2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11247613

HFSENet: Hierarchical Fusion Semantic Enhancement Network for RGB-T Semantic Segmentation in Annealing Furnace Operation Area

Haoyu Yuan[1], Lin Zhang[1], Runjiao Bao[1], Jinge Si[1], Shoukun Wang[1] and Tianwei Niu[1]

***Abstract*— Regular temperature measurement of critical parts of an annealing furnace has always been a difficult task. Due to the harsh environment of high temperature, high noise, and darkness in the annealing furnace operation area, unmanned vehicles equipped with the RGB-T semantic segmentation model are usually adopted in most factories for inspection. However, existing RGB-T semantic segmentation models usually rely on good lighting or thermal conditions, which are generally difficult to fulfill in annealing furnace operation areas. In this paper, we propose a new hierarchical fusion-based semantic enhancement network, HFSENet. We first adopt the two-stream structure and the siamese structure to extract the low-level and high-level features of unimodal modalities, respectively. Then, considering the differences between the features in different hierarchical levels, we introduce a novel low-level feature spatial fusion module and a high-level feature channel fusion module to perform the multi-modal feature hierarchical fusion. On this basis, we also propose the semantic feature complementary enhancement module, which utilizes the appearance information set and object information set extracted from RGB and thermal infrared (TIR) branches to enhance the fused features and give them more semantic information. Finally, segmentation results with refined edges are obtained by an edge refinement decoder that includes a local search extraction module. The unmanned inspection vehicle we built with the proposed HFSENet has successfully passed the test, and the recognition performance of the four targets exceeds the current state-of-the-art (SOTA) method on our homemade annealing furnace operation area dataset.**

I. INTRODUCTION

In the steel smelting annealing process, regularly monitoring the temperature of critical parts of annealing furnaces is necessary to ensure high-quality production and a safety requirement. However, the annealing furnace environment is typically characterized by high temperatures, darkness, intense noise, confined spaces, and risks of collapse. Currently, inspections primarily rely on manual operations, where workers use handheld infrared thermometers to measure temperature at specific places, resulting in low efficiency, insufficient accuracy, and posing health risks to the workers. With the development of unmanned systems and deep learning technologies, robots are gradually replacing humans in inspection tasks. Robots capture images of the annealing furnace, use semantic segmentation models to locate critical parts, and then measure the temperature using infrared cameras, enabling automated inspections. Traditional semantic segmentation primarily targets RGB images [1], but its performance

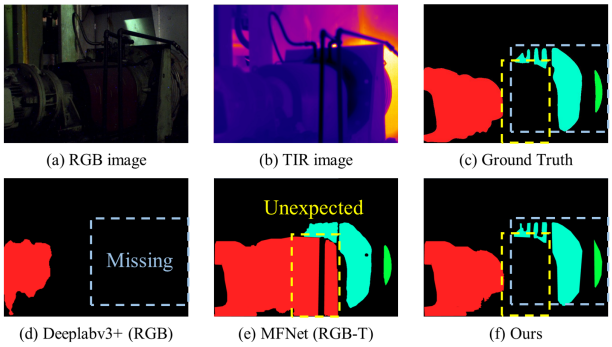

Fig. 1. The work area of an annealing furnace under poor lighting and thermal conditions, with the TIR image presented in pseudo-color. (d) The segmentation result misses a large portion of the actual regions (Blue dashed box). (e) The segmentation result contains some regions that should not be present (Yellow dashed box).

significantly degrades under the harsh lighting conditions of an annealing furnace (see Fig. 1(d)). RGB-T semantic segmentation technology enhances the model's segmentation capabilities in low-light environments by introducing thermal infrared images (TIR) as a complement to the RGB images [2], [3], whose core lies in designing efficient multimodal fusion modules [4], [5], [6], [7].

However, the fusion of multimodal information also brings new challenges. On the one hand, as pointed out by several studies [3], [8], [9], most existing methods use the same fusion strategy to process both low-level and high-level features from a single modality, neglecting the differences between features at different layers [10], [11], [12], [13]. This leads to redundant features and insufficient utilization of cross-modal information. Consequently, this method wastes valuable computational resources for inspection robots, as these systems also need to allocate computational power for navigation and motion control. On the other hand, many studies [9], [14] have shown that RGB images provide rich texture and color information under good lighting conditions, while thermal images capture more target information in poor lighting conditions. By leveraging complementary learning from both modalities, the robustness of the model can be improved [15], [16]. However, in complex scenes like shown in Fig. 1(b), structures such as pipelines and protective covers block each other, and the temperatures are similar. The RGB images lack distinct textures, and the TIR images have blurred boundaries. This can easily cause the model to detect areas that should not be measured (see Fig. 1(e)), affecting the accuracy of the inspection. Moreover, if the

*This study was supported by the National Natural Science Foundation of China (62473044) and BIT Research Innovation Project (2024YCXY037)

[1]Tianwei Niu (Corresponding author), Haoyu Yuan, Lin Zhang, Runjiao Bao, Jinge Si, Shoukun Wang are with School of Automation, Beijing Institute of Technology, 100081, China. Email: ntwbit@bit.edu.cn

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

nature of multimodal data is not considered, the model is prone to suboptimal solutions, i.e., over-reliance on a single modality. Therefore, achieving robust and efficient semantic segmentation under the potential degradation of both RGB and TIR data remains a critical challenge for annealing furnace inspection robots.

