Contents lists available at ScienceDirect

Biomimetic Intelligence and Robotics

journal homepage: www.keaipublishing.com/en/journals/biomimetic-intelligence-and-robotics/

Research Article

HAPNet: Toward superior RGB-thermal scene parsing via hybrid, asymmetric, and progressive heterogeneous feature fusion

Jiahang Li [a], Peng Yun [b], Yang Xu [c], Ye Zhang [c], Mingjian Sun [d,e,f,g], Qijun Chen [a], Ilin Alexander [h], Rui Fan [a,*]

[a] College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, and State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai 201804, China
[b] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China
[c] MSU-BIT-SMBU Joint Research Center of Applied Mathematics, Shenzhen MSU-BIT University, Shenzhen 518172, China
[d] Qingdao Innovation and Development Base, Harbin Institute of Technology, Qingdao 266109, China
[e] Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China
[f] Harbin Institute of Technology, Weihai 264200, China
[g] Suzhou Research Institute, Harbin Institute of Technology, Suzhou 215000, China
[h] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow 119991, Russia

ARTICLE INFO

Article history:
Received 9 September 2025
Revised 30 November 2025
Accepted 5 January 2026
Available online 26 February 2026

Keywords:
Data-fusion
Thermal
Scene parsing
Heterogeneous feature
Vision foundation model

ABSTRACT

Data-fusion networks have shown significant promise for RGB-thermal scene parsing. However, the majority of existing studies have relied on symmetric duplex encoders for heterogeneous feature extraction and fusion, paying inadequate attention to the inherent differences between RGB and thermal modalities. Recent progress in vision foundation models (VFMs), which leverage self-supervised learning on large-scale unlabeled datasets, has exhibited superior capabilities in extracting informative, general-purpose features compared to supervised encoders. However, their potential has yet to be fully leveraged in the domain. In this study, we take one step toward this new research area by exploring a feasible strategy to fully exploit VFM features for RGB-thermal scene parsing. Specifically, we delve deeper into the unique characteristics of RGB and thermal modalities, thereby designing a hybrid, asymmetric encoder that incorporates both a VFM and a cross-modal spatial prior descriptor (CSPD), enabling enhanced extraction of complementary heterogeneous features. The extracted features undergo dual-path feature fusion through our proposed progressive heterogeneous feature integrators. Moreover, we introduce an auxiliary task to further enrich the local semantics of fused features, thereby improving the overall performance of RGB-thermal scene parsing. Our proposed HAPNet, incorporating all these components, delivers superior performance under challenging illumination conditions. Extensive experiments demonstrate that HAPNet outperforms all other state-of-the-art methods, with improvements of 0.1%, 1.0%, and 2.4% in mIoU on three public RGB-thermal scene parsing datasets: MFNet, PST900, and KP Day-Night, respectively. Additionally, our method exhibits exceptional generalizability for RGB-HHA scene parsing. We believe this new paradigm has opened up new opportunities for future developments in data-fusion scene parsing approaches. The source code is publicly available at https://mias.group/HAPNet/.

1. Introduction

1.1. Background

Unity in diversity strengthens perception. RGB-thermal (often abbreviated as RGB-T) scene parsing has emerged as a crucial feature in autonomous vehicles and mobile robots [1]. RGB images capture visible light, while thermal images capture heat signatures. The fusion of these two modalities of data, especially during the feature encoding stage, has been proven to dramatically enhance the reliability and robustness of scene parsing [2–4]. Nevertheless, current data-fusion approaches [5–12] generally employ duplex [13–17] encoders (two identical pre-trained hierarchical backbone networks with independent trainable parameters) to indiscriminately extract heterogeneous features from the RGB-T data, limiting their ability to fully exploit the advantages of both data modalities [18]. Therefore, this study aims to explore a more ingenious and effective strategy for heterogeneous feature extraction and fusion, with a specific focus on leveraging VFMs, as exemplified by BEiT series [19,20] and DINOv2 [21], which

* Corresponding author.
E-mail address: rui.fan@ieee.org (R. Fan).

have garnered considerable attention in recent computer vision research endeavors.

1.2. Existing challenges and motivation

The backbone networks in duplex encoders are commonly pre-trained on the ImageNet database [22] (containing millions of annotated natural images) through fully supervised learning. Although the ImageNet database has enabled these networks to closely approximate real-world data distributions, its limited dataset size restricts these networks from fully exploiting the vast amount of unlabeled data available worldwide [19]. Unfortunately, designing encoders based on VFMs, pre-trained in a self/un-supervised manner on extensive unlabeled data, for heterogeneous feature extraction remains largely unexplored in the domain of RGB-T scene parsing.

Additionally, while symmetric duplex encoders [9,23,24] are prevalently used, they may not be the optimal architectures for heterogeneous feature extraction. This limitation arises from the substantial inherent modality differences [25] between the RGB and thermal images, as well as the unique advantages provided by vision Transformers (ViTs) and convolutional neural networks (CNNs). RGB images, rich in global semantic cues [26], are well-suited for learning through ViTs, while both RGB and thermal images provide gradient-related details (local semantics) [8], which can be more effectively learned through CNNs. Therefore, developing a novel asymmetric duplex encoder, with ViT and CNN employed separately in each branch, for RGB-T scene parsing, stands as an underexplored area of research deserving greater attention.

Moreover, the heterogeneous feature fusion strategy is of considerable importance in RGB-T scene parsing [27]. The simplistic and indiscriminate strategies used in prior arts [3,4,24,28] often cause conflicting feature representations and erroneous scene parsing results [18,29]. Taking MFNet [4] as an instance, its feature fusion relies solely on the element-wise concatenation of the heterogeneous features extracted from the duplex encoders at their final stage, treating the features from both modalities equally in the decoder's input. Such a hard-coded feature fusion strategy overlooks the inherent differences in RGB-T feature characteristics and their respective reliability [8], resulting in unsatisfactory performance, particularly under poor illumination conditions. Therefore, another motivation for this article is to design an effective strategy for heterogeneous feature fusion, especially when these features are extracted from different modalities using an asymmetric duplex encoder.

1.3. Contributions

To address the aforementioned limitations, we propose the novel integration of VFMs into RGB-T scene parsing, thus unleashing the representational power derived from self-supervised pre-training on extensive unlabeled data. Given the inherent differences between RGB and thermal images, we first develop a cross-modal spatial prior descriptor (CSPD) based on a CNN to capture robust local spatial patterns from both RGB and thermal images. We then combine the CSPD with a VFM to construct a hybrid, asymmetric encoder, which fuses the heterogeneous features (global context extracted from RGB images and spatial prior captured from RGB-T data) using a progressive heterogeneous feature integrator (PHFI). To achieve this goal, two key components based on multi-scale deformable attention [30] are employed within each encoding stage in PHFIs. These components enable the effective integration of multi-scale spatial priors with single-scale global context. This dual-path, progressive strategy enables a deeper fusion of heterogeneous features. Additionally, inspired by the deep supervision approaches widely utilized in single-modal scene parsing tasks [31], we introduce an auxiliary task to implicitly enrich the local semantics of fine-grained fused features. This approach effectively reduces noisy and semantically irrelevant features, thereby further improving the overall performance of RGB-T scene parsing. Our proposed Hybrid, Asymmetric, and Progressive Network (HAPNet) outperforms all existing state-of-the-art (SoTA) RGB-T scene parsing methods that primarily leverage symmetric architectures. Specifically, HAPNet achieves improvements of 0.1%, 1.0%, and 2.4% in mean intersection over union (mIoU) on the MFNet [4], PST900 [1], and KP Day-Night [32] datasets, respectively. Furthermore, experiments on the NYU-Depth V2 [33] dataset demonstrate HAPNet's generalizability for RGB-HHA scene parsing. These results demonstrate that HAPNet is particularly effective in handling a wide range of challenging scenarios, especially those characterized by poor illumination conditions or cluttered backgrounds.

In a nutshell, this study makes the following contributions:

- We investigate the application of VFMs for RGB-T scene parsing, demonstrating that with appropriate strategies, methods in this domain can significantly benefit from VFMs.
- By revisiting the inherent characteristics of RGB and thermal modalities, we propose a hybrid, asymmetric encoder architecture as well as an asymmetric input design, which can fully exploit the complementary strengths of both modalities.
- We propose an auxiliary task to further enrich the local semantics of fine-grained fused features.
- Our proposed HAPNet achieves SoTA performance on three publicly available RGB-T scene parsing datasets, and demonstrates promising generalizability for RGB-HHA scene parsing.

1.4. Outline

The remainder of this article is structured as follows: Section 2 reviews related works. Section 3 details our proposed HAPNet. Section 4 compares HAPNet with other SoTA methods and presents the ablation study results. Finally, Section 5 concludes this article and discusses possible future work.

2. Related work

2.1. Single-modal scene parsing

The advent of the fully convolutional network (FCN) [34] catalyzed a wave of research on CNN-based scene parsing networks, which typically utilize convolutions for pixel-wise classification and explore various modules and architectures to enhance overall performance. For instance, the DeepLab series [35,36] leverages the atrous spatial pyramid pooling module to capture multi-scale context information, while the feature pyramid network (FPN) [37] effectively fuses multi-scale features in a fully convolutional fashion for robust dense prediction. Nonetheless, CNNs suffer from inherent limitations, *e.g.*, constrained receptive fields, which impede their effectiveness in comprehensively modeling global context.

