Segment Any RGB-Thermal Model with Language-aided Distillation

Dong Xing[1], Xianxun Zhu[2], Wei Zhou[3], Qika Lin[4],Hang Yang[1], Yuqing Wang[1†]

[1] Changchun Institute of Optics, Fine Mechanicsand Physics, Chinese Academy of Sciences University,

[2]School of Computing, Macquarie University, NSW 2109, Australia,

[3]School of Computer Science and Informatics, Cardiff University, United Kingdom

[4]Saw Swee Hock School of Public Health, National University of Singapore

Abstract—The recent Segment Anything Model (SAM) demonstrates strong instance segmentation performance across various downstream tasks. However, SAM is trained solely on RGB data, limiting its direct applicability to RGB-thermal (RGB-T) semantic segmentation. Given that RGB-T provides a robust solution for scene understanding in adverse weather and lighting conditions, such as low light and overexposure, we propose a novel framework, SARTM, which customizes the powerful SAM for RGB-T semantic segmentation.Our key idea is to unleash the potential of SAM while introduce semantic understanding modules for RGB-T data pairs. Specifically, our framework first involves fine tuning the original SAM by adding extra LoRA layers, aiming at preserving SAM's strong generalization and segmentation capabilities for downstream tasks. Secondly, we introduce language information as guidance for training our SARTM. To address cross-modal inconsistencies, we introduce a Cross-Modal Knowledge Distillation(CMKD) module that effectively achieves modality adaptation while maintaining its generalization capabilities. This semantic module enables the minimization of modality gaps and alleviates semantic ambiguity, facilitating the combination of any modality under any visual conditions. Furthermore, we enhance the segmentation performance by adjusting the segmentation head of SAM and incorporating an auxiliary semantic segmentation head, which integrates multi-scale features for effective fusion. Extensive experiments are conducted across three multi-modal RGBT semantic segmentation benchmarks: MFNET, PST900, and FMB. Both quantitative and qualitative results consistently demonstrate that the proposed SARTMsignificantly outperforms state-of-the-art approaches across a variety of conditions.

Index Terms—RGB-Thermal Semantic Segmentation, Segment Anything Model, Language Distillation

I. INTRODUCTION

In the field of autonomous driving, robust and reliable semantic scene understanding is critical for ensuring the safety of vehicle operations [1]–[4]. RGB-thermal (RGBT) semantic
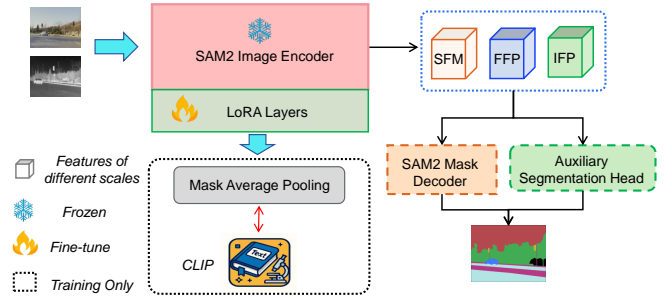
[†]: Corresponding Author

Dong Xing, Hang Yang, Yuqing Wang are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences University, Changchun, 130033, China (e-mail: xingdong24@mails.ucas.ac.cn, yanghang@ciomp.ac.cn, wyq7903@163.com).

Xianxun Zhu is with the School of Computing, Macquarie University, NSW 2109, Australia. (e-mail: xianxun.zhu@mq.edu.au).

Wei Zhou is with the School of Computer Science and Informatics, Cardiff University, United Kingdom.

Qika Lin is with Saw Swee Hock School of Public Health, National University of Singapore, 119077, Singapore

Fig. 1. Overall framework of our proposed SARTM, consists of original SAM2 components and proposed language distillation part,including the Semantic Feature Map (SFM), Fine-Grained Feature Pyramid (FFP), and Intermediate-Resolution Feature Pyramid (IFP).

segmentation technology provides a significant avenue for addressing the scene understanding challenges in adverse weathers and lighting conditions. For instance, during foggy or low-light scenarios, RGB cameras struggle with object recognition due to low visibility, while thermal infrared cameras can effectively detect objects by utilizing their thermal signatures [5]–[8]. By fusing information from both modalities, more accurate and robust semantic segmentation can be achieved in complex environments [9], [10]. As a result, RGBT semantic segmentation has garnered considerable attention in recent years, with notable advancements in the field [11], [12].

Recently, the Segment Anything Model (SAM) [13] represents a breakthrough, particularly in RGB image segmentation. Despite SAM's impressive performance in single-modality segmentation tasks, its application in multi-modal segmentation remains a ***unique challenge***. Specifically, in multi-modal scenarios such as thermal infrared, the data from different modalities exhibit significant differences, and how to effectively integrate complementary information is still an open problem. The recently proposed SAM2 model [14] extends SAM's capabilities by incorporating the temporal dimension to address challenges in video segmentation, such as motion, deformation, occlusion, and lighting changes. These advancements expand SAM's applicability to dynamic and multi-modal environments, but further research is needed to explore ***how to effectively fuse cross-modal information while***

retaining SAM2's strong generalization ability. Additionally, SAM not directly perform pixel-level fine-grained predictions, as it focuses more on segmenting object regions rather than precise pixel-level labeling.

Although image fusion technology has wide applications, existing methods heavily rely on visual features, such as texture, contrast, and pixel registration, while neglecting deeper semantic information [15]–[18]. As a result, deeper semantic information inherent within the images is often neglected. A pressing challenge is to effectively leverage the deeper, non-visual semantic features present in the images. Inspired by the successes of multi-modal vision-language model [19]–[29], we explore a novel approach under the guidance of MVLM,*i.e.*CLIP [22], addressing the challenge by learning modality-agnostic representations and opening a new direction in RGBT fusion.

To address the aforementioned challenges including: ① adapt SAM2 to multi-modal domain, ② introduce language as guidance for RGBT, and ③ equip SAM2 semantic understanding ability, we present a novel framework SARTMbased on the SAM2 architecture. As illustrated in Figure1, The SAM2 architecture requires modality-specific fine-tuning to improve segmentation performance across various modalities. Built upon SAM2 model, we further propose: ① to first adapt SAM2 which is sorely trained with RGB data, we apply the Low rank adaptation (LoRA) layers to the image encoder to facilitates effective modality-adaptation fine-tuning while preserving the generalization ability of SAM2's pre-trained knowledge; The image encoder processes the input visual modality and generates a semantic feature map (SFM). This map is then passed through the mask decoder's convolutional module, which creates two additional feature pyramids: the fine-grained feature pyramid (FFP) and the intermediate-resolution feature pyramid (IFP).These pyramids, along with the SFM, enhance the model's *spatial and semantic representation capacity*. ② aiming at utilizing the pretrained knowledgeable large vision language models, we integrate language information as guidance into the language-aided distillation module. This module allows the model to *better understand the intrinsic semantic correlation within the RGBT data*, and finally improve the semantic segmentation accuracy in even complex and challenging environments. ③ to introduce the SAM2 which is only capable of instance segmentation to semantic segmentation, we modify the original mask decoder of SAM2 and further incorporate a auxiliary semantic decoder. Aiming at *effectively integrating multi-scale features* is essential for improving segmentation accuracy in dynamic and diverse environments, we upgrade the SAM2 segmentation pipeline by incorporating multi-scale feature extraction and fusion mechanisms. Specifically, we augment the original segmentation head with an auxiliary head to leverage complementary information across multiple scales, thereby improving segmentation accuracy.

Extensive experiments are conducted on three benchmark datasets, including MFnet [6], PST900 [8], and FMB [7], demonstrate the outstanding performance of our framework in the RGB-T semantic segmentation task. As shown in Figure 1, our method achieves significant improvements on the PST900

dataset, while also demonstrating notable enhancements on the MFnet dataset. Our contributions are as follows:

(i) We modify the SAM2 framework by integrating the LoRA layers into the multi-modal semantic segmentation task, investigating their potential for RGB-T image segmentation and adapting the framework to the multimodal domain.

(ii) We incorporate language descriptions into the training process by embedding explicit (language model-derived) textual guidance into the image fusion algorithm. This allows the model to better grasp the semantic context, enhancing segmentation accuracy in complex scenarios to support semantic segmentation.

(iii) We redesign the SAM2 segmentation pipeline by merging modified segmentation heads for multi-modal inputs and introducing an auxiliary segmentation head. This configuration enables efficient multi-scale feature fusion, leading to a significant improvement in segmentation precision.

(iv) Our method achieves state-of-the-art performance on three widely used multi-modal benchmarks, spanning synthetic to real-world scenarios, outperforming existing methods in segmentation accuracy and generalization across various dimensions.

