PMTSeg: Prompt-Driven Multimodal Transformer for Task-Adapted Remote Sensing Image Segmentation

Kejun Liu, Xuesong Yan, Yuanyuan Liu, *Member, IEEE*, Chang Tang, *Senior Member, IEEE*, Yibing Zhan, *Member, IEEE*, Wei Luo, Wujie Zhou, *Senior Member, IEEE*, and Hongyan Zhang, *Senior Member, IEEE*

***Abstract*— Multimodal remote sensing image segmentation (MRSIS) is important for intelligent remote sensing (RS) image interpretation, which encompasses three distinct tasks: semantic segmentation, instance segmentation, and panoptic segmentation. Existing methods typically address individual tasks with specialized models, limiting generalization and real-world applicability. Multitask learning (MTL) approaches have introduced separate task heads to unify tasks, yet we identify two key challenges when directly applying them to MRSIS: 1) the *modality gap*, arising from semantic discrepancies and granularity discrepancies across RS modalities and 2) the *task gap*, due to varying preferences in learning different segmentation tasks. To overcome these challenges, we propose *PMTSeg*, a novel *prompt-driven multimodal Transformer* for task-adapted MRSIS. PMTSeg integrates three key components: 1) task-common multimodal affinity approximation (TMAA); 2) task-common multiscale semantic fusion (TMSF); and 3) a unified prompt-driven segmentation head (PSH). First, TMAA addresses the *modality gap* by approximating intermodal affinity matrices, extracting task-common features across modalities, and aligning semantic information. Then, TMSF further integrates these features using the scale-matched fusion (SF) at multiple scales to produce enriched, multiscale task-common features. Moreover, to address the task gap, the PSH leverages task-adapted text prompts and task-adapted contrastive loss to model relationships across tasks, enabling adaptive optimization for robust and universal MRSIS performance. Extensive experiments on three MRSIS datasets—VALID, SEMCITY TOULOUSE, and UBCV2—demonstrate that PMTSeg significantly surpasses state-of-the-art methods in all three segmentation tasks, offering a unified and accurate solution to MRSIS.**

Received 28 November 2024; revised 27 May 2025; accepted 29 June 2025. Date of publication 7 July 2025; date of current version 23 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62076227, in part by the Natural Science Foundation of Hubei Province under Grant 2023AFB572, and in part by the Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2022-B10. *(Corresponding author: Yuanyuan Liu.)*

Kejun Liu, Xuesong Yan, Yuanyuan Liu, Chang Tang, and Hongyan Zhang are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: liukejun@cug.edu.cn; yanxs@cug.edu.cn; liuyy@cug.edu.cn; tangchang@cug.edu.cn; zhanghongyan@whu.edu.cn).

Yibing Zhan is with JD Explore Academy (JD.com), Beijing 100176, China (e-mail: zhanybjy@gmail.com).

Wei Luo is with the Department of Innovation Center, China Ship Development and Design Center, Wuhan 430074, China (e-mail: 171772014@qq.com).

Wujie Zhou is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310000, China (e-mail: wujiezhou@163.com).

Digital Object Identifier 10.1109/TGRS.2025.3586620

***Index Terms*— Multimodal image segmentation, multiscale semantic fusion, multitask segmentation, prompt-driven learning, remote sensing (RS).**

I. Introduction

MULTIMODAL remote sensing image segmentation (MRSIS) is a crucial task in remote sensing (RS) image interpretation, which has been widely applied in the fields of land-use classification [1], [2], urban planning [3], agriculture [4], and ecology [5]. It involves assigning semantic or instance labels to each pixel by leveraging different modalities. MRSIS encompasses three distinct segmentation tasks: RS semantic segmentation, RS instance segmentation, and RS panoptic segmentation, each serving different application needs. RS semantic segmentation focuses mainly on labeling pixels as different categories of land cover, for example, forests, grasslands, or water bodies. RS instance segmentation distinguishes each pixel as belonging to different individual instances, for example, segmenting specific building instances. RS panoptic segmentation needs to differentiate both the category of each pixel (including the foreground and background) and the instances, for example, tree1, tree2, grasslands, and water bodies.

Currently, many advanced MRSIS techniques are designed to address specific segmentation tasks, requiring distinct models for each task, which limits their versatility and reusability in practical applications [4]. For example, in RS semantic segmentation, Fan et al. [6] used a two-level fusion encoder for red, green, blue (RGB) and digital mapping system (DMS) data to enhance spatial detail and performance. For RS instance segmentation, Zhang et al. [1] utilized symmetric compact multimodal fusion to extract complementary information from RGB and near-infrared (NIR) data, to discern small objects in open backgrounds, thereby refining instance detection accuracy. For RS panoptic segmentation, Fernando et al. [4] proposed a multimodal teacher network with an attention-based fusion strategy to transfer knowledge to a unimodal student model. Despite the advancements, these methods focus exclusively on individual RS segmentation tasks, lacking the ability to handle multiple segmentation tasks in a unified framework. These task-specific methods limit their applicability in real-world environments that require simultaneous processing of diverse RS segmentation tasks.

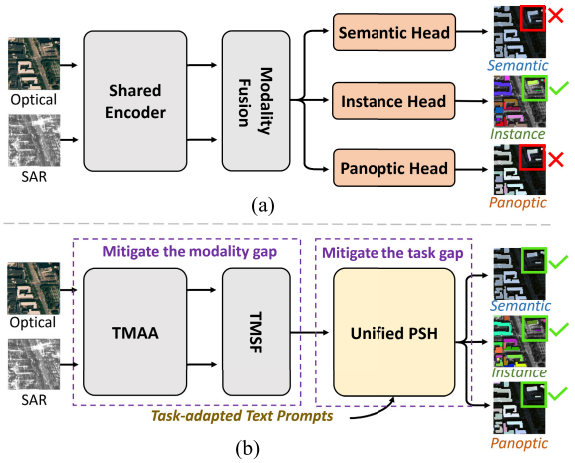

Fig. 1. Motivation of our PMTSeg method. (a) Traditional multitask methods rely on multiple task heads and a shared encoder to handle multimodal data, overlooking the modality and task gaps and resulting in suboptical performance. (b) Our PMTSeg employs three key components: TMAA, TMSF, and unified PSH, to bridge both the modality and task gaps in MRSIS, thus improving the performance of multiple segmentation tasks. Notably, the optical and SAR images serve merely as examples and can be extended to any modality of data.

To address the diverse demands of MRSIS, recent approaches have explored multitask learning (MTL) methods for MRSIS [3], [7]. For instance, Liu et al. [7] proposed an efficient Transformer-based network for multiple RS tasks through cross-attention learning of shared representations between RGB and DSM images. Zhou et al. [3] proposed a multitask perceptual network (MTANet) that incorporates hierarchical multimodal fusion for RGB-T urban scene understanding, and they employed MTL to optimize the parameters of the MTANet. However, current MTL methods typically employ separate task heads to process fused multimodal features for each task, without adequately exploring the relationships between different tasks. As shown in Fig. 1(a), the MTL approach often overlooks the *modality gap* and *task gap*, leading to suboptimal performance. More specifically, we identify these gaps as follows.

1) *Modality gap* arises from both semantic discrepancies and granularity discrepancies across various RS data types. Different modalities capture distinct semantic features and operate at different granularities. For instance, RGB images emphasize color, texture, and shape at high resolutions, while DSM images focus on terrain height at lower resolutions [8]. Direct fusion may not fully leverage the strengths of each modality, leading to information loss. These discrepancies exacerbate the modality gap, complicating effective data integration.
2) *Task gap* arises from the differing requirements across segmentation tasks, including variations in annotation standards, task complexity, and objectives. These differences lead to modality preferences for various tasks. For instance, instance segmentation labels demand precise object boundaries, favoring DSM images for their height information, while semantic segmentation labels emphasize region classification, making optical images more suitable for their rich color and texture [1].

These gaps—both modality and task—lead to the prioritization of different RS modalities by each segmentation task. Unfortunately, many current MTL methods fail to effectively address these gaps, limiting their ability to capture task-specific requirements and impeding generalization across tasks. This results in suboptimal performance for complex MRSIS tasks.

To address both the modality and task gaps in MRSIS, we propose the prompt-driven multimodal Transformer for task-adapted RS image segmentation (PMTSeg), which ensures robust and effective performance. As shown in Fig. 1(b), PMTSeg integrates three key components: task-common multimodal affinity approximation (TMAA), task-common multiscale semantic fusion (TMSF), and the prompt-driven segmentation head (PSH). Specifically, *TMAA* addresses the semantic discrepancies in the modality gap by approximating intermodal affinity matrices using Jensen–Shannon (JS) divergence [9], generating semantic-aligned multimodal features. Then, *TMSF* uses the scale-matched fusion (SF) to integrate these features at multiple scales, mitigating the granularity discrepancies in the modality gap and creating multiscale task-common features enriched with complementary information. Moreover, instead of multiple task heads, *PSH* employs a unified task-adapted text prompting with contrastive loss to capture the relationship between different tasks, bridging the task gap. The integration of these three components within a unified training framework enhances performance across multiple segmentation tasks by effectively addressing both the modality and task gaps. Our main innovations and contributions of this study are summarized as follows.

1) *PMTSeg* is the first unified multimodal approach for task-adapted MRSIS, capable of addressing semantic, instance, and panoptic segmentation tasks in a single model. To the best of our knowledge, this is the first effective method for handling all three MRSIS tasks simultaneously.
2) To mitigate the *modality gap*, we introduce the *TMAA* module to approximate intermodal affinity matrices via JS divergence, generating semantic-aligned multimodal features. We then propose *TMSF*, which integrates these features using the SF at multiple scales to create multiscale task-common features that contain complementary information from various modalities. This approach enables efficient feature learning across multiple tasks by addressing the modality gap.
3) To bridge the *task gap*, we propose a unified *PSH* that adapts to various segmentation tasks. In contrast to traditional MTL, which uses multiple task heads, we introduce task-adapted text prompts and a task-adapted contrastive loss. This enables PMTSeg to dynamically capture intertask relationships.
4) We conducted extensive experiments on three publicly available MRSIS datasets (VALID, SEMCITY TOULOUSE, and UBCV2). The results demonstrate that *PMTSeg* outperforms existing methods across all three segmentation tasks. Specifically, we achieve an average relative improvement of 8.12% for mean

intersection over union (mIoU) on the semantic segmentation task, 12.78% for mean average precision (mAP) on the instance segmentation task, and 23.39% for panoptic quality (PQ) on the panoptic segmentation task, confirming the effectiveness and generalizability of our method for MRSIS.

