FuseForm: Multimodal Transformer for Semantic Segmentation

Justin McMillen
University of South Florida
Tampa, Florida, USA
jmcmillen@usf.edu

Yasin Yilmaz
University of South Florida
Tampa, Florida, USA
yasiny@usf.edu

Abstract

For semantic segmentation, integrating multimodal data can vastly improve segmentation performance at the cost of increased model complexity. We introduce FuseForm, a multimodal transformer for semantic segmentation, which can effectively and efficiently fuse a large number of homogeneous modalities. We demonstrate its superior performance on 5 different multimodal datasets ranging from 2 to 12 modalities and comprehensively analyze its components. FuseForm outperforms existing methods through two novel features, a hybrid multimodal fusion block and a transformer-based decoder. It leverages a multimodal cross-attention module for global token fusion, alongside convolutional filters' ability to fuse local features. Global and local fusion modules together enable enhanced multimodal semantic segmentation. We also introduce a decoder based on a mirrored version of the encoder transformer, which outperforms a popular decoder when tuned sufficiently on the dataset.

1. Introduction

Semantic segmentation, a popular area of computer vision, involves the detailed pixel-wise classification of images into distinct semantic categories. This process is fundamental to a variety of applications, ranging from autonomous driving [14] and medical imaging [35] to video surveillance [26] and environmental monitoring [22, 28]. Recent advancements in deep learning, particularly with Convolutional Neural Networks (CNNs) [9, 10, 24, 36], U-Net architectures [35], and the integration of skip connections [13], have propelled the field forward, enabling more accurate and efficient segmentation models.

However, the scope of semantic segmentation extends beyond conventional RGB imagery, embracing multimodal data sources that enrich the models' understanding of environments [14, 34, 43]. On one hand, multiple modalities can significantly enhance the environmental understanding of a model, leading to more accurate and detailed segmentation results. On the other hand, the integration of such varied modalities presents unique challenges [17]. While simply stacking or concatenating these data types can offer performance boosts, it is the nuanced interactions between modalities that hold the key to unlocking new levels of segmentation accuracy [2, 4, 29, 34, 40, 43, 44]. Nevertheless, not all fusion strategies are equal; some may inadvertently hinder performance by failing to meaningfully integrate features from different sources [27].

In response to these challenges, we propose a novel, hybrid transformer-based architecture for multimodal semantic segmentation. By leveraging the global attention mechanisms of transformers, along with the high resolution local feature extraction power of convolutional filters, the proposed architecture surpasses state-of-the-art models in deciphering complex interactions between different data modalities, thereby enhancing segmentation performance.

To validate the effectiveness of our architecture, we conduct experiments across various multimodal datasets with homogeneous modalities, demonstrating its advantages in diverse scenarios. Homogeneous modalities refer to the case where the input data has the same structure. For a homogeneous dataset with two modalities, $\mathcal{D} = \{(x_1^{(1)}, x_1^{(2)}), ..., (x_n^{(1)}, x_n^{(2)})\}$, both x_1 and x_2 belong to the same feature space $\mathcal{R}^d$. Whereas, a heterogeneous dataset may have $x^{(1)} \in \mathcal{R}^{d_1}$ and $x^{(2)} \in \mathcal{R}^{d_2}$.

The contributions of our research can be summarized as follows:

- A novel multimodal transformer model, FuseForm, is introduced, which can efficiently work with a large number of homogeneous modalities.
- A hybrid data fusion block based on multimodal cross attention and convolutions to facilitate passing both global and local representation of multimodal features to the decoder.
- A transformer-based decoder, which allows for enhanced feature integration over traditional methods while maintaining a lightweight implementation.

- FuseForm achieves state-of-the-art performance on five datasets from distinct applications in both single-modality and multimodal implementations (2-12 modalities).

2. Related Works

2.1. Semantic Segmentation

Semantic segmentation is a more detailed form of image classification, where each pixel in the image is assigned one or more class labels, producing a segmentation map the same size as the input image [35]. Traditional deep learning architectures achieve this through Convolutional Neural Networks (CNNs) [3, 9, 20, 28, 35, 36]. Fully Convolutional Networks (FCNs) pioneered this approach by performing pixel-wise segmentation from end to end, replacing the dense layers with convolutional layers [36]. U-Net [35], notable for its symmetric encoder-decoder architecture, achieved new levels of accuracy in semantic segmentation of medical images, and subsequently influenced many developments in semantic segmentation.

2.2. Vision Transformers

The introduction of Vision Transformer (ViT) [12] has applied the benefits of Natural Language Processing (NLP) transformers to images, leading to transformer-based architectures replacing CNNs in state-of-the-art computer vision tasks. The use of transformers as feature extractors in sequence-to-sequence networks was demonstrated in [46], showcasing their ability to capture long-range dependencies and contextual information, thus enhancing semantic segmentation tasks. The Pyramid Vision Transformer (PVT) [39] integrates a hierarchical structure to combine the strengths of transformers and pyramid designs common in segmentation models, effectively capturing multi-scale features for dense classification [7, 32, 35, 40–44].

Building on these advancements, SegFormer [42] refined the transformer semantic segmentation pipeline, using PVT's feature extractor with a lightweight, MLP-based decoder to achieve exceptional segmentation performance while maintaining computation efficiency. SegFormer has since become the foundation of many new implementations, showcasing how versatile the architecture is [34, 40, 43, 44].

