Spatial Gated Multi-Layer Perceptron for Land Use and Land Cover Mapping

Ali Jamali,* Swalpa Kumar Roy,* *Member, IEEE*, Danfeng Hong, *Senior Member, IEEE*, Peter M Atkinson, and Pedram Ghamisi, *Senior Member, IEEE*

***Abstract*—Due to its capacity to recognize detailed spectral differences, hyperspectral data have been extensively used for precise Land Use Land Cover (LULC) mapping. However, recent multi-modal methods have shown their superior classification performance over the algorithms that use single data sets. On the other hand, Convolutional Neural Networks (CNNs) are models extensively utilized for the hierarchical extraction of features. Vision transformers (ViTs), through a self-attention mechanism, have recently achieved superior modeling of global contextual information compared to CNNs. However, to harness their image classification strength, ViTs require substantial training datasets. In cases where the available training data is limited, current advanced multi-layer perceptrons (MLPs) can provide viable alternatives to both deep CNNs and ViTs. In this paper, we developed the SGU-MLP, a deep learning algorithm that effectively combines MLPs and spatial gating units (SGUs) for precise Land Use Land Cover (LULC) mapping using multi-modal data from multi-spectral, LiDAR, and hyperspectral data. Results illustrated the superiority of the developed SGU-MLP classification algorithm over several CNN and CNN-ViT-based models, including HybridSN, ResNet, iFormer, EfficientFormer, and CoAtNet. The SGU-MLP classification model consistently outperformed the benchmark CNN and CNN-ViT-based algorithms. The code will be made publicly available at https://github.com/aj1365/SGUMLP**

***Index Terms*—Attention mechanism, image classification, spatial gating unit (SGU), vision transformers.**

I. Introduction

Land use and land cover (LULC) change is one of the most significant indicators of anthropogenic interaction with the natural environment. Massive growth in land use because of forest destruction, urbanization, and soil erosion has altered the global landscape and increased stress on natural ecosystems across the world [1]. Analysis of urban growth, including intense growth in urban areas known as urban sprawl, is essential for understanding its environmental consequences, as well as promoting the adoption of more sustainable forms of urban expansion. Hyperspectral (HS) data have been utilized widely for accurate LULC mapping due to their ability to distinguish subtle spectral differences [2]. However, recent research on the use of multi-modal models, such as multi-modal fusion transformer (MFT) network has proven their superior classification performance compared to the models that utilize only hyperspectral data [3].

It has been shown that due to the complex characteristics of HS data, conventional machine learning models, such as the random forests, struggle to accurately classify HSI [2]. Furthermore, traditional models do not take spatial information. Additionally, hyperspectral imaging often involves a naturally nonlinear interaction between the corresponding ground classes and the acquired spectral information [2]. On the other hand, deep learning models have been used increasingly for HS classification in recent years. In particular, Convolutional Neural Networks (CNNs) are widely used models because of their ability for automatic hierarchical feature extraction. To address the limitation of CNNs in capturing global contextual information, vision transformers (ViTs) have been successfully employed for HSI classification [4]. ViTs use self-attention mechanisms to obtain global contextual information more effectively than CNNs, significantly increasing the accuracy of HS classification [3].

To fully benefit from current CNNs, a significant number of reference data are needed, while ViTs require even larger training datasets to maximize image classification accuracy. On the other hand, where fewer training data are available, current advanced Multi-layer Perceptrons (MLPs) can be used as an alternative to both deep CNNs and ViTs [5]. Employing MLP models requires far fewer reference data compared to CNNs and ViTs due to fewer parameters needing to be trained. Generally, similar to traditional classifiers (e.g., the random forest), MLPs utilize solely spectral information and ignore the spatial interaction between pixels and their surroundings, resulting in lower classification accuracy. Thus, in this paper, we develop and propose the SGU-MLP, a deep learning classifier that employs MLPs and a spatial gating unit (SGU) for accurate LULC modeling utilizing hyperspectral, multi-spectral, and LiDAR data. The SGU concept enables the algorithm to efficiently characterize complex spatial interactions across input data tokens without the use of positional information embedding as utilized in popular ViTs. The SGU-MLP model's final layer employs a structure entirely composed of MLPs, eliminating the requirement for CNNs or ViTs and,

This research was funded by the Institute of Advanced Research in Artificial Intelligence (IARAI). (*Corresponding author: Pedram Ghamisi*)

A. Jamali is with the Department of Geography, Simon Fraser University, 8888 University Dr, Burnaby, BC V5A 1S6, Canada (e-mail: alij@sfu.ca).

