A Comparative Study of Swin-Based Enhanced Remote Sensing Image Classifications

BERFİN KURTOĞLU*

Department of Computer Engineering, Harran University, Şanlıurfa, Turkey kurtogluberfin@gmail.com

SERDAR ÇİFTÇİ

Department of Computer Engineering, Harran University, Şanlıurfa, Turkey serdarciftci@harran.edu.tr

In image classification methods, the quality of the input image plays an important role in improving classification performance. However, sometimes the low resolution and sharpness of remote sensing images can cause various problems in image analysis. Therefore, improving and correcting the quality of remote sensing images is of great importance for the classification of remote sensing images. In this study, five man-made and five natural field images were selected from the RSI-Cb remote sensing dataset. The corresponding images were super-resolutioned using the Swin-based HST, Swin2SR and SwinIR transformers. The classifications were performed using the pre-trained architectures DenseNet121, Xception and EfficientV2_B3 and their performance was compared. The results of the experiments show that the classification accuracy was improved by using super-resolution methods rather than the absence of super-resolution methods. It was found that better results were achieved especially for images of natural areas. The best classification performance was achieved when the images were super-resolutioned using the HST algorithm and classification was performed using the Xception architecture, as the classification accuracy increased from 99.18% to 99.59%.

CCS CONCEPTS • Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Object recognition;

Additional Keywords and Phrases: Image classification, Remote Sensing, Swin Based Transformer

1 INTRODUCTION

Remote sensing is the process of obtaining information about a desired location with technological devices that we place at a selected location from a certain distance and analyzing, displaying and monitoring it in spatial, spectral, radiometric and temporal resolution with measurements from any distance without physical contact [1]. Remote sensing is used in the fields of cartography, hydrology, geology, forestry, agriculture, defense, security and space. There are platforms with data sets such as Sentinel, Landsat, Maxar, Planet, UC Merced, EuroSAT, PatternNet, SpaceNet, and Google Earth Engine. Improvements have been made in image processing and data mining techniques to solve both the problem of providing big data and analyzing the data [2], and the SatlasPretrain [3] dataset is one of the big data sets that have been used.

Image classification is an important part of computer vision, i.e. the recognition of identical data. It is widely used in areas such as healthcare, security, operational efficiency, autonomous systems, agriculture, industry and engineering. The most commonly used methods for classification in machine learning are algorithms such as support vector machine [4], random forest [5], logistic regression [6], k-nearest neighbor [7]. Bansal et al. provided results on the advantages and disadvantages of machine learning algorithms in their study [8]. Furthermore, Ouchra et al. [9] showed in a comparison of machine learning methods for the classification of remotely sensed satellite images that the classification performance is improved depending on which machine learning method is used in the datasets.

With the introduction of deep learning methods in the field of classification, studies have shown that they outperform machine learning techniques [10]. Deep learning is more powerful than machine learning because it uses more parameters and more layers to increase the performance of the representation. Powerful computers are required due to high computing costs and the difficulty of storing high quality data. Transfer learning promises to overcome this problem in computer vision. A model trained for one purpose can serve as a reference for another problem [11]. Transfer learning uses the weights and features of previously trained models and is becoming increasingly popular due to its time and cost savings. In this study, we used DenseNet121 [12], Xception [13] and EfficientV2_B3 [14] architectures as transfer learning for image classification.

The better the quality of the input image during classification, the better the performance [15]. In this study, the concept of super resolution and sharpness is used to improve the image quality. Super-resolution is a high-resolution version of a low-resolution image. It involves increasing the number of pixels in the image by scaling the image. Image processing [16], machine learning [17] and deep learning [18] techniques are widely used in super-resolution applications. Wang et al. [18] have shown in their study that deep learning outperforms super-resolution and is widely used in convolutional networks [19], adversarial generative networks [20] and vision transformers [21].

