This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

MMSFormer: Multimodal Transformer for Material and Semantic Segmentation

Md Kaykobad Reza[1], Ashley Prater-Bennette[2], and M. Salman Asif[1], Senior Member, IEEE

[1]University of California, Riverside, CA 92508 USA
[2]Air Force Research Laboratory, Rome, NY 13441 USA

Corresponding author: M. Salman Asif (email: sasif@ucr.edu).

This work is supported in part by AFOSR award FA9550-21-1-0330.

ABSTRACT Leveraging information across diverse modalities is known to enhance performance on multimodal segmentation tasks. However, effectively fusing information from different modalities remains challenging due to the unique characteristics of each modality. In this paper, we propose a novel fusion strategy that can effectively fuse information from different modality combinations. We also propose a new model named Multi-Modal Segmentation TransFormer (MMSFormer) that incorporates the proposed fusion strategy to perform multimodal material and semantic segmentation tasks. MMSFormer outperforms current state-of-the-art models on three different datasets. As we begin with only one input modality, performance improves progressively as additional modalities are incorporated, showcasing the effectiveness of the fusion block in combining useful information from diverse input modalities. Ablation studies show that different modules in the fusion block are crucial for overall model performance. Furthermore, our ablation studies also highlight the capacity of different input modalities to improve performance in the identification of different types of materials. The code and pretrained models will be made available at https://github.com/csiplab/MMSFormer.

INDEX TERMS multimodal image segmentation, material segmentation, semantic segmentation, multimodal fusion, transformer

I. Introduction

Image segmentation [1], [2] methods assign one class label to each pixel in an image. The segmentation map can be used for holistic understanding of objects or context of the scene. Image segmentation can be further divided into different types; examples include semantic segmentation [3], [4], instance segmentation [5], [6], panoptic segmentation [7], [8] and material segmentation [9], [10]. Each of these segmentation tasks are designed to address specific challenges and applications.

Multimodal image segmentation [11], [12] aims to enhance the accuracy and completeness of the task by leveraging diverse sources of information, and potentially leading to a more robust understanding of complex scenes. In contrast to single-modal segmentation [13], the multimodal approach [14] is more complex due to the necessity of effectively integrating heterogeneous data from different modalities. Key challenges arise from variations in data quality and attributes, distinct traits of each modality, and need to create models capable of accurately and coherently segmenting with the fused information.

Most of the existing multimodal segmentation methods are designed to work with specific modality pairs, such as RGB-Depth [15]–[17], RGB-Thermal [18]–[20], and RGB-Lidar [21]–[23]. As they are designed for specific modality combinations, most of them generally do not work well with modality combinations different from the ones used in the original design. Recently, CMX [24] introduced a technique to fuse information from RGB and one other supplementary modality, but it is incapable of fusing more than two modalities at the same time. Some recent models have proposed techniques to fuse more than two modalities

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

[9], [25], [26]. However, they either use very complex fusion strategies [24], [25] or require additional information like semantic labels [9] for performing underlying tasks.

In this paper, we propose a novel fusion block that can fuse information from diverse combination of modalities. We also propose a new model for multimodal material and semantic segmentation tasks that we call MMSFormer. Our model uses transformer based encoders [27] to capture hierarchical features from different modalities, fuses the extracted features with our novel fusion block and utilizes MLP decoder to perform multimodal material and semantic segmentation. In particular, our proposed fusion block uses parallel convolutions to capture multi-scale features, channel attention to re-calibrate features along the channel dimension and linear layer to combine information across multiple modalities. Such a design provides a simple and computationally efficient fusion block that can handle an arbitrary number of input modalities and combine information effectively from different modality combinations. We compare our fusion block with some of the existing fusion methods in terms of number of parameters and GFLOPs in Table 9.

To evaluate our proposed MMSFormer and fusion block, we focus on multimodal material segmentation on MCubeS [9] dataset and multimodal semantic segmentation on FMB [29] and PST900 [30] datasets. MCubeS dataset consists of four different modalities: RGB, angle of linear polarization (AoLP), degree of linear polarization (DoLP) and near-infrared (NIR). FMB dataset includes RGB and infrared modalities, while PST900 dataset comprises RGB and thermal modalities. We show the overall and per-class performance comparison in Table 1-5 for these datasets. A series of experiments highlight the ability of the proposed fusion block to effectively combine features from different modality combinations, resulting in superior performance compared to current state-of-the-art methods. Ablation studies show that different input modalities assist in identifying different types of material classes as shown in Table 8. Furthermore, as we add new input modalities, overall performance increases gradually highlighting the ability of the fusion block to incorporate useful information from new modalities. We summarize the results in Table 4 and 6 for FMB and MCubeS datasets respectively.

Main contributions of this paper can be summarized as follows.

- We propose a new multimodal segmentation model called MMSFormer. The model incorporates a novel fusion block that can fuse information from arbitrary (heterogeneous) combinations of modalities.
- Our model achieves new state-of-the-art performance on three different datasets. Furthermore, our method achieves better performance for all modality combinations compared to the current leading models.
- A series of ablation studies show that each module on the fusion block has an important contribution towards the overall model performance and each input modality assists in identifying specific material classes.

Rest of the paper is structured as follows. Section II presents a brief review of related work. We describe our model and fusion block in detail in Section III. Section IV presents experimental results and ablation studies on multimodal material and semantic segmentation tasks with qualitative and quantitative analysis.

II. Related Work

Image segmentation has witnessed significant evolution, spurred by advancements in machine learning and computational capabilities. A significant improvement in this evolution came with the inception of fully convolutional networks (FCNs) [31], [32], which enabled pixel-wise predictions through the utilization of hierarchical features within convolutional neural networks (CNNs). This led to the development of a variety of CNN based models for different images segmentation tasks. U-Net [33] is one such model that utilizes skip-connections between the lower-resolution and corresponding higher-resolution feature maps. DeepLabV3+ [34] introduced dilated convolutions (atrous convolutions) into the encoder allowing the expansion of the receptive field without increasing computational complexity significantly. PSPNet [35] introduced global context modules that enable the model to gather information from a wide range of spatial scales, essentially integrating both local and global context into the segmentation process.

Recently, Transformer based models have proven to be very effective in handling complex image segmentation tasks. Some of the notable transformer-based models are Pyramid Vision Transformer (PVT) [36], SegFormer [27], and Mask2Former [37]. PVT [36] utilizes transformer based design for various computer vision tasks. SegFormer [27] utilizes efficient self-attention and lightweight MLP decoder for simple and efficient semantic segmentation. Mask2Former [37] uses masked-attention along with pixel decoder and transformer decoder for any segmentation task. Their success demonstrates the capacity of these models to provide state-of-the-art solutions in various segmentation tasks.