Transformers [38], originally proposed for natural language processing, have revolutionized computer vision due to their capability to model long-range dependencies. As a representative example, segmentation Transformer (SETR) [39] introduces a ViT-based encoder for scene parsing, enabling long-range dependency modeling via self-attention mechanisms. Another approach, SegFormer [40], employs a hierarchical encoder to achieve multi-scale feature encoding within the Transformer framework, outperforming SETR in handling objects of varying sizes. Additionally, drawing upon earlier works [41,42], the MaskFormer [43,44]

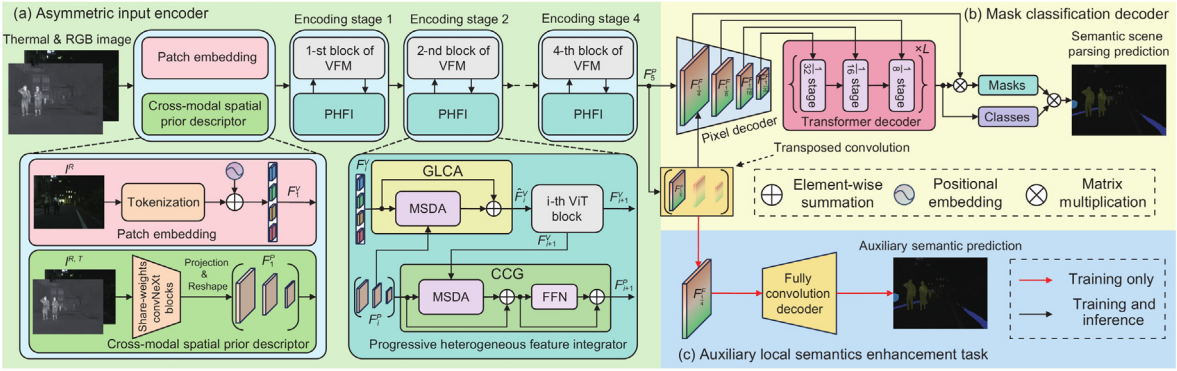

Fig. 1. An overview of our proposed HAPNet architecture.

series formulates scene parsing as a mask classification problem. These networks utilize a Transformer decoder to refine object queries and generate pixel-level predictions by decoding these queries into a set of masks. Given the impressive performance demonstrated by the MaskFormer architecture in dense prediction tasks, we adopt a similar network design for scene parsing in this study.

2.2. Data-fusion scene parsing

The aforementioned single-modal networks often face challenges in handling complex scenes, especially under poor illumination and adverse weather conditions [27,45]. This is primarily because they rely solely on color and texture information present in RGB images [8]. Therefore, researchers have explored multi-modal/source data-fusion techniques (with duplex encoders) to overcome these limitations. MFNet [4] pioneers the use of duplex encoders for RGB-T scene parsing. It extracts heterogeneous features using two lightweight backbones and realizes feature fusion via element-wise concatenation, achieving a balance between speed and accuracy. Most subsequent studies generally employ element-wise summation to achieve heterogeneous feature fusion and focus more on encoder and decoder designs for improved performance. For example, RTFNet [3] introduces skip connections in the decoder to preserve clear boundaries and fine details. Similarly, FuseSeg [28] replaces the ResNet [13] backbone in RTFNet with the stronger DenseNet [46], enabling richer multi-scale feature extraction. Despite their improved performance, these networks employ simplistic heterogeneous feature fusion strategies that limit their ability to fully exploit the complementary information in RGB-T data.

To address this drawback, recent works [5,7–9,23,24,47] have resorted to more advanced, learnable feature fusion strategies. For instance, FEANet [24] and GMNet [8] introduce attention-based modules to dynamically select important heterogeneous features. The subsequent research endeavor CMX [47], building upon the SegFormer [40] architecture, employs average and max pooling operations along with the attention modules to recalibrate heterogeneous features at the spatial and channel dimensions. Afterwards, its subsequent version CMNeXt [48] utilizes similar modules based on CNN and pooling operations for feature fusion, and further expands the architecture of CMX to accommodate an arbitrary number of modalities. Similarly, CAINet [9] employs different attention modules to model global context and aggregate local semantics across various spatial scales of RGB-T feature maps. Moreover, the ABMDRNet series [5,23] introduces novel modality-wise bridging-then-fusing frameworks, capable of reconstructing pseudo images of one modality using the features extracted from the other modality. This approach successfully reduces modality discrepancy and generates more discriminative fused features. Additionally, inspired by the masked autoencoder (MAE) [49], CRM-RGBTSeg [7] randomly masks fixed-size regions within one modality while simultaneously providing complementary regions from another modality to duplex encoders, thus preventing over-reliance on either modality. In this article, we employ a novel asymmetric duplex encoder to capture both local spatial priors and global context from the RGB-T data. These heterogeneous features are progressively fused for more robust scene parsing.

3. Methodology

This section details HAPNet, a hybrid, asymmetric network capable of progressively fusing heterogeneous features extracted from RGB-T data to perform scene parsing. As illustrated in Fig. 1, HAPNet comprises three key components:

- An asymmetric, hybrid encoder that extracts heterogeneous features from RGB-T data using a VFM (based on ViT) and a CSPD. In the meantime, these features are progressively fused via PHFI;
- A Transformer-based mask classification decoder that leverages the fused heterogeneous features to produce semantic predictions;
- An auxiliary task to further enhance the local semantics of fused heterogeneous features.

3.1. Overall workflow

Unlike previous works that utilize symmetric duplex encoder structures, we innovatively introduce an asymmetric, hybrid architecture (based on both a Transformer-based VFM and a lightweight CNN). Compared to the indiscriminate feature extraction strategies utilized by symmetric duplex architectures, our HAPNet effectively leverages the complementary strengths of both modalities. It extracts rich global semantic cues using a VFM, while learning cross-modal local semantics with a CNN. Furthermore, HAPNet conducts effective global-local semantic integration within our proposed PHFIs, resulting in more distinguishable fused features. Consequently, our decoder can more effectively utilize the global context for accurate classification, while maintaining sensitivity to the local details of small-sized objects. This process starts by evenly dividing the ViT (with L encoder layers) into four blocks, each of which contains $\frac{L}{4}$ encoder layers, and inserting PHFIs into each block to form four encoding stages. The RGB image $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$ and its corresponding thermal image $\boldsymbol{I}^T \in \mathbb{R}^{H\times W\times 3}$ are first fed into the CSPD, extracting the cross-modal spatial prior $\boldsymbol{F}_1^P \in \mathbb{R}^{(\sum_{i=2}^{4}\frac{HW}{S_i^2})\times D}$,

which serves as one of the inputs to following stages, where $S_i = 2^{i+1}$ ($i \in [2,4] \cap \mathbb{Z}$) denotes the corresponding stride number. Afterwards, $\boldsymbol{I}^R$ is tokenized to form the context feature $\boldsymbol{F}_1^V \in \mathbb{R}^{\frac{HW}{16^2} \times D}$, which serves as the other input to following stages. $\boldsymbol{F}_1^V$ and $\boldsymbol{F}_1^P$ are then fused in a dual-path, progressive manner using the two key components of our developed PHFI: (1) the global-local context aggregator (GLCA) and (2) the complementary context generator (CCG). Meanwhile, fused features undergo global context encoding in four ViT blocks. This design enables the RGB features to retain fine-grained local semantics while effectively capturing global context through the powerful VFM. Eventually, the fused multi-scale features $\mathcal{F}^F = \{\boldsymbol{F}^F_{\frac{1}{S_j}} \in \mathbb{R}^{\frac{H}{S_j} \times \frac{W}{S_j} \times D}\}$, where $S_j = 2^{i+1}$ ($j \in [1,4] \cap \mathbb{Z}$) from the four encoding stages are obtained and then fed into the subsequent mask classification decoder to generate semantic predictions $\boldsymbol{M}^P \in \mathbb{R}^{H \times W}$, which are also utilized by an auxiliary local semantics enhancement task during model training.

3.2. Hybrid, asymmetric encoder

3.2.1. Encoding pipeline

We evenly divide $\boldsymbol{I}^R$ into non-overlapping patches (resolution: 16×16 pixels) and project them into D-dimensional tokens to form the RGB context feature $\boldsymbol{F}_1^V$. In the ith encoding stage, the input context features $\boldsymbol{F}_i^V$ and spatial prior $\boldsymbol{F}_i^P$ are first fused in the GLCA to generate the local semantics-enhanced context feature $\hat{\boldsymbol{F}}_i^V \in \mathbb{R}^{\frac{HW}{16^2} \times D}$. Subsequently, $\hat{\boldsymbol{F}}_i^V$ undergoes context encoding in the ViT block, generating the output $\boldsymbol{F}_{i+1}^V$, which is then fed into the CCG to perform another fusion with $\boldsymbol{F}_i^P$. This step complements spatial prior $\boldsymbol{F}_i^P$ with updated global context from $\boldsymbol{F}_{i+1}^V$ to generate $\boldsymbol{F}_{i+1}^P$. By repeating such a dual-path feature fusion across four encoding stages, we obtain the fused spatial prior $\boldsymbol{F}_5^P$, which is then recovered to its original three resolutions, forming $\{\boldsymbol{F}^F_{\frac{1}{8}}, \boldsymbol{F}^F_{\frac{1}{16}}, \boldsymbol{F}^F_{\frac{1}{32}}\}$. A 2×2 transposed convolution is employed to directly create the $\frac{1}{4}$-scale features $\boldsymbol{F}^F_{\frac{1}{4}}$ from $\boldsymbol{F}^F_{\frac{1}{8}}$. This design avoids the heavy computational overhead of the attention operations required for creating $\boldsymbol{F}^F_{\frac{1}{4}}$. Finally, these fused features across four scales constitute $\mathcal{F}^F$, which is compatible with current SoTA scene parsing network architectures [43,50].

Solely employing RGB images as the ViT input stems from our hypothesis that explicitly encoding thermal images using VFM may cause convergence issues and representation shift, as discussed in [51]. Furthermore, thermal data often has limited capability in capturing global context, as objects with similar geometries tend to have similar heat signatures. For example, distinguishing between a soccer ball and a basketball or between a cat and a dog is challenging in thermal images. We validate the effectiveness and superior performance of this hybrid, asymmetric network design through ablation studies in Section 4.5.

