ESIF: A Novel Frequency and Texture Aware Network for Enhanced Remote Sensing Scene Classification

Russo Ashraf, *Member, IEEE*, Kang-Hyun Jo, *Member, IEEE*

***Abstract*—**

***Index Terms*— Remote Sensing, Scene Classification, Texture Analysis, Convolutional Neural Network (CNN), Self-Attention**

I. Introduction

EARTH observation via remote sensing techniques constitutes a research domain that encompasses the measurement of signals originating from diverse physical phenomena, acquired by instruments deployed on both spaceborne and airborne platforms. This technology offers versatile utilization prospects, serving either for the precise quantification and estimation of geo-bio-physical parameters or for material identification through the analysis of acquired signals. These objectives can be realized due to the fundamental behavior of materials within a scene, where they interact with electromagnetic radiation by reflecting, absorbing, and emitting radiation contingent on their molecular composition and geometric characteristics. Remote sensing strategically leverages these fundamental principles, enabling the acquisition of information pertaining to a scene or specific object situated at varying proximities from the sensor, spanning short, medium, or long distances [1]–[3]. Among the multitude of data products derivable from remote sensing imagery, classification maps represent a notably consequential category [4], [5]. The problem of remote sensing image classification stands as a formidable challenge, given the imperative role of land-cover and land-use maps in multitemporal investigations and their invaluable contribution to diverse domains, including climate change modeling, oceanic current analysis, arctic research, and post-catastrophe response efforts.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Department of Commerce under Grant BS123456.”

The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

In the realm of machine learning (ML), historical approaches to addressing these tasks have predominantly adhered to two paradigms: pixel-level [6] and object-level classification [7]. Pixel-level classification pertains to the assignment of a semantic label to each individual pixel within an image. While effective for certain applications, these approaches often exhibit limitations when applied to high-resolution imagery. More critically, they may struggle to capture higher-level spatial patterns spanning multiple pixels. In contrast, object-level classification methods center their analysis on discernible and meaningful objects within an image, considering them as collections of pixels rather than isolated entities. This paradigm generally offers enhanced scalability and performance; however, it can encounter difficulties when faced with images containing diverse and less distinguishable objects, a common scenario in high-resolution remote sensing data. Approaches based on both pixel-level and object-level paradigms have demonstrated commendable performance and remain the subject of active research, often manifesting as instances of image segmentation and object detection tasks. More recently, a novel paradigm of scene-level classification [8], [9] has emerged, showcasing notable performance enhancements. This paradigm emphasizes the acquisition of semantically meaningful representations for intricate patterns within an image by harnessing the capabilities of deep learning.

Remote sensing images exhibit distinctive characteristics compared to conventional images, primarily attributable to their unique acquisition mode. These images typically encompass extensive geographical areas, offering an overhead perspective that incorporates a diverse array of objects and features. As illustrated in Fig. 1, it becomes evident that not all spatial information within these images holds equal significance. Consequently, the task of discerning and prioritizing critical image components while disregarding less informative ones assumes paramount importance. Regrettably, the prevailing approach in many prior studies has been to construct a global representation of the entire image, affording equal weight to all regions [10], [11], [12]. This approach neglects the detrimental impact of redundant and inconsequential areas, undermining the potential to extract meaningful insights.

In general, remote sensing scenes can be categorized into specific thematic classes, such as segments of a forest, parking lots, agricultural fields, and more. For such classification tasks, supervised learning techniques are commonly employed [13]. This approach involves the initial representation of a scene im-

age as a feature vector, which is subsequently utilized for both training and testing a learning machine, as depicted in Fig. 1(c). Within the context of feature-based image representation, beyond the pivotal steps of feature extraction and feature coding, the process of feature selection assumes considerable significance. The application of an effective feature-selection method can yield substantial improvements in the ultimate performance outcomes [14], [15]. Consequently, the research and development of proficient feature-selection methodologies hold significant import.