To address the above problems, we propose a novel hierarchical fusion semantic enhancement network (HFSENet), which aims to fully utilize feature information from different modal data and different fusion stages. In short, we perform a hierarchical fusion of multimodal features using the low-level feature spatial fusion (LFSF) and high-level feature channel fusion (HFCF) modules, respectively, and then utilize the semantic feature complementary enhancement (SFCE) module to enhance the semantic information of the fused features and the robustness of the model through the bootstrapping of object information set (OIS) and appearance information set (AIS). Finally, an edge refinement decoder (ERD) generates a refined edge segmentation result. To the best of our knowledge, we are the first RGB-T segmentation model designed for complex annealing furnace operating areas with harsh lighting and thermal conditions, and it has now been successfully implemented in actual production.

Our contributions are summarized as follows:

- We propose the HFSENet for RGB-T semantic segmentation of critical parts in annealing furnaces. Extensive experiments demonstrate its superior accuracy and performance compared to state-of-the-art methods.
- We introduce the LFSF and HFCF modules for fusing low- and high-level multimodal features, fully exploiting the complementarity of multimodal data and reducing information redundancy.
- The SFCE module enhances fused features through semantic guidance from OIS and AIS, eliminating noise and avoiding reliance on a single modality.
- We propose an edge refinement decoder that includes the local search extraction (LSE) module, further refining the edges of segmented objects.

II. RELATED WORK

A. RGB Semantic Segmentation

In recent years, with the introduction of Fully Convolutional Networks (FCN) [1], CNN-based semantic segmentation models have achieved remarkable success. Ronneberger et al. [17] proposed U-Net, which employs skip connections to fuse features from the encoder and decoder for precise semantic segmentation. Subsequently, Chen et al. [18] introduced Atrous Spatial Pyramid Pooling (ASPP) in DeepLabV2, utilizing dilated convolutions to expand the receptive field. To address the diversity of object scales, Zhao et al. [19] proposed PSPNet, which captures global contextual information through multiple pooling layers. Yu et al. [20] developed BiSeNet, which preserves spatial information via a spatial path while expanding the receptive field through a context path. Recently, the introduction of self-attention mechanisms [21] has further enhanced segmentation performance. For instance, Fu et al. [22] proposed DANet, which constructs spatial and channel attention modules. Li et al. [23] interactively explore spatial and channel contextual information to capture semantic features better. Additionally, studies emphasizing multi-scale and local-global feature aggregation [24], [25], [26] have provided valuable insights into the advancement of semantic segmentation. Despite the significant progress in RGB-based semantic segmentation methods, their performance often deteriorates in complex backgrounds and under challenging lighting conditions. Consequently, researchers have explored multimodal semantic segmentation as a promising solution.

B. RGB-T Semantic Segmentation

RGB-thermal (RGB-T) semantic segmentation networks have been successfully proposed to overcome the limitations of RGB segmentation networks. Currently, most RGB-T fusion networks focus on designing multimodal fusion modules, which can be broadly classified into three categories: naive feature-level fusion [10], [27], multi-scale feature fusion [6], [11], [13], [28], and attention-weighted fusion [4], [5], [14], [29]. Among the most representative methods, MFNet [2] was the first network to use a two-stream structure to fuse features from both RGB and thermal infrared progressively modalities. Yuan et al. [30] proposed a strategy that separately enhances the fusion of low-level and high-level unimodal features. Wu et al. [29] designed a cascaded cross-modal fusion module to adaptively select and process complementary information from RGB and thermal features. Dong et al. [31] incorporated prior edge information further to refine object boundaries in the semantic segmentation map. However, most of these methods fail to utilize the complementary information between different modalities effectively. During training, the network may overly rely on one modality, thereby suppressing meaningful representations from the other modality. Therefore, fully utilizing complementary information between modalities is crucial for achieving accurate and robust object segmentation.

III. METHOD

Fig. 2 illustrates the overall structure of the proposed HFSENet. We employ two parallel backbone networks to extract multi-level features from RGB and thermal images, respectively. Subsequently, the LFSF and HFCF modules perform hierarchical fusion of these features. Under the semantic guidance of AIS and OIS, the SFCE module further enhances the fused features. Finally, the ERD module generates the segmentation results and refines the boundary contours. The details will be further described in the following sections.

A. Unimodal Feature Extraction

As shown in Fig. 2, the proposed HFSENet adopts a symmetric encoder architecture to extract features from both thermal infrared (TIR) and visible light (RGB) modalities. The encoder is based on ResNet-50 [32], with the initial layer serving as the first layer for feature extraction, and the first three residual convolutional blocks used as the 2nd to 4th layers. Considering the significant differences