In the context of multimodal image segmentation, fusion of data from diverse sources [12] has gained traction as a means to extract richer information and improve accuracy. A variety of models and fusion strategies have been proposed for RGB-Depth segmentation tasks. FuseNet [16] model integrates depth feature maps into RGB feature maps, while SA-Gate [15] employs Separation-and-Aggregation Gating to mutually filter and recalibrate RGB and depth modalities before fusion. Attention Complementary Module has been proposed by ACNet [17] that extracts weighted RGB and depth features for fusion. The domain of RGB-Thermal image segmentation has also gained prominence. Recent models include RTFNet [20] that achieves fusion through elementwise addition of thermal features with RGB, RSFNet [18] proposing Residual Spatial Fusion module to blend

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812.

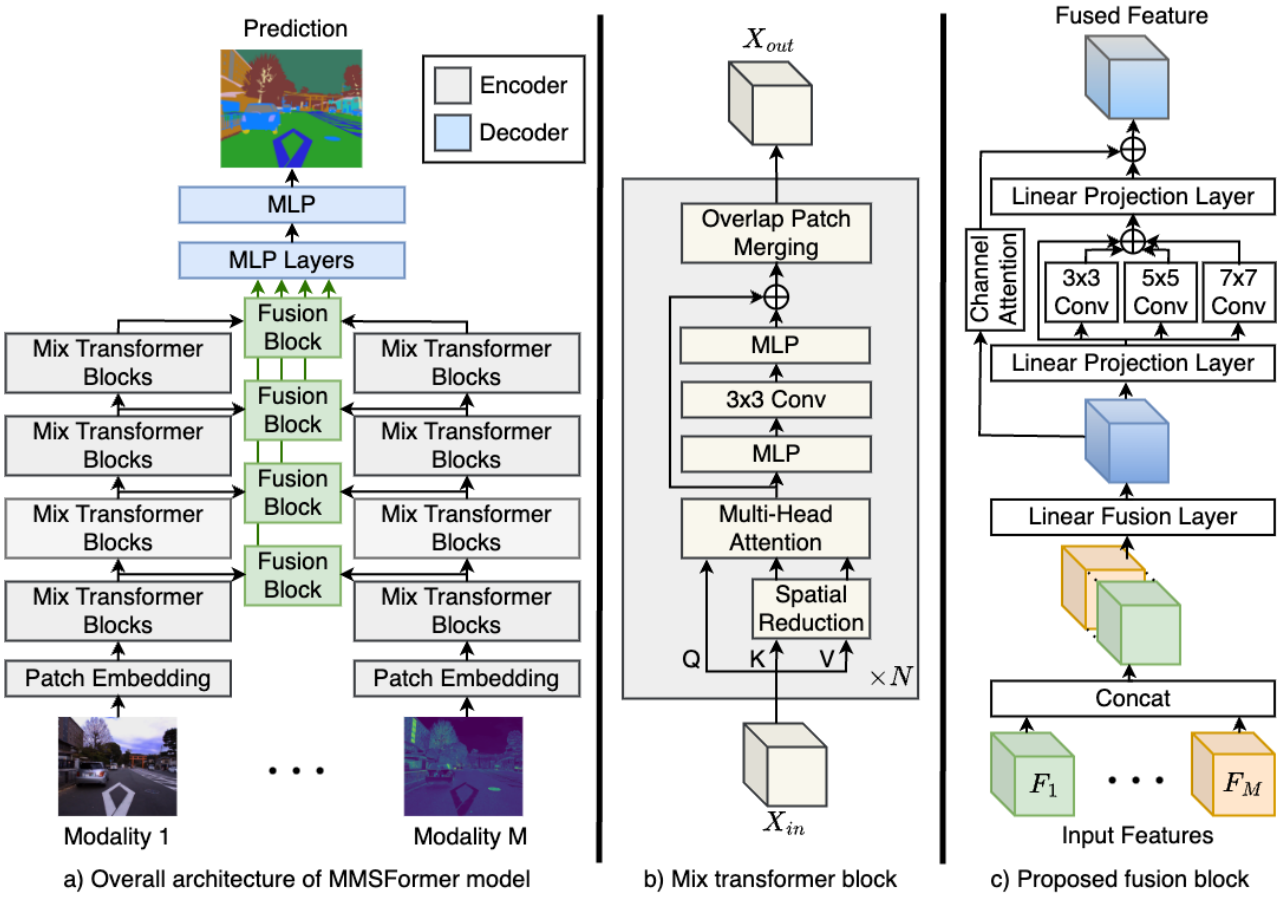


FIGURE 1: (a) Overall architecture of MMSFormer model. Each image passes through a modality-specific encoder where we extract hierarchical features. Then we fuse the extracted features using the proposed fusion block and pass the fused features to the decoder for predicting the segmentation map. (b) Illustration of the mix transformer [27] block. Each block applies a spatial reduction before applying multi-head attention to reduce computational cost. (c) Proposed multimodal fusion block. We first concatenate all the features along the channel dimension and pass it through linear fusion layer to fuse them. Then the feature tensor is fed to linear projection and parallel convolution layers to capture multi-scale features. We use Squeeze and Excitation block [28] as channel attention in the residual connection to dynamically re-calibrate the features along the channel dimension.

RGB and Thermal modalities, and EAEFNet [19] utilizing attention interaction and attention complement mechanisms to merge RGB and Thermal features. A number of methods also focus on fusing RGB-Lidar data that include TransFuser [22] that employs Transformer blocks, whereas LIF-Seg [23] relies on coarse feature extraction, offset learning, and refinement for effective fusion.

While the previously mentioned studies focus on specific pairs of modalities, some recent research has demonstrated promising results in the fusion of arbitrary modalities. CMX [24] introduces cross-modal feature rectification and fusion modules to merge RGB features with supplementary modalities. For multimodal material segmentation, MCubeSNet [9] model is proposed, which can seamlessly integrate four different modalities to enhance segmentation accuracy. In the context of arbitrary modal semantic segmentation, CMNeXt [26] introduces Self-Query Hub and Parallel Pooling Mixer modules, offering a versatile approach for fusing diverse modalities. Additionally, HRFuser [25] employs multi-window cross-attention to fuse different modalities at various resolutions, thereby enriching model performance.

Though some of these models can fuse different modalities, they either use very complex fusion strategies [24]–[26] or requires additional information [9] to perform underlying task. We aim to design a simple fusion module that can handle arbitrary number of input modalities and able to effectively fuse information from diverse modality combinations.

III. Proposed Model

The overall architecture of our proposed MMSFormer model and the fusion block is shown in Figure 1. The model has three modules: (1) Modality specific encoder; (2) Multimodal

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

fusion block; and (3) Shared MLP decoder. We use mix transformer [27] as the encoder of our model. We choose mix transformer for various reasons. First, it can provide hierarchical features without positional encoding. Second, it uses spatial reduction before attention that reduces the number of parameters significantly [27], [36]. Third, it also works well with simple and lightweight MLP decoder [27].