2.3. Multimodal Data Fusion

Multimodal data fusion integrates information from multiple data sources, enhancing system performance by providing a more robust and comprehensive understanding than single-modal data. The primary challenge lies in efficiently combining different types of data, such as images, text, or audio, which often vary in nature, processing requirements, and information sparsity.

Recent works in the field of multimodal computer vision have investigated rigorous data fusion techniques at every layer in the model. Transformer blocks can integrate global contextual information from different modalities at each layer in the feature extractor [33]. In MFTransNet [47], they use a hybrid CNN-Transformer architecture with multi-headed self-attention for data fusion in the encoder, augmenting RGB data with Digital Surface Maps (DSMs) information to enhance complementary information. C3Net [5], a cross modal feature re-calibration module is used to learn from multiple modalities while minimizing noise impact. TokenFusion [40] adapts the SegFormer architecture for multimodal data by using an auxiliary network to dynamically replace low information tokens with high information tokens from other modalities, improving performance.

CMNeXT [44] employs asymmetric branches for multimodal dense prediction tasks, utilizing SegFormer for the optical modality and a unique self-query hub to select informative features from supplementary modalities, which are fused before the decoder. MMSFormer [34] uses modality specific encoders to extract different hierarchical features and combines them using a lightweight multimodal fusion strategy, improving performance as more modalities are added.

3. Methodology

Since the intricate dependencies in multimodal data cause a more complex learning problem than the better understood unimodal image segmentation task, with FuseForm we aim to develop more effective multimodal data fusion and decoding techniques for extracted multimodal features. Existing models typically utilize either convolution-based or transformer-based fusion techniques. The theoretical motivation behind our proposed multimodal data fusion module is to combine the local information processing of convolution operation through its inductive bias, such as local receptive field at different resolutions and weight sharing, and the global information processing of multimodal cross-attention operation through discovering correlations beyond the local connectivity assumption of convolution. As for the proposed decoder, our theoretical motivation is the hypothesis that the multimodal input data would benefit from further information processing in the decoding phase in addition to feature encoding and data fusion, which the existing methods focus on. For developing an effective decoder, we were motivated by the theoretical underpinnings of attention and convolution operations similar to the ones discussed for the fusion module.

3.1. Encoder

Each modality is processed with a separate encoder to ensure extraction of relevant features. We adopt a hierar-

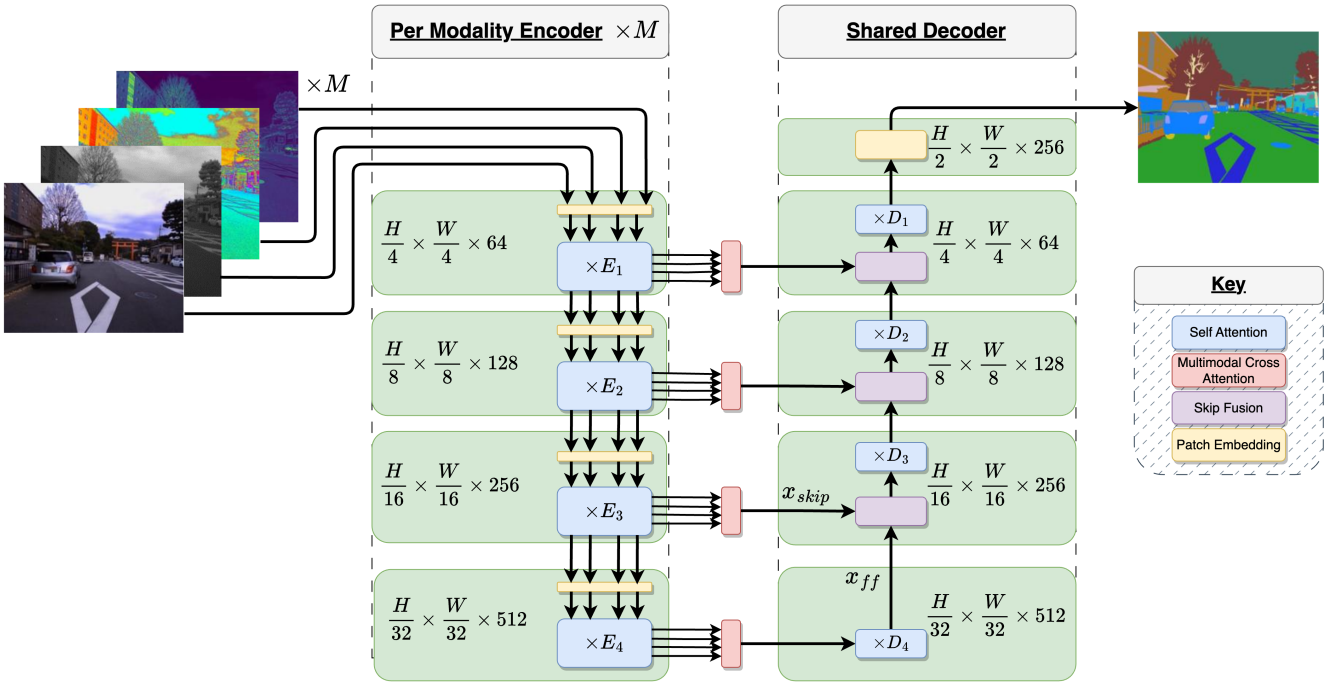

Figure 1. FuseForm architecture.