II. Related Work

A. RS Image Segmentation

Image segmentation is a pivotal task in image interpretation, where substantial advancements have been achieved. RS data present more complex scenes compared to natural scenes. Image segmentation mainly includes semantic, instance, and panoptic segmentation tasks. Semantic segmentation involves assigning a semantic category to each pixel, facilitating pixel-level semantic understanding. Niu et al. [10] introduced a pyramid semantic segmentation framework with decoupled learning to explicitly model foreground and boundary objects in RS images. Instance segmentation, on the other hand, identifies and delineates individual objects, providing unique identifiers for accurate localization. Wei et al. [11] proposed the concentric loop convolutional neural network (CLP-CNN), specifically designed for instance segmentation in RS, to precisely segment and distinguish objects in complex scenes. Panoptic segmentation combines both pixel-level classification and instance identification, offering a holistic view of the scene. Garnot and Landrieu [12] developed an end-to-end, single-stage method for panoptic segmentation of satellite image time series, extracting adaptive multiscale spatiotemporal features to improve segmented scene interpretation over time. Recently, prompt-based methods have also been introduced to enhance segmentation performance in low-data regimes or specific tasks. Bi et al. [13] proposed a PAT approach that leverages category-aware dynamic enhancement to improve few-shot segmentation. While effective, their method relies heavily on category-specific prompts and lacks general adaptability across diverse tasks. Similarly, Shang et al. [14] introduced a Prompt-RIS strategy, which enhances instance segmentation via instance-level contrast. However, their design primarily focuses on single-task optimization and does not consider task transferability. Despite these advances, existing segmentation methods are often tailored for specific RS tasks, limiting their flexibility. Moreover, relying solely on unimodal data restricts the semantic richness necessary for fine-grained classification, hindering overall segmentation performance.

B. Multimodal RS Image Segmentation

Unimodal RS data often fall short in capturing the comprehensive information required for accurate image interpretation. Consequently, there is a growing focus on leveraging multimodal data to enhance RS analysis. MRSIS integrates data from multiple sensors to assign semantic or instance labels to each pixel, improving the accuracy and interpretability of RS data. Liu et al. [15] introduced a pyramid-style context-guided fusion module for semantic segmentation, utilizing complementary information from RGB and depth features to address semantic inconsistencies across modalities. Sharma et al. [16] proposed a framework for instance extraction in multimodal RS images, integrating different modalities to harness complementary information. Zhang et al. [1] combined RGB and NIR data using auxiliary super-resolution (SR) learning, enabling high-resolution instance extraction of multiscale targets while balancing detection accuracy and computational efficiency. Garnot et al. [17] leveraged optical and radar time series, employing spatial and temporal encoders in a unified multimodal sequence to extract robust spatiotemporal features for panoptic segmentation. In spite of advancements, these methods remain tailored to specific RS segmentation tasks rather than being universally applicable across diverse segmentation scenarios.

C. Multimodal Multitask RS Image Segmentation

To meet the complex demands of real-world scenarios, an increasing number of researchers have explored multimodal multitask methods in RS image segmentation. The objective of multimodal multitask image segmentation is to simultaneously address multiple related tasks from different modalities to achieve various segmentation outcomes. Wang et al. [18] proposed an adaptive channel exchange network (CEN) to accomplish multiple dense image prediction tasks, utilizing both optical and depth images. Liu et al. [7] employed optical and DSM images to simultaneously perform semantic and height change detection, modeling multimodal relationships using consistency constraints. Cheng et al. [19] proposed a multitask, multisource information fusion method for semantic labeling in RS, utilizing data from multiple sources and intertask dependencies for co-learning to enhance performance. Despite the progress made, these methods mainly focus on some specific tasks within RS scenes and lack a comprehensive approach to handle multiple segmentation tasks. Furthermore, while these methods utilize multiple modalities to obtain diverse task results, they seldom address the combined impact of modality gap and task gap on multimodal multitask methods.

III. Proposed Approach

A. Problem Definition and Overview

We first define the problem of the MRSIS task. The training set is denoted as $D_{\text{train}} = \{(x_r^i, x_s^i, y_{\text{sem}}^i, y_{\text{ins}}^i, y_{\text{pan}}^i)\}$, where i indexes over samples, x_r and x_s represent RS images of different modalities (such as the optical image and the synthetic aperture radar (SAR) image), y_{sem}^i represents the corresponding semantic label, y_{ins}^i represents the corresponding instance label, and y_{pan}^i represents the corresponding panoptic label. The test set is defined as $D_{\text{test}} = \{(x_r^j, x_s^j)\}$, where j indexes over test samples.

For MRSIS, we propose a novel PMTSeg. Our PMTSeg unifies semantic, instance, and panoptic segmentation tasks, leveraging one unified PSH, with great generalization and adaptation. Fig. 2 illustrates the general framework of the PMTSeg, which comprises three main modules: the TMAA, the TMSF, and the PSH. First, taking the multimodal RS data, for example, the optical image and SAR image, as the input,

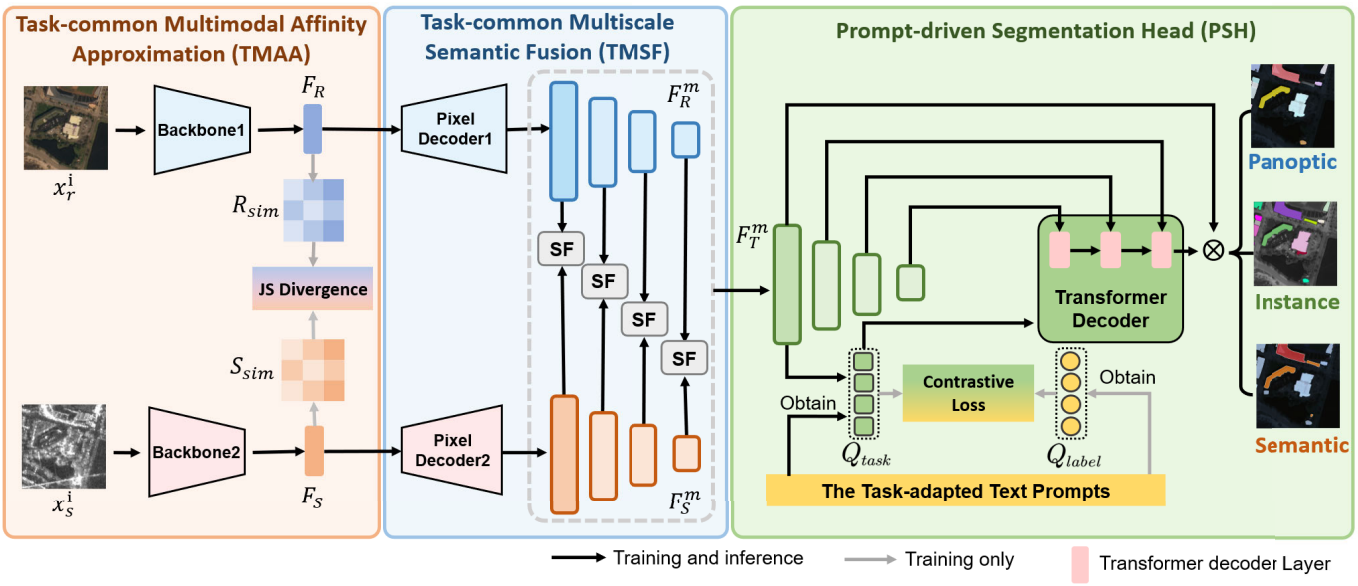

Fig. 2. Overall pipeline of our proposed PMTSeg. Given multimodal RS data, such as optical and SAR images, as input, the TMAA module first generates semantic-aligned multimodal representations by approximating intermodal affinity matrices. Next, the TMSF module extracts the multiscale task-common features, capturing relevant semantic information across modalities through the SF at multiple scales. Finally, the PSH, guided by the task-adapted text prompts, enables unified training to deliver various segmentation task outcomes.

in TMAA, we use the backbones to extract different unimodal features, and then the intermodal affinity matrices generated from these features are approximated leveraging JS divergence, thus obtaining the semantic-aligned multimodal features. Next, TMSF further integrates these aligned features by introducing the SF to explore valid semantic information from various modalities in both low- and high-resolution features, yielding the multiscale task-common features enriched with complementary information. Finally, the multiscale task-common features are refined by the PSH, which uses task-adapted text prompts and task-adapted contrastive loss to generate a query that captures intertask relationships. The Transformer decoder then leverages this query and the semantic features to produce the segmentation results.

By combining these three modules for universal training, PMTSeg efficiently alleviates the modality gap and task gap, enabling the performance of panoptic, semantic, and instance segmentation with high robustness and effectiveness.

B. TMAA Module

The modality gap across various RS data brings the semantic discrepancies, thus making cross-modal fusion and analysis more challenging [20]. Therefore, to mitigate the semantic discrepancies and extract task-common features across modalities, we introduce the TMAA, which captures consistent affinity relationships between different modalities to align semantic information, obtaining the semantic-aligned multimodal features.

Formally, given the optical image x_r^i and the SAR image x_s^i, the backbones (Backbone1 and Backbone2) are used to extract the optical feature F_R and the SAR feature F_S, respectively, which can be written as

$$F_R = \text{Backbone1}\left(x_r^i\right),\ F_S = \text{Backbone2}\left(x_s^i\right). \tag{1}$$

To capture the intermodal affinity relationship, we compute the intermodal normalized cosine similarity leveraging F_R and F_S. Mathematically, we define the optical affinity matrix R_{sim} and the SAR affinity matrix S_{sim} as

$$R_{\text{sim}} = \frac{r_{n1} \cdot r_{n2}}{\|r_{n1}\|_2^2 \cdot \|r_{n2}\|_2^2}, \quad \{r_{n1}, r_{n2}\} \in F_R \tag{2}$$

$$S_{\text{sim}} = \frac{s_{n1} \cdot s_{n2}}{\|s_{n1}\|_2^2 \cdot \|s_{n2}\|_2^2}, \quad \{s_{n1}, s_{n2}\} \in F_S \tag{3}$$

where r_{n1} and r_{n2} denote the different feature points of F_R, and s_{n1} and s_{n2} denote the different feature points of F_S. To obtain the semantic-aligned multimodal features, we design an affinity-based distribution loss L_{TMAA}, which utilizes the JS divergence [9] to approximate the affinity distributions across modalities. Specifically, R_{sim} and S_{sim} are first converted into probability distributions, followed by computing their JS divergence to derive L_{TMAA}. This process can be formally expressed as

$$R_{\text{dis}} = \frac{\exp\left(\frac{r_k'}{\tau}\right)}{\sum_{k=1}^{\text{HW}} \exp\left(\frac{r_k'}{\tau}\right)}, \quad r_k' \in R_{\text{sim}} \tag{4}$$

$$S_{\text{dis}} = \frac{\exp\left(\frac{s_k'}{\tau}\right)}{\sum_{k=1}^{\text{HW}} \exp\left(\frac{s_k'}{\tau}\right)}, \quad s_k' \in S_{\text{sim}} \tag{5}$$

$$L_{\text{TMAA}} = \text{JS}(R_{\text{dis}}, S_{\text{dis}}) \tag{6}$$

where HW represents the total number of R_{sim} or S_{sim}, r_k' denotes the feature point of R_{sim}, and s_k' denotes the feature point of S_{sim}.

Leveraging the TMAA, the semantic discrepancies in the modality gap between different RS modalities are alleviated through the affinity-based distribution constraint, producing the semantic-aligned optical feature F_R and the semantic-aligned SAR feature F_S.