II. RELATED WORKS

A. RGB-T Semantic Segmentation.

Multi-modal semantic segmentation often enhances performance by integrating the RGB modality with additional visual modalities containing complementary scene information, such as thermal and depth modalities. These supplementary modalities provide critical information for vision systems in diverse scenarios. In RGB-thermal semantic segmentation, Ha *et. al.* [6] introduced MFNet as a pioneering method for RGB-T semantic segmentation, employing a dual-stream architecture for feature extraction and cascading the fusion of these features. Sivakumar *et. al.* [8] proposed a two-stage framework where the output of the first stage was fused with thermal and color images. However, these approaches primarily rely on simple fusion strategies, such as element-wise summation or concatenation, to capture cross-modal features, which can lead to redundancy by overlooking differences between cross-modal information. To address this, several studies have focused on designing specialized feature fusion operations. For instance, Lv *et. al.* [12] proposed the Context-Aware Interaction Network (CAINet), which establishes complementary relationships between multi-modal features and long-term contexts in spatial and channel dimensions. Zhang *et. al.* [30] introduced MRFS, where different modalities provide learning priors to each other. By contrast, *we improve RGB-thermal semantic segmentation by leveraging SAM for robust segmentation and LoRA for fine-tuning, enhancing cross-modal feature alignment and reducing redundancy*.

B. SAM for Semantic Segmentation.

The Segment Anything Model (SAM) [13] is a versatile segmentation framework that achieves unprecedented performance in image segmentation by leveraging a combination of large-scale, diverse datasets and robust vision models. SAM

has catalyzed advancements across various subfields of computer vision [31]–[33]. Its architecture consists of three core components: an image encoder, a prompt encoder, and a mask decoder. SAM has demonstrated its effectiveness in multiple domains, including medical imaging [34]–[37], remote sensing segmentation [38]–[40], object segmentation [41]–[44],object tracking, [45], detection [46], [47] and 3D reconstruction [48].

In our work, we pioneer the application of SAM to this domain by training it on RGB-X semantic segmentation tasks, marking the first exploration of SAM in this context. We adpat SAM to RGB-T semantic segmentation by fine-tuning it with LoRA, enhancing cross-modal feature alignment and improving segmentation accuracy.

C. Vision-Language Model.

Recently, vision-language multi-modal learning [25], [29], [49]–[54], has emerged as a prominent research focus. Vision-language models such as CLIP [22], DALL-E [55], and GPT-4 [56]have demonstrated remarkable performance across a variety of downstream tasks. These large-scale models provide external knowledge for image description, enabling the generation of strong and explicit prompts. For instance, Zhang *et. al.* [57] proposed an interactive model that facilitates the understanding of video-sentence queries and captures semantic correspondences. Shang *et. al.* [58] achieved more comprehensive vision-language interactions and fine-grained text-to-pixel alignment through bidirectional prompting. Furthermore, Wang *et. al.* [59] introduced a CLIP-Driven Referring Image Segmentation framework (CRIS), which effectively transfers multi-modal knowledge to achieve text-to-pixel alignment. In this paper, we use clip text encoder as a teacher model to *extract semantic embeddings of class names and transfer cross-modal knowledge to the segmentation model to improve the performance* of the model.

III. METHOD

A. Framework Overview

Building upon the SAM2 framework, we introduce a customized SAM2 architecture, referred to as the SARTM framework, which is specifically optimized for RGB-T semantic segmentation tasks, as shown in Figure 2. The customization starts by freezing the pre-trained image encoder and fine-tuning it through a LoRA layer. This strategy allows the encoder to adapt to the new visual modality while retaining the valuable knowledge learned during the pre-training phase. The image encoder processes the input visual modality I and produces a semantic feature map (SFM) F_n^m. These feature maps are then passed through the convolutional module of the mask decoder to generate two supplementary feature pyramids: the fine-grained feature pyramid (FFP) F_0^m and the intermediate-resolution feature pyramid (IFP) F_1^m. Combined with the SFM, these pyramids further enhance the model's spatial and semantic representation capacity.

To facilitate effective cross-modal feature fusion, we propose a framework that performs a weighted average of the cross-modal representations (SFM, FFP, and IFP), resulting in the feature representation $\overline{F}_i$, where $i \in \{0, 1, n\}$. To further enhance the quality of image fusion, we introduce a language-guided fusion module, which leverages language information to guide the feature fusion process and generate the weighted feature maps $\hat{F}_i$. These fused feature maps are subsequently combined into a unified feature representation $\tilde{F}_i$, which serves as the input for downstream semantic segmentation tasks.

For the semantic segmentation task, we adopt a dual-path prediction strategy to improve segmentation performance. The first path processes the fused features by feeding them into the SAM2 mask decoder, which employs the frozen Transformer block to derive mask labels from the SFM. These labels then interact with the fine-grained and intermediate-resolution pyramid features, facilitating the construction of high-resolution feature representations. These enhanced high-resolution representations are subsequently processed by the hypernetwork to generate precise segmentation masks, denoted as $\tilde{\mathbf{S}}_0$.

In the second path, the fusion features are passed to the auxiliary segmentation head, which uses bilinear interpolation to upsample the backbone features, and fuses them with the medium resolution feature pyramid pixel by pixel, then upsamples to the fine-grained pyramid size for fusion, and finally the merged features are processed by the 3×3 convolution layer to enhance the context information, and then upsamples to the target resolution, denoted as $\tilde{\mathbf{S}}_1$.

In addition, we extract the semantic embedding of class names through CLIP's text encoder as a teacher model to guide the training of segmentation models. In order to realize cross-modal knowledge transfer, we propose an implicit relational transfer mechanism, which distills knowledge by aligning self-similar matrices of image features and text features. By calculating the similarity difference between the student model and the teacher model, the student model can distill the cross-modal semantic knowledge from the teacher model and improve the model performance.

B. LoRA Layers in SAM2's image encoder

We define the input set for multi-modal data as:$I = \{I^m \in \mathbb{R}^{H \times W \times C} \mid m \in [1, M]\}$, where H, W, and C represent the height, width, and number of channels of each modality, respectively. The index m denotes a specific modality, such as RGB or Thermal. Each modality is independently processed through the hierarchical backbone network of Hiera to extract multi-scale features.

Initially, a patch embedding operation transforms each input I^m into an embedded feature map:

$$P(I^m) = I^m W_e + b_e, \tag{1}$$

where $W_e \in \mathbb{R}^{C \times d}$ is the weight matrix, $b_e \in \mathbb{R}^d$ is the bias vector, and d denotes the embedding dimension. The height and width of the downsampled feature map are given by:

$$H_0 = H/s_0, \quad W_0 = W/s_0. \tag{2}$$

The SAM2 backbone progressively reduces the spatial resolution while increasing feature dimensionality across n hierarchical stages:

$$\{I_i^m \in \mathbb{R}^{C_i \times H_i \times W_i} \mid i \in [0, n], \, m \in [1, M]\}, \tag{3}$$
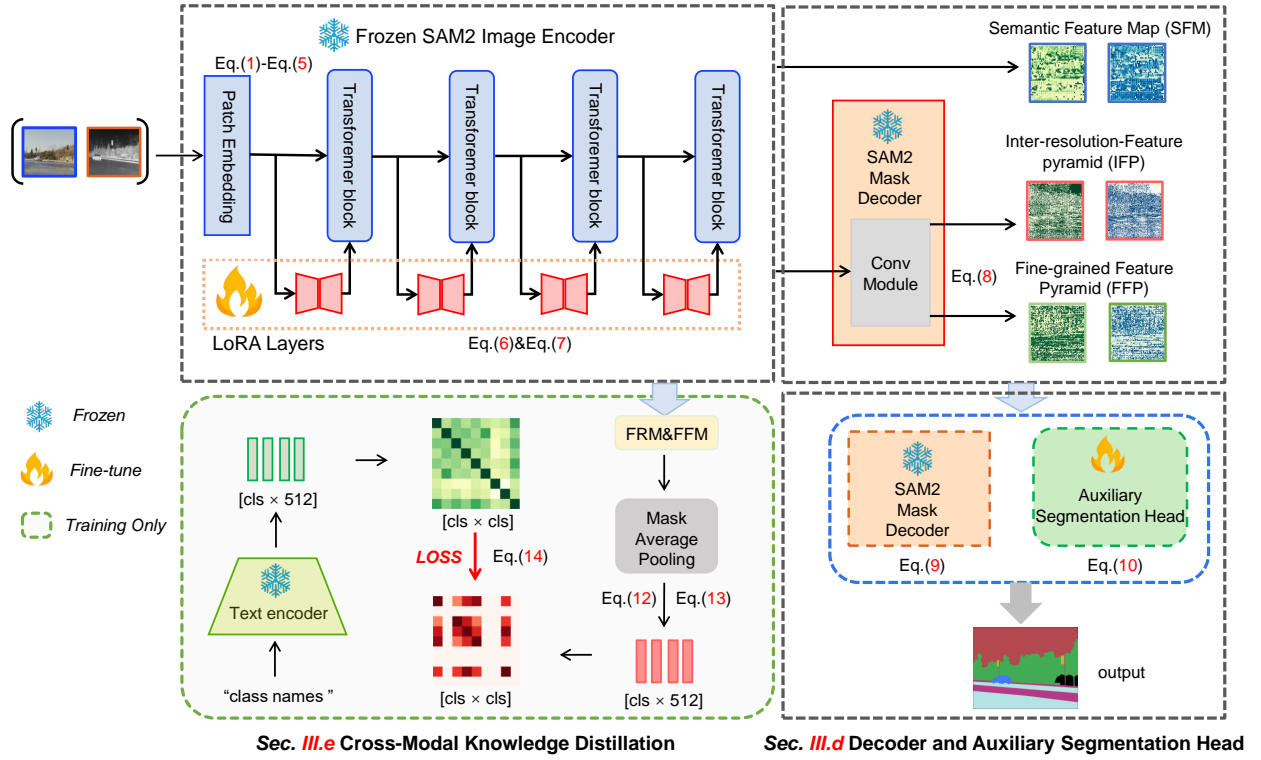
Fig. 2. Illustration of the proposed **SARTM** framework for multi-modal semantic segmentation. The architecture combines multi-scale features from a frozen image encoder fine-tuned with LoRA layers.