3.2.2. Cross-modal spatial prior descriptor

Compared to Transformers [9,52], CNNs have demonstrated superior performance in capturing local semantics, underscoring their significance in computer vision applications that necessitate clear object boundaries. ConvNeXt [16] has shown exceptional performance in capturing rich, robust visual features in comparison with other hierarchical CNN architectures such as ResNet [13] and its variants [53], as evidenced by recent endeavors using ConvNeXt for multi-modal scene parsing [54,55]. Given these advantages, we employ basic ConvNeXt blocks to construct our CSPD. Furthermore, incorporating RGB data as a complementary input enables the CSPD to extract more comprehensive local spatial patterns than using thermal images alone. Consequently, we construct our CSPD using two identical, weight-sharing subnetworks, enabling efficient and effective extraction of cross-modal spatial priors from the RGB-T data.

Specifically, $\boldsymbol{I}^R$ and $\boldsymbol{I}^T$ are independently fed into a series of weight-sharing ConvNeXt blocks, generating multi-scale features $\mathcal{P}^R = \{\boldsymbol{P}_2^R, \boldsymbol{P}_3^R, \boldsymbol{P}_4^R\}$ and $\mathcal{P}^T = \{\boldsymbol{P}_2^T, \boldsymbol{P}_3^T, \boldsymbol{P}_4^T\}$, respectively, where $\boldsymbol{P}_i^{R,T} \in \mathbb{R}^{\frac{H}{S_i} \times \frac{W}{S_i} \times C_i}$ represents the features at ith stage, C_i and $S_i = 2^{i+1}$ ($i \in [2,4] \cap \mathbb{Z}$) denote the corresponding channel and stride numbers, respectively. We then combine $\boldsymbol{P}_i^R$ and $\boldsymbol{P}_i^T$ via element-wise summation, and reduce the results to D channels through 1×1 convolutions, yielding heterogeneous features $\mathcal{P}^H = \{\boldsymbol{P}_2^H, \boldsymbol{P}_3^H, \boldsymbol{P}_4^H\}$, where $\boldsymbol{P}_i^H \in \mathbb{R}^{\frac{H}{S_i} \times \frac{W}{S_i} \times D}$. The three-scale features are then flattened and concatenated to form the cross-modal spatial prior $\boldsymbol{F}_1^P$ which is used for the following encoding stages. We present detailed ablation studies in Section 4.5 to analyze the impact of different CSPD construction blocks and data input strategies on the overall performance.

3.2.3. Progressive heterogeneous feature integrator

The GLCA and CCG, two attention-based components of our PHFI, are responsible for the dual-path feature fusion process before and after each ViT block, respectively. Given $\boldsymbol{F}_i^V$ and $\boldsymbol{F}_i^P$ of the ith encoding block ($i \in [1,4] \cap \mathbb{Z}$), we first feed them into the GLCA to obtain the local semantics-enhanced input $\hat{\boldsymbol{F}}_i^V$. Specifically, $\boldsymbol{F}_i^V$ serves as the query, while $\boldsymbol{F}_i^P$ acts as the key and value within the GLCA. This process can be formulated as follows:

$$\hat{\boldsymbol{F}}_i^V = \boldsymbol{F}_i^V + \kappa_i \mathrm{MHA}(\mathrm{LN}(\boldsymbol{F}_i^V), \mathrm{LN}(\boldsymbol{F}_i^P)) \tag{1}$$

where $\mathrm{LN}(\cdot)$ represents the layer normalization operation, $\mathrm{MHA}(\cdot,\cdot)$ represents the multi-head attention operation. Following common practices in attention modules such as [54,56,57], we introduce a learnable coefficient κ_i to dynamically adjust the weight of the MHA, enabling flexible fusion of the local semantics-enhanced features into the VFM. After the ith ViT block, the output feature maps $\boldsymbol{F}_{i+1}^V$ with enriched global context are yielded. $\boldsymbol{F}_{i+1}^V$ is then fed into the CCG to fuse its global context with the spatial prior $\boldsymbol{F}_i^P$. In this process, $\boldsymbol{F}_i^P$ serves as the query, while $\boldsymbol{F}_{i+1}^V$ acts as the keys and values to perform MHA within the CCG as follows:

$$\hat{\boldsymbol{F}}_i^P = \boldsymbol{F}_i^P + \mathrm{MHA}(\mathrm{LN}(\boldsymbol{F}_i^P), \mathrm{LN}(\boldsymbol{F}_{i+1}^V)) \tag{2}$$

After the MHA operation, a feed-forward network (FFN) is applied to further process the fused features, resulting in the updated spatial prior $\boldsymbol{F}_{i+1}^P$, which has already incorporated rich local and global semantics and will be inputted to the next encoding stage. In the above two components, the MHA processes are implemented based on multi-scale deformable attention (MSDA) [30] to reduce computations. Please refer to the supplemental material for more detailed mathematical derivations of MSDA.

3.3. Mask classification decoder

Following [44], our mask classification decoder consists of a pixel decoder and a Transformer decoder. The former improves the fused heterogeneous features $\{\boldsymbol{F}^F_{\frac{1}{8}}, \boldsymbol{F}^F_{\frac{1}{16}}, \boldsymbol{F}^F_{\frac{1}{32}}\}$, producing $\hat{\mathcal{F}}^P = \{\hat{\boldsymbol{F}}^F_{\frac{1}{8}}, \hat{\boldsymbol{F}}^F_{\frac{1}{16}}, \hat{\boldsymbol{F}}^F_{\frac{1}{32}}\}$, and generates the pixel embedding $\boldsymbol{E}^P \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$ from $\boldsymbol{F}^F_{\frac{1}{4}}$; The latter employs $\hat{\mathcal{F}}^P$ to refine a fixed size of queries, thus producing the mask embedding $\boldsymbol{E}^M \in \mathbb{R}^{Q \times C}$, where Q and C represent the numbers of object queries and feature channels, respectively. $\boldsymbol{E}^M$ is multiplied by $\boldsymbol{E}^P$ to yield mask predictions $\boldsymbol{M}^M \in \mathbb{R}^{Q \times \frac{H}{4} \times \frac{W}{4}}$. By resizing $\boldsymbol{M}^M$ and associating it with the class predictions $\boldsymbol{E}^C \in \mathbb{R}^{Q \times N}$ learned from queries, we obtain the semantic predictions $\boldsymbol{M}^P$, where N denotes the number of classes (including a "no object" class). Readers can refer to [44] for more details on this decoding process.

3.4. Auxiliary local semantics enhancement task

In our mask classification decoder, $\boldsymbol{E}^P$ generated from $\boldsymbol{F}^F_{\frac{1}{4}}$ typically provides rich local (per-pixel) semantics of the scene, while $\boldsymbol{E}^M$ generated from the other three scales of features provides information on category-related cluster centers [58]. To enhance the robustness and discriminability of local semantics utilized by our mask classification decoder, we incorporate deep supervision into $\boldsymbol{F}^F_{\frac{1}{4}}$, effectively reducing noisy and ambiguous patterns within it. Specifically, we attach $\boldsymbol{F}^F_{\frac{1}{4}}$ with a lightweight fully convolutional decoder head and directly supervise this auxiliary network to produce semantic predictions. After two convolutional layers, the auxiliary semantic predictions with $(N-1)$ channels and the same resolution as $\boldsymbol{F}^F_{\frac{1}{4}}$ are generated (excluding the "no object" category). This design enables the corresponding encoder layers to produce a $\boldsymbol{F}^F_{\frac{1}{4}}$ with more distinguishable local semantics, further ensuring that the pixel decoder can generate a more distinguishable $\boldsymbol{E}^P$, and ultimately leading to improved scene parsing performance without increasing network parameters or computational complexity.

3.5. Loss function

We employ the cross-entropy (CE) loss $\mathcal{L}_{\text{ce}}$ for the class prediction of object queries, along with the binary cross-entropy (BCE) loss $\mathcal{L}_{\text{bce}}$ and dice loss $\mathcal{L}_{\text{dice}}$ for mask prediction. $\mathcal{L}_{\text{ce}}$ can be formulated as follows:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_c^i \log(\boldsymbol{p}_c^i) \tag{3}$$

where N denotes the total number of object queries, $\boldsymbol{y}_c^i$ denotes the one-hot encoded ground-truth vector for c categories, and $\boldsymbol{p}_c^i$ stores values indicating the prediction probabilities over c categories. $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ for mask classification are formulated as follows:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{M}\sum_{i=1}^{M}(y_b^i \log(p_b^i) + (1-y_b^i)\log(1-p_b^i)) \tag{4}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{M} y_b^i p_b^i}{\sum_{i=1}^{M} y_b^i + \sum_{i=1}^{M} p_b^i} \tag{5}$$

where M denotes the total number of pixels, $y_b^i \in \{0, 1\}$ denotes the binary ground truth of one foreground pixel, and $p_b^i \in [0, 1]$ is a scalar indicating the corresponding prediction probability. For our local semantics enhancement task, we directly calculate another CE loss $\mathcal{L}_{\text{aux}}$ over its semantic predictions. The overall loss function can be expressed as follows:

$$\mathcal{L} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} \tag{6}$$

We follow [36,44] to set $\lambda_{\text{bce}} = 5.0$, $\lambda_{\text{dice}} = 5.0$, $\lambda_{\text{ce}} = 2.0$, and $\lambda_{\text{aux}} = 0.4$, respectively, and we set $\lambda_{\text{ce}} = 0.1$ for the "no object" category. Additionally, we employ bipartite matching [59] to determine the least-cost assignment during model training.

4. Experiments

We conduct extensive experiments in this article to evaluate the performance of our developed HAPNet both quantitatively and qualitatively. We further quantify the generalizability of HAPNet for RGB-depth/HHA (RGB-D/HHA) scene parsing. The following subsections provide details on the datasets and experimental setup, the evaluation metrics, as well as the comprehensive evaluation of our proposed method.

