Sigma 🐍: Siamese Mamba Network for Multi-Modal Semantic Segmentation

Zifu Wan[1]   Pingping Zhang[2]   Yuhao Wang[2]   Silong Yong[1]
Simon Stepputtis[1]   Katia Sycara[1]   Yaqi Xie[1]

[1]Robotics Institute, Carnegie Mellon University, USA
[2]School of Future Technology, Dalian University of Technology, China

Abstract

Multi-modal semantic segmentation significantly enhances AI agents' perception and scene understanding, especially under adverse conditions like low-light or overexposed environments. Leveraging additional modalities (X-modality) like thermal and depth alongside traditional RGB provides complementary information, enabling more robust and reliable prediction. In this work, we introduce Sigma, a Siamese Mamba network for multi-modal semantic segmentation utilizing the advanced Mamba. Unlike conventional methods that rely on CNNs, with their limited local receptive fields, or Vision Transformers (ViTs), which offer global receptive fields at the cost of quadratic complexity, our model achieves global receptive fields with linear complexity. By employing a Siamese encoder and innovating a Mamba-based fusion mechanism, we effectively select essential information from different modalities. A decoder is then developed to enhance the channel-wise modeling ability of the model. Our proposed method is rigorously evaluated on both RGB-Thermal and RGB-Depth semantic segmentation tasks, demonstrating its superiority and marking the first successful application of State Space Models (SSMs) in multi-modal perception tasks. Code is available at https://github.com/zifuwan/Sigma.

1. Introduction

Semantic segmentation, aiming to assign a semantic label for each pixel within an image, has been increasingly significant for AI agents to accurately perceive their environment [4, 18]. However, current vision models still struggle in challenging conditions like low light or obstructive elements such as sun glare and fire [4, 58, 59]. With the goal of enhancing segmentation under such challenging conditions, additional modalities like thermal and depth are utilized to enhance vision system robustness. With the supplementary information, the robustness and capabilities of vision pipelines can be improved [23, 30, 60]. However, utilizing multiple modalities introduces additional chal-
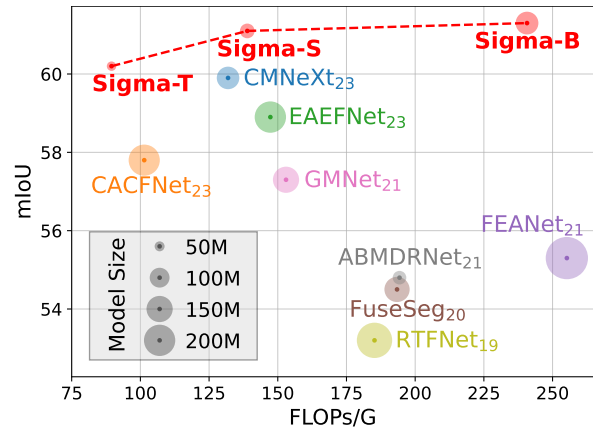

Figure 1. Computation and model size comparison of Sigma and other methods on MFNet dataset [19]. The size of each circle indicates the model size (parameters).

lenges, namely the alignment and fusion of the information provided through these additional channels [11].

Previous approaches in multi-modal semantic segmentation rely on Convolutional Neural Networks (CNN) or Vision Transformers (ViT). While CNN-based approaches [12, 22] are known for their scalability and linear complexity, they suffer from the small receptive field limited by the kernel size, leading to the local reductive bias. Besides, CNNs utilize a weight-sharing mechanism across different parts of the input, limiting its flexibility in adapting to unseen or low-quality images. In contrast, ViT-based methods [1, 23, 42] offer enhanced visual modeling by leveraging global receptive fields and dynamic weights. However, their self-attention mechanism leads to quadratic complexity in terms of input sizes [14], raising efficiency concerns. Attempts to improve this efficiency by reducing the dimensions or strides of the processing windows compromise the extent of the receptive fields [54].

To address the aforementioned limitations, Selective Structured State Space Models (S6), or Mamba [14] have gained more popularity due to their global receptive field coverage and dynamic weights with linear complexity. Mamba has shown outstanding effectiveness in tasks in-

volving long sequence modeling, notably in natural language processing [14]. Furthermore, more studies have explored its potential for vision-related applications, such as image classification [26], medical image segmentation [31, 32, 45, 49, 53], and 3D scene comprehension [21]. Inspired by these benefits, we introduce **Sigma**🦭, a Siamese Mamba network for multi-modal information fusion and apply it to the challenging domain of semantic segmentation.

Our Sigma integrates a Siamese encoder for modality-specific feature extraction, fusion modules for aggregating information from different modalities, and a decoder for enhancing both spatial and channel-specific information. The encoder backbone utilizes cascaded *Visual State Space Blocks (VSSB)* with downsampling to extract multi-scale information from various modalities. Subsequently, the extracted features are inputted to a fusion module at each level, where multi-modal features initially interact through a *Cross Mamba Block (CroMB)* to augment cross-modal information. Then, the enhanced features are processed by a *Concat Mamba Block (ConMB)*, which employs an attention mechanism to select pertinent information from each modality. Finally, the fused features are sent to multi-level *Channel-Aware Visual State Space Blocks (CAVSSB)* to effectively capture multi-scale long-range information.

Comprehensive experiments on both RGB-Thermal [19, 36] and RGB-Depth datasets [37, 39] demonstrate that Sigma outperforms state-of-the-art models in both accuracy and efficiency, as depicted in Fig. 1.

Our contributions can be summarized as follows:

- To our best knowledge, our work marks the first successful application of state space models, specifically Mamba, in multi-modal semantic segmentation.

- We introduce a Mamba-based fusion mechanism alongside a channel-aware decoder, to efficiently extract information across different modalities and integrate them seamlessly.

- Comprehensive evaluations in RGB-Thermal and RGB-Depth domains showcase our method's superior accuracy and efficiency, setting a new benchmark for future investigations into Mamba's potential in multi-modal learning.

2. Related Work

Multi-Modal Semantic Segmentation. Multi-modal semantic segmentation typically integrates the RGB modality with other complementary modalities such as thermal, depth, and LiDAR [11, 62]. These supplementary modalities provide crucial information to the vision system across various scenarios. To optimize the utilization of these additional modalities, the development of effective feature ex-

tractors and fusion mechanisms has been of paramount importance.