II. Related Works

A. Earth Observation

Satellite-based earth observation has evolved into an indispensable instrument for comprehending and surveilling global environmental transformations, encompassing phenomena such as deforestation, urban expansion, and climate fluctuations [16]. Within this context, satellite image classification assumes a pivotal role, exerting a profound impact across a spectrum of applications, notably land use and land cover mapping, agricultural surveillance, disaster mitigation, and urban planning initiatives [17], [18]. Furthermore, satellite image classification finds pertinence in the domain of disaster management, where it expedites damage assessment and bolsters disaster response endeavors [19]. To augment classification precision, amalgamating data from diverse sources, including satellite imagery, climatic data, and ground-level observations, proves instrumental [20]. In a comprehensive study, the authors of [21] delve into an extensive examination involving 22 datasets, exploring numerous amalgamations of deep learning models while conducting a rigorous comparative analysis of their efficacy.

B. Efficient CNNs for Classification

In recent years, there has been a notable surge in research interest surrounding the efficiency of convolutional neural networks (CNNs). A pivotal milestone in this pursuit was the introduction of Depthwise-Separable Convolution by Howard et al. [22], which gave birth to the Xception architecture. This groundbreaking approach significantly reduces the parameters and computational operations (FLOPs) associated with conventional convolutions while retaining robust feature-capturing capabilities. Subsequently, MobileNets [23] built upon this concept, ushering in a family of efficient CNN architectures meticulously designed for expeditious performance on mobile and embedded devices. Another noteworthy contribution in this domain was made by Zhang et al. [24] with the inception of ShuffleNet. This innovative CNN architecture harnesses channel shuffling techniques and pointwise group convolutions to achieve commendable accuracy while maintaining a low computational burden. EfficientNets, introduced by Tan and Le [25], represent yet another significant advancement. These CNN architectures leverage a novel compound scaling method to attain state-of-the-art performance metrics, all the while substantially reducing the number of parameters and computational expenses. SqueezeNet, pioneered by Iandola et al. [26], offers a distinct approach. This CNN architecture employs a combination of 1x1 and 3x3 convolutions to effectively curtail the parameter count while upholding high precision in classification tasks. Furthermore, Wu et al. [27] brought forth ProxylessNAS, a groundbreaking neural architecture search method. This approach enables the direct optimization of CNN architectures tailored to specific hardware and tasks, yielding highly efficient and accurate models. Additionally, the research community witnessed innovations such as RTM-Det [28], which introduced a modification of the renowned darknet-53 architecture. This adaptation incorporates large-kernel depthwise-separable convolutions, further contributing to the realm of efficient CNN architectures.

C. Texture Analysis for Scene Classification Task

While existing CNN-based methodologies have exhibited promise in the realm of Scene Classification tasks, they primarily rely on pure RGB images and may fall short in capturing intricate high-level texture attributes. To address this limitation and augment the texture characteristics inherent in facial expressions, classical texture features have been harnessed as supplementary inputs within a parallel neural network framework [29]. For instance, the Local Binary Pattern (LBP) was amalgamated with features extracted from CNN, employing an attentional selective fusion strategy [30]. Additionally, Liu et al. [31] introduced the application of the gray-level co-occurrence matrix to preprocess facial images, subsequently extracting deep texture features. In light of these advancements, our study centers on the development of a texture-aware feature enrichment module. This module is adept at leveraging a spectrum of texture extraction techniques, thereby providing a wealth of texture information, particularly beneficial for the characterization of challenging land cover classes.

III. Methodology

A. Efficient SpectroFormer Block (ESFB)

B. Texture Feature Alignment Block (TFAB)

Initially, we introduce the Local Binary Pattern (LBP) operator as a means to characterize the local texture attributes of an image. LBP offers notable advantages, including rotational and grayscale invariance. Specifically, we employ a 3×3 sliding window to traverse the entirety of the facial image, extracting texture features. Within this window, the central pixel serves as the threshold against which the other eight neighboring pixels are compared. In this context, a value of ”1” signifies a pixel intensity higher than the threshold, while ”0” designates a lower intensity. Subsequently, an eight-bit binary number is derived by encoding these comparisons in a clockwise manner, commencing from the top-left corner. To encapsulate the texture information within this window, we further convert the binary representation into a decimal pixel value.