A. Overall model Architecture

Our overall model architecture is shown in Figure 1a. Assume we have M distinct modalities. Given a set of modalities as input, each modality-specific encoder captures distinctive features from each input modality by mapping the corresponding image into modality-specific hierarchical features as

$$F_m = \text{Encoder}_m(I_m), \tag{1}$$

where $I_m \in \mathbb{R}^{H\times W\times 3}$ represents input image for modality $m \in \{1, 2, \ldots, M\}$ and $\text{Encoder}_m(\cdot)$ denotes the encoder for that modality. The encoder generates four feature maps at $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of the input image resolution. We represent them as $F_m = \{F_m^1, F_m^2, F_m^3, F_m^4\}$. For simplicity we denote the shape of the feature map for the i^{th} encoder stage as $(H_i \times W_i \times C_i)$ where $i \in \{1, 2, 3, 4\}$.

We use four separate fusion blocks, one corresponding to each encoder stage, to fuse the features from each stage of the encoder. We pass the extracted features F_m^i for all modalities to the i^{th} fusion block as

$$F^i = \text{FusionBlock}^i(\{F_m^i\}_m). \tag{2}$$

Each fusion block fuses the features extracted from all the modalities to generate a combined feature representation $F = \{F^1, F^2, F^3, F^4\}$, where F^i denotes the fused feature at i^{th} stage. Finally, we pass the combined features F to the MLP decoder [27] to predict the segmentation labels.

B. Modality Specific Encoder

We use mix transformer encoder [27] to capture hierarchical features from the input modalities. Each input image I_m goes through patch embedding layer where it is divided into 4×4 patches following [27] and then fed to the mix transformer blocks. The design of mix transformer block is shown in Figure 1b. We denote the input to any mix transformer block as $X_{in} \in \mathbb{R}^{H_i\times W_i\times C_i}$ that is reshaped to $N_i \times C_i$ (with $N_i = H_i W_i$) and used as query Q, key K, and value V.

To reduce the computational overhead, spatial reduction is applied following [36] using a reduction ratio R. K and V are first reshaped into $\frac{N_i}{R} \times C_i R$ matrices and then mapped to $\frac{N_i}{R} \times C_i$ matrices via linear projection. A standard multi-head self-attention (MHSA) maps Q, K, V to intermediate features as

$$\begin{aligned} \text{MHSA}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \\ \text{head}_j &= \text{Attention}(QW_j^Q, KW_j^K, VW_j^V), \end{aligned} \tag{3}$$

where h represents the number of attention heads, $W_j^Q \in \mathbb{R}^{C_i\times d_K}$, $W_j^K \in \mathbb{R}^{C_i\times d_K}$, $W_j^V \in \mathbb{R}^{C_i\times d_V}$, and $W^O \in \mathbb{R}^{hd_V\times C_i}$ are the projection matrices, d_K, d_V represent dimensions of K, V, respectively. We can formulate the Attention function as

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_K}})V, \tag{4}$$

where Q, K, and V are the input query, key, and value matrices. MHSA is followed by a mixer layer (with two MLP and one 3×3 convolution layer). The convolution layer provides sufficient positional encoding into the transformer encoder for optimal segmentation performance [27]. This layer can be written as

$$\begin{aligned} \hat{X}_{in} &= \text{MHSA}(Q, K, V), \\ X_{out} &= \text{MLP}(\text{GELU}(\text{Conv}_{3\times 3}(\text{MLP}(\hat{X}_{in})))) + \hat{X}_{in}, \end{aligned} \tag{5}$$

Finally, overlap patch merging is applied to X_{out} following [27] to generate the final output.

C. Multimodal Fusion Block

After extracting hierarchical features, we fuse them using our proposed fusion block. The fusion block shown in Figure 1c is responsible for fusing the features extracted from the modality specific encoders. We have one fusion block for each of the four encoder stages. For the i^{th} fusion block, let us assume the input feature maps are given as $F_m^i \in \mathbb{R}^{H_i\times W_i\times C_i}\ \forall m \in \{1, 2, \ldots, M\}$. First, we concatenate the feature maps from M modalities along the channel dimension to get the combined feature map $F^i \in \mathbb{R}^{H_i\times W_i\times MC_i}$. Then we pass the features through a linear fusion layer that combines the features and reduces the channel dimension to C_i. Let us denote the resulting features as $\hat{F}^i \in \mathbb{R}^{H_i\times W_i\times C_i}$. We represent the operation as

$$\hat{F}^i = \text{Linear}(F_1^i || \cdots || F_M^i). \tag{6}$$

Here $||$ represents concatenation of features along the channel dimension and the linear layer takes an MC_i dimensional input and generates a C_i dimensional output.

After the linear fusion layer, we added a module for capturing and mixing multi-scale features. The module consists of two linear projection layers having parallel convolution layers in between them. First we apply a linear transformation on $\hat{F}^i$ along the channel dimension by passing it through the first linear projection layer. It refines and tunes the features from different channels. Then we apply 3×3, 5×5, and 7×7 convolutions to effectively capture diverse features across different spatial contexts. By employing convolutions with different sizes, the fusion block can attend to local patterns as well as capture larger spatial structures, thereby enhancing its ability to extract meaningful features from the input data. Finally we apply another linear transformation along the channel dimension using the second linear project layer to consolidate the information captured by the parallel convolutions, promoting feature consistency and enhancing the discriminative power of the fused features. These steps can be performed as

$$\tilde{F}^i = \text{Linear}(\hat{F}^i), \tag{7}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812.

TABLE 1: Performance comparison on Multimodal Material Segmentation (MCubeS) dataset [9]. Here A, D, and N represent angle of linear polarization (AoLP), degree of linear polarization (DoLP), and near-infrared (NIR) respectively.

Method	Modalities	% mIoU
DRConv [38]	RGB-A-D-N	34.63
DDF [39]	RGB-A-D-N	36.16
TransFuser [22]	RGB-A-D-N	37.66
DeepLabv3+ [34]	RGB-A-D-N	38.13
MMTM [40]	RGB-A-D-N	39.71
FuseNet [16]	RGB-A-D-N	40.58
MCubeSNet [9]	RGB-A-D-N	42.46
CBAM [41]	RGB-A-D-N	51.32
CMNeXt [26]	RGB-A-D-N	51.54
MMSFormer (Ours)	RGB-A-D-N	**53.11**

TABLE 2: Performance comparison on FBM [29] dataset. We show performance for different methods from already published works.

Methods	Modalities	% mIoU
CBAM [41]	RGB-Infrared	50.1
GMNet [42]	RGB-Infrared	49.2
LASNet [43]	RGB-Infrared	42.5
EGFNet [44]	RGB-Infrared	47.3
FEANet [45]	RGB-Infrared	46.8
DIDFuse [46]	RGB-Infrared	50.6
ReCoNet [47]	RGB-Infrared	50.9
U2Fusion [48]	RGB-Infrared	47.9
TarDAL [49]	RGB-Infrared	48.1
SegMiF [29]	RGB-Infrared	54.8
MMSFormer (Ours)	RGB-Infrared	**61.7**

$$F^i = \text{Linear}(\tilde{F}^i + \sum_{k\in\{3,5,7\}} \text{Conv}_{k\times k}(\tilde{F}^i)). \quad (8)$$

We found that using 3 parallel convolution layers with sizes 3×3, 5×5, and 7×7 provide optimal performance. Increasing the convolution kernel size reduces performance which we show in Table 7. As larger kernels reduce performance, we did not add more than 3 parallel convolutions in our model.