where the spatial dimensions at stage i are given by:

$$H_i = H/s_i, \quad W_i = W/s_i, \quad s_i = 2^{i+2}. \tag{4}$$

A window-based multi-head self-attention mechanism is employed at each stage:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{5}$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k is the key dimensionality.

To improve efficiency and adaptation specific to the modality, we introduce a LoRA layer to update the query and value projections, as shown in Eq. (6).In this formulation, $W_a^Q, W_a^V \in \mathbb{R}^{d \times r}$ and $W_b^Q, W_b^V \in \mathbb{R}^{r \times d}$ are low-rank matrices, with $r \ll d$ serving as the rank parameter. These updates result in augmented projections, as defined in Eq. (7). The LoRA parameters are modality-specific and trained independently, while the backbone parameters remain frozen, ensuring efficient cross-modal adaptation.

$$\Delta Q^m = W_a^Q W_b^Q, \quad \Delta V^m = W_a^V W_b^V \tag{6}$$

$$Q'^m = Q^m + \Delta Q^m, \quad V'^m = V^m + \Delta V^m \tag{7}$$

C. Feature Pyramid Network

The hierarchical features are further refined using a Feature Pyramid Network (FPN), which integrates lateral and top-down pathways to enhance multi-scale feature representations. At each stage i, the input feature map I_i^m undergoes a lateral

convolution operation, yielding a refined modality-specific feature map $Z_i^m \in \mathbb{R}^{d \times H_i \times W_i}$. This operation reduces the channel dimensionality to d, while maintaining the spatial dimensions H_i and W_i, thus ensuring consistency in spatial resolution and compatibility for subsequent fusion operations within the FPN.

Let $\mathcal{L}$ denote the set of layers where top-down fusion is applied. For each layer $i \in \mathcal{L}$, top-down fusion combines feature representations from deeper layers with those at the current stage, producing the fused feature map F_i^m. This fusion process is mathematically defined in Eq. (8).

$$F_i^m = \begin{cases} \frac{Z_i^m + \text{Upsample}(F_{i+1}^m)}{2}, & i \in \mathbf{L} \\ Z_i^m, & i \notin \mathbf{L}. \end{cases} \tag{8}$$

Here, $F_i^m \in \mathbb{R}^{d \times H_i \times W_i}$ denotes the fused feature map in stage i, which integrates the modality-specific feature map Z_i^m with the upsampled feature map from the subsequent layer, F_{i+1}^m. The Upsample operation adjusts the spatial resolution of F_{i+1}^m to match that of Z_i^m, ensuring accurate feature integration. The hierarchical refinement that underlies the multi-scale feature representation of the FPN is central to this fusion process.

The Feature Pyramid Network (FPN) is employed to generate three separate feature maps for each modality, each adapted to capture semantic and spatial details at different resolutions. These feature maps include the *SFM* ($F_n^m \in \mathbb{R}^{d \times H_n \times W_n}$), the *FFP* ($F_0^m \in \mathbb{R}^{d \times H_0 \times W_0}$), and the *IFP* ($F_1^m \in \mathbb{R}^{d \times H_1 \times W_1}$). To improve the representational power of the higher resolution maps (F_0^m and F_1^m), 1x1 convolutional layers are applied, reducing their channel dimensions while preserving their spa-
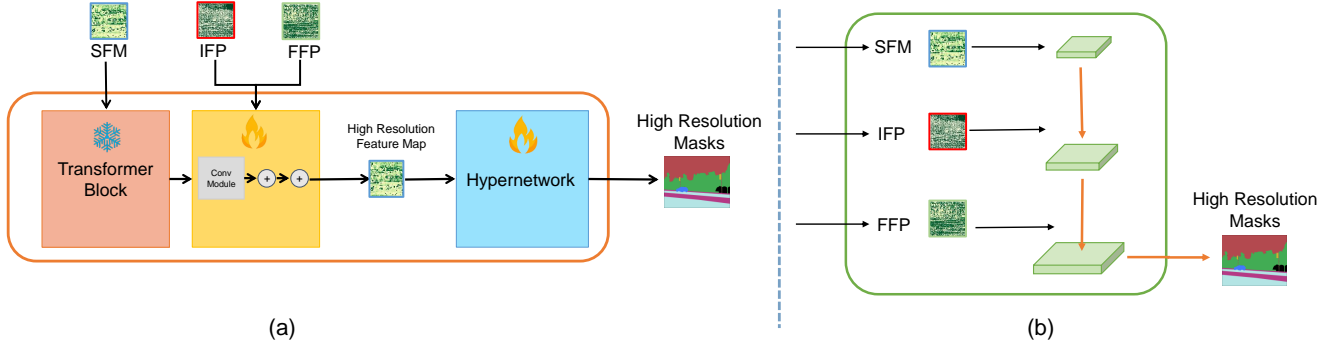
Fig. 3. (a) Hierarchical Feature Fusion for Cross-Scale Information Aggregation using Semantic Feature Map (SFM), Fine-Grained Feature Pyramid (FFP), and Intermediate-Resolution Feature Pyramid (IFP) for high-resolution feature map generation. (b) Hierarchical Refinement Pathway for High-Resolution Embedding leveraging SFM, FFP, and IFP to refine the high-resolution masks.

tial resolution. Consequently, the dimensions are adjusted so that $F_0^m \in \mathbb{R}^{d/8 \times H_0 \times W_0}$ and $F_1^m \in \mathbb{R}^{d/4 \times H_1 \times W_1}$, leading to a more compact and efficient representation, suitable for the subsequent fusion.

D. Decoder and Auxiliary Segmentation Head

Next, we employ a dual-pathway mask prediction strategy on the unified feature map $\tilde{F}$ to generate high-resolution segmentation masks.

In the first pathway shown in Figure 3(a), we extend SAM2's mask decoder to produce high-resolution multimasks. This involves generating high-resolution segmentation logits, denoted as $\tilde{\mathbf{S}}_0 \in \mathbb{R}^{\mathcal{C} \times H_0 \times W_0}$, through a structured multi-scale fusion process. Here, $\mathcal{C}$ represents the number of segmentation categories. The backbone features $\tilde{\mathbf{F}}_n \in \mathbb{R}^{d \times H_n \times W_n}$, which encapsulate the global semantic context, are processed via a transformer-based decoder f_{dec}, producing low-resolution logits. These logits are iteratively refined by incorporating spatially detailed features from intermediate resolution feature maps $\tilde{\mathbf{F}}_1 \in \mathbb{R}^{d/4 \times H_1 \times W_1}$ and fine-grained feature maps $\tilde{\mathbf{F}}_0 \in \mathbb{R}^{d/8 \times H_0 \times W_0}$. This hierarchical refinement process is mathematically described as Eq (9), where f_{dec} denotes the transformer-based decoding operation applied to $\tilde{\mathbf{F}}_n$, Upsample performs bilinear upsampling to match spatial resolutions, and Conv is a 1×1 convolution for channel alignment.

$$\begin{aligned} \mathbf{S}_{\text{low}} &= f_{\text{dec}}(\tilde{\mathbf{F}}_n) \\ \mathbf{S}_{\text{inter}} &= \text{Upsample}(\mathbf{S}_{\text{low}}) + \text{Conv}(\tilde{\mathbf{F}}_1) \\ \tilde{\mathbf{S}}_0 &= \text{Upsample}(\mathbf{S}_{\text{inter}}) + \text{Conv}(\tilde{\mathbf{F}}_0) \end{aligned} \quad (9)$$