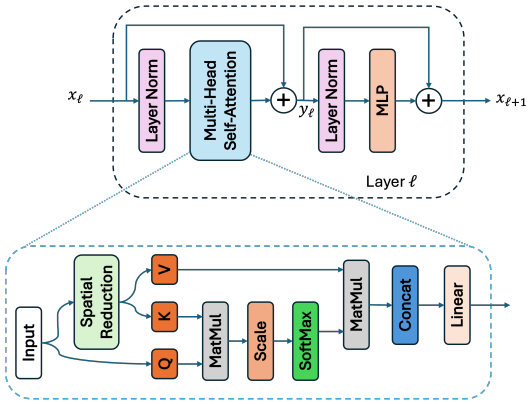

Figure 2. Transformer block with E_n layers.

chical design for the feature encoder similar to the Mix-Transformer [42]. Taking an image of size $H \times W \times C$, we split the image into $\frac{HW}{s_1}$ patches, where s_1 denotes the stride in stage one. The patches are flattened and embedded into $\frac{HW}{s_1}$ tokens with channel depth C_1 using a single convolutional layer. This process effectively downsamples the image resolution by a factor of $\sqrt{s_n}$ at each stage $n \in \{1, 2, 3, 4\}$.

In each stage n, after patch embedding, there is a transformer block with E_n layers. Denoting the input to layer ℓ with x_ℓ, as shown in Fig. 2, multi-head self-attention (MHSA) is applied to x_ℓ after layer normalization (LN) with a residual connection. Then, after another LN, a multilayer perceptron (MLP) is applied with a residual connection to yield the input to the next layer $x_{\ell+1}$:

$$x_{\ell+1} = MLP(LN(y_\ell)) + y_\ell, \;\; y_\ell = MHSA(LN(x_\ell)) + x_\ell.$$

In MHSA, spatial reduction is applied as in PVT (Fig. 2).

3.2. Multimodal Data Fusion

As each encoder produces its own output for each modality, to reduce model size and complexity, we introduce a data fusion block which intelligently combines informative features from each modality before sending them to a shared decoder. The information flow for the model is shown in Fig. 1. To facilitate this multimodal data fusion, we propose a hybrid attention-convolution mechanism. The multimodal fusion block, which consists of two main modules, Global Fusion and Local Fusion, is illustrated in Fig. 3. Global Fusion is based on a novel multimodal cross-attention (MMCA) mechanism, whereas Local Fusion uses a mixture of different size convolution operations.

The data flow through the multimodal fusion block with input $x_m, \;\; m \in M$, is represented through the following equation.

$$\begin{aligned} x'_m &= x_m + GF(x_m) + LF(x_m), \;\; m \in M \\ x &= LN(Linear(Concat(x'_m|_m^M))), \end{aligned} \tag{1}$$

where GF and LF are the Global Fusion and Local Fusion modules introduced in Sec. 3.2.1, and Sec. 3.2.2. Each

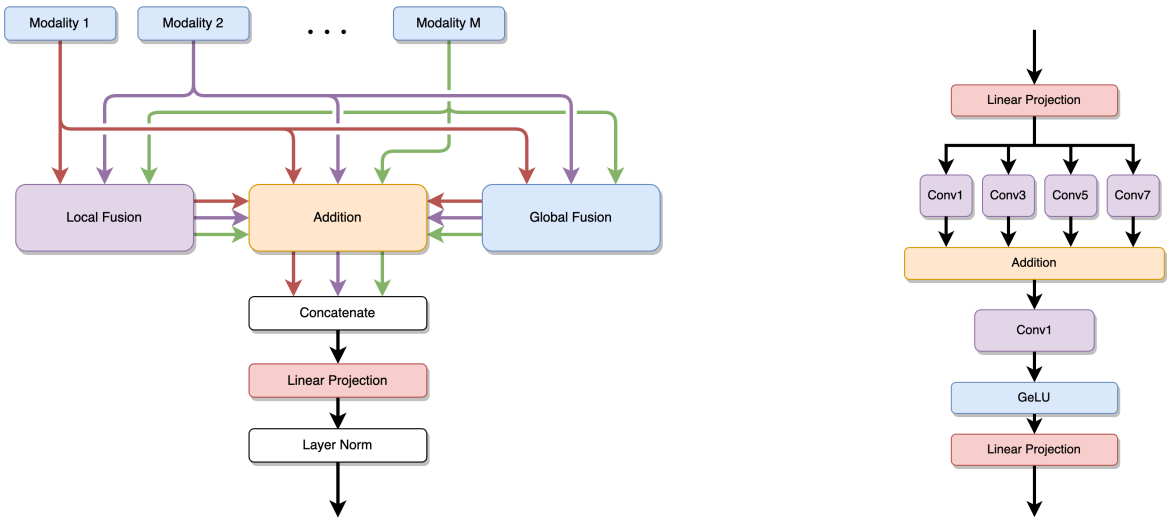

Figure 3. **(Left)** Overall data flow of the proposed Multimodal Fusion block. Single modality tokens are fed into both the Local and Global Fusion modules, residually added, then passed then projected and normalized to provide the decoder with a highly informative representation of the input image. **(Right)** Local Fusion module graphical representation. Data is first linearly projected, then fed into parallel convolutions before being combined and projected again.

modality is updated with its GF output. Then, the modalities are linearly projected from a $C_n \times M$ dimensional space to C_n dimensional space. The homogenized features are then fed into the Local Fusion module, which extracts relevant local features. The highly informative output feature vector passes through a layer normalization before being sent to the decoder.

