2024 5th International Conference on Computer Engineering and Application (ICCEA) | 979-8-3503-8677-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCEA62105.2024.10604222

EsccNet: A Hybrid CNN and Transformers Model for the Classification of Whole Slide Images of Esophageal Squamous Cell Carcinoma

Zhaoxin Kang [1,a]
[1]College of Computer and Data Science
Fuzhou University
Fuzhou, China
[a]221020033@fzu.edu.cn

Hejun Zhang [2,b]
[2]Department of Pathology
Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital
Fuzhou, China
[b]1483196228@qq.com

Mingqiu Chen [3,c*]
[3]Department of Radiation
Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital
Fuzhou, China
[c]drchenmingqiu@163.com

Xiangwen Liao [4,d*]
[4]College of Computer and Data Science
Fuzhou University
Fuzhou, China
[d]liaoxw@fzu.edu.cn

***Abstract*—This study presents a novel approach in the application of deep learning for the classification of esophageal squamous cell carcinoma (Escc) using whole-slide images (WSIs). Our methodology uniquely combines Convolutional Neural Network (CNN) with Transformer, leveraging the strengths of both architectures to enhance the accuracy and efficiency of cancer detection and classification in histopathological images. In this research, we first preprocess a substantial dataset of WSI samples, annotated by expert pathologists, to train and validate our model. The CNN component effectively extracts detailed local features from the high-resolution images, while the Transformer, known for its capability in handling sequential data, adeptly manages the global context, addressing the challenges posed by the complex and heterogeneous nature of WSIs. The accuracy, F1 score, recall, and precision of our proposed model on the dataset provided by Fujian Cancer Hospital are 94.71%, 94.32%, 94.68%, and 94.08%, respectively, which are significantly better than other models. This study not only assists pathologists in analyzing esophageal squamous carcinoma WSIs but also paves the way for further research into the combined application of CNN and Transformer in the diagnosis of other types of cancer.**

Keywords-WSIs, esophageal squamous cell carcinoma (Escc), CNN, Transformer

I. INTRODUCTION

Esophageal cancer is the ninth most common cancer worldwide and the sixth leading cause of cancer death[1]. In China, esophageal cancer represents a significant burden in terms of incidence and mortality rates among male cancers, ranking among the top five types of cancer following lung cancer[2]. Esophageal squamous cell carcinoma (Escc) is a more common type of Esophageal carcinoma. Known for its aggressive nature and high mortality rates, Escc often leads to late-stage discovery and exhibits resistance to conventional treatments. These factors contribute to the low survival rates, reinforcing the urgent need for improved diagnostic techniques that enable early and precise identification of Escc. This situation underscores the urgent need for more effective diagnostic methods to ensure early and accurate detection of the disease.

Traditionally, the diagnosis of cancers like Escc largely relies on pathologists manually examining Whole-Slide Images (WSIs). Pathologists have to painstakingly analyze extensive tissue areas, which can be time-consuming and subjective, leading to varying diagnoses for the same case by different pathologists. Moreover, the rising caseload in remote and underdeveloped regions, coupled with a shortage of experienced pathologists, exacerbates this issue. Therefore, there is acritical need to develop a deep learning system to classify pathological images of esophageal squamous cell carcinoma, assisting pathologists in improving diagnostic efficiency and establishing objective standards.

In recent years, with the gradual maturity of deep learning, especially convolutional neural networks, CNN still has a place in the field of computer vision. CNN can learn morphological features. This method provides a new technical means for the automatic analysis of tumor pathological images. Evidence of their effectiveness can be seen in various medical challenges where CNNs have been employed. Byungjae Lee and Kyunghyun Paeng[3] published the results of four classification of pathological stage on the entire pathological image based on the Camelyon16 and Camelyon17 datasets, which reached or even exceeded human recognition accuracy in terms of accuracy. Alom et al.[4] proposed a Recurrent Convolutional Neural Network (IRRCNN) with Inception module, which combines the advantages of Inception Network (Inception v4), Residual Network, and Recurrent Convolutional Neural Network. On the data set of 2015 Break Cancer Classification Challenge, the accuracy of the second and multi classification of breast cancer was 99.1% and 98.1% respectively. Zeiser et al.[5] introduced a convolutional neural network (CNN) based model that utilizes cascaded CNN and U-Net to provide fine segmentation of cancer WSI. Soldatov et al.[6] applied deep learning methods to diagnose six types of colon mucosal lesions using convolutional

This work was partially supported by the National Natural Science Foundation of China (No. 82173051) and the Provincial Natural Science Foundation of Fujian (No. 2023J011287).

neural networks (CNNs), and developed an algorithm for WSI automatic segmentation in colon biopsy based on ResNet and EfficientNet[7].