4.1. Datasets and experimental setup

We first compare HAPNet with other SoTA RGB-T scene parsing networks on the following three public datasets:

4.1.1. MFNet [4]

This is an urban driving scene parsing dataset. An InfReC R500 camera was utilized to capture 1569 pairs of synchronized RGB and thermal images at a resolution of 640 × 480 pixels. The dataset provides semantic annotations of nine classes: bike, person, car, road lanes, guardrail, car stop, bump, color cone, and background. We adopt the same dataset splitting strategy as introduced in [4].

4.1.2. PST900 [1]

This dataset contains 894 pairs of RGB and thermal images captured in challenging underground environments (originally released on the DARPA Subterranean Challenge). The RGB images were acquired using a Stereolabs ZED Mini stereo camera, while the thermal images were captured using a FLIR Boson 320 camera, both at a resolution of $1,280 \times 720$ pixels. The dataset has five categories of semantic annotations: background, fire extinguisher, backpack, hand drill, and survivor. We adopt the same data splitting strategy[1] as described in [1].

4.1.3. KP day-night [32]

This dataset contains 950 well-rectified RGB-T image pairs (resolution: 640 × 512 pixels) captured in urban driving scenes. This study [32] provides semantic annotations of 19 categories (identical to the CityScapes [60] dataset). We adopt the same data splitting strategy presented in [7] to train and evaluate our method.

Furthermore, we conduct an additional experiment on the NYU-Depth V2 dataset [33] to evaluate the generalizability of our network for RGB-D/HHA scene parsing.

4.1.4. NYU-Depth V2 [33]

This dataset has been widely used for the evaluation of indoor scene parsing performance. It provides 1449 pairs of RGB and depth images (resolution: 480 × 640 pixels), and semantic annotations for 40 categories. We adopt the same dataset splitting strategy as presented in [51].

4.2. Evaluation metrics

We employ two widely used evaluation metrics: accuracy (Acc) and IoU, to quantify the scene parsing performance with respect to each category. The mean values of these two metrics across all categories (denoted as mIoU and mAcc, respectively) are also computed to quantify the comprehensive network performance.

4.3. Implementation details

Our model was trained for 200 epochs on an NVIDIA RTX 3090 GPU, utilizing the AdamW optimizer [61] with an initial learning rate of 2×10^{-4} and a weight decay of 5×10^{-2}. Following the common practice employed in [19], we apply a layer-wise learning rate decay of 0.9 for our VFM encoder. For the experiments conducted on the NYU dataset, we employ RGB images and the HHA encoding of depth images as model input. We conduct all ablation studies on the widely utilized MFNet dataset, with the auxiliary task omitted. By comparing the results

[1] The number of image pairs being published is fewer than what was reported in publication [1].

Table 1
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. The values of Acc and IoU for the "background" category are omitted in the table, but they are still included in the calculation of the corresponding mean values.

Methods	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [4]	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	8.5	30.3	25.2	30.0	27.7	45.1	39.7
RTFNet [3]	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
FuseSeg [28]	93.1	87.9	81.4	71.7	78.5	64.6	68.4	44.8	29.1	22.7	63.7	6.4	55.8	46.9	66.4	47.9	70.6	54.5
EGFNet [62]	**95.8**	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	33.6	7.0	65.3	48.3	71.1	47.1	72.7	54.8
ABMDRNet [5]	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
ECGFNet [63]	89.4	83.5	85.2	72.1	72.9	61.6	62.8	40.5	44.8	30.8	45.2	11.1	57.2	49.7	65.1	50.9	69.1	55.3
FEANet [24]	93.3	87.8	82.7	71.1	76.7	61.1	65.5	46.5	26.6	22.1	**70.8**	6.6	**66.6**	55.3	77.3	48.9	73.2	55.3
SFAF-MA [64]	94.0	88.1	82.5	73.0	73.9	61.3	63.6	45.6	37.5	29.5	42.2	5.5	57.9	45.7	74.4	53.8	69.6	55.5
ABMDRNet+ [23]	95.2	87.1	**92.5**	69.8	76.2	60.9	**72.0**	47.8	42.3	34.2	66.8	8.2	64.8	50.2	63.5	55.0	74.7	56.8
LLE-Seg [65]	91.8	88.6	81.3	73.2	73.7	64.8	62.5	46.8	33.7	30.0	49.1	8.8	55.7	52.5	72.9	**62.4**	71.6	58.4
CAINet [9]	93.0	88.5	74.6	66.3	**85.2**	**68.7**	65.9	**55.4**	34.7	31.5	65.6	9.0	55.6	48.9	**85.0**	60.7	73.2	58.6
EAEFNet [2]	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	14.2	62.7	52.4	71.9	58.3	**75.1**	58.9
CMX [47]	–	90.1	–	75.2	–	64.5	–	50.2	–	35.3	–	8.5	–	54.2	–	60.6	–	59.7
CMNeXt [48]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–		59.9
CRM-RGBTSeg [7]	–	90.0	–	75.1	–	67.0	–	45.2	–	**49.7**	–	**18.4**	–	54.2	–	54.4	–	61.4
HAPNet (Ours)	95.1	**90.6**	85.5	**75.4**	79.0	67.2	67.9	51.1	**59.5**	48.4	16.0	4.2	65.0	**59.1**	80.3	59.1	70.3	**61.5**

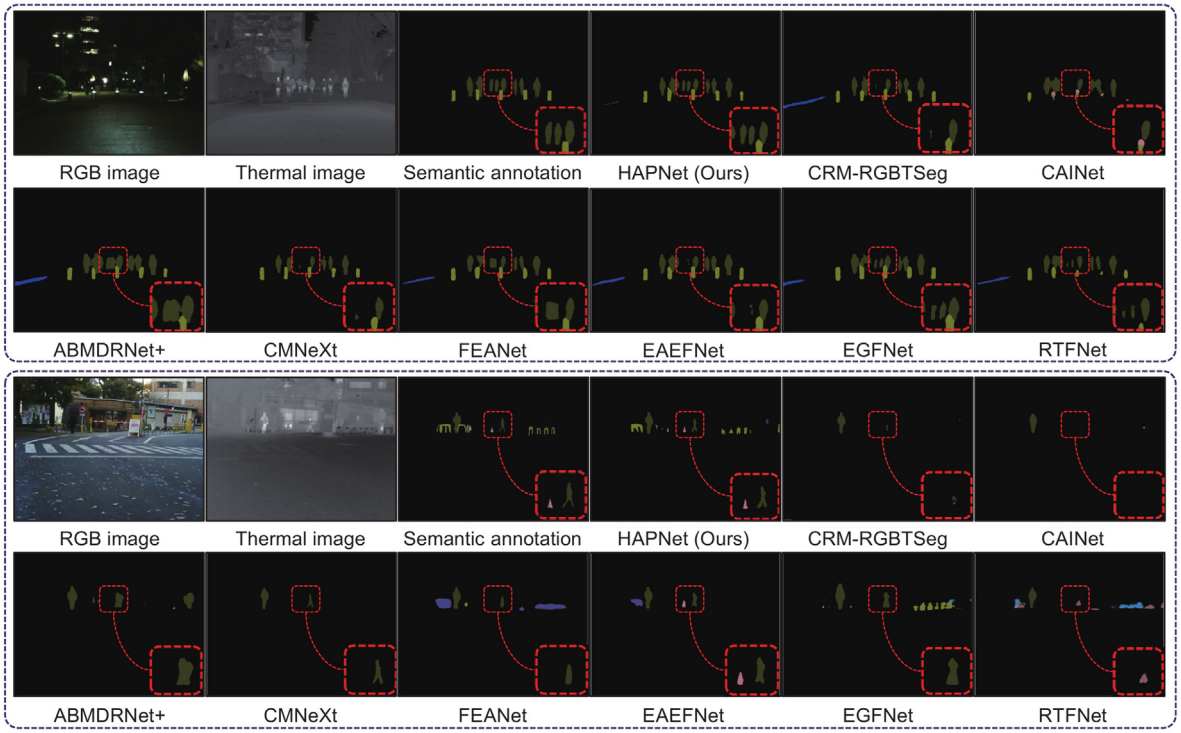

Fig. 2. Qualitative comparisons with the SoTA RGB-T scene parsing networks on the MFNet test set, where significantly improved regions are shown with red dashed boxes.

in Tables 1 and 8, we observe that the performance improvement attributed to the auxiliary task is 0.6% in mIoU (calculated as an average across multiple experiments and subject to a fluctuation of approximately ±0.2%). These results demonstrate the effectiveness of the auxiliary task.

4.4. Comparisons with SoTA scene parsing networks

We conduct quantitative comparisons with 15, 10, and four SoTA RGB-T scene parsing networks on the MFNet [4], PST900 [1], and KP Day-Night [32] datasets, respectively. The results are detailed in Tables 1, 2, and 3, respectively. Additionally, we present the qualitative comparisons on these three datasets, as shown in Figs. 2, 3, and 4, respectively.