Transformers are first used in natural language processing to capture long-range prior information [22]. Self-attention layers are the core components of transformers and use the keys, values, and query metrics to compute how a piece of information interacts with all other indices in the sequence. In image processing, the mechanism works by segmenting images into specific parts, flattening the segmented parts, embedding them as a low-dimensional vector, adding position information, and then evaluating this vector as a patch array. Image transformers have achieved successful results in areas such as super-resolution, object detection, image enhancement, and segmentation [23]. Liu et al. proposed the Swin transformer [24], a hierarchical hybrid structure that divides the image into windows based on patches. Instead of processing the whole image simultaneously, patch-based processing divides the image into non-overlapping parts and processes them independently. They have multi-head attention mechanisms to capture the relationship between patches in the image and also use shifted windows to effectively deal with each other. In this way, global context is captured while maintaining computational efficiency. In addition to these advantages, features are hierarchically aggregated at the image patch level. SwinIR [25], HST [26], Swin2SR [27] are some of the Swin-based works that have improved the performance of super-resolution, and we have used these methods in our study. Our contributions to this work are listed below:

- We applied three different swin-based super-resolution methods to the sampled RSI-Cb [28] remote sensing dataset and classified them using pre-trained state-of-the-art classification methods to determine the best combinations.

- We tested our study on five man-made structured image-set and five natural field image-set and found that the natural images are improved better than man-made images.
- Our experiments show that the use of super-resolution as pre-processing in image classification increases classification accuracy.

2 RELATED WORK

Remote sensing is used in geographic information systems, aerospace, mining, engineering and many other fields. Zhang et al. [29] describe the development and challenges of remote sensing techniques. In addition, Richards and Jia [30] have done extensive work on enhancing, correcting and analyzing remote sensing images. With the popularization of machine learning, it has been observed that the performance of remote sensing images and data is increasing to make them meaningful [31, 32]. Cengiz and Avcı [33] compared the performance of machine learning methods with the super-resolution method of satellite images. Various studies on the RSI-Cb dataset, which we used in this study, have also performed classifications using deep learning methods [34, 35].

In image classification, the improvement of the image through image preprocessing affects the performance. Therefore, when classifying remote sensing images, the data should be of high quality. To improve image quality, various deep learning methods have been used to increase classification accuracy [36,37]. Swin-based transformers are increasingly used to improve the resolution and sharpness of super-resolution images. They are also used in healthcare [38], metrology [39], agriculture [40] and many other fields. Jannat and Willis [41], who were the first to use the Swin Transformer, a Vision Transformer (VIT), for classification, were more successful in classification performance compared to traditional CNN models using the EuroSat, NWPU-RESISC45 and AID datasets. Sentinel-1 and Sentinel-2 features using Random Forest, Support Vector Machine, VGG-16, 3D CNN, and Swin Transformer were compared with their proposed coastal wetland classification model built from Google Earth Engine (GEE) images and LIDAR data called DEM using QGIS software and the LAS tool [42]. They proposed a 3-layer model with a modified version of the VGG-16 model, a 3D CNN and a Swin transformer and obtained better results. They developed a Swin Unet model for segmenting Sentinel-2 MSI (Multi-Spectral Imager) images into 10 categories and an image segmentation for preprocessing [43]. Swin UNet consists of an encoder, a bottleneck, a decoder and a skip connection based on the skip connection of UNet to reduce semantic information loss. The model has shown promising results compared to other CNN-based models, including DeepLabV3+ and U-Net, as well as VGG, ResNet50, MobileNet and Xception. They have proposed a new Swin Transformer-based model for contrastive self-supervised learning (Swin-TCSSL) using the CIFAR-10, Snapshot Serengeti, Stanford Dogs, Animals with Attributes, and ImageNet datasets [44]. Swin Transformer [24], introduced as Tiny Swin-T with respect to C = 96, layer numbers = {2, 2, 6, 2} Swin -TCSSL is a self-supervised learning method coupled with input images. Swin -TCSSL achieves good accuracy while reducing computation time and cost compared to other methods. 3D Swin T (3DSwinT-HCL), a method developed for classifying hyperspectral images with fine details, is an alternative to supervised learning with self-supervised learning (SSL) [45]. Another method, SpectralSWIN, developed a Swin Spectral Module (SSM) that effectively represents spectral-spatial features in hyperspectral images [46]. P-Swin [47] proposed a parallel window-based transformation network that better extracts contextual information from remote sensing data. In this study, we used HST [26], SwinIR [25], and Swin2SR [27] for super-resolution.