However, CNNs, due to their excessive focus on local features of images, tend to overlook the global context of the images. To address this, researchers have introduced the Vision Transformer (VIT)[8]. VIT treats an image as a sequence of patches and applies the self-attention mechanism to them. This process allows the model to dynamically focus on the most informative parts of the image, regardless of their spatial location, thereby compensating for the limitations of CNN[9]. Consequently, we propose the hybrid model called EsccNet, combining CNNs and Transformers. Our model is designed to capitalize on the traditional CNN's sensitivity to local information, while incorporating the global and contextual awareness insights provided by VIT for the classification of Esophageal Squamous Cell Carcinoma WSI images. It can automatically classify pathological images of esophageal squamous cell carcinoma into mediators (nom), tumor tissues (tumor), twisted and necrotic (tn), and squamous epithelial tissues (es). In order to obtain richer image features, EsccNet has two channels. One is based on AFF-ResNet and an improved VIT Transformer to obtain global contextual features, known as the Global Information Feature Extraction Module (GFE). Another method also uses a structure called Local Information Feature Extraction Module (LFE) to obtain local information features. Our work makes the following contribution:

(1) We propose a new hybrid network to integrate global and local information for better performance.

(2) Our model successfully achieved the best performance on a dataset labeled by multiple professional pathologists from Fujian Cancer Hospital.

(3) At present, there are relatively few automated analysis methods used for esophageal squamous cell carcinoma, and our article supplements this point.

The rest of this paper is divided into three sections. Section 2 details the proposed EsccNet model. Section 3 describes the corresponding experiments; Section 4 presents the final conclusions and future prospects.

II. Method

A. *EsccNet Deep Learning Architecture*

EsccNet is a deep learning model that integrates both CNN and transformer architectures. Its overall framework is depicted in Figure 1. EsccNet comprises two main parts: the Global Feature Extraction (GFE), which adopts the transformer technique to learn global features of esophageal squamous carcinoma pathology slides, and the Local Feature Extraction (LFE), which uses traditional convolution to extract local features from the pathology slides. Before processing, Whole Slide Images (WSIs) must be cropped into computationally manageable sizes (e.g., 256x256).

Within the GFM, images are initially processed by a Conv Layer composed of 64 filters of 7x7 size. This is followed by three Residual modules, the first two of which are adapted from two modules within ResNet34[10]. The third Residual module introduces the Attention Fusion Feature (AFF) module[11], which fuses feature maps based on attention weights, enhancing the traditional addition operations with the AFF mechanism. The resulting features pass through a Patch Embedding layer and are then processed in an Efficient Transformer Block to extract global features. Meanwhile, in the LFE module, traditional convolution operations coupled with the AFF attention fusion mechanism effectively extract local information. Subsequently, the features extracted from the LFE and GFE are fused through a convolution layer of size 1x1 and a fully connected layer, and finally, the probability of each category is obtained through the softmax function. The highest probability indicates the classification of the category.

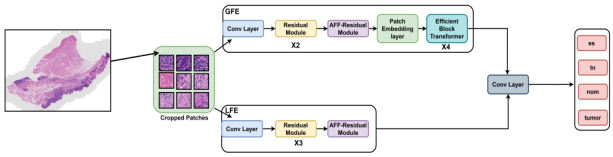

Figure 1: The structure of EsccNet. This network has two channels, the upper channel is the global information feature extraction module GFE, and the lower channel is the local feature extraction module LFE.

