https://doi.org/10.2352/EI.2023.35.1.VDA-407

VIT BASED COVID-19 DETECTION AND CLASSIFICATION FROM CXR IMAGES

Muhammad Saeed[1], *Mohib Ullah*[2], *Sultan D. Khan*[3],
Faouzi Alaya Cheikh[2], *Muhammad Sajjad* [2]

[1] Islamia College University Peshawar, 25000 Peshawar, Pakistan
[2] Norwegian University of Science and Technology, 2815 Gjøvik, Norway
[3] Department of computer science, National University of Technology, 44000 Islamabad, Pakistan

ABSTRACT

The COVID-19 virus induces infection in both the upper respiratory tract and the lungs. Chest X-ray are widely used to diagnose various lung diseases. Considering chest X-ray and CT images, we explore deep-learning-based models namely: AlexNet, VGG16, VGG19, Resnet50, and Resnet101v2 to classify images representing COVID-19 infection and normal health situation. We analyze and present the impact of transfer learning, normalization, resizing, augmentation, and shuffling on the performance of these models. We explored the vision transformer (ViT) model to classify the CXR images. The ViT model incorporates multi-headed attention to disclose more global information in constrast to CNN models at lower layers. This mechanism leads to quantitatively diverse features. The ViT model renders consolidated intermediate representations considering the training data. For experimental analysis, we use two standard datasets and exploit performance metrics: accuracy, precision, recall, and F1-score. The ViT model, driven by self-attention mechanism and long-range context learning, outperforms other models.

***Index Terms*—** COVID-19,Deep Learning,CNN,CXR Images,ViT.

1. INTRODUCTION

An unknown disease was first introduced to humans in late 2019, in China, some people were infected with the disease in Wuhan city, China. The disease was completely unknown at first, but specialists diagnose its symptoms similar to coronavirus infection and flu [1–4]. The specific cause of this was initially unknown, but after the laboratory examination and analysis of positive sputum by real-time polymerase chain reaction (PCR) test, confirmed the infection and named it "COVID-19" upon the recommendation of the World Health Organization (WHO). In a short period, the COVID-19 epidemic spread out geographically with the devastating infection on the health, economies, and welfare of the global population [1–5]. The early detection of COVID-19 is essential not only for patient care but also for public health to ensure the patients' isolation and control of the pandemic. Commonly three methods are used for the diagnosis of COVID-19, which are blood tests, viral tests, and medical imaging. The clinical features of the blood test for COVID-19 include respiratory symptoms, fever, cough, dyspnea, and pneumonia, however, these symptoms do not always indicate COVID-19, and are observed as pneumonia in many cases, leading to a diagnostic problem for physicians [2, 6]. The reliability of blood tests for COVID-19 is low as 2% or 3%. Another common test for COVID-19 is a viral test. A commonly used viral test is the reverse transcription-polymerase chain reaction (RT-PCR) test. RT-PCR is a gold standard diagnosis for COVID-19. However, the sensitivity of this test ranges from 50-62% [7] found by many studies. The third commonly used method for COVID-19 is medical imaging, because COVID-19 targets the respiratory system, therefore, chest radiology scans are important tools for diagnosis and early management. Radiology scans give effective results in the detection of lung conditions along with other illnesses. Radiologists have a range of abnormalities in COVID-19 patients. Deep learning (DL) techniques have been used in medical imaging to improve the performance of image analysis significantly. Convolution neural network (CNN) is mostly used for medical imaging [8, 9]. 2018)]; it has various architectures and applications. We proposed a vision transformer model for this work. Various diagnostic tests have been adopted for SARS-CoV-2 based on serological, molecular, and technological techniques. Despite the availability of all these diagnostic techniques, a correct diagnosis of COVID-19 infection can only be established considering the test to be used, the type of sample to be analyzed, and the timing of the test itself. Therefore, it is necessary to perform the correct test, at the correct time in the correct biological sample [10, 11].