Remote sensing image classification has been extensively studied and CNN-based classification [48] compared the results of research with different models such as VGG-16, U-Net using CNN model for different datasets. Gargees and Scott [49] developed a chip-based change detection method to extract features from the images of RSI-CB256

and CoMo dataset, and performed classification in the deep visual model by orthogonal feature reduction using ResNet50, which is known as deep convolutional neural network (DCNN) and belongs to transfer learning methods. The chip-based method facilitates change detection by making it easier to analyze the level of detail in the images. They used soft clustering based on the fuzzy C-Means algorithm to combine the images. Cluster analysis, geospatial analysis and metric change analysis were used to analyze the images. On the RSI-CB256 dataset, the validation and test performances were 98.72% and 99.31%, respectively. Jayasree et al. [50] classified a large aerial image dataset generated by Google Earth using AID and RSI-CB256 by running it through a CNN with the EfficientNetB7, MobileNetV2 and ResNet50 models. They achieved an accuracy of 94% for the AID dataset and an accuracy of 96.53% for the RSI-CB256. Wijaya et al. [51] compared the performance of MobileNet V-2, ResNet50 and VGG-16 models used for limited resources using RSI-CB256 satellite imagery and contributed that they can be used for computationally intensive devices. Huang [52], who developed the RSIC model fusion method based on deep transfer learning and multi-feature networks, classified remote sensing images using the VGG16, Inception V3, ResNet50, and MobileNet models. The source domain is trained with one of the four models using parameters trained with the CNN ImageNet, and then a model is created by binary fusion by transferring information to the target domain TL-CNN. In the study [52] using the UC land use dataset and the RSIC benchmark dataset RSI-CB, the best model was the Transfer Learning ResNet50-MobileNet (TL-RM) with an average accuracy of 96.8%. To extract the best features from the SAT-4, SAT-6, and RSI-CB datasets, VGG19 and ResNet50 architectures were used for decision tree, K-nearest neighbor (K-NN), and modified random forest with empirical loss function for classification with the combination of the separately extracted features and an accuracy of 97% in decision tree, 89.05% in K-NN, and 99.89% in modified RF [53].

3 MATERIAL AND METHOD

3.1 Dataset

The RSI-Cb [28] remote sensing image dataset has a size of more than 24,000 256x256 pixels with 35 subclasses from 6 categories. In our experiments, we used 10 subclasses. These are airplane, bare land, city building, container, desert, forest, marina, mountain, parking lot and river. These classes are used in the datasets as 5 man-made structures and 5 natural structures. The dataset consists of a total of 7247 images, of which 6514 are used for training the model, 721 for validation and 733 for testing the accuracy of the model.

3.2 Methods for Super-resolution and Image Sharpening

3.2.1 SwinIR

SwinIR [25] is a method that promises to increase the super-resolution of the image, clean up the image and reduce JPEG compression artifacts by using a Swin-based transformer algorithm. SwinIR consists of three parts. These are shallow feature extraction, deep feature extraction and high-quality image reconstruction. The most important module of the three parts, deep feature extraction, is the Residual Swin Transformer Block (RSTB) with additional residual connections to multiple Swin Transformer layers. This residual block establishes an identity-based connection with the reconstruction so that features from different blocks can be aggregated. This method has

shown more promising results than other super-resolution methods. For super-resolution in this work, we used the trained model from the SwinIR authors' Github[1] and chose scale 4.