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

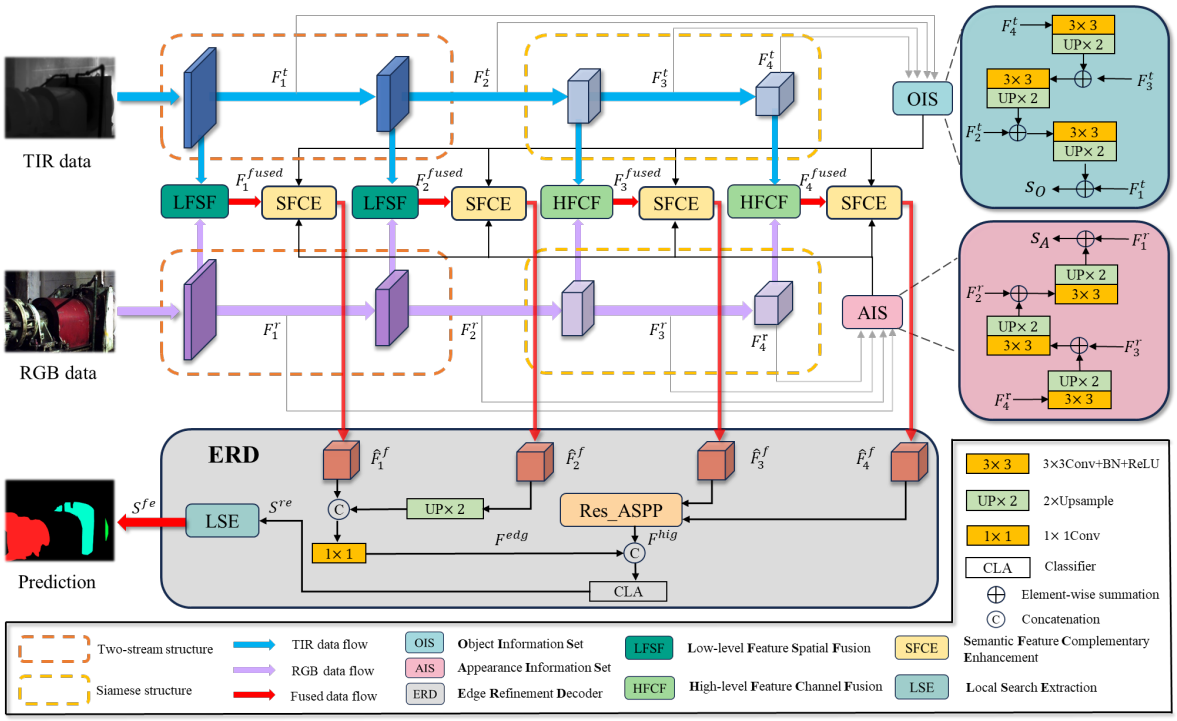

Fig. 2. General framework of our proposed HFSENet. The LFSF and HFCF modules are used to fuse unimodal features at different levels. the SFCE module enhances the fused features using OIS and AIS. The ERD containing the LSE module is used to generate segmentation results with refined edges.

between low-level RGB and TIR features and the stronger semantic consistency in high-level features, the model uses a two-stream structure during the low-level feature extraction stages (layer1 and layer2). In this structure, the RGB and TIR branches share the same architecture but with different parameters, allowing the capture of unique information from each modality. A siamese structure is adopted in the high-level feature extraction stages (layer3 and layer4), where the RGB and TIR branches share the same architecture and parameters. This design yields two sets of low-level features (RGB: $\{\mathbf{F}_i^r|i = 1,2\}$, TIR: $\{\mathbf{F}_i^t|i = 1,2\}$) and two sets of high-level features (RGB: $\{\mathbf{F}_i^r|i = 3,4\}$, TIR: $\{\mathbf{F}_i^t|i = 3,4\}$). This approach captures complementary information at the low-level stages and reduces feature redundancy at the high-level stages, thereby enhancing model efficiency and reducing complexity.

B. Cross-Modal Feature Fusion

As described in [11], [12], in paired RGB-T images, shallow features contain rich spatial details, while deep features carry category-discriminative information. Based on this, we propose the LFSF and HFCF modules to fuse unimodal features at different levels along the spatial and channel dimensions, respectively, to better capture spatial details and semantic category cues from multimodal inputs. As shown Fig. 3, both the LFSF and HFCF modules consist of two parts: two different modality fusion sections (light blue and light green backgrounds) and two identical distillation and refinement sections (pink backgrounds).

1) Modal Fusion: The LFSF module (Fig. 3(a)) operates solely on low-level RGB and TIR features. Specifically, we

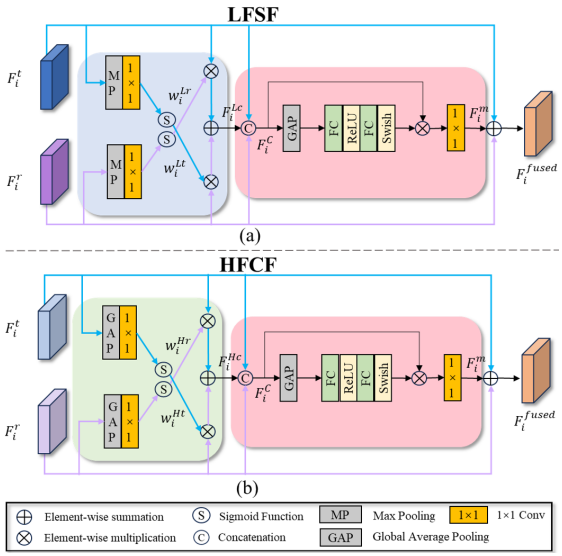

Fig. 3. Structure of our proposed LFSF and HFCF modules. LFSF is used to fuse low-level features, and HFCF is used to fuse high-level features. The light blue and light green regions represent different fusion parts, and the pink region represents the same distillation and purification parts.

first input the low-level TIR features $\{\mathbf{F}_i^t|i = 1,2\}$ into a max pooling (MP) layer to extract the most important features at each spatial location. Then, we calculate the spatial activation weights $\mathbf{w}_i^{Lt}$ of the current low-level TIR features through a convolution layer and a Sigmoid function. A similar approach is applied to the RGB features, obtaining the weights $\mathbf{w}_i^{Lr}$. Next, using a cross-multiplication method,

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

$\mathbf{w}_i^{Lt}$ is applied to activate the low-level RGB spatial features, and $\mathbf{w}_i^{Lr}$ is used to activate the low-level TIR spatial features. Finally, the activated and enhanced features are summed element-wise to obtain the low-level cross-modal fused intermediate feature $\mathbf{F}_i^{Lc}$. Through this cross-modal mutual activation approach along the spatial dimension, the low-level RGB and TIR features gradually absorb each other's useful spatial details to form a complement.