1.1. Vision Transformer

Vision Transformer [12] is a deep learning model, introduced in 2015. Vision transformer is based on transformer model [13], which totally relies on the self-attention mechanism. this is the only network that outperforms CNN [14]. Vision transformer is used for image classification [15–18], detection [19–21], segmentation [22–25], image enhancement [23,26], image generation [27], video processing [28,29], and 3D point cloud processing [30]. Moreover, vision transformer has been also used for medical imaging, it has been used for the classification of breast ultrasound images [17], and vision transformer has also been used for the COVID-19 detection from CXR and CT images [19, 21, 31–33]. Various methods have been used for the detection of COVID-19, like RT-PCR, rapid antigen and antibody-based tests, imaging, machine learning methods, and deep learning methods. These all have done a tremendous amount of work for the detection of COVID-19, which lead the world to tackle COVID-19. But still, there are required improvements for the detection of COVID-19 that lead us to better results for COVID-19 detection and classification. Therefore, in this study, a deep learning model is proposed based on a vision transformer for the automatic detection and classification of COVID-19. For this work, CXR images have been used to be trained the model. The model shows a better result in terms of accuracy, sensitivity, and specificity. Overall, the contributions of this paper can be summarized as follows:

- We proposed a framework that incorporates multi-headed attention to disclose more information. This mechanism leads to quantitatively diverse features to identify COVID-19 from CXR images.
- The proposed framework is evaluated against metrics like accuracy, precision, recall, and f1-score.
- Comparison with state-of-the-art CNN models shows superior performance.

COVID-19 is a critical disease that spread rapidly and can lead to death, especially for aged people or those whose immune system is weak. The COVID-19 detection and classification were harsh at the beginning but have been improved somehow, in order to make it more accurate, easy, and affordable. In this study, the deep learning method has been used to detect COVID-19 from CXR images that boost the result, in form of accuracy, struggle, and expenditure.

2. PROPOSED METHODOLOGY

2.1. Pre Processing

Pre-processing is the process of transferring raw data into useful data. We have taken the above datasets and applied some pre-processing methods in order to make them compatible with the model which leads to a fast and accurate result. The given datasets contain various size images which are large too and can be time-consuming and computationally expensive, for this purpose each image of the dataset has been resized to 100X100, in order to be best compatible, with the given data with the used model. In addition to this, the normalization took place to the range [0,1], because of the large variability

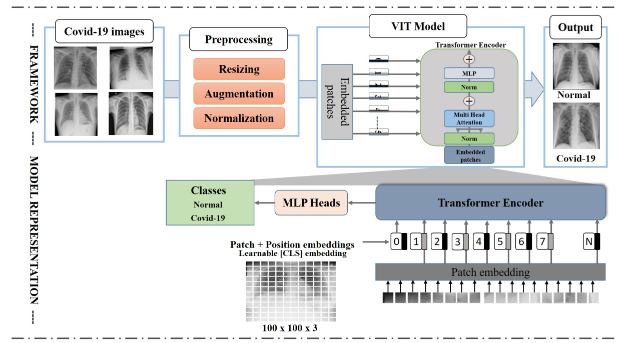

Fig. 1: Propsed Framework

in the appearance of images depending on different factors, like source acquisition. In addition to this, data augmentation has been applied in this work which is flipping both horizontally and vertically on each image in the existing dataset, the purpose of this augmentation is to introduce some new variations to the dataset. Random shuffling is a standard procedure in all machine learning pipelines, the classification of images is not an exception, and its purpose is to break possible biases during data preparation. In this project, we have shuffled the images randomly in order that picks images from different classes and labels them.