Specifically, our proposed HAPNet achieves the highest mIoU across all datasets, outperforming SoTA methods by 0.1–21.8% on the MFNet dataset, 1.0–32.0% on the PST900 dataset, and 2.4–33.6% on the KP day-night dataset, respectively. These results demonstrate the robustness and effectiveness of HAPNet for RGB-T scene parsing in various scenarios, ranging from complex urban driving scenes to challenging underground scenes. Notably, despite performance variations on rare categories (e.g., "guardrail") in the MFNet dataset, our approach consistently outperforms CRM-RGBTSeg [7], a state-of-the-art Transformer-based method using the mask classification paradigm. This demonstrates the efficacy of our hybrid asymmetric architecture for heterogeneous feature fusion. Compared to CMNeXt [48], a general-purpose multi-modal data-fusion method that also employs an asymmetric encoder architecture, HAPNet achieves an improvement of

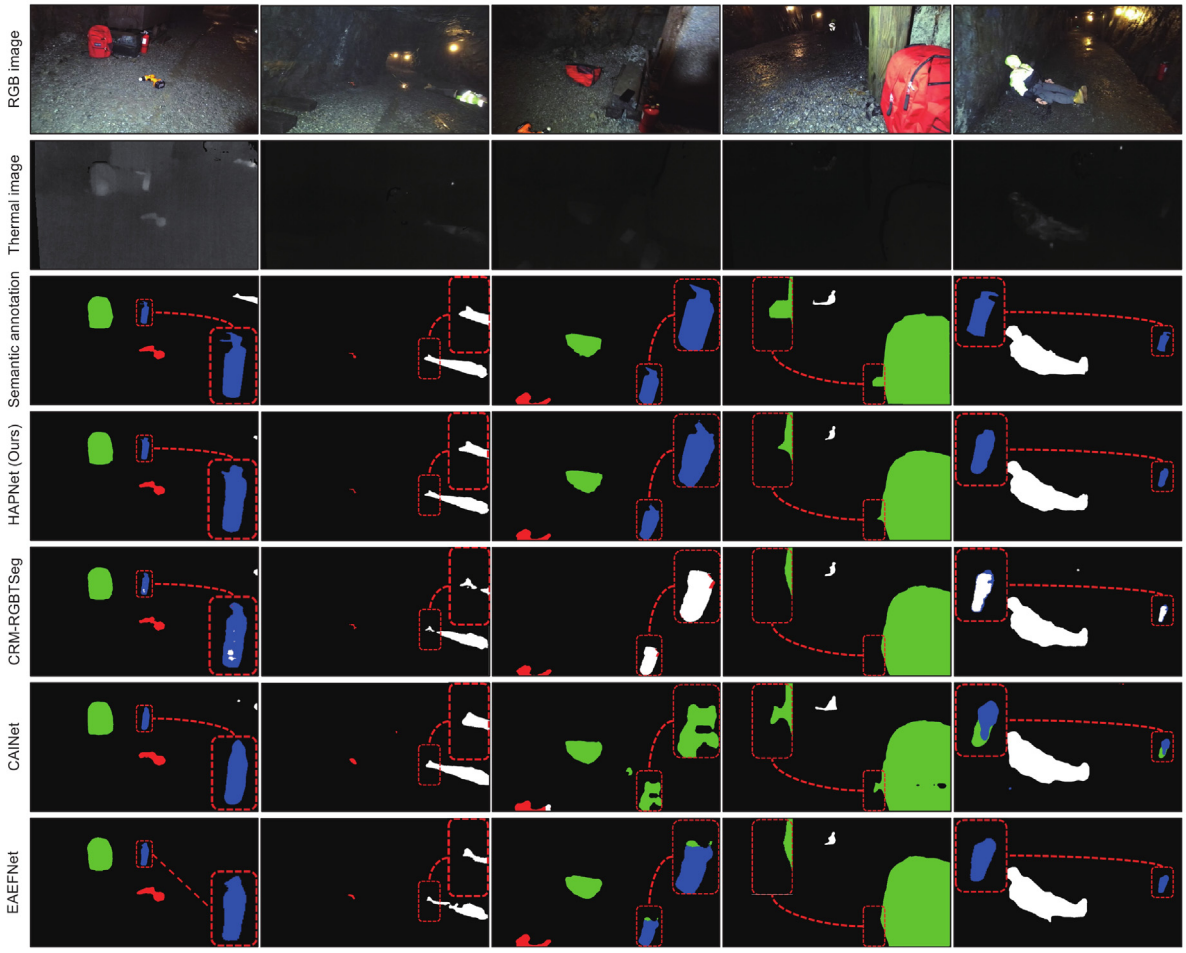

Fig. 3. Qualitative comparisons with the SoTA RGB-T scene parsing networks on the PST900 test set, where significantly improved regions are shown with red dashed boxes.

Table 2
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		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [4]	–	98.6	–	41.1	–	64.2	–	60.3	–	20.7	–	57.0
RTFNet [3]	–	98.9	–	36.4	–	75.3	–	52.0	–	25.3	–	57.6
EGFNet [62]	–	99.2	–	74.3	–	83.0	–	71.2	–	64.6	–	78.5
ABMDRNet [5]	–	98.7	–	24.1	–	72.9	–	54.9	–	57.6	–	67.3
FEANet [24]	–	–	–	–	–	–	–	–	–	–	91.4	85.5
DBCNet [45]	–	98.9	–	62.3	–	71.1	–	52.4	–	40.6	–	74.5
CAINet [9]	99.6	99.5	**95.8**	80.3	**96.0**	88.0	88.3	77.2	91.3	78.6	**94.2**	84.7
EAEFNet [2]	**99.8**	99.5	92.2	80.4	91.0	87.7	93.0	83.9	79.3	75.6	91.1	85.4
GMNet [8]	**99.8**	99.4	90.2	**85.1**	89.0	83.8	88.2	73.7	80.8	78.3	89.6	84.1
CRM-RGBTSeg [7]	–	**99.6**	–	79.5	–	89.6	–	89.0	–	82.2	–	88.0
HAPNet (Ours)	**99.8**	**99.6**	93.9	81.3	95.1	**92.0**	**95.5**	**89.3**	**85.6**	**82.4**	94.0	**89.0**

1.6% in mIoU on the MFNet. This performance improvement can be attributed to our method's utilization of a more powerful VFM and a more advanced multi-scale cross-attention mechanism that integrates global context from the VFM and multi-scale spatial priors provided by a CNN, in contrast to CMNeXt's reliance on CNN and pooling-based feature fusion strategies. Consequently, HAPNet generates more discriminative fused features for RGB-T scene parsing. The distinct parsing results in Figs. 2, 3, and 4 also demonstrate the superior capability of HAPNet for visual perception under poor illumination conditions.

In addition, HAPNet demonstrates promising generalizability for RGB-D/HHA scene parsing. As shown in Table 4, our method outperforms all other RGB-T scene parsing networks on the NYU Depth V2 dataset. Specifically, HAPNet achieves a 3.5% higher mIoU than ECGFNet [63], and a 5.9% higher mIoU than RTFNet [3]. However, these results suggest that HAPNet still lags behind SoTA methods, developed specifically for RGB-D/HHA data fusion, including Omnivore [66], DFormer-L [51], and OmniVec [67] (mIoU is lower by 2.2–5.8%). Additionally, it underperforms CMX and CMNeXt [47,48], two SoTA general-purpose multi-modal scene parsing methods by 1.9% in mIoU. We hypothesize that this performance gap stems from thermal images providing richer complementary features than depth/HHA images. While depth/HHA images primarily convey geometric

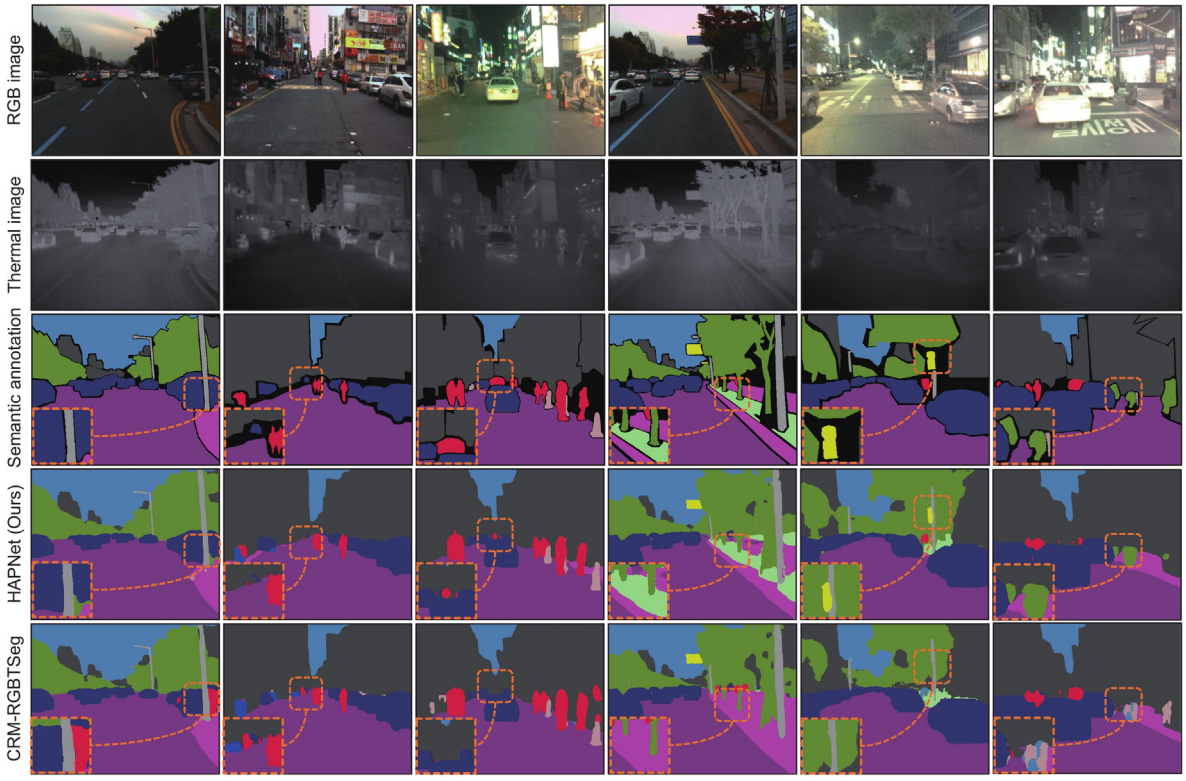

Fig. 4. Qualitative comparisons with the SoTA RGB-T scene parsing networks on the KP Day-Night test set, where significantly improved regions are shown with orange dashed boxes.

Table 3
Quantitative comparisons (%) with the SoTA RGB-T scene parsing methods on the KP Day-Night test set. The best results are presented in bold font. Although results for Acc and IoU for certain categories are omitted in the table, they are still included in the calculations of the corresponding mean values.