As shown in Figure 3(b), the second pathway utilizes a feature fusion mechanism to integrate multi-scale features into a unified high-resolution embedding. ierarchical feature integration framework that aggregates multi-scale features into a unified high-resolution embedding The backbone features $\tilde{\mathbf{F}}_n \in \mathbb{R}^{d \times H_n \times W_n}$ is upsampled to match the spatial dimensions of intermediate resolution feature maps $\tilde{\mathbf{F}}_1 \in \mathbb{R}^{d/4 \times H_1 \times W_1}$ through bilinear interpolation.Then, The upsampled $\tilde{\mathbf{F}}_n \in \mathbb{R}^{d \times H_n \times W_n}$ and $\tilde{\mathbf{F}}_1 \in \mathbb{R}^{d/4 \times H_1 \times W_1}$ are combined via pixel-wise summation to generate intermediate feature fpn merge. Finlly,fpn merge is further upsampled and fused

with $\tilde{\mathbf{F}}_0 \in \mathbb{R}^{d/8 \times H_0 \times W_0}$ through additive merging, producing the final fused feature final merge. The fused feature final merge is then processed by a 3×3 convolutional layer to enhance contextual awareness, followed by upsampling to the target resolution, denoted as $\tilde{\mathbf{S}}_1 \in \mathbb{R}^{\mathcal{C} \times H_0 \times W_0}$. This dual-path integration scheme enables effective cross-scale information aggregation, significantly improving the segmentation model's accuracy and robustness.

$$\tilde{\mathbf{S}}_1 = \text{Conv}\left(\text{Upsample}\left(\tilde{\mathbf{F}}_1 + \text{Upsample}\left(\tilde{\mathbf{F}}_n\right)\right) + \tilde{\mathbf{F}}_0\right) \quad (10)$$

The proposed framework optimizes a composite loss function that strategically combines Cross-Entropy (CE) loss with multi-scale supervision. Given ground truth segmentation labels $\mathbf{L} \in \mathbb{R}^{H_t \times W_t}$ are defined such that $\mathbf{L}(i,j) \in \{0, 1, \ldots, \mathcal{C} - 1, 255\}$ (255 denoting ignored regions), the CE loss for prediction $\tilde{\mathbf{S}}$ is formulated as:

$$\mathcal{L}_{\text{CE}}(\tilde{\mathbf{S}}, \mathbf{L}) = \frac{1}{n_{\min}} \sum_{i \in \mathcal{H}} \mathcal{L}_{\text{CE}}(\tilde{\mathbf{S}}(i), \mathbf{L}(i)) \quad (11)$$

where $\mathcal{L}_{\text{CE}}$ is the pixel-wise cross-entropy loss, and $\mathcal{H}$ represents the set of hardest pixels, selected based on prediction difficulty. The normalization factor $n_{\min} = \max(|\mathcal{H}|, n_{\text{threshold}})$ ensures that a sufficient number of complex examples are included, where $n_{\text{threshold}} = n_{\text{total}}/16$, and n_{total} is the total number of valid pixels in the image.

E. Cross-Modal Knowledge Distillation

CLIP [22] optimizes the distance between image and text features in a shared embedding space, aiming to minimize the distance between matched image-text pairs for modality gap mitigation .

As illustrated in Fig 2, we employ CLIP text encoder to extract semantic embeddings T_e from category names (e.g., ["Person", ...]) as supervision signals for the segmentation model

At the fusion step, the feature $F_n^m \in \mathbb{R}^{d \times H_n \times W_n}$ will be further fused with RGB feature by the cross-modal Feature Rectification Module (FRM) [49] and Feature Fusion Module (FFM) [49],termed as f.
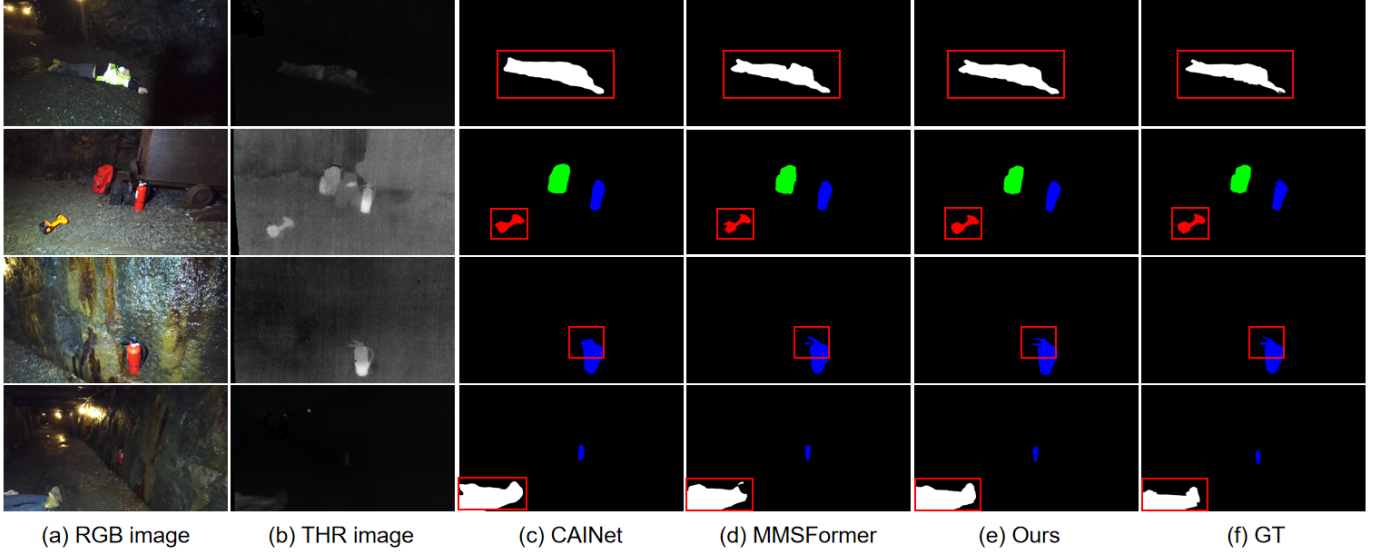
Fig. 4. Qualitatively compared with the SoTA RGB-T scene resolution network on the PST900 test set, where areas of significant improvement are shown in red boxes.

The category representations are computed through label $y \in Y$ and segmentation head outputs using Mask Average Pooling (MAP).

$$\{f_0, \ldots, f_K\} = MAP(f, y). \tag{12}$$

$$\mathcal{L}_{cr} = KL(f, y). \tag{13}$$

The MAP procedure comprises two key operations: Class-specific feature extraction using y as class masks, followed by average pooling over spatial dimensions to obtain final semantic representations $\{f_0, \ldots, f_K\}$. To enable cross-modal knowledge transfer, we propose an implicit relation transfer mechanism that distills intra-modality semantic knowledge by aligning the self-similarity matrices of T_e and $\{f_0, \ldots, f_K\}$.

$$\mathcal{L}_{se} = KL(Cos(\{f_0, \ldots, f_K\}, \{f_0, \ldots, f_K\}^T), Cos(T_e, T_e^T)). \tag{14}$$

where KL denotes the Kullback-Leibler Divergence. After minimizing modality gaps among visual modalities, we strive to reduce the semantic ambiguity by transferring the intra-modal semantic knowledge.

F. Ovearll Training Objectives

The overall loss function incorporates the OhemCrossEntropy loss applied to both $\tilde{\mathbf{S}}_0$ and $\tilde{\mathbf{S}}_1$, as defined in Eq. (15).

$$\mathcal{L} = w_0 \cdot \mathcal{L}_{CE}(\tilde{\mathbf{S}}_0, \mathbf{L}) + w_1 \cdot \mathcal{L}_{CE}(\tilde{\mathbf{S}}_1, \mathbf{L}) + w_2 \cdot \mathcal{L}_{cr} + w_3 \cdot \mathcal{L}_{se} \tag{15}$$

where $w_0, w_1, w_2, w_3 > 0$ are scalar weights that control the relative importance of each loss term.

TABLE I
QUANTITATIVE COMPARISONS (%) WITH THE SoTA RGB-T SCENE PARSING METHODS ON THE PST900 TEST SET. THE SYMBOL "-" DENOTES MISSING DATA IN THE ORIGINAL PUBLICATION, AND THE BEST RESULTS ARE PRESENTED IN BOLD FONT.

Methods	Background	Fire-Extinguisher	Backpack	Hand-Drill	Survivor	mIoU
MFNet [6]	98.6	41.1	64.2	60.3	20.7	57.0
RTFNet [10]	98.9	36.4	75.3	52.0	25.3	57.6
EGFNet [60]	99.2	74.3	83.0	71.2	64.6	78.5
ABMDRNet [61]	98.7	24.1	72.9	54.9	57.6	67.3
FEANet [62]	-	-	-	-	-	85.5
DBCNet [63]	98.9	62.3	71.1	52.4	40.6	74.5
CAINet [5]	99.5	80.3	88.0	77.2	78.6	84.7
EAEFNet [64]	99.5	80.4	87.7	83.9	75.6	85.4
GMNet [65]	99.4	85.1	83.8	73.7	78.3	84.1
MRFS [30]	99.6	81.5	89.8	79.6	76.7	87.5
MMSFormer [66]	99.6	81.5	89.8	79.6	76.7	87.5
CRM-RGBTSeg [67]	99.6	79.5	89.6	89.0	82.2	88.0
HAPNet [68]	99.6	81.3	92.0	**89.3**	82.4	89.0
SARTM (Ours)	**99.7**	**88.92**	**92.97**	83.6	**84.19**	**89.88**

IV. EXPERIMENTS

A. Datasets

Datasets. To verify the effectiveness of SARTM, we conduct extensive experiments on three publicly available RGB-Thermal (RGB-T) semantic segmentation datasets, namely MFNet [6],FMB [7],PST900 [8].