In RGB-Thermal semantic segmentation, early approaches commonly employ encoder-decoder architectures with shortcut connections [19, 36, 41], dense connections [41, 50, 64], and dilated convolutions [51, 63]. To address the limited context understanding of CNNs, many methods apply attention mechanisms during feature fusion [22, 41, 64]. With the rise of Transformers, methods like CMX [23] utilize SegFormer [48] for feature extraction and incorporate cross-attention modules for feature fusion. Building on CMX, CMNeXt [60] introduces a self-query hub to select informative features from various modalities. SegMiF [25] employs a cascading structure with a hierarchical interactive attention mechanism to ensure precise mapping across modalities.

In RGB-Depth semantic segmentation, approaches effective in RGB-Thermal semantic segmentation, such as CMX [23] and CMNeXt [60], also show impressive performance. Additionally, self-supervised pre-training is being explored in RGB-Depth perception. MultiMAE [1] uses a masked autoencoder with pseudo-labeling, while DFormer [57] integrates RGB and Depth modalities within the pre-training architecture to learn transferable multi-modal representations.

Although the aforementioned Transformer-based methods have shown promising results in RGB-X semantic segmentation, the quadratic scaling nature of the self-attention mechanism limits the length of input sequences. In contrast, our proposed Sigma processes very long sequences, preserving all valuable information while requiring significantly less computation.

State Space Models. State Space Models (SSM) [15, 38], inspired by linear time-invariant systems, are considered as efficient sequence-to-sequence models. Recently, the Structured State-Space Sequence model (S4) [15] has emerged as a pioneering work in capturing long-range dependencies. Furthermore, with the selective mechanism introduced into S4, Mamba [14] surpasses Transformers and other advanced architectures. Due to the remarkable performance, researchers have extended it to the field of computer vision. For example, Vision Mamba [67] integrates SSM with bidirectional scanning, making each patch related to another. VMamba [26] extends scanning in four directions to fully capture interrelations among image patches. Besides, state space models have been extended to medical image segmentation [31, 32, 45, 49], image restoration [17] and point cloud analysis [21], all showing competitive results with lower complexity. However, these recent works directly employ SSM as a plug-and-play module, without the in-depth design tailored to specific tasks. Additionally, there is a lack of exploration of SSM in multi-modal tasks. Thus, we propose a Mamba-based fusion mechanism for
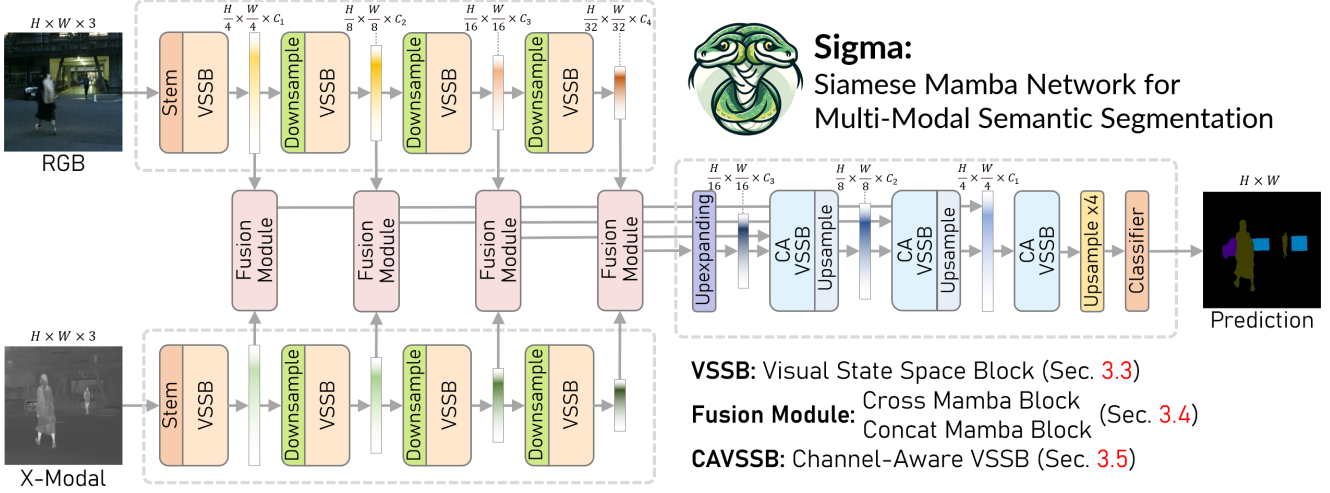
Figure 2. Overall architecture of the proposed Sigma.

cross-modality exchange and concatenated feature enhancement, along with a channel-aware decoder for channel information selection. By leveraging the specialized design of SSM for multi-modal tasks, our approach attains enhanced accuracy while maintaining low complexity.

3. Sigma: Siamese Mamba Network

In this section, we give a detailed illustration of our proposed **Siamese Mamba Network (Sigma)** for multi-modal semantic segmentation. Firstly, we provide basic information of State Space Models. Subsequently, we present an overview of our Sigma architecture, followed by in-depth discussions of the encoder, fusion module, and decoder.

3.1. Preliminaries

State Space Models. State Space Models (SSM) [15, 16, 38] represent a class of sequence-to-sequence modeling systems characterized by constant dynamics over time, a property also known as linear time-invariant. With linear complexity, SSM can effectively capture the inherent dynamics through an implicit mapping to latent states, which can be defined as:

$$y(t) = Ch(t) + Dx(t), \dot{h}(t) = Ah(t) + Bx(t). \quad (1)$$

Here, $x(t) \in \mathbb{R}, h(t) \in \mathbb{R}^N$, and $y(t) \in \mathbb{R}$ denotes the input, hidden state, and the output, respectively. N is the state size and $\dot{h}(t)$ refers to the time derivative of $h(t)$. Additionally, $A \in \mathbb{R}^{N \times N}, B \in \mathbb{R}^{N \times 1}, C \in \mathbb{R}^{1 \times N}$, and $D \in \mathbb{R}$ are the system matrices. To process discrete sequences like image and text, SSMs adopt Zero-Order Hold (ZOH) discretization [15] to map the input sequence $\{x_1, x_2, ..., x_K\}$ to the output sequence $\{y_1, y_2, ..., y_K\}$. Specifically, suppose $\Delta \in \mathbb{R}^D$ is the pre-defined timescale parameter to map continuous parameters A, B into a discrete space, the dis-

cretization process can be formulated as:

$$\overline{A} = \exp(\Delta A), \ \overline{B} = (\Delta A)^{-1}(\exp(A) - I)\Delta B, \ \overline{C} = C, \quad (2)$$

$$y_k = \overline{C}h_k + \overline{D}x_k, h_k = \overline{A}h_{k-1} + \overline{B}x_k. \quad (3)$$