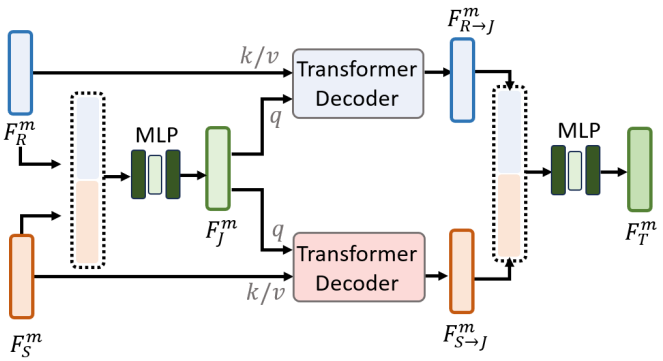

Fig. 3. Architecture of the mth SF in TMSF. The two Transformer decoders aggregate the semantic information of the joint feature (F_J^m) and different modality features (F_R^m and F_S^m), generating the m-scale task-common feature F_T^m.

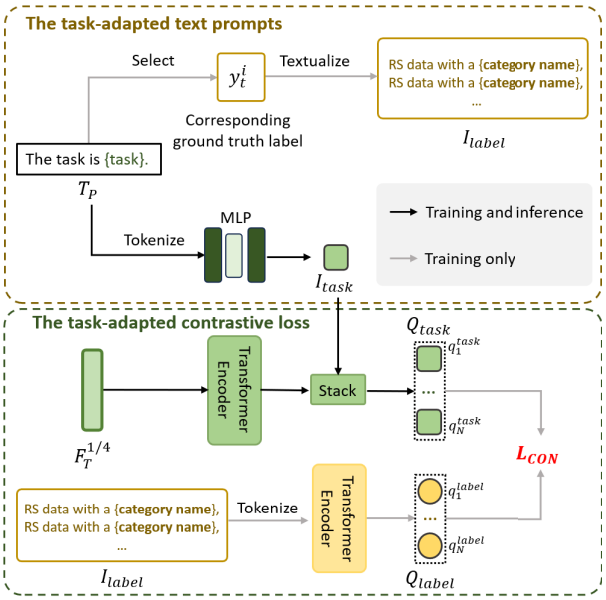

Fig. 4. Details of the task-adapted text prompts and the task-adapted contrastive loss. The task-adapted text prompts derive task-level information I_{task} and label-level information I_{label} from T_p. The task-adapted contrastive loss is applied to capture the relationship between different tasks, utilizing the task-level object query Q_{task} and the label-level object query Q_{label}. {task}∈{semantic, instance, panoptic}.

C. TMSF Module

To further mitigate the granularity discrepancies within the modality gap, with the obtained F_R and F_S, we introduce TMSF with the SF to aggregate semantic-aligned information at multiple scales, maximizing the strengths of each modality to obtain the multiscale task-common features.

Specifically, to capture different granularities across various modalities, we first feed F_R and F_S into the pixel encoders that are denoted as in [21], generating the multiscale optical feature $\{F_R^m,\ m = 1/4, 1/8, 1/16, 1/32\}$ and the multiscale SAR feature $\{F_S^m,\ m = 1/4, 1/8, 1/16, 1/32\}$.

Then, to further integrate valid semantic information at different granularities across various modalities, we employ the SF at each scale. The SF in m-scale is shown in Fig. 3. For the m-scale optical feature F_R^m and the m-scale optical feature F_S^m, we first concentrate the F_R^m and F_S^m to get the m-scale joint feature F_J^m through a two-layer MLP, which can be written as

$$F_J^m = \text{MLP}\big(\big[F_R^m; F_S^m\big]\big). \tag{7}$$

Next, to capture the task-common semantic information based on the intra- and intermodal relationships, we feed F_J^m to the Transformer decoder for different modalities. For the Transformer decoder of optical image, F_J^m serves as the query, while F_R^m is the key/value, producing the optical-enhanced feature $F_{R\to J}^m$. Similarly, for the Transformer decoder of SAR image, F_J^m is used as the query, with F_S^m as key/value, to generate the SAR-enhanced feature $F_{S\to J}^m$. Following the typical formulas for Transformer structures, $Trans(\cdot)$, $F_{R\to J}^m$ and $F_{S\to J}^m$ can be expressed as

$$F_{R\to J}^m = \text{Trans}\big(q = F_J^m, k/v = F_R^m\big) \tag{8}$$

$$F_{S\to J}^m = \text{Trans}\big(q = F_J^m, k/v = F_S^m\big). \tag{9}$$

Finally, $F_{S\to J}^m$ and $F_{R\to J}^m$ are concatenated and refined through a two-layer MLP to produce the m-scale task-common features F_T^m, which can be expressed as

$$F_T^m = \text{MLP}\big(\big[F_{R\to J}^m; F_{S\to J}^m\big]\big). \tag{10}$$

Through the TMSF, we aggregate valid semantic information from different granularities across modalities to obtain the multiscale task-common features $\{F_T^m,\ m = 1/4, 1/8, 1/16, 1/32\}$, which encode comprehensive segmentation information that further mitigates the granularity discrepancies in the modality gap.

D. PSH Module

After obtaining the multiscale task-common features $\{F_T^m,\ m = 1/4, 1/8, 1/16, 1/32\}$, traditional multitask methods usually use separate heads for various tasks. However, the varying modality preferences across tasks introduce a task gap, which can lead to suboptimal performance. To address the task gap, we propose a novel unified PSH that leverages task-adapted text prompts and a task-adapted contrastive loss. This approach captures the relationships between different tasks and alleviates the modality preference issues across tasks. By doing so, the PSH enables effective semantic, instance, and panoptic segmentation simultaneously. The detailed architecture of the PSH is illustrated in Fig. 4.

1) Task-Adapted Text Prompts: We first define the task-adapted text prompts as T_p, to indicate the segmentation task being performed. We randomly assign a prompt from a predefined textual set based on a randomly generated number $r \in [0, 1]$, as follows:

$$T_p = \begin{cases} \text{'The task is semantic.'} & \text{if } r < 0.3 \\ \text{'The task is panoptic.'} & \text{if } r > 0.6 \\ \text{'The task is instance.'} & \text{Otherwise.} \end{cases} \tag{11}$$

Once the prompt T_p is selected, we extract task-level and label-level information from it. As illustrated in Fig. 4, we process the task-level information and label-level information separately. For the *task-level information*, we use a two-layer MLP to transform the prompt T_p into a task feature I_{task}.

For the *label-level information*, by accessing the corresponding ground-truth labels y_t^i, we obtain the label-level information I_{label} in a structured text format, such as "RS data with a *category name.*," which helps to encode fine-grained task-specific details.

2) Task-Adapted Contrastive Loss: To further capture the relationship between various tasks, we introduce task-adapted contrastive learning. Specifically, with the high-resolution task-common feature $F_T^{1/4}$, which retains rich detailed information, we first utilize a two-layer Transformer encoder to extract high-level semantic features. These features are then concatenated with the task-level information I_{task} to produce the task-level object query $Q_{\text{task}} = \{q_1^{\text{task}}, q_2^{\text{task}}, \ldots, q_N^{\text{task}}\}$, where N denotes the length of the query. The initialization of Q_{task} is crucial for enabling PMTSeg to learn multiple tasks effectively, as it encodes richer semantic content specific to each task. Meanwhile, we use another two-layer Transformer encoder to generate a label-level object query to generate a label-level object query $Q_{\text{label}} = \{q_1^{\text{label}}, q_2^{\text{label}}, \ldots, q_N^{\text{label}}\}$, which captures the specific label content for the task. By aligning the task-level object query Q_{task} and label-level object query Q_{label}, we can effectively model the intertask relationships. To achieve this alignment, we define the task-adapted contrastive loss L_{CON} as

$$L_{Q_{\text{task}} \to Q_{\text{label}}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(q_i^{\text{task}} \odot q_i^{\text{label}} / \tau\right)}{\sum_{j=1}^{N} \exp\left(q_i^{\text{task}} \odot q_j^{\text{label}} / \tau\right)} \tag{12}$$

$$L_{Q_{\text{label}} \to Q_{\text{task}}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(q_i^{\text{label}} \odot q_i^{\text{task}} / \tau\right)}{\sum_{j=1}^{N} \exp\left(q_i^{\text{label}} \odot q_j^{\text{task}} / \tau\right)} \tag{13}$$

$$L_{\text{CON}} = L_{Q_{\text{task}} \to Q_{\text{label}}} + L_{Q_{\text{label}} \to Q_{\text{task}}} \tag{14}$$

where i and j are the indices of Q_{task} and Q_{label}, respectively, and τ denotes the temperature parameter. During training, L_{CON} encourages a closer match between Q_{task} and Q_{label}, further improving task consistency and the effectiveness of MTL.

3) Segmentation Head: Instead of the traditional approach of using multiple heads for different tasks in MTL, we propose using a single three-layer Transformer decoder as the final segmentation head to produce three segmentation results. Specifically, we take Q_{task} as the query, and the Transformer decoder progressively attends to multiscale task-common features $\{F_T^m, m = 1/4, 1/8, 1/16, 1/32\}$ as the key/value at different layers. This enables the model to generate precise segmentation outcomes by effectively integrating semantic information across multiple granularities. To optimize this unified segmentation head, we introduce the prompt-driven segmentation loss L_{PSH}, which combines three essential loss components: the standard categorical cross-entropy loss (L_{cls}), the binary cross-entropy loss (L_{bce}), and the dice loss (L_{dice}) [21]. This loss function can be written as

$$L_{\text{PSH}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{bce}} L_{\text{bce}} + \lambda_{\text{dice}} L_{\text{dice}} \tag{15}$$

where we empirically set the weights of each loss as $\lambda_{\text{cls}} = 2$, $\lambda_{\text{bce}} = 5$, and $\lambda_{\text{dice}} = 5$ according to [21].

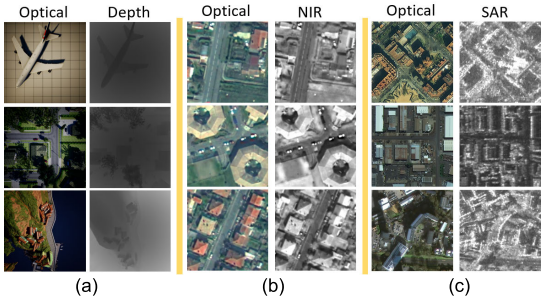

Fig. 5. Sampled image pairs from the (a) VALID dataset, (b) SEMCITY TOULOUSE dataset, and (c) UBCV2 dataset.

E. Overall Objective Function

The proposed model, incorporating the three modules, is trained sequentially in an end-to-end manner. The overall objective function consists of three primary components: 1) the affinity-based distribution loss L_{TMAA}; 2) the task-adapted contrastive loss L_{CON}; and 3) the prompt-driven segmentation loss L_{PSH}. Formally, the total optimization objective for training is defined as

$$L = \alpha L_{\text{TMAA}} + \beta L_{\text{CON}} + L_{\text{PSH}} \tag{16}$$

where α and β are the weighting factors of losses to balance the training and are set as 5 and 0.5, respectively, in this study. Detailed discussion about these factors can be seen in Section IV-E.