We apply Squeeze-and-Excitation block [28] as channel attention in the residual connection. The final fused feature can be represented as

$$F^i = \text{ChannelAttention}(\hat{F}^i) + F^i. \quad (9)$$

Channel attention re-calibrates interdependence between channels and allows the model to select the most relevant features or channels while suppressing less important ones [28]. This leads to more effective feature representations and thus better performance on the underlying task.

D. Shared MLP Decoder

The fused features generated from all the 4 fusion blocks are sent to the shared MLP decoder. We use the deocder design proposed in [27]. The decoder shown in Figure 1a can be represented as the following equations:

$$\begin{aligned} F^i &= \text{Linear}(F^i), \quad \forall i \in \{1,2,3,4\} \\ F^i &= \text{Upsample}(F^i), \quad \forall i \in \{1,2,3,4\} \\ F &= \text{Linear}(F^1||\cdots||F^4), \\ P &= \text{Linear}(F). \end{aligned} \quad (10)$$

The first linear layers take the fused features of different shapes and generate features having the same channel dimension. Then the features are up-sampled to $\frac{1}{4}^{th}$ of the original input shape, concatenated along the channel dimension and passed through another linear layer to generate the final fused feature F. Finally F is passed through the last linear layer to generate the predicted segmentation map P.

IV. Experiments and Results

We evaluated our model and proposed fusion block on multiple datasets and with different modality combinations for multimodal semantic and material segmentation tasks. We also compared our methods with existing baseline methods both qualitatively and quantitatively. We report results from already published works whenever possible. $*$ indicates that we have used the code and pretrained models from the papers to generate the results.

A. Datasets

Multimodal material segmentation (MCubeS) dataset [9] contains 500 sets of images from 42 street scenes having four modalities: RGB, angle of linear polarization (AoLP), degree of linear polarization (DoLP), and near-infrared (NIR). It provides annotated ground truth labels for both material and semantic segmentation and divided into training set with 302 image sets, validation set with 96 image sets, and test set with 102 image sets. This dataset has 20 class labels corresponding to different materials.

FMB dataset [29] is a new and challenging dataset with 1500 pairs of calibrated RGB-Infrared image pairs. The training and test set contains 1220 and 280 image pairs respectively. The dataset covers a wide range of scenes under different lighting and weather conditions (Tyndall effect, rain, fog, and strong light). It also provides per pixel ground truth annotation for 14 different classes.

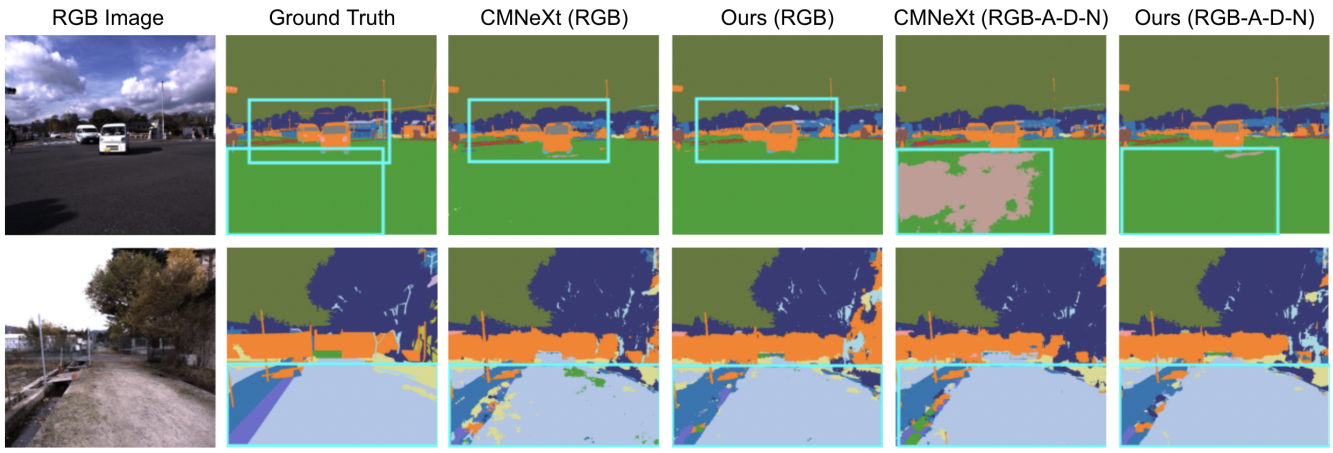
PST900 dataset [30] contains 894 pairs of synchronized RGB-Thermal image pairs. The dataset is divided into training and test sets with per pixel ground truth annotation for five different classes.

B. Implementation Details

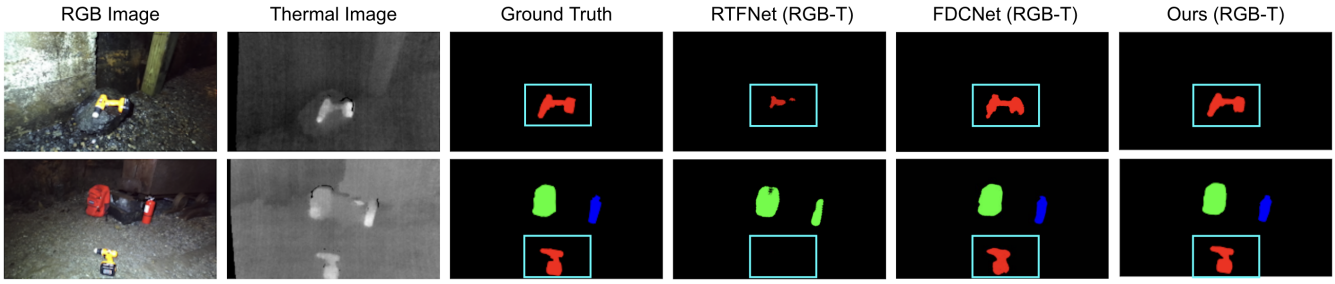
To ensure a fair comparison with prior models, we followed the same data preprocessing and augmentation strategies employed in previous studies [9], [26], [29]. We used

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812



(a) Visualization of predictions on MCubeS dataset



(b) Visualization of predictions on PST900 dataset

FIGURE 2: Visualization of predictions on MCubeS and PST900 datasets. Figure 2a shows RGB and all modalities (RGB-A-D-N) prediction from CMNeXt [26] and our model on MCubeS dataset. For brevity, we only show the RGB image and ground truth material segmentation maps along with the predictions. Figure 2b shows predictions from RTFNet [20], FDCNet [50] and our model for RGB-thermal input modalities on PST900 dataset. Our model shows better predictions on both of the datasets.

the Mix-Transformer (MiT) [27] encoder pretrained on the ImageNet [63] dataset as the backbone for our model to extract features from different modalities. Each modality has a separate encoder. We used a shared MLP decoder introduced in SegFormer [27] and used random initialization for it. We trained and evaluated all our models using two NVIDIA RTX 2080Ti GPUs and used PyTorch for model development.