3.2.1 Global Fusion:

Cross attention originates from the NLP transformer [38], where the query from one language's token vector is compared against the keys and values from another language. We extend this to facilitate multimodal data fusion by comparing a modality's query to the keys and values of all other modalities. Over time, this allows for high information tokens from all modalities to be emphasized while overall lower information tokens to be ignored. Our experiments show that our implementation using MMCA improves fusion module performance versus other MHSA based implementations [25, 34]. Fig. 4 summarizes the MMCA mechanism. Taking u_m to be an input token vector from modality m, we first linearly map each u_m to its query, (Q_m). We perform spatial reduction from PVT [39] on each u_m separately before passing it through the key (K_m) and value (V_m) linear projections. To compute the MMCA for modality m, its query Q_m is multiplied with all key vectors from other modalities to obtain the multimodal attention weight matrix through the softmax function. The resulting weights are multiplied with the value vectors of other modalities and linearly projected to obtain a fused token vector.

3.2.2 Local Fusion:

While the attention mechanism of transformers excels at aggregating global features and enhancing feature extraction, there may still exist localized correlations within the multimodal data, which can enhance overall model performance. Other implementations, [4, 6, 11, 32, 47], rely only on attention based mechanisms for feature fusion, which can skip over these local correlations. To extract these features, we enhance the global feature representation of MMCA with important local information. Our Local Fusion module, shown in Fig. 3, performs parallel convolutions of different kernel sizes over the tokens to extract different scale features. We utilize kernels of size 1×1, 3×3, 5×5, and 7×7 before passing through a shared 1×1 kernel convolution layer. The resulting output is combined with the Global Fusion output in the Multimodal Fusion Block.

3.3. Decoder

One of our closely related works, SegFormer [42], and many of its multimodal implementations [34, 40, 43, 44] utilize a hierarchical transformer encoder with a lightweight decoder. While the lightweight decoder is aimed at enabling real-time segmentation of single modality imagery, recognizing the limitations in scenarios demanding superior segmentation accuracy, our research introduces a new, high-performance transformer decoder. Our decoder consists of two main parts which are explained in the following sections.

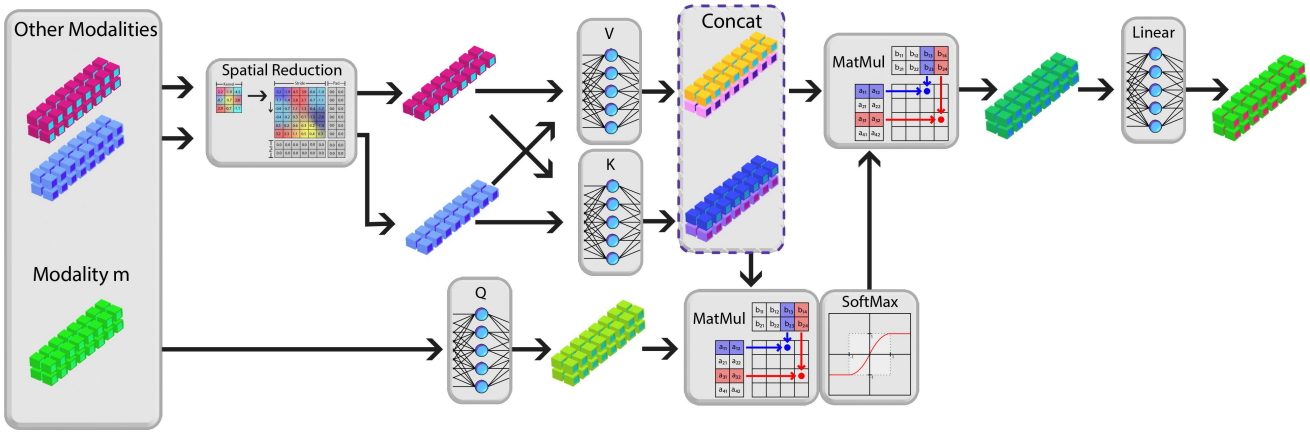

Figure 4. Data flow through the Global Fusion module, which performs multimodal cross-attention (MMCA) between different modalities.

3.3.1 Skip Fusion:

After data is fused through the multimodal fusion blocks in the encoder at stage n, it is sent directly to the decoder at stage n. To facilitate merging this skip connection information with the feed-forward data, we implement a Skip Fusion block. The Skip Fusion block is modeled using the following equation,

$$x_\ell = Conv_{1\times1}((Proj(Conv_{1\times1}(z)))), \tag{2}$$

where z is the concatenation of $\omega_1 x_{ff}$ and $\omega_2 x_{skip}$, ω_1 and ω_2 are learnt coefficients to weight the inputs. As shown in Fig. 1, x_{ff} is the feedforward output of the previous decoder stage, and x_{skip} is the output of the Multimodal Fusion module in the same stage. $Conv_{1\times1}$ are 1×1 kernel convolutions. In between each $Conv_{1\times1}$, there is a $Proj$ sequential network which consists of $Linear$, $ReLU$, $Linear$, and $ReLU$ layers to project the convolved input into a feature space of $C_n \times 2$ dimensions before projecting back to C_n dimensions. The final combined feature representation is passed through batch normalization and $ReLU$ activation before being sent to the decoder blocks in the same stage n of the decoder.