3.2.2 Hierarchical Swin Transformer (HST)

HST [26], which extracts features in a hierarchical structure, is processed using the divide-and-conquer technique, which gives the network representability. Therefore, it has improved the difficult image distortions in the image in terms of parameters and performance. The HST architecture is a hierarchical architecture with three branches. Pang et al. [54] developed a FAN technique that has three different convolutional layers with hierarchically different kernel sizes and levels to extract hierarchical features at three scales. They used multiple RSTBs for feature extraction. To upscale the image from a low scale to a high scale, they combined the pixel shuffle technique [55] and applied a feature fusion module. They achieved better results with the proposed method than with the SwinIR method. We opted for scale 4 and used the checkpoint_comp40_x4 model trained by HST. For the HST model, we used the code from the authors' Github[2].

3.2.3 Swin2SR

Swin2SR [27] is an improved version of the Swin transformer. Swin2SR uses the Residual Transformer Block (RSTB) and the new SwinV2 transformer [56]. By using the SwinV2 transformer, the feature variance of the deeper layers is improved by using post-normalization instead of pre-normalization. They upscaled the method with bicubic interpolation [57]. In this way, no significant structural information is lost. Good results were obtained in removing JPEG compression errors, super-resolving the image and removing image distortions. In this paper, we use the Swin2SR_CompressedSR_X4_48 (ClassicalSR_X2) model as scale 4, which is pre-trained in the Swin2SR super-resolution method. Swin2SR was run with the code available on Github[3].

3.3 Methods for Image Classifications

3.3.1 DenseNet-121

DenseNet [58], also known as Densely Connected Convolutional Networks, is an architecture that provides maximum information flow, with each layer forwarding all other layers, and has excellent performance. In DenseNet, there are layers with bottlenecks and dense blocks as transitions. The critical point of this model is the bottleneck layer, which reduces the number of parameters by reusing features by forwarding each layer to the other layers. The bottleneck layer consists of two convolutional layers and a batch normalization layer. The first layer reduces the number of input feature maps, while the second layer generates 3x3 feature maps for output. The output of this layer is merged with the second convolution layer of the input feature maps. The transition layer consists of a 1x1 convolution layer, a batch normalization layer and a 2x2 mean pooling layer. These layers reduce the number of feature maps and reduce the spatial dimensionality of the maps. The result is high accuracy through efficient use of parameters and memory. DenseNet121 [12] states that DenseNet consists of 121 connected convolutional layers that contain a 1000-unit layer up to the last output layer in addition to the DenseNet features.

[1] https://github.com/JingyunLiang/SwinIR/releases/download/v0.0/003_realSR_BSRGAN_DFOWMFC_s64w8_SwinIR-L_x4_GAN.pth

[2] https://github.com/lixinustc/HST-for-Compressed-Image-SR

[3] https://github.com/mv-lab/swin2sr

3.3.2 Xception

Xception [13] is a convolutional neural network with 71 layers. It is an extension of the Inception V3 architecture, replacing the Inceptionv3 [59] modules with depth-separable convolutions. While Inception applies different filter sizes to the convolutional layers, Xception applies depth-separable convolutions to the convolutional layers. This makes it possible to achieve more effective results with fewer parameters. There is also an intermediate ReLU nonlinearity in the middle layer. Xception has 3 streams. These are the input stream, the middle stream and the output stream. The input stream consists of two block convolution layers and a ReLU activation function. These layers are followed by depth-separable convolution layers, a maximum pooling layer and hopping connections. The middle stream consists of ReLu and depth-separable convolutional layers. This middle stream is updated 8 times. Finally, the output stream has global average pooling as a mixture of the architectures of the input and middle streams.

3.3.3 EfficientV2_B3

EfficientNetV2 [14] is an extension of EfficentNet [60]. In addition to the inverted bottleneck residual blocks of MobileNetv2 [61], EfficientNet is an architecture that scales all depth, width and resolution dimensions with a uniform coefficient. The uniform scaling method uses AutoML [62] to adjust the layers, depth, and resolution of the network according to the size of the input image, which requires a size-aware channel. This fine-tuning has been found to improve performance. EfficientNetV2 adds both MBConv [63] and fused-MBConv [63] to EfficentNet, a smaller expansion ratio for MBConv, a preference for many times smaller 3x3 kernel sizes, and the elimination of the stride-1 step in the last step. These changes have improved both the performance and speed of training.

