MULTIMODAL FUSION METHODS WITH VISION TRANSFORMERS FOR REMOTE SENSING SEMANTIC SEGMENTATION

Veronica Grazia Morelli, Mirko Paolo Barbato, Flavio Piccoli, Paolo Napoletano

Department of Informatics, Systems and Communications (DISCo), University of Milano-Bicocca
Viale Sarca 336, 20126 Milan, Italy

ABSTRACT

This paper presents a comparative analysis of transformer-based fusion methods applied to a novel multimodal dataset for remote sensing semantic segmentation. This investigation evaluates the impact of several fusion methods on the accuracy of the results. In particular, for early fusion, we investigate the *Early Concatenation*. For middle fusion, we investigate four methods, namely the *Token Patch Embedding*, *Channel Patch Embedding*, *Token Fusion at Attention Level*, and *Cross-Attention*. Finally, as a representative of late fusion, we investigate the use of *Late Concatenation*. The methods presented here are specifically designed to operate effectively with all modalities under investigation. Experiments conducted on the Ticino dataset show that *Late Concatenation* outperforms the best single modality RGB method of 4.04%, 2.24% and 3.47% respectively on accuracy, precision and mIoU. This study provides an opportunity to further explore fusion methods utilizing transformers, thereby enhancing our understanding of the potential of data fusion.

***Index Terms*—** Remote sensing, Semantic Segmentation, Multimodal fusion, Vision Transformers

1. INTRODUCTION

Multimodal Learning (MML) represents a versatile approach to constructing AI models capable of extracting and correlating information from various data sources, which are commonly referred to as modalities [1]. In Remote Sensing (RS), each modality is often associated with a specific sensor, serving as a distinct information source characterized by its own unique statistical attributes [2]. The fusion of different RS data sources captured within the same geographic area holds great promise as a strategy to enhance material identification on the Earth's surface [3]. This approach exploits the information present in diverse data sources, enabling more detailed and precise scene understanding, particularly in challenging scenarios where individual modalities may struggle to differentiate between similar surface categories. One of the central tasks in RS is semantic segmentation. This method involves classifying each individual pixel within an image, thus yielding an output map with the same spatial extent as the input image. In this map, pixels are grouped into areas that share the same semantic class [4]. Semantic segmentation plays a crucial role in various remote sensing applications, including precision agriculture, environmental monitoring, land-use planning, ecosystem-oriented natural resource management, food supply management, nature conservation, and numerous other essential domains. Moreover, semantic segmentation, due to its inherent complexity, poses a unique challenge in the context of multimodal learning. Semantic segmentation, like several other computer vision tasks, has advanced significantly with the introduction of deep learning techniques. Notably, this progress is exemplified by the emergence of two key neural architectures: convolutional neural networks [5] and vision transformers [6]. Transformer-based architectures have emerged as a prominent choice in MML research, but their utilization for semantic segmentation of RS images remains relatively underexplored [7]. Transformers present two critical challenges: the need for large amounts of data, and high computational complexity due to the quadratic nature of the self-attention mechanism that characterizes them. To address these concerns, the Shifted-Window Transformer (Swin) was introduced to specifically resolve issues related to computational complexity [8], while data-efficient transformers were proposed to mitigate the demands for extensive training data [9].

In this work, we present a comprehensive analysis of multimodal fusion methods for semantic segmentation of RS images based on the use of Swin-UperNet transformers [8]. We experimented with several early, middle and late fusion methods on the newly introduced Ticino dataset [10]. This dataset consists of 1502 tiles covering an area of 1332 km^2, featuring three modalities: RGB, Hyperspectral (HS), and Digital Terrain Model (DTM); as well as a Land-Cover pixel-wise labeling. Results show the effectiveness of multimodal approaches when compared with single modality approaches, as well as they reveal that middle and late fusion methods achieve greater accuracy than an early fusion method. Moreover, insights regarding the computational complexity of the fusion methods will be discussed.