2.2. Convolutional Neural Network (CNN)

CNN is composed of several convolution layers which use learnable filters or kernels to extract features from images such as points, edges, textures, color, and shapes. Furthermore, a gradient decent-base optimizer is used to learn the appropriate filters, and CNN can capture spatial and temporal connections in an image. They hierarchically constructed high-level features from low-level features which help CNN to properly discriminate among the various object present in the image. CNN also shares its parameters. CNN reused its parameters (filters) to compute specific features in different positions of an image, which lead to the reduction of parameters. Convolution layers are commonly used activation functions that introduce non-linearity between layers, which abilities the network to capture the complicated relationship between the input features. Rectified linear unit (ReLU) is a commonly used activation function. In addition to this, it uses a pooling layer to reduce the size of feature representation as we propagate deeper into the network. For classification, in the final layer (fully connected layer) a function like softmax or sigmoid is used to generate the final result.

2.3. Transfer Learning

Transfer learning is the application learned from one problem and applying them to a new, similar problem to be solved, which is based on CNN. In this work, various pre-trained models have been adopted to be trained on both COVID-19 datasets. These models are AlexNet, VGG16, VGG19, Resnet50, and Resnet101 each of them has a different number of layers and parameters.

2.4. Vision Transformer (ViT)

For image classification, the Vision transformer directly applies to the sequence of image patches. The transformer design is originally followed by ViT as possible. The overview of the model is shown in figure 1. The standard transformer receives as input a 1D sequence of token embedding. To work with 2D images, we split the image $X \in R^{H \times W \times C}$ into a sequence of flattened 2D patches $X_p \in R^{N \times (P^2.C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which also provide as the effective input sequence length for the transformer. Constant latent vector size D was used by the transformer through all of its layers, so the patches were flattened and mapped to D dimensions with a trainable linear projection represent in equation (1). We refer to the output of this projection as the patch embedding.

$$z_0 = [X_c lass; X_p^1 \mathbf{E}; X_p^2 \mathbf{E}; ....; X_p^N \mathbf{E};] + \mathbf{E}_{pos}, \\ \mathbf{E}^{(P^2.C)\times D}, \mathbf{E}_{pos} \in R^{(N+1)\times D} \tag{1}$$

As the original BERT's token, a learnable embedding is applied to the sequence of embedded patches ($z_0^0 = X_{class}$), whose state at the output of the transformer encoder (z_L^0) serves as the image representation y represent in equation (4). Both during pre-training and fine-tuning, a classification head is attached to z_L^0. The classification head is implemented by an MLP with one hidden layer. Position embedding is added to the patch embedding to retain positional information. We use standard learnable 1D position embedding since we have not observed significant performance gains from using more advanced 2D-aware position embedding. The resulting sequence of embedding vectors serves as input to the encoder.

2.5. ViT Transformer Encoder

The transformer encoder [13] includes alternating layers of multi-headed self-attention layer (MSP) which concatenates all the attention outputs linearly to the right dimension. and multi-layer perceptron (MLP) blocks which contain two-layer with GELU represent in equations (2,3). In order to improve training time and performance, the layer-norm (LN) is applied before every block, and residual connections after every block [34, 35].

$$z_l' = MSA(LN(z_l - 1)) + z_l - 1, \qquad l = 1...L \tag{2}$$

$$z_l = MLP(LN(z_l' - 1)) + z_l', \qquad l = 1...L \tag{3}$$

$$y = LN(z_L^0) \tag{4}$$

3. EXPERIMENTAL

First, we will go through the experimental setup, then the datasets utilized in the model's training and assessment, and the evaluation matrix, ablation analysis, and real-time testing. Details are given in the following subsections.

3.1. Experimental Setup

All experiments were carried out in a Python 3.7-based virtual environment installed on a PC with the specifications of Windows 10 OS, having GTX GeForce TITAN 1070 graphic processing unit (GPU) with a memory of GB, the processor of intel ® X5560, and clock speed of 2.80GH. Further, different frameworks and libraries are utilized during training; the proposed framework is utilized for training as a back-end TensorFlow-GPU and a frontend Keras-GPU of versions 2.4 and 2.9, respectively. Moreover, we trained the proposed model on a mini-batch of size 32 for 100 iterations of epochs, which took almost two hours to complete the training of our proposed framework.