IV. Experiments and Analysis

A. Datasets

To evaluate the proposed method, we conducted experiments on three publicly available MRSIS datasets: VALID [22], SEMCITY TOULOUSE [23], and UBCV2 [24]. Examples from each dataset are presented in Fig. 5.

1) VALID Dataset: The VALID dataset [22] is a virtual aerial image dataset that provides panoptic segmentation annotation. It contains 6690 paired optical and depth images captured across six virtual environments under five different conditions: sunlight, dusk, night, snow, and fog. Each image features pixel-level panoptic annotations and was taken at three altitudes (20, 50, and 100 m) with a resolution of 1024 × 1024. The dataset includes 30 categories, divided into 17 thing categories and 13 stuff categories. For experimentation, the dataset is split into 3345 image pairs for training, 1115 for validation, and 2230 for testing.

2) SEMCITY TOULOUSE Dataset: The SEMCITY TOULOUSE dataset [23] is a high-resolution multispectral dataset covering 50 km^2 of the Toulouse city center in France. It includes instance segmentation annotations for one category and semantic segmentation annotations for six categories. The dataset comprises 16 multispectral images with a resolution of 0.5 m, each sized 876 × 876. Given that only four images contain building instance annotations, we segmented these four images into smaller patches of 128 × 128 and extracted the corresponding optical and NIR images. Additionally, we combined the instance and semantic annotations to create

panoptic annotations. Following the official split for building instances, the dataset was divided into 1404 pairs for training and 1404 pairs for testing.

3) UBCV2 Dataset: The UBCV2 dataset [24] is a diverse collection sourced from 20 cities around the world, representing a wide range of landforms and architectural styles. It includes multimodal data—optical and SAR images—available for 17 of the cities. UBCV2 features 12 roof categories, including flat roofs, hillside roofs, gable roofs, row roofs, multieave roofs, quadruple slope roofs (v1 and v2), pitched roofs, pyramid roofs, gabled roofs, rotating roofs, and other types. The dataset contains 7314 image pairs, each with a resolution of 512 × 512 at a 0.5-m spatial scale. These are split into 4008 pairs for training, 1681 for validation, and 1625 for testing. We also consolidated the annotations for the 12 roof categories and the background into panoptic annotations using the multimodal subset of UBCV2.

B. Evaluation Protocol

To comprehensively assess the proposed method, we employ a range of widely used evaluation metrics tailored to segmentation tasks. For panoptic segmentation, we report three metrics: PQ, segmentation quality (SQ), and recognition quality (RQ) [20]. RQ evaluates detection accuracy and is defined as

$$\mathrm{RQ} = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|} \tag{17}$$

where TP represents true positive, FP represents false positive, and FN represents false negative. SQ measures segmentation accuracy, calculated as the mean IoU of correctly matched segments, which can be written as follows:

$$\mathrm{SQ} = \frac{\sum_{(i,j)\in \mathrm{TP}} \mathrm{IoU}(i,j)}{|\mathrm{TP}|} \tag{18}$$

where i and j denote the segmentation detected in true positive. PQ, the primary metric for panoptic segmentation, integrates segmentation and detection quality

$$\mathrm{PQ} = \mathrm{SQ} \times \mathrm{RQ}. \tag{19}$$

For the instance segmentation, we utilize three standard metrics: mAP, AP@0.5, and AP@0.75 [25]. AP@0.5 and AP@0.75 indicate performance at IoU thresholds of 0.5 and 0.75, respectively. IoU [26], which quantifies the overlap between predicted and true segments, is defined as

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{20}$$

The mAP metric computes the average precision across a range of IoU thresholds, offering a holistic evaluation of accuracy

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{21}$$

For the semantic segmentation, we employ mIoU and overall accuracy (OA) [1]. mIoU calculates the average IoU across all categories, while OA measures pixel-level accuracy as the ratio of correctly predicted pixels to the total number of pixels

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{22}$$

where TN indicates true negative. These metrics provide a detailed evaluation of the proposed method's performance across various segmentation tasks.

C. Experimental Settings and Implementation Details

The experiments were conducted on a Linux platform using NVIDIA GeForce RTX 3090 and RTX 2080 Ti GPUs, with implementation in the PyTorch framework. We utilized Detectron2 [27] with a training configuration based on [21], leveraging an updated Mask R-CNN with the AdamW optimizer and uniform-interval learning rate scheduling. During training, the initial learning rate was set to 0.0001, with a weight decay of 0.05. The model was trained for 300 epochs with a batch size of 4. To manage GPU memory constraints, images from the VALID dataset were resized to 512 × 512, while SEMCITY TOULOUSE and UBCV2 datasets retained their original resolutions. Data augmentation included large-scale dithering with random sampling parameters ranging from 0.1 to 2.0, followed by cropping to maintain the original input size. During inference, we used the standard Mask R-CNN configuration. To ensure fair and consistent evaluations, all experiments were performed under standardized conditions.

The backbones (Backbone1 and Backbone2) were based on the ResNet-50 architecture [25], pretrained on the ImageNet dataset [21]. For the pixel decoders (PixelDecoder1 and PixelDecoder2), we employed the architecture based on multiscale deformable transformers (MSDeformAttn) [28]. In addition, we defined the Transformer encoder and the Transformer decoder following the structure of a conventional Transformer [29].

D. Analysis of Experimental Results

Our experiments encompass three tasks: semantic (SEM), instance (INS), and panoptic (PAN) segmentation. We compared our PMTSeg with several recent MRSIS methods, including NNCNet [30], CAFE [31], CMFNet [32], FTransUNet [33], SuperYolo [1], ICAFusion [34], RGBDInst [35], CalibNet [36], and others, which primarily focus on semantic and instance segmentation. Given the limited availability of panoptic segmentation methods for MRSIS, we applied multimodal data to unimodal panoptic segmentation methods (e.g., Mask2Former-EF [21], YOSO-EF [37], CoMFormer-EF [20], OneFormer-EF [28], etc.), using the early fusion (EF) strategy to obtain panoptic results. Additionally, we compared several unimodal segmentation methods, such as Mask2Former [21], OneFormer [28], Mask R-CNN [25], QueryInst [26], SparseInst [38], and others. For all experiments, optical images were used as the input for unimodal methods, and the default configurations of each method were employed for both training and testing.

1) Experiments on the VALID Dataset: We conducted comprehensive experiments on the VALID test set to evaluate the performance of our PMTSeg and compare it with existing state-of-the-art segmentation methods. Detailed comparisons were performed across three tasks, that is, semantic, instance, and panoptic segmentation. The experimental results

TABLE I

COMPARISON RESULTS ON THE VALID DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Task	Modality	Semantic Metrics		
			mIoU	mACC	OA
FCN [39]	SEM	Optical	21.6	-	74.0
PSPNet [40]	SEM	Optical	53.8	-	92.7
OneFormer [28]	SEM	Optical	68.7	78.0	96.4
OTSeg [41]	SEM	Optical	60.5	69.5	95.5
NNCNet [30]	SEM	Optical+Depth	72.1	78.2	95.8
CAFE [31]	SEM	Optical+Depth	71.2	81.7	96.0
CMFNet [32]	SEM	Optical+Depth	74.4	80.5	96.3
FTransUNet [33]	SEM	Optical+Depth	76.1	82.2	96.4
Masks2Former-EF (baseline) [21]	SEM	Optical+Depth	77.5	83.7	96.9
Ours	**SEM**	**Optical+Depth**	**82.4**	**87.0**	**97.7**
Method	Task	Modality	Instance Metrics		
			AP	AP0.5	AP0.75
Mask R-CNN [25]	INS	Optical	28.9	52.5	28.4
QueryInst [26]	INS	Optical	36.7	54.9	41.5
SparseInst [38]	INS	Optical	34.6	53.5	37.0
OneFormer [28]	INS	Optical	39.7	53.7	43.8
SAPNet [42]	INS	Optical	35.3	48.4	37.1
SuperYolo [1]	INS	Optical+Depth	40.3	58.1	44.1
ICAFusion [34]	INS	Optical+Depth	41.1	58.6	44.7
RGBDInst [35]	INS	Optical+Depth	43.8	55.4	46.9
CalibNet [36]	INS	Optical+Depth	40.9	55.2	44.6
Mask2Former-EF (baseline) [21]	INS	Optical+Depth	43.2	56.8	47.4
Ours	**INS**	**Optical+Depth**	**51.9**	**63.8**	**54.8**
Method	Task	Modality	Panoptic Metrics		
			PQ	SQ	RQ
Mask2Forme [21]	PAN	Optical	38.7	70.8	48.8
OneFormer [28]	PAN	Optical	38.8	68.8	48.9
YOSO-EF [37]	PAN	Optical+Depth	35.8	70.0	45.1
CoMFormer-EF [20]	PAN	Optical+Depth	36.8	68.3	46.4
OneFormer-EF [28]	PAN	Optical+Depth	38.1	70.5	48.1
Mask2Former-EF (baseline) [21]	PAN	Optical+Depth	38.5	70.9	48.6
Ours	**PAN**	**Optical+Depth**	**43.4**	**72.8**	**54.1**

TABLE II

COMPARISON RESULTS ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Task	Modality	Semantic Metrics		
			mIoU	mACC	OA
Mask2Former [21]	SEM	Optical	56.9	72.2	76.4
OneFormer [28]	SEM	Optical	57.6	71.1	76.7
OTSeg [41]	SEM	Optical	60.1	77.0	75.9
NNCNet [30]	SEM	Optical+NIR	57.3	73.9	74.0
CAFE [31]	SEM	Optical+NIR	57.5	72.8	73.7
CMFNet [32]	SEM	Optical+NIR	59.7	74.8	74.7
FTransUNet [33]	SEM	Optical+NIR	54.4	70.3	73.1
Mask2Former-EF (baseline) [21]	SEM	Optical+NIR	59.1	73.2	75.8
Ours	**SEM**	**Optical+NIR**	**65.0**	**78.5**	**78.2**
Method	Task	Modality	Instance Metrics		
			AP	AP0.5	AP0.75
Mask R-CNN [25]	INS	Optical	34.3	64.7	33.3
QueryInst [26]	INS	Optical	38.0	68.2	39.1
SparseInst [38]	INS	Optical	34.7	66.9	33.4
OneFormer [28]	INS	Optical	37.0	69.2	36.3
SAPNet [42]	INS	Optical	37.2	65.9	37.9
SuperYolo [1]	INS	Optical+NIR	38.4	69.7	38.8
ICAFusion [34]	INS	Optical+NIR	39.4	70.0	39.1
RGBDInst [35]	INS	Optical+NIR	38.7	69.9	39.7
CalibNet [36]	INS	Optical+NIR	39.3	68.8	40.0
Mask2Former-EF (baseline) [21]	INS	Optical+NIR	37.1	67.0	36.9
Ours	**INS**	**Optical+NIR**	**40.3**	**70.3**	**41.4**
Method	Task	Modality	Panoptic Metrics		
			PQ	SQ	RQ
Mask2Forme [21]	PAN	Optical	43.0	71.5	59.7
OneFormer [28]	PAN	Optical	43.6	69.7	61.0
YOSO-EF [37]	PAN	Optical+NIR	32.9	69.7	46.5
CoMFormer-EF [20]	PAN	Optical+NIR	33.2	69.4	47.8
OneFormer-EF [28]	PAN	Optical+NIR	42.8	70.4	59.6
Mask2Former-EF (baseline) [21]	PAN	Optical+NIR	34.9	70.6	49.0
Ours	**PAN**	**Optical+NIR**	**46.7**	**72.4**	**64.2**