TABLE II
QUANTITATIVE COMPARISONS (%) WITH THE SOTA RGB-T SCENE PARSING METHODS ON THE MFNET TEST SET. THE SYMBOL '-' DENOTES MISSING DATA IN THE ORIGINAL PUBLICATION, AND THE BEST RESULTS ARE PRESENTED IN BOLD FONT.

Methods	Unlabeled	Car	Person	Bike	Curve	Car Stop	Color Cone	Bump	mIoU
MFNet [6]	96.9	65.9	58.9	42.9	29.9	9.9	25.2	27.7	39.7
RTFNet [10]	98.5	87.4	70.3	62.7	45.3	29.8	29.1	55.7	53.2
FuseSeg [11]	97.6	87.9	71.7	64.6	44.8	22.7	46.9	47.9	54.5
EGFNet [60]	-	87.6	69.8	58.8	42.8	33.8	48.3	47.1	54.8
ABMDRNet [61]	**98.6**	84.8	69.6	60.3	45.1	33.1	47.4	50.0	54.8
FEANet [62]	98.3	87.8	71.1	61.1	46.5	22.1	55.3	48.9	55.3
SegMiF+ [70]	98.1	87.8	71.4	63.2	47.5	31.1	48.9	50.3	56.1
CAINet [5]	-	88.5	66.3	**68.7**	**55.4**	31.5	48.9	**60.7**	58.6
EAEFNet [64]	-	87.6	72.6	63.8	48.6	35.0	52.4	58.3	58.9
CMX [71]	98.3	90.1	75.2	64.5	50.2	35.3	54.2	60.6	59.7
SARTM (Ours)	98.4	**91.32**	**75.42**	65.22	49.97	**49.47**	**54.96**	55.53	**60.03**

The details of these datasets are as follows.

- **MFNet dataset** consists of 820 daytime and 749 nighttime RGB-T images, each with a resolution of 640×480. This dataset covers eight common object classes typically encountered in driving scenarios.
- **PST900 dataset** provides a total of 597 and 288 calibrated RGB-T images with a resolution of 1280×720 for training and validation, respectively. Collected from the DARPA Subterranean Challenge, the dataset is annotated with four object classes.
- **FMB dataset** contains 1500 image pairs, each consisting of infrared and visible images, annotated with 15 pixel-level categories. The training set comprises 1220 pairs, while the test set includes 280 pairs.

Evaluation Metrics. We utilize two commonly adopted evaluation metrics, accuracy (Acc) and intersection over union (IoU), to assess the scene parsing performance for each category. Additionally, the mean values of these metrics across all categories, referred to as mean IoU (mIoU) and mean accuracy (mAcc), are calculated to provide an overall assessment of the network's performance.

Training Settings. We use the AdamW optimizer [69] with an initial learning rate 1e-4 and weight decay 0.01. The model is trained with a batch size of 8 for 500 epochs. We use cross-entropy loss function.

B. Comparisons with SoTA Scene Parsing Networks

We perform quantitative comparisons with 12, 12, and 11 state-of-the-art (SoTA) RGB-T scene parsing networks on the PST900 [8], MFNet [6], and FMB [7] datasets, respectively. The results are presented in Tables I, II, and III. Furthermore, we provide qualitative comparisons on these three datasets, as illustrated in Figures 4,Figures 5,Figures 6 .

The experimental results of our proposed SARTM on the PST900 underground dataset are shown in Table I, demonstrating the superior performance of our framework in key safety perception for underground environments. Our method achieves a state-of-the-art mIoU of 89.88%, representing a 0.88% absolute improvement over HAPNet [68], with particularly significant advancements in the life-critical categories.

Furthermore, the proposed method is applicable to various datasets, demonstrating strong performance across different scenarios. Specifically, as shown in Table II and Table III, on the MFnet dataset, although our method slightly underperforms compared to the best-performing method in terms of mIoU, the difference is marginal, indicating that our framework achieves near-optimal performance on this dataset. On the FMB dataset, our method outperforms the comparison methods, further validating the effectiveness of SARTM in diverse datasets.

Therefore, despite minor discrepancies on some datasets, our method shows broad applicability and robust performance across different environments, making it highly effective for real-world applications.

C. Qualitative Analysis

In addition to the quantitative analysis, we conducted a qualitative evaluation of the predicted segmentation maps. As shown in Fig 4, we compare prediction results on the PST900 [8] dataseg with those of CAINet [12] and MMS-Former [66]. The figure displays the input RGB image, thermal image, ground truth segmentation map and the model's predictions. As highlighted by the rectangular bounding boxes, our model demonstrates superior accuracy in detecting ocjects with more precise contours compared to the other two metods.

Fig 5 presents the material segmenation results predicted by CMNext [72],CAINet [12] and our SARTM model. As

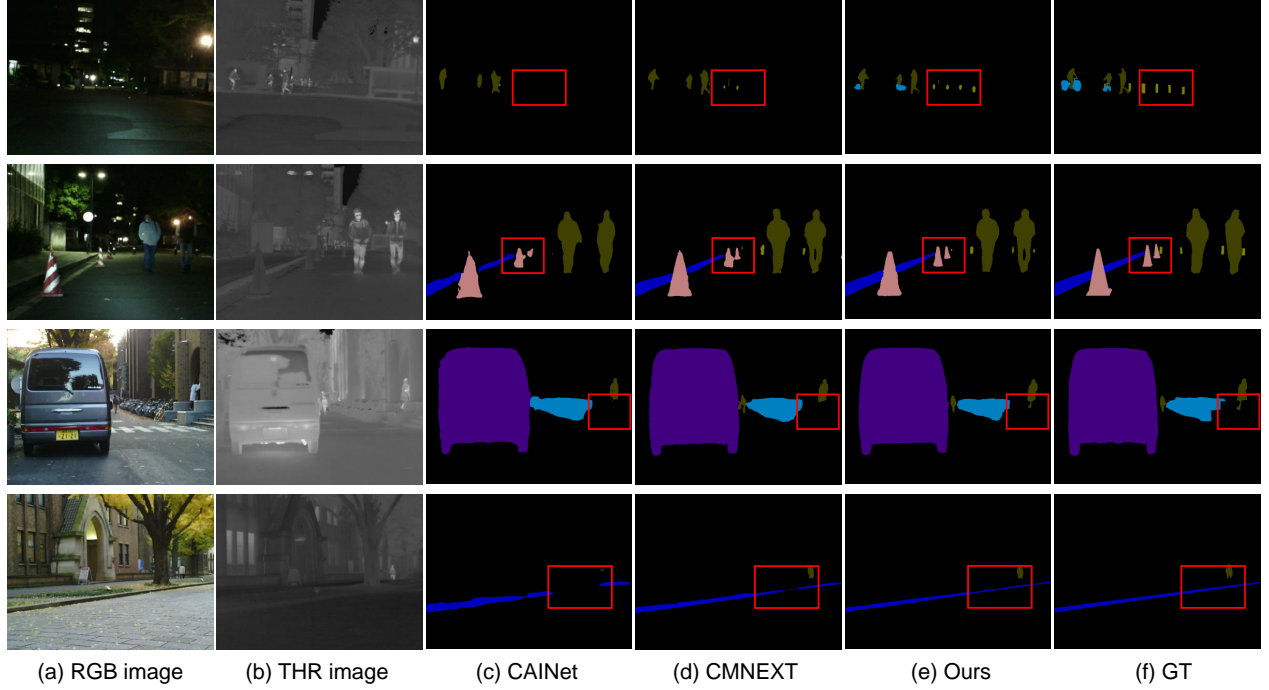| (a) RGB image | (b) THR image | (c) CAINet | (d) CMNEXT | (e) Ours | (f) GT |

Fig. 5. Qualitatively compared with the SoTA RGB-T scene resolution network on the MFNet test set, where areas of significant improvement are shown in red boxes.

emphasized within the rectangulra borders, our propsed model exhibtis strong segmentation performance, particularly for objects such as color cone, fine details of figure, and Curve.

Fig 6 illustrates the capability of effectively utilizing thermal data to generate more coherent segmentations. For instance, for people in nighttime conditions, the segmentation performance of the baseline model is poor due to lighting issues.However, our method successfully identifies the person as a single entity, highlighting its remarkable ability to leverage thermal data for segmentation. Compared to other methods, our approach exhibits strong segmentation performance across categories such as bus, person, and pole.