3.3.2 Decoder Block:

The decoder block is a modified version of the Multimodal Fusion block shown in Fig. 3. As there is only a single, fused modality input to the decoder block, we utilize standard MHSA for global feature extraction rather than Global Fusion introduced in Section 3.2.1. Our decoder block is modeled according to the following equation

$$\begin{aligned} y_\ell &= x_\ell + Proj(Concat(LF(x_\ell) + (MHSA(x_\ell))) \\ x_{\ell+1} &= y_\ell + MLP(BN(y_\ell)), \end{aligned} \tag{3}$$

where LF is Local Fusion as described in Section 3.2.2, and $Proj$ is a linear projection to reduce the dimensionality from $C_n \times 2$ to C_n. In the presented experiments, we utilize two decoder blocks in each layer, i.e., $D_1 = D_2 = D_3 = D_4 = 2$. After the final stage in the decoder, we utilize a linear projection to increase the spatial resolution of the decoded image to be closer to the input size. The final expansion layer projects the input from C_1 to $C_1 \times 16$, then subsequently rearranges the output from $[H/4, W/4, C_1 \times 16]$ to $[H/2, W/2, C_1 \times 4]$. This expansion allows for increasing the resolution and detail of segmentation maps, by allowing the model to attend to more pixels with each token in the decoder.

4. Experiments

We test the efficacy of our proposed FuseForm model on five multimodal semantic segmentation datasets. We compare our model's results to the state-of-the-art on each dataset. We also test our model's ability for RGB-only semantic segmentation, showcasing its versatility for many different tasks.

4.1. Datasets

The Whu-Opt-SAR [28] dataset is a terrain classification dataset which consists of 100 images of dimension 5556×3704 pixels spanning 51,488 km^2 across 7 classes. It combines satellite and drone imagery across optical (RGB), near-infrared (N), and Synthetic Aperture Radar (SAR) modalities.

The MCubeS [29] dataset contains 500 images, each of 1024×1224 pixels resolution, spanning 20 material classes. The dataset integrates four modalities, including optical (RGB), Angle of Linear Polarization (A), Degree of Linear Polarization (D) and Near-Infrared (N).

MFNet [18] is an urban street dataset containing 1,569 image sets total of both optical (RGB) and thermal (T) imagery. The resolution for each image set is 640×480 with 8 total classes for semantic segmentation. Out of the 1,569

images, 820 of them were collected during daytime and the remaining 749 were taken at nighttime.

The DeLiVER [44] dataset is a large-scale multimodal semantic segmentation dataset containing 4 modalities: optical (RGB), depth (D), LiDAR (Li), and event (E) views. The dataset consists of 3,983/2,005/1,897 image sets for training, validation, and test respectively, for a total of 7,885 image sets of resolution 1042×1042. The dataset contains 25 classes ranging from buildings and pedestrians to 4 different classes of vehicles.

The Next Day Wildfire Spread dataset [22] is a multimodal next frame prediction dataset aimed at predicting wildfire spread for the next day. The input consists of 12 inputs derived from seven modalities: Fire masks, topography (elevation), weather (temperature, wind, precipitation, humidity), drought indices, vegetation indices, and population density. The dataset comprises 18,545 samples with 64×64 resolution and an 80:20 split for training and testing.

4.2. Implementation Details

Each dataset features a number of spatially aligned modalities, which we feed into separate encoder networks. We trained FuseForm on various hardware, ranging from one 4090M laptop GPU, two 4090 desktop GPUs, to six H100 GPUs. Model performance is not affected by hardware selection, only training and inference speed. We train for 200 (DeLiVER, Whu-Opt-SAR, Next Day Wildfire) / 500 (MFNet, MCubeS) epochs in total with 10 epochs of warm-up, where the learning rate linearly ramps from 6×10^{-7} to 6×10^{-6}. We use AdamW optimizer with an epsilon of 10^{-8} and weight decay of 10^{-2}, with a poly strategy (power 0.9) loss scheduler. The minimum learning rate is capped to 10^{-9}. We utilize a batch size of 2 on each GPU, and perform image augmentations through random resized cropping, color jitter, flipping, and Gaussian blur. We utilize Mix-Transformer (MiT) [42] pretrained on ImageNet dataset as the backbone for our encoder and a random initialization for our transformer decoder. For each dataset, we choose the size of our backbone to be equivalent to other methods which use Mix-Transformer.

4.3. Comparison with Existing Methods

Results on Whu-Opt-SAR: As seen in Table 1, compared to other methods, our FuseForm model obtained the highest median Intersection over Union (mIoU) score of 48.4%, outperforming MCANet [28] by 5.5% and OPTSARMSNet [20] by 3.2% when using MiT-B4 [42] as the encoder backbone for the RGB branch and MiT-B2 for the N and SAR branches. Our single modality RGB only result is still 3.3% higher than MCANet and 1.3% higher than OPTSARMSNet, both which all modalities. Our single modality implementation of FuseForm outperforms task-specific multimodal models on this dataset, highlighting the effectiveness of our advanced transformer decoder at extracting and decoding relevant information.