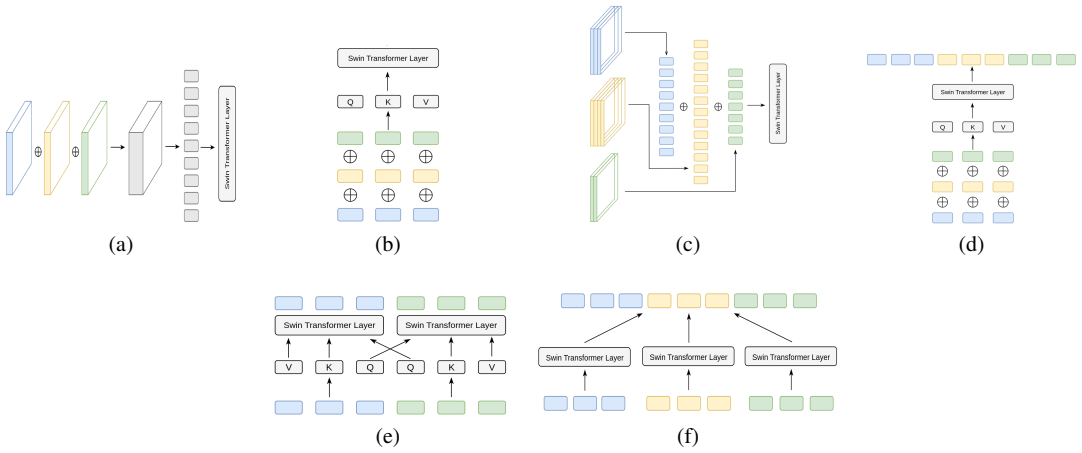

Fig. 1. Fusion methods schemes considered in this work: (a) Early concatenation; (b) Token Patch Embedding; (c) Channel Patch Embedding; (d) Token Fusion at Attention Level; (e) Cross-Attention; and (f) Late Concatenation.

2. METHODOLOGY

All the fusion techniques we experimented with, are based on U-shaped neural architecture composed of an encoder and a decoder module. The encoder is a hierarchical shifted window-based vision transformer (Swin) [8], while the decoder is a UperNet with skip connections [11], which is a powerful semantic segmentation model known for its effectiveness in capturing intricate spatial relationships and high-level context. We deployed and compared six multimodal fusion techniques: (1) *Early Concatenation* (EC); (2) *Token Patch Embedding* (TPE); (3) *Channel Patch Embedding* (CPE); (4) *Token Fusion at Attention Level* (TFA); (5) *Cross-Attention* (CA); (6) *Late Concatenation* (LC). A generalized schematic representation of these fusion methods can be seen in Figure 1. They are categorized into three classes based on where the fusion occurs: early fusion at the input level Figure 1(a), middle fusion at an intermediate point within the encoder Figure 1(b)-(e) and late fusion after the encoder's processing Figure 1(f). The investigated methodologies can vary considerably in terms of complexity, performance capabilities and computation requirements. For both *Early Concatenation* and *Late Concatenation* methods, we needed to modify the Swin-UperNet to accommodate all three modalities presented in the dataset used in our experiments. In the case of the middle fusion methods, we had to devise suitable strategies for integrating these three modalities, drawing inspiration from prior research in multimodal fusion.

2.1. Swin-based encoder

The encoder is based on the canonical Swin transfomer architecture [8], consisting of 4 stages $\{S_i\}_{i=1}^{4}$. Each Stage, apart from the first one, is characterized by a Patch Merging module and a Swin Transformer Block (STB). Each Block includes at least a pair of consecutive Window Multi-head Self Attention (W-MSA) and Shifted Window Multi-head Self Attention (SW-MSA) modules. The first stage S_1 consists of a Linear Embedding layer and a STB. At the beginning, the image is divided into N patches $\{p_i\}_{i=1}^{N}$ that are then introduced into the first Stage S_1. Here, each patch p_i is projected by the Linear Embedding ($embedding()$) layer into a token z_i. All tokens $Z = \{z_i\}_{i=1}^{N}$ enter into the STB and consequently in the self-attention modules that extract the new tokens and give them to the stage S_2. Each stage, from the second to the last, starts with the Patch Merging module that reduces the number of patches grouping them 2 by 2 and then giving them to the STB. Given the U-shape of our encoder-decoder model, intermediate representations produced after each stage of the Swin encoder are subsequently fed into the symmetric UperNet decoder using skip connections.

2.2. Fusion techniques

Let's consider the case of fusing three modalities $\{X_i\}_{i=1,2,3}$ (RGB, HS and DTM in this paper). Z_i denotes the respective set of token embeddings of the modality X_i and Z the input of the STB derived by the previous operations.