are shown in Table I. From the results, our proposed PMTSeg achieved the best performance in all segmentation tasks. For the SEM task, PMTSeg demonstrated superior performance. Our PMTSeg achieved 82.4% mIoU, representing a 6.3% relative improvement over the baseline. This demonstrates that our PMTSeg effectively captures semantic information across multiple modalities by leveraging the SF. For the INS task, compared with the existing state-of-the-art method, RGBInst, the proposed PMTSeg also achieved a relative improvement of 15.2% on AP@0.5, 16.9% on AP@0.75, and 18.5% on mAP. For the PAN task, our PMTSeg led all evaluation metrics. Compared to the current optimal method, OneFormer, the proposed model relatively improved PQ and SQ by 11.9% and 5.8%, respectively. Interestingly, the results showed that unimodal methods performed better than some multimodal methods for panoptic segmentation, which may be attributed to the EF strategy not being optimal for capturing multimodal semantic information. In addition, our method achieves significantly higher scores in both semantic and instance segmentation compared to OTSeg and SAPNet. While OTSeg leverages multiprompt attention for semantic tasks, its single-modality input limits performance in complex scenes. Similarly, SAPNet's point-prompted instance segmentation shows limitations in instance-level precision. In contrast, our approach benefits from multimodal input and task-specific optimization, leading to consistent improvements across all metrics. Overall, these findings confirm that PMTSeg offers superior generalization performance across all tasks.

2) Experiments on the SEMCITY TOULOUSE Dataset: We performed experiments on the SEMCITY TOULOUSE test set to compare our proposed PMTSeg with existing state-of-the-art methods. The experimental results are shown in Table II. From the results, our proposed PMTSeg achieved the best performance in all segmentation tasks. For the SEM task, compared with the baseline, the PMTSeg achieved a relative improvement of 10.0% on mIoU, 7.2% on Macc, and 3.2% on OA. This shows that the affinity-based constraint can effectively mitigate the modality gap, thus improving the performance. For the INS task, the proposed PMTSeg achieved a relative improvement of 2.3%/0.4%/5.9% over the current optimal method ICAFusion at mAP/AP@0.5/AP@0.75, respectively. It indicates that our PMTSeg is more suitable for extracting the exact contours of the instances, leveraging the multiscale features. In addition, for the PAN task, we observe a relative increase of 33.8%, 2.5%, and 31.0% in SQ, RQ, and PQ for the baseline, respectively. Moreover, our PMTSeg continues to outperform OTSeg and SAPNet. Both OTSeg and SAPNet rely on prompt-based strategies for guiding segmentation tasks, which is why we chose them for comparison. However, our PMTSeg integrates task-specific and label information into the prompt mechanism. This allows PMTSeg to better adapt to various tasks, leading to superior performance across multiple segmentation challenges.

3) Experiments on the UBCV2 Dataset: Experiments on the UBCV2 dataset also demonstrate results similar to those derived from the VALID and SEMCITY TOULOUSE datasets. As presented in Table III, our proposed PMTSeg outperformed other methods on all segmentation tasks. For the SEM task, the proposed PMTSeg outperformed other methods on all evaluation indicators, obtaining the best mIoU of 37.5%, the best mACC of 47.5%, and the best OA of 92.7%.

TABLE III

COMPARISON RESULTS ON THE UBCV2 DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Task	Modality	Semantic Metrics		
			mIoU	mACC	OA
Mask2Former [21]	SEM	Optical	34.4	41.9	91.9
OneFormer [28]	SEM	Optical	35.5	45.8	91.9
OTSeg [41]	SEM	Optical	34.1	44.4	91.3
NNCNet [30]	SEM	Optical+SAR	26.1	44.7	90.8
CAFE [31]	SEM	Optical+SAR	29.4	45.6	91.1
CMFNet [32]	SEM	Optical+SAR	25.7	44.3	90.7
FTransUNet [33]	SEM	Optical+SAR	26.5	45.9	91.0
Mask2Former-EF (baseline) [21]	SEM	Optical+SAR	34.7	45.1	91.4
Ours	**SEM**	**Optical+SAR**	**37.5**	**47.5**	**92.7**
Method	Task	Modality	Instance Metrics		
			AP	AP0.5	AP0.75
Mask R-CNN [25]	INS	Optical	13.0	22.9	-
Mask2Former [21]	INS	Optical	13.8	22.0	15.2
QueryInst [26]	INS	Optical	11.9	19.3	12.9
SparseInst [38]	INS	Optical	10.6	18.0	11.6
OneFormer [28]	INS	Optical	14.1	23.3	15.0
SAPNet [42]	INS	Optical	13.8	24.5	14.6
Mask R-CNN [25]	INS	Optical+SAR	7.9	15.5	-
SuperYolo [1]	INS	Optical+SAR	10.2	17.0	10.3
ICAFusion [34]	INS	Optical+SAR	14.4	24.0	15.4
RGBDInst [35]	INS	Optical+SAR	8.8	14.5	9.5
CalibNet [36]	INS	Optical+SAR	8.0	15.6	7.8
Mask2Former-EF (baseline) [21]	INS	Optical+SAR	7.3	12.5	7.5
Ours	**INS**	**Optical+SAR**	**16.9**	**27.9**	**18.4**
Method	Task	Modality	Panoptic Metrics		
			PQ	SQ	RQ
Mask2Forme [21]	PAN	Optical	29.3	80.8	34.9
OneFormer [28]	PAN	Optical	30.0	80.7	35.8
YOSO-EF [37]	PAN	Optical+SAR	31.6	81.0	37.7
CoMFormer-EF [20]	PAN	Optical+SAR	28.8	74.6	34.4
OneFormer-EF [28]	PAN	Optical+SAR	30.5	81.3	36.2
Mask2Former-EF (baseline) [21]	PAN	Optical+SAR	27.9	74.9	33.3
Ours	**PAN**	**Optical+SAR**	**34.5**	**81.8**	**41.2**

The performance amounted to a relative increase of 8.1%, 5.3%, and 1.4% compared to the baseline. For the INS task, compared with the suboptimal method, ICAFusion, the proposed PMTSeg also achieved a relative improvement of 17.4% on mAP, 16.3% on AP@0.5, and 19.5% on AP@0.75. For the PAN task, the corresponding PQ, SQ, and RQ were 34.5%, 81.8%, and 41.2%, respectively, which corresponds to relatively increases of 9.2%, 1.0%, and 9.3%, over YOSO-EF. In addition, OTSeg and SAPNet use optical modalities for segmentation, and these methods rely on prompt mechanisms to guide the model in completing the task. While they demonstrate certain advantages on optical data, our PMTSeg, by combining optical and SAR modalities, not only enhances the accuracy of semantic and instance segmentation but also effectively overcomes the limitations of a single modality.

The results indicate that substantial improvements were achieved for semantic, instance, and panoptic segmentation tasks compared to other state-of-the-art methods.

E. Ablation Studies and Analysis

1) Effect of Different Modules: To evaluate the individual contributions of each component, we performed a comprehensive modular ablation experiment on the SEMCITY TOULOUSE dataset, the results are shown in Table IV. Compared to the baseline, adding the TMAA module relatively improved the results by 2.0% in mIoU on the SEM task, by 5.1% in mAP on the INS task, and by 15.2% in PQ on the PAN task, respectively. It shows that the model efficiently extracts the semantically aligned features, leveraging the affinity-based constraint. In addition, after adding the TMSF module, the performance further improved, which proves that the TMSF captures the task-common information based on the semantic-aligned features from various modalities. The usage of the PSH makes our PMTSeg achieve unified training across various tasks by leveraging one task head and improves the performance of the TMAA and TMSF for MRSIS by leveraging the relationship across different tasks.

TABLE IV

EFFECT OF DIFFERENT MODULES ON MODEL PERFORMANCE ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Base.	PSH	TMAA	TMSF	Semantic Metrics			Instance Metrics			Panoptic Metrics		
				mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
✓				59.1	73.2	75.8	37.1	67.0	36.9	34.9	70.6	49.0
✓		✓		60.3	72.6	76.0	39.0	70.1	40.8	40.2	71.6	55.9
✓			✓	60.1	72.5	76.4	38.7	69.7	38.7	38.8	71.0	54.3
✓		✓	✓	61.3	73.1	77.0	39.6	70.1	40.3	42.2	71.6	58.1
✓	✓			59.1	72.4	76.2	37.1	66.4	37.7	42.8	70.4	59.6
✓	✓	✓		61.5	74.5	77.2	39.0	68.7	40.1	44.6	71.6	61.8
✓	✓		✓	60.8	74.2	76.9	39.2	69.1	40.3	44.7	71.4	62.1
✓	✓	✓	✓	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

TABLE V

EFFECT OF DIFFERENT TASK HEADS ON MODEL PERFORMANCE IN THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD

Type	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
Separate	62.3	74.8	77.4	39.1	67.8	40.4	46.2	**72.6**	63.3
Unified (Ours)	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	72.4	**64.2**

2) Effect of Different Task Heads: To analyze the effect of different task heads on our PMTSeg performance, we compared the separate task heads and the unified task head of three segmentation tasks on the SEMCITY TOULOUSE dataset, and the results as shown in Table V. Compared to using separate task heads, our unified task head in PMTSeg achieved relative improvements of 4.3%, 3.1%, and 1.1% in mIoU, mAP, and PQ, respectively. It indicates that the unified task head achieved better performance across the three segmentation tasks. Additionally, the result reveals that while the separate task heads yielded superior performance in panoptic segmentation compared to semantic and instance segmentation, the variation between their results and those of the unified task head is relatively small, with slightly higher SQ metrics. This may be attributed to the task gap, which hinders the ability of separate task heads to effectively balance all tasks, making the model more prone to certain tasks. In contrast, our PMTSeg employs the unified task head that effectively mitigates the task gap, further enhancing performance across all tasks.