Here, all the matrices keep the same dimension as the operation iterates. Notably, $\overline{D}$, serving as a residual connection, is often discarded in the equation:

$$y_k = \overline{C}h_k. \quad (4)$$

Besides, following Mamba [14], the matrix $\overline{B}$ can be approximated by the first-order Taylor series:

$$\overline{B} = (\exp(A) - I)A^{-1}B \approx (\Delta A)(\Delta A)^{-1}\Delta B = \Delta B. \quad (5)$$

Selective Scan Mechanism. While SSMs are effective for modeling discrete sequences, they encounter invariant parameters regardless of differences in the input. To address this limitation, the Selective State Space Model (S6, *a.k.a* Mamba) [14] is introduced, ensuring the input-dependency. In Mamba, the matrices $B \in \mathbb{R}^{L \times N}, C \in \mathbb{R}^{L \times N}$, and $\Delta \in \mathbb{R}^{L \times D}$ are derived from the input $x \in \mathbb{R}^{L \times D}$, enabling the model to be contextually aware of the input. With this selection mechanism, Mamba is capable of effectively modeling the complex interactions present in long sequences.

3.2. Overall Architecture

As illustrated in Fig. 2, our proposed method comprises a siamese mamba encoder (Sec. 3.3), a fusion module (Sec. 3.4), and a channel-aware mamba decoder (Sec. 3.5). During the encoding phase, four Visual State Space Blocks (VSSB) with downsampling operations are sequentially cascaded to extract multi-level image features. The two encoder branches share weights to reduce computational

complexity. Subsequently, features from each level, derived from the two branches, are processed through a fusion module. In the decoding phase, the fused features at each level are further enhanced by a Channel-Aware Visual State Space Block (CAVSSB) with an upsampling operation. Ultimately, the final feature is forwarded to a classifier to generate the prediction.
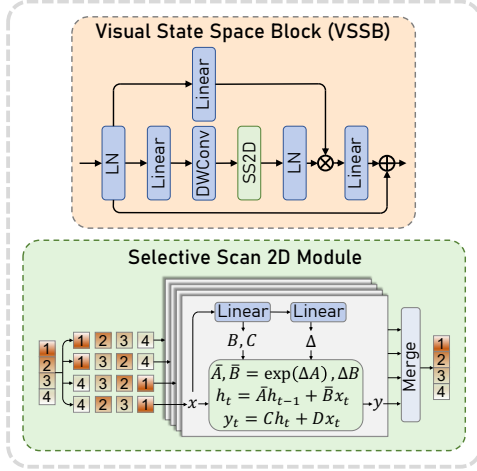
3.3. Siamese Mamba Encoder

Figure 3. The Visual State Space Block (VSSB) used in the encoder.

The encoder has two weight-sharing branches which take an RGB image and an X-modality image as input, denoted as $I_{\text{rgb}}, I_{\text{x}} \in \mathbb{R}^{H \times W \times 3}$. Here, H and W represent the height and width of the input modalities. The encoder starts with a stem module similar to ViT [10] which partitions the input into patches, generating feature maps $\mathcal{F}_i^1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$, where $i \in \{\text{rgb}, \text{x}\}$ refers to the RGB or X modality. The features are continually processed by three sets of downsampling and VSSB, yielding multi-scale features $\{\mathcal{F}_i^2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}, \mathcal{F}_i^3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}, \mathcal{F}_i^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}\}$. The details of VSSB are introduced as follows.

Visual State Space Block (VSSB). Following VMamba [26], we implement the VSSB with Selective Scan 2D (SS2D) modules. As shown in Fig. 3, the input feature is processed by a sequence of linear projection (Linear), Depth-wise Convolution (DWConv). Then, an SS2D module with a residual connection is used to model long-range spatial information.

SS2D Module. Within the SS2D module, the input feature is first flattened to four sequences from four directions (top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right). Then four distinctive selective scan modules [14] are used to extract
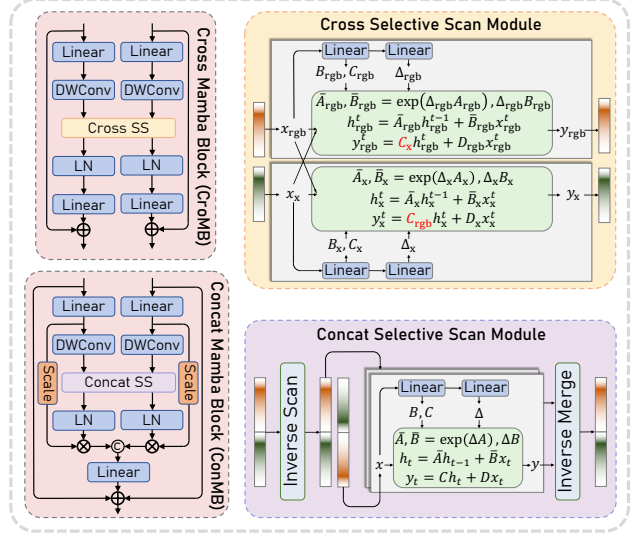

Figure 4. The left part shows the Cross Mamba Block (CroMB) and Concat Mamba Block (ConMB). The right part shows the Cross Selective Scan (Cross SS) module and Concat Selective Scan (Concat SS) module.

multi-direction information, where each of them captures the long-range dependencies of the sequence with the operation in Eq. 3. Lastly, the four sequences are reversed to the same direction and summed.

3.4. Fusion Module

The detailed architecture of the feature fusion module is illustrated in Fig. 4. It consists of a Cross Mamba Block (CroMB) and a Concat Mamba Block (ConMB). Specifically, CroMB employs a cross-multiplication mechanism to enhance the features with one another, and ConMB applies the selective scan mechanism to concatenated features to obtain the fusion result. Suppose the features from the k-th Siamese encoder block are represented as $\mathcal{F}_{\text{rgb}/\text{x}}^k \in \mathbb{R}^{H_k \times W_k \times C_k}$, then the entire fusion process can be represented as:

$$\hat{\mathcal{F}}_{\text{rgb}}^k, \ \hat{\mathcal{F}}_{\text{x}}^k = \texttt{CroMB}(\mathcal{F}_{\text{rgb}}^k, \ \mathcal{F}_{\text{x}}^k), \qquad (6)$$

$$\mathcal{F}_{\text{fuse}}^k = \texttt{ConMB}(\hat{\mathcal{F}}_{\text{rgb}}^k, \ \hat{\mathcal{F}}_{\text{x}}^k). \qquad (7)$$

Here, $\hat{\mathcal{F}}_{\text{rgb}}^k, \hat{\mathcal{F}}_{\text{x}}^k$, and $\mathcal{F}_{\text{fuse}}^k$ remains the original dimension as $\mathbb{R}^{H_k \times W_k \times C_k}$. Details of CroMB and ConMB are as follows.