Methods	Road	Sidewalk	Building	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Motorcycle	Bicycle	mIoU
MFNet [4]	93.5	23.6	75.1	0.1	9.1	0.0	0.0	69.3	0.2	90.4	24.0	0.0	69.6	**0.3**	0.3	0.0	0.6	24.0
RTFNet [3]	94.6	39.4	86.6	0.6	0.0	0.0	0.0	81.7	3.7	92.8	58.4	0.0	87.7	0.0	0.0	0.0	0.5	28.7
CMX [47]	97.7	53.8	90.2-	47.1	46.2	10.9	45.1	87.2	34.3	93.5	74.5	0.0	91.6	0.0	59.7	46.1	0.2	46.2
CRM-RGBTSeg [7]	**99.0**	**61.9**	**91.8**	**58.7**	50.6	39.2	55.3	**89.2**	23.2	94.3	**85.2**	2.9	**95.3**	0.0	**80.5**	**66.2**	**54.6**	55.2
HAPNet (Ours)	98.6	59.3	91.5	57.0	**58.0**	**41.0**	**56.8**	87.8	**30.4**	**94.8**	81.8	**21.3**	94.2	0.0	69.7	49.9	43.8	**57.6**

information, thermal images capture both fine-grained spatial priors and semantic cues through heat signatures (e.g., distinguishing animate from inanimate objects). These characteristics are particularly well-suited to our proposed CSPD and CCG modules. Consequently, these inherent distributional differences between thermal and depth/HHA modalities limit the applicability of our architecture to RGB-D/HHA scene parsing.

Despite achieving SoTA performance, HAPNet demonstrates competitive real-time performance. Our network contains 169.1M parameters and achieves an inference speed of $\sim$ 23 frames per second (FPS) when processing images at a resolution of 480 $\times$ 480 pixels on an NVIDIA RTX 4090-D GPU. This performance nearly meets the real-time processing requirements for autonomous driving systems.

Table 4
Quantitative comparisons (%) with the SoTA data-fusion scene parsing networks on the NYU-Depth V2 test set. The symbol "–" denotes missing data in the original publication, and the best results are presented in bold font.

Input Data	Methods	Pixel Acc	mAcc	mIoU	Rank
RGB-D	OmniVec [67]	–	–	**60.8**	**1**
	Omnivore [66]			56.8	8
	TokenFusion [68]	79.0	66.9	54.2	16
	DFormer-L [51]	–	–	57.2	5
	AsymFormer [69]	78.5	–	54.1	17
	RTFNet [3]	–	64.8	49.1	59
	ECGFNet [63]	–	65.2	51.5	37
RGB-HHA	CMX [47]	–	–	56.9	6
	CMNeXt [48]	–	–	56.9	6
	SA-Gate [70]	–	–	52.4	31
	HAPNet (Ours)	**79.2**	**68.8**	55.0	15

4.5. Ablation study

We first explore the performance of RGB-T scene parsing with respect to different input data combination strategies. Given the asymmetric two-branch architecture, the RGB-T data can be fed into our HAPNet using nine different strategies to the VFM and CSPD, respectively, as shown in Table 5. Among them, strategies E and I can be considered as single-modal versions (only RGB or thermal images are encoded by the network), yielding mIoU scores of only 53.9% and 51.5%, respectively. On the other hand, strategies A-D, and F-H, which employ both RGB and thermal images, achieve superior performance, validating the importance of

Table 5
An ablation study (%) on different data input strategies on the MFNet test set.

Strategies	VFM	CSPD	IoU								mAcc	mIoU
			Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
A	RGB + T	RGB + T	89.3	73.9	66.1	48.0	32.4	4.4	55.7	62.6	74.3	58.9
B	RGB + T	RGB	88.3	69.1	64.3	44.3	32.3	4.3	52.9	44.5	67.0	55.3
C	RGB + T	Thermal	85.5	72.7	54.6	39.3	28.5	**8.2**	46.1	55.4	71.5	54.2
D (Ours)	RGB	RGB + T	89.1	**76.0**	**67.2**	**50.6**	**41.3**	7.9	**58.6**	58.7	**75.1**	**60.9**
E	RGB	RGB	87.9	62.5	63.8	40.7	28.2	5.6	52.1	47.1	66.7	53.9
F	RGB	Thermal	86.7	73.9	60.3	39.1	33.0	7.3	53.6	**59.0**	74.6	56.7
G	Thermal	RGB + T	**89.6**	74.6	65.1	48.9	33.9	1.5	53.9	56.8	68.8	58.1
H	Thermal	RGB	88.4	69.3	65.3	43.2	29.7	4.4	53.4	44.4	67.4	55.1
I	Thermal	Thermal	84.5	71.8	53.8	35.0	24.3	3.8	38.4	54.5	66.6	51.5

Table 6
An ablation study (%) on different CSPD construction blocks on the MFNet test set.

Encoders	IoU								mAcc	mIoU
	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
Duplex ResNet-101 [13] sharing weights	66.3	48.0	38.4	19.9	9.1	0.0	29.6	23.6	42.3	36.8
Duplex Swin-Transformer-S [14] sharing weights	**89.7**	74.1	67.2	47.6	31.0	6.3	52.7	57.0	74.6	58.2
Duplex MiT-B4 sharing weights [40] sharing weights	87.8	74.7	64.9	48.4	40.9	**13.1**	51.4	48.1	73.2	58.6
Duplex ConvNeXt-S [16] with separate weights	89.6	74.6	66.6	46.3	**45.7**	10.2	56.6	**62.5**	73.6	**61.2**
Duplex ConvNeXt-S [16] sharing weights (Ours)	89.1	**76.0**	**67.2**	**50.6**	41.3	7.9	**58.6**	58.7	**75.1**	60.9

Table 7
An ablation study (%) on the effectiveness of GLCA and CCG on the MFNet test set. When both components are removed, the spatial priors extracted using CSPD and the global context extracted using ViT are fused via element-wise summation after resolution alignment.

GLCA	CCG	Fusion Strategies	IoU								mAcc	mIoU
			Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
		Element-wise Summation	89.2	73.5	65.7	50.0	33.0	4.1	52.0	55.4	71.1	57.9
✓		Local Semantics Enhanced	**89.9**	74.7	67.0	50.0	37.0	5.8	52.9	55.9	72.9	59.1
	✓	Global Context Updated	89.6	74.5	67.0	46.6	36.9	2.9	53.3	53.3	72.4	58.0
✓	✓	**Both (Ours)**	89.1	**76.0**	**67.2**	**50.6**	**41.3**	**7.9**	**58.6**	**58.7**	**75.1**	**60.9**

Table 8
An ablation study (%) on different symmetric and asymmetric encoder architectures on the MFNet test set.

Methods (encoder)	IoU								mAcc	mIoU
	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
Duplex BEiTv2-B/16 [20] concatenation	65.7	57.4	12.7	27.7	8.5	0.0	14.7	20.6	40.9	33.7
Duplex BEiTv2-B/16 [20] summation	65.4	57.1	23.3	28.0	8.5	0.0	19.9	16.2	42.6	35.0
Duplex ConvNeXt-S [16] concatenation	89.7	74.1	66.4	45.3	29.8	4.8	52.8	55.5	69.2	57.4
Duplex ConvNeXt-S [16] summation	**90.7**	74.6	66.9	46.8	32.9	6.6	56.6	52.7	72.0	58.5
HAPNet CMNeXt-B4 [48]	88.5	74.0	62.5	44.9	41.1	**12.0**	53.3	**59.0**	71.6	59.3
HAPNet (Ours)	89.1	**76.0**	**67.2**	**50.6**	**41.3**	7.9	**58.6**	58.7	**75.1**	**60.9**

Table 9
An ablation study (%) on different VFMs on the MFNet test set. "MM" represents multi-modal pre-training. BEiT and BEiTv2 are trained via a self-supervised learning strategy named masked image modeling, while DINOv2 is trained via discriminative self-supervised learning strategy.

VFM	Pre-training Strategy	Dataset	IoU								mAcc	mIoU
			Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
DeiT-B/16 [71]	Supervised	ImageNet-1K	90.4	75.2	66.6	47.6	36.3	0.7	**59.9**	**60.0**	68.3	59.4
AugReg-B/16 [72]	Supervised	ImageNet-22K	90.3	74.9	66.8	48.2	35.8	6.0	56.6	56.5	74.7	59.3
BEiT-B/16 [19]	Self-Supervised	ImageNet-22K	**90.7**	75.0	**67.6**	48.6	39.1	4.2	55.0	59.0	72.5	59.7
DINOv2-B/14 [21]	Self-Supervised	LVD-142M (MM) [21]	90.0	74.7	66.8	45.1	**42.1**	**9.7**	56.2	59.7	**76.4**	60.3
BEiTv2-B/16 [20]	Self-Supervised	ImageNet-22K (MM)	89.1	**76.0**	67.2	**50.6**	41.3	7.9	58.6	58.7	75.1	**60.9**

modality complementation. Notably, strategy D, employing RGB images as inputs to the VFM branch while using a combination of RGB-T data in the CSPD, achieves the highest mIoU of 60.9%. This result demonstrates that encoding thermal images using the VFM may deteriorate features and degrade the overall performance, which verifies our hypothesis stated in Section 1.

Our CSPD extracts multi-scale spatial priors from both RGB and thermal images, which are subsequently fused with the VFM context features to implicitly enrich local semantics. To further explore the most suitable structure for the CSPD, we replace the basic ConvNeXt blocks with other mainstream hierarchical encoder structures. As detailed in Table 6, ConvNeXt demonstrated superior performance, achieving a mIoU of 60.9%, outperforming established architectures such as ResNet, Swin-Transformer, and the Mix Vision Transformer (MiT-B4), the latter being a heavy structure previously employed in another RGB-X scene parsing network [48].