3) Effect of Different Affinity-Based Distribution Loss in TMAA: To evaluate the effect of different affinity-based distribution losses used in the TMAA module for mitigating the modality gap, Table VI shows the segmentation results of using different L_{TMAA} on the SEMCITY TOULOUSE dataset. We conducted experiments using three commonly used distribution evaluation functions: Euclidean distance (ED) [43], KL divergence (KL) [44], and JS divergence (JS) [9], respectively. Among the three different functions, our PMTSeg achieved the best results using the JS divergence. It shows

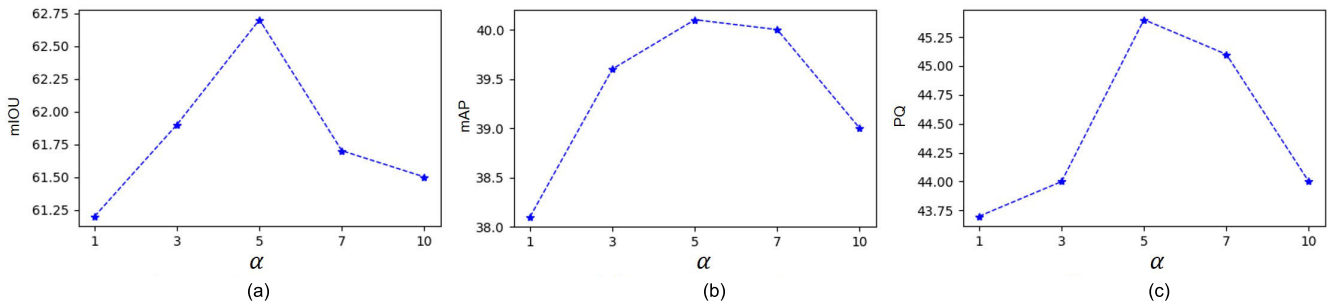

Fig. 6. Effect of different weight parameter values on the performance in the SEMCITY TOULOUSE dataset. (a) Semantic segmentation. (b) Instance segmentation. (c) Panoptic segmentation.

TABLE VI

EFFECT OF DIFFERENT PROBABILITY DISTRIBUTION APPROXIMATION LOSSES IN TMAA ON MODEL PERFORMANCE IN THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Type	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
ED	61.8	73.9	76.9	39.2	69.2	39.7	43.8	71.0	61.3
KL	61.1	74.2	76.3	39.0	69.2	39.5	44.4	71.5	61.7
JS	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

TABLE VII

EFFECT OF DIFFERENT LAYER NUMBERS OF THE CROSS-ATTENTION MODULE APPLIED IN TMSF ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Layer Num.	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
3	59.4	75.8	76.4	39.6	70.3	40.3	43.7	71.2	61.1
2	61.4	76.2	77.2	39.1	70.3	39.3	44.4	71.5	61.6
1	**63.9**	**76.3**	**77.4**	**40.3**	**70.9**	**41.7**	**46.6**	**72.6**	**63.8**

that JS divergence effectively evaluates the similarity between two probability distributions.

4) Effect of Different TMSF Application Layers: On the SEMCITY TOULOUSE dataset, we compared the effect of different numbers of layers in the cross-attention module applied in TMSF, and the results are shown in Table VII. We observe that as the number of layers increases, the model's segmentation performance decreases, which may be attributed to the small size of the RS dataset, making it easy to overfit.

5) Effect of the Key Parameters: To assess the effect of key parameters on our PMTSeg, we adjusted different values for the loss weighting coefficient α on the performance (see (16)). As shown in Fig. 6, the results on all three tasks show the same trend. We observe that as the coefficient values increased, the corresponding mIoU, mAP, and PQ first increased and then decreased. The performance in all tasks achieved the best value on all evaluation metrics when α was set to 5. It demonstrates that our TMAA module can facilitate the semantic alignment across different modalities.

6) Effect of Different Modalities: To analyze the effect of each modality, we conducted a detailed ablation study by selectively removing individual modalities and evaluating the performance on the SEMCITY TOULOUSE dataset. This approach allowed us to isolate and quantify the contribution of each modality to the overall segmentation performance. The results are shown in Table VIII. Compared to the multimodal results, training with only the NIR image resulted in a relative reduction of 16.9%/11.7%/13.1% at mIoU/mAP/PQ, respectively. The results show that removing either the optical image or the NIR image significantly affects the segmentation accuracy of the model, especially in specific scenes. In contrast, the model that combines optical and NIR images shows the best overall performance. It demonstrates that our

TABLE VIII

EFFECT OF DIFFERENT MODALITIES ON MODEL PERFORMANCE ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Modality	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
Optical	57.6	71.1	76.7	38.1	68.2	38.4	43.1	71.6	60.1
NIR	53.1	66.4	74.3	35.6	67.2	34.5	40.5	70.9	56.8
Optical+NIR	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

PMTSeg can capture the valid semantic information from various modalities, thus achieving better performance.

7) Effect of Multiscale Design in TMSF: To further validate the effectiveness of the multiscale design in the TMSF module, we conducted an ablation study by comparing different scale combinations and analyzing their contributions to task performance. Specifically, we progressively increased the number of input scales in the TMSF module—from a single scale (1/4) to up to four scales (1/4, 1/8, 1/16, and 1/32)—and evaluated the model's performance on semantic segmentation, instance segmentation, and panoptic segmentation tasks on the SEMCITY TOULOUSE dataset. The results are presented in Table IX. The findings show a consistent performance improvement as more scales are incorporated. When using only the 1/4 scale, the model yields limited performance across all metrics. Adding the 1/8 scale leads to significant gains, while further including the 1/16 and 1/32 scales continues to enhance the model's capability, eventually achieving the best performance on all tasks. These results highlight the advantages of the multiscale design in capturing fine-grained semantics and facilitating multimodal feature alignment. In summary, the multiscale strategy plays a critical role in the TMSF module. It effectively mitigates the granularity gap across modalities

TABLE IX

EFFECT OF DIFFERENT MULTISCALE DESIGNS ON MODEL PERFORMANCE ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

1/4 scale	1/8 scale	1/16 scale	1/32 scale	Semantic Metrics			Instance Metrics			Panoptic Metrics		
				mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
✓				56.6	74.9	75.9	37.7	67.4	38.5	42.7	71.5	59.5
✓	✓			59.8	72.7	76.4	38.9	68.5	40.1	44.7	72.3	61.7
✓	✓	✓		62.6	76.0	77.1	39.3	68.8	40.0	45.4	72.1	62.3
✓	✓	✓	✓	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

TABLE X

EFFECT OF TASK-ADAPTIVE CONTRASTIVE LOSS L_{CON} ON MODEL PERFORMANCE ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD

Type	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
w/o L_{CON}	62.3	75.3	77.2	39.5	70.1	39.9	45.6	71.5	63.5
with L_{CON}	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

and enhances the model's adaptability and generalization across diverse tasks.

8) Effect of Task-Adaptive Contrastive Loss in PSH: To evaluate the impact of the task-adaptive contrastive loss on model performance, we conducted experiments on the SEMCITY TOULOUSE dataset by comparing models with and without this loss. As shown in Table X, incorporating the task-adaptive contrastive loss leads to consistent performance improvements across all evaluation metrics. Notably, the model shows significant gains in semantic metrics (e.g., mIoU, mACC, and OA) and instance-level metrics (e.g., AP, AP0.5, and AP0.75). In addition, the contrastive loss also enhances the model's performance in panoptic metrics (e.g., PQ, SQ, and RQ), further validating its effectiveness in modeling cross-task relationships. These results demonstrate that the task-adaptive contrastive loss effectively facilitates information alignment and relational modeling across tasks, thereby enhancing overall performance on complex multitask scenarios.

9) Effect of Task Information Encoding in PSH: In the PSH module, the method of encoding task information is crucial for MTL performance. Currently, the mainstream task prompting methods can be categorized into two types: 1) learnable task category embedding (LV) [45], [46], where a unique learnable vector is assigned to each task. This method is simple to implement and ensures stable training but has limited semantic expressiveness, making it difficult to capture the commonalities and differences between tasks and 2) natural language task encoding (NT) [13], [14], which encodes task descriptions through language models, introducing some semantic priors. However, the subjectivity and uncertainty in textual representations, particularly in visual tasks, may lead to semantic guidance biases and unstable performance. To comprehensively evaluate the effectiveness of different task prompting strategies, we designed representative comparative experiments, covering the aforementioned methods and introducing our joint task prompting and label-level semantic modeling strategy. This strategy, building upon task-level semantic prompting, further integrates label-level semantics, enhancing the model's ability to perceive fine-grained task structures and facilitating both cross-task semantic sharing and differentiation. As shown in Table XI, our method outperforms the others in key metrics, such as mIoU, mACC, AP, and PQ, across semantic segmentation, instance segmentation, and panoptic segmentation tasks. Compared to the LV method, we improve task modeling expressiveness by incorporating external semantic information. Compared to the NT method, our approach uses label-level semantic constraints to clarify task scope, alleviating the vagueness caused by language descriptions and improving the model's stability and discriminative power. Overall, our method demonstrates stronger adaptability, discriminability, and generalization capability, particularly in complex multitask scenarios.

TABLE XI

EFFECT OF TASK INFORMATION ENCODING METHODS ON MODEL PERFORMANCE IN THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Semantic Metrics			Instance Metrics			Panoptic Metrics		
	mIoU	mACC	OA	AP	AP0.5	AP0.75	PQ	SQ	RQ
LV	62.1	76.2	76.4	38.9	68.6	39.1	44.7	72.2	61.8
NT	61.2	75.5	76.9	38.2	69.4	38.3	44.5	71.8	61.7
PSH	**65.0**	**78.5**	**78.2**	**40.3**	**70.3**	**41.6**	**46.7**	**72.4**	**64.2**

TABLE XII

COMPUTATIONAL EFFICIENCY COMPARISON ON THE SEMCITY TOULOUSE DATASET. THE BEST RESULTS ARE IN BOLD. THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Task	#Params(M)	#Speed(s)	Semantic Metrics		
				mIoU	mACC	OA
CMFNet [32]	SEM	**106.4**	0.0502	59.7	74.8	74.7
FTransUNet [33]	SEM	204.1	0.0393	54.4	70.3	73.1
Ours	**SEM**	226.3	**0.0263**	**65.0**	**78.5**	**78.2**
Method	Task	#Params(M)	#Speed(s)	Instance Metrics		
				AP	AP0.5	AP0.75
ICAFusion [34]	INS	229.4	0.0621	39.4	70.0	39.1
CalibNet [36]	INS	237.4	0.0862	39.3	68.8	40.0
Ours	**INS**	**226.3**	**0.0263**	**40.3**	**70.3**	**41.4**
Method	Task	#Params(M)	#Speed(s)	Panoptic Metrics		
				PQ	SQ	RQ
CoMFormer-EF [20]	PAN	**132.2**	0.193	33.2	69.4	47.8
OneFormer-EF [28]	PAN	212.7	0.077	42.8	70.4	59.6
Ours	**PAN**	226.3	**0.0263**	**46.7**	**72.4**	**64.2**

10) Analysis of Computational Efficiency: To comprehensively evaluate the computational efficiency of our method, we conducted comparison experiments on the SEMCITY TOULOUSE dataset in terms of inference time and model parameters. As shown in Table XII, although our method involves a larger number of parameters compared to other approaches, it demonstrates highly competitive inference speed, comparable to or even faster than existing baselines. Specifically, our method achieves significantly reduced inference time while maintaining high accuracy across semantic, instance, and panoptic segmentation metrics. In terms of model size, while our approach requires more parameters than some alternatives, it achieves a balanced tradeoff between performance and computational cost. This indicates that, despite the larger model capacity, our method can deliver superior performance without compromising inference efficiency. These results highlight the effectiveness of our approach in achieving both high accuracy and efficient inference.