4 EXPRIMENTS AND EVALUATION

4.1 Configurations

The experiments were performed on a GoogleColab [64] A100GPU processor using the PyTorch library. Pre-trained DenseNet-121, Xception and EfficientV2_B3 were used as architectures and trained with global average pooling followed by 3 times relu, batch normalization, dropout and in the fully connected layer. A learning parameter of 0.001 and the Adam optimization algorithm were used for optimization. The experiments were performed for 50 epochs with the RSI-Cb dataset and with super-resolution using the HST, Swin2SR and SwinIR architectures.

4.2 Evaluation Metrics

The accuracy and F1 score metrics were used to evaluate the performance and effectiveness of the DenseNet121, Xception and EfficientV2_B3 models at super-resolution of remote sensing image data using the HST, Swin2SR and SwinIR transformers. Accuracy determines how often the model makes an accurate prediction. The F1 score metric is the harmonic mean of precision and recall. The accuracy and F1 scores obtained after 50 epochs of training with the DenseNet121, Xception and EfficientV2_B3 models for remote sensing images acquired with the HST, Swin2SR and SwinIR transformers on the validation set. Table 1 shows that the classification performance without super-resolution is lower than that with super-resolution. When classifying with the DenseNet121 architecture, SwinIR has the highest accuracy and F1-score. When classifying with the Xception architecture, HST has the highest

accuracy and F1 score. In the classification with the EfficientV2_B3 architecture, Swin2SR has the highest accuracy and F1 score.

Table 1: Results from the validation set using DenseNet121, Xception and the EfficientV2_B3 architecture model with transfer learning technique.

	DenseNet121		**Xception**		**EfficientV2_B3**	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Raw Image	0.9891	0.9894	0.9918	0.9916	0.9864	0.9884
HST	0.9918	0.9908	**0.9959**	**0.9960**	0.9905	0.9908
SwinIR	0.9932	0.9932	0.9932	0.9929	0.9918	0.9929
Swin2SR	0.9918	0.9919	0.9932	0.9934	0.9932	0.9947

Table 2: Results of the average F1 score values for man-made and natural classes using the DenseNet121, Xception and EfficientV2_B3 architecture models with transfer learning technique.

	DenseNet121		**Xception**		**EfficientV2_B3**	
	Man-made	Natural	Man-made	Natural	Man-made	Natural
Raw Image	0.99182	0.98688	0.99434	0.98898	0.99794	0.97882
HST	0.99366	0.98796	0.99794	0.99404	0.99584	0.98586
SwinIR	0.99434	0.99200	0.99182	0.99404	0.99794	0.98792
Swin2SR	0.99182	0.99200	0.99794	0.98894	1.0000	0.98948
Average	0.99327	0.99065	0.99590	0.99234	0.99793	0.98775

The accuracy rates of other studies for the entire RSI-Cb dataset were 95.13% with the VGG-16 architecture in [48], 99.31% with the ResNet50 architecture in [49], 96.53% with the ResNet50 architecture in [50], 98.94% in [51], 96.8% with the TL-RM architecture in [52], and 99.89% with the Random Forest architecture in [53]. In our study, it was found that the classification accuracy of the 10 subclasses selected from the RSI-Cb dataset increased from 99.18% to 99.59% after super-resolution with the HST method and classification with the Xception model. The F1 values of the images created with the Swin-Transformer are evaluated as man-made and natural in Table 2. It can be seen from the results in Table 2 that natural images provide better results.

4.3 Generated Images

An image of the container class we used in the dataset, whose resolution and sharpness was improved using the HST, Swin2SR and SwinIR algorithms, is shown in Figure 1.