B. *Efficient Transformer Block*

As illustrated in Figure 2, the left side represents a standard Vision Transformer (VIT) block, usually comprised of MLP (Multilayer Perceptron), Norm (Normalization), and MSA (Multi-Head Self-Attention) modules. The MLP module is responsible for processing the sequence of embedded patches. The Norm module standardizes the data to stabilize the learning process. The MSA module focuses on capturing the complex dependencies between different parts of the image. Inspired by Resnet, the introduction of residual connections helps to mitigate the issue of vanishing gradients and enables deeper architectural designs by allowing gradients to flow directly through the network layers. For the token input x, the output z of the corresponding transformer block is:

$$y=x+MSA(Norm(x)) \text{ and } z=y+MLP(Norm(x)) \quad (1)$$

However, MSA has two significant drawbacks: 1) The scale of MSA computation is tied to the dimensions of the input tokens, either dm or n2, leading to substantial training and inference overheads; 2) In MSA, each head is responsible for only a subset of input tokens, which can potentially impact network performance. To address these two issues, we have introduced the method by Zhang et al[12]. As depicted in Figure 3, researchers modified the MSA module by: 1) Developing a memory-efficient multi-head self-attention algorithm. This algorithm compresses memory through simple depth wise convolutions and, while maintaining the diversity of multiple heads, projects interactions onto the dimensions of attention heads, significantly reducing overhead. 2) Using up-sampling operations to reconstruct the mid and high-frequency information lost due to down-sampling operations. These two methods have achieved notable success, and their application in the identification of esophageal squamous cell carcinoma in Whole Slide Images (WSI) has also been effective. The mathematical definition of EMS is as follows:

$$EMS(Q,K,V)=Softmax(\frac{QK^T}{\sqrt{d_k}})V+Up(V) \quad (2)$$

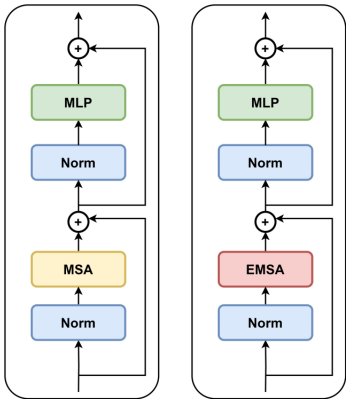

Figure 2: Examples of MSA blocks and EMSA. Left: A Standard Transformer Block. Right: The Efficient Transformer Block. The sole distinction from the standard Transformer block lies in substituting the Multi-Head Self-Attention (MSA) with an Efficient Multi-head Self-Attention (EMSA).

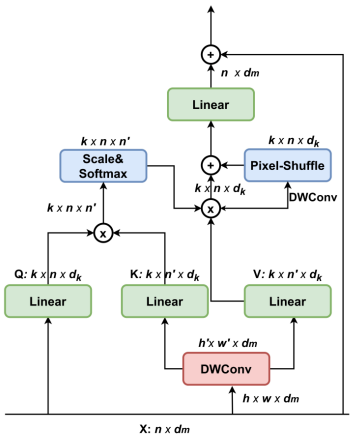

Figure 3: The structure of EMSA. In EMSA, DWConv and Pixel-Shuffle are respectively two corresponding effective solutions.

C. *AFF Residual Module*

In both the Global Feature Extraction (GFE) and Local Feature Extraction (LFE) modules, we have adopted the AFF Residual Module, as shown in Figure 4. The AFF Residual Module differs from the traditional Residual Module in that it incorporates the AFF (Adaptive Feature Fusion) module instead of the conventional addition operation. The AFF module comprises two branches, each utilizing different scales to extract channel attention weights. One branch adopts Global Average Pooling to extract attention for global features. This approach emphasizes important features by pooling global information and disregarding less significant aspects. The other branch directly uses point-wise convolution to extract channel attention for local features, focusing on local details and variations within the image. Finally, the features from these two branches are fused according to their respective weights. This combination of global and local features aims to enhance the network's sensitivity and discriminative power towards different regions in the image, thereby improving overall feature extraction and recognition capabilities.