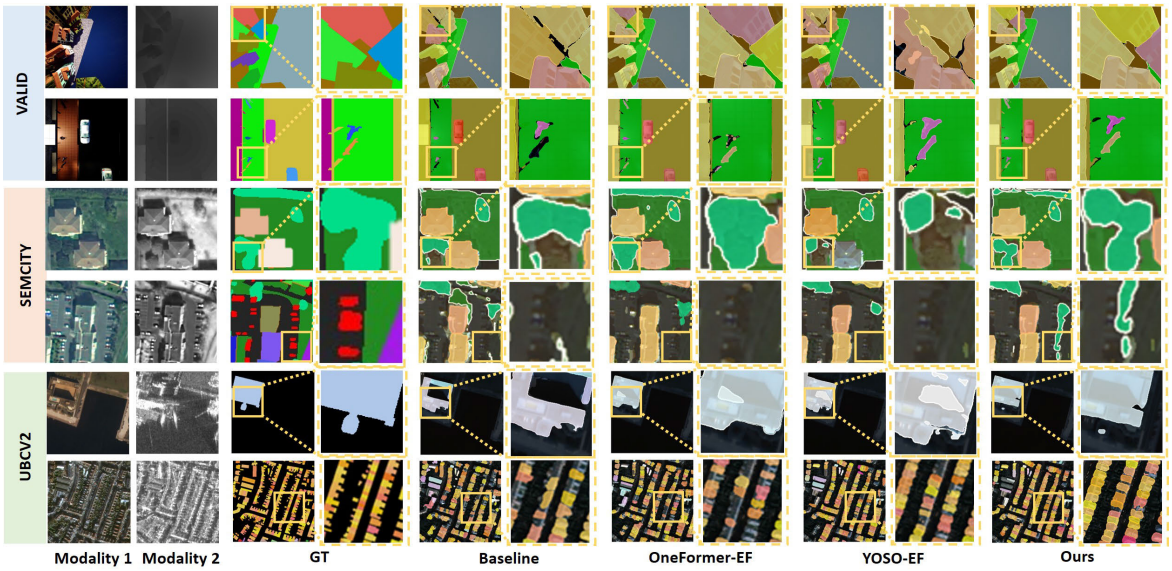

Fig. 7. Visualization comparison of different methods in the panoptic segmentation task.

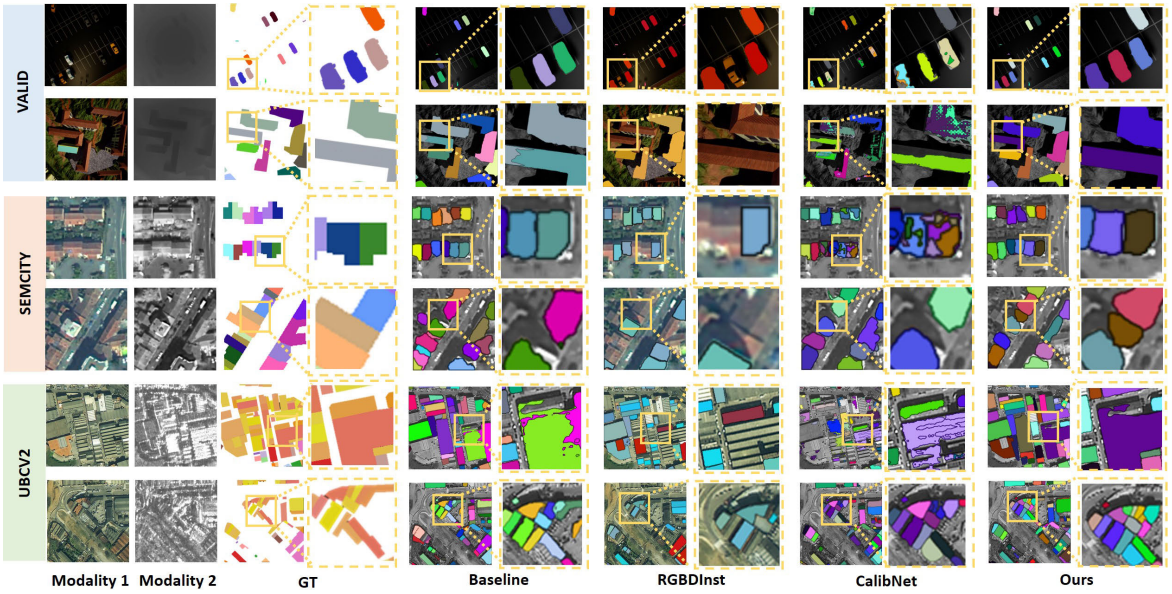

Fig. 8. Visualization comparison of different methods in the instance segmentation task.

F. Visualization Analysis

1) Visualization Results on Panoptic Segmentation: Fig. 7 shows the visualization examples of the panoptic results obtained by our PMTSeg and other state-of-the-art methods. It is observed that RS images are more complex than natural images. Obviously, the proposed PMTSeg can identify complex edges with accurate results, providing a complete solution with fewer independent points. Specifically, the baseline performed poorly in dense segmentation, leading to incomplete results (see the sixth row of Fig. 7). Compared with the baseline, OneFormer [28] reduced the incompleteness degree but had more misclassified pixels (see the third row of Fig. 7). In addition, we observe that YOSO-EF [37] produces rounded edges without obvious corners and segment small objects difficultly (see the sixth row of Fig. 7). Our PMTSeg performs more accurate contours of stuff categories and thing categories than other methods, by effectively eliminating the task gap between various tasks, thus enhancing the panoptic result. However, the model's performance degrades in scenarios with very small and densely packed objects (see the fourth row of Fig. 7). We analyze that the primary reason for this phenomenon lies in the difficulty of capturing the details of small objects. Due to the limited number of pixels for small objects in RS images, the model may not have adequately learned the features of these objects during training, thus impacting its ability to recognize them. Despite this, our model's superior ability to handle complex and dense scenarios, as shown in the other examples, confirms its robustness and efficacy across a wide range of segmentation tasks.

2) Visualization Results on Instance Segmentation: For the instance segmentation task, we compared PMTSeg with other state-of-the-art segmentation methods as shown in Fig. 8.

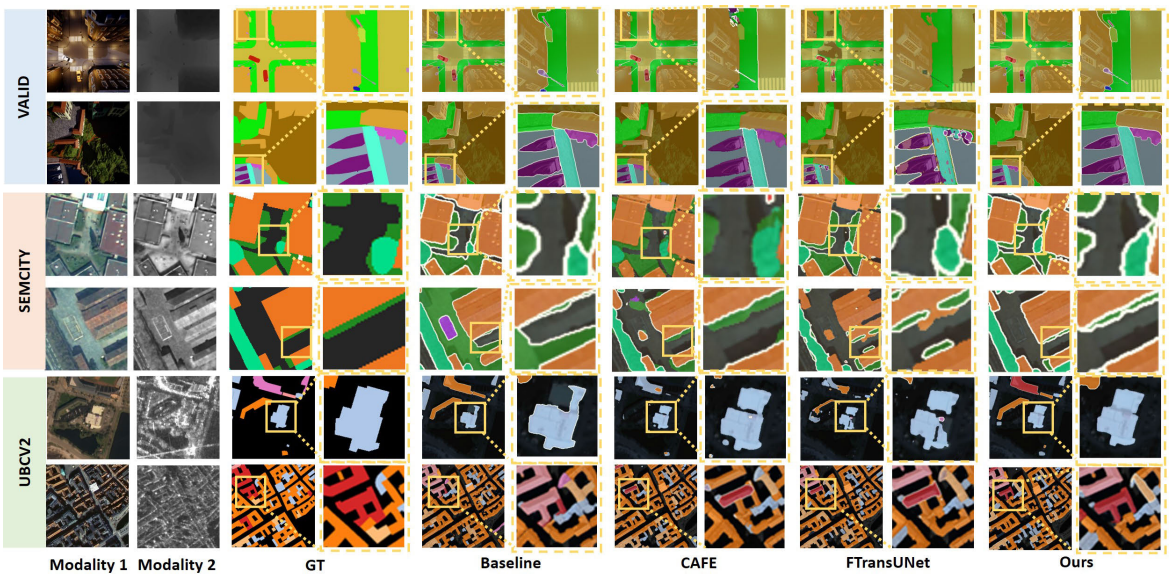

Fig. 9. Visualization comparison of different methods in the semantic segmentation task.

The first two rows show the extraction results on the VALID dataset, the middle two rows show the extraction results on the SEMCITY TOULOUSE dataset, and the last two rows show the extraction results on the UBCV2 dataset. From the 3–6 rows of Fig. 8, we observe that the baseline missed and had incorrect instance results on both the VALID and UBCV2 datasets. RGBDInst [35] produced clearer contours compared to the baseline, but missed more instances. CalibNet [36] reduced the number of missed results but performed poorly in dense building extraction, resulting in interconnections between instances. Our PMTSeg recognized more correct instances with clear contours, benefiting from retaining the details of objects with various scales through the TMSF. However, we also observed a failure case in the sixth row of Fig. 8. In densely populated urban scenes with multiple buildings and complex terrain, our model encountered multiple detections. The primary reason for this issue lies in the model's difficulty in effectively distinguishing the boundaries between closely positioned, similar structures or adjacent objects, leading to misdetections. Despite this, the model performed robustly in other scenarios, demonstrating its strong performance and stability across various tasks.

3) Visualization Results on Semantic Segmentation: Fig. 9 shows the comparison of segmentation results for different methods on VALID, SEMCITY TOULOUSE, and UBCV2 datasets. We observe that the baseline had a significantly lower number of correctly recognized pixels compared to the ground truth. CAFE [31], although recognizing pixel distributions with more numbers than the baseline, still fell short compared to the ground truth. The overall segmentation result of FTransUNet [33] was better, but the boundary was blurred. Our PMTSeg could more accurately classify objects like buildings and clearly generate the edge of the semantic segmentation compared to other methods. This was achieved by the semantic-aligned features constrained by the intermodal affinity, thus integrating the modality gap between various modalities. Nevertheless, we identified a failure case in the sixth row of Fig. 9. In scenes where buildings are connected or adjacent, our model sometimes produces correct contours but assigns incorrect semantic labels. This misclassification is primarily due to the high similarity of appearance between adjacent buildings and the subtle differences in their contextual information, which can confuse the model. Despite such occasional failures, our PMTSeg still demonstrates superior robustness and accuracy across various datasets and segmentation challenges.

4) Fine-Grained Segmentation Visualization: To analyze the ability of our PMTSeg to recognize different categories, we compared the fine-grained results of different methods on the SEMCITY TOULOUSE dataset for three segmentation tasks, as shown in Fig. 10. Compared to other state-of-the-art methods, our PMTSeg performed significantly improvement in all categories, especially in the SV, BD, and WA categories for the SEM task. It demonstrates that our PMTSeg is beneficial and robust across various tasks in all categories.

5) Visualization on Semantic Feature Through Progressive Module Integration: To analyze the effect of different modules on the results, Fig. 11 visualizes the semantic features obtained by progressively integrating different modules on the SEMCITY TOULOUSE dataset. After the features were processed by the TMAA and TMSF, we observed that the semantic features are more focused on the meaningful areas instead of the fragmented areas, compared with the baseline. In addition, after processing by the PSH, the semantic features exhibited the information that easily adapted to various segmentation tasks. Through these three modules, our PMTSeg significantly enhances the edge features of objects and provides valid semantic features for the MRSIS task.