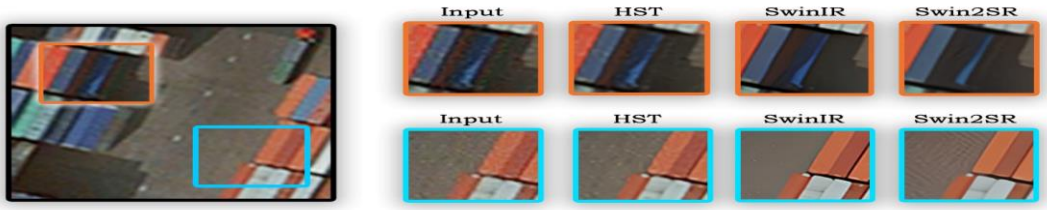

Figure 1: Cropped regions from the container image with the original input, and the results of SwinIR, HST and Swin2SR.

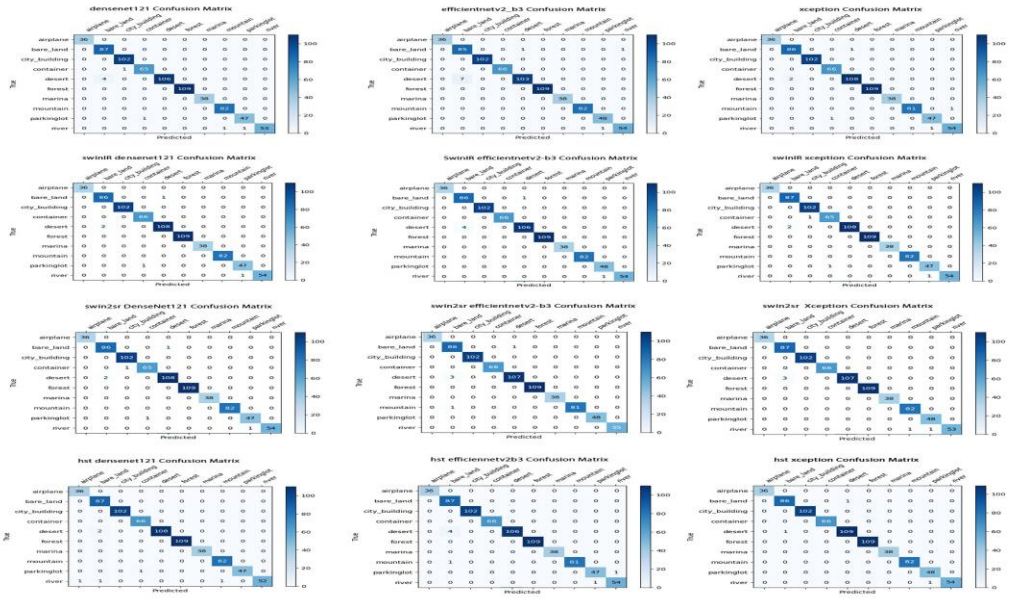

Figure 2: Confusion matrices for the validation part of the RSI-Cb dataset using the images from Input, SwinIR, HST and Swin2SR on the pre-trained models DenseNet121, Xception and EfficientV2_B3. The final (bottom right) confusion matrix is the HST Super-resolutioned Xception classification confusion matrix, which gives the best result.

5 CONCLUSIONS

This study investigated the change in image classification performance with increasing image resolution and sharpness. We evaluated the performance of remote sensing image data acquired with HST, Swin2SR and SwinIR transformers. We evaluated the performance using the pre-trained models DenseNet121, Xception and EfficientV2_B3 and assessed their effectiveness using the metrics accuracy and F1-score. The results show that increasing image quality significantly improves classification accuracy and F1-score. It was also found that the performance was better for the images with natural structures than for the images with man-made structures. Since the quality of the image improves the classification, it can be said that the preprocessing of image enhancement improves the classification performance. In the future, it is planned to use transformer-based image enhancement and color preservation techniques in addition to the super-resolution techniques used.

REFERENCES

[1] Bing, Z. (2017). Current Status and Future Prospects of Remote Sensing. Bulletin of Chinese Academy of Sciences (Chinese Version), 32(7), 774-784.Jason Jerald. 2015. The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan & Claypool.