In addition, we investigate the impact of employing separate weights for the two modalities, and this strategy yields a slightly better result. However, considering the trade-off between accuracy and model complexity, we still adopt the weight-sharing strategy, as the weight-separating strategy only achieves an improvement of 0.3% in terms of mIoU while incurring a substantial increase in the modal parameters.

Furthermore, we evaluate the effectiveness of GLCA and CCG. As presented in Table 7, removing either of them leads to a significant performance deterioration of HAPNet. The baseline setup, which could only perform cross-modal feature fusion via element-wise summation between spatial priors and global context, exhibits a 3.0% lower mIoU compared to the complete HAPNet.

To validate the effectiveness of our proposed asymmetric architecture, we first construct baseline architectures incorporating symmetric duplex encoders based on BEiTv2 [20] and ConvNeXt [16], respectively. As shown in Table 8, such baseline architectures fail to achieve performance comparable to that of our developed HAPNet. Specifically, the duplex, symmetric ConvNeXt and BEiTv2 architectures fall short of HAPNet, with a decrease in mIoU by 3.5–2.4% and 27.2–25.9%, respectively. These findings lend further support to our hypothesis stated in Section 1, suggesting that applying asymmetric architectures to RGB and thermal modalities can play to their own strengths. Importantly, the disappointing performance of the symmetric duplex BEiTv2 encoder suggests that directly applying VFMs for the RGB-T scene parsing task might not be suitable. We further compare the encoder of HAPNet with another best-performing asymmetric encoder proposed in [48]. In order to fairly compare the performance of the two encoder architectures, we combined CMNeXt's encoder with HAPNet's decoder, abbreviated as HAPNet CMNeXt-B4, as shown in Table 8, and the mIoU exhibits a decrease of 1.6%, further underscoring the superiority of our encoder design compared to other architectures for RGB-T scene parsing.

Moreover, we also evaluate the compatibility of HAPNet with various SoTA VFMs, including those trained through traditional supervised learning (DeiT [71] and AugReg [72]) as well as the most recent models developed based on self-supervised pretraining (BEiT [19], BEiTv2 [20], and DINOv2 [21]). As shown in Table 9, BEiTv2 achieves the highest mIoU of 60.9%, outperforming all other models on the RGB-T scene parsing task. Given BEiTv2's superior performance, we empirically determine it to be the most suitable VFM for this task, thereby adopting BEiTv2 as the default VFM within HAPNet.

5. Conclusion and future work

In this article, we revisited the design of heterogeneous feature extraction and fusion strategies. Drawing upon the inherent characteristics of RGB and thermal data, we developed a novel hybrid, asymmetric network to capitalize on their unique strengths, leveraging the powerful VFM for this specific task. We first developed a cross-modal spatial prior descriptor to capture local semantics jointly from RGB-T data. Additionally, we designed a progressive heterogeneous feature integrator, consisting of a global-local context aggregator and a complementary context generator, to fuse heterogeneous features more effectively. We also introduced an auxiliary task to further enhance the local semantics of fused features, leading to improved overall performance. Extensive experiments demonstrate that HAPNet not only achieves state-of-the-art performance for RGB-T scene parsing across three public datasets but also shows great potential to generalize well for RGB-HHA scene parsing in indoor scenarios. Despite its superior performance over other existing approaches, the generalizability of HAPNet for RGB-X (where "X" represents, but is not limited to, depth, surface normal, and so forth) can be further improved. Additionally, considering scene parsing is a common functionality embedded in autonomous cars, mobile robots, and drones, the real-time performance of HAPNet is crucial, which is also a direction for possible future work.

CRediT authorship contribution statement

Jiahang Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Peng Yun:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Yang Xu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Ye Zhang:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Mingjian Sun:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis. **Qijun Chen:** Writing – review & editing, Visualization, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition. **Ilin Alexander:** Writing – review & editing, Visualization, Resources, Investigation. **Rui Fan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (62473288 and 62233013), the Fundamental Research Funds for the Central Universities, China, and the Xiaomi Young Talents Program.

Appendix A. Mathematical details

This section provides details on the mathematical derivations of the multi-head self-attention (MHSA) [73] and multi-scale deformable cross-attention (MSDA) [30], which are integral components of our proposed HAPNet. MHSA is employed within the vision foundation model, while MSDA is utilized in the progressive heterogeneous feature aggregators.

A.1. Multi-head attention

Let $\boldsymbol{X}_q \in \mathbb{R}^{N_q \times C}$ be the query, and let $\boldsymbol{X}_f \in \mathbb{R}^{N_f \times C}$ be the key and value, where N_q and N_f are the respective sequence lengths. The multi-head attention (MHA), including both MHSA and multi-head cross-attention, can be formulated as follows:

$$\boldsymbol{Q}^i, \boldsymbol{K}^i, \boldsymbol{V}^i = \boldsymbol{X}_q\boldsymbol{U}_q^i,\ \boldsymbol{X}_f\boldsymbol{U}_k^i,\ \boldsymbol{X}_f\boldsymbol{U}_v^i,\quad i \in [1, n] \cap \mathbb{Z} \tag{A.1}$$

$$\boldsymbol{A}^i = \mathrm{Softmax}(\frac{\boldsymbol{Q}^i(\boldsymbol{K}^i)^\top}{\sqrt{C_v}}) \tag{A.2}$$

$$\mathrm{MHA}(\boldsymbol{X}_q, \boldsymbol{X}_f) = \mathrm{Concat}([\boldsymbol{A}^1\boldsymbol{V}^1; \ldots; \boldsymbol{A}^n\boldsymbol{V}^n])\boldsymbol{U}_m \tag{A.3}$$

where $\boldsymbol{U}_q^i \in \mathbb{R}^{C \times C_v}$, $\boldsymbol{U}_k^i \in \mathbb{R}^{C \times C_v}$ and $\boldsymbol{U}_v^i \in \mathbb{R}^{C \times C_v}$ denotes the linear projections of the query, key, and value in the ith attention head, respectively, $\boldsymbol{A}^i$ stores attention weights, and $\boldsymbol{U}_m \in \mathbb{R}^{n \cdot C_v \times C}$ represents the linear projection of n attention operations ($C_v = \frac{C}{n}$ by default). $\mathrm{Concat}([\cdot; \ldots; \cdot])$ denotes the feature concatenation operation performed on the results from n attention operations.

Table B.1
Backbone architectures, model complexities (in total parameters, Params), and inference speeds (in frames per second, FPS) for all evaluated networks, with all performance benchmarks conducted on an NVIDIA RTX 3090 GPU. "-" denotes that the corresponding results are unavailable.

Method	Backbone	Params (M)	FPS	Method	Backbone	Params (M)	FPS
MFNet [4]	Mini-Inception	0.72	194.0	CMX [47]	MiT-B4	139.9	11.72
RTFNet [3]	ResNet-152	254.51	26.18	CMNeXt [48]	MiT-B4	119.6	12.6
FuseSeg [28]	DenseNet-161	141.52	23.85	CRM-RGBTSeg [7]	Swin-B	193.0	–
EGFNet [62]	ResNet-152	62.82	8.58	GMNet [8]	ResNet-50	149.8	15.0
ABMDRNet [5]	ResNet-50	64.60	43.46	Omnivore [66]	Swin-B	95.7	–
ECGFNet [63]	MobileNetV2	26.06	38.37	TokenFusion [68]	MiT-B3	45.9	–
FEANet [24]	Modified ResNet	255.21	21.13	DFormer-L [51]	DFormer-L	39.0	–
SFAF-MA [64]	ResNet-152	156.0	21.8	LLE-Seg [65]	ConvNeXt V2	–	–
ABMDRNet+ [23]	ResNet-50	64.60	40.14	AsymFormer [69]	Modified MiT-B0	33.0	–
CAINet [9]	MobileNetV2	12.16	61.44	SA-Gate [70]	ResNet-50	110.85	18.36
EAEFNet [2]	ResNet-152	200.4	21.0	OmniVec [67]	ViT varients	95.7	–

A.2. Deformable attention

Compared to standard MHA, the deformable attention (DA) only attends a small set of points from the feature map as the reference points instead of the whole feature map, which reduces the computation complexity while maintaining competitive performance. The sampling process of DA can be achieved by predicting the sampling offsets and attention weights $\boldsymbol{A}^i$ through linear projections of the query $\boldsymbol{X}_q$. Let $\boldsymbol{p}_q = (h, w)$ index a 2-D reference point on the resized $\boldsymbol{X}_q$, where $\boldsymbol{X}_q$ has been restored to its original 2-D spatial dimensions. At this reference point, $\boldsymbol{X}_q(\boldsymbol{p}_q) \in \mathbb{R}^{1\times C}$ represents the feature vector corresponding to position $\boldsymbol{p}_q$. The DA operation can be formulated as:

$$\mathrm{DA}^i(\boldsymbol{X}_q(\boldsymbol{p}_q), \boldsymbol{X}_f) = \sum_{k=1}^{K} \boldsymbol{A}^i_{k,\boldsymbol{p}_q} \cdot \boldsymbol{X}_f(\boldsymbol{p}_q + \Delta^i_{k,\boldsymbol{p}_q})\boldsymbol{U}^i_v \tag{A.4}$$

where k indexes the sampled keys, and K is the total number of sampled keys. $\Delta^i_{k,\boldsymbol{p}_q}$ and $\boldsymbol{A}^i_{k,\boldsymbol{p}_q}$ denote the sampling offsets and attention weights of the kth sampling point on $\boldsymbol{p}_q$ in the ith attention head, respectively. The attention weights $\boldsymbol{A}^i_{k,\boldsymbol{p}_q}$ are normalized to satisfy $\sum_{k=1}^{K} \boldsymbol{A}^i_{k,\boldsymbol{p}_q} = 1$.

