Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing 125 (2022) 109111

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

FCF: Feature complement fusion network for detecting COVID-19 through CT scan images

Shu Liang [a], Rencan Nie [a,b,*], Jinde Cao [c,d], Xue Wang [a], Gucheng Zhang [a]

[a] School of Information Science and Engineering, Yunnan University, Kunming, 650500, Yunnan, China
[b] School of Automation, Southeast University, Nanjing, 210096, Jiangsu, China
[c] School of Mathematics, Southeast University, Nanjing, 210096, Jiangsu, China
[d] Yonsei Frontier Lab, Yonsei University, Seoul, 03722, South Korea

ARTICLE INFO

Article history:
Received 5 January 2022
Received in revised form 12 May 2022
Accepted 26 May 2022
Available online 6 June 2022

Keywords:
COVID-19 detecting
Deep Learning
Feature complement fusion
Weakly supervised learning

ABSTRACT

COVID-19 spreads and contracts people rapidly, to diagnose this disease accurately and timely is essential for quarantine and medical treatment. RT-PCR plays a crucial role in diagnosing the COVID-19, whereas computed tomography (CT) delivers a faster result when combining artificial assistance. Developing a Deep Learning classification model for detecting the COVID-19 through CT images is conducive to assisting doctors in consultation. We proposed a feature complement fusion network (FCF) for detecting COVID-19 through lung CT scan images. This framework can extract both local features and global features by CNN extractor and ViT extractor severally, which successfully complement the deficiency problem of the receptive field of the other. Due to the attention mechanism in our designed feature complement Transformer (FCT), extracted local and global feature embeddings achieve a better representation. We combined a supervised with a weakly supervised strategy to train our model, which can promote CNN to guide the VIT to converge faster. Finally, we got a 99.34% accuracy on our test set, which surpasses the current state-of-art popular classification model. Moreover, this proposed structure can easily extend to other classification tasks when changing other proper extractors.

1. Introduction

Since Dec 2019, a coronavirus storm turned out in one night, and outbreaking spread all over the world rapidly. This type of coronavirus called Coronavirus Disease 2019 (COVID-19) became an ongoing epidemic engulfing the globe. COVID-19 was firstly reported in Wuhan, Hubei province of China [1]. Whereas, extensive transportation and population mobility were still operating during the Chinese Spring Festival in China, which aggravated the epidemic when people were unaware of its human-to-human transmission trait [2]. Some reports have indicated that this type of coronavirus has a low potential for sustained community transmission [3]. However, its spread speed is far more what people can imagine. According to the report, this COVID-19 epidemic follows an exponential growth [2,3]. And up to November 2021, over 256,966,237 people have been detected infection, and deaths-cumulative have achieved 5,151,643 total [4]. This disease is highly contagious and may lead to acute respiratory distress or multiple organ failure in severe cases [5]. People infected with COVID-19 act fever and cough as the most common symptoms [6]. But these cannot examine the COVID-19 as the principal argument. Recent theoretical developments have revealed that the nearest clinical diagnosis regards the real-time reverse transcriptase-polymerase chain reaction (RT-PCR) as the golden standard to diagnose COVID-19 [7]. However, a scientifically synthetic diagnosis must combine Chest X-ray (CXR) or computed tomography (CT) result on patients. Both results show the same characteristic in COVID-19 symptoms, demonstrating bilateral peripheral consolidation in their lungs [8]. CXR and CT contribute to early diagnosing the viral disease, although nucleic acid detection with RT-PCR remains the standard Ref. [9]. Baseline CXR performed a sensitivity of 69% compared to 91% of RT-PCR [8]. Nevertheless, the CT scan receives a better sensitivity in contrast to the RT-PCR [9,10]. Ground-glass opacity (GGO) and consolidation are two principal higher confidence characteristics for diagnosis, which display more distinctively on the CT scan than the CXR [8,11]. Therefore CT scan displays a unicorn role in COVID-19 diagnose mission.

With the development of computer science nowadays, applications based on the deep learning (DL) approach are extensively used around our life, especially in image processing for feature extraction purposes. Like object detection and identification

* Corresponding author.
E-mail address: rcnie@ynu.edu.cn (R. Nie).

https://doi.org/10.1016/j.asoc.2022.109111
1568-4946/

based on local calculate service, or deploys feature extraction models on users and edges relying on cloud computing so-called federated learning [12].

And numerous industries began to add the DL technique to combine their original design. Convolutional neural network (CNN) especially shows its powerful ability in feature extraction work, which has dominated the DL field in recent decades. This type of network has even influenced the field of Clinical Medicine Science because its tremendous ability has been adopted widely in finding small details people overlooked [13]. Such as infection diagnosis: CNN assists the doctor in making the decision more precisely, which can focus on details people may neglect. It will deliver a prediction on the infection category or locate the position of the lesion [14–16]; lesion and organ segmentation: doctors once need segment the lesion and organ manually until CNN was involved in and support doctors to segment their interest part automatically for the succeeding diagnose [17–19]; multimodal medical image fusion: image imaging from different apparatus stores various information, whereas CNN fused images can reduce redundant data and preserve the essential data from both source images [20–22].

Nearly, Vision Transformer (ViT) appeared in the computer vision (CV) field, which migrates the algorithm used in natural language processing (NLP). This action impacts the status dominated by CNN and exerts a profound influence on CV researchers. In the subsequent time, ViT presented, researchers applied this structure to explore more facet applications. Like U-Transformer does a great framework on the task of complex organ segmentation [23]. And RTMIC can caption the CXR image automatically and generate a medical diagnosis [24].

A massive of methods of DL assist diagnosis occurred in this anti-COVID storm. Inspired by those works mentioned above, we further put forward our framework. In contrast to existing methods, the main contribution of our approach can be listed as follows:

- Different from CNN-based local feature extraction or ViT-based global feature extraction, our method employs two ones to extract local and global features of CT, respectively, and further integrated these features to give more effective discrimination for COVID-19.
- We construct a feature fusion block named FCT to fuse the local and global features. Due to the usage of the Transformer, our FCT can produce an embedding vector resulting from attention mechanism to present better feature representation.
- We designed a hybrid loss combined supervised learning with a weakly supervised learning strategy training our model efficiently to boost ViT structure converge faster in our proposed FCF.
- Our proposed FCF defeats other models on the benchmark and provides a reliable pre-diagnose to assist clinic surgeons doing the final diagnosis.

The rest of the paper is organized as follows. In Section 2, we introduced the recent feature extraction works and feature fusion works. Section 3 included details of our algorithm methods. Dataset choice, training details, and extended experiments are arranged in Section 4. The conclusion is finally given in Section 5.