We utilized a polynomial learning rate scheduler with a power of 0.9 to dynamically adjust the learning rate during training. The first 10 epochs were designated as warm-up epochs with a learning rate of 0.1 times the original rate. For loss computation, we used the cross-entropy loss function. Optimization was performed using the AdamW [64] optimizer with an epsilon value of 10^{-8} and weight decay set to 0.01. For CBAM [41], we use the same encoder, decoder and hyperparameters used in our experiments and replace our fusion block with CBAM[1] module. To be specific, after extracting the feature maps from each input modality using the modality specific encoders, we add them (sum them up) and pass the combined feature map to the CBAM module.

[1]https://github.com/luuuyi/CBAM.PyTorch

C. Performance Comparison with Existing Methods

We conducted a rigorous performance evaluation of our model compared to established baseline models for three datasets. The comprehensive results are summarized in Tables 1–6. We report results for CBAM from our experiments. Other results are taken from published literature.

Results on MCubeS Dataset. Table 1 shows the overall performance comparison between our model and existing baseline models for MCubeS dataset. Our model achieves a mean intersection-over-union (mIoU) of 53.11%, surpassing the current state-of-the-art model. It shows 1.57% improvement over CMNeXt [26], 1.79% improvement over CBAM [41] and 10.65% improvement over MCubeSNet [9] models. To further analyze the performance of our model, we conducted a per-class IoU analysis and presented in Table 3. Our model performs better in detecting most of the material classes compared to the current state-of-the-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812.

TABLE 3: Per-class % IoU comparison on Multimodal Material Segmentation (MCubeS) [9] dataset. Our proposed MMSFormer model shows better performance in detecting most of the classes compared to the current state-of-the-art models. $*$ indicates that the code and pretrained model from the authors were used to generate the results.

Methods	Asphalt	Concrete	Metal	Road marking	Fabric	Glass	Plaster	Plastic	Rubber	Sand	Gravel	Ceramic	Cobblestone	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
MCubeSNet [9]	85.7	42.6	47.0	59.2	12.5	44.3	3.0	10.6	12.7	66.8	67.1	27.8	65.8	36.8	54.8	39.4	73.0	13.3	0.0	94.8	42.9
CBAM [41]	85.7	47.7	55.4	70.4	27.6	54.7	**0.9**	30.9	26.5	61.6	63.0	28.0	71.1	41.8	58.6	47.4	76.7	**56.3**	25.9	96.5	51.3
CMNeXt [26] $*$	84.3	44.9	53.9	**74.5**	32.3	54.0	0.8	28.3	**29.7**	**67.7**	66.5	27.7	68.5	42.9	58.7	**49.7**	75.4	55.7	18.9	96.5	51.5
MMSFormer (Ours)	**88.0**	**48.3**	**56.2**	72.2	**35.4**	**54.9**	0.5	**34.6**	29.4	67.2	**69.0**	**29.9**	**73.4**	**44.7**	**59.5**	47.8	**77.1**	50.5	**26.9**	**96.6**	**53.1**

TABLE 4: Per-class % IoU comparison on FMB [29] dataset for both RGB only and RGB-infrared modalities. We show the comparison for 8 classes (out of 14) that are published. T-Lamp and T-Sign stand for Traffic Lamp and Traffic Sign respectively. Our model outperforms all the methods for all the classes except for the truck class.

Methods	Modalities	Car	Person	Truck	T-Lamp	T-Sign	Building	Vegetation	Pole	% mIoU
SegMiF [29]	RGB	78.3	46.6	**43.4**	23.7	64.0	77.8	82.1	41.8	50.5
MMSFormer (Ours)	RGB	**80.3**	**56.7**	42.1	**31.6**	**77.8**	**77.9**	**85.4**	**48.1**	**57.2**
CBAM [41]	RGB-Infrared	71.9	49.3	20.9	25.8	67.1	75.8	80.9	19.7	50.1
GMNet [42]	RGB-Infrared	79.3	60.1	22.2	21.6	69.0	79.1	83.8	39.8	49.2
LASNet [43]	RGB-Infrared	72.6	48.6	14.8	2.9	59.0	75.4	81.6	36.7	42.5
EGFNet [44]	RGB-Infrared	77.4	63.0	17.1	25.2	66.6	77.2	83.5	41.5	47.3
FEANet [45]	RGB-Infrared	73.9	60.7	32.3	13.5	55.6	79.4	81.2	36.8	46.8
DIDFuse [46]	RGB-Infrared	77.7	64.4	28.8	29.2	64.4	78.4	82.4	41.8	50.6
ReCoNet [47]	RGB-Infrared	75.9	65.8	14.9	34.7	66.6	79.2	81.3	44.9	50.9
U2Fusion [48]	RGB-Infrared	76.6	61.9	14.4	28.3	68.9	78.8	82.2	42.2	47.9
TarDAL [49]	RGB-Infrared	74.2	56.0	18.8	29.6	66.5	79.1	81.7	41.9	48.1
SegMiF [29]	RGB-Infrared	78.3	65.4	**47.3**	43.1	74.8	82.0	85.0	49.8	54.8
MMSFormer (Ours)	RGB-Infrared	**82.6**	**69.8**	44.6	**45.2**	**79.7**	**83.0**	**87.3**	**51.4**	**61.7**

art models. Notably, our model demonstrates a substantial improvement in the detection of plastic (+3.7%), fabric (+3.1%), asphalt (+2.3%), and cobblestone (2.3%) classes while maintaining competitive or better performance in other classes. This led to the overall better performance and sets new state-of-the-art for this dataset.

Results on FMB Dataset. Performance Comparison for FMB dataset is shown on Table 2. Our model shows a significant improvement of 6.9% mIoU compared to the current state-of-the-art model. Per-class IoU analysis for this dataset is shown on Table 4. For a fair comparison, we only compare the performance on 8 classes (out of 14) that are published in literature. T-Lamp and T-Sign represent Traffic Lamp and Traffic Sign, respectively. Our model shows an overall performance improvement of 6.7% mIoU for RGB only predictions compared to the most recent SegMiF [29] model. Alongside this, our model also shows superior performance in detecting all of the classes except for the truck class for both RGB only and RGB-Infrared semantic segmentation tasks. Performance on RGB-Infrared input modalities is much better than RGB only performance for all the classes, which demonstrates the ability of the fusion block to effectively fuse information from the input modalities.