CroMB. As demonstrated in the upper part of Fig. 4, two input features are first processed by linear layers and depth-wise convolutions, then sent to the Cross Selective Scan (Cross SS) module. According to the selection mechanism of Mamba mentioned in Sec. 3.1, the system matrices B, C and Δ are generated to enable the context-aware ability of the model. Here, linear projection layers are utilized to generate the matrices. According to Eq. 4, matrix

C is used to decode the information from the hidden state h_k to obtain the output y_k. Inspired by the cross-attention mechanism [3], we facilitate information exchange among multiple selective scan modules. In particular, the process can be represented as:

$$\overline{A}_{\mathsf{rgb}} = \exp(\Delta_{\mathsf{rgb}}A_{\mathsf{rgb}}), \ \overline{A}_{\mathsf{x}} = \exp(\Delta_{\mathsf{x}}A_{\mathsf{x}}), \quad (8)$$

$$\overline{B}_{\mathsf{rgb}} = \Delta_{\mathsf{rgb}}B_{\mathsf{rgb}}, \ \overline{B}_{\mathsf{x}} = \Delta_{\mathsf{x}}B_{\mathsf{x}}, \quad (9)$$

$$h_{\mathsf{rgb}}^t = \overline{A}_{\mathsf{rgb}}h_{\mathsf{rgb}}^{t-1} + \overline{B}_{\mathsf{rgb}}x_{\mathsf{rgb}}^t, \ h_{\mathsf{x}}^t = \overline{A}_{\mathsf{x}}h_{\mathsf{x}}^{t-1} + \overline{B}_{\mathsf{x}}x_{\mathsf{x}}^t, \quad (10)$$

$$y_{\mathsf{rgb}}^t = C_{\mathsf{x}}h_{\mathsf{rgb}}^t + D_{\mathsf{rgb}}x_{\mathsf{rgb}}^t, \ y_{\mathsf{x}}^t = C_{\mathsf{rgb}}h_{\mathsf{x}}^t + D_{\mathsf{x}}x_{\mathsf{x}}^t, \quad (11)$$

$$y_{\mathsf{rgb}} = [y_{\mathsf{rgb}}^1, y_{\mathsf{rgb}}^2, \ldots, y_{\mathsf{rgb}}^l], \ y_{\mathsf{x}} = [y_{\mathsf{x}}^1, y_{\mathsf{x}}^2, \ldots, y_{\mathsf{x}}^l]. \quad (12)$$

Here, $x_{\mathsf{rgb/x}}^t$ represents the input at time step t, and $y_{\mathsf{rgb/x}}$ denotes the selective scan output. C_{x} and C_{rgb} are the cross-modal matrices used for recovering the outputs at each time step from the hidden states.

ConMB. In CroMB, the features from two modalities interact with each other and obtain the cross-modal-enhanced features. To further obtain a fused feature containing vital information from both modalities, we propose ConMB to integrate the outputs from CroMB.

Specifically, the outputs $\hat{\mathcal{F}}_{\mathsf{rgb}}^k, \hat{\mathcal{F}}_{\mathsf{x}}^k \in \mathbb{R}^{H_k \times W_k \times C_k}$ from CroMB are first processed by linear and depth-wise convolution layers, then sent to the Concat Selective Scan (Concat SS) module. Within the Concat SS module, the two features are first flattened and then concatenated in the sequence length dimension. This provides a sequence $\mathcal{S}_{\mathsf{Concat}}^k$ of shape $\mathbb{R}^{(2 \times H_k \times W_k) \times C_k}$. Besides, to comprehensively capture information from two modalities, we inversely scan the concatenated sequence $\mathcal{S}_{\mathsf{Concat}}^k$ to obtain an additional sequence $\mathcal{S}_{\mathsf{Inverse}}^k \in \mathbb{R}^{(2 \times H_k \times W_k) \times C_k}$. Afterwards, each sequence is processed to capture long-range dependencies, obtaining $\hat{\mathcal{S}}_{\mathsf{Concat}}^k$ and $\hat{\mathcal{S}}_{\mathsf{Inverse}}^k$. Then the inversed sequence output is flipped back and added with the processed sequence. The summed sequence is separated to recover two outputs. This process can be represented as:

$$\widetilde{\mathcal{F}}_{\mathsf{rgb}}^k = \mathtt{DWConv}(\mathtt{Linear}(\hat{\mathcal{F}}_{\mathsf{rgb}}^k)), \quad (13)$$

$$\widetilde{\mathcal{F}}_{\mathsf{x}}^k = \mathtt{DWConv}(\mathtt{Linear}(\hat{\mathcal{F}}_{\mathsf{x}}^k)), \quad (14)$$

$$\mathcal{S}_{\mathsf{Concat}}^k = \mathtt{Concat}(\widetilde{\mathcal{F}}_{\mathsf{rgb}}^k, \widetilde{\mathcal{F}}_{\mathsf{x}}^k), \quad (15)$$

$$\mathcal{S}_{\mathsf{Inverse}}^k = \mathtt{Inverse}(\mathcal{S}_{\mathsf{Concat}}^k), \quad (16)$$

$$\hat{\mathcal{S}}_{\mathsf{Concat}}^k, \hat{\mathcal{S}}_{\mathsf{Inverse}}^k = \mathtt{SSM}(\mathcal{S}_{\mathsf{Concat}}^k), \mathtt{SSM}(\mathcal{S}_{\mathsf{Inverse}}^k), \quad (17)$$

$$\overline{\mathcal{F}}_{\mathsf{RGB}}^k, \overline{\mathcal{F}}_{\mathsf{X}}^k = \mathtt{Seperate}(\hat{\mathcal{S}}_{\mathsf{Concat}}^k + \mathtt{Inverse}(\hat{\mathcal{S}}_{\mathsf{Inverse}}^k)). \quad (18)$$

After obtaining the scanned features $\overline{\mathcal{F}}_{\mathsf{rgb}}^k$ and $\overline{\mathcal{F}}_{\mathsf{x}}^k$, they are multiplied with two scaling parameters and concatenated in the channel dimension. Finally, a linear projection layer is used to reduce the feature shape to $\mathbb{R}^{H_k \times W_k \times C_k}$.