2. Related work

2.1. Feature extractors

CNN has been used widely in the DL involved task for feature extraction since VGG [25] opened up the era of the deep neural network. Subsequently, add identity mapping from the shallower layer into the deep layer first appeared in ResNet [26]. It was broadly adopted to maintain the feature in CNN encoders helpfully and overcome the gradient vanish in the training process [27]. Another framework to efficiently reuse features is DenseNet [28]. This novel approach leverage violent concatenate operation to link all features from the front layer to the current layer. It further reduces the parameters and accelerates the propagation efficiency in the deep neural network. Because of the development of the neural network, artificial intelligence for the COVID-19 detecting task joins following this popular stream as well. Yujin Oh et al. exploit the feature reuse by dense architecture to receive a clear edge result on COVID-19 CT lung segmentation [29]. People began to shift their attention to researching what information should be most careful when extracting information from images meets its bottleneck. For this reason, attention-based algorithms become embedded in blocks to lead crucial information into a better representation [30,31]. Attention mechanisms combined with the residual block, which emerges henceforth, further activate to focus and preserve more on core information importance [32,33]. But in recent years, a popular model stemmed from the natural language processing (NLP) domain called Transformer transplant into the computer vision domain successfully [34]. This kind of model named Vision Transformer (ViT) [35] is a fully-attention structure and inherent the conspicuous characteristic of BERT [36], which applies a token ahead of the embedding for final decision. However, the original ViT structure is hard to train economically without a large base dataset supported even if it surpasses the remarkable performance created by CNN [37]. And the converge situation of the Transformer model heavily depends on the training batch and multi-parallel GPUs [38]. Consequently, people began to add absolute position attribution on tokens, a different approach from the position embedding (PE). T2T-ViT compresses the tokens by re-structurization them like input images [37]. This approach provides a piece of new position information between tokens. Besides, the NesT creates tokens by block aggregation service to contribute sturdy position information. Besides, the NesT also emphasizes the importance of the absolute position by deploying block aggregation service on tokens [39]. In addition, tokens in the Swin-Transformer serve as different scales for acquiring extra position information [40]. These optimizations of ViT establish closer connections on position information, which made its long-range relationship characteristics involved orderly.

2.2. Feature fusion

Vary embeddings from the feature extractor represent different detailed information learned by the network. UNet fuse the same scale feature in the encoder and decoder network to reinforce the context feature [41]. Whereas feature pyramid network (FPN) fuse the different scale feature embeddings in the encoding process to acquire a better feature representation [42]. Another way is utilizing a fully-connected layer to process the concatenated embeddings, which also can conserve essential instances of them all [43]. In [44], Chao et al. encode multi-type data and concatenate the representation embeddings for synthesis prediction on ICU admission through the random forest. Anunay Gupta et al. proposed InstaCovNet obtains a more reliable accuracy performance by associating features from five different feature extractors together for the final decision [45].

3. Method

3.1. Motivation

Current works only applied CNN or ViT for feature extraction. But CNN is only good at capturing local features, which cannot

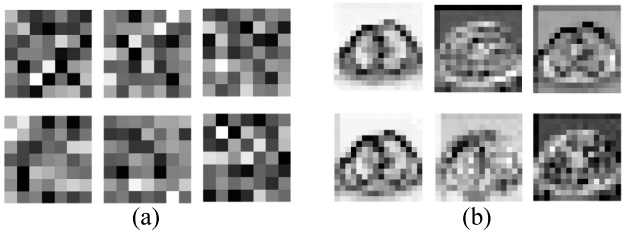

Fig. 1. Respective Field: (a) respective field of CNN, (b) respective field of ViT.

extract global features causes of its small convolution kernels. In contrast, ViT benefits from its particular way to get embeddings from patches, which obtain a large receptive field but lack of detail information. So, we consider that integrating the feature from CNN and ViT can complement the feature deficiency of each other. The respective field of CNN and ViT were visualized in Fig. 1 (a) and Fig. 1(b) severely. According to the trait of their respective field, we define features from CNN as local features and features from ViT as global features. In Fig. 1(a), we have already cannot recognize what is meaning in the filter. Whereas we can clearly be aware of an outline with some details of a lung was shown in the filter as Fig. 1(b).

And we also consider that merging local feature embedding and global feature embedding can complement the data deficiency of each other. But operating a linear layer only to process the concatenated embeddings cannot emphasize the complement information specifically. We thus proposed a feature complement Transformer (FCT) to strengthen the expression of the concatenated feature embeddings.

Due to our architecture is consists of a ViT, and the full-Transformer model needs a large dataset to train. Otherwise, it will exhibit an inferior performance compared to the identical size CNN model without pretraining [37]. And CT scan images are grayscale images, which cannot transfer learning on the three-channel input pretraining model. It also converges slowly when training on a single GPU without transfer learning. To resolve this problem mentioned above, we provided a weakly supervised module to guide the ViT extractor gaining better performance.

3.2. Network

According to our motivation in Section 3.1, we proposed a feature complement fusion network (FCF) contributing to providing a brand new view to get various aspect feature embeddings. And further combining the CNN and ViT hybrid embeddings integrated for prediction efficiently. The overview of our FCF structure is shown in Fig. 2. Pseudocodes of the whole FCF architecture are shown in 1. Specifically, we separate this model into three parts:

- Feature extractor: this module consists of a CNN branch and a parallel ViT branch to extract CT images features into feature embeddings.
- Weakly supervised module: it produces a weak label to weakly supervise the ViT extractor, which accelerates its convergence rate meanwhile.
- Feature complement fusion block: features come from the feature extraction part gathered here, then make a complement fusion for the final decision.

3.2.1. Feature extractor

Features extracted from the feature extractor have to involve local information and global information from the object. We select CNN and ViT to extract local information and global information severally.

CNN Extractor In our model, the CNN extractor is based on a backbone network, e.g., the ResNet [26] and DenseNet [28]. We will discuss how to select a specific one in the "Ablation Study" and "Compared to Classic SOTA" sections. Where to balance the compute efficiency and the performance, we finally adopt ResNet-50 as an extractor to extract local features with spatial information consisting of abundant texture and edge [37,46] in the CNN branch. Generally, a ResNet-50 includes one 7×7 convolutional layer, 16 residual blocks, and a linear layer. Each residual block consists of 3 convolutional layers of kernel size 1×1, 3×3, and 1×1 respectively. A ReLU activate function is employed between every convolution layer. More details can see in [26]. Finally, we retain all convolutional blocks and rectify the output dimension of the final linear layer to generate a 512-D feature.

ViT Extractor Concerning the limitation of the training condition mentioned in the "motivation" section, we utilized T2T-ViT [37] in the ViT extractor to extract global features by its long-range dependency. We make two modifications to the T2T-ViT so that it can fit COVID-19 classification tasks. Firstly, we truncated the inner Transformer layer into 5. Secondly, we adjust its hidden dimension to 512 formulating a 512-D feature directly.