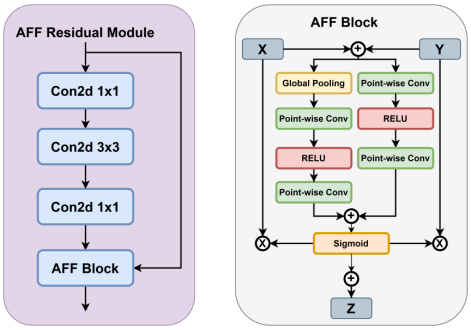

Figure 4: The structure of AFF Residual Module and AFF Block. Left: AFF Residual Module. Right: AFF Block. AFF Residual Module. In the AFF Residual Module, the input values X and Y of the AFF block are respectively the input to the AFF residual module and the output after convolution processing.

III. Experiments and Result

In this experiment, we used 73 WSI images of esophageal squamous cell carcinoma from 50 cases provided by Fujian Cancer Hospital. In this paper, the dataset is referred to as the FCH dataset. These images were selected and annotated by professional pathologists, and then reviewed by 2-3 different pathologists. We randomly selected 13 WSI images from 7 cases as the test set and processed them into 256*256 patches, excluding patches where more than 70% of the area is background. The remaining 50 WSI images were processed in the same way, and then randomly divided into training and validation sets in a 7:3 ratio. Table 1 shows the number of patches in each set after the division.

Table 1: The distribution of the FCH dataset (patches).

Class	Tumor	Nom	Tn	Es	Total
Train	55605	55846	12286	21894	145631
Val	23995	23754	5332	9332	62413
Test	17625	18664	3853	7626	47768
Total	97225	98264	21471	38852	255812

A. *Experimental Details*

The experiments in this paper were conducted on a Tesla T4, setting the batch size to 64 and using the Adam optimizer. Initial learning rate was set to 0.0001, with a halving of the learning rate every 10 epochs, and each experiment was set for 50 epochs. Due to the issue of class imbalance among the data, we introduced focal loss to address this problem. Below is the definition of the focal loss function[13]:

$$FL(p_t)= -(1-p_t)^{\gamma}\log(p_t) \quad (3)$$

γ is a parameter that ranges from [0, 5]. When γ is 0, it reverts to the original CE (Cross Entropy) loss function. The factor $(1-p_t)^{\gamma}$ can reduce the loss contribution of easily classified samples, thereby increasing the loss proportion for difficult-to-classify samples. When p_t approaches 1, it indicates

that the sample is easily distinguishable, and at this time, the modulation factor $(1-p_t)^\gamma$ tends to be 0, which means it contributes less to the loss, thus reducing the loss proportion for easily distinguishable samples. Therefore, focal loss addresses the issue of data imbalance.

B. *Experimental Results*

We selected Accuracy, F1-Score, Precision and Recall as evaluation metrics to compare the performance of our model, EsccNet, with other models on the test set. We used pre trained models on ImageNet in all comparative experiments. Our model, in comparison with VGG16, ResNet50, ResNet34, MobileNetV3[14], EfficientNet_b0 and other models, achieved the results with 0.9471 Accuracy, 0.9432 F1-score, 0.9408 Precision, and 0.9468 Recall. All classification results are shown in Table 2. The model we proposed outperforms other models in terms of Accuracy, F1-Score, Precision, and Recall.

Table 2: Classification results of EsccNet and other comparison models on the test set of FCH dataset.

Model	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
VGG16	90.68	90.20	89.02	91.69
EfficientNet_b0	91.89	90.80	89.63	92.35
MobileNetV3	91.93	91.22	89.84	92.97
ResNet34	92.22	91.91	90.86	93.22
AFF_ResNet34	92.66	92.07	90.95	93.50
ResNet50	92.79	92.24	91.22	93.51
AFF_ResNet50	93.27	92.88	92.23	93.70
VIT	85.15	82.38	81.06	85.35
EsccNet	94.71	94.32	94.08	94.68

C. *Ablation Studies*

In order to analyze the impact of the number of AFF modules and Efficient Transformer Block modules in EsccNet on performance, we fixed the random seed and designed two ablation experiments, ensuring all other conditions were equal. Due to the identical structure of the AFF modules in GFE and LFE, we only analyze the AFF modules in GFE. These experiments included varying the number of AFF modules and Efficient Transformer Block modules.