As shown in Fig. 3(b), the HFCF module operates on high-level RGB and TIR features. Unlike LFSF, the activation of high-level features is conducted along the channel dimension. Through this channel-wise cross-modal mutual activation, high-level RGB and TIR features can effectively embed their semantic information into each other.

2) Distillation and Purification: Although the cross-modal mutual activation along the spatial and channel dimensions fully utilizes the characteristics of different modalities, it may introduce noise in many cases [10], [14]. Therefore, after obtaining the cross-modal fused intermediate features, purifying the original and fused features is necessary. Specifically, as shown in the pink area of Fig. 3, we first concatenate the three inputs, $\mathbf{F}_i^{Hc}(\mathbf{F}_i^{Lc})$, $\mathbf{F}_i^t$, and $\mathbf{F}_i^r$, to obtain the concatenated feature $\mathbf{F}_i^c$. Then, $\mathbf{F}_i^c$ is re-weighted and modulated by element-wise multiplication with itself. The modulated feature $\mathbf{F}_i^m$, obtained via a 1×1 convolution, is then directly added to the original $\mathbf{F}_i^t$ and $\mathbf{F}_i^r$ to produce the final multimodal fused feature $\mathbf{F}_i^{\text{fused}}$.

C. Semantic Feature Complementary Enhancement Module

To accurately locate targets and minimize noise interference in $\mathbf{F}_i^{\text{fused}}$ by leveraging the rich appearance and object information present in both TIR and RGB images, we propose the Semantic Feature Complementary Enhancement (SFCE) module. Specifically, we first utilize the Object Information Set (OIS) and the Appearance Information Set (AIS) to extract distinct semantic information from TIR and RGB images, respectively. The structures of OIS and AIS are shown in the upper right corner of Fig. 2. These modules progressively convolve and upsample the unimodal features extracted from each encoder layer of TIR and RGB, ultimately adding them to $\mathbf{F}_1^t$ and $\mathbf{F}_1^r$, resulting in the object feature $\mathbf{S}_O$ and the appearance feature $\mathbf{S}_A$.

Since $\mathbf{S}_O$ and $\mathbf{S}_A$ contain foundational and rich semantic information about the same target, they are used to guide $\mathbf{F}_i^{\text{fused}}$, generated in the previous stage, to achieve stable feature-level semantic enhancement. As shown in Fig. 4, the SFCE module first applies convolution and normalization to the three inputs—$\mathbf{S}_O$, $\mathbf{S}_A$, and $\mathbf{F}_i^{\text{fused}}$—independently, followed by activation using the Sigmoid function to obtain their respective activation weights. These weights are then used for pairwise mutual modulation among the three, allowing for full integration of their features and enabling balanced semantic guidance between the TIR and RGB modalities. This ensures the fusion network is not overly dependent on a single modality. Once we obtain the enhanced features $\mathbf{F}_i^O$ (containing object semantics) and $\mathbf{F}_i^A$ (containing appearance semantics), we draw inspiration from Mix-FFN [33] and

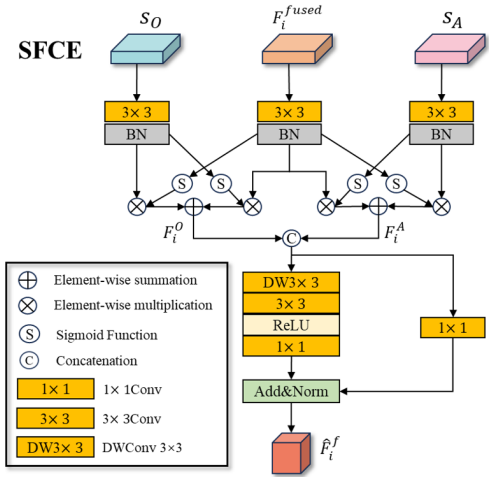

Fig. 4. Structure of the proposed semantic feature complementary enhancement module (SFCE).

ConvMLP [34] by using a simple channel embedding to fuse the features from the two pathways. We also incorporate a depth-wise convolution layer (DWConv 3×3) to construct a skip connection structure, which captures information from the surrounding regions to achieve more robust segmentation. Finally, we generate four distinct enhanced features $\{\hat{F}_i^f | i = 1, 2, 3, 4\}$, which are then used for feature decoding.

D. Edge Refinement Decoder

In detecting annealing furnaces, regions with different temperatures may be near each other. If the edge segmentation of the object is not fine enough, the measured temperature may have a significant error. Therefore, we improve on Res_ASPP [35] and propose an edge refinement decoder (ERD) to obtain a larger receptive field and finer edges. Specifically, our ERD receives four enhanced features of the encoder output. First, we fuse $\hat{\mathbf{F}}_4^f$ with $\hat{\mathbf{F}}_3^f$ to obtain $\mathbf{F}^{\text{hig}}$ via the original Res_ASPP. Next, we utilize $\hat{\mathbf{F}}_2^f$ and $\hat{\mathbf{F}}_1^f$, which contain detailed texture information, to obtain the low-level edge feature $\mathbf{F}^{\text{edg}}$ via edge supervision. Then, $\mathbf{F}^{\text{hig}}$ is connected to $\mathbf{F}^{\text{edg}}$ and input into the classifier, which outputs the rough edge segmentation graph $\mathbf{S}^{\text{re}}$. We re-input the obtained $\mathbf{S}^{\text{re}}$ into the newly designed Local Search Extraction (LSE) module to refine the edges further. First, we perform a local search, and for the vector $p\,(i, j, \cdot)$ with fixed i and j in the semantic segmentation output $\mathbf{S}^{\text{re}}$, we use the Softmax function to obtain the confidence of the predicted target $P\,(i, j, \cdot)$. For a fixed threshold T, the conditions for determining whether a pixel $\hat{P}$ is an edge based on the confidence $P\,(i, j, \cdot)$ are as follows:

$$\hat{P}\,(i,j) := \begin{cases} 1 & \text{Case1, 2, 3 holds} \\ \arg\max\limits_{k \in [0,C]} P\,(i,j,k) & \text{Otherwise} \end{cases}$$

$$\text{where} \quad \begin{aligned} &\text{Case 1}: \arg\max_{k \in [0,C]} P\,(i,j,k) = 0; \\ &\text{Case 2}: \arg\max_{k \in [1,C]} P\,(i,j,k) = 1; \\ &\text{Case 3}: \max_{0 \le k \le C} P\,(i,j,k) \le T. \end{aligned} \tag{1}$$