According to the feature extractor we chose above, a pair of 512-D feature embeddings generated from ResNet-50 and T2T-ViT-5, which represents local and global features severally.

3.2.2. Feature fusion block

We note that using concatenated feature embedding from CNN extractor and ViT extractor directly cannot reflect the importance proportion derived from different feature extractors. For this reason, we designed a feature complement Transformer (FCT) shown in Fig. 3 to conduct the merging features from the different feature extractors. FCT adopts the multi-head self-attention as the core mechanism, which gets the ability to concentrate on the essential value automatically. After that, we further introduced a single-layer feed-forward network (FFN) to reinforce the expression capability of the FCT structure.

Let $\boldsymbol{F} = \{x_1, x_2, x_3 \ldots x_n\} \in \mathbb{R}^L$ denote the concatenated feature embedding from the CNN and ViT extractor: where L is the length of the concatenated feature embedding. To meet the requirement of calculating the self-attention, we thus add an extra dimension to this embedding to get $\boldsymbol{F}' \in \mathbb{R}^{1\times L}$. This approach makes the embedding match the size of the FCT required only, which does not influence the data and its structure. To balance the computational efficiency and model performance, we finally utilize an 8-head self-attention structure, such that each instance project to embeddings Q_i, K_i, and V_i of each $head_i$ can represent as:

$$Q_i = \boldsymbol{F}' \cdot \boldsymbol{w}_{\boldsymbol{q}_i} \tag{1}$$

$$K_i = \boldsymbol{F}' \cdot \boldsymbol{w}_{\boldsymbol{k}_i} \tag{2}$$

$$V_i = \boldsymbol{F}' \cdot \boldsymbol{w}_{\boldsymbol{v}_i} \tag{3}$$

where $\boldsymbol{w}_{\boldsymbol{q}} \in \mathbb{R}^{L\times \frac{L}{8}}$, $\boldsymbol{w}_{\boldsymbol{k}} \in \mathbb{R}^{L\times \frac{L}{8}}$, and $\boldsymbol{w}_{\boldsymbol{v}} \in \mathbb{R}^{L\times \frac{L}{8}}$ denote the embedding weight matrix of each head. The process of calculating the attention of the whole feature embedding of each head is then given by:

$$Attention\,(Q_i, K_i, V_i) = softmax\left(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}}\right) \cdot V_i \tag{4}$$

where $d_k = \frac{d_F}{head_{num}} = \frac{L}{8}$ is the embedding dimension of each head. The symbol $(\cdot)$ represents the dot product operation. For

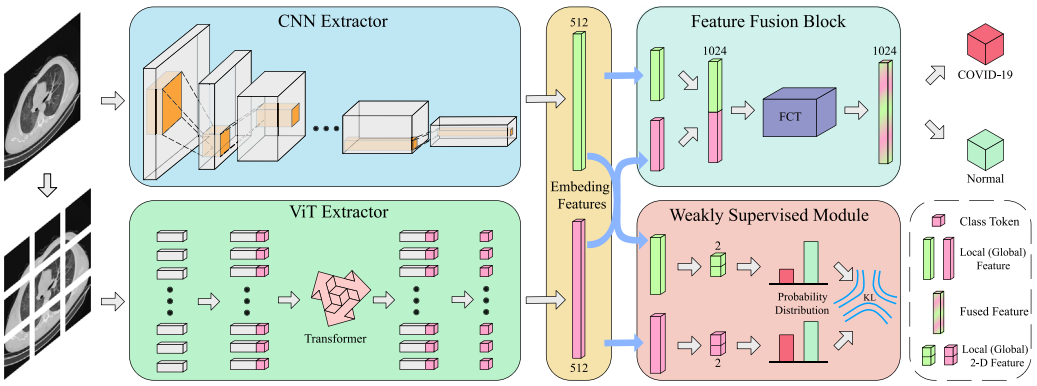

Fig. 2. The architecture of our proposed FCF-Net includes two feature extractors, a feature fusion block, and a weakly supervised module. A CT scan image passed two various feature extractors to generate 512-D local and global feature embeddings. Feature fusion block concatenates these embeddings and processes them within a feature complement Transformer (FCT), then makes the final prediction. The weakly supervised module first maps the various embedding to 2-D feature. These 2-D features will translate into probability distribution through a softmax layer. Finally, we calculate the KL divergence of these two distributions to optimize the ViT extractor in a weakly supervised way.

each of these, we refer to the final attention embedding $F'_{Attention}$ as:

$$F_{Attention} = Concate(head_1, head_2, \ldots head_i) \quad (5)$$
$$where\ head_i = Attention\,(Q_i, K_i, V_i)$$

$$FFN = MLP_1(GELU(MLP_2(F_{Attention}))) \quad (6)$$

$$F'_{Attention} = FFN(F_{Attention}) \quad (7)$$

where MLP represents the multilayer perceptron and GELU denotes the activation function. The output dimension of MLP_1 and MLP_2 is set to L and $2L$ correspondingly. Then restore its shape from $\boldsymbol{F}'_{Attention} \in \mathbb{R}^{1\times L}$ to $\boldsymbol{F}''_{Attention} \in \mathbb{R}^{L}$. Cause the attention mechanism effect, the processed feature embedding retains the crucial instance and attaches more importance to its staple feature from the different feature extractor. On the other side, the weakness instance, meanwhile, recedes its representation through the attention process. That means local feature and global feature reinforce their representation and integrate more efficiently. In addition, we add an FFN as the sub-layer to apply local and translationally equivariant that do not supply in the attention mechanisms. It further strengthens the representation power of the whole FCT structure.

As other neural network classification tasks do, we adopt a linear classifier and a softmax layer to make the final prediction $\hat{y}$ as:

$$\hat{y} = argmax\,(softmax\,(Linear(F''_{Attention}))) \quad (8)$$

3.2.3. Weakly supervised module

To overcome the convergence problem in the ViT structure, we designed this weakly supervised module. DL model for classification task final mapping the feature embedding into a probability distribution, no matter the CNN model or the ViT model. And considering the CNN model has tremendous generalization ability, we adopt the probability distribution created by the CNN extractor to be a weak label. We employ this weak label to inform the ViT where the probable location of proper distribution is, although it cannot provide a definite distribution of the ground truth. First, this module leverage the embedding features from two different extractors into 2-D features by a linear layer. Then we use a softmax layer to render it into a probability distribution of the final prediction. We use these distributions to calculate KL divergence to optimize the ViT extractor in training. It will detail more in the modified loss section.