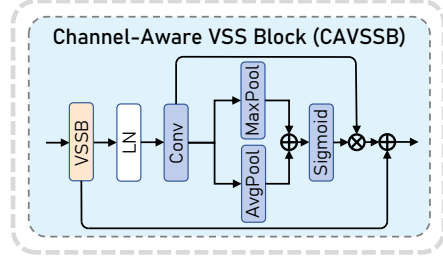## 3.5. Channel-Aware Mamba Decoder

Figure 5. The Channel-Aware Visual State Space Block (CAVSSB) used in the Sigma decoder.

To obtain the final prediction, we introduce a channel-aware Mamba decoder, as shown in Figure 2. The decoder consists of three groups of Channel-Aware Visual State Space Blocks (CAVSSB) and upsampling operations. Finally, the decoded features are classified by MLP layers for semantic segmentation prediction. Details of the CAVSSB are shown in Figure 5.

The CAVSSB incorporates a VSSB to process the fused features. While VSSBs are adept at extracting global spatial context, they fall short in learning inter-channel information. To address this issue, we apply channel attentions to select vital channel-wise representations. Specifically, our channel-attention operation consists of the average pooling and max pooling. This straightforward implementation offers effective channel selection while maintaining efficiency. Additionally, two residual connections are used to acquire the final output. In this manner, we form a spatial-channel-aware scheme for robust decoding.

4. Experiments

4.1. Experimental Settings

Datasets. To verify the effectiveness of Sigma, we conduct extensive experiments on two publicly available RGB-Thermal (RGB-T) semantic segmentation datasets, namely MFNet [19] and PST900 [36]. In addition, to better understand the generalizability of Sigma, we conduct experiments on two RGB-Depth (RGB-D) datasets, including NYU Depth V2 [37] and SUN RGB-D [39]. The details of these datasets are as follows.

- **MFNet dataset** contains 820 daytime and 749 nighttime RGB-T images with a resolution of 640×480. The dataset includes eight common classes of objects in driving scenarios. We follow the training/testing split of [23].

- **PST900 dataset** provides 597 and 288 calibrated RGB-T images with a resolution of 1280×720 for training and validation. This dataset is collected from

Method	Backbone	Params (M)	FLOPs (G)	Unlabeled	Car	Person	Bike	Curve	Stop	Guardrail	Cone	Bump	mIoU
MFNet[17] [19]	–	–	–	96.9	65.9	58.9	42.9	29.9	9.9	0.0	25.2	27.7	39.7
RTFNet[19] [40]	ResNet-152	245.7	185.2	98.5	87.4	70.3	62.7	45.3	29.8	0.0	29.1	55.7	53.2
PSTNet[20] [36]	ResNet-18	105.8	123.4	97.0	76.8	52.6	55.3	29.6	25.1	**15.1**	39.4	45.0	48.4
FuseSeg[20] [41]	DenseNet-161	141.5	193.4	97.6	87.9	71.7	64.6	44.8	22.7	6.4	46.9	47.9	54.5
U2Fusion[20] [50]	VGG-16	–	–	97.7	82.8	64.8	61.0	32.3	20.9	–	45.2	50.2	50.8
AFNet[21] [51]	ResNet-50	–	–	98.0	86.0	67.4	62.0	43.0	28.9	4.6	44.9	56.6	54.6
ABMDRNet[21] [61]	ResNet-50	64.6	194.3	**98.6**	84.8	69.6	60.3	45.1	33.1	5.1	47.4	50.0	54.8
FEANet[21] [7]	ResNet-152	337.1	255.2	98.3	87.8	71.1	61.1	46.5	22.1	6.6	55.3	48.9	55.3
GMNet[21] [64]	ResNet-50	149.8	153.0	97.5	86.5	73.1	61.7	44.0	42.3	14.5	48.7	47.4	57.3
TarDAL[22] [24]	–	297	–	97.6	80.7	67.1	60.1	34.9	10.5	–	38.7	45.5	48.6
CMX[22] [23]	MiT-B4	139.9	134.3	98.3	90.1	75.2	64.5	50.2	35.3	8.5	54.2	60.6	59.7
EAEFNet[23] [22]	ResNet-152	200.4	147.3	–	87.6	72.6	63.8	48.6	35.0	14.2	52.4	58.3	58.9
CACFNet[23] [63]	ConvNeXt-B	198.6	101.4	–	89.2	69.5	63.3	46.6	32.4	7.9	54.9	58.3	57.8
PAIF[23] [28]	–	260	–	–	88.1	72.4	60.8	–	–	–	56.0	57.2	56.5
CENet[23] [12]	ResNet-50	–	–	–	85.8	70.0	61.4	46.8	29.3	8.7	47.8	56.9	56.1
SegMiF[23] [25]	MiT-B3	–	–	98.1	87.8	71.4	63.2	47.5	31.1	–	48.9	50.3	56.1
CMNeXt[23] [60]	MiT-B4	119.6	131.9	98.4	**91.5**	75.3	67.6	50.5	40.1	9.3	53.4	52.8	59.9
Sigma (Ours)	VMamba-T	48.3	89.5	98.4	90.8	75.2	66.6	48.2	38.0	8.7	55.9	60.4	60.2
Sigma (Ours)	VMamba-S	69.8	138.9	98.5	91.5	**75.8**	67.8	49.6	41.8	9.6	54.8	60.4	61.1
Sigma (Ours)	VMamba-B	121.4	240.7	98.5	91.1	75.2	68.0	50.8	**43.0**	9.7	**57.6**	57.9	**61.3**

(a) Quantitative comparisons on MFNet day-night evaluation set [19] (9 classes)

Method	Backbone	Background	Extinguisher	Backpack	Hand-Drill	Survivor	mIoU
MFNet[17] [19]	–	98.6	60.4	64.3	41.1	20.7	57.0
RTFNet[19] [40]	ResNet-152	98.9	52.0	75.3	25.4	36.4	57.6
PSTNet[20] [36]	ResNet-18	98.9	70.1	69.2	53.6	50.0	68.4
ABMDRNet[21] [61]	ResNet-50	99.0	66.2	67.9	61.5	62.0	71.3
GMNet[21] [64]	ResNet-50	99.4	73.8	83.8	85.2	78.4	84.1
CCFFNet[22] [46]	ResNet-101	99.4	**82.8**	75.8	79.9	72.7	82.1
EGFNet[23] [8]	ResNet-101	**99.6**	80.0	**90.6**	76.1	80.9	85.4
CACFNet[22] [63]	ConvNeXt-B	**99.6**	82.1	89.5	80.9	80.8	86.6
CAINet[24] [30]	MobileNet-V2	99.5	80.3	88.0	77.2	78.7	84.7
Sigma (Ours)	VMamba-T	**99.6**	81.9	89.8	88.7	**82.7**	**88.6**
Sigma (Ours)	VMamba-S	**99.6**	79.4	88.7	**90.2**	81.2	87.8

(b) Quantitative comparisons on PST900 evaluation set [36] (5 classes)

Table 1. Quantitative comparisons for semantic segmentation of RGB-T images on MFNet [19] and PST900 [36] datasets. The best and second best results in each block are highlighted in **bold** and underline, respectively.