In addition to LBP, we incorporate characteristics derived from the Gray-Level Co-occurrence Matrix (GLCM) to augment our texture analysis. Specifically, we utilize both the contrast ratio and relevance metrics as supplementary texture

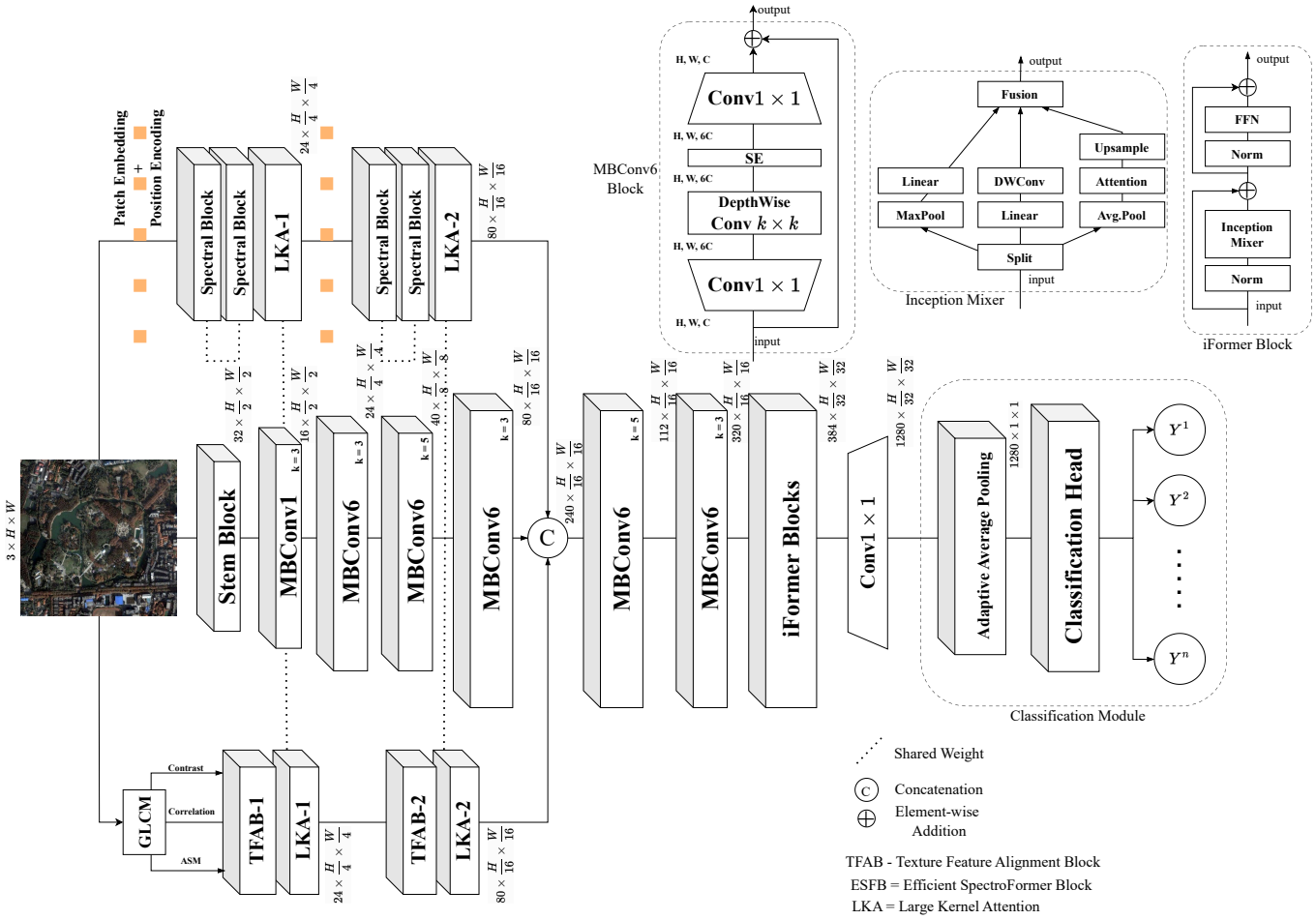

Fig. 1. Overall Architecture of the proposed Network comprised of the following modules-TAFEM,EAAM,FEM and CM.