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

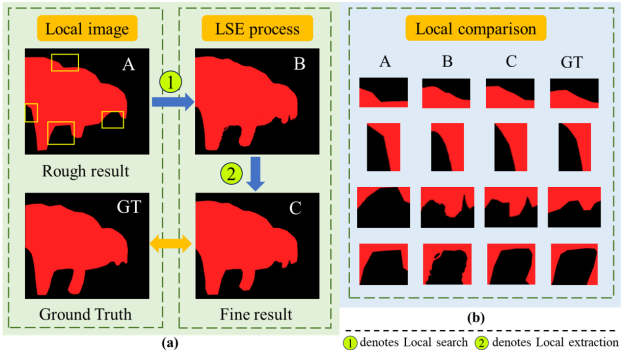

Fig. 5. Qualitative experiments on the effectiveness of the LSE module. (a) Visualization of the LSE module application process. (b) It shows edge comparisons of four typical sites at different processing stages.

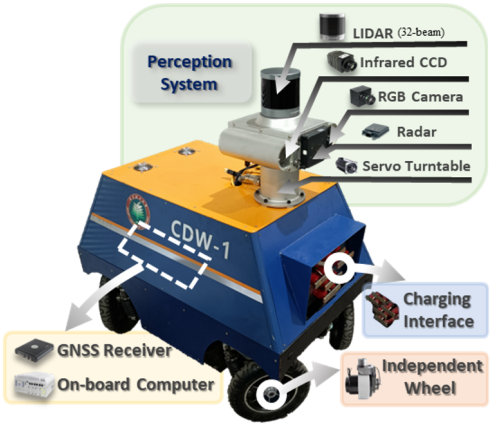

Fig. 6. Component structure of inspection robot.

Due to the high sensitivity of local search, pixels belonging to invisible edges and their neighboring pixels are also easily captured together. Therefore, we further removed the noise through local extraction. We introduce a local square window in each edge region to refine the local search results using the Otsu algorithm [36]. Subsequently, the results after binary segmentation are combined, and small noisy regions and holes are removed by a closure operation and connected-domain filtering, ultimately producing the refined segmentation result $\mathbf{S}^{\mathrm{fe}}$. Empirically, we set T to 0.7. The application process of LSE is shown in Fig. 5, which shows that the LSE module further improves the accuracy of edge segmentation based on the original segmentation accuracy.

E. Loss Function

Our overall loss function consists of two parts. The first part is the semantically supervised loss L_{IS} for the information sets OIS and AIS. We use the same weighted cross-entropy loss as in [37], denoted as:

$$L_{\mathrm{IS}} = \left(L_{\mathrm{wce}}\left(\mathbf{S}_O, \mathbf{S}^{\mathrm{true}}\right) + L_{\mathrm{wce}}\left(\mathbf{S}_A, \mathbf{S}^{\mathrm{true}}\right)\right)/2, \quad (2)$$

where L_{wce} is the weighted cross-entropy loss, and $\mathbf{S}^{\mathrm{true}}$ is the ground truth. The second part is the supervised loss of L_{ERD} to the ERD. It is worth noting that we only supervise the $\mathbf{S}^{\mathrm{re}}$ before edge refinement, i.e.:

$$L_{\mathrm{ERD}} = L_{\mathrm{lovasz}}\left(\mathbf{S}^{\mathrm{re}}, \mathbf{S}^{\mathrm{true}}\right), \quad (3)$$

where L_{lovasz} denotes the Lovasz-softmax loss. Therefore, the total loss function can be expressed as:

$$L_{\mathrm{total}} = \alpha L_{\mathrm{IS}} + \beta L_{\mathrm{ERD}}. \quad (4)$$

We empirically set $\{\alpha, \beta\}$ to $\{1, 3\}$ to make the network more focused on the output coarse segmentation results.

IV. EXPERIMENTS

A. Experimental Setup

1) Inspection Robots and Data Sets: The inspection robot we designed is shown in Fig. 6, which is a four-wheeled omnidirectional mobile AGV platform equipped with multiple sensors. Using the CMOS camera and infrared camera onboard the robot, we collected 700 pairs of raw images with a resolution of 1280×1024 from different furnace sections in Baosteel's No. 1 plant in Shanghai. Subsequently, we aligned the TIR images with the RGB images using the calibration method from [27] and manually annotated the ground truth labels for each RGB-T image pair at the pixel level using the Roboflow platform. After removing low-quality images, we constructed an annealing furnace dataset consisting of 676 aligned RGB-T image pairs, covering four object classes. The dataset was then split into training, validation, and test sets in a 7:2:1 ratio, with representative RGB-T image pairs and their ground truth labels shown in Fig. 7.