6) Visualization on Multiple Task Heads Versus Our PSH: To significantly analyze the effect of different task heads on the results, we compare the heatmap visualizations generated by multiple task heads and our PSH across multiple

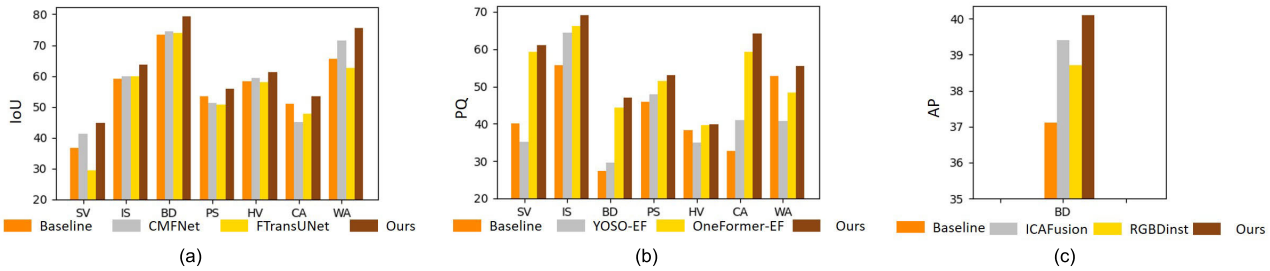

Fig. 10. Visualization results of different methods to segment 7 fine-grained categories with IoU on the SEM task of SEMCITY TOULOUSE dataset. The seven categories are: sport venues (SV), impervious surface (IS), building (BD), pervious surface (PS), high vegetation (HV), car (CA), and water (WA). (a) Semantic segmentation. (b) Panoptic segmentation. (c) Instance segmentation.

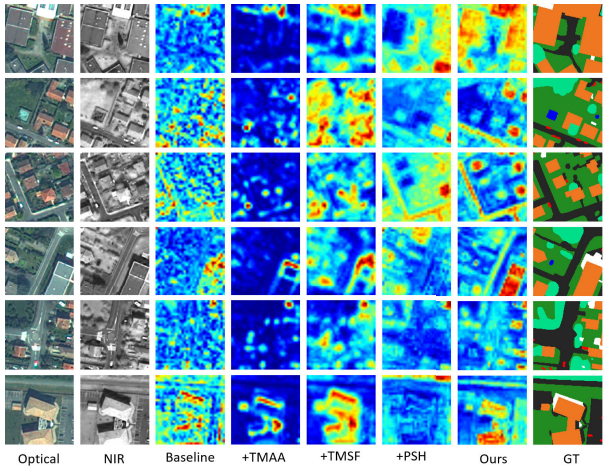

Fig. 11. Comparison of heatmaps generated by different models on the SEMCITY TOULOUSE dataset. We gradually add the components based on the baseline model to evaluate the effectiveness.

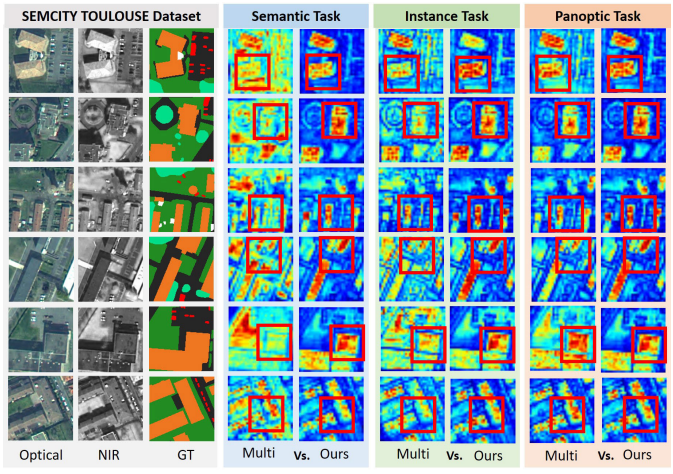

Fig. 12. Comparison of heatmaps generated by multiple task heads (Multi) and our PSH (Ours) of three segmentation tasks on SEMCITY TOULOUSE.

segmentation tasks. From Fig. 12, we observe that, compared to the heatmaps generated by multiple task heads, our PSH exhibits a closer alignment of the attended semantics with the ground truth across all three segmentation tasks. Moreover, the results are consistent with those shown in Table V, indicating that the attention semantics of multiple task heads are more dispersed in semantic and instance segmentation. However, in panoptic segmentation, the attention semantics are significantly closer to the ground truth. This further demonstrates that our PSH effectively mitigates the task gap, thereby improving performance across multiple segmentation tasks.

V. Conclusion

In this article, we propose a novel PMTSeg, which effectively captures the task-common semantic information across different modalities for multiple tasks using one task head, thus suppressing the modality gap and the task gap. To alleviate the modality gap, PMTSeg first uses the TMAA to obtain the semantic-aligned multimodal features by approximating the intermodal affinity matrices across modalities. This effectively reduces the semantic discrepancies for various RS data. Then, PMTSeg employs the TMSF to further integrate the semantic-aligned information of different modalities at different granularities to obtain the multiscale task-common features, further mitigating the granularity discrepancies within the modality gap. To bridge the task gap, the PSH is used to guide the PMTSeg to adaptively capture the relationship between different tasks, thus achieving the unified training. Experimental results on three MRSIS datasets demonstrate the effectiveness, robustness, and generalization of our proposed approach on semantic, instance, and panoptic segmentation tasks. Despite the effectiveness of our method, we found that there is still some room that our PMTSeg could be improved. For example, our approach primarily focuses on validating the effectiveness of combining multimodal RS data to enhance the performance of various segmentation tasks. In the future, we will consider exploring more advanced techniques to integrate additional tasks in RS, rather than being limited to segmentation tasks.

References

[1] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605415.

[2] Y. Wang et al., "Emotion-oriented cross-modal prompting and alignment for human-centric emotional video captioning," *IEEE Trans. Multimedia*, vol. 27, pp. 3766–3780, 2025.

[3] W. Zhou, S. Dong, J. Lei, and L. Yu, "MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 48–58, Jan. 2023.

[4] T. Fernando, C. Fookes, H. Gammulle, S. Denman, and S. Sridharan, "Toward on-board panoptic segmentation of multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402312.

[5] Y. Liu et al., "Sample-cohesive pose-aware contrastive facial representation learning," *Int. J. Comput. Vis.*, vol. 133, no. 6, pp. 3727–3745, Jun. 2025.

[6] J. Fan, J. Li, Z. Hua, F. Zhang, and C. Zhang, "Elevation information-guided multimodal fusion robust framework for remote sensing image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[7] B. Liu, H. Chen, K. Li, and M. Y. Yang, "Transformer-based multimodal change detection with multitask consistency constraints," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102358.

[8] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[9] P. W. Lamberti, A. P. Majtey, M. Madrid, and M. E. Pereyra, "Jensen-Shannon divergence: A multipurpose distance for statistical and quantum mechanics," in *Proc. AIP Conf.*, 2007, vol. 913, no. 1, pp. 32–37.

[10] R. Niu, X. Sun, Y. Tian, W. Diao, Y. Feng, and K. Fu, "Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5611918.

[11] S. Wei, T. Zhang, and S. Ji, "A concentric loop convolutional neural network for manual delineation-level building boundary segmentation from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407511.

[12] V. S. Fare Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4852–4861.

[13] H. Bi et al., "Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 131–148, Jan. 2025.

[14] C. Shang, Z. Song, H. Qiu, L. Wang, F. Meng, and H. Li, "Prompt-driven referring image segmentation with instance contrasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4124–4134.

[15] H. Liu, L. Guo, Z. Zhou, and H. Zhang, "Pyramid-context guided feature fusion for RGB-D semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2022, pp. 1–6.

[16] M. Sharma et al., "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021.

[17] V. Sainte Fare Garnot, L. Landrieu, and N. Chehata, "Multi-modal temporal attention models for crop mapping from satellite time series," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 294–305, May 2022.

[18] Y. Wang, F. Sun, W. Huang, F. He, and D. Tao, "Channel exchanging networks for multimodal and multitask dense image prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5481–5496, May 2023.

[19] X. Cheng, Y. Zheng, J. Zhang, and Z. Yang, "Multitask multisource deep correlation filter for remote sensing data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3723–3734, 2020.

[20] F. Cermelli, M. Cord, and A. Douillard, "CoMFormer: Continual learning in semantic and panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3010–3020.

[21] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[22] L. Chen, F. Liu, Y. Zhao, W. Wang, X. Yuan, and J. Zhu, "VALID: A comprehensive virtual aerial image dataset," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 2009–2016.

[23] R. Roscher, M. Volpi, C. Mallet, L. Drees, and J. D. Wegner, "SemCity Toulouse: A benchmark for building instance segmentation in satellite images," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 5, pp. 109–116, Aug. 2020.

[24] X. Huang et al., "Urban building classification (UBC) V2—A benchmark for global building detection and fine-grained classification from satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5620116.

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

[26] Y. Fang et al., "Instances as queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6890–6899.

[27] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2019). *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2

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

[29] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[30] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.

[31] A. Zheng, J. He, M. Wang, C. Li, and B. Luo, "Category-wise fusion and enhancement learning for multimodal remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4416212.

[32] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3463–3474, 2022.

[33] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5403215.

[34] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang, "ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109913.

[35] W. Ye, W. Zhang, W. Lei, W. Zhang, X. Chen, and Y. Wang, "Remote sensing image instance segmentation network with transformer and multi-scale feature representation," *Expert Syst. Appl.*, vol. 234, Dec. 2023, Art. no. 121007.

[36] J. Pei et al., "CalibNet: Dual-branch cross-modal calibration for RGB-D salient instance segmentation," *IEEE Trans. Image Process.*, vol. 33, pp. 4348–4362, 2024.

[37] J. Hu, L. Huang, T. Ren, S. Zhang, R. Ji, and L. Cao, "You only segment once: Towards real-time panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17819–17829.

[38] T. Cheng et al., "Sparse instance activation for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4423–4432.

[39] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

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

[41] K. Kim, Y. Oh, and J. C. Ye, "OTSeg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 200–217.

[42] Z. Wei, P. Chen, X. Yu, G. Li, J. Jia, and Z. Han, "Semantic-aware SAM for point-prompted instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3585–3594.

[43] P.-E. Danielsson, "Euclidean distance mapping," *Comput. Graph. Image Process.*, vol. 14, no. 3, pp. 227–248, Nov. 1980.

[44] J. M. Joyce, *Kullback–Leibler Divergence*. Berlin, Germany: Springer, 2011.

[45] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, and Z. Chen, "Open-set video-based facial expression recognition with human expression-sensitive prompting," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 5722–5731.

[46] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "MMA-DFER: MultiModal adaptation of unimodal models for dynamic facial expression recognition in-the-wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4673–4682.