DARPA Subterranean Challenge and annotated with four classes.

- **NYU Depth V2 dataset** contains 1449 RGB-D images annotated with 40 semantic classes with a resolution of 640×480. We divide them into 795/654 for training/testing following previous work [13, 20].

- **SUN RGB-D dataset** incorporates 10335 RGB-D images with 37 classes. We follow the common setting [20] to split 5285/5050 for training/testing, and reshape the images to 640×480.

Evaluation Metrics. Following previous work [20, 60], we report the Intersection over Union (mIoU) averaged across the semantic classes for evaluation.

Training Settings. We follow [23] to use the AdamW optimizer [29] with an initial learning rate $6e^{-5}$ and weight decay 0.01. The model is trained with a batch size of 8 for 500 epochs. We utilize the ImageNet-1K [33] pre-trained model provided by VMamba [26] for the Siamese image encoder, leading to three different sizes of models (Sigma-Tiny, Sigma-Small, and Sigma-Base). More details of the experimental settings are described in the supplementary material.

4.2. Quantitative and Qualitative Results

RGB-T Semantic Segmentation. Table 1(a) shows the per-class semantic segmentation results, alongside comparisons of the model size and computational complexity, on the MFNet dataset. It is observed that our tiny model surpasses other compared methods with fewer model parameters and FLOPs, and our base model achieves a 1.1% performance improvement compared to the tiny variant. In addition, as shown in Table 1(b), our method outperforms other methods by more than 2% on the PST900 dataset, demonstrating the superiority of our proposed method.

The qualitative analysis in Fig. 6 reveals that our model outperforms the baseline models by generating more comprehensive segmentations and accurate classifications, notably in identifying intricate features like tactile paving and bollards. This indicates that our method can extract vital features and detailed information to improve the prediction.

RGB-D Semantic Segmentation. In Table 2, we compare Sigma against various RGB-D methods to validate its generalization capability. Remarkably, our small model surpasses the performance of CMNeXt [60] while employing only 69.8M parameters, which is 49.8M less than CMNeXt. This demonstrates our method's superior balance between
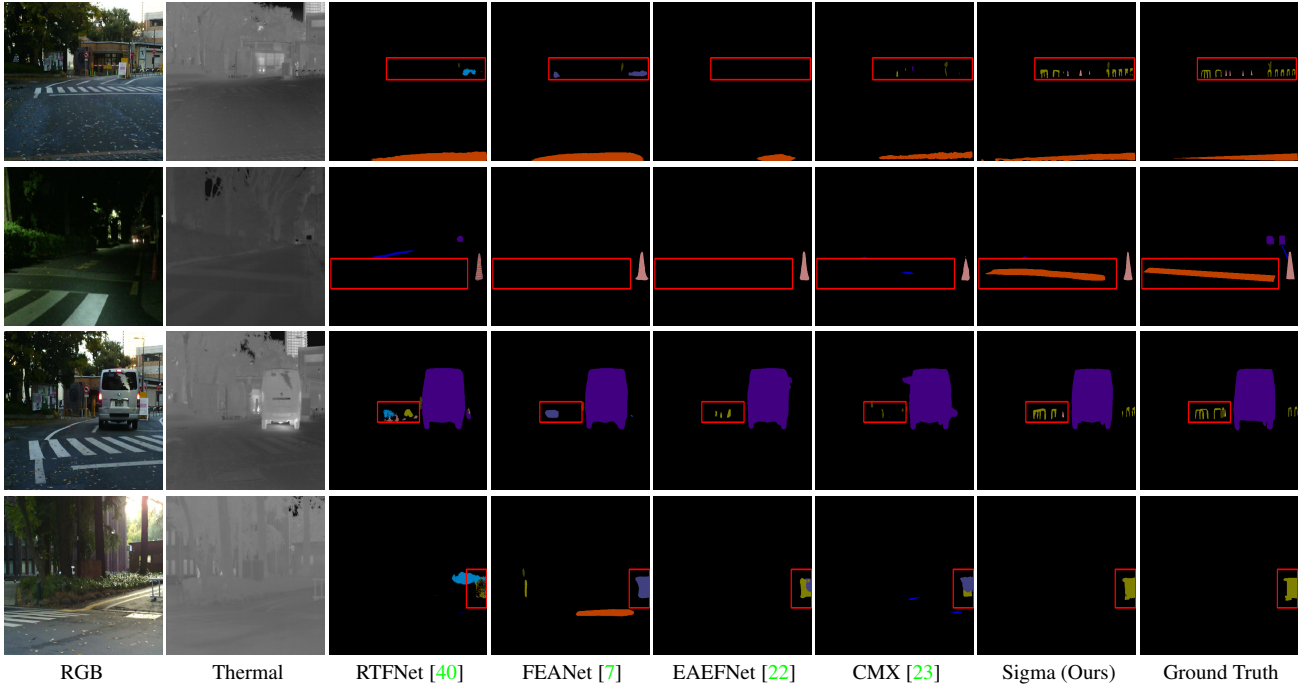
Figure 6. Qualitative comparison on MFNet [19] dataset. More results can be found in the supplementary material.

| RGB | Thermal | RTFNet [40] | FEANet [7] | EAEFNet [22] | CMX [23] | Sigma (Ours) | Ground Truth |