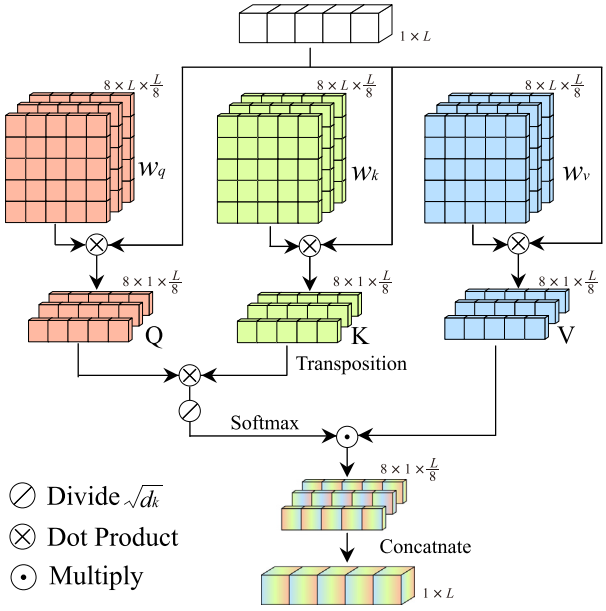

Fig. 3. FCT structure.

3.3. Modified loss

We utilize the categorical cross-entropy loss as the majority loss to supervise this classification model.

$$\mathscr{L}_{CE} = -\frac{1}{N}\sum_i \hat{y}_i\, logp_i + \left(1-\hat{y}_i\right)\log\left(1-p_i\right) \quad (9)$$

where p_i is the actual probability of the image, $\hat{y}$ denotes the predict label of the model.

Consider our model incorporated with ViT structure for adding global features. And Transformer-based model gets a worse performance when it lack of large dataset to do pretraining on it, in contrast to the same size CNN model counterpart [37,47]. To address this problem, we created a weakly supervised module to accelerate the rate of convergence of ViT by a weakly supervised learning manner.

Due to these two distributions from various feature extractors needing to describe the same prediction result, we expected the prediction distribution from these feature embeddings to be

Algorithm 1 Feature Complement Fusion Network (FCF)

Input: A Batch of Image Data: X_i;
Label of A Batch of Image: L_i;
Initialization: $Loss = 0$;
for X_i in Dataset **do**
  $E_c, E_v \leftarrow$ **CNN**(X_i), **ViT**(X_i)
  % Acquire embeddings from CNN and ViT networks
  $E_{c_{2-D}}, E_{v_{2-D}} \leftarrow$ **Linear(**E_c**)**, **Linear(**E_v**)**
  % Obtain 2-D feature embeddings
  $D_c, D_v \leftarrow$ **Softmax**$(E_{c_{2-D}})$, **Softmax**$(E_{v_{2-D}})$
  % Translate 2-D feature embeddings into probability distributions
  $F \leftarrow$ **Concate**(E_c, E_v)
  % Fuse feature embeddings
  $F'' \leftarrow$ **FCT**(F)
  % emphasizes the representation of F
  $\mathscr{L}_{CE} \leftarrow$**CrossEntropy**$(F'', L_i)$
  $\mathscr{L}_{KL} \leftarrow$ **KL Divergence**(D_c, D_v)
  $\mathscr{L}_{Total} \leftarrow \mathscr{L}_{CE} + \lambda * \mathscr{L}_{KL}$
  $Loss \leftarrow Loss + \mathscr{L}_{Total}$
end for
Backward *Loss* to **optimize** the parameters in FCF

similar. We consequently introduced a KL divergence as a regularization term loss to minimize the distance of two distributions:

$$\mathscr{L}_{KL}(D_c \parallel D_t) = \sum_{x \in X} D_c(x) log \frac{D_c(x)}{D_t(x)} \tag{10}$$

where x here denotes each input from the dataset X, $D_c(x)$ means the distribution of $x \in X$ generated from the CNN feature embedding. $D_t(x)$ can be obtained in the same way by the ViT extractor. We set $D_c(x)$ as the pseudo label, let $D_t(x)$ learn information from the pseudo label to optimize the ViT extractor. That means the ViT extractor can learn the distribution generated from the CNN extractor to accelerate its convergence rate and train the ViT extractor more efficiently.

The total loss can be written as follow:

$$\mathscr{L}_{total} = \mathscr{L}_{CE} + \lambda * \mathscr{L}_{KL} \tag{11}$$

where λ represents an impact factor to restrain how strength this term intervenes in the ViT extractor optimization process. λ will be discussed in the following parameter analysis section.

4. Experiment

4.1. Implementation details

Dataset. The dataset we used is a subset of COVID-CTset, which was introduced in a COVID-19 pneumonia detecting framework by Mohammad Rahimzadeh [48]. COVID-CTset was gathered from Negin medical center in Iran, they adopt a SOMATOM Scope model and syngo CT VC30-easyIQ software version for capturing and visualizing the lung HRCT radiology images from the patients. It uses 16-bit rather than 8-bit grayscale images to reserve more detail for experts to diagnose low-level inflation precisely. These kinds of images in this dataset perfectly evade the data leakage problem by compressing the image into other formats mentioned in [49]. CT images are not in color, and these raw data is only one channel instead of RGB (R=G=B) [49], which is more scientific for the later training process. This subset contains 4562 images which controlled the positive and negative samples as the same proportion strictly. And then we split the selected images into train-set, val-set and test-set as a proportion of 8:1:1. We show the dataset configuration in Fig. 4 and sample details in Fig. 5.

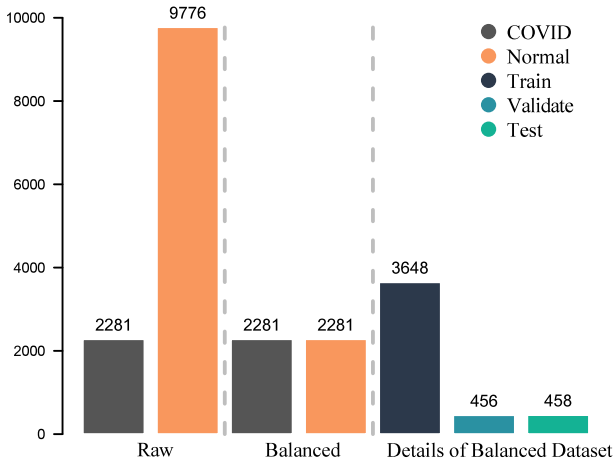

Fig. 4. Dataset Configuration.

