COVID19 to Pneumonia: Multi Region Lung Severity Classification using CNN Transformer Position-Aware Feature Encoding Network

Jong Bub Lee[1 [0009−0002−0586−3262]], Jung Soo Kim[2,3], and Hyun Gyu Lee[1,3,*]

[1] Department of Electrical and Computer Engineering, Inha University, Republic of Korea
[2] Division of Critical Care Medicine, Department of Hospital Medicine, Inha University Hospital, Republic of Korea
[3] College of Medicine, Inha University, Republic of Korea
bub3690@inha.edu, acecloer31@gmail.com, hglee@inha.ac.kr

Abstract. This study investigates utilizing chest X-ray (CXR) data from COVID-19 patients for classifying pneumonia severity, aiming to enhance prediction accuracy in COVID-19 datasets and achieve robust classification across diverse pneumonia cases. A novel CNN-Transformer hybrid network has been developed, leveraging position-aware features and Region Shared MLPs for integrating lung region information. This improves adaptability to different spatial resolutions and scores, addressing the subjectivity of severity assessment due to unclear clinical measurements. The model shows significant improvement in pneumonia severity classification for both COVID-19 and heterogeneous pneumonia datasets. Its adaptable structure allows seamless integration with various backbone models, leading to continuous performance improvement and potential clinical applications, particularly in intensive care units.

Keywords: Position aware feature · Weakly supervised learning · Portability on heterogeneous dataset · Transformer

1 Introduction

During the COVID-19 pandemic, many studies have been proposed to use retrospective analyses to build a database of COVID-19 images and use learning-based approaches based on them [1–4]. In particular, studies using chest x-rays (CXRs) have been effective in shaping treatment strategies for patients with COVID-19 pneumonia in intensive care units(ICU) [1].

After the end of the pandemic, these assessment methods will still be needed in ICUs with severe pneumonia patients. However, despite the considerable accumulation of data on COVID-19 pneumonia patients, the lack of clear clinical measures and the subjectivity of raters lead to label noise, which significantly degrades the performance of learning-based algorithms [3]. In addition, the lack of standardized criteria for staging and geographic boundaries of severity complicates the integration of data from different datasets. Therefore, a deeper ex-

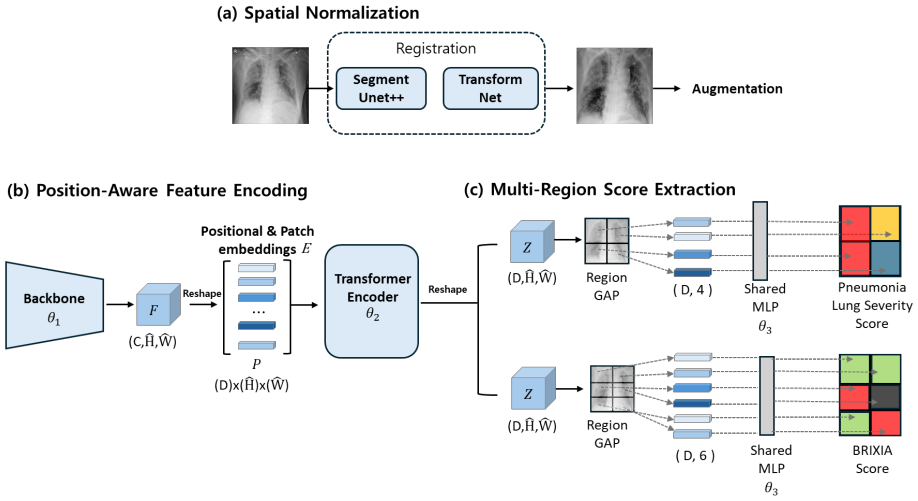

Fig. 1. Workflow and structure of multi-region CNN Transformer hybrid networks with position-aware feature encoding for lung severity classification. (a) Spatial Normalization is a preprocess stage which consists of segmentation, transform, and augmentation. (b) Position-Aware Feature Encoding is performed with CNN, patch embedding, positional embedding, and transformer. (c) Multi-Region Score Extraction is a process of extracting ROIs for each task and computing region-specific labels by region shared MLPs.

ploration of the portability of scores between datasets with different criteria is needed.

Goal In this study, we explore the potential value of chest X-ray (CXR) data from COVID-19 patients for addressing the challenge of classifying pneumonia severity. Our goal is to improve prediction performance on COVID-19 datasets and achieve robust results in severity classification on heterogeneous pneumonia data by developing architectures specifically for multi-region classification.