Early fusion. The simplest fusion strategy is the *Early Concatenation* (EC), where the images from multiple modalities are concatenated ($concat()$) at input level on channel dimension and then processed by one Swin-based Encoder:

$$X_{(1,2,3)} = concat(X_1, X_2, X_3)$$

$$Z = embedding(X_{(1,2,3)}).$$

Middle fusion. A middle fusion solution is the *Token Patch Embedding* (TPE) concatenation in which the token embedding sequences from multiple modalities are concatenated and fed into the Swin Transformer layers of the first STB [6]:

$$Z_i = embedding(X_i)\ with\ i = 1, 2, 3$$

$$Z = concat(Z_1, Z_2, Z_3).$$

Another middle fusion method is the *Channel Patch Embedding* (CPE) [12], which involves generating individual token embeddings for each channel within every modality. These embeddings are then concatenated and fed as input to the first STB. For example for hyperspectral data, this would correspond to the individual spectral bands, while for RGB data, to the different color channels. Formally:

$$Z_{i,j} = embedding(X_{i,j})$$

where i is for the modality and j for the channel of the modality. Then, for $i = 1, 2, 3$ and each channel j:

$$Z = concat(Z_{i,j}).$$

The *Token Fusion at Attention Level* (TFA) method involves processing the three modalities separately within three distinct Swin Transformer encoders, alternating one and three streams throughout the process. It has been designed by us as a variant of the *Token Patch Embedding* where the concatenation is done at the token level at each stage. Before computing W-MSA and before SW-MSA in each transformer block, the tokens generated by the three modalities up to that point are concatenated, allowing for joint attention computation. After attention computation, the outputs are divided ($split()$) and processed individually by the three encoders until the next attention module. In this particular case, let's also consider Y_i^l as the tokens of the i-th modality at stage l (in the first stage it will be equal to Z_i) and Y^l as the input of the Transformer Block in S_l. For each stage l the operations are as follows:

$$Y^l = concat(Y_1^l, Y_2^l, Y_3^l)$$

$$Y_{1,w}^l, Y_{2,w}^l, Y_{3,w}^l = split(WMSA(Y^l))$$

$$Y_w^l = concat(Y_{1,w}^l, Y_{2,w}^l, Y_{3,w}^l)$$

$$Y_1^{l+1}, Y_2^{l+1}, Y_3^{l+1} = split(SWMSA(Y_w^l)).$$

These operations are computed for every self-attention operation in every STB. Every Y^l is fed into the decoder through skip connections.

Cross Attention (CA) is a method used in two-stream Transformers [13], to facilitate cross-modal interactions by exchanging query embeddings between modalities. In this case, we leveraged the third modality following the idea outlined by Dufter et al. [14], and utilizing it as positional embedding. Considering Q_i, K_i, V_i, the query, key and value of the canonical self-attention technique for the i-th modality and MSA as self-attention operator valid for both W-MSA and SW-MSA, the cross attention between only X_1 and X_2 is computed as:

$$\begin{cases} M_1 = MSA(Q_2, K_1, V_1) \\ M_2 = MSA(Q_1, K_2, V_2) \end{cases}$$

where M_1 and M_2 are the token outputs for the first stream of modality 1 and the second stream of modality 2. Cross-attention allows for cross-modal interactions, highlighting the importance of considering self-attention within each modality for a more comprehensive understanding.

Late fusion. *Late Concatenation* (LC) works in a multi-stream mode. It involves processing the three modalities separately in three distinct Swin Transformer encoders. The output of each stage is then concatenated on the channel dimension and into the UperNet decoder. Formally, let's consider the output at each stage l for each i-th modality as O_i^l:

$$O^j = concat(O_1^l, O_2^l, O_3^l).$$

Each O^j is then used in the skip connection with the correspondent layer of the UperNet decoder.

3. EXPERIMENTS

The pansharpened version of the Ticino dataset [10] was used in our experiments. The data cover an area in the South of Milan of around 1332 km^2 with a total of 1502 tiles. This version of the dataset consists of three modalities: RGB, Pansharpenend Hyperspectral (HS↑) and DTM. These modalities offer intrinsic advantages defined by their nature. In particular, RGB is mainly considered for spatial information, HS↑ for spectral information and DTM for morphological structure. RGB has 3 color bands and a spatial resolution of about 2m/px. HS↑ has a spatial resolution of 5m/px and 182 spectral bands cleaned from corrupted bands [10] that covers both VNIR and SWIR components of the spectrum (400-2500nm), having more discriminating power with materials. DTM image consists of a single band with a spatial resolution of 5m/px and an elevation range from 51.86 to 124.75 meters. Finally, the labeling considered in our experiments is the Land Cover from the Ticino dataset. It consists of 8 semantic classes: *Background*, *Building*, *Road*, *Residential*, *Industrial*, *Forest*, *Farmland*, and *Water*.