Table 1. Results on Whu-Opt-SAR dataset [28]. The dataset contains RGB, Near-Infrared (N), and Synthetic Aperature Radar (SAR) modalities.

Method	Modals	mIoU
SegNet [3]	RGB-N-SAR	37.4
DeeplabV3+ [9]	RGB-N-SAR	41.2
MCANet [28]	RGB-N-SAR	42.9
OPTSARMSNet [20]	RGB-N-SAR	45.2
FuseForm	RGB	46.5
FuseForm	**RGB-N-SAR**	**48.4**

Table 2. Results on MCubeS dataset [29]. The dataset contains RGB, Angle of Linear Polarization (A), Degree of Linear Polarization (D), and Near-Infrared (N) modalities.

Method	Modals	mIoU
MCubeSNet [29]	RGB	33.70
CMNeXt [44]	RGB	48.16
MMSFormer [34]	RGB	50.44
FuseForm	**RGB**	**52.08**
DRConv [8]	RGB-A-D-N	34.63
DDF [48]	RGB-A-D-N	36.16
TransFuser [32]	RGB-A-D-N	37.66
MMTM [23]	RGB-A-D-N	39.71
FuseNet [19]	RGB-A-D-N	40.58
MCubeSNet [29]	RGB-A-D-N	42.86
CMNeXt [44]	RGB-A-D-N	51.54
MMSFormer [34]	RGB-A-D-N	53.11
FuseForm	**RGB-A-D-N**	**54.70**

Results on MCubeS: Table 2 presents a comprehensive comparison of various methods evaluated on the MCubeS dataset. Our model demonstrates exceptional performance across both RGB and RGB-A-D-N modalities, significantly outperforming all other methods listed. Using all 4 modalities and MiT-B4 [42] as the backbone for RGB and MiT-B2 for accompanying modalities, we see an improvement in mIoU of 11.84% over MCubeSNet [29], 3.16% over CMNeXt [44], and 1.59% over MMSFormer [34]. These results demonstrate the effectiveness of our fusion module as well as our transformer decoder. Using only the RGB modality, we still see an improvement of 18.38% over the RGB implementation of MCubeSNet [29], 3.92% over CMNeXt [44], and a 1.64% improvement over MMSFormer [34].

Results on MFNet: The results for MFNet can be seen in Table 3. Using MiT-B4 [42] as the encoder backbone for the RGB branch and MiT-B2 for the Thermal branch, our model outperforms other general segmentation models on this dataset by a margin of 1.0% in terms of mIoU. FuseForm also outperforms SegFormer in RGB only segmentation by a margin of 0.8%.

Results on DeLiVER: As shown in Table 4, we con-

Table 3. Results on MFNet dataset [18]. The dataset contains both RGB and Thermal (T) images.

Method	Modals	mIoU
SwinT [30]	RGB	49.0
SegFormer [42]	RGB	52.0
FuseForm	**RGB**	**52.8**
ACNet [21]	RGB-T	46.3
FuseSeg [37]	RGB-T	54.5
ABMDRNet [45]	RGB-T	54.8
LASNet [25]	RGB-T	54.9
FEANet [11]	RGB-T	55.3
MFTNet [47]	RGB-T	57.3
GMNet [49]	RGB-T	57.3
DooDLeNet [16]	RGB-T	57.3
CMX [43]	RGB-T	59.7
CMNeXt [44]	RGB-T	59.9
FuseForm	**RGB-T**	**60.9**

Table 4. Results on DeLiVER dataset [44]. The dataset contains RGB, Depth (D), Events (E) and LiDAR (Li) modalities.

Method	Modals	mIoU
HRFuser [4]	RGB	47.95
Segformer [42]	RGB	57.20
FuseForm	**RGB**	**57.93**
HRFuser [4]	RGB-D	49.32
TokenFusion [40]	RGB-D	60.25
CMNeXt [44]	RGB-D	63.58
FuseForm	**RGB-D**	**68.34**
HRFuser [4]	RGB-E	42.22
TokenFusion [40]	RGB-E	45.63
CMNeXt [44]	RGB-E	57.48
FuseForm	**RGB-E**	**57.53**
HRFuser [4]	RGB-Li	43.13
TokenFusion [40]	RGB-Li	53.01
CMNeXt [44]	**RGB-Li**	**58.04**
FuseForm	RGB-Li	58.00
HRFuser [4]	RGB-D-E-Li	52.97
CMNeXt [44]	RGB-D-E-Li	66.30
FuseForm	**RGB-D-E-Li**	**68.49**

duct in depth testing to show how FuseForm compares to the state-of-the-art methods on this dataset. We utilize MiT-B2 [42] as the encoder backbone for all branches. We improve segmentation performance over state-of-the-art in single, dual, and quad modality tests. We outperform CMNeXt [44] by 4.76% in RGB-D tests, which also outperforms their quad-modality result by 2.04%. With all modalities, this gap increases to 2.19%.

Results on Next Day Wildfire Spread: The performance of FuseForm is shown in Table 5 compared to various U-Net implementations [1, 15, 22, 31] and the Wildfire Prediction Network [15]. We initialize MiT-B0 [42] for each modality and train for 200 epochs. FuseForm outperforms other implementations on this dataset by a wide margin. Our model improves on Wildfire Prediction Network in precision by 14.2%, recall by 4.0% (equivalent to R2U-Net [1]), F1 by 0.03, and mIoU by 11.8%. This dataset, while low resolution, shows how our model can combine a large number of modalities effectively to enhance performance over existing methods.