Solution The patterns of 'haziness' associated with pneumonia severity can manifest as characteristics of other structures in different lung regions. In consideration of this, we have adopted a CNN-Transformer hybrid structure to incorporate the positional information of lung regions. Furthermore, scoring pneumonia severity is a weakly supervised learning problem due to the lack of clinically defined rules, making the selection of regions subjective based on the dataset and specific tasks [3]. For this reason, a flexible network without spatial constraints, suitable for various downstream tasks, is necessary. Our approach utilizes Region of Interest (ROI) pooling followed by Region Shared Multi-Layer Perceptrons (MLPs), allowing for the substitution of MLPs in downstream tasks to achieve both flexibility and superior performance.

Contribution We have achieved two main contributions: **i)** We proposed a CNN-Transformer hybrid method for predicting lung severity scores. This method integrates lung region information and can be flexibly applied to various

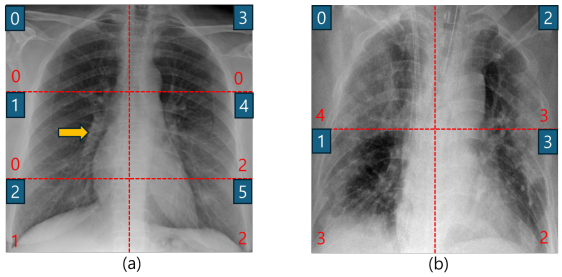

Fig. 2. Normalized CXR image of two lung severity datasets. (a) BrixIA COVID-19 dataset, (b) Inha ICU pneumonia severity dataset; the number in the blue box depict position index, red font number depict severity label.

downstream tasks. Additionally, this adaptable structure can be seamlessly integrated with various backbones. **ii)** We explored the potential utility of COVID-19 data to enhance the performance of pneumonia severity classification models. As a result, we significantly improved classification performance not only in the COVID-19 dataset BrixIA but also in pneumonia data from Inha University Hospital ICU.

The code is publicly available at
https://github.com/bub3690/Multi-Region-Lung-Severity-PAFE

2 Method

The goal of the proposed approach is for a single model to perform learning on labels with different regions and to also incorporate the positional information of each location in the lungs. To achieve this goal, we have defined the method in three stages as depicted in Fig. 1. In the Spatial Normalization(SN) stage, we normalized the lung positions of all images. In the Position-Aware Feature Encoding(PAFE) stage, we extracted local features, applied patch embedding, and added positional embedding to the normalized positions. Then we used the transformer to reflect the relationships between local features. In the Multi-Region Score Extraction process, we dynamically extracted ROIs according to the labels and computed region-specific labels.

2.1 Spatial Normalization

The lung severity classification can be considered weakly supervised learning because it involves predicting the labeling of lung regions, as shown in Fig. 2, without explicit labels for lung pixel-level segmentation. To solve this problem, BrixIA Score network(BS-Net)[3] used Spatial Transformer network(STN)[5] to align features in the CNN. However, because STN only execute spatial transformations, they lack the capacity, in most cases, to align the feature maps of a transformed image with those of its original. Consequently, STN is incapable

of ensuring invariance during the transformation of CNN feature maps[6]. And also it's not easy to apply on complex networks like DenseNet[7]. Instead, we used pre-trained STN to normalize input images. The STN was initially trained to compute affine transformation matrices using CXR segmentation maps as input, where the target image was the CXR lung segmentation, and the moving image was its augmented counterpart. This pre-training allows the STN to effectively predict affine transformation matrices when presented with new CXR images[3]. The normalization process, in Fig. 1.a, involves using UNET++ to extract masks[8], and then using the masks as input values for the Spatial Transformer Network to predict affine transformation matrices for enlargement and alignment. Through this explicit normalization process, it becomes possible to determine the positions of features outputted by CNNs, making it easier to designate ROIs.