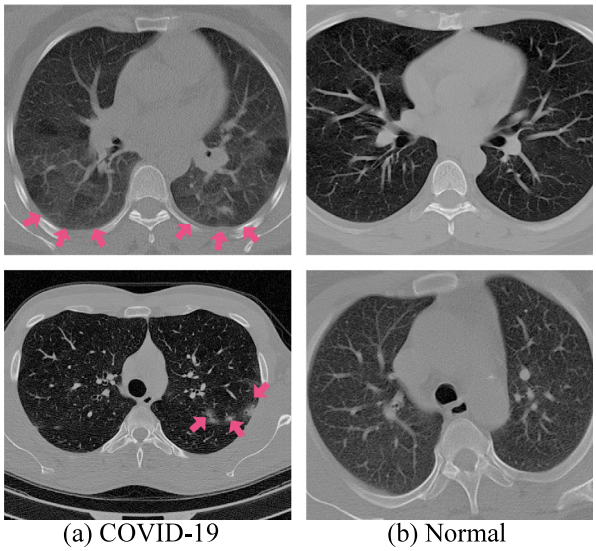

Fig. 5. Samples in COVID-CTset: (a) COVID-19 samples, where red arrows denote the infection area. (b) Normal samples.

Training details. The implementation of the whole FCF architecture software is based on the Pytorch [50] framework. The hardware information to train this model is on an i9-10900F CPU and a single RTX 3090 GPU. Our proposed work is an end-to-end model, which represents this model requires a CT image only for the final result. We train this network end-to-end for 200 epochs using Adam [51] optimizer with a constant learning rate of 7e-5 through our proposed loss. The parameters in the Adam optimizer are setting as $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Assessment criteria. And to better critic the performance of our proposed model, we applied four evaluation metrics to judge our model synthetically.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

Table 1
Parameter Analysis.

Para.	Value	Accuracy	Recall	Precision	F1-Score
λ	0.4	97.81	98.25	97.39	97.82
	0.5	98.68	98.25	99.12	98.68
	0.6	**99.34**	99.56	**99.13**	**99.34**
	0.7	97.81	96.93	98.66	97.79
	0.8	97.15	96.49	97.78	97.13
	1.0	98.90	**100.00**	97.85	98.92
	1.2	98.46	97.81	99.11	98.45

$$F1\text{-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

where

- True Positive (TP): The model forecasts the COVID-19 category to be the COVID-19 class.
- False Positive (FP): The model predicts the initially Normal class into the COVID-19 class.
- True Negative (TN): The model forecasts the Normal category to be the Normal class.
- False Negative (FN): The model predicts the COVID-19 class into the Normal class.

4.2. Parameter analysis

We introduced a hyperparameter λ to describe the impact force on the regularization term in Eq. (11). This parameter constrains how similar these distributions are, generated from the different extractors. We experiment with seven different values of this hyperparameter from 0.4 to 1.2 to analyze the performance of the COVID-19 detection model. To guarantee fairness in the experiment, we initial the training model at the same weight. And we introduced four metric methods from Eq. (15) to judge the result synthetically. These four metric methods show different variations when changing the parameter λ. We can see the best accuracy shows up s.t. λ = 0.6 in Table 1. It indicates that this model can make the minimum error decision in the prediction process. The highest recall emerged when λ = 1.0, which represents the model acquired an excellent capability to cover the fully COVID-19 samples in the test set. Besides, the precision reaches the apogee when λ = 0.6. It demonstrates our model gets the best ability to predict the COVID-19 class using this hyperparameter. However, greedy much on the recall or precision of the model is not a wise choice. The comprehensive critical ability of the focal infection prediction model cannot be neglected. Thus we compared their F1-Score, which employs an average judgment on recall and precision, to determine the best accuracy and F1-Score model. We finally choose λ = 0.6 to be the best optimization hyperparameter.

4.3. Ablation study

We design various ablation experiments to illustrate the effectiveness of our proposed methods in this brand-new architecture. We show the ablation result of the whole architecture used ResNet-50 to be the CNN extractor and T2T-ViT-5 as the ViT extractor in Table 2. All these experiments are based on the same initial weight to ensure fairness. We first focus on how the KL divergence in the loss function affects the performance of the whole architecture. Then, we remove the KL divergence in the loss function and add our FCT structure to test the influence of this proposed method. After that, we merged them into the model to see its bonus performance.

KL divergence. We introduced KL divergence in the loss function to assist the ViT extractor in learning from the CNN extractor. This approach aims to train the model more efficiently and overcome the low convergence defect of the Transformer-based model. It raises the accuracy performance in the final prediction as well. Although ablation on KL divergence only model gets the same performance as the FCT only model. As shown in Fig. 6, the model adding KL divergence to the loss function achieves a higher speed in the convergence course.

Table 2
Ablation on FCF-Res model.

FCT	KL	Accuracy	Recall	Precision	F1-Score
		97.81	97.37	98.23	97.80
✓		98.46	98.68	98.25	98.47
	✓	98.49	98.68	98.25	98.47
✓	✓	**99.34**	**99.56**	**99.13**	**99.34**

Table 3
Ablation on FCF-Dense model.

FCT	KL	Accuracy	Recall	Precision	F1-Score
		98.46	99.12	97.84	98.47
✓		98.90	99.12	98.69	98.91
	✓	98.68	**99.56**	97.85	98.70
✓	✓	**99.34**	99.12	**99.56**	**99.34**

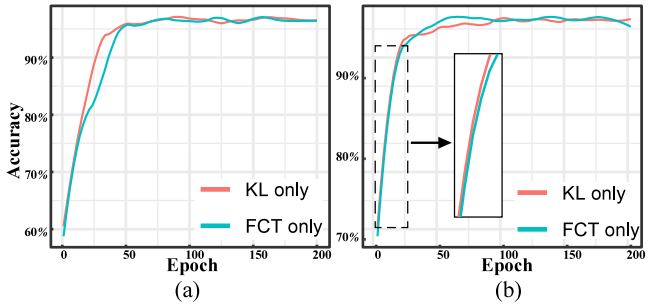

Fig. 6. The variation of accuracy in the training process: (a) FCF-Res model, (b) FCF-Dense model.

FCT. Next, we dissect the result in Table 2 and note that the whole model performed well when adding our elegant fusion strategy — FCT into it. Given result demonstrates that using FCT rather than mere concatenated feature embeddings can make a 0.25% rise in prediction accuracy. It also developed the recall indices, which represent the whole model receive a strong ability in detecting disease through the FCT.

We do the same ablation experiment when replacing the ResNet-50 with DenseNet-121 in CNN branch to prove the validity of our framework. The result in Table 3 corroborates the effectiveness of our thought and the correctness of our framework as well.

4.4. Compared to classic SOTA

We do experiments to compare our network with the state-of-the-art classic classification algorithms, including CNN models, ViT models, and feature fusion models in Table 4.