descriptors, leveraging inputs from GLCM. In this configuration, we employ sub-windows of size 3 × 3 and set the number of gray levels to eight. Angular Second Momentum (ASM) emerges as a valuable metric for discerning the depth of textures and patterns. A higher ASM value signifies the presence of more pronounced textures and deeper patterns, while a lower value corresponds to a blurred visual representation with shallower textures. The calculation of ASM is outlined as follows:

$$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2 \tag{1}$$

where N is the size of GLCM and $P(i,j)$ is the probability density of the corresponding pixel. Relevance is the similarity degree of GLCM elements in directions of line and row, which denotes the relevant degree of some gray levels in facial images. The relevance value will be larger with equal matrix element values and can be defined by:

$$CORR = \sum_{i,j=0}^{N-1} P_{i,j} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \tag{2}$$

where μ_i,μ_j and σ_i,σ_j refer to mean and variance of $P_x(i)$ and $P_y(j)$ respectively. Finally, three texture feature maps: $x_{LBP} \in \mathbb{R}^{HxW}$, $x_{ASM} \in \mathbb{R}^{HxW}$, $x_{CORR} \in \mathbb{R}^{HxW}$ are obtained with above equations. The final output of the TAFEM is produced by:

$$TAFEM(i_{rgb}) = MBConv4(CAT(x_{LBP}, x_{CORR}, x_{ASM}) \tag{3}$$

C. Inception Transformer Block (iFB)

We propose an Inception mixer that combines the powerful capability of Convolutional Neural Networks (CNNs) for extracting high-frequency representations with Transformers. The detailed architecture of the mixer is depicted in Figure 3. We use the name "Inception" since the token mixer is highly inspired by the Inception module with multiple branches. Instead of directly feeding image tokens into the Multi-Head Self-Attention (MSA) mixer, the Inception mixer first splits the input feature along the channel dimension and then respectively feeds the split components into a high-frequency mixer and a low-frequency mixer. Here, the high-frequency mixer consists of a max-pooling operation and a parallel convolution operation, while the low-frequency mixer is implemented by a self-attention mechanism.

$$Y_{h1} = FC(MaxPool(X_{h1}) \tag{4}$$

$$Y_{h2} = DwConv(FC(X_{h2}) \tag{5}$$

$$Y_l = Upsample(MSA(AvePoool(X_l))) \quad (6)$$

$$Y_c = Concat(Y_l, Y_{h1}, Y_{h2}) \quad (7)$$

$$ITM(Y) = FC(Y_c + DwConv(Y_c) \quad (8)$$

$$X = X + ITM(LN(X)) \quad (9)$$

$$H = X + FFN(LN(X)) \quad (10)$$

Technically, given the input feature map $X \in R^{N \times C}$, we factorize X into $X_h \in R^{N \times C_h}$ and $X_l \in R^{N \times C_l}$ along the channel dimension, where $C_h + C_l = C$. Then, X_h and X_l are assigned to the high-frequency mixer and the low-frequency mixer, respectively. The high-frequency mixer is designed to learn the high-frequency components of the input feature map. Considering the sharp sensitiveness of the maximum filter and the detail perception of convolution operation, we propose a parallel structure to learn the high-frequency components. We divide the input X_h into $X_{h1} \in R^{N \times C_h/2}$ and $X_{h2} \in R^{N \times C_h/2}$ along the channel. As shown in Figure 3, X_{h1} is embedded with a max-pooling and a linear layer , and X_{h2} is fed into a linear and a depthwise convolution layer. The outputs of the high-frequency mixers are denoted by Y_{h1} and Y_{h2}. Finally, the outputs of the low- and high-frequency mixers are concatenated along the channel dimension. The upsample operation in Eq. (7) selects the value of the nearest point for each position to be interpolated regardless of any other points, which results in excessive smoothness between adjacent tokens. We design a fusion module to elegantly overcome this issue, i.e., a depthwise convolution exchanging information between patches, while keeping a cross-channel linear layer that works per location like in previous Transformers. Like the vanilla Transformer, our iFormer is equipped with a feed-forward network (FFN), and differently it also incorporates the above Inception token mixer (ITM); LayerNorm (LN) is applied before ITM and FFN. Low-frequency mixer. We use the vanilla multi-head self-attention to communicate information among all tokens for the low-frequency mixer. Despite the strong capability of the attention for learning global representation, the large resolution of feature maps would bring large computation cost in lower layers. We therefore simply utilize an average pooling layer to reduce the spatial scale of X_l before the attention operation and an upsample layer to recover the original spatial dimension after the attention. This design largely reduces the computational overhead and makes the attention operation focus on embedding global information. Here, Y_l is the output of low-frequency mixer. Note that the kernel size and stride for the pooling and upsample layers are set to 2 only at the first two stages.