[2] Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., ... & Blaschke, T. (2019). Big Earth data: disruptive changes in Earth observation data management and analysis? International Journal of Digital Earth.

[3] Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., & Kembhavi, A. (2023). SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16772-16782).

[4] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[5] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[6] Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301.

[7] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[8] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.

[9] Ouchra, H., Belangour, A., & Erraissi, A. (2022, October). Machine learning for satellite image classification: A comprehensive review. In 2022 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 1-5). IEEE.

[10] Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing, 28(5), 823-870.

[11] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[14] Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International conference on machine learning (pp. 10096-10106). PMLR.

[15] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

[16] Van Ouwerkerk, J. D. (2006). Image super-resolution survey. Image and vision Computing, 24(10), 1039-1052.

[17] Nasrollahi, K., & Moeslund, T. B. (2014). Super-resolution: a comprehensive survey. Machine vision and applications, 25, 1423-1468.

[18] Wang, Z., Chen, J., & Hoi, S. C. (2020). Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3365-3387.

[19] Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307.

[20] Tian, C., Zhang, X., Lin, J. C. W., Zuo, W., Zhang, Y., & Lin, C. W. (2022). Generative adversarial networks for image super-resolution: A survey. arXiv preprint arXiv:2204.13620.

[21] Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., & Zeng, T. (2022). Transformer for single image super-resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 457-466).

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[25] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844).

[26] Li, B., Li, X., Lu, Y., Liu, S., Feng, R., & Chen, Z. (2022, October). Hst: Hierarchical swin transformer for compressed image super-resolution. In European Conference on Computer Vision (pp. 651-668). Cham: Springer Nature Switzerland.

[27] Conde, M. V., Choi, U. J., Burchi, M., & Timofte, R. (2022, October). Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In European Conference on Computer Vision (pp. 669-687). Cham: Springer Nature Switzerland.

[28] Li, H., Dou, X., Tao, C., Hou, Z., Chen, J., Peng, J., ... & Zhao, L. (2017). RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data. arXiv preprint arXiv:1705.10450.

[29] Zhang, B., Wu, Y., Zhao, B., Chanussot, J., Hong, D., Yao, J., & Gao, L. (2022). Progress and challenges in intelligent remote sensing satellite systems. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1814-1822.

[30] Richards, J. A., Jia, X. (2022). Remote sensing digital image analysis (Vol. 5). Berlin/Heidelberg, Germany: springer.

[31] Meraj, G., Kanga, S., Ambadkar, A., Kumar, P., Singh, S. K., Farooq, M., ... & Sahu, N. (2022). Assessing the yield of wheat using satellite remote sensing-based machine learning algorithms and simulation modeling. Remote Sensing, 14(13), 3005.

[32] Shirmard, H., Farahbakhsh, E., Müller, R. D., & Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration. Remote Sensing of Environment, 268, 112750.

[33] Cengiz, A., & Avcı, D. (2023). The Effect of Super Resolution Method on Classification Performance of Satellite Images. Turkish Journal of Science and Technology, 18(2), 331-344.

[34] Cheng, G., Xie, X., Han, J., Guo, L., & Xia, G. S. (2020). Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3735-3756.

[35] Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516.

[36] Ansith, S., & Bini, A. A. (2022). Land use classification of high resolution remote sensing images using an encoder based modified GAN

architecture. Displays, 74, 102229.

[37] Zeng, Y., Guo, Y., & Li, J. (2022). Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning. Neural Computing and Applications, 34(4), 2691-2706.

[38] Sun, R., Pang, Y., & Li, W. (2023). Efficient Lung Cancer Image Classification and Segmentation Algorithm Based on an Improved Swin Transformer. Electronics, 12(4), 1024.

[39] Wei, L., Zhu, T., Guo, Y., & Ni, C. (2023). MMST: A Multi-Modal Ground-Based Cloud Image Classification Method. Sensors, 23(9), 4222.

[40] Xie, J., Hua, J., Chen, S., Wu, P., Gao, P., Sun, D., ... & Lu, J. (2023). HyperSFormer: A transformer-based end-to-end hyperspectral image classification method for crop classification. Remote Sensing, 15(14), 3491.