FCF reaches 99.34% accuracy, which is 0.44% higher when compared to CNN models. This performance can defeat any CNN model, but it receives more parameters and FLOPs than ResNet-50 and DenseNet all series. As we declared in the related work, a fully-Transformer structure will get a poor performance if it encounters a data scarcity condition or training on a single GPU. Hence, we only exhibit the result of base ViT and the T2T-ViT-5 we used in our framework. As for FPN and UNet, they both gathered features to enhance the representation for the final prediction. FPN concatenates every embedding from different fusion

Table 4
Compared to Classic SOTA.

Model type	Model	Params. (M)	FLOPs (G)	Accuracy	Recall	Precision	F1-Score
CNN	ResNet 50	23.5	10.6	97.81	97.37	98.23	97.80
	ResNet 101	42.5	20.2	98.03	98.25	97.82	98.03
	ResNet 152	58.1	30.0	98.68	**99.56**	97.84	98.70
	DenseNet 121	6.9	7.2	98.46	99.12	97.84	98.47
	DenseNet 161	26.5	20.0	98.46	**99.56**	97.42	98.48
	DenseNet 169	12.5	8.6	98.90	99.12	98.69	98.91
	DenseNet 201	18.1	11.2	98.68	**99.56**	97.84	98.70
ViT	ViT-b	15.8	8.2	81.36	88.16	77.60	82.55
	T2T-ViT-5	10.9	6.0	94.74	92.11	97.22	94.59
Feature Fusion	FPN(ResNet 50)	26.3	17.8	98.03	97.81	98.24	98.02
	UNet	17.2	80.0	**99.34**	**99.56**	**100.0**	**99.34**
Feature Complement Fusion	FCF-Res(ours)	40.0	16.6	**99.34**	**99.56**	99.13	**99.34**
	FCF-Dense(ours)	24.2	13.2	**99.34**	99.12	99.56	**99.34**

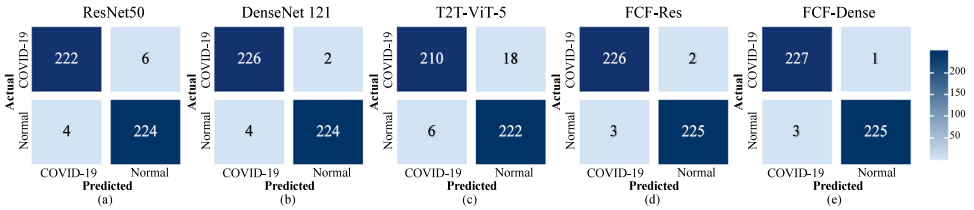

Fig. 7. Confusion matrix of FCF and their source feature extractor.

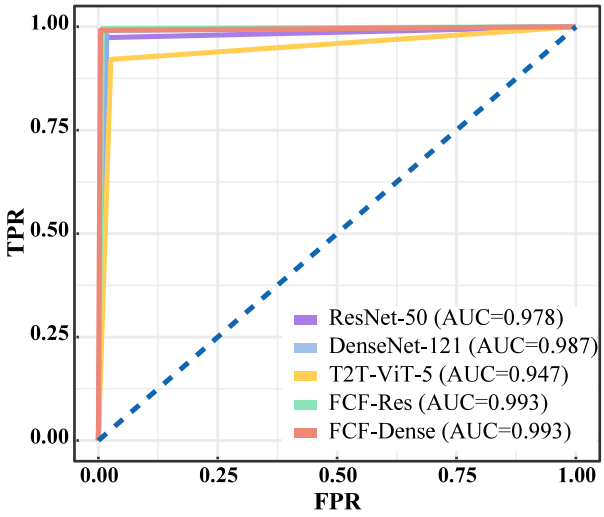

Fig. 8. ROC of FCF and their source feature extractor.

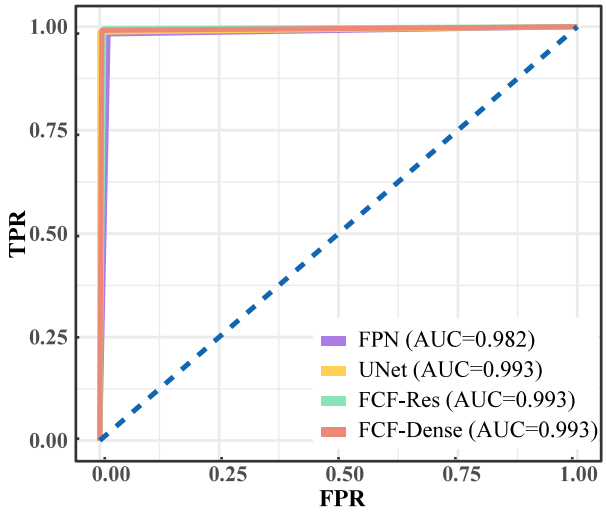

Fig. 9. ROC of FCF and other feature fusion methods.

layers in the pyramid network. And every sub-layer under the top layer applied a feature enhancement by adding an up sample feature of its up-neighbor layer. This process increases 40% FLOPs than normal ResNet-50, although it gets a better accuracy performance when compared to ResNet-50. UNet attains the same high accuracy with 99.34%, according to its characteristic U-shape structure. This architecture applied huge feature maps in the skip connection process to do feature enhancement roughly. However, the FLOPs run up rapidly along with that process corresponding. A four times increase in FLOPs to get an extra 0.22% F1-Score is not a wise choice. In contrast, our provided framework makes a good trade-off on prediction performance and computation efficiency. We sacrifice the training parameters and bring our model size into the middle-level to trade fewer FLOPs for refining our model. The confusion matrix we post in Fig. 7 shows the detailed performance of each model we applied in our feature complement fusion framework. We compare each class that was predicted, including every engaged sample. Although T2T-ViT-5 performs inferior to ResNet-50 and DenseNet-121, it provides useful property features fed into FCT and promotes prediction performance. Compare the data shown in Fig. 7 (a) and Fig. 7 (d), we can conclude that global features generated from T2T-ViT-5 compensate the local features extracted by the ResNet-50. The same conclusion can be easily obtained when comparing the Fig. 7 (b) and Fig. 7 (e). The AUC of ROC curve results in Fig. 8 and Fig. 9 also shores up the effectiveness of our proposed model exactly.