Results on PST900 Dataset. We also tested our model on PST900 [30] dataset and summarized the result in Table 5. Experiments show that our model outperforms existing baseline models for RGB-Thermal semantic segmentation on this dataset. It outperforms the most recent CACFNet [62] model by 0.89% mIoU. Our model also shows better performance in detecting 3 out of the 5 classes available in the dataset and competitive performance in other two classes.

D. Performance Comparison for Incremental Modality Integration

A critical aspect of this work involves evaluating the effectiveness of our proposed fusion block in combining valuable information from diverse modalities. To analyze this effect, we trained our model with various combinations of modalities on the MCubeS dataset. The results are presented in

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

TABLE 5: Performance comparison on PST900 [30] dataset. We show per-class % IoU as well as % mIoU for all the classes.

Methods	Modalities	Background	Fire-Extinguisher	Backpack	Hand-Drill	Survivor	% mIoU
ACNet [17]	RGB-Thermal	99.25	59.95	83.19	51.46	65.19	71.81
CCNet [51]	RGB-Thermal	99.05	51.84	66.42	32.27	57.50	61.42
Efficient FCN [52]	RGB-Thermal	98.63	39.96	58.15	30.12	28.00	50.98
RTFNet [20]	RGB-Thermal	99.02	51.93	74.17	7.07	70.11	60.46
PSTNet [30]	RGB-Thermal	98.85	70.12	69.20	53.60	50.03	68.36
EGFNet [53]	RGB-Thermal	99.26	71.29	83.05	64.67	74.30	78.51
MTANet [54]	RGB-Thermal	99.33	64.95	87.50	62.05	79.14	78.60
MFFENet [55]	RGB-Thermal	99.40	66.38	81.02	72.50	75.60	78.98
GMNet [42]	RGB-Thermal	99.44	73.79	83.82	85.17	78.36	84.12
CGFNet [56]	RGB-Thermal	99.30	71.71	82.00	59.72	77.42	78.03
GCNet [57]	RGB-Thermal	99.35	77.68	79.37	82.92	73.58	82.58
GEBNet [58]	RGB-Thermal	99.39	73.07	85.93	67.14	80.21	81.15
GCGLNet [59]	RGB-Thermal	99.39	77.57	81.01	81.90	76.31	83.24
DHFNet [60]	RGB-Thermal	99.44	78.15	87.38	71.18	74.81	82.19
MDRNet+ [61]	RGB-Thermal	99.07	63.04	76.27	63.47	71.26	74.62
FDCNet [50]	RGB-Thermal	99.15	71.52	72.17	70.36	72.36	77.11
CBAM [41]	RGB-Thermal	99.43	73.81	82.75	80.00	69.60	81.12
EGFNet [44]	RGB-Thermal	99.55	79.97	90.62	76.08	**80.88**	85.42
CACFNet [62]	RGB-Thermal	99.57	**82.08**	89.49	80.90	80.76	86.56
MMSFormer (Ours)	RGB-Thermal	**99.60**	81.45	**89.86**	**89.65**	76.68	**87.45**

TABLE 6: Performance comparison (% mIoU) on Multimodal Material Segmentation (MCubeS) [9] dataset for different modality combinations. Here A, D, and N represent angle of linear polarization (AoLP), degree of linear polarization (DoLP), and near-infrared (NIR) respectively.

Modalities	MCubeSNet [9]	CMNeXt [26]	MMSFormer (Ours)
RGB	33.70	48.16	**50.44**
RGB-A	39.10	48.42	**51.30**
RGB-A-D	42.00	49.48	**52.03**
RGB-A-D-N	42.86	51.54	**53.11**

Table 6. Our model exclusively trained on RGB data provided an mIoU score of 50.44%, which is 2.28% grater than the current state-of-the-art model. We observe progressive improvement in performance as we incorporated additional modalities: AoLP, DoLP, and NIR. The integration led to incremental performance gains, with the mIoU increasing from 50.44% to 51.30%, then to 52.03%, and ultimately reaching to 53.11%. These findings serve as a compelling evidence that our fusion approach effectively leverages and fuses valuable information from different combination of modalities, resulting in a notable enhancement in segmentation performance.

Furthermore, our model consistently outperforms the current state-of-the-art benchmark across all modality combinations. This consistent superiority underscores the robustness and versatility of our fusion block, demonstrating its ability to adapt and excel regardless of the specific modality combination provided.

E. Qualitative analysis of the Predictions

Apart from quantitative analysis, we also perform qualitative analysis of the predicted segmentation maps. We show material segmentation results predicted by CMNeXt [26] model and our proposed MMSFormer model in Figure 2a. For brevity, we only show RGB images and ground truth material segmentation maps in the illustrations. We show RGB only predictions and all modalities (RGB-A-D-N) predictions for both of the models. As highlighted in the rectangular bounding boxes, our proposed MMSFormer model identifies asphalt, sand and water with greater accuracy than CMNeXt [26] model for both RGB only and all modalities (RGB-A-D-N) predictions.

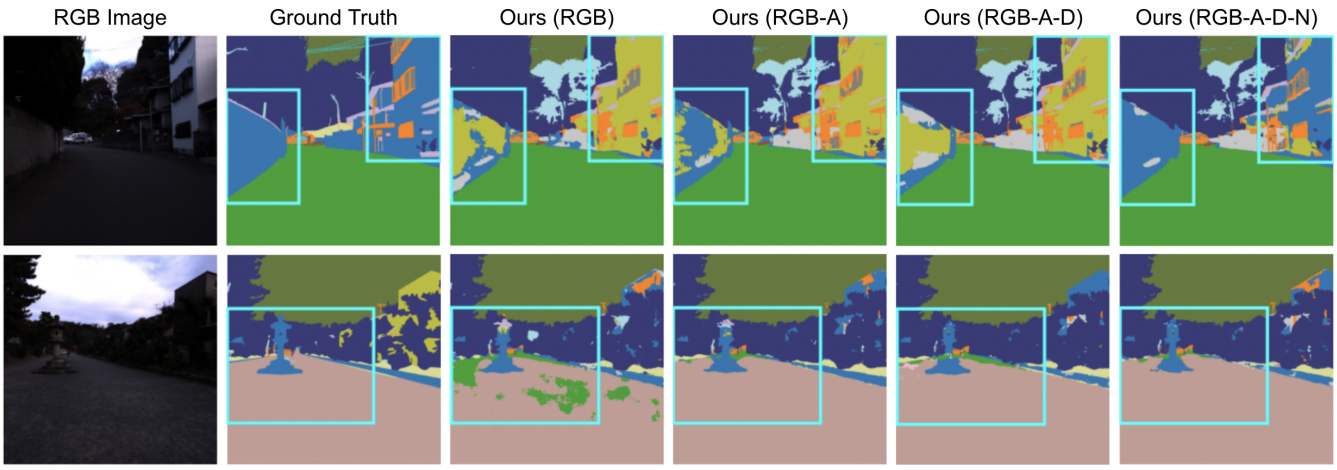
We also compare our prediction on PST900 [30] dataset with RTFNet [20] and FDCNet [50] on Figure 2b. We show the input RGB image, thermal images, ground truth segmentation maps and prediction form the models. As highlighted by the rectangular bounding boxes, our model shows better accuracy in detecting objects with more precise contours compared to the other two methods.