2) Training Setup: The experiments were conducted on an NVIDIA RTX 4060 GPU using the PyTorch framework, running on the Ubuntu 20.04 operating system with CUDA version 11.8. All models were trained and tested on the self-constructed annealing furnace dataset. During training, the input image size was set to 640×480 with a batch size of 4. We optimized the models using the AdamW optimizer [38], with an initial learning rate of 1e-4 and a weight decay rate of 1e-2. Additionally, we applied data augmentation techniques such as random color jittering, random horizontal flipping, and random cropping, and trained the models for 100 epochs. Mean accuracy (mAcc) and mean intersection over union (mIoU) were used as evaluation metrics.

B. Comparison With State-of-the-Art Methods

In this section, we compare our model with previous state-of-the-art RGB-T semantic segmentation networks (MFNet [2], RTFNet [10], PSTNet [27], FuseSeg [11], AFNet [4], ABMDRNet [5], FEANet [14], EGFNet [9], LASNet [16], SGFNet [39], MDRNet+ [13], MCOFNet [40]).

1) Qualitative and Quantitative Comparisons: The quantitative comparison results are shown in Table I. Compared with the previous state-of-the-art method (MDRNet+), our method improves the performance of mIoU metrics by 1.2%, proving the effectiveness of our proposed HFSENet. In addition, our methods achieve the best or second best performance on the Acc and IoU metrics for all four categories.

We also visualize the segmentation results of all methods, as shown in Fig. 7. our model is significantly enhanced in

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

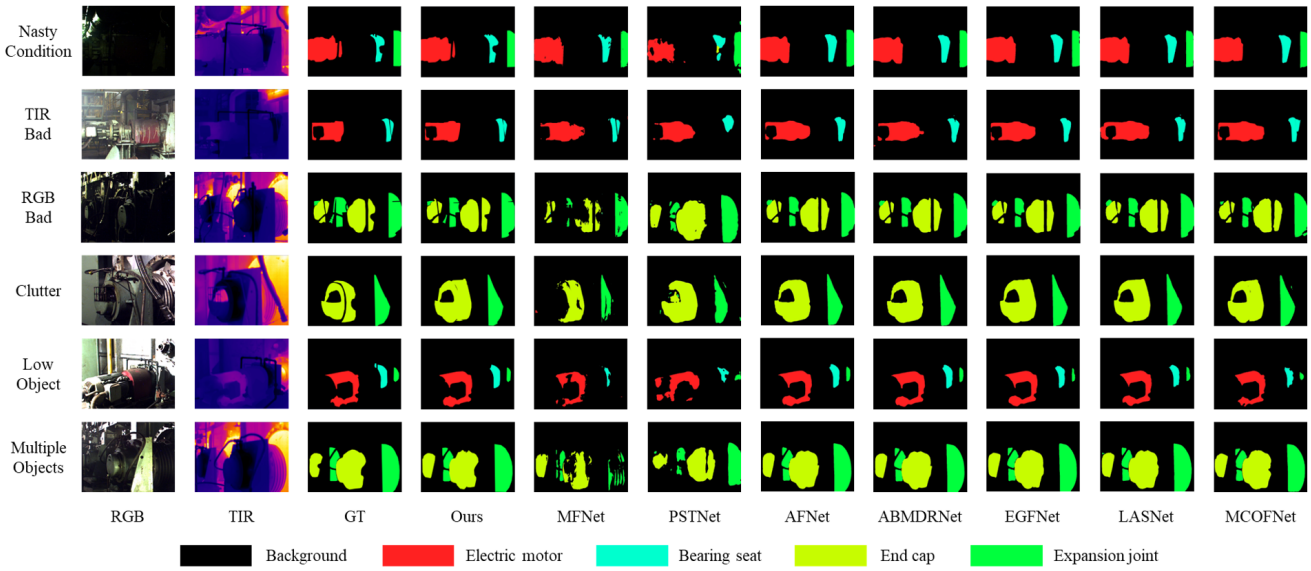

Fig. 7. Visual comparison of different models on the annealing furnace dataset. The TIR images are shown in pseudo-color.

TABLE I: Comparison of quantitative experimental results (%) for the annealing furnace data set. The best results in each column are marked in bold. The subscript of each method indicates its publication year.

Methods	Bearing seat		End Cap		Electric motor		Expansion joint		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$MFNet_{17}$ [2]	41.3	36.4	33.1	22.6	18.7	13.2	75.2	58.0	53.1	45.3
$RTFNet_{19}$ [10]	78.4	64.5	56.5	37.0	46.4	32.0	89.8	79.7	73.9	62.2
$PSTNet_{20}$ [27]	64.7	54.8	53.9	45.2	30.5	21.4	90.4	70.2	67.4	57.7
$FuseSeg_{20}$ [11]	71.3	66.1	61.8	51.5	54.6	35.6	88.7	80.1	74.9	66.1
$AFNet_{21}$ [4]	68.5	61.7	62.4	54.6	**62.3**	36.5	91.6	76.3	76.6	65.3
$ABMDRNet_{21}$ [5]	71.8	65.4	74.3	58.4	56.4	35.3	90.2	76.5	78.3	66.8
$FEANet_{21}$ [14]	76.0	65.4	78.5	57.1	53.7	37.7	95.1	81.2	80.4	67.9
$EGFNet_{22}$ [9]	72.8	65.9	61.6	55.6	58.2	35.1	87.9	77.5	75.8	66.4
$LASNet_{23}$ [16]	74.1	65.2	74.4	56.9	60.1	35.8	91.3	78.4	79.7	66.9
$SGFNet_{23}$ [39]	73.1	65.5	73.7	58.4	59.6	35.7	**95.3**	81.7	80.1	67.9
$MDRNet+_{24}$ [13]	74.2	66.3	**79.9**	56.9	57.5	38.1	93.4	81.2	80.8	68.1
$MCOFNet_{24}$ [40]	72.3	66.1	72.1	57.6	55.8	36.2	92.6	80.3	78.2	67.6
HFSENet(Ours)	**78.5**	**67.0**	74.8	**58.9**	61.8	**38.6**	94.3	**83.4**	**81.7**	**69.3**

segmenting key details and objects compared to previous methods. In particular, under both harsh lighting and thermal conditions (line 1), our method still segments the target object more completely with fewer missing parts. In addition, both in poor lighting and thermal environments (line 2, 3), our method can accurately capture the segmented objects without including too many irrelevant parts, which proves the robustness of the model. Finally, in many complex and special scenes (the last three rows), our method also separates different objects well and presents finer edges.