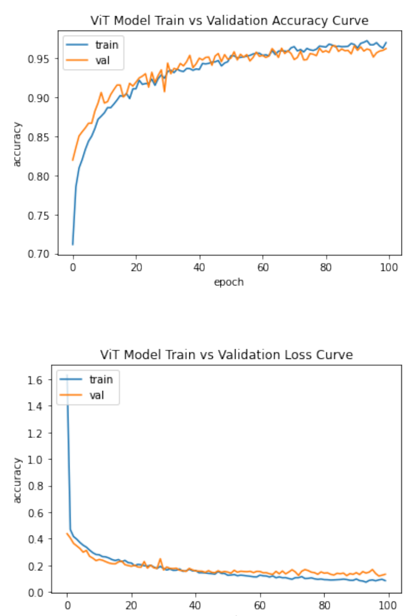

Fig. 2: Accuracies and losses of the proposed model.

3.2. Dataset

In this work two datasets have been used which are COVID-19-Dataset and COVID-19_Radiography_Dataset, both are publically available on [36] and [37]. The first dataset named Covid-19-Dataset contains X-ray and CT images of COVID and Non-COVID. This dataset had been augmented by the team who collected this dataset and generated 17099 X-ray and CT images. This dataset has two main folders, one is CT and another is X-ray, these are further divided into two subfolders of COVID and Non-COVID. The X-ray folder has 5500 of Non-COVID images shown in figure 4 and 4044 COVID images shown in figure 3. The CT folder has 2628 Non-COVID images and 5427 COVID images.

In addition to this, the second dataset named COVID-19_Radiography_Datase is collected by a team of researchers from Qatar university, Dhaka university, Doha and along with the collaboration of Pakistani students. This dataset contains 3616 COVID-19 images shown in figure 3 positives cases along with 10192 Normal images shown in figure 4, 6012 Lung Opacity, and 1345 Viral Pneumonia CXR images. For this proposed work we have only used COVID-19 and Normal images. The train part includes 70% of images along with 20% validation data and 10% testing data of the X-ray images.

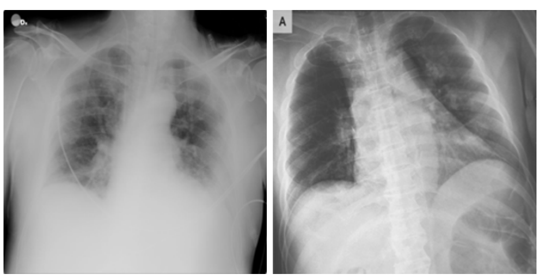

Fig. 3: COVID-19 Infected Images

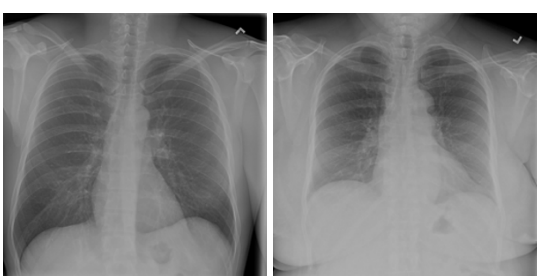

Fig. 4: Normal Images

3.3. Experimental Results

We applied various pre-trained models on each dataset and each model has some variation in the result in terms of metrics, accuracy, recall, precision, and f1-score on each dataset. Accuracy is the number of correctly predicted data points among all the predicted points.

$Accuracy = \frac{True_{pos}+True_{neg}}{True_{pos}+False_{pos}+True_{neg}+False_{neg}}$ -Precision is the ratio of true positive with anything that was predicted as positive.

$Precision = \frac{True_{pos}}{True_{pos}+False_{pos}}$ -Recall is the ratio of ture positive with anything that should have been predicted as positive.