Table 5. Results on Next Day Wildfire Prediction Dataset [22].

Method	Precision	Recall	F1 Score	mIoU
U-Net [35]	28.2	47.8	0.36	15.0
R2U-Net [1]	25.9	**48.8**	0.34	14.6
Attn U-Net [31]	30.3	44.8	0.36	14.8
R2AttnU-Net [15]	28.3	47.6	0.36	14.9
WPN [15]	30.3	44.8	0.36	14.4
FuseForm	**43.5**	**48.8**	**0.39**	**26.2**

Table 6. Computational complexity and performance comparison. Tested with a 512×512 resolution, with 1 and 4 modalities (shown in parentheses). The encoder backbone chosen is MiT-B4 for all methods. IPS denotes the number of processed images per second. The number of parameters is given in millions.

Model	GFLOPs	#Params(M)	IPS
CMNeXt (1)	72.2	62.4	30.8
MMSFormer (1)	72.3	62.4	31.6
FuseForm (1)	87.9	100.0	20.5
CMNeXt (4)	124.0	117.0	11.9
MMSFormer (4)	136.0	138.0	14.3
FuseForm (4)	153.0	202.0	10.4

4.4. Computational Complexity

We compare the computational efficiency of FuseForm with other state-of-the-art methods in Table 6 using the MCubeS dataset. Our method significantly outperforms the existing methods at the cost of a reasonable increase in the computational complexity. This result indicates that our method can be preferred over other methods for systems that can tolerate some more computations (e.g., $\sim$12% for four modalities).

We further analyze in Table 7 the size and computational complexity of the novel parts in FuseForm. Looking at the fusion block, our model falls on the lower end in terms of number of parameters, but is the most computationally intensive method. This is primarily due to the use of full QKV-based cross attention in our block. We also compare our decoder with two other popular decoders, the decoder of SegFormer [42] (Used in [34, 40, 43, 44], among others) and U-Net [35] (Used in [1, 15, 22, 31], among others). Our decoder is roughly $2\times$ in size to the convolution-based U-Net while needing 75% more GFLOPs. Although our decoder has $4\times$ more parameters than the SegFormer decoder with an embedding depth of 768, it is $3\times$ more computa-

Table 7. Comparison of number of parameters and computational efficiency. Tested on the MCubeS dataset at $3 \times 512 \times 512$ resolution input with 4 modalities present.

Method	#Params(M)	GFLOPs
Fusion Module		
- CMNeXt [44]	16.63	6.47
- MCubeSNet [29]	7.41	12.10
- HRFuser [4]	**1.72**	17.50
- MMSFormer [34]	3.23	**2.47**
- FuseForm	5.67	19.19
Decoder		
- Segformer [42]	**3.15**	40.29
- U-Net [35]	6.54	**7.59**
- FuseForm	13.97	13.46

Table 8. Ablation study of FuseForm's multimodal data fusion module and decoder. Tested on the MCubeS dataset with 4 modalities. #Params is given for the fusion module and decoder.

Setup	#Params(M)	mIoU
FuseForm	19.64	54.70
Fusion Module		
- without local fusion (Eq. (1))	18.01	53.52
- without global fusion (Eq. (1))	17.16	53.85
- replace CA with SA	18.01	52.16
Decoder		
- without local fusion (Eq. (3))	15.68	50.65
- without MHSA (Eq. (3))	13.87	54.22
- without skip fusion (Eq. (2))	18.76	51.02
- SegFormer decoder [42]	8.82	52.42

tionally efficient. The SegFormer decoder uses a large convolutional layer of the full output resolution of the model with $768 \times 4 = 3072$ filters, which amounts to over 38 GFLOPs alone. These results show that a vast majority of FuseForm's size and computations come from the standard encoder, which is an important future research direction.

4.5. Ablation Study

In this section, we investigate individual contributions of key components within our FuseForm architecture. Shown in Table 8, we investigate the impacts of local and global fusion modules, attention types, and decoder type on the performance of our model as a whole.

Multimodal Data Fusion: Our multimodal fusion block is a combination of both high-resolution local features and overall global context. To determine the contribution of both modules, we disable each sequentially and test FuseForm on the MCubeS dataset by loading our baseline weights and training the model for 200 iterations with the new configuration. Table 8 illuminates how vital each module is to performing effective data fusion. By having local fusion only, the model is missing vital global context, which causes a large drop in mIoU to 53.85%. Similarly, the drop in performance due to the model only being able to see the global context but not local details such as texture is 53.52%. In a surprise, swapping cross attention (CA) for self attention (SA) in our fusion block sees a drop to 52.16%, more than when removing global fusion. A possible explanation is that without any attention mechanism, global context is completely lost in the fused features. With SA however, we are still able to extract modality specific context, which is sometimes contrasting and causes information loss.