1) *Impact of AFF Modules*

AFF modules can impact the performance of the model. Appropriate numbers of AFF modules enable the model to better extract contextual features. If the number of AFF modules is too high, this will increase the computational cost of the model. Conversely, too few AFF modules may prevent the model from adequately extracting contextual features. Therefore, we set the number of AFF modules in the GFE module, which contains three residual blocks. We numbered these residual blocks, where b=0, 1, 2, 3, where b=3 indicates the use of AFF only on the third residual block and b=2 indicates the use of AFF modules on residual blocks 2 and 3. As shown in Table 3, we observed that as b increases, metrics like accuracy slightly improve. Considering overall performance and other factors, we chose to use AFF modules only on the third residual block.

Table 3: Ablation study on the number of AFF modules.

AFF modules	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
0	93.62	93.07	92.61	93.76
1	93.74	93.15	92.70	93.84
2	94.02	93.38	92.99	94.01
3	94.24	93.55	93.17	94.15

2) *Impact of Efficient Transformer Blocks*

The number of Efficient Transformer Blocks determines the model's ability to extract global features. Similar to AFF modules, an appropriate number of Efficient Transformer Blocks can effectively enhance the model's performance. Therefore, we experimented with 2, 4, and 6 modules at the same location. As shown in Table 4, after comprehensive consideration, we ultimately chose 4 modules for extracting global information

Table 4: Ablation study on the number of Efficient Transformer Blocks.

Efficient-Transfor-mer Blocks	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
2	94.24	93.55	93.17	94.15
4	94.71	94.32	94.08	94.68
6	94.32	94.07	93.75	94.51

D. *Visualization*

We predicted the classification map of the trained model, and Figure 5 shows the predicted classification map of the EsccNet model for the test set image, as well as the corresponding thumbnail of the original image. Each pixel point in the thumbnail represents the prediction result of 256*256 in the WSI original image (level=0), with red representing tumor, green representing nom, blue representing Tn, and yellow representing es.

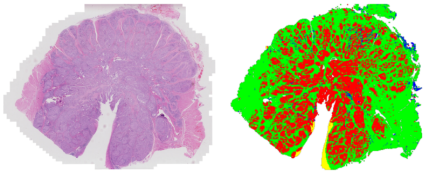

Figure 5: The classification map of the EsccNet model on the test set. Left:Thumbnail of the Original Image; Right: Prediction classification map by EsccNet model.

We used Gradient-weighted Class Activation Mapping (Grad-CAM) [15]to interpret the Esccnet model. Grad-CAM is a type of class activation heatmap that visualizes the most significant results considered by the model through a heatmap. In the heatmap, deep blue and deep red areas respectively represent the contribution to the classification result of this patch as less and high strength. Figure 6 shows the heatmap generated by the convolutional layer after our model fuses features. We randomly selected 3 small patches from different WSIs, which are respectively the boundary between the tumor and normal areas, the boundary between the tumor and Tn areas, and a patch selected within the tumor. It can be observed that our model focuses on the aggregation area of cancerous cells

when extracting the tumor part, which astonishingly coincides with the judgment method of professional pathologists. Thus, the output is more human-understandable and avoids the black-box nature of deep learning models.

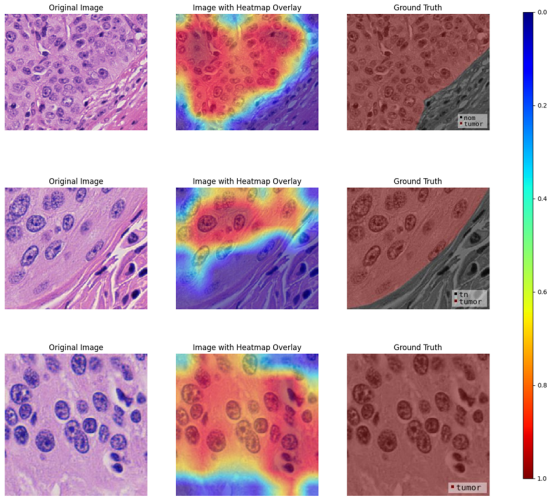

Figure 6: Heat maps generated by WSI patch.