F. Ablation Study on the Fusion Block

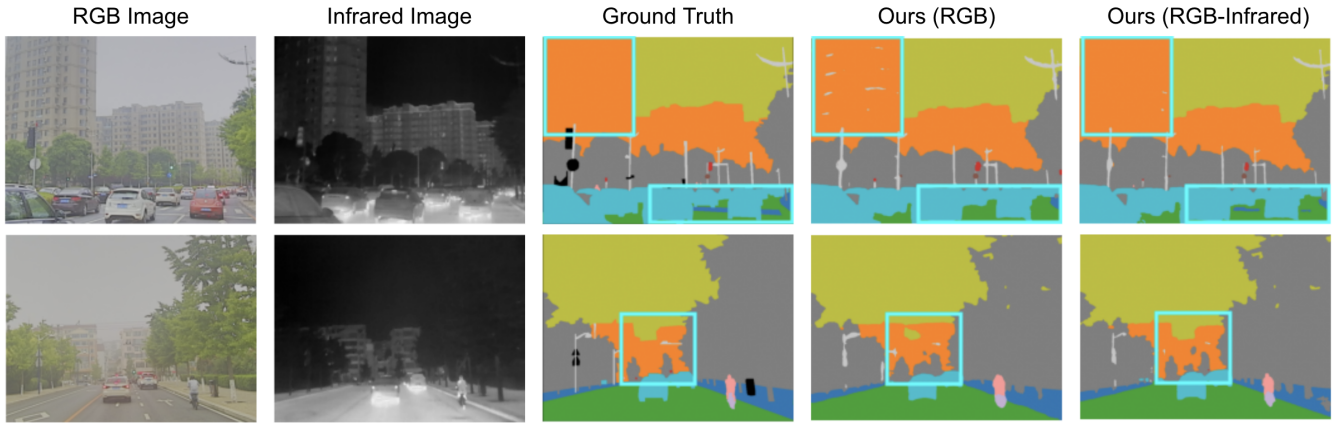
We conducted a number of ablation studies aimed at investigating the contributions of individual components within the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812.



(a) Visualization of predictions on MCubeS dataset for different modality combinations



(b) Visualization of predictions on FMB dataset for different modality combinations

FIGURE 3: Visualization of predicted segmentation maps for different modality combinations on MCubeS [9] and FMB [29] datasets. Both figures show that prediction accuracy increases as we incrementally add new modalities. They also illustrate the fusion block's ability to effectively combine information from different modality combinations.

fusion block to the overall model performance. The findings, as detailed in Table 7, shed light on the critical importance of these components. We used both RGB and infrared modalities of the FMB dataset during training and testing in these experiments. First, we observed that the absence of channel attention in the residual connection had a negative impact, resulting in a reduction in performance by 3.36%. This indicates that feature calibration along channel dimension plays an important role in capturing and leveraging crucial information effectively. Additionally, while comparing larger convolution kernel sizes (3×3, 7×7, and 11×11) to the originally employed (3×3, 5×5, and 7×7), we noted a decrease in performance by 5.36%. This result underscores the significance of the carefully chosen convolution kernel sizes within the fusion block.

Furthermore, completely removing the parallel convolutions from the block led to a performance decline of 4.51%, emphasizing their substantial contribution in capturing multi-scale features and overall model performance. Finally, if we only use the linear fusion layer to fuse the features and remove the parallel convolutions and channel attention from the fusion block, performance drop significantly by 9.25%. These studies demonstrate that multi-scale feature capturing via parallel convolutions and channel-wise feature calibration using channel attention is extremely important in learning better feature representation and thus crucial to overall model performance. These comprehensive ablation studies collectively underscore the significance of every component within the fusion block, revealing that each module plays a distinct and vital role in achieving the overall performance of our model.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

TABLE 7: Ablation study of the Fusion Block on FMB [29] dataset. Both RGB and infrared input modalities were used during training and testing. The table shows the contribution of different modules in fusion block in overall model performance.

Structure	Parameter Count (M)	% mIoU (Change)
MMSFormer	61.26	61.68
- without channel attention	61.21	58.32 (-3.36)
- without parallel convolutions	61.17	57.17 (-4.51)
- with 3x3, 7x7 and 11x11 convolutions	61.36	56.32 (-5.36)
- only linear fusion	59.57	52.43 (-9.25)

TABLE 8: Per class % IoU comparison on Multimodal Material Segmentation (MCubeS) [9] dataset for different modality combinations. As we add modalities incrementally, overall performance increases gradually. This table also shows that specific modality combinations assist in identifying specific types of materials better.

Modalities	Asphalt	Concrete	Metal	Road marking	Fabric	Glass	Plaster	Plastic	Rubber	Sand	Gravel	Ceramic	Cobblestone	Brick	Grass	Wood	Leaf	Water	Human	Sky	Mean
RGB	83.2	44.2	52.1	70.4	31.0	51.6	1.3	26.2	21.8	65.0	61.8	31.3	72.5	**45.0**	55.4	46.0	74.7	**56.0**	22.7	96.4	50.4
RGB-A	86.5	46.5	55.9	**73.0**	35.3	**56.0**	0.8	27.3	27.8	66.2	67.0	28.6	69.6	43.0	57.6	**49.6**	76.4	53.8	8.4	96.6	51.3
RGB-A-D	86.0	44.0	55.5	68.1	31.9	54.8	**2.3**	30.0	**29.7**	**69.4**	**73.7**	**32.2**	69.4	41.4	59.2	48.3	76.6	50.6	20.9	**96.7**	52.0
RGB-A-D-N	**88.0**	**48.3**	**56.2**	72.2	**35.4**	54.9	0.5	**34.6**	29.4	67.2	69.0	29.9	**73.4**	44.7	**59.5**	47.8	**77.1**	50.5	**26.9**	96.6	**53.1**

TABLE 9: Comparison of number of parameters in the fusion block and GFLOPs for MCubeS dataset having 4 input modalities with an input shape of $(3 \times 512 \times 512)$ for each modality. Our fusion block shows significantly lower complexity compared to existing methods.

Methods	**Fusion Block Parameters (M)**	**GFLOPs**
CMNeXt [26]	16.63	6.47
MCubeSNet [9]	7.41	12.10
HRFuser [25]	**1.72**	17.50
CMX [24]	16.59	6.41
DDF (Resnet-101) [39]	28.10	4.10
MMSFormer (Ours)	3.23	**2.47**

G. Ablation Study on Different Modality Combinations

To analyze the contributions of different modalities in the identification of distinct materials, we conducted a series of ablation studies, focusing on per-class IoU for different modality combinations. The insights are summarized in Table 8. As we progressively integrate new modalities, performance gradually increases for specific classes, which include grass, leaf, asphalt, cobblestone and plastic classes. Particularly noteworthy is the assistance provided by NIR data in classifying asphalt, concrete, plastic, cobblestone, and human categories, leading to significant performance gains in these classes as NIR was added as an additional modality.