$Recall = \frac{True_{pos}}{True_{pos}+False_{pos}}$ -F1-score is the harmonic mean of precision and recall.

$F1-Score = \frac{Precision X Recall}{2(Precision+Recall)}$ The result of each model for COVID-19-Dataset are given in the table 1 and COVID-19_Radiography_Dataset are given in the table 2.

Architecture	Accuracy%	Precision%	Recall%	F1-Score%
AlexNet [38]	70	99	29	44
VGG16 [39]	83	79	82	81
VGG19 [39]	86	86	84	85
Resnet50 [40]	65	60	55	57
Resnet101 [40]	85	88	79	84
ViT	88	85	87	86

Table 1: Result of Various Models on COVID-19-Dataset.

Architecture	Accuracy%	Precision%	Recall%	F1-Score%
AlexNet [38]	93	88	85	86
VGG16 [39]	26	26	58	42
VGG19 [39]	90	86	74	79
Resnet50 [40]	88	88	83	86
Resnet101 [40]	89	72	93	81
ViT	97	98	91	94

Table 2: Result of Various Models on COVID-19_Radiography_Dataset.

3.4. Ablation Study

In order to ensure that our proposed model is optimal, an ablation study had conducted on binary classification datasets. For this purpose, we conduct different experiments with variations in layer, dense layer, activation function, and dropout. First of all, we change the transformer layers from 8 to 6, and the accuracy of the model fall to 95%, and also checked on 10 layers which gives 96% with a little over fitting, so for this purpose, we leave it on 8 layers because that gives us high accuracy and non overfitted result. Moreover, we apply RELU instead of GELU which also gives 96% accuracy with a little overfitting, this shows that GELU is slightly more effective than RELU. In addition to this, we also add an extra MLP layer of function GELU with 0.3 dropouts, the output was overfitted, and fall the accuracy to 95%.

4. CONCLUSION

In this work, we proposed a framework for COVID-19 detection and classification from CXR images, various pre-trained models are used for comparative analysis in terms of performance. For the detection of COVID-19, we consider the [36] and [37] datasets which are publically available. Further, we applied different models and compare their results and parameters. Finally, we get a transformer model which has outperformed the pre-trained models in terms of accuracy, precision, recall, and f1-score. COVID-19 is a big challenge to the world. That dreadfully affect the health of people, live, and business. To overcome this problem the identification of the disease is crucial, to resolve this problem various methods have been used to detect the virus. With the advancement in technology and AI, the deep learning method has been widely used for the detection of COVID-19. The advancement of deep learning is enough accurate, inexpensive, and avoids the limitation of experts. In the future, our aim is to make our model more accurate by close to 100%. Furthermore, various other lung diseases are to be included and classify them with accurate results. In addition to this, we have made our project much easier that can be easily adopted by health workers. .

5. REFERENCES

[1] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[2] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al., "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[3] Claudio Márcio Amaral de Oliveira Lima, "Information about the new coronavirus disease (covid-19)," 2020.

[4] Thomas Struyf, Jonathan J Deeks, Jacqueline Dinnes, Yemisi Takwoingi, Clare Davenport, Mariska Mg Leeflang, René Spijker, Lotty Hooft, Devy Emperador, Julie Domen, et al., "Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19," *Cochrane Database of Systematic Reviews*, , no. 2, 2021.

[5] Jiaqiang Liao, Shibing Fan, Jing Chen, Jianglin Wu, Shunqing Xu, Yuming Guo, Chunhui Li, Xianxiang Zhang, Chuansha Wu, Huaming Mou, et al., "Epidemiological and clinical characteristics of covid-19 in adolescents and young adults," *The Innovation*, vol. 1, no. 1, pp. 100001, 2020.

[6] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al., "Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia," *New England journal of medicine*, 2020.