D. Ablation Study

1) Impact of Rank Size on Performance in LoRA Layers: LoRA is utilized because it provides a lightweight method to adapt pre-trained models to our task without retraining all the parameters, which is computationally expensive. In this study, we investigate the impact of LoRA rank on the performance of SARTM. We tested various LoRA ranks (2, 4, 16, 32, and 64) . Table IV shows the performance of SARTM with different LoRA ranks. Our results indicate that, within a certain range of ranks, the performance of SAM steadily improves. However, when the rank exceeds 16, performance begins to degrade. This trend suggests that returns diminish once the LoRA rank surpasses the optimal point, with Rank 16 yielding the best overall segmentation performance.

2) Effectiveness of Key Components: We conducted a series of ablation studies to investigate the contribution of each component within the fusion module to the overall model performance. As shown in Table V, the results highlight the importance of these components.Initially, In the RGB and Thermal modalities of the FMB dataset,we observed that the absence of language guidance resulted in a 3% performance degradation, indicating that aligning knowledge across modalities through the calculation of feature similarities significantly aids in effective feature extraction. Furthermore, removing the auxiliary segmentation head led to a 4% performance drop. The FPN module, renowned for its ability to aggregate contextual information at multiple scales, substantially enhances the model's capacity to capture and integrate multi-scale contextual information, which is crucial for achieving precise segmentation. Similar effects were observed on the other two datasets, further validating the effectiveness of our approach. These comprehensive ablation studies collectively underscore the importance of each component in the fusion module, revealing that every module plays a unique and vital role in achieving the overall performance of the model.

3) Effectiveness of Loss Functions: We conducted an ablation study on the proposed loss functions, and the results presented in the table demonstrate that each of the loss functions contributes positively to the improvement of model performance. Specifically, the experimental results highlight the distinct contributions of each loss function to the overall performance. Initially, without any additional loss functions, the model achieved a mean Intersection over Union (mIoU) of 84.95, an F1 score of 90.43, and an accuracy of 87.48. After incorporating the Cross-Entropy (CE) loss, the model's performance improved by 0.84 in mIoU, 0.89 in F1 score, and 0.79 in accuracy, indicating a significant impact of the

TABLE III
QUANTITATIVE COMPARISONS (%) WITH THE SoTA RGB-T SCENE PARSING METHODS ON THE FMB TEST SET. THE SYMBOL "-" DENOTES MISSING DATA IN THE ORIGINAL PUBLICATION, AND THE BEST RESULTS ARE PRESENTED IN BOLD FONT.

Methods	Car	Person	Truck	T-Lamp	T-Sign	Building	Vegetation	Pole	mIoU
GMNet [73]	79.3	60.1	22.2	21.6	69.0	79.1	83.8	39.8	49.2
LASNet [74]	72.6	48.6	14.8	2.9	59.0	75.4	81.6	36.7	42.5
EGFNet [75]	77.4	63.0	17.1	25.2	66.6	77.2	83.5	41.5	47.3
FEANet [62]	73.9	60.7	32.3	13.5	55.6	79.4	81.2	36.8	46.8
DIDFuse [76]	77.7	64.4	28.8	29.2	64.4	78.4	82.4	41.8	50.6
ReCoNet [77]	75.9	65.8	14.9	34.7	66.6	79.2	81.3	44.9	50.9
U2Fusion [78]	76.6	61.9	14.4	28.3	68.9	78.8	82.2	42.2	47.9
TarDAL [79]	74.2	56.0	18.8	29.6	66.5	79.1	81.7	41.9	48.1
SegMiF [70]	78.3	65.4	47.3	43.1	74.8	82.0	85.0	49.8	54.8
U3M [80]	**82.3**	65.97	41.93	46.22	**81.0**	81.3	86.76	48.76	60.76
MRFS [30]	76.1	**71.3**	34.4	**50.0**	75.8	**85.4**	**86.9**	**53.64**	61.1
SARTM (Ours)	78.3	65.4	**47.3**	43.1	74.8	82.0	85.0	49.8	**61.57**

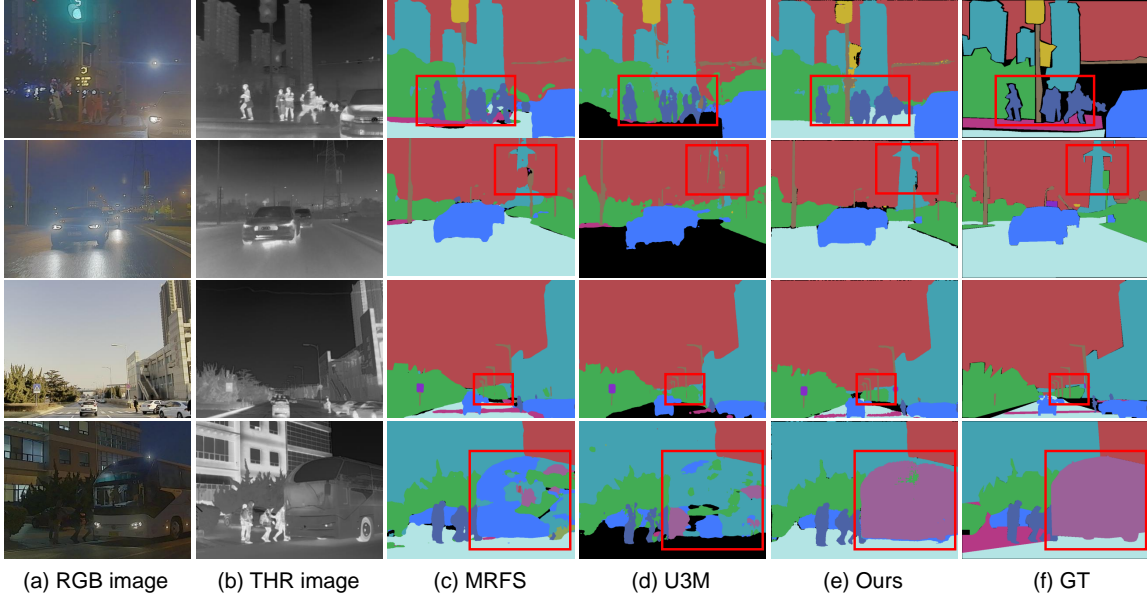(a) RGB image   (b) THR image   (c) MRFS   (d) U3M   (e) Ours   (f) GT

Fig. 6. Qualitatively compared with the SoTA RGB-T scene resolution network on the FMB test set, where areas of significant improvement are shown in red boxes.

TABLE IV
ABLATION STUDY ON THE RANK SIZE ON THE LoRA LAYER

Rank size	mAcc	mIoU
4	92.01	87.68
8	92.55	88.11
16	**93.71**	**89.88**
32	92.73	89.01
64	91.95	88.8

TABLE V
ABLATION STUDY OF THE FUSION BLOCK ON THREE DATASET. THE TABLE SHOWS THE CONTRIBUTION OF DIFFERENT MODULES IN OVERALL MODEL PERFORMANCE.

Structure	FMB	MFNet	PST900
SARTM	**61.57**	**60.03**	**89.88**
- without language	58.85 (-2.72)	59.32 (-0.71)	86.24 (-3.64)
- without Aux_Seg_Head	57.12 (-4.4)	58.67 (-1.36)	84.92 (-4.96)

CE loss on performance enhancement. The best performance was achieved when all loss functions were combined, further validating their effectiveness in optimizing the model.