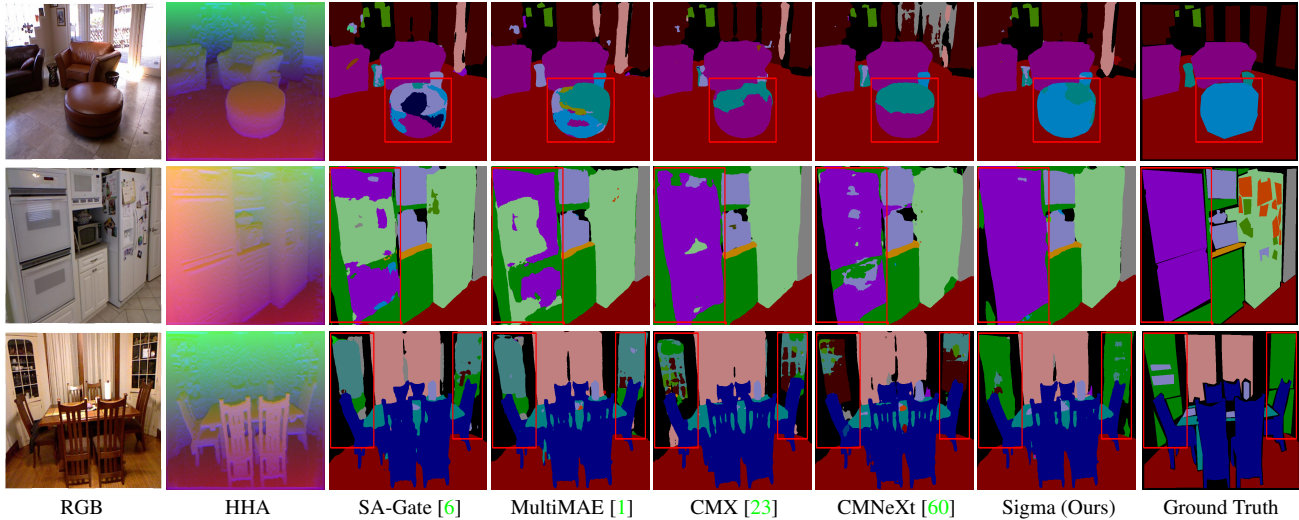| RGB | HHA | SA-Gate [6] | MultiMAE [1] | CMX [23] | CMNeXt [60] | Sigma (Ours) | Ground Truth |

Figure 7. Qualitative comparison on NYU Depth V2 [37] dataset. We use HHA images for better visualization of depth modality. More results can be found in the supplementary material.

accuracy and efficiency.

Fig. 7 shows the capability to generate more coherent segmentation by utilizing depth information effectively. For instance, in the case of the round chair adjacent to the sofa, shadows cause baseline models to fragment the chair into multiple segments. Our method successfully recognizes it as a singular entity, highlighting its proficiency in leveraging depth data for segmentation.

Qualitative Analysis of Dual-Modality Learning. In Figure 8, we examine the contributions of both RGB and thermal modalities to the final prediction. By leveraging information from both modalities, our model achieves

more comprehensive segmentation and more precise boundary delineation. The enhanced outcomes stem from fusing valuable information in both RGB and thermal modalities. Specifically, RGB enhances color distinction, whereas thermal imaging excels in texture differentiation. Integrating these modalities results in improved segmentation accuracy.

4.3. Ablation Studies

Effectiveness of Key Components. As detailed in Table 3, we carried out ablation studies with Sigma-Tiny on the MFNet [19] dataset. Compared to the complete model, eliminating CroMB and ConMB individually results in
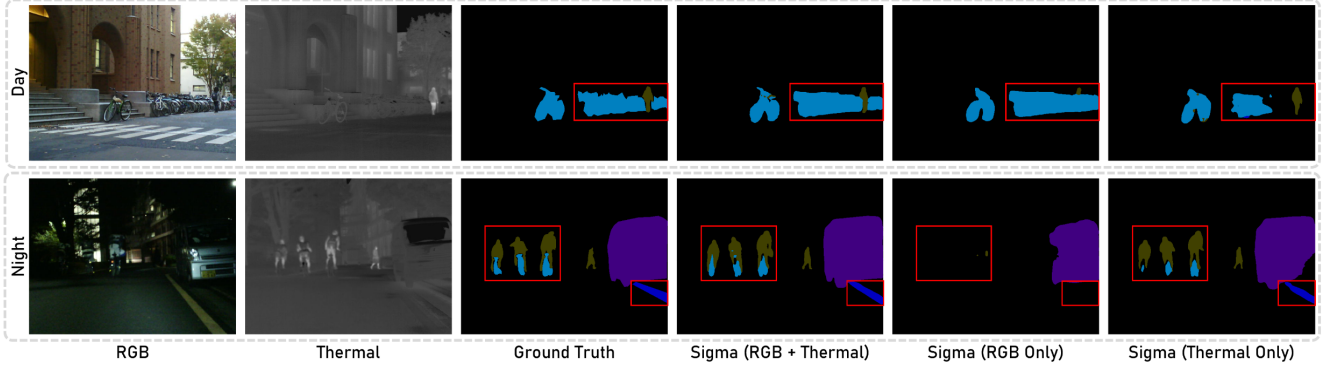
Figure 8. Comparative analysis of semantic segmentation results: single-modal vs. multi-modal approach.

Method	Backbone	Params	NYU Depth V2		SUN RGB-D	
			Input Size	mIoU	Input Size	mIoU
ACNet[20]	ResNet-50	116.6	480×640	48.3	530×730	48.1
SA-Gate[6]	ResNet-101	110.9	480×640	52.4	530×730	49.4
CEN[44]	ResNet-101	118.2	480×640	51.7	530×730	50.2
CEN[44]	ResNet-152	133.9	480×640	52.5	530×730	51.1
SGNet[5]	ResNet-101	64.7	480×640	51.1	530×730	48.6
ShapeConv[2]	ResNext-101	86.8	480×640	51.3	530×730	48.6
ESANet[35]	ResNet-34	31.2	480×640	50.3	480×640	48.2
FRNet[66]	ResNet-34	85.5	480×640	53.6	530×730	51.8
PGDENet[65]	ResNet-34	100.7	480×640	53.7	530×730	51.0
EMSANet[34]	ResNet-34	46.9	480×640	51.0	530×730	48.4
TokenFusion[43]	MiT-B2	26.0	480×640	53.3	530×730	50.3
TokenFusion[43]	MiT-B3	45.9	480×640	54.2	530×730	51.0
MultiMAE[1]	ViT-B	95.2	640×640	56.0	640×640	51.1
Omnivore[13]	Swin-B	95.7	480×640	54.0	–	–
PDCNet[52]	ResNet-101	–	480×480	53.5	480×480	49.6
InvPT[55]	ViT-L	423[†]	480×640	53.6	–	–
T-Prompter[56]	T-Prompter-L	401[†]	480×640	55.3	–	–
SMMCL[9]	SegNeXt-B	–	480×640	55.8	–	–
CMNeXt[60]	MiT-B4	119.6	480×640	56.9	530×730	51.9
Sigma (Ours)	VMamba-T	48.3	480×640	53.9	480×640	50.0
Sigma (Ours)	VMamba-S	69.8	480×640	**57.0**	480×640	**52.4**

Table 2. Comparison of RGB-D semantic segmentation on NYU Depth V2 [37] and SUN RGB-D [39]. [†] indicates the parameters for multi-task learning reported from [56].