[41] Jannat, F. E., & Willis, A. R. (2022, March). Improving classification of remotely sensed images with the swin transformer. In SoutheastCon 2022 (pp. 611-618). IEEE.

[42] Jamali, A., & Mahdianpari, M. (2022). Swin transformer and deep convolutional neural networks for coastal wetland classification using sentinel-1, sentinel-2, and LiDAR data. Remote Sensing, 14(2), 359.

[43] Yao, J., & Jin, S. (2022). Multi-category segmentation of Sentinel-2 images based on the Swin UNet method. Remote Sensing, 14(14), 3382.

[44] Agilandeeswari, L., & Meena, S. D. (2023). SWIN transformer based contrastive self-supervised learning for animal detection and classification. Multimedia Tools and Applications, 82(7), 10445-10470.

[45] Huang, X., Dong, M., Li, J., & Guo, X. (2022). A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.

[46] Ayas, S., & Tunc-Gormus, E. (2022). SpectralSWIN: a spectral-swin transformer network for hyperspectral image classification. International Journal of Remote Sensing, 43(11), 4025-4044.

[47] Wang, D., Yang, R., Zhang, Z., Liu, H., Tan, J., Li, S., ... & Su, P. (2023). P-Swin: Parallel Swin transformer multi-scale semantic segmentation network for land cover classification. Computers & Geosciences, 175, 105340.

[48] Jarrallah, Z. H. (2022, November). Satellite Images Classification Using CNN: A Survey. In 2022 International Conference on Data Science and Intelligent Computing (ICDSIC) (pp. 111-116). IEEE.

[49] Gargees, R. S., & Scott, G. J. (2021). Large-scale, multiple level-of-detail change detection from remote sensing imagery using deep visual feature clustering. Remote Sensing, 13(9), 1661.

[50] Jayasree, J., Madhavi, A. V., & Geetha, G. (2023, April). Multi-Label Classification On Aerial Images Using Deep Learning Techniques. In 2023 International Conference on Networking and Communications (ICNWC) (pp. 1-6). IEEE.

[51] Wijaya, B. A., Gea, P. J., Gea, A. D., Sembiring, A., & Hutagalung, C. M. S. (2023). Satellite Images Classification using MobileNet V-2 Algorithm. Sinkron: jurnal dan penelitian teknik informatika, 8(4), 2316-2326.

[52] Huang, X. (2023). High Resolution Remote Sensing Image Classification Based on Deep Transfer Learning and Multi Feature Network. IEEE Access.

[53] Pazhanikumar, K., & KuzhalVoiMozhi, S. N. (2023). Remote sensing image classification using modified random forest with empirical loss function through crowd-sourced data. Multimedia Tools and Applications, 1-23.

[54] Pang, Y., Li, X., Jin, X., Wu, Y., Liu, J., Liu, S., & Chen, Z. (2020). FAN: frequency aggregation network for real image super-resolution. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16 (pp. 468-483). Springer International Publishing.

[55] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1874-1883).

[56] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12009-12019).

[57] Yang, R., Timofte, R., Li, X., Zhang, Q., Zhang, L., Liu, F., ... & Peng, L. (2022, October). Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In European Conference on Computer Vision (pp. 174-202). Cham: Springer Nature Switzerland.

[58] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[59] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2021). Rethinking the Inception Architecture for Computer Vision. arXiv: 151200567; 2015.

[60] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[61] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[62] He, Y., Lin, J., Liu, Z., Wang, H., Li, L. J., & Han, S. (2018). Amc: Automl for model compression and acceleration on mobile devices. In Proceedings of the European conference on computer vision (ECCV) (pp. 784-800).

[63] Gupta, S., & Akin, B. (2020). Accelerator-aware neural network design using automl. arXiv preprint arXiv:2003.02838.

[64] Bisong, E., & Bisong, E. (2019). Google colaboratory. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners, 59-64.