4) Impact of losses weight on Performance in SARTM: The impact of the weight of different loss functions on the performance of SARTM was evaluated through an ablation
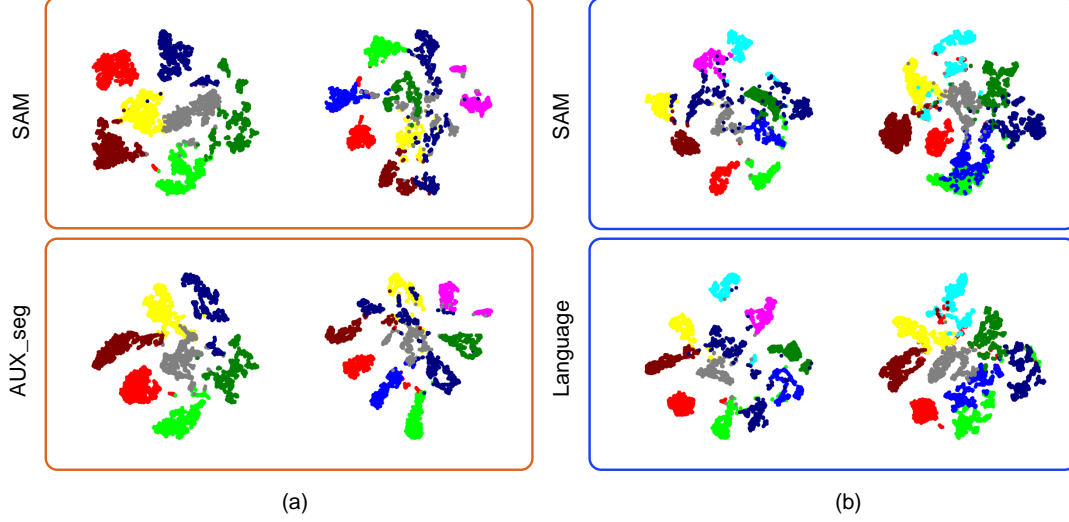
Fig. 7. t-SNE visualization based on the FMB dataset. (a)Comparison between SAM and the addition of an auxiliary segmentation head. (b)Comparison between SAM and the introduction of language-aided. These panels help to illustrate the distribution and separation of features in two-dimensional space, providing insight into the discriminative ability of features extracted under different input conditions
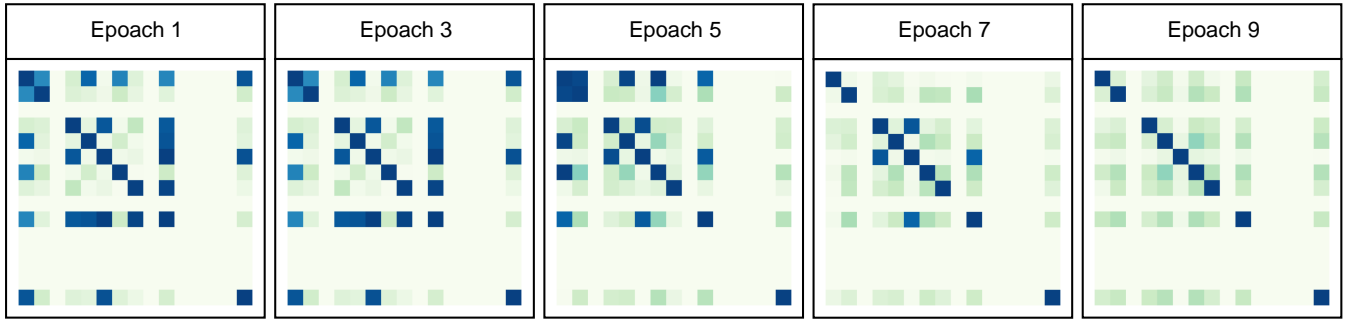

Fig. 8. Visualization of the similarity matrix at different epochs, showing the transition from a dispersed pattern at Epoch 1 to a more structured and unified pattern by Epoch 9, indicating the model's improved understanding of the underlying data structure.

TABLE VI
ABLATION OF DIFFERENT LOSSES ON PST900 WITH SARTM.

$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{cr}$	$\mathcal{L}_{se}$	mIoU	Δ	F1	Δ	Acc	Δ
✓				84.95		90.43		87.82	
✓	✓			85.79	+0.84	91.32	+0.89	89.27	+1.45
✓	✓	✓		87.14	+1.35	92.87	+1.55	91.01	+1.74
✓	✓	✓	✓	**89.88**	+1.84	**94.57**	+1.70	**93.7**	+2.29

study on the MFNet dataset. The results, shown in Table VII, indicate that varying the weights of the loss terms significantly influences the model's performance. For $w1$, the highest mIOU value of 58.67 was achieved when $w1$ was set to 0.008, indicating an optimal balance between the loss terms. Increasing $w1$ to higher values, such as 0.03 or 0.05, led to a decrease

in mIOU, with values of 54.53 and 57.81, respectively. This suggests that excessively large or small weights for $w1$ may negatively impact the model's performance. For $w2$, a weight of 10000 achieved the highest mIOU of 59.32, which was slightly lower than the peak performance of 60.03 when $w3$ was set to 100. These findings highlight the importance of properly tuning the loss weights to maximize performance, as extreme values for any of the loss terms could lead to suboptimal results. In summary, the optimal configuration for maximizing mIOU is achieved by setting $w1$ to 0.008, $w2$ to 10000, and w_3 to 100, resulting in a final mIOU of 60.03, demonstrating the importance of fine-tuning the loss weight parameters in SARTM.

5) Impact of Auxiliary Segmentation Head on Performance:
To further examine the impact of the auxiliary segmentation head on the performance of the fusion block, we compared three different segmentation heads: FPN, DeepLab, and Seg-Former. The experimental results, presented in Table VIII, show that the FPN segmentation head outperforms the others, achieving an mIoU of 58.85%. In contrast, the DeepLab

TABLE VII
ABLATION STUDY OF THE LOSSES WEIGHT ON MFNET DATASET.

w_1	mIOU	w_2	mIOU	w_3	mIOU
0.05	57.81	10	57.42	1	58.05
0.03	54.53	100	57.16	10	57.56
0.008	**58.67**	1000	57.42	50	59.16
0.004	57.46	**10000**	**59.32**	**100**	**60.03**
0.002	56.91	100000	58.64	1000	57.63

TABLE VIII
IMPACT OF DIFFERENT AUXILIARY SEGMENTATION HEADS
PERFORMANCE IN FMB [7] DATASET.

Auxiliary Segmentation Head	% mIoU (Change)
fpn	58.85
deeplab	57.58
segformer	56.94

segmentation head yields an mIoU of 57.58%, and the Seg-Former segmentation head achieves 56.94%. These findings indicate that the FPN segmentation head provides superior performance in this experiment, significantly enhancing the overall performance of the fusion block.

E. t-SNE visualization

The feature clustering results extracted by SARTM on the FMB dataset, as shown in Figure 7.Specifically, as depicted in Figure 7(a), incorporating the auxiliary segmentation head improves feature discrimination. Furthermore, as shown in Figure 7(b), the introduction of language further significantly improves the feature discriminability, enabling the model to learn more discriminative features. This analysis further substantiates the role of the auxiliary segmentation head and language-aided. The experimental results highlight the value of the proposed method in semantic segmentation tasks, particularly in achieving more precise and robust segmentation outcomes.

1) langue-aided distillation visualization: We visualized the language module, and as shown in Fig 8, at epoch 1, the patterns in the similarity matrix were highly dispersed, with many off-diagonal values, indicating that the model had not yet fully captured the underlying structure of the data. As training progressed, the similarity matrix became more organized, with clearer and stronger focal elements. By epoch 9, the matrix displayed distinct patterns, with prominent diagonal values, and the structure became more cohesive. This progression reflects an improvement in the alignment and consistency across the language-aided modalities.

V. CONCLUSION

This paper presents SARTM, a novel adaptation of the SAM2 architecture specifically tailored for RGB-T semantic segmentation tasks. SARTM integrates LoRA-based adaptability, which enhances the SAM model to better align with the requirements of our task. Furthermore, we introduce a dual-channel prediction mechanism and multi-scale fusion features, which significantly improve segmentation accuracy. Finally, we utilize language as the guiding modality. The experimental results demonstrate that, with the language-aided, the model can more effectively achieve the alignment between the category modality and the feature modality. This allows for a deeper understanding of the context and enhances the model's capability in category discrimination, enabling the model to identify and process information of different categories more accurately, thereby improving the model's performance. Extensive experiments demonstrate that SARTM exhibits superior performance on RGB-T semantic segmentation benchmarks.

REFERENCES

[1] X. Zheng, Y. Liu, Y. Lu, T. Hua, T. Pan, W. Zhang, D. Tao, and L. Wang, "Deep learning for event-based vision: A comprehensive survey and benchmarks," *arXiv preprint arXiv:2302.08890*, 2023. 1

[2] X. Zheng and L. Wang, "Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 448–17 458. 1

[3] J. Zhou, X. Zheng, Y. Lyu, and L. Wang, "Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 633–18 643. 1

[4] W. Zhou, R. Zhang, L. Li, G. Yue, J. Gong, H. Chen, and H. Liu, "Dehazed image quality evaluation: From partial discrepancy to blind perception," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3843–3858, 2024. 1

[5] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Transactions on Multimedia*, 2024. 1, 6, 7

[6] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "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)*, 2017, pp. 5108–5115. 1, 2, 6, 7

[7] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8115–8124. 1, 2, 6, 7, 11

[8] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9441–9447. 1, 2, 6, 7

[9] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing," *IEEE Transactions on Multimedia*, vol. 24, pp. 2526–2538, 2022. 1

[10] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019. 1, 6, 7

[11] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2020. 1, 7

[12] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6348–6360, 2024. 1, 2, 7

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 4015–4026. 1, 2

[14] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint*

arXiv:2408.00714, Aug. 2024. [Online]. Available: http://arxiv.org/abs/2408.00714v2 1

[15] J. Wen, F. Qin, J. Du, M. Fang, X. Wei, C. L. P. Chen, and P. Li, "Msgfusion: Medical semantic guided two-branch network for multi-modal brain image fusion," *IEEE Transactions on Multimedia*, vol. 26, pp. 944–957, 2024. 2

[16] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5906–5916. 2

[17] J. Chen, X. Li, L. Luo, and J. Ma, "Multi-focus image fusion based on multi-scale gradients and image matting," *IEEE Transactions on Multimedia*, vol. 24, pp. 655–667, 2022. 2