#	Encoder	CroMB	ConMB	Decoder	mIoU (∇)
1	VMamba-T	✓	✓	CMD	60.5 (0.0)
2	VMamba-T	✗	✓	CMD	59.9 (-0.6)
3	VMamba-T	✓	✗	CMD	59.7 (-0.8)
4	VMamba-T	✗	✗	CMD	58.4 (-2.1)
5	VMamba-T	✓	✓	MLP	59.4 (-1.1)
6	VMamba-T	✗	✗	MLP	57.5 (-3.0)
7	VMamba-T	✓	✓	Swin	59.6 (-0.9)
8	VMamba-T	✓	✓	UperNet	59.0 (-1.5)
9	Swin-T	✗	✗	Swin	56.3 (-4.2)

Table 3. Effectiveness of different components. We report the mIoU metric and the relative decrease (∇) on the MFNet [19] dataset. When both CroMB and ConMB show ✗ marks, we use a simple feature summation operation to obtain the fused feature.

decreases of 0.6% and 0.8% in performance, respectively. The removal of both blocks leads to a performance decline of 2.1%, highlighting the effectiveness of our proposed fusion module. Additionally, we evaluate our proposed decoder against a simple Multi-Layer Perceptron (MLP) decoder, a Swin-Transformer decoder [27], and an UperNet

decoder [47], discovering that our decoder surpasses them by 1.1%, 0.9%, and 1.5%, respectively. To further assess the effectiveness, we substitute our encoder with a pre-trained Swin Transformer [27] Tiny backbone and apply Swin Transformer blocks in the decoder. The results suggest that our design with SSM can be more effective than a simple integration of Transformers.

#	Encoder	Cross Attention Matrix	mIoU (∇)
1	VMamba-T	C	60.5 (0.0)
2	VMamba-T	B	59.3 (-1.2)
3	VMamba-T	D	59.6 (-0.9)
4	VMamba-T	B & C	59.7 (-0.8)
5	VMamba-T	C & D	59.8 (-0.7)

Table 4. Comparison of attention matrix used in CroMB.

Comparison of Attention Matrix in CroMB. In CroMB, we exchange matrix C of two modalities for cross-modal information interaction. For comparison, we conduct experiments to use matrix C, D, $B\&C$, and $C\&D$ in CroMB. As shown in Table 4, using the matrix C achieves the optimal performance. This supports our claim that matrix C is used to decode the information from the hidden state and can guide the reconstruction for complementary modality.

5. Conclusion

In this work, we propose Sigma, a novel Siamese Mamba Network for multi-modal semantic segmentation, which explores the application of state space models in multi-modal scene understanding for the first time. A Siamese backbone is first applied to extract robust global long-range dependencies with linear complexity. Then we introduce a fusion module that incorporates a cross-selective scan and a concat-selective scan. Finally, we design a channel-aware Mamba decoder to extract essential information from the fused features for predictions. Extensive experiments on RGB-T and RGB-D semantic segmentation benchmarks show the superiority of our method in both accuracy and efficiency.

Acknowledgement

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 1, 2, 7, 8

[2] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7088–7097, 2021. 8

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 5

[4] Leo Chen, Benjamin Boardley, Ping Hu, Yiru Wang, Yifan Pu, Xin Jin, Yongqiang Yao, Ruihao Gong, Bo Li, Gao Huang, et al. 2023 low-power computer vision challenge (lpcvc) summary. *arXiv preprint arXiv:2403.07153*, 2024. 1

[5] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Transactions on Image Processing*, 30:2313–2324, 2021. 8

[6] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 561–577. Springer, 2020. 7, 8

[7] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun 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)*, pages 4467–4473. IEEE, 2021. 6, 7

[8] Shaohua Dong, Wujie Zhou, Caie Xu, and Weiqing Yan. Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 6

[9] Xiaoyu Dong and Naoto Yokoya. Understanding dark scenes by contrasting multi-modal observations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 840–850, 2024. 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020. 1, 2

[12] Zhen Feng, Yanning Guo, and Yuxiang Sun. Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images. *IEEE Robotics and Automation Letters*, 8(4):2205–2212, 2023. 1, 6

[13] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 6, 8

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 3, 4

[15] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 3

[16] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 3

[17] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model, 2024. 2

[18] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. 1

[19] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multispectral scenes. pages 5108–5115, 2017. 1, 2, 5, 6, 7, 8

[20] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 6, 8

[21] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis, 2024. 2

[22] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 2023. 1, 2, 6, 7

[23] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1, 2, 5, 6, 7

[24] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual

adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 6

[25] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023. 2, 6

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 2, 4, 6

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021. 8

[28] Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu. Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3706–3714, 2023. 6

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[30] Ying Lv, Zhi Liu, and Gongyang Li. Context-aware interaction network for rgb-t semantic segmentation. *IEEE Transactions on Multimedia*, 2024. 1, 6

[31] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2

[32] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[34] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2022. 8

[35] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13525–13531. IEEE, 2021. 8

[36] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 2, 5, 6

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th Eu-*

ropean Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2, 5, 7, 8

[38] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 2, 3

[39] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2, 5, 8

[40] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. 4(3):2576–2583, 2019. 6, 7

[41] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Trans. on Automation Science and Engineering (TASE)*, 2020. 2, 6

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 1

[43] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022. 8

[44] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33:4835–4845, 2020. 8

[45] Ziyang Wang and Chao Ma. Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation, 2024. 2

[46] Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022. 6

[47] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 8

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[49] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024. 2

[50] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020. 2, 6

[51] Jiangtao Xu, Kaige Lu, and Han Wang. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146:179–184, 2021. 2, 6

[52] Jun Yang, Lizhi Bai, Yaoru Sun, Chunqi Tian, Maoyu Mao, and Guorun Wang. Pixel difference convolutional network

for rgb-d semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 8

[53] Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation, 2024. 2

[54] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding, 2023. 1

[55] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 8

[56] Hanrong Ye and Dan Xu. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023. 8

[57] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *The Twelfth International Conference on Learning Representations*. 2

[58] Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, and Yaqi Xie. Robust hierarchical scene graph generation. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023. 1

[59] Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, and Yaqi Xie. Hiker-sgg: Hierarchical knowledge enhanced robust scene graph generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[60] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 1, 2, 6, 7, 8

[61] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong 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*, pages 2633–2642, 2021. 6

[62] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021. 2

[63] Wujie Zhou, Shaohua Dong, Meixin Fang, and Lu Yu. Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing. *IEEE Transactions on Intelligent Vehicles*, 2023. 2, 6

[64] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 2, 6

[65] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. *IEEE Transactions on Multimedia*, 2022. 8

[66] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing*, 16(4):677–687, 2022. 8

[67] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2