D. MBConv Based ESIF Baseline

E. ESIF: Efficient Spectral Inception Former

IV. Experiments

A. Implementation and Dataset Details

We evaluate our proposed model in two remote sensing scene classification datasets- WHU-RS19 and Optimal-31. WHU-RS19 contains 19 classes of satellite images of 600x600 dimension, each class containing at least 50 images and in a total of 1005 images. Optimal-31 is also a scene classification dataset, but it's more difficult as it contains 31 classes and an image dimension of 256x256. Each class has at least 60 images and the total number of images is 1860. Evaluation metrics for WHU-RS19 are accuracy, precision, and F1 score, and for Optimal-31 accuracy, F1 score and mean IoU is used. All the ablation experiments are done in the WHU-RS19 dataset. We utilize the AITLAS toolbox for Earth Observation from [1] to train and evaluate our models. Training Split is 60% for training, 20% for validation, and 20% for testing, for both datasets. All models are trained on one NVIDIA Tesla V-100 GPU with 32 GB of memory. The batch size was set to 64 for training. Rectified Adam or RAdam [27] is used as the optimizer. We use learning rate .0001 for WHU-RS19 and .001 for Optimal-31, learning is reduced by factor of 0.1 when validation loss stops improving. Each model is trained for 300 epochs, as we train models from scratch higher iterations were necessary. We use input size of 224x224 by default as "1x" in Table 1 and 2. Inputs are first resized to 256x256 and then center-cropped to 224x224, horizontal and vertical flips are used as data augmentations.

B. Evaluation on Optimal-31

Table.2 shows the evaluation details on the Optimal-31 dataset. In a similar fashion as the previous dataset, DenseNet161 the baseline performs better than ResNet152 and ResNet50. Here, our EITF Network already outperforms the baseline at 1x input by increasing the accuracy by almost 1% while having similar F1, mIoU scores, and training time. At 1.14x input accuracy improves by around 2% and at 2x accuracy is improved by 4.5%, F1 and mIoU also improve by a similar amount. One disadvantage of the larger input size is the training time increase, which is a much bigger jump in this case than in the previous dataset.

C. Evaluation on WHURS-19

We use the baseline DenseNet161, ResNet152, ResNet50[28], and the proposed EITF in three input settings to evaluate the WHU-RS19 dataset. The experimental results are shown in Table. 1, DenseNet161 contains 26.51m parameters and 7.82 GFlops at input size 224x224. The baseline has much better accuracy than ResNet152 and ResNet50, while ResNet152 has much higher parameters and Flops, it suffers from overfitting, ResNet50 the smaller variant performs better. At 1x input, the proposed EITF has 20% fewer parameters and 79% fewer GFlops, but it performs very close to the baseline only decreasing the accuracy by 1.7%. The training time is also lower than baseline. At 1.5x input size while still having 53% fewer GFlops our model can already outperform the baseline. And, at 2x input, it improves the accuracy by more than 4%