2.2 Position-Aware Feature Encoding

In the position-aware feature encoding of the proposed method, the normalized input image $x \in \mathbb{R}^{H \times W}$ is first passed through the backbone θ_1, resulting in the extraction of local feature $F \in \mathbb{R}^{C \times \hat{H} \times \hat{W}}$.

$$F = \beta_{\theta_1}(x) \tag{1}$$

In Fig. 2.a, as indicated by the arrows, lung images exhibit various internal structures depending on the region. Therefore, relying solely on local feature F for predictions can lead to prediction errors. Thus, a process is conducted to incorporate positional information into the local feature F. The local feature F is reshaped as shown in Equation (2). Perform patch embedding E and 1x1 convolution as shown in Equation (3), then combine with positional embedding E_{pos}. Patch embedding kernel E, positional embedding E_{pos} serve as learnable parameters, allowing the model to learn the local lung region from normalized positions autonomously.

$$F = [f_1, f_2, f_3, \cdots, f_l], \quad l = \hat{H} \times \hat{W} \tag{2}$$

$$P = [f_1 E, f_2 E, f_3 E, \cdots, f_l E] + E_{pos}, \quad E \in \mathbb{R}^{(1 \cdot C) \times D}, E_{pos} \in \mathbb{R}^{(l) \times D} \tag{3}$$

Furthermore, ROI pooling alone may not adequately represent surrounding information. Since diseases can metastasize within the same lung, incorporating surrounding information can lead to better performance. Therefore, as shown in Equation (4), we input a patch set P that reflects positional information into the Transformer encoder θ_2. Unlike conventional ViT and BERT models, we do not add a Class token to the patch set [9–11]. This is because, unlike traditional image classification tasks, we operate in a weakly supervised setting where labels exist for each region. In this study, only one Transformer Encoder is used.

$$Z = \tau_{\theta_2}(P) \tag{4}$$

Table 1. Comparison of the severity classification performances and Abaltion study on BrixIA dataset. Our method with Resnet34 achived SOTA Result on BrixIA. The Consensus Test set underwent consensus among four labelers. * : Published score.

Method	**Train ACC**	**Test ACC**	**Consensus Test ACC (MAE)**
CXR Clip Resnet50[14]	**95.1**	51.3	56.2 (0.50)
BS-NET Ensemble*	-	-	57.1 (0.42)
Ours	61.0	**58.5**	**67.1 (0.35)**
Resnet18 w/ SN	58.9	56.9	63.0 (0.39)
w/ PAFE	58.5	58.7	65.7 (0.37)
Resnet34 w/ SN	61.4	58.7	63.0 (0.40)
w/ PAFE	**61.0**	58.5	**67.1 (0.35)**
Resnet50 w/ SN	52.1	53.3	56.4 (0.52)
w/ PAFE	51.5	49.1	55.0 (0.55)

Through this process, feature embedding, $Z \in \mathbb{R}^{D \times l}$ is computed. By utilizing both local features through CNN and long-range relationships through Transformer, the advantages of the hybrid structure are expected to be beneficial for modeling lung severity.

2.3 Multi-Region Score Extraction

To make it scalable for labels with different regions, ROI pooling should be performed without additional parameters or separate models. Thanks to the Spatial Normalization process, feature extraction for the respective label is achieved through Region pooling without the need for a separate segmentation network. Region Pooling divides the feature embedding $Z \in \mathbb{R}^{D \times \hat{H} \times \hat{W}}$ into intended areas of the spatial dimension($\hat{H} \times \hat{W}$)[3]. ROI Pooling may not accurately reflect vertical separation as intended by the labeler. Therefore, features learned through the attention process can be more flexibly incorporated. After ROI Pooling, scores are computed using Region Shared MLP_{θ_3}. By utilizing a Region Shared MLP, it is expected to effectively capture global features even with diverse regions in future downstream tasks.

3 Experimental Result

3.1 Dataset and Training Details

The model was trained and evaluated on the publicly available BrixIA dataset. Additionally, the pneumonia severity dataset was used to assess the portability of the model trained on the COVID-19 dataset. This dataset includes patients with acute respiratory failure due to pneumonia and was collected from the

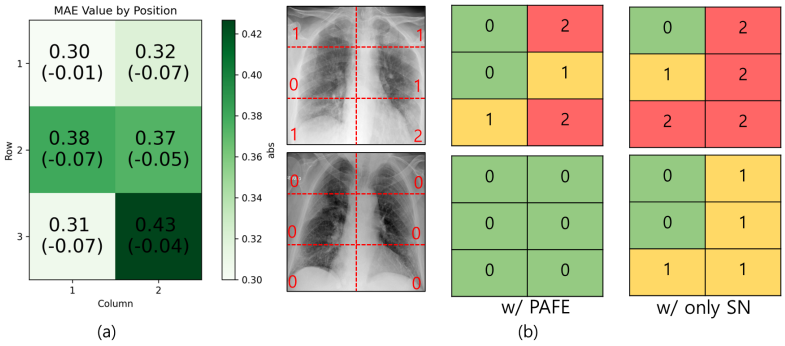

Fig. 3. Results on the consensus test set of the BrixIA dataset. (a) MAE per position for RenNet34 with PAFE method, (b) Sample prediction result. Values within parentheses indicate the performance improvement achieved after Position-Aware Feature Encoding compared to with only Spatial Normalization(SN). Red text within the CXR images represents the labels.