To address the curse of dimensionality problem, typical of HS data, and to adapt it to our experiments and resources, we applied Principal Component Analysis to the pansharpened image, extracting spectrally homogeneous regions. The first four principal components were retained, accounting for 99% of the variation and resulting in a revised HS* with four spectral bands. To train and test the models, the 1502 tiles were split into 1051 training, 225 validation, and 226 test images [10].

3.1. Implementation Details

All fusion methods were implemented using three modalities: RGB, HS*, and DTM except for Cross-Attention in which we employed RGB and HS* as main modalities and DTM as positional embedding. Before training, a data augmentation strategy based on the Albumentations library [15] was

Table 1. Land Cover overall results for each method, dividing single modality and multimodality fusion methods. The bold and underlined values represent the best and the second-best performance achieved for each metric, respectively.

	Method	Acc	Pr	mIoU	Macs	Pars
Single	RGB	58.82	63.66	48.75	9.65	39.28
	HS*	50.84	54.87	39.82	9.65	39.28
	DTM	18.30	20.55	31.81	9.65	39.28
Multi	Early Conc. (EC)	59.12	64.01	49.05	9.68	39.29
	Tok. Pat. Emb. (TPE)	60.28	64.71	50.04	16.40	60.60
	Cha. Pat. Emb. (CPE)	56.89	62.28	47.12	65.43	241.96
	Tok. Fus. Att. (TFA)	60.49	64.99	50.32	16.14	38.74
	Cross-Att. (CA)	**62.87**	<u>65.38</u>	<u>51.99</u>	37.86	111.61
	Late Conc. (LC)	<u>62.86</u>	**65.90**	**52.23**	16.14	63.29

applied. It includes random cropping, rotation, and horizontal and vertical flipping to images resized to 256x256 pixels and normalized. All models employed a Swin encoder configuration with a patch size of 4, a window size of 7, and a depth specified as 2, 2, 6, 2 along with attention heads set to 3, 6, 12, 24 and expansion layer. Due to limitations in computational resources, the embedding dimension for each Swin Transformer was adjusted accordingly: 96 for Early Concatenation, Token Patch Embedding and Cross-Attention, 48 for Token Fusion at the Attention Level and Late Concatenation, and 24 for the Channel Patch Embedding method. All models were subject to a stochastic depth regularization of 0.3. For training, we employed Adam optimizer and trained for 250 epochs with an initial learning rate of 10^{-3} and weight decay 10^{-4}. A learning rate scheduler was also applied to reduce it. The cross-entropy loss was used for training. All experiments were run on NVIDIA GTX 1070 GPU with 8GB of RAM.

3.2. Results

This section presents the outcomes of the experiments, which have been assessed and examined using three evaluation metrics averaged on classes: Accuracy (Acc), Precision (Pr) and mean Intersection over Union (mIoU). We also measured the computational complexity of each method using the number (Million) of parameters of the neural model (Pars) and the number (Giga) of multiply–accumulate operations (Macs). Table 1 reports the results achieved by every tested method, comparing both single modality and multimodality approaches. Figure 2 shows examples of visual results from each method. As expected, due to the higher spatial resolution, RGB mode performs better among single-mode approaches, achieving superior performance to HS* and DTM on all metrics.

Comparing multimodal and single modality approaches, it is possible to note that, apart from *Channel Patch Embedding*, all the multimodal methods outperform RGB alone. Among these methods, *Cross-Attention* and *Late Concatenation* achieve the best results. They are both comparable,

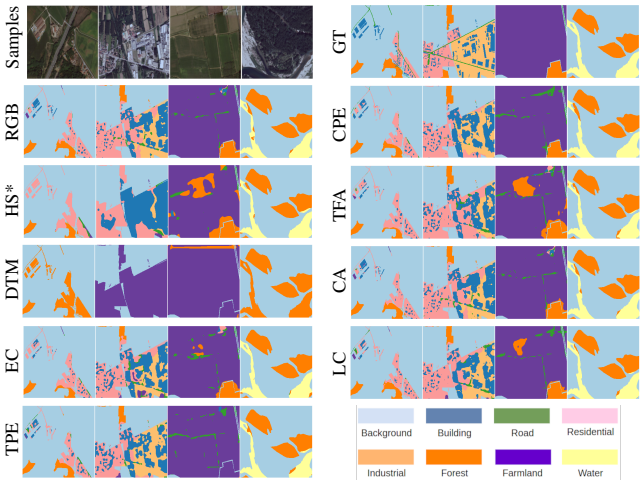

Fig. 2. RGB samples at top-left row; Ground-truth (GT) at top-right row; single modality results: RGB, HS* and DTM; multimodality results: EC, TPE, CPE, TFA, CA and LC.