[7] Jian-Long He, Lin Luo, Zhen-Dong Luo, Jian-Xun Lyu, Ming-Yen Ng, Xin-Ping Shen, and Zhibo Wen, "Diagnostic performance between ct and initial real-time rt-pcr for clinically suspected 2019 coronavirus disease (covid-19) patients outside wuhan, china," *Respiratory medicine*, vol. 168, pp. 105980, 2020.

[8] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[10] Raquel M Martinez, "Clinical samples for sars-cov-2 detection: review of the early literature," *Clinical Microbiology Newsletter*, vol. 42, no. 15, pp. 121–127, 2020.

[11] G Gopal Rao, Ashwini Agarwal, and Deepak Batura, "Testing times in coronavirus disease (covid-19): a tale of two nations," *Medical Journal, Armed Forces India*, vol. 76, no. 3, pp. 243, 2020.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Mohib Ullah, Mohammed Ahmed Kedir, and Faouzi Alaya Cheikh, "Hand-crafted vs deep features: A quantitative study of pedestrian appearance model," in *2018 Colour and Visual Computing Symposium (CVCS)*. IEEE, 2018, pp. 1–6.

[15] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 357–366.

[16] Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, pp. 516, 2021.

[17] Behnaz Gheflati and Hassan Rivaz, "Vision transformer for classification of breast ultrasound images," *arXiv preprint arXiv:2110.14731*, 2021.

[18] Ahmed Kedir, Mohib Ullah, and Jacob Renzo Bauer, "Spectranet: A deep model for skin oxygenation measurement from multi-spectral data," *Electronic Imaging*, vol. 2020, no. 15, pp. 83–1, 2020.

[19] Debaditya Shome, T Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar, "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, pp. 11086, 2021.

[20] Habib Ullah, Ahmed B Altamimi, Muhammad Uzair, and Mohib Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[21] Mohamad Mahmoud Al Rahhal, Yakoub Bazi, Rami M Jomaa, Ahmad AlShibli, Naif Alajlan, Mohamed Lamine Mekhalfi, and Farid Melgani, "Covid-19 detection in ct/x-ray imagery using vision transformers," *Journal of Personalized Medicine*, vol. 12, no. 2, pp. 310, 2022.

[22] Habib Ullah, Mohib Ullah, and Muhammad Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, vol. 31, pp. 7317–7333, 2019.

[23] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12299–12310.

[24] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[25] Ullah Mohib Nordbø Øyvind Mamadou, Keita and Faouzi Alaya Cheikh, "Multi-encoder convolution block attention model for binary segmentation," in *IEEE 19th International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2022.

[26] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5791–5800.

[27] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran, "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.

[28] Yanhong Zeng, Jianlong Fu, and Hongyang Chao, "Learning joint spatial-temporal transformations for video inpainting," in *European Conference on Computer Vision*. Springer, 2020, pp. 528–543.

[29] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[30] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16259–16268.

[31] Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla, and AP Prathosh, "xvitcos: Explainable vision transformer based covid-19 screening using radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–10, 2021.

[32] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye, "Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus," *arXiv preprint arXiv:2103.07055*, 2021.

[33] Ahmed Mohammed, Congcong Wang, Meng Zhao, Mohib Ullah, Rabia Naseem, Hao Wang, Marius Pedersen, and Faouzi Alaya Cheikh, "Weakly-supervised network for detection of covid-19 in chest ct scans," *IEEE Access*, vol. 8, pp. 155987–156000, 2020.

[34] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.

[35] Alexei Baevski and Michael Auli, "Adaptive input representations for neural language modeling," *arXiv preprint arXiv:1809.10853*, 2018.

[36] Walid El-Shafai, ""extensive covid-19 x-ray and ct chest images," 12 June 2020.

[37] Preet. Viradiya, "Preet. "covid-19 radiography dataset." kaggle, 22 may 2021,," *https://www.kaggle.com/preetviradiya/covid19-radiography-dataset.*

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[39] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.