ICU at Inha University Hospital. The BrixIA dataset comprises 4,695 samples with six regions within a score range of 4 grades. The Inha ICU pneumonia severity dataset consists of 611 samples with four regions within a score range of 5 grades. Therefore, the situation with the pneumonia severity classification dataset is more challenging, and baseline training does not converge without fine-tuning. Particularly, due to subjectivity in labeling and label noise in both datasets, training is extremely difficult. To mitigate memorization effects and overfitting, a slow learning rate of 10^{-4} and SGD Optimizer with momentum 0.9 were used for 300 epochs[12]. Additionally, contrast/brightness distortion, random affine transform, sharpness, and rotation augmentation were applied.

3.2 Performance evaluation for BrixIA

To evaluate the severity prediction performance, we used BrixIA's Consensus Test, which uses the voting results of four doctors and contains relatively less label noise, and a test set not included in BrixIA's training data, which contains more label noise. Therefore, the consensus test metric is a more reliable dataset than the validation/test set.

In Table 1, our approach demonstrated the best performance when using ResNet34 as the backbone with a hybrid structure. BS-NET serves as a baseline model by incorporating STN and Feature Pyramid Network (FPN) into the ResNet18 backbone and adding a RetinaNet classifier during ROI pooling [3, 19]. When our method with spatial normalization was applied to ResNet-18 and ResNet 34, we found a 5.9 improvement in accuracy and a 0.025 reduction in MAE compared to BS-NET, suggesting that the region shared MLP provided assistance in generalization performance compared to the RetinaNet classifier and FPN layer. Moreover, when the proposed method is applied along with

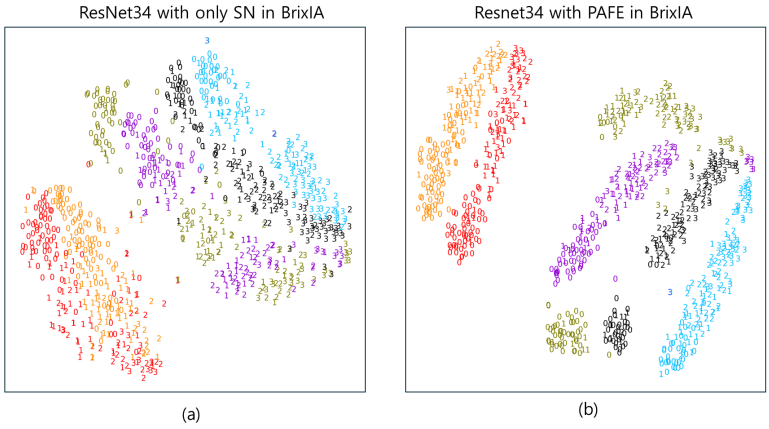

Fig. 4. T-SNE visualiztion of embedding on BrixIA consensnus testset (a) Visualization of Resnet34 with only Spatial Normalization, (b) Resnet34 with Position-Aware Feature Encoding. Colors represent 6 positions of lung and numbers for labels.

spatial normalization, the accuracy is improved by 10.0 and the MAE is reduced by 0.06 compared to BS-NET. On the consensus test set, the labelers had an MAE of 0.528[3], indicating that our method was able to find better consensus on the training dataset despite the presence of larger label noise in the training data.

ResNet50 showed overall poor performance due to the insufficient amount of training data used in the experiments compared to the complexity of the model. Also, CXR-Clip showed a tendency of overfitting and underperformance compared to ResNet50 pre-trained on ImageNet[13, 14, 18], even though it was trained on a large chest X-ray dataset with language-image pre-training[15–17]. Many current state-of-the-art methods use large-scale models for training, which makes them unsuitable for small and medium-sized datasets such as severity classification.

The position analysis of the proposed method with ResNet34 on the consensus test set also shows a reduction in MAE for all positions (see Fig. 3.a). As shown in Fig. 3.b, we can see that the tendency to overestimate the severity of pneumonia at certain locations according to the 'haziness' pattern is reduced in the proposed method as opposed to the comparison method.

Additionally, the t-SNE embeddings of the features at each position demonstrated more clustering in our proposed method[20], indicating that the PAFE was effectively trained to more accurately represent the 'haziness' features at each position(See Fig. 4).

3.3 Portability test on Pneumonia Severity Result

We evaluated the model portability from the COVID-19 dataset (BrixIA) to a heterogeneous pneumonia severity dataset (Inha ICU pneumonia severity dataset),

Table 2. Portability test on Inha ICU pneumonia severity dataset. The metric is Accuracy on Test set. The values in parentheses represent the improved accuracy of the methods that achieved the highest performance using the PAFE.

Method	Scratch	Fine-tuning	Linear-probing
Resnet18 w/ PAFE	27.0	45.3	**51.4** (+5.5)
Resnet34 w/ PAFE	**32.9** (+0.6)	**52.8** (+2.8)	51.2
Resnet50 w/ PAFE	29.7	36.4	29.7

which has differences in the division of regions and severity criteria, as shown in Table 2. Scratch requires fine-tuning an Imagenet pretrained model, Fine-tuning involves retraining all model parameters after training on the BrixIA dataset, and Linear-probing [21] requires freezing all parameters except for the linear layer. The only modifications needed in the model are creating a new Shared MLP and adjusting the region pooling scope. When trained on Scratch, the model achieves an average test accuracy of 26.3%, indicating that it struggles to classify most of the situations within the 5 classes. On fine-tuning, we observed an average improvement of approximately 14%, indicating notable learning progress within the model. In particular, ResNet34 performed the best with a 20% increase in accuracy, which tended to match the BrixIA dataset, suggesting that the PAFE works well in transfer learning scenarios. The results of Linear-probing revealed that sufficiently good embedding were learned when pre-training, with an average difference of 2% compared to fine-tuning. As a result, the performance of our model on the Inha ICU pneumonia severity dataset was comparable to that on BrixIA, showing that the PAFE has effective portability across heterogeneous pneumonia datasets.

4 Conclusion

We proposed a CNN-Transformer hybrid network for multi region lung severity classification. By leveraging attention mechanisms on position-aware features and region-shared MLPs, this model effectively integrates lung region information, enabling easy incorporation into various downstream tasks with diverse label spatial resolutions and offering enhanced flexibility and adaptability. As a result, the model shows significant improvements on both COVID-19 pneumonia datasets and heterogeneous pneumonia datasets. In addition, the structure of the adaptive transformer we proposed can be seamlessly combined with different backbones, allowing us to continuously improve the performance of the model through the evolution of backbones suitable for small and medium-sized data. Our next goals are to demonstrate further performance improvements in combination with different backbone CNNs, and to provide useful clinical applications through time series analysis of ICU patients.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No. 2022-0-00641, XVoice: Multi-Modal Voice Meta Learning], [No. RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1F1A1071574).

Disclosure of Interests. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Cohen, Joseph Paul, et al. “Predicting COVID-19 pneumonia severity on chest x-ray with deep learning." Cureus 12.7 (2020).
2. Rubin, Geoffrey D., et al. “The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society." Radiology 296.1 (2020): 172-180.
3. Signoroni, Alberto, et al. “BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset." Medical Image Analysis 71 (2021): 102046.
4. Toussie, Danielle, et al. “Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19." Radiology 297.1 (2020): E197-E206.
5. Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. “Spatial transformer networks." Advances in neural information processing systems 28 (2015).
6. Finnveden, Lukas, Ylva Jansson, and Tony Lindeberg. “Understanding when spatial transformer networks do not support invariance, and what to do about it." 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.
7. Huang, Gao, et al. “Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
8. Zhou, Zongwei, et al. “Unet++: A nested u-net architecture for medical image segmentation." Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018.
9. Dosovitskiy, Alexey, et al. “An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
10. Vaswani, Ashish, et al. “Attention is all you need." Advances in neural information processing systems 30 (2017).
11. Devlin, Jacob, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
12. Rolnick, David, et al. “Deep learning is robust to massive label noise." arXiv preprint arXiv:1705.10694 (2017).
13. Radford, Alec, et al. “Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
14. You, Kihyun, et al. “Cxr-clip: Toward large scale chest x-ray language-image pre-training." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.

15. Johnson, Alistair EW, et al. “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific data 6.1 (2019): 317.
16. Irvin, Jeremy, et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
17. Wang, Xiaosong, et al. “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
18. He, Kaiming, et al. “Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
19. Lin, Tsung-Yi, et al. “Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.
20. Van der Maaten, Laurens, and Geoffrey Hinton. “Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).
21. Chen, Ting, et al. “A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.