The above results show that our proposed network is able to fuse semantic features from different modalities well, exhibits strong robustness even in the challenging annealing furnace workspace scenario, and achieves accurate semantic segmentation performance from both RGB and TIR.

TABLE II: Comparison of different model parametric quantities and FLOPS. – denotes not provided.

Methods	FLOPs (G)	Params(M)	mIoU (%)
$MFNet_{17}$ [2]	8.39	0.72	45.3
$RTFNet_{19}$ [10]	290.61	254.51	62.2
$PSTNet_{20}$ [27]	129.37	20.38	57.7
$FuseSeg_{20}$ [11]	193.40	141.52	66.1
$ABMDRNet_{21}$ [5]	194.33	64.60	66.8
$LASNet_{23}$ [16]	233.81	93.58	66.9
$SGFNet_{23}$ [39]	147.73	125.25	67.9
$MDRNet+_{24}$ [13]	194.32	64.60	68.1
$MCOFNet_{24}$ [40]	-	258.04	67.6
HFSENet(Ours)	178.46	89.73	**69.3**

2) Comparison of Model Complexity: As shown in Table II, we give parametric quantities and performance comparisons between several prior models and our model. The image input size for all models is 640×480. Compared with the previous mainstream methods, our model achieves a proper compromise between improving segmentation accuracy and reducing model complexity (only 89.73M). This shows that our proposed model can better conserve the limited arithmetic power and ensure the proper operation of other inspection functions.

C. Ablation Studies

We conducted extensive ablation experiments on our self-constructed annealing furnace dataset to verify the effectiveness of each component in HFSENet. The ablation study includes the following three aspects.

1) The Effectiveness of LFSF, HFCF, and SFCE Modules: The experimental results are presented in Table III. *BS* represents the baseline model, which is constructed by removing the LFSF, HFCF, and SFCE modules from our proposed HFSENet while retaining the ERD. As shown in Table III, both the LFSF and HFCF modules for hierarchical feature

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

fusion and the SFCE module for enhancing complementary information contribute to improving the performance of the baseline model. Furthermore, the results of NO.2 and NO.3 indicate that feature fusion has a more significant impact on semantic segmentation compared to feature enhancement. With the combined effect of the LFSF, HFCF, and SFCE modules, our model achieves a 4.5% improvement in mIoU compared to the baseline model.

TABLE III: Ablation study results of different modules.

No.	Methods	mAcc(%)	mIoU(%)
1	BS	71.5	64.8
2	$BS + SFCE$	74.6	66.5
3	$BS + LFSF + HFCF$	78.4	67.6
4	$BS + LFSF + HFCF + SFCE$	**81.7**	**69.3**

TABLE IV: Ablation study results of different feature fusion structures. $*^L$ and $*^H$ represent modules used for low-level and high-level feature fusion, respectively. Sum denotes simple element-wise summation.

No.	Methods	mAcc(%)	mIoU(%)
1	$BS + Sum^L + Sum^H$	71.5	64.8
2	$BS + LFSF^L + Sum^H$	78.6	67.8
3	$BS + Sum^L + LFSF^H$	75.3	66.9
4	$BS + HFCF^L + Sum^H$	76.2	67.1
5	$BS + Sum^L + HFCF^H$	75.9	67.5
6	$BS + LFSF^L + LFSF^H$	80.4	68.4
7	$BS + HFCF^L + HFCF^H$	79.8	68.2
8	$BS + LFCF^L + HFCF^H$	**81.7**	**69.3**

TABLE V: Ablation study results for different components of each module. w/o denotes "without".

Module	Methods	mAcc(%)	mIoU(%)
LFSF	$w/o\ SF$	78.8	68.0
	$w/o\ DP$	77.9	68.2
HFCF	$w/o\ CF$	79.3	68.1
	$w/o\ DP$	78.2	68.4
SFCE	$w/o\ OIS$	80.3	68.9
	$w/o\ AIS$	79.5	68.4
LRD	$w/o\ LSE$	80.4	68.5
	MDRNet+ + LSE	81.5	68.6
	MCOFNet + LSE	79.6	68.2
	MDRNet+	80.8	68.1
	MCOFNet	78.2	67.6
	HFSENet(Ours)	**81.7**	**69.3**

2) The Effectiveness of the Hierarchical Feature Fusion Structure: We conducted a comprehensive quantitative comparison of different feature fusion strategies. The experimental results are shown in Table IV. As seen from NO.2 and NO.3, the LFSF module, which activates across the spatial dimension, is more suitable for fusing cross-modal low-level features. Similarly, from NO.4 and NO.5, it can be observed that the HFSF module, which activates across the channel dimension, is more suitable for fusing cross-modal high-level features. Additionally, NO.6, NO.7, and NO.8 further demonstrate that using improper activation methods for low-level and high-level features can impair the model's overall performance. In contrast, the hierarchical feature fusion method we proposed, "$BS + LFCF^L + HFCF^H$", fully considers the differences between high- and low-level multimodal features, enabling more effective extraction of useful information from features at different levels.

