



Emb-trattunet: a novel edge loss function and transformer-CNN architecture for multi-classes pneumonia infection segmentation in low annotation regimes

Fares Bougourzi¹ · Fadi Dornaika^{2,3} · Amir Nakib¹ · Abdelmalik Taleb-Ahmed⁴

Accepted: 31 January 2024 / Published online: 14 March 2024
© The Author(s) 2024

Abstract

One of the primary challenges in applying deep learning approaches to medical imaging is the limited availability of data due to various factors. These factors include concerns about data privacy and the requirement for expert radiologists to perform the time-consuming and labor-intensive task of labeling data, particularly for tasks such as segmentation. Consequently, there is a critical need to develop novel approaches for few-shot learning tasks in this domain. In this work, we propose a Novel CNN-Transformer Fusion scheme to segment Multi-classes pneumonia infection from limited CT-scans data. In total, there are three main contributions: (i) CNN-Transformer encoders fusion, which allows to extract and fuse richer features in the encoding phase, which contains: local, global and long-range dependencies features, (ii) Multi-Branched Skip Connection (MBSC) is proposed to extract and fuse richer features from the encoder features then integrate them into the decoder layers, where MBSC blocks extract higher-level features related to the finer details of different infection types, and (iii) a Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss function is proposed to deal with fuzzy boundaries, enhance the separability between classes and give more attention to the minority classes. The performance of the proposed approach is evaluated using two evaluation scenarios and compared with different baseline and state-of-the-art segmentation architectures for Multi-classes Covid-19 segmentation. The obtained results show that our approach outperforms the comparison methods in both Ground-Glass Opacity (GGO) and Consolidation segmentation. On the other hand, our approach shows consistent performance when the training data is reduced to half, which proves the efficiency of our approach in few-shot learning. In contrast, the performance of the comparison methods drops in this scenario. Moreover, our approach is able to deal with imbalanced data classes. These advantages prove the effectiveness and efficiency of the proposed EMB-TrAttUnet approach in a pandemic scenario where time is critical to save patient lives.

Keywords Covid-19 · Transformer · Convolutional neural network · Deep learning · Segmentation · Unet

1 Introduction

Over the past three years, the world has been facing a global crisis due to the spread of the Covid-19 pandemic, which has affected all world's countries (Amyar et al., 2020; Bougourzi et al., 2022). By November 2023, more than 770 million cases and about than 7 million deaths as declared by World Health Organization (WHO) WHO (2023). The Covid-19 pandemic is caused by the infectious SARS-CoV-2 virus, which mainly affects the lungs and spreads to other organs (Parasher 2021). Since the start of the pandemic, medical imaging has been widely used for Covid-19 analysis, including: Covid-19 diagnosis (Vantaggiato et al. 2021), segmentation (Pezzano et al. 2021; Bougourzi et al. 2022), and severity estimation and prediction (Li et al. 2021; Bougourzi et al. 2021). Medical scan modalities include: X-ray, Computerized Tomography (CT) and Ultrasound (Vantaggiato et al. 2021; Bougourzi et al. 2021; Roy et al. 2020). In fact, CT-scans are more efficient for Covid-19 analysis because they provide more details about the spread and severity of the infection, which allows to following-up the patient state and taking the right decision to save the patient's life (Shi et al. 2020; Zhang et al. 2020). In particular, Covid-19 infection segmentation from CT-scans is a very important step in detecting, quantifying and evaluating the infection and predicting its severity (Shi et al. 2020; Zhang et al. 2020).

In the last decade, Deep Learning (DL) approaches have become the dominant trend in many machine learning and computer vision tasks (Van der Velden et al. 2022; Bougourzi et al. Apr. 2022; Goceri 2023). In particular, DL approaches have been widely studied for automatically analysing various infections and diseases using different medical imaging modalities, which can assist or even replace experienced physicians and radiologists in diagnosing and analysing the diseases (Xie et al. 2023; Bougourzi et al. 2023; Soomro et al. 2022). However, In the case of the Covid-19 pandemic, the number of cases in a short period of time is enormous, which overwhelming the resources and personnel of hospitals, including physicians and radiologists. Therefore, it is necessary to develop automatic machine learning solutions to support the control of this disease and reduce the workload of medical staff (Vantaggiato et al. 2021; Roy et al. 2020; Rehman et al. 2023).

In the last decade, semantic segmentation tasks have been extensively studied and numerous deep learning segmentation approaches have been proposed (Ghosh et al. 2019; Minaee et al. 2021). In particular, the segmentation of Covid-19 infections has been widely investigated, which is a very challenging task due to two main factors. First, the available labelled data are very limited due to the significant time and effort required for the labelling process, with physicians and radiologists being overwhelmed due to the pandemic (Wang et al. 2020; Pezzano et al. 2021; Yao et al. 2021). The second factor is related to the nature of Covid-19 infection, which progress differently from one case to case. Moreover, Covid-19 infection exhibits high variability in terms of intensity, shape, position and type depending on the stage of infection (early vs. advanced), symptoms (asymptomatic vs. symptomatic patients), and severity (Shi et al. 2020; Zhang et al. 2020). In fact, the literature works have concentrated on segmenting Covid-19 infection as a binary task (background or infection) Pezzano et al. (2021); Paluru et al. (2021); Cong et al. (2022a, 2022b); Yao et al. (2021), as the available labelled data for Covid-19 infection types (Multi-classes) are very limited. However, segmentation of Covid-19 infections into multiple classes provides more details about infection progression, stage, and severity (Hefeda 2020; Salehi et al. 2020).

In this paper, a new CNN-Transformer based approach is proposed for segmenting Covid-19 infections into multiple classes from limited data. In detail, the proposed EMB-TrAttUnet consists of two main components: (i) CNN-Transformer architecture

(MB-TrAttUnet) and (ii) Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss function. The MB-TrAttUnet is an Encoder-Decoder architecture that follows the Unet architecture family (Ronneberger et al. 2015). The encoder of the proposed MB-TrAttUnet architecture combines CNN and Transformer layers to extract local, global, and sequential features in the encoding phase. Instead of the traditional Unet skip connection, a Multi Branches Skip Connection (MBSC) is proposed to extract high-level features, interpret the encoder features, and select the most important features. The decoder of the proposed MB-TrAttUnet is similar to the encoder of the Att-Unet architecture (Oktay et al. 2018), where the attention gates are better exploited because the encoder layers and the MBSC extract higher and richer features for Multi-classes Covid-19 Segmentation. The MBA-CE Loss function is designed to give more attention to the classes boundaries to enhance boundaries segmentation, the separability between classes and minority classes decision. The main contributions can be summarized as follows:

- We propose a novel hybