mIoU	mAcc.	mIoU	OA	AA	Kappa
Foundation	–	76.9	61.1	49.3	72.8	54.6	68.1	65.2	56.5
FFT	112.59M	78.5	65.4	53.1	74.9	56.3	70.8	68.3	58.9
VPT-Shallow	0.54M	77.0	61.9	50.2	73.2	54.5	68.4	65.7	57.1
VPT-Deep	3.39M	78.3	64.2	52.8	74.1	55.7	69.8	67.0	58.3
GoPT-Shallow	0.68M	78.1	65.0	52.9	74.5	56.0	70.3	67.5	58.1
GoPT	0.97M	**79.8**	**67.5**	**54.1**	**77.1**	**57.4**	**72.9**	**69.3**	**60.7**
GoPT (w/o ACP)	–	78.6	65.4	53.7	75.3	56.4	70.9	67.8	58.4
GoPT (w/o CUP)	–	78.5	65.7	53.9	75.6	56.8	71.2	67.9	58.3

Table 4: Ablation studies on various downstream tasks. "Params" refers to the parameters that need to be trained.

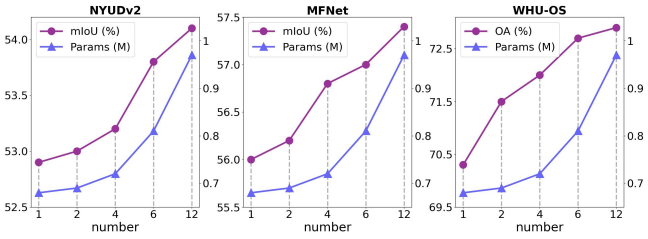

Figure 6: Impact of different number of grouping prompters.

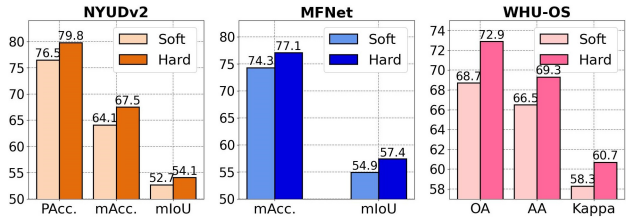

Figure 7: Impact of hard vs. soft assignment on grouping.

choose element-wise multiplication instead of the original concatenation in VPT to process prompt tokens. table 4 provides a comparison of the amount of trainable parameters and performance. Obviously, the proposed GoPT achieves the best balance in accuracy and efficiency.

Effectiveness of Prompter Structure. The proposed grouping prompter consists of CUP and ACP, which are analyzed for ablation in table 4. Compared with VPT, the introduction of CUP (*row* 7) promotes the accuracy of almost all metrics, verifying the effectiveness of promoting intra-modal semantic propagation. We also try to apply ACP directly to the foundation model (*row* 8), even with such a simple setup, it still brings performance improvements. We argue that the grouping structure in prompters can aid in mining the underlying modality-common distribution.

Impact of Multi-Modal Information. To explore the impact of multi-modal information on segmentation, we quantitatively evaluate the performance of uni-modal input (*i.e.*, RGB-based foundation segmenter) and multi-modal input, where the difference between GoPT and full fine-tuning (FFT) for the multi-modal case is also compared. FFT is extended from the foundation model by adding auxiliary modal branches, and the design of the specific structure is inspired by (Wang et al. 2022c). As shown in table 4, FFT combined with auxiliary modal input achieves a significant improvement over the foundation model, with significant increases in all three tasks. Even so, GoPT still has more performance advantages, and the amount of parameters required for training is less than 1% of FFT.

Number of Grouping Prompter. GoPT achieves multi-modal segmentation by inserting grouping prompters into different layers of the foundation model. We study the impact of the number of prompters on performance by setting different insertion intervals, *i.e.*, inserting prompters at every 1, 2, 4, 6 and 12 layers of the foundation model. As shown in fig. 6, the results on different tasks all show that the segmentation performance is positively correlated with the number of prompters.

Hard vs. Soft Assignment. We assign image tokens to prompts via hard or soft assignment in each grouping prompter. For soft assignment, we use the original $\mathbf{A}_m^l$ matrix instead of $\hat{\mathbf{A}}_m^l$ for hard assignment to compute eq. (10). As shown in fig. 7, we find that hard assignment improves by a large margin than soft one. We conjecture that in the case of soft assignment, since there are no zero values in $\mathbf{A}_m^l$, the features of the new image tokens are likely to be more correlated with each other. Thus, the information of the same image token may be assigned to different prompts, increasing ambiguity. However, in the case of hard assignment, the affinity matrix works in a mutually exclusive manner, making prompts more discriminative.

Conclusion

In this paper, we propose a novel parameter-efficient visual tuning framework for multi-modal image segmentation, *i.e.*, GoPT, by introducing explicit semantic grouping into prompt learning to adapt the frozen pre-trained foundation model to various downstream multi-modal segmentation tasks. Specifically, a class-aware uni-modal prompter is designed to balance intra- and inter-modal semantic propagation by grouping modality-specific class tokens. Furthermore, an alignment-induced cross-modal prompter is introduced to aggregate class-aware representations and assist in modeling common statistics. Extensive experiments on various downstream tasks demonstrate the superiority and generalization of the proposed GoPT.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Erhan, D.; Courville, A.; Bengio, Y.; and Vincent, P. 2010. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 201–208. JMLR Workshop and Conference Proceedings.

Gao, Y.; Li, W.; Zhang, M.; Wang, J.; Sun, W.; Tao, R.; and Du, Q. 2021. Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115. IEEE.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, 213–228. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Hosseinaee, Z.; Abbasi, N.; Pellegrino, N.; Khalili, L.; Mukhangaliyeva, L.; and Haji Reza, P. 2021. Functional and structural ophthalmic imaging using noncontact multimodal photoacoustic remote sensing microscopy and optical coherence tomography. *Scientific Reports*, 11(1): 1–11.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Lee, H.; and Kwon, H. 2017. Going deeper with contextual CNN for hyperspectral image classification. *IEEE Transactions on Image Processing*, 26(10): 4843–4855.

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

Li, P.; Chen, J.; Lin, B.; and Xu, X. 2023. Residual Spatial Fusion Network for RGB-Thermal Semantic Segmentation. *arXiv preprint arXiv:2306.10364*.

Li, W.; Chen, C.; Su, H.; and Du, Q. 2015. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7): 3681–3693.

Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; and Zhang, L. 2022. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102638.

Lin, G.; Liu, F.; Milan, A.; Shen, C.; and Reid, I. 2020. RefineNet: Multi-Path Refinement Networks for Dense Prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(5): 1228–1242.

Liu, S.-A.; Xie, H.; Xu, H.; Zhang, Y.; and Tian, Q. 2022a. Partial class activation attention for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16836–16845.

Liu, W.; Shen, X.; Pun, C.-M.; and Cun, X. 2023. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19434–19445.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68.

Long, J.; et al. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Melgani, F.; and Bruzzone, L. 2004. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42(8): 1778–1790.

Ordun, C. 2023. Multimodal Deep Generative Models for Remote Medical Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16127–16128.

Park, S.-J.; et al. 2017. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 4980–4989.

Roelfsema, P. R.; and Houtkamp, R. 2011. Incremental grouping of image elements in vision. *Attention, Perception, & Psychophysics*, 73: 2542–2572.

Sandler, M.; Zhmoginov, A.; Vladymyrov, M.; and Jackson, A. 2022. Fine-tuning image transformers using learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12155–12164.

Shivakumar, S. S.; Rodrigues, N.; Zhou, A.; Miller, I. D.; Kumar, V.; and Taylor, C. J. 2020. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, 9441–9447. IEEE.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760. Springer.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Sun, Y.; Zuo, W.; and Liu, M. 2019. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3): 2576–2583.

Valada, A.; et al. 2020. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5): 1239–1285.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, J.; Li, W.; Gao, Y.; Zhang, M.; Tao, R.; and Du, Q. 2022a. Hyperspectral and SAR Image Classification via Multiscale Interactive Fusion Network. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Y.; Chen, X.; Cao, L.; Huang, W.; Sun, F.; and Wang, Y. 2022b. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12186–12195.

Wang, Y.; Sun, F.; Huang, W.; He, F.; and Tao, D. 2022c. Channel exchanging networks for multimodal and multitask dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5481–5496.

Wang, Y.; Sun, F.; Lu, M.; and Yao, A. 2020. Learning deep multimodal feature representation with asymmetric multilayer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3902–3910.

Watt, R. J.; and Phillips, W. A. 2000. The function of dynamic grouping in vision. *Trends in cognitive sciences*, 4(12): 447–454.

Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; and Zhang, B. 2017. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2): 937–949.

Zhang, Y.; Sidibé, D.; Morel, O.; and Mériaudeau, F. 2021. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105: 104042.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhou, T.; Ruan, S.; and Canu, S. 2019. A review: Deep learning for medical image segmentation using multimodality fusion. *Array*, 3: 100004.

Zhou, W.; Dong, S.; Lei, J.; and Yu, L. 2022a. MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding. *IEEE Transactions on Intelligent Vehicles*, 8(1): 48–58.

Zhou, W.; Dong, S.; Xu, C.; and Qian, Y. 2022b. Edge-aware guidance fusion network for RGB–thermal scene parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3571–3579.

Zhou, W.; Lin, X.; Lei, J.; Yu, L.; and Hwang, J.-N. 2021. MFFENet: Multiscale feature fusion and enhancement network for RGB–Thermal urban road scene parsing. *IEEE Transactions on Multimedia*, 24: 2526–2538.