3) The Effectiveness of Different Components within Each Module: To verify the effectiveness of the specially designed components within each module, we conducted ablation experiments on different parts of the modules. The experimental results are presented in Table V. The first two columns of the table show four variants of the LFSF and HFCF modules: removing the spatial fusion component ($w/o\ SF$), removing the channel fusion component ($w/o\ CF$), and removing the distillation purification component ($w/o\ DP$) separately. The results indicate that removing the fusion components significantly degrades model performance (-1.3% mIoU). This may be because the model can no longer effectively utilize low-level spatial details and high-level semantic details. Furthermore, removing the distillation purification component weakens the ability to eliminate redundant features, leading to a performance drop.

The third column of the table compares different information sets within the SFCE module. The results demonstrate that both OIS and AIS are essential components, as they provide additional semantic information for object segmentation. Notably, AIS, which supplies appearance information, proves more effective for segmentation than OIS, which provides object-related information.

The fourth column of the table validates the role of the LSE module in LRD. After removing the LSE module results in a 0.8% decrease in mIoU, highlighting the importance of refining boundary delineation for improving segmentation performance. Additionally, we integrated the LSE module into other models for testing, and the results show that adding the LSE module consistently enhances the performance of each model, further demonstrating the generality and effectiveness of the LSE module.

V. CONCLUSIONS

In this paper, a novel RGB-T semantic segmentation method, HFSENet, is proposed for the annealing furnace operation area. To address the variability of modal features at different layers, LFSF and HFCF modules are designed for feature fusion, respectively. To cope with environments with undesirable lighting and thermal conditions, the SFCE module based on OIS and AIS guidance is designed for feature enhancement. In addition, an edge refinement decoder containing the LSE module was designed to obtain fine-edged segmentation results. Extensive experiments on our self-constructed annealing furnace operating area dataset show that HFSENet achieves optimal segmentation performance compared to 12 state-of-the-art methods.

Finally, in real industrial scenarios, many critical objects have limited sample availability or are difficult to fully enumerate, making traditional learning methods that rely on large amounts of manually annotated data less effective. Therefore, in the future, we will further explore few-shot and zero-shot learning approaches to enhance the robot's autonomous recognition capability in complex environments.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] 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)*. IEEE, 2017, pp. 5108–5115.

[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[4] J. Xu, K. Lu, and H. Wang, "Attention fusion network for multi-spectral semantic segmentation," *Pattern Recognition Letters*, vol. 146, pp. 179–184, 2021.

[5] 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*, 2021, pp. 2633–2642.

[6] S. Zhao and Q. Zhang, "A Feature Divide-and-Conquer Network for RGB-T Semantic Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2892–2905, 2023.

[7] X. He, M. Wang, T. Liu, L. Zhao, and Y. Yue, "SFAF-MA: Spatial Feature Aggregation and Fusion With Modality Adaptation for RGB-Thermal Semantic Segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023.

[8] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, "Encoder-decoder with cascaded CRFs for semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1926–1938, 2020.

[9] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3571–3579.

[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.

[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.

[12] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "GCNet: Grid-like context-aware network for RGB-thermal semantic segmentation," *Neurocomputing*, vol. 506, pp. 60–67, 2022.

[13] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating Modality Discrepancies for RGB-T Semantic Segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 9380–9394, 2024.

[14] 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.

[15] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[16] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T Semantic Segmentation With Location, Activation, and Sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1223–1235, 2023.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[20] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[21] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[22] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 646–662.

[23] Z. Li, Y. Sun, L. Zhang, and J. Tang, "CTNet: Context-Based Tandem Network for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9904–9917, 2022.

[24] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-Supervised Learning for Time Series Analysis: Taxonomy, Progress, and Prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6775–6794, 2024.

[25] J. Gao, L. Zhao, and X. Li, "NWPU-MOC: A Benchmark for Fine-Grained Multicategory Object Counting in Aerial Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[26] J. Gao, H. Yang, D. Zhang, Y. Yuan, and X. Li, "Imbalanced Aircraft Data Anomaly Detection," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–11, 2024.

[27] 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)*. IEEE, 2020, pp. 9441–9447.

[28] 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*, 2023.

[29] W. Wu, T. Chu, and Q. Liu, "Complementarity-aware cross-modal feature fusion network for RGB-T semantic segmentation," *Pattern Recognition*, vol. 131, p. 108881, 2022.

[30] J. Yuan, T. Wang, G. Huo, R. Jin, and L. Wang, "Semantic segmentation algorithm fusing infrared and natural light images for automatic navigation in transmission line inspection," *Electronics*, vol. 12, no. 23, p. 4810, 2023.

[31] 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.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[34] J. Li, A. Hassani, S. Walton, and H. Shi, "Convmlp: Hierarchical convolutional mlps for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6307–6316.

[35] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and unimodal features for RGB-D salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 2428–2441, 2020.

[36] N. Ostu, "A threshold selection method from gray-level histograms." *IEEE Trans SMC*, vol. 9, p. 62, 1979.

[37] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

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

[39] Y. Wang, G. Li, and Z. Liu, "SGFNet: Semantic-Guided Fusion Network for RGB-Thermal Semantic Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7737–7748, 2023.

[40] J. Zhang, R. Zhang, W. Yuan, and Y. Liu, "RGB-T semantic segmentation based on cross-operational fusion attention in autonomous driving scenario," *Evolving Systems*, vol. 15, no. 4, pp. 1429–1440, 2024.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 28,2025 at 14:37:55 UTC from IEEE Xplore. Restrictions apply.