Decoder: The ablation study on the decoder of the FuseForm architecture shows the critical role of different fusion mechanisms in enhancing semantic segmentation performance. Removing local fusion from the decoder (Eq. (3)) results in a notable drop in mIoU to 50.65%, highlighting its importance in capturing fine-grained details. Similarly, the absence of MHSA in the decoder causes a decrease in mIoU to 54.22%, which shows it also contributes to overall performance, albeit less than local fusion. To test the effectiveness of the Skip Fusion module, we replace Eq. (2) with a simple addition of x_{ff} and x_{skip}. The resulting mIoU declines to 51.02%, displaying the contribution of the Skip Fusion module to integrating multi-scale features effectively. Notably, replacing the FuseForm decoder with the decoder of SegFormer [42], leads to a reduced mIoU of 52.42%. This suggests that the sophisticated fusion strategies within the FuseForm decoder are essential for achieving superior segmentation accuracy.

5. Future Work and Conclusion

In this paper, we introduce FuseForm, a hybrid convolutional-transformer fusion module alongside a novel transformer-based decoder, replacing SegFormer's [42] widely-used decoder. Our comprehensive experiments show that FuseForm effectively merges data from various modalities, leading to state-of-the-art performance on five major datasets: DeLiVER [44], MCubeS [29], MFNet [18], Whu-Opt-SAR [28], and Next Day Wildfire Spread [22]. We perform detailed ablation studies to understand the contributions of each component within the fusion module and decoder to the FuseForm's overall performance. Nevertheless, a limitation of FuseForm is the need for modality-specific encoders, which increases computational complexity with the number of modalities. Future research will focus on exploring new architectures which can leverage a shared encoder, reducing model size, as well as extending the model's capabilities to other multimodal tasks.

Acknowledgment

This research was funded by US Army Corps of Engineers ERDC W9132V-22-2-0001 and Chih Foundation.

References

[1] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, 2018. 7

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks, 2017. 1

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2, 6

[4] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection, 2023. 1, 4, 7, 8

[5] Zhiying Cao, Wenhui Diao, Xian Sun, Xiaode Lyu, Menglong Yan, and Kun Fu. C3net: Cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images. *Remote Sensing*, 13(3), 2021. 2

[6] Zhiying Cao, Wenhui Diao, Xian Sun, Xiaode Lyu, Menglong Yan, and Kun Fu. C3net: Cross-modal feature recalibrated, cross-scale semantic aggregated and compact network for semantic segmentation of multi-modal high-resolution aerial images. *Remote Sensing*, 13(3), 2021. 4

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. 2

[8] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution, 2021. 6

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. 1, 2, 6

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 1

[11] 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, 2021. 4, 7

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[13] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation, 2016. 1

[14] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, 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, 2021. 1

[15] Jack Fitzgerald, Ethan Seefried, James E Yost, Sangmi Pallickara, and Nathaniel Blanchard. Paying attention to wildfire: Using u-net with attention blocks on multimodal data for next day prediction. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 470–480, New York, NY, USA, 2023. Association for Computing Machinery. 7

[16] Oriel Frigo, Lucien Martin-Gaffé, and Catherine Wacongne. Doodlenet: Double deeplab enhanced feature fusion for thermal-color semantic segmentation, 2022. 7

[17] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities, 2022. 1

[18] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 5, 7, 8

[19] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer, 2017. 6

[20] Wei Hu, Xinhui Wang, Feng Zhan, Lu Cao, Yong Liu, Weili Yang, Mingjiang Ji, Ling Meng, Pengyu Guo, Zhi Yang, and Yuhang Liu. Opt-sar-msnet: A multi-source multi-scale siamese network for land object classification using remote sensing images. *Remote Sensing*, 16:1850, 05 2024. 2, 6

[21] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation, 2019. 7

[22] Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 1, 6, 7, 8

[23] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion, 2020. 6

[24] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, 2017. 1

[25] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023. 4, 7

[26] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *CoRR*, abs/2201.04676, 2022. 1

[27] Wangbin Li, Kaimin Sun, Wenzhuo Li, Jinjiang Wei, Shunxia Miao, Song Gao, and Qinhui Zhou. Aligning semantic distribution in fusing optical and sar images for land use classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:272–288, 2023. 1

[28] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022. 1, 2, 5, 6, 8

[29] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19800–19808, 6 2022. 1, 5, 6, 8

[30] 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, 2021. 7

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 7

[32] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving, 2021. 2, 4, 6

[33] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection, 2022. 2

[34] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Mmsformer: Multimodal transformer for material and semantic segmentation. *IEEE Open Journal of Signal Processing*, 5:599–610, 2024. 1, 2, 4, 6, 7, 8

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1, 2, 7, 8

[36] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2016. 1, 2

[37] 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 Transactions on Automation Science and Engineering*, 18(3):1000–1011, 2021. 7

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 4

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021. 2, 4

[40] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers, 2022. 1, 2, 4, 7

[41] Zongwei WU, Zhuyun Zhou, Guillaume Allibert, Christophe Stolz, Cédric Demonceaux, and Chao Ma. Transformer fusion for indoor rgb-d semantic segmentation. *Available at SSRN 4251286*, 2022. 2

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021. 2, 3, 4, 6, 7, 8

[43] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers, 2023. 1, 2, 4, 7

[44] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation, 2023. 1, 2, 4, 6, 7, 8

[45] 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 (CVPR)*, pages 2633–2642, June 2021. 7

[46] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021. 2

[47] Heng Zhou, Chunna Tian, Zhenxi Zhang, Qizheng Huo, Yongqiang Xie, and Zhongbo Li. Multispectral fusion transformer network for rgb-thermal urban scene semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2, 4, 7

[48] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks, 2021. 6

[49] 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. 7