S. K. Roy is with the Department of Computer Science and Engineering, Alipurduar Government Engineering and Management College, West Bengal 736206, India (e-mail: swalpa@cse.jgec.ac.in).

D. Hong is with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, 100049 Beijing, China. (e-mail: hongdf@aircas.ac.cn).

P. M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster, U.K. (e-mail:pma@lancaster.ac.uk).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and is also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

(* indicates these two authors contributed equally to the work.)

consequently, minimizing the necessity for extensive training data.

This letter introduces the SGU-MLP in Section II, illustrates the experiments and analyses the results in Section III, and highlights the concluding remarks in Section IV.

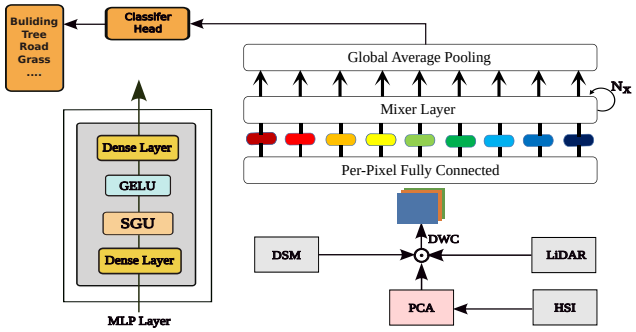

Fig. 1: Graphical representation of spatial gated multi-layer perceptron framework for land use and land cover classification. The MLP-Mixer layer includes two MLPs to extract spatial information. $\odot$ represents channel-wise concatenation.

II. Proposed Classification Framework

As illustrated in Fig. 1, the SGU-MLP, is developed for image classification using a small number of training data. For efficient application of the multi-scale representation in the classification task, we incorporated a computationally light and straightforward depth-wise CNN-based architecture. As presented in Fig. 2, the MLP-Mixer layer of the developed model includes two different types of layers: (i) MLPs utilized across image patches for extraction of spatial information and (ii) MLPs utilized individually to extract per-location features from image inputs. In addition, in each MLP block, the SGU is utilized to enable the developed algorithm to effectively learn intricate spatial relationships among the tokens of the input data.

A. Depth-wise Convolution Block (DWC):

The DWC architecture is light and straightforward and is based on CNNs. With so many variables and the limited available training data, a higher probability of overfitting exists during the training process. Hence, to address the overfitting issue and capture multi-scale feature information, we incorporated three depth-wise convolutions in parallel. These convolutions consist of 20 outputs channels with kernel (k) sizes of 1×1, 3×3, and 5×5, respectively. Feature maps X with a size of $9 \times 9 \times d$ are the input for the DWC block that produces output D_Z, where d is the number of bands.

$$D_Z = \sum_{j=1,3,5} \mathrm{DWConv2D}_{(\mathrm{k}\times\mathrm{k})}(X) \tag{1}$$

The output maps of the three depth-wise CNNs are added and fed to the MLP-Mixer blocks.

B. Spatial gating unit (SGU):

The SGU is designed to extract complex spatial interaction across tokens. Unlike, the current ViT models, the SGU does not necessitate the use of positional embedding. In other words, the positional embedding information is obtained through the use of spatial depth-wise convolutions [6] similar to inverted bottlenecks employed in MobileNetV2 [7]. Considering the dense layer of D (i.e., input feature) in the MLP block , as illustrated in Fig. 1, the SGU uses a linear projection layer that benefits from a contraction operation across the spatial dimension of the cross-tokens interaction as defined by:

$$f_{W,b}(D) = WD + b \tag{2}$$

where $W \in R^{n\times n}$ defines a matrix that has a size equal to the input sequence length, while n and b present the sequence length and biases of the tokens. It should be highlighted that the spatial projection matrix of W is not dependent on the input data, contradicting the self-attention models where $W(D)$ is created dynamically from the D. The SGU can be formulated as:

$$S(D) = D \cdot f_{W,b}(D) \tag{3}$$

where element-wise multiplication is represented by $(\cdot)$. The SGU equation can be improved by dividing D into $D1$ and $D2$ along the channel dimension. Thus, the SGU can be formulated as:

$$S(D) = D1 \cdot f_{W,b}(D2) \tag{4}$$

The output map of the DWC block is flattened and fed to the MLP-Mixer layer. Considering a dense layer of size 256×256, The $D1$ and $D2$ both have sizes of 256×128. The $f_{W,b}(D2)$ has a size of 256×128, where the $S(D)$ has a size of 256×128.

C. Multi-layer Perceptron Mixer Block (MLP-Mixer):

In current advanced deep vision architectures, layers combine features in one or more of the following ways: first, at a given spatial location, second, among various spatial locations, or third, both operations simultaneously, with a kernel of $k \times k$ convolutions (for $k > 1$) and pooling operations (i.e., second operation), incorporated in CNNs. Convolutions with kernel size 1×1 perform only the first operation, whereas convolutions with larger kernels accomplish both the first and second operations. Self-attention layers in ViTs and other attention-based structures include the first and second operations, while models based on MLPs perform only the first operation. The objective of the MLP-Mixer architecture is to distinguish between cross-location (height and width mixing) operations and per-location (channel-mixing) operations, as presented in Fig. 2 [5]. A series of non-overlapping patches of images E from the output feature of the DWC block D_Z are the input to the MLP-Mixer that is projected to a given hidden dimension of C, resulting in two-dimensional table of $\mathbf{M} \in \mathcal{R}^{E\times C}$. The output features of the DWC block are first flattened and then fed to the MLP-Mixer layers. Given the input image of size $H \times W$, and patches of $F \times F$, the number of patches would be $E = \frac{H\times W}{F^2}$, where all resulting patches of images are projected into the same projection matrix. For instance, considering the input image size of 9×9, the reshaped feature has a size of $9 \times 9 = 81$. As we set the dimension of

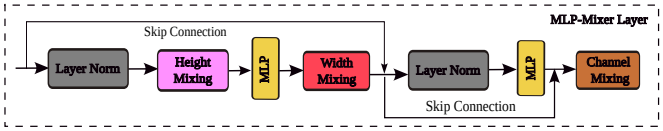

Fig. 2: Graphical representation of MLP-Mixer layer.

the token-mixing MLP to 256, the output feature map has a dimension of 81×256. The MLP-Mixer consists of several layers of identical size (i.e., 4 layers), where each layer has two MLP blocks. The first token-mixing is applied to column of the M table (i.e, it is applied to the transposed input M^T), while the second MLP block (i.e., channel mixing) is applied on the rows of the M table. Two fully connected layers are in each MLP block, and a non-linearity function is applied independently to each row of the input image tensors. As such, each MLP-Mixer can be formulated as:

$$U_{\iota,i} = M_{\iota,i} + W_2\xi(W_1 LN(M)_{\iota,i})), i = 1, ..., C \quad (5)$$

$$Y_{j,\iota} = U_{j,\iota} + W_4\xi(W_3 LN(U)_{\iota,i})), j = 1, \ldots, E \quad (6)$$

Where ξ illustrates the element-wise non-linearity function, while LN presents layer normalization. Notably, the MLP-Mixer has a linear computation complexity, which distinguishes it from vision transformers with quadratic computation complexity and, consequently, exhibits a high level of computational efficiency.

D. Spatial Gating Unit Multi-layer Perceptron (SGU-MLP):

Let us consider three data modalities: X_1, X_2, and X_3. From these datasets, image patches with a size of 9×9 are extracted and then concatenated. It should be noted that a Principal Component Analysis (PCA) algorithm is utilized for dimension reduction of the HS data. After applying the PCA algorithm, the channel numbers for the Berlin, Augsburg, and Houston data benchmarks are 15, 12, and 12, respectively. In this study, we utilized concatenation-based models, which stack the source multimodal remote sensing imagery before passing it through a particular feature extractor or learner to generate the combined features. As depicted in Fig. 1, the concatenated layer is fed into the DWC layer. After passing through the DWC block, the input images of size $9 \times 9 \times B$ result in feature maps of equal size, i.e., $9 \times 9 \times B$, where B represents the number of bands. The resulting feature map is then flattened and passed on to the MLP-Mixer blocks. The MLP-Mixer comprises four blocks with patch sizes of 4 (i.e., input feature), a token-mixing dimension of 256, and a channel-mixing dimension of 256. As discussed, in each MLP block, the SGU is employed to extract complex spatial interactions between the tokens before the activation function (i.e., GELU) is applied. Finally, the last layer of the MLP-Mixer is a dense layer with a softmax activation function. The size of the last layer is determined by the number of existing classes in each study area.

III. Experimental Results

A. Experimental Data

Houston dataset: This dataset was captured over the University of Houston campus and the neighboring urban area. It consists of a co-registered hyperspectral and multispectral dataset containing 144 and 8 bands, respectively, with 349×1905 pixels. More information can be found at [8].

Berlin dataset: This dataset has a spatial resolution of 797×220 pixels and contains 244 spectral bands over Berlin. The Sentinel-1 dual-Pol (VV-VH) single-look complex (SLC) product represents the SAR data. The processed SAR data have a spatial resolution of 1723×476 pixels. The HS data are interpolated through the nearest neighbor algorithm, as for the Houston dataset, to provide the same image size as the SAR data [9].

Augsburg dataset: This scene over the city of Augsburg, Germany includes three distinct datasets: a spaceborne HS dataset and a dual-Pol PolSAR image. All image spatial resolutions were down-scaled to a single 30 m GSD. The scene describes four features from the dual-Pol (VV-VH) SAR image, 180 spectral bands for the HS dataset of 332×485 pixels [10].

B. Classification Results

The classification capability of the developed SGU-MLP was evaluated against several CNN-based and cutting-edge CNN-ViT algorithms, including HybridSN [11], ResNet [12], iFormer [13], EfficientFormer [14], and CoAtNet [15]. In the Augsburg dataset, as shown in Table I, the developed SGU-MLP algorithm demonstrated superior classification performance with an average accuracy of 65.75% compared to ResNet (43.57%), CoAtNet (49.9%), EfficientFormer (52.81%), iFormer (52.96%), and HybridSN (55.76%). The developed SGU-MLP classifier significantly increased the classification accuracy of the CNN-ViT-based algorithms, iFormer, EfficientFormer, and CoAtNet, by about 13, 13, and 16 percentage points in terms of average accuracy, as illustrated in Table I and Fig. 3.

In the Berlin study area, the SGU-MLP classifier with an average accuracy of 65.89% considerably increased the classification accuracy of the other CNN-ViT algorithms iFormer (62.89%), CoAtNet (60.53%), and Efficientformer (60.05%) by approximately 3, 5 and 6 percentage points, respectively, as shown in Table II and Fig. 4. Moreover, as shown in Table III and Fig. 5, with an average accuracy of 87.25%, the SGU-MLP algorithm noticeably surpassed the classification performance of the ResNet (71.42%), iFormer (72.86%), Efficientformer (70.69%), CoAtNet (75.62%), and HybridSN (76.44%), respectively, in the Houston pilot site. The developed SGU-MLP

classification model outperformed the other CNN and CNN-ViT-based algorithms of the HybridSN, CoAtNet, Efficientformer, iFormer, and ResNet by about 11, 14, 15, 15, and 19 percentage points, respectively, in terms of average accuracy, as demonstrated in Table III.

TABLE I: Classification results of Augsburg dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	HybridSN	ResNet	iFormer	Efficientformer	CoAtNet	SGU-MLP
Forest	0.88	0.83	0.91	0.88	0.85	**0.93**
Residential	0.89	0.83	0.89	0.9	0.87	**0.96**
Industrial	0.43	0.15	0.35	0.4	0.22	**0.59**
Low Plants	0.87	0.88	0.88	0.88	**0.98**	0.96
Allotment	0.13	0.1	0.13	0.11	0.09	**0.27**
Commercial	0.04	0.05	0.1	0.11	0.16	**0.29**
Water	0.35	0.19	0.21	0.25	0.19	**0.55**
OA×100	82.28	79.07	82.82	82.72	81.32	**91.13**
AA×100	55.76	43.57	52.96	52.81	49.9	**65.75**
κ×100	74.85	69.34	75.37	75.24	73.12	**87.24**
Training time (min)	6	3	34	7	13	4

TABLE II: Classification results of Berlin dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	HybridSN	ResNet	iFormer	Efficientformer	CoAtNet	SGUMLP
Forest	0.71	0.64	0.69	**0.73**	0.65	0.72
Residential	0.80	0.81	**0.82**	0.81	0.76	0.81
Industrial	**0.49**	0.39	0.35	0.32	0.32	0.39
Low Plants	0.59	0.35	0.72	0.70	0.59	0.70
Soil	0.65	0.72	0.70	0.67	**0.75**	0.72
Allotment	textbf0.44	0.28	0.34	0.29	0.30	**0.44**
Commercial	**0.45**	0.25	0.29	0.24	0.29	0.27
Water	**0.65**	0.53	0.49	0.38	0.28	0.50
OA×100	66.81	63.7	68.6	68.17	63.14	**70.56**
AA×100	62.67	58.23	62.84	60.05	60.53	**65.89**
κ×100	55.84	47.61	55.28	54.32	49.21	**57.85**
Training time (min)	6	3	29	7	13	4

TABLE III: Classification results of Houston dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	HybridSN	ResNet	iFormer	Efficientformer	CoAtNet	SGUMLP
Healthy Grass	0.85	0.88	0.86	0.89	**0.90**	**0.90**
Stressed Grass	0.84	**0.90**	0.87	0.87	0.88	**0.90**
Synthetic Grass	0.84	0.78	0.5	0.58	0.72	**0.97**
Tree	0.87	0.89	0.92	0.91	**0.93**	0.92
Soil	0.96	0.94	0.93	0.95	0.85	**1**
Water	**0.73**	0.71	0.29	0.39	0.25	0.33
Residential	0.69	0.72	0.68	0.6	**0.79**	**0.79**
Commercial	0.69	0.39	0.68	0.56	0.6	**0.81**
Road	0.7	0.57	0.75	0.77	0.82	**0.85**
Highway	0.58	0.52	0.45	0.54	0.54	**0.83**
Railway	0.7	0.54	0.67	0.57	0.67	**0.82**
Parking Lot1	0.74	0.42	0.48	0.71	0.55	**0.97**
Parking Lot2	**0.94**	0.61	0.72	0.78	0.58	0.86
Tennis Court	0.84	0.77	0.74	0.73	0.56	**1**
Running Track	0.64	0.82	0.83	0.61	0.92	**0.95**
OA×100	75.62	68.16	71.03	71.66	72.67	**85.34**
AA×100	76.44	71.42	72.86	70.69	75.62	**87.25**
κ×100	73.59	65.49	68.71	69.25	70.56	**84.17**
Training time (min)	4	2	20	5	10	3

C. Ablation study

An ablation study was performed to better understand the contribution and significance of different parts of the developed SGU-MLP classification algorithm. As seen in Table IV, the inclusion of the DWC block and SGU block increased the classification accuracy of the MLP-Mixer model by approximately 2 and 3 percentage points, respectively, in terms of average accuracy for the Augsburg dataset. The highest classification accuracy was achieved by the inclusion of both the DWC and SGU blocks with an average accuracy of 65.75%, increasing the classification accuracy of the MLP-Mixer algorithm by about 5 percentage points.

In the Berlin dataset, as illustrated in Table V, the inclusion of the SGU block and DWC block increased the classification accuracy of the MLP-Mixer algorithm by about 1 and 2 percentage points, respectively, in terms of Kappa index. By incorporating both the DWC and SGU blocks, the highest classification was attained with a Kappa index of 57.85%. This increased the accuracy of the MLP-Mixer classifier by approximately 3 percentage points.

As demonstrated in Table VI, the inclusion of the DWC block and SGU block increased the accuracy of the MLP-Mixer algorithm by approximately 2 and 1 percentage points, respectively, in terms of average accuracy for the Houston dataset. By the inclusion of both the DWC and SGU blocks, the MLP-Mixer's classification accuracy was increased by approximately 7 percentage points to 87.25%.

TABLE IV: Classification results of Augsburg dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	MLP	SGU + MLP	DWC + MLP	SGUMLP
Forest	0.87	0.92	0.91	**0.93**
Residential	0.91	0.93	0.92	**0.96**
Industrial	0.36	0.52	0.55	**0.59**
Low Plants	0.95	0.96	0.95	**0.98**
Allotment	0.20	0.20	0.21	**0.27**
Commercial	0.15	0.20	0.18	**0.29**
Water	0.55	**0.57**	0.54	0.55
OA×100	87.64 ± (0.61)	88.90 ± (0.45)	89.48 ± (0.52)	**91.13 ± (0.30)**
AA×100	60.96 ± (1.16)	62.36 ± (1.59)	63.59 ± (1.25)	**65.75 ± (0.42)**
κ×100	82.12 ± (0.90)	84.01 ± (0.66)	84.83 ± (0.78)	**87.24 ± (0.41)**

TABLE V: Classification results of Berlin dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	MLP	SGU + MLP	DWC + MLP	SGUMLP
Forest	**0.74**	0.72	0.72	0.72
Residential	0.81	0.80	**0.82**	0.81
Industrial	**0.40**	0.39	**0.40**	0.39
Low Plants	0.68	0.66	0.68	**0.70**
Soil	0.71	0.67	0.67	**0.72**
Allotment	**0.44**	0.43	**0.44**	**0.44**
Commercial	0.25	0.26	0.24	**0.27**
Water	**0.50**	0.44	0.45	**0.50**
OA×100	68.43 ± (0.83)	69.12 ± (0.65)	70.03 ± (0.17)	**70.56 ± (0.58)**
AA×100	65.16 ± (0.52)	65.20 ± (0.50)	64.70 ± (1.04)	**65.89 ± (0.26)**
κ×100	55.25 ± (0.93)	56.06 ± (0.75)	56.95 ± (0.25)	**57.85 ± (0.58)**

TABLE VI: Classification results of Houston dataset in terms of F-1 score where κ = Kappa index, OA = Overall Accuracy, AA = Average Accuracy, respectively.

Class	MLP	SGU + MLP	DWC + MLP	SGUMLP
Healthy Grass	0.89	**0.90**	**0.90**	**0.90**
Stressed Grass	0.90	**0.91**	0.90	0.90
Synthetic Grass	0.43	**0.97**	0.98	**0.97**
Tree	0.89	**0.94**	**0.94**	0.92
Soil	0.96	**1**	**1**	**1**
Water	**0.46**	0.17	0.22	0.33
Residential	0.79	0.78	**0.80**	0.79
Commercial	0.66	0.69	**0.81**	**0.81**
Road	0.82	0.84	0.81	**0.85**
Highway	0.62	0.59	0.62	**0.83**
Railway	0.73	**0.83**	0.80	0.82
Parking Lot1	0.75	0.93	0.94	**0.97**
Parking Lot2	0.79	0.69	**0.88**	0.86
Tennis Court	0.88	**1**	**1**	**1**
Running Track	0.82	**0.96**	0.95	0.95
OA×100	78.27 ± (1.53)	82.45 ± (0.92)	84.22 ± (0.81)	**85.34 ± (0.91)**
AA×100	80.53 ± (1.46)	85.03 ± (0.68)	86.38 ± (0.73)	**87.25 ± (0.68)**
κ×100	76.53 ± (1.65)	81.08 ± (0.98)	82.99 ± (0.88)	**84.17 ± (0.96)**

D. Computation cost

As illustrated in Table I, the proposed model required the least computation cost in terms of training time (4 min) in the Augsburg data benchmark compared to other ViT-based models of iFormer (34 min), CoAtNet (13 min), and Efficient Former (7 min). Moreover, in the Berlin dataset, the SGU-MLP algorithm with a required training time of 4 min demonstrated better computation efficiency over the other ViTs of iFormer (29 min), CoAtNet (13 min), and Efficient Former (7 min) (see Table II). In addition, as seen in Table III, in the benchmark of the Houston data, the computational complexity of the SGU-MLP model was much less in terms of training time (3 min) compared to the other implemented ViTs, including iFormer (20 min), CoAtNet (10 min), and Efficient Former (5 min). It is worth mentioning that an RTX 2070 MAX-Q GPU and Intel core-i7 CPU were utilized. The optimizer, loss function, batch size, and learning rate were set to Adam, Sparse Categorical Cross Entropy, 100, and 0.001, respectively, in all of the implemented models.

IV. Conclusion

In this study, we developed the SGU-MLP algorithm based on advanced MLP models and a spatial gating unit for land use and land cover mapping which demonstrated superior classification accuracy compared to several CNN and CNN-ViT-based models. The obtained results illustrated that the utilized MLP-Mixer architecture could obtain greater cross-location

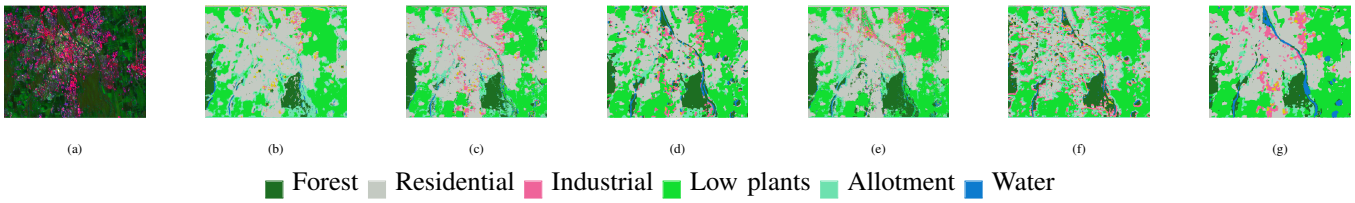

Fig. 3: Classification Maps over the Augsburg dataset using a) Study image, b) CoAtNet, c) EfficientFormer, d) HybridSN, e) iFormer, f) ResNet, and g) the SGU-MLP.

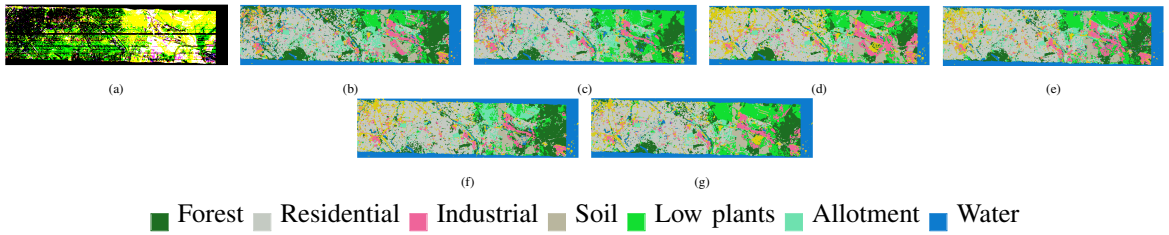

Fig. 4: Classification Maps over the Berlin dataset using a) Study image, b) CoAtNet, c) EfficientFormer, d) HybridSN, e) iFormer, f) ResNet, and g) the SGU-MLP.

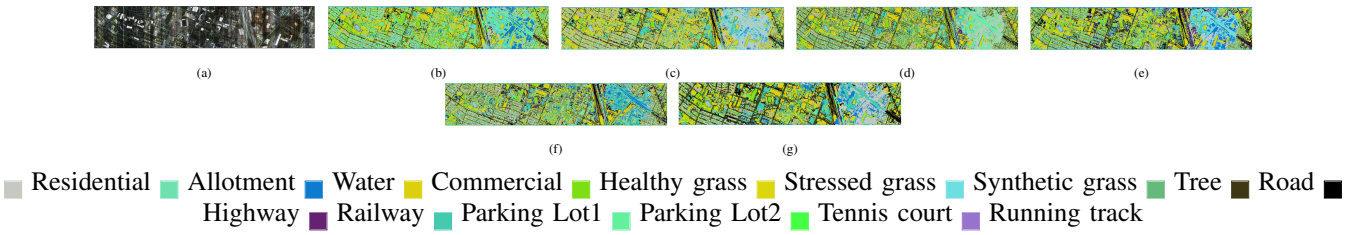

Fig. 5: Classification Maps over the Houston dataset using a) Study image, b) CoAtNet, c) EfficientFormer, d) HybridSN, e) iFormer, f) ResNet, and g) the SGU-MLP.

(height and width) and per-location (channel) information compared to the current advanced ViTs. Additionally, the SGU increased the classification accuracy by efficiently acquiring complex spatial interaction across image tokens. Moreover, the SGU-MLP algorithm was demonstrated to be much more computationally efficient in terms of training time compared to other implemented ViT-based models of iFormer, Efficient Former, and the state-of-the-art ViT model of CoAtNet.

REFERENCES

[1] J. Yang, A. Guo, Y. Li, Y. Zhang, and X. Li, "Simulation of landscape spatial layout evolution in rural-urban fringe areas: a case study of ganjingzi district," *GIScience & Remote Sensing*, vol. 56, no. 3, pp. 388–405, 2019. [Online]. Available: https://doi.org/10.1080/15481603.2018.1533680

[2] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[3] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[4] H. Yan, E. Zhang, J. Wang, C. Leng, A. Basu, and J. Peng, "Hybrid convvit network for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[5] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24 261–24 272.

[6] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 9204–9215. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/4cc05b35c2f937c5bd9e7d41d3686fff-Paper.pdf

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.

[9] A. Okujeni, S. van der Linden, and P. Hostert, "Berlin-urban-gradient dataset 2009 - an enmap preparatory flight campaign," 2016.

[10] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621001362

[11] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11 106–11 115, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17325

[14] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," 2022.

[15] Z. Dai, H. Liu, Q. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems 34*, 2021.