Conversely, certain classes such as water and brick exhibited a gradual performance decline as we introduced additional modalities. This suggests that RGB data alone suffices for accurately identifying these classes, and the inclusion of more modalities potentially introduces noise or redundancy that negatively impacts performance. Moreover, AoLP appears to be helpful in enhancing the recognition of materials like road markings, glass and wood. Similarly, DoLP improved performance for classes like plaster, rubber, sand, gravel and ceramic. These findings underscore the relationship between different imaging modalities and the unique characteristics of different types of materials, demonstrating that specific modalities excel in detecting particular classes based on their distinctive traits.

In Figure 3a, we presents some examples to show how adding different modalities help improve performance of segmentation. We show predictions for RGB, RGB-A, RGB-A-D and RGB-A-D-N inputs from our proposed MMSFormer model. As we add new modalities, the predictions become more accurate as shown in the bounding boxes. The illustrations show that the identification of concrete and gravel becomes more accurate with additional modalities. Figure 3b shows predictions for RGB and RGB-Infrared from FMB dataset. As highlighted by the bounding boxes, adding new modality helps improve performance in detecting building, road and sidewalks. This also illustrates the capability of the fusion block to effectively fuse information from different modality combinations.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812.

H. Computational Complexity of the Fusion Block

In addition to better performance, our fusion block is also computationally efficient compared to most of the fusion blocks proposed for these datasets. We show a comparison in terms of the number of parameters in the fusion block and GFLOPs for some recent models on Table 9 for MCubeS dataset having 4 input modalities with an input shape of $(3 \times 512 \times 512)$ for each modality. As observed from the table, our proposed fusion strategy demonstrates significantly lower complexity in terms of both the number of parameters and GFLOPs compared to existing methods. HRFuser [25] has a lower parameter count than ours but it requires more than $7\times$ GFLOPs. Other methods require significantly more parameters $(2.3\times - 8.7\times)$ and GFLOPs $(1.6\times - 7\times)$ compared to our fusion strategy. Our comparison only includes models for which these results are available in the published literature.

V. CONCLUSION

In this paper, we introduce a novel fusion module designed to combine useful information from various modality combinations. We also propose a new model called MMSFormer that integrates the proposed fusion block to accomplish multimodal material and semantic segmentation tasks. Experimental results illustrate the model's capability to efficiently fuse information from different combination of modalities, leading to new state-of-the-art performance on three different datasets. Experiments also show that the fusion block can extract useful information from different modality combinations that helps the model to consistently outperform current state-of-the-art models. Starting from only one input modality, performance increases gradually as we add new modalities. Several ablation studies further highlight how different components of the fusion block contribute to the overall model performance. Ablation studies also reveal that different modalities assist in identifying different types of material classes. However, one limitation of the proposed model is the use of modality specific encoders and the number of encoders grows with number of modalities. Future work will include exploring the possibility and effectiveness of using a shared encoder for all the modalities, investigating and extending the model's capability with other modalities and multimodal tasks.

REFERENCES

[1] H.-D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[2] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[3] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[4] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*. Ieee, 2018, pp. 1451–1460.

[5] W. Gu, S. Bai, and L. Kong, "A review on 2d instance segmentation based on deep neural networks," *Image and Vision Computing*, vol. 120, p. 104401, 2022.

[6] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International journal of multimedia information retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[7] O. Elharrouss, S. A. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, and Y. Himeur, "Panoptic segmentation: A review," *ArXiv*, vol. abs/2111.10250, 2021.

[8] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[9] Y. Liang, R. Wakaki, S. Nobuhara, and K. Nishino, "Multimodal material segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 800–19 808.

[10] P. Upchurch and R. Niu, "A dense material segmentation dataset for indoor and outdoor scene parsing," in *European Conference on Computer Vision*. Springer, 2022, pp. 450–466.

[11] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[12] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[13] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[14] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[15] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 561–577.

[16] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, November 2016.

[17] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1440–1444.

[18] P. Li, J. Chen, B. Lin, and X. Xu, "Residual spatial fusion network for rgb-thermal semantic segmentation," *arXiv preprint arXiv:2306.10364*, 2023.

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

[20] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, July 2019.

[21] J. Li, H. Dai, H. Han, and Y. Ding, "Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2023, pp. 21 694–21 704.

[22] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[23] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation," *IEEE Transactions on Multimedia*, pp. 1–11, 2023.

[24] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[25] T. Broedermann, C. Sakaridis, D. Dai, and L. Van Gool, "Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[26] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmenta-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Open Journal of Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJSP.2024.3389812

tion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1136–1147.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

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

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

[32] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv preprint arXiv:1903.11816*, 2019.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

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

[36] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[37] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[38] J. Chen, X. Wang, Z. Guo, X. Zhang, and J. Sun, "Dynamic region-aware convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8064–8073.

[39] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6647–6656.

[40] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 289–13 299.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

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

[43] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1223–1235, 2023.

[44] S. Dong, W. Zhou, C. Xu, and W. Yan, "Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2023.

[45] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 2021, p. 4467–4473.

[46] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, and P. Li, "Didfuse: Deep image decomposition for infrared and visible image fusion," in *IJCAI*. ijcai.org, 2020, pp. 970–976.

[47] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "Reconet: Recurrent correction network for fast and efficient multi-modality image fusion," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 539–555.

[48] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[49] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.

[50] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for rgb-t semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2892–2905, 2023.

[51] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.

[52] J. Liu, J. He, J. Zhang, J. S. Ren, and H. Li, "Efficientfcn: Holistically-guided decoding for semantic segmentation," in *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 1–17.

[53] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 3571–3579.

[54] W. Zhou, S. Dong, J. Lei, and L. Yu, "Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 48–58, 2023.

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

[56] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "Cgfnet: Cross-guided fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2949–2961, 2022.

[57] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "Gcnet: Grid-like context-aware network for rgb-thermal semantic segmentation," *Neurocomput.*, vol. 506, no. C, p. 60–67, sep 2022.

[58] S. Dong, W. Zhou, X. Qian, and L. Yu, "Gebnet: Graph-enhancement branch network for rgb-t scene parsing," *IEEE Signal Processing Letters*, vol. 29, pp. 2273–2277, 2022.

[59] T. Gong, W. Zhou, X. Qian, J. Lei, and L. Yu, "Global contextually guided lightweight network for rgb-thermal urban scene understanding," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105510, 2023.

[60] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, "Dhfnet: dual-decoding hierarchical fusion network for rgb-thermal semantic segmentation," *The Visual Computer*, pp. 1–11, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256320037

[61] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for rgb-t semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[62] W. Zhou, S. Dong, M. Fang, and L. Yu, "Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing," *IEEE Transactions on Intelligent Vehicles*, pp. 1–10, 2023.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/