4.5. Compared to COVID-19 SOTA

We also compared our FCF-Res with other COVID-19 state-of-art works on three public datasets to verify its generalization ability and effectiveness. As shown in Table 5, InstaCovNet-19 and our FCF-Res emerge with the same accuracy of 99.34% on

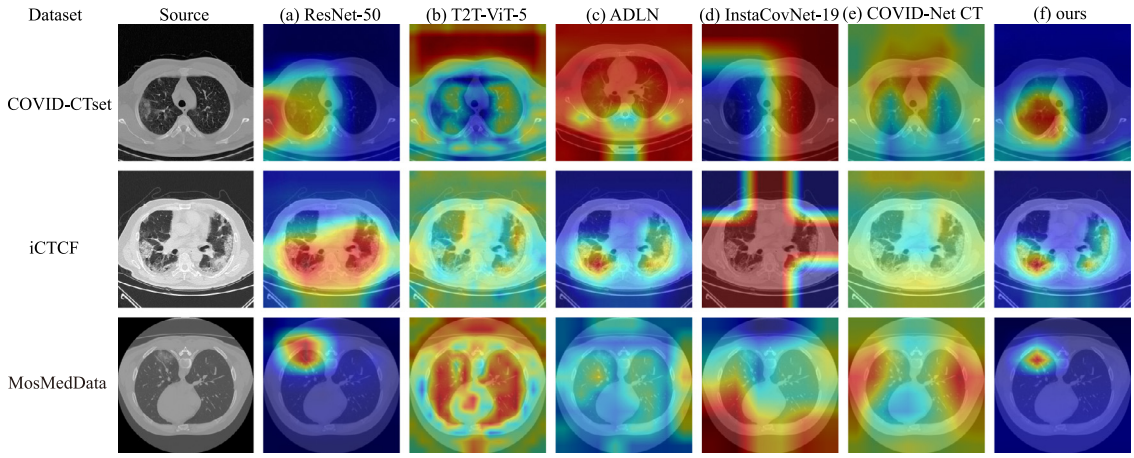

Fig. 10. Grad-CAM on comparison methods.

Table 5
Accuracy performance of COVID-19 SOTA methods on COVID-CTset [48], iCTCF [52] and MosMedData [53].

Method	Params. (M)	FLOPs (G)	COVID-CTset[48]	iCTCF [52]	MosMedData [53]
ADLN [48]	26.2	17.8	96.05	97.56	72.40
InstaCovNet-19 [45]	165.8	69.2	**99.34**	**99.31**	78.73
COVID-Net CT [54]	15.2	0.9	78.82	97.00	66.86
FCF-Res(ours)	40.0	16.6	**99.34**	98.19	**84.50**

the COVID-CTset, but our calculation efficiency is more economical. Whereas InstaCovNet-19 integrated five baseline models to get the final prediction. Due to InstaCovNet-19 integrating five baseline models to get the final prediction, it boosts parameters and FLOPs dramatically, although it achieves the best performance on iCTCF [52]. It boosts parameters and FLOPs increasing to a dramatic extent, although it achieves the best performance. Four times of resource usage but only increasing the accuracy by 1.12% is not a cost-effective strategy. COVID-Net applied the minimum parameters, which perform well on iCTCF while lacking generalization ability performing worse on other datasets. Even on the MosMedData [53], our model shows the best accuracy performance and verified its robustness and generalization ability.

We also handle Grad-CAM [55] to visualize the result to illustrate where the gradient actually focused in Fig. 10. ADLN shows a similar result on iCTCF (Fig. 10 (c)) as ours (Fig. 10 (f)) but appears weak gradient on MosMedData. Although InstaCovNet-19 achieves the best accuracy, it cannot locate the gradient of lesion areas in visualization results. The gradient of COVID-Net CT cannot focus on lesion areas in all three datasets, which also explains its poor generalization ability. The gradient was relatively dispersed in the T2T-ViT-5 or made misregistration of lesion areas in ResNet-50. Ours FCF integrated the advantage characteristic of ResNet-50 and T2T-ViT-5. It not only possesses broader receptive fields to guarantee a wide range of gradients but also preserves effective gradients on the lesion areas precisely.

5. Conclusion

In this paper, we introduce a detailed study of feature complement fusion network (FCF) to detect COVID-19 through CT images. This structure involves the advantage of the receptive field in both CNN and ViT models to extract complement features. And a weakly supervised module successfully assists the ViT structure to converge better. We further provided an FCT architecture to merge the characteristics from feature extractors economically. Extensive ablation experiments demonstrate that our proposed method specifically worked in the weakly supervised module and feature fusion block. Moreover, these algorithms perform well in the whole construction. It emphasizes the priority in different features and enables them fused more efficiently.

Although we alleviate the converge slow question caused by ViT in our architecture, our model is training unstable cause of its inner layer normalization structure. It also needs a long time to train for convergence on a large-scale dataset compared to the pure CNN model. We are going to overcome the problems mentioned above in our future works.

CRediT authorship contribution statement

Shu Liang: Conceptualization, Software, Writing – original draft and editing. **Rencan Nie:** Supervision, Writing – reviewing. **Jinde Cao:** Supervision. **Xue Wang:** Software. **Gucheng Zhang:** Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding

This work was supported by National Natural Science Foundation of China under Grants 61966037, 61833005, 61463052, and 62066047, National Key Research and Development Project of China under Grant 2020YFA0714301, and China Postdoctoral Science Foundation under Grant 2017M621586.

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (10223) (2020) 497–506.

[2] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong, et al., Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia, N. Engl. J. Med. (2020).

[3] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study, Lancet 395 (10225) (2020) 689–697.

[4] WHO, Who coronavirus (covid-19) dashboard, 2021, Website, https://covid19.who.int/ (Accessed 15 Nov 2021).

[5] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: evaluation of the diagnostic accuracy, Radiology 296 (2) (2020) E65–E71.

[6] F. Song, N. Shi, F. Shan, Z. Zhang, J. Shen, H. Lu, Y. Ling, Y. Jiang, Y. Shi, Emerging 2019 novel coronavirus (2019-ncov) pneumonia, Radiology 295 (1) (2020) 210–217.

[7] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, J. Liu, Chest ct for typical coronavirus disease 2019 (covid-19) pneumonia: relationship to negative rt-pcr testing, Radiology 296 (2) (2020) E41–E45.

[8] H.Y.F. Wong, H.Y.S. Lam, A.H.-T. Fong, S.T. Leung, T.W.-Y. Chin, C.S.Y. Lo, M.M.-S. Lui, J.C.Y. Lee, K.W.-H. Chiu, T.W.-H. Chung, et al., Frequency and distribution of chest radiographic findings in patients positive for covid-19, Radiology 296 (2) (2020) E72–E78.

[9] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest ct for covid-19: comparison to rt-pcr, Radiology 296 (2) (2020) E115–E117.

[10] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases, Radiology 296 (2) (2020) E32–E40.

[11] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z.A. Fayad, et al., Ct imaging features of 2019 novel coronavirus (2019-ncov), Radiology 295 (1) (2020) 202–207.

[12] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S.K. Lo, F.-Y. Wang, Dynamic-fusion-based federated learning for covid-19 detection, IEEE Internet Things J. 8 (21) (2021) 15884–15891.