Let $\mathcal{X}_f = \{\boldsymbol{X}^l_f\}_{l=1}^{L}$ be the multi-scale feature maps, where $\boldsymbol{X}^l_f \in \mathbb{R}^{H_l\times W_l\times C}$, and let $\hat{\boldsymbol{p}}_q \in [0, 1]^2$ be the normalized coordinate of the reference point for the query $\boldsymbol{X}_q$. The DA process can be extended to its MSDA form as follows:

$$\mathrm{MSDA}^i(\boldsymbol{X}_q(\hat{\boldsymbol{p}}_q), \mathcal{X}_f) = \sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q} \cdot \boldsymbol{X}^l_f(\phi_l(\hat{\boldsymbol{p}}_q) + \Delta^{i,l}_{k,\boldsymbol{p}_q})\boldsymbol{U}^i_v \tag{A.5}$$

where l indexes the feature map level, and k indexes the sampling point. $\Delta^{i,l}_{k,\boldsymbol{p}_q}$ and $\boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q}$ denote the corresponding sampling offsets and attention weights, respectively, for the kth sampling point in the lth feature level and the ith attention head. The attention weights $\boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q}$ are normalized to satisfy $\sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q} = 1$. The transformation $\phi_l(\cdot)$ rescales the $\hat{\boldsymbol{p}}_q$ in (A.5) to the original spatial size of the lth feature level.

Appendix B. Supplementary content for tables

This section provides the detailed backbone configurations, model complexities, and inference speeds for all methods compared with HAPNet in this study, as shown in Table B.1.

References

[1] S.S. Shivakumar, et al., PST900: RGB-Thermal calibration, dataset and segmentation network, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2020, pp. 9441–9447.

[2] M. Liang, et al., Explicit attention-enhanced fusion for RGB-Thermal perception tasks, IEEE Robot. Autom. Lett. 8 (7) (2023) 4060–4067.

[3] Y. Sun, et al., RTFNet: RGB-Thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.

[4] Q. Ha, et al., 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.

[5] Q. Zhang, et al., 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, 2021, pp. 2633–2642.

[6] S. Guo, et al., LIX: Implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving, IEEE Trans. Image Process. 34 (2025) 7250–7263, http://dx.doi.org/10.1109/TIP.2025.3618378.

[7] U. Shin, et al., Complementary random masking for RGB-Thermal semantic segmentation, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 11110–11117.

[8] W. Zhou, et al., GMNet: Graded-feature multilabel-learning network for RGB-Thermal urban scene semantic segmentation, IEEE Trans. Image Process. 30 (2021) 7790–7802.

[9] Y. Lv, et al., Context-aware interaction network for RGB-T semantic segmentation, IEEE Trans. Multimed. (2024) DOI:10.1109/TMM.2023.3349072.

[10] Z. Wu, et al., S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving, IEEE Trans. Intell. Veh. 9 (2) (2024) 3940–3951, http://dx.doi.org/10.1109/TIV.2024.3357056.

[11] W. Zhou, et al., CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing, IEEE Trans. Intell. Veh. 9 (1) (2023) 1919–1929.

[12] J. Huang, et al., DepthMatch: Semi-supervised RGB-D scene parsing through depth-guided regularization, IEEE Signal Process. Lett. 32 (2025) 2549–2553, http://dx.doi.org/10.1109/LSP.2025.3575640.

[13] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[14] Z. Liu, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022.

[15] G. Tang, et al., TiCoSS: Tightening the coupling between semantic segmentation and stereo matching within a joint learning framework, IEEE Trans. Autom. Sci. Eng. 22 (2025) 18646–18658, http://dx.doi.org/10.1109/TASE.2025.3586286.

[16] Z. Liu, et al., A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 11976–11986.

[17] M.-J. Lee, et al., SG-RoadSeg+: End-to-end freespace detection upgraded at data, feature, and loss levels, IEEE Trans. Instrum. Meas. 74 (2025) 1–9, http://dx.doi.org/10.1109/TIM.2025.3579733.

[18] Y. Feng, et al., SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection, IEEE Trans. Instrum. Meas. 74 (2025) 1–9.

[19] H. Bao, et al., BEiT: BERT pre-training of image transformers, in: International Conference on Learning Representations, ICLR, 2022.

[20] Z. Peng, et al., BEiT v2: Masked image modeling with vector-quantized visual tokenizers, 2022, arXiv preprint arXiv:2208.06366.

[21] M. Oquab, et al., DINOv2: Learning robust visual features without supervision, Trans. Mach. Learn. Res. (2023).

[22] A. Krizhevsky, et al., ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (NeurIPS) 25 (2012) 1097–1105.

[23] S. Zhao, et al., Mitigating modality discrepancies for RGB-T semantic segmentation, IEEE Trans. Neural Networks Learn. Syst. (2023) http://dx.doi.org/10.1109/TNNLS.2022.3233089.

[24] F. Deng, et al., 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.
[25] J. Liu, et al., Revisiting modality-specific feature compensation for visible-infrared person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 7226–7240.
[26] D. Seichter, et al., Efficient RGB-D semantic segmentation for indoor scene analysis, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2021, pp. 13525–13531.
[27] Q. Zhang, et al., RGB-T salient object detection via fusing multi-level CNN features, IEEE Trans. Image Process. 29 (2019) 3321–3335.
[28] S. Y, et al., FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion, IEEE Trans. Autom. Sci. Eng. 18 (3) (2020) 1000–1011.
[29] J. Huang, et al., RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion, IEEE Trans. Intell. Veh. 10 (5) (2025) 3156–3165.
[30] X. Zhu, et al., Deformable DETR: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, ICLR, 2020.
[31] S. Sun, et al., ReMaX: Relaxing for better training on efficient panoptic segmentation, Adv. Neural Inf. Process. Syst. (NeurIPS) 36 (2024).
[32] Y.-H. Kim, et al., MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation, IEEE Robot. Autom. Lett. 6 (4) (2021) 6497–6504.
[33] N. Silberman, et al., Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, ECCV, Springer, 2012, pp. 746–760.
[34] J. Long, et al., Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3431–3440.
[35] L.-C. Chen, et al., DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[36] L. Chen, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.
[37] T.-Y. Lin, et al., Feature pyramid networks for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2117–2125.
[38] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (NeurIPS) 30 (2017) 5998–6008.
[39] S. Zheng, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 6881–6890.
[40] E. Xie, et al., SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. (NeurIPS) 34 (2021) 12077–12090.
[41] J. Dai, et al., Convolutional feature masking for joint object and stuff segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3992–4000.
[42] B. Hariharan, et al., Simultaneous detection and segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2014, pp. 297–312.
[43] B. Cheng, et al., Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Process. Syst. (NeurIPS) 34 (2021) 17864–17875.
[44] B. Cheng, et al., Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 1290–1299.
[45] W. Zhou, et al., DBCNet: Dynamic bilateral cross-fusion network for RGB-T urban scene understanding in intelligent vehicles, IEEE Trans. Syst. Man, Cybern.: Syst. 53 (12) (2023) 7631–7641.
[46] G. Huang, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 4700–4708.
[47] J. Zhang, et al., CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023) 14679–14694.
[48] Z. J, et al., Delivering arbitrary-modal semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 1136–1147.
[49] K. He, et al., Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 16000–16009.
[50] T. Xiao, et al., Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 418–434.
[51] B. Yin, et al., DFormer: Rethinking RGBD representation learning for semantic segmentation, Int. Conf. Learn. Represent. (ICLR) (2024).
[52] K. Li, et al., UniFormer: Unifying convolution and self-attention for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 12581–12600.
[53] S. Xie, et al., Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 1492–1500.
[54] J. Li, et al., RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing, IEEE Trans. Intell. Veh. 9 (7) (2024) 5163–5172.
[55] R. Bachmann, et al., MultiMAE: Multi-modal multi-task masked autoencoders, in: European Conference on Computer Vision, ECCV, Springer, 2022, pp. 348–367.
[56] S. Huang, et al., FaPN: Feature-aligned pyramid network for dense image prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 864–873.
[57] Z. Chen, et al., Vision transformer adapter for dense predictions, in: The Eleventh International Conference on Learning Representations, ICLR, 2023.
[58] X. Yang, et al., PolyMaX: General dense prediction with mask transformer, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2024, pp. 1050–1061.
[59] N. Carion, et al., End-to-end object detection with transformers, in: European Conference on Computer Vision, ECCV, Springer, 2020, pp. 213–229.
[60] M. Cordts, et al., The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 3213–3223.
[61] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, ICLR, 2018.
[62] W. Zhou, et al., Edge-aware guidance fusion network for RGB–thermal scene parsing, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (3) 2022, pp. 3571–3579.
[63] W. Zhou, et al., Embedded control gate fusion and attention residual learning for RGB–thermal urban scene parsing, IEEE Trans. Intell. Transp. Syst. 24 (5) (2023) 4794–4803.
[64] X. He, et al., SFAF-MA: Spatial feature aggregation and fusion with modality adaptation for RGB-Thermal semantic segmentation, IEEE Trans. Instrum. Meas. 72 (2023) 1–10.
[65] X. Guo, et al., Low-light enhancement and global-local feature interaction for RGB-T semantic segmentation, IEEE Trans. Instrum. Meas. (2025).
[66] R. Girdhar, et al., OMNIVORE: A single model for many visual modalities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 16102–16112.
[67] S. Srivastava, G. Sharma, OmniVec: Learning robust representations with cross modal sharing, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2024, pp. 1236–1248.
[68] Y. Wang, et al., Multimodal token fusion for vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 12186–12195.
[69] S. Du, et al., AsymFormer: Asymmetrical cross-modal representation learning for mobile platform real-time RGB-D semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 7608–7615.
[70] X. Chen, et al., Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: European Conference on Computer Vision, ECCV, Springer, 2020, pp. 561–577.
[71] H. Touvron, et al., Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, ICML, PMLR, 2021, pp. 10347–10357.
[72] A.P. Steiner, et al., How to train your ViT? Data, augmentation, and regularization in vision transformers, Trans. Mach. Learn. Res. (2022).
[73] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, Int. Conf. Learn. Represent. (ICLR) (2020).