nonetheless, the former is the best on Acc, while the latter reaches better results on Pr and mIoU. Nevertheless, when taking into account Pars and Macs in the analysis, it becomes evident that the *Late Concatenation* method exhibits a significantly lower number of Pars compared to the *Cross-Attention* approach. In contrast, the *Cross-Attention* method substantially increases the complexity of the RGB network by about twice. The same argument is valid for the Macs where the *late Concatenation* has less than half Macs than *Cross-Attention*. Therefore, we consider *Late Concatenation* as the best methods, outperforming RGB of 4.04%, 2.24% and 3.47% on Acc, Pr and mIoU, respectively. Figure 3 shows a comparison of all methods (excluding HS* and DTM models). Ideally, the best method is the one in the upper right part of the plot with a small circle that indicates the number of Pars, confirming the conclusion that *Late Concatenation* is the method that overall performs better. It is worth noting that *Token Fusion at Attention Level* represents an excellent trade-off between performance and resources used, since it is superior to RGB and, at the same time, has fewer Pars with comparable complexity in terms of Mac (equal to *Late Concatenation*).

4. CONCLUSION

In this paper, we presented a comprehensive analysis of multimodal fusion methods for RS semantic segmentation using vision transformers. We considered the multimodal dataset Ticino, with three modalities (RGB, HS↑, and DTM), and the transformer models Swin-UperNet. We compared single modality and multimodal approaches, considering six different fusion methods. In the comparison, performance, Macs, and parameters have been evaluated. Excluding *Channel*

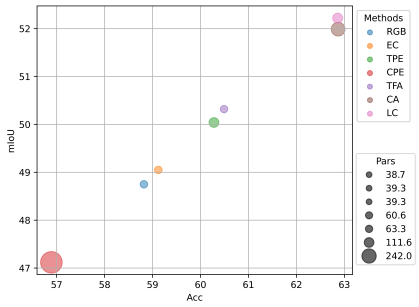

Fig. 3. Comparison of the performance fusion methods based on Acc (x), mIoU (y) and Parameters (area of the circles).

Patch Embedding, five of six multimodal approaches outperformed RGB and consequently any other single modality approach. In particular, compared with RGB, two methods distinguished themselves. The *Token Fusion at Attention Level* revealed to be the best compromise in terms of performance (outperforming RGB) and memory (parameters). The *Late Concatenation* method proved to be the best multimodal method. Results demonstrate that a multimodal approach is more efficient in terms of performance while keeping the consumption of resources comparable with single modality methods.

5. ACKNOWLEDGEMENTS

The authors are grateful to Prof. Begüm Demir for the valuable comments and stimulating discussions.

6. REFERENCES

[1] Dhanesh Ramachandram and Graham W Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[2] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[3] Qibin He, Xian Sun, Wenhui Diao, Zhiyuan Yan, Fanglong Yao, and Kun Fu, "Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling," *IEEE Transactions on Image Processing*, vol. 32, pp. 1474–1487, 2023.

[4] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, pp. 114417, 2021.

[5] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[6] Peng Xu, Xiatian Zhu, and David A Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] Abdulaziz Amer Aleissaee, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, pp. 1860, 2023.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of the IEEE/CVF Int. conference on computer vision*, 2021, pp. 10012–10022.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *Int. conference on machine learning*. PMLR, 2021, pp. 10347–10357.

[10] Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano, "Ticino: A multi-modal remote sensing dataset for semantic segmentation," *Avail. at SSRN 4535928*, 2023.

[11] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun, "Unified perceptual parsing for scene understanding," in *Proc. of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[12] David Hoffmann, Kai Norman Clasen, and Begüm Demir, "Transformer-based multi-modal learning for multi label remote sensing image classification," *arXiv preprint arXiv:2306.01523*, 2023.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[14] Philipp Dufter, Martin Schmitt, and Hinrich Schütze, "Position information in transformers: An overview," *Computational Linguistics*, vol. 48, no. 3, pp. 733–763, 2022.

[15] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, pp. 125, 2020.