[18] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X.-J. Wu, "Ifsepr: A general framework for image fusion based on separate representation learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 608–623, 2023. 2

[19] J. Han, R. Zhang, W. Shao, P. Gao, P. Xu, H. Xiao, K. Zhang, C. Liu, S. Wen, Z. Guo *et al.*, "Imagebind-llm: Multi-modality instruction tuning," *arXiv preprint arXiv:2309.03905*, 2023. 2

[20] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023. 2

[21] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023," *arXiv preprint arXiv:2305.06500*, vol. 2, 2023. 2

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 2, 3, 5

[23] W. Zhang, L. Wu, Z. Zhang, T. Yu, C. Ma, X. Jin, X. Yang, and W. Zeng, "Unleash the power of vision-language models by visual attention prompt and multi-modal interaction," *IEEE Transactions on Multimedia*, 2024. 2

[24] K. Zhang, Y. Yang, J. Yu, H. Jiang, J. Fan, Q. Huang, and W. Han, "Multi-task paired masking with alignment modeling for medical vision-language pre-training," *IEEE Transactions on Multimedia*, vol. 26, pp. 4706–4721, 2024. 2

[25] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96:1–96:6, 2007. 2, 3

[26] X. Zheng, Y. Luo, P. Zhou, and L. Wang, "Distilling efficient vision transformers from cnns for semantic segmentation," *Pattern Recognition*, vol. 158, p. 111029, 2025. 2

[27] J. Zhou, X. Zheng, Y. Lyu, and L. Wang, "Eventbind: Learning a unified representation to bind them all for event-based open-world understanding," in *European Conference on Computer Vision*. Springer, 2024, pp. 477–494. 2

[28] X. Zheng, Y. Lyu, L. Jiang, J. Zhou, L. Wang, and X. Hu, "Magic++: Efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection," *arXiv preprint arXiv:2412.16876*, 2024. 2

[29] X. Zheng, H. Xue, J. Chen, Y. Yan, L. Jiang, Y. Lyu, K. Yang, L. Zhang, and X. Hu, "Learning robust anymodal segmentor with unimodal and cross-modal distillation," *arXiv preprint arXiv:2411.17141*, 2024. 2, 3

[30] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma, "Mrfs: Mutually reinforcing image fusion and segmentation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26964–26973. 2, 6, 9

[31] C. Zhu, B. Xiao, L. Shi, S. Xu, and X. Zheng, "Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts," *arXiv preprint arXiv:2412.04220*, 2024. 3

[32] J. Zhao, F. Teng, K. Luo, G. Zhao, Z. Li, X. Zheng, and K. Yang, "Unveiling the potential of segment anything model 2 for rgb-thermal semantic segmentation with language guidance," *arXiv preprint arXiv:2503.02581*, 2025. 3

[33] D. Zhong, X. Zheng, C. Liao, Y. Lyu, J. Chen, S. Wu, L. Zhang, and X. Hu, "Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation," *arXiv preprint arXiv:2503.07098*, 2025. 3

[34] Y. Xu, J. Tang, A. Men, and Q. Chen, "Eviprompt: A training-free evidential prompt generation method for adapting segment anything model in medical images," *IEEE Transactions on Image Processing*, vol. 33, pp. 6204–6215, 2024. 3

[35] Y. Zhang and Z. Shen, "Unleashing the potential of sam2 for biomedical images and videos: A survey," *arXiv preprint arXiv:2408.12889*, 2024. 3

[36] Y. Zhang and R. Jiao, "Towards segment anything model (sam) for medical image segmentation: a survey," *arXiv preprint arXiv:2305.03678*, 2023. 3

[37] "Segment anything model for medical image segmentation: Current applications and future directions," *Computers in Biology and Medicine*, vol. 171, p. 108238, 2024. 3

[38] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3

[39] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023. 3

[40] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[41] Z. Zhang, Z. Wei, S. Zhang, Z. Dai, and S. Zhu, "Uvosam: A mask-free paradigm for unsupervised video object segmentation via segment anything model," *ArXiv*, vol. abs/2305.12659, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258833355 3

[42] X. Lan, J. Lyu, H. Jiang, K. Dong, Z. Niu, Y. Zhang, and J. Xue, "Foodsam: Any food segmentation," *IEEE Transactions on Multimedia*, pp. 1–14, 2023. 3

[43] Y. He, W. Chen, S. Wang, T. Liu, and M. Wang, "Recalling unknowns without losing precision: An effective solution to large model-guided open world object detection," *IEEE Transactions on Image Processing*, vol. 34, pp. 729–742, 2025. 3

[44] P. Liu, J. Deng, L. Duan, W. Li, and F. Lv, "Segmenting anything in the dark via depth perception," *IEEE Transactions on Multimedia*, pp. 1–12, 2025. 3

[45] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, "Segment and track anything," *arXiv preprint arXiv:2305.06558*, 2023. 3

[46] Y. Mei, J. Sun, Z. Peng, F. Deng, G. Wang, and J. Chen, "Rogsam: A language-driven framework for instance-level robotic grasping detection," *IEEE Transactions on Multimedia*, pp. 1–13, 2025. 3

[47] S. Zhang, D. Kong, Y. Xing, Y. Lu, L. Ran, G. Liang, H. Wang, and Y. Zhang, "Frequency-guided spatial adaptation for camouflaged object detection," *IEEE Transactions on Multimedia*, vol. 27, pp. 72–83, 2025. 3

[48] X. Wei, R. Zhang, J. Wu, J. Liu, M. Lu, Y. Guo, and S. Zhang, "Nto3d: Neural target object 3d reconstruction with segment anything," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 20352–20362. 3

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International conference on machine learning (ICML)*, 2021, pp. 8748–8763. 3

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International conference on machine learning (ICML)*, 2015, pp. 2048–2057. 3

[51] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18392–18402. 3

[52] Y. Lyu, X. Zheng, J. Zhou, and L. Wang, "Unibind: Llm-augmented unified and balanced representation space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26752–26762. 3

[53] Y. Lyu, X. Zheng, D. Kim, and L. Wang, "Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all," *arXiv preprint arXiv:2405.16108*, 2024. 3

[54] J. Fu, W. Zhou, Q. Jiang, H. Liu, and G. Zhai, "Vision-language consistency guided multi-modal prompt learning for blind ai generated image quality assessment," *IEEE Signal Processing Letters*, vol. 31, pp. 1820–1824, 2024. 3

[55] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, and M. N. Penchalaiah, "Dall-e: Creating images from text," *UGC Care Group I Journal*, vol. 8, no. 14, pp. 71–75, 2021. 3

[56] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774 3

[57] Z. Zhang, X. Han, X. Song, Y. Yan, and L. Nie, "Multi-modal interaction graph convolutional network for temporal language localization in videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 8265–8277, 2021. 3

[58] C. Shang, Z. Song, H. Qiu, L. Wang, F. Meng, and H. Li, "Prompt-driven referring image segmentation with instance contrasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4124–4134. 3

[59] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 676–11 685. 3

[60] S. Dong, W. Zhou, C. Xu, and W. Yan, "Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 657–669, 2024. 6, 7

[61] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2633–2642. 6, 7

[62] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 4467–4473. 6, 7, 9

[63] W. Zhou, T. Gong, J. Lei, and L. Yu, "Dbcnet: Dynamic bilateral cross-fusion network for rgb-t urban scene understanding in intelligent vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 12, pp. 7631–7641, 2023. 6

[64] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for rgb-thermal perception tasks," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 4060–4067, 2023. 6, 7

[65] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021. 6

[66] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Mmsformer: Multimodal transformer for material and semantic segmentation," *IEEE Open Journal of Signal Processing*, pp. 1–12, 2024. 6, 7

[67] U. Shin and J. Lee, Kyunghyun and, "Complementary random masking for rgb-thermal semantic segmentation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 110–11 117. 6

[68] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for rgb-thermal semantic segmentation," 2024, pp. 11 110–11 117. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10611200 6, 7

[69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019. 7

[70] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *International Conference on Computer Vision*, 2023. 7, 9

[71] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 14 679–14 694, 2023. 7

[72] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 1136–1147. 7

[73] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE TIP*, vol. 30, pp. 7790–7802, 2021. 9

[74] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE TCSVT*, 2022. 9

[75] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb thermal scene parsing," *AAAI*, 2022. 9

[76] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "Didfuse: Deep image decomposition for infrared and visible image fusion," *IJCAI*, 2020. 9

[77] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "Reconet: Recurrent correction network for fast and efficient multi-modality image fusion," in *ECCV*. Springer, 2022, pp. 539–555. 9

[78] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE TPAMI*, 2020. 9

[79] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *IEEE/CVF CVPR*, 2022, pp. 5802–5811. 9

[80] B. Li, D. Zhang, Z. Zhao, J. Gao, and X. Li, "U3m: Unbiased multiscale modal fusion model for multimodal semantic segmentation," *arXiv preprint arXiv:2405.15365*, 2024. 9