D. Evaluation on AID

E. Evaluation on UC-MERCED

F. Evaluation on RSSCN

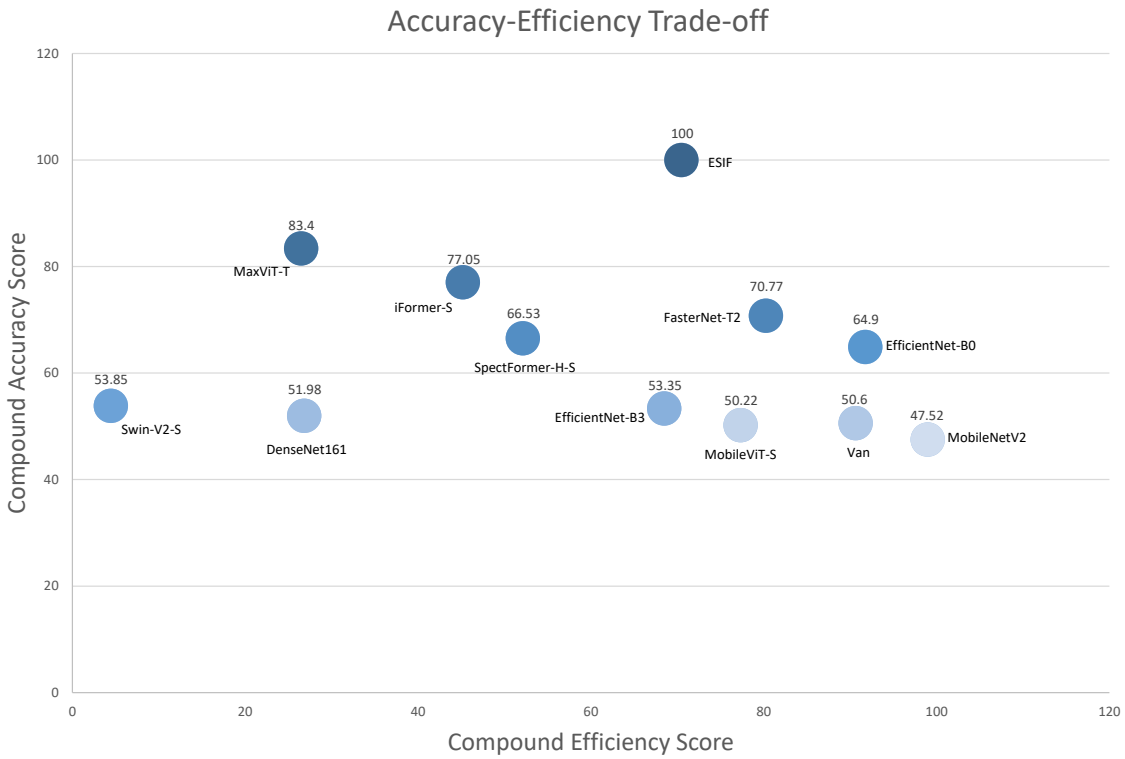

Fig. 2. Accuracy and Efficiency Trade-off for each model.

Model Name	Model Composition	Params. (M)	FLOPs (G)	BA	AA	Model Size (MB)	Memory Access (GB)	Training Time(h)	Inf. Speed (FPS)	AETS
MobileNetV2 (2018)	CNN	2.2	0.3	79.83	77.68	**18.4**	1.47	**0.46**	**240**	73.25
EfficientNet-B0 (2019)	CNN	4	0.4	80.1	78.75	32.7	1.57	0.65	226	78.33
Van (2022)	CNN	4.1	0.9	73.11	72.75	31.1	1.43	0.87	208	70.62
EfficientNetB3 (2019)	CNN	10.7	1	79.56	78.84	86.6	1.77	1.56	186	60.92
MobileViT-S (2022)	Hybrid	5	1.8	68.81	68.63	40.4	1.71	1.06	192	63.77
FasterNet-T2 (2023)	CNN	13.7	1.9	76.07	75.89	110.1	**1.46**	**0.46**	200	75.51
SpectFormer-H-S (2023)	Transformer	20.2	3.9	80.37	78.84	171	1.73	1.44	163	59.33
iFormer-S (2022)	Hybrid	18.9	4.5	76.88	76.88	156.2	1.74	2.16	145	61.12
MaxViT-T (2022)	Transformer	30.3	5.4	77.15	77.15	244.4	2.05	1.6	129	54.94
SwinV2-S (2021)	Transformer	33.2	5.8	78.22	76.07	393.1	2.08	3.21	110	29.15
DenseNet161 (2017)	CNN	26.5	7.8	80.91	80.64	213.7	1.75	2.45	127	39.4
ESIF(Ours)	Hybrid	9	1.1	**85.21**	**84.67**	75.2	1.61	1.23	144	**85.24**

Model Name	Params. (M)	FLOPs (G)	UC-Merced		RSSCN7		SIRI-WHU		WHU-RS19		AID
			BA	AA	BA	AA	BA	AA	BA	AA	BA
MobileNetV2 (2018)	2.2	0.3	92.85	92.13	90.17	88.32	91.45	90.27	92.53	87.55	
EfficientNet-B0 (2019)	4	0.4	95	94.04	92.67	90.55	93.33	92.91	86.06	84.57	92.4
Van (2022)	4.1	0.9	91.9	91.34	89.64	89.05	93.12	92.84	88.55	86.89	
EfficientNet-B3 (2019)	10.7	1	92.38	88.72	93.57	91.72	93.54	92.29	77.11	76.11	
MobileViT-S (2022)	5	1.8	90.47	90.39	90.71	90.65	92.29	92.29	87.56	87.56	
FasterNet-T2 (2023)	13.7	1.9	93.57	92.77	91.25	91.13	93.54	93.19	92.03	92.03	90
SpectFormer-H-S (2023)	20.2	3.9	92.61	92.29	90.71	90.23	93.54	93.05	90.04	89.71	
iFormer-S (2022)	18.9	4.5	92.61	92.61	92.14	92.14	93.95	93.95	90.54	90.20	
MaxViT-T (2022)	30.3	5.4	93.33	92.77	93.21	92.97	94.37	94.16	91.04	91.04	
SwinV2-S (2021)	33.2	5.8	82.61	82.16	91.6	91.48	92.91	92.91	88.05	88.05	90.1
DenseNet161 (2017)	26.5	7.8	95.47	94.75	86.7	85.86	92.5	92.08	93.53	92.7	
ESIF(Ours)	9	1.1	**95.71**	**95.15**	**94.1**	**93.62**	**95**	**94.58**	**94.02**	**93.02**	**93.4**