[13] E.T. Hastuti, A. Bustamam, P. Anki, R. Amalia, A. Salma, Performance of true transfer learning using cnn densenet121 for covid-19 detection from chest x-ray images, in: 2021 IEEE International Conference on Health, Instrumentation & Measurement, and Natural Sciences, InHeNce, IEEE, 2021, pp. 1–5.

[14] H. Kang, L. Xia, F. Yan, Z. Wan, F. Shi, H. Yuan, H. Jiang, D. Wu, H. Sui, C. Zhang, et al., Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning, IEEE Trans. Med. Imaging 39 (8) (2020) 2606–2614.

[15] S.M.J. Jalali, M. Ahmadian, S. Ahmadian, A. Khosravi, M. Alazab, S. Nahavandi, An oppositional-cauchy based gsk evolutionary algorithm with a novel deep ensemble reinforcement learning strategy for covid-19 diagnosis, Appl. Soft Comput. 111 (2021) 107675.

[16] A. Saygılı, A new approach for computer-aided detection of coronavirus (covid-19) from ct and x-ray images using machine learning methods, Appl. Soft Comput. 105 (2021) 107323.

[17] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, S. Zhang, A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images, IEEE Trans. Med. Imaging 39 (8) (2020) 2653–2663.

[18] Q. Yao, L. Xiao, P. Liu, S.K. Zhou, Label-free segmentation of covid-19 lesions in lung ct, IEEE Trans. Med. Imaging (2021).

[19] W. Xie, C. Jacobs, J.-P. Charbonnier, B. Van Ginneken, Relational modeling for robust and efficient pulmonary lobe segmentation in ct scans, IEEE Trans. Med. Imaging 39 (8) (2020) 2664–2675.

[20] H. Xu, J. Ma, Emfusion: An unsupervised enhanced medical image fusion network, Inf. Fusion 76 (2021) 177–186.

[21] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.

[22] X. Li, X. Guo, P. Han, X. Wang, H. Li, T. Luo, Laplacian redecomposition for multimodal medical image fusion, IEEE Trans. Instrum. Meas. 69 (9) (2020) 6880–6890.

[23] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, ArXiv preprint https://arxiv.org/abs/2102.04306.

[24] Y. Xiong, B. Du, P. Yan, Reinforced transformer for medical image captioning, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 673–680.

[25] K.S. andAndrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May (2015), in: Conference Track Proceedings, 2015, pp. 7–9, http://arxiv.org/abs/1409.1556.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[27] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, C. Zheng, A weakly-supervised framework for covid-19 classification and lesion localization from chest ct, IEEE Trans. Med. Imaging 39 (8) (2020) 2615–2625.

[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[29] Y. Oh, S. Park, J.C. Ye, Deep learning covid-19 features on cxr using limited training data sets, IEEE Trans. Med. Imaging 39 (8) (2020) 2688–2700.

[30] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, W. Zhang, Accurate screening of covid-19 using attention-based deep 3d multiple instance learning, IEEE Trans. Med. Imaging 39 (8) (2020) 2584–2594.

[31] Y. Chen, H. Zhang, Y. Wang, Y. Yang, X. Zhou, Q.J. Wu, Mama net: Multi-scale attention memory autoencoder network for anomaly detection, IEEE Trans. Med. Imaging 40 (3) (2020) 1032–1041.

[32] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-net: Automatic covid-19 lung infection segmentation from ct images, IEEE Trans. Med. Imaging 39 (8) (2020) 2626–2637.

[33] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song, et al., Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia, IEEE Trans. Med. Imaging 39 (8) (2020) 2595–2605.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 2021, 2021, pp. 3–7, https://openreview.net/forum?id=YicbFdNTTy.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, [BERT]: PRe-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, in: Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

[37] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021, ArXiv preprint https://arxiv.org/abs/2101.11986.

[38] M. Popel, O. Bojar, Training tips for the transformer model, 2018, ArXiv preprinthttps://arxiv.org/abs/1804.00247.

[39] Z. Zhang, H. Zhang, L. Zhao, T. Chen, T. Pfister, Aggregating nested transformers, 2021, ArXiv preprinthttps://arxiv.org/abs/2105.12723.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021, ArXiv preprinthttps://arxiv.org/abs/2103.14030.

[41] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[43] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, D. Qian, Prior-attention residual learning for more discriminative covid-19 screening in ct images, IEEE Trans. Med. Imaging 39 (8) (2020) 2572–2583.

[44] H. Chao, X. Fang, J. Zhang, F. Homayounieh, C.D. Arru, S.R. Digumarthy, R. Babaei, H.K. Mobin, I. Mohseni, L. Saba, et al., Integrative analysis for covid-19 patient outcome prediction, Med. Image Anal. 67 (2021) 101844.

[45] A. Gupta, S. Gupta, R. Katarya, et al., Instacovnet-19: A deep learning classification model for the detection of covid-19 patients using chest x-ray, Appl. Soft Comput. 99 (2021) 106859.

[46] T. Durand, T. Mordan, N. Thome, M. Cord, Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 642–651.

[47] Z. Dai, H. Liu, Q.V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, 2021, ArXiv preprinthttps://arxiv.org/abs/2106.04803.

[48] M. Rahimzadeh, A. Attar, S.M. Sakhaei, A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset, Biomed. Signal Process. Control 68 (2021) 102588.

[49] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, P. Biecek, Checklist for responsible deep learning modeling of medical images based on covid-19 detection studies, Pattern Recognit. 118 (2021) 108035.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[51] D.P. Kingma, J. Ba, [Adam: A] method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May (2015), in: Conference Track Proceedings, 2015, pp. 7–9, http://arxiv.org/abs/1412.6980.

[52] W. Ning, S. Lei, J. Yang, Y. Cao, P. Jiang, Q. Yang, J. Zhang, X. Wang, F. Chen, Z. Geng, L. Xiong, H. Zhou, Y. Guo, Y. Zeng, H. Shi, L. Wang, Y. Xue, Z. Wang, Ictcf: an integrative resource of chest computed tomography images and clinical features of patients with covid-19 pneumonia, 2020, http://dx.doi.org/10.21203/rs.3.rs-21834/v1, https://europepmc.org/article/PPR/PPR141530.

[53] S. Morozov, A. Andreychenko, I. Blokhin, A. Vladzymyrskyy, P. Gelezhe, V. Gombolevskiy, A. Gonchar, N. Ledikhova, N. Pavlov, V. Chernina, Mosmeddata: Chest ct scans with covid-19 related findings, 2020, Website, https://www.kaggle.com/datasets/mathurinache/mosmeddata-chest-ct-scans-with-covid19.

[54] H. Gunraj, L. Wang, A. Wong, Covidnet-ct: A tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images, Front. Med. (2020) 1025.

[55] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.