IV. Conclusion

In this study, we designed a model called EsccNet that combines CNN and Transformer. This model combines the advantages of CNN and Transformer to effectively integrate local feature detection and global context. We have pioneered the application of a model integrating CNN and Transformer in the classification of pathological images of esophageal squamous cell carcinoma and demonstrated the feasibility of this approach. This model was used to classify four types of esophageal squamous cell carcinoma: tumor, nom, es, and Tn. On the dataset provided by Fujian Cancer Hospital, this model significantly outperformed other traditional CNN models and achieved the best results. However, validation of a single dataset is not sufficient. Our future work will mainly involve validating datasets provided by different institutions. In general, our research aids pathologists in the diagnosis of diseases and in establishing objective standards.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. 82173051) and the Provincial Natural Science Foundation of Fujian (No. 2023J011287). The study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics Committee of Fujian Cancer Hospital.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. (2020) "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394-424. doi: 10.3322/caac.21609.

[2] W. Cao, H. D. Chen, Y. W. Yu, N. Li, and W. Q. Chen. (2021) "Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020." (in eng). Chinese Medical Journal, vol. 134, no. 7, pp. 783-791. doi: 10.1097/cm9.0000000000001474.

[3] B. Lee and K. Paeng. (2018) "A Robust and Effective Approach Towards Accurate Metastasis Detection and pN-stage Classification in Breast Cancer." In: MICCAI 2018. Granada, Spain. pp. 841-850. doi: 10.48550/arXiv.1805.12067.

[4] M. Z. Alom, C. Yakopcic, M. S. Nasrin, T. M. Taha, and V. K. Asari. (2019) "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network." (in eng). Journal of Digital Imaging, vol. 32, Art no. 4, pp. 605-617. doi: 10.1007/s10278-019-00182-7.

[5] F. A. Zeiser, C. A. da Costa, G. D. Ramos, H. C. Bohn, I. Santos, and A. V. Roehe (2021) "DeepBatch: A hybrid deep learning model for interpretable diagnosis of breast cancer in whole-slide images." Expert Systems with Applications, vol. 185. doi: 10.1016/j.eswa.2021.115586.

[6] S. A. Soldatov, D. M. Pashkov, S. A. Guda, N. S. Karnaukhov, A. A. Guda, and A. V. Soldatov. (2022) "Deep Learning Classification of Colorectal Lesions Based on Whole Slide Images." Algorithms, vol. 15, no. 11, pp. 398-413, Art no. 398. doi: 10.3390/a15110398.

[7] M. Tan and Q. V. Le. (2019) "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In: The 36th International Conference on Machine Learning (ICML 2019) . Long Beach, California, USA. pp. 6105-6114. doi: 10.48550/arXiv.1905.11946.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby. (2020) "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: The 9th International Conference on Learning Representations(ICLR 2021). doi: 10.48550/arXiv.2010.11929.

[9] B. K. Fu, M. D. Zhang, J. J. He, Y. Cao, Y. C. Guo, and R. P. Wang, Jun. (2022) "StoHisNet: A hybrid multi-classification model with CNN and Transformer for gastric pathology images." Computer Methods and Programs in Biomedicine, vol. 221. doi: 10.1016/j.cmpb.2022.106924.

[10] K. He, X. Zhang, S. Ren, and J. Sun. (2016) "Deep Residual Learning for Image Recognition." In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas, NV, USA. pp. 770-778, doi: 10.48550/arXiv.1512.03385.

[11] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard. (2021) "Attentional Feature Fusion." In: Winter Conference on Applications of Computer Vision(WACV 2021). Waikoloa, HI, USA. pp. 3560-3569. doi: 10.48550/arXiv.2009.14082.

[12] Q. L. Zhang and Y. B. Yang. (2022) "ResT V2: Simpler, Faster and Stronger." In: The 36th Annual Conference on Neural Information Processing Systems(NeurIPS 2022). pp. 36440–36452. doi: 10.48550/arXiv.2204.07366.

[13] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. (2017) "Focal Loss for Dense Object Detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318 – 327. doi: 10.48550/arXiv.1708.02002.

[14] A. G. Howard et al. (2017) "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Hawaii, USA. doi: 10.48550/arXiv.1704.04861.

[15] R. R. Selvaraju et al. (2017) "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." In: The 16th IEEE International Conference on Computer Vision (ICCV 2017). Venice, ITALY. pp. 618-626. doi: 10.1109/ICCV.2017.74.