TFAB	iFB	ESFB	Params (M)	FLOPs (G)	Mem. Access	BA	AA
x	x	x	1.64	0.41		81.45	80.82
Tick	x	x	2	0.64		82.79(+1.34)	82.32(+1.5)
Tick	Tick	x	8.48	0.89		83.33(+0.54)	82.79(+0.47)
			8.57	0.94			
	Tick	MHSA				85.21	84.94
Tick +LKA	Tick	LKA	9	1.1		85.21	84.67

V. Conclusion

Appendixes, if needed, appear before the acknowledgment.

Acknowledgment

References

[1] S. Liang, *Quantitative remote sensing of land surfaces*. John Wiley & Sons, 2005.

[2] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[3] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal processing magazine*, vol. 19, no. 1, pp. 12–16, 2002.

[4] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[5] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE signal processing magazine*, vol. 31, no. 1, pp. 45–54, 2013.

[6] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, "A review of remote sensing image classification techniques: The role of spatio-contextual information," *European Journal of Remote Sensing*, vol. 47, no. 1, pp. 389–411, 2014.

[7] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[8] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[9] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[10] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2013.

[11] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[12] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2889–2901, 2017.

[13] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, 2007.

[14] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 5, pp. 3023–3034, 2013.

[15] J. A. Piedra-Fernández, M. Cantón-Garbín, and J. Z. Wang, "Feature selection in avhrr ocean satellite images by means of filter methods," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 12, pp. 4193–4203, 2010.

[16] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland *et al.*, "High-resolution global maps of 21st-century forest cover change," *science*, vol. 342, no. 6160, pp. 850–853, 2013.

[17] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[18] S. Fei, M. A. Hassan, Y. Xiao, X. Su, Z. Chen, Q. Cheng, F. Duan, R. Chen, and Y. Ma, "Uav-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat," *Precision Agriculture*, vol. 24, no. 1, pp. 187–212, 2023.

[19] S. Dotel, A. Shrestha, A. Bhusal, R. Pathak, A. Shakya, and S. P. Panday, "Disaster assessment from satellite imagery by analysing topographical features using deep learning," in *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, 2020, pp. 86–92.

[20] X. Liu, L. Jiao, L. Li, X. Tang, and Y. Guo, "Deep multi-level fusion network for multi-source image pixel-wise classification," *Knowledge-Based Systems*, vol. 221, p. 106921, 2021.

[21] I. Dimitrovski, I. Kitanovski, D. Kocev, and N. Simidjievski, "Current trends in deep learning for earth observation: An open-source benchmark arena for image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 18–35, 2023.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[25] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[26] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[27] H. Cai, L. Zhu, and S. Han, "Proxylessnas: Direct neural architecture search on target task and hardware," *arXiv preprint arXiv:1812.00332*, 2018.

[28] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "Rtmdet: An empirical study of designing real-time object detectors," *arXiv preprint arXiv:2212.07784*, 2022.

[29] Y. Li, W. Cui, M. Luo, K. Li, and L. Wang, "Epileptic seizure detection based on time-frequency images of eeg signals using gaussian mixture model and gray level co-occurrence matrix features," *International journal of neural systems*, vol. 28, no. 07, p. 1850003, 2018.

[30] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2021.

[31] Z. Xi, Y. Niu, J. Chen, X. Kan, and H. Liu, "Facial expression recognition of industrial internet of things by parallel neural networks combining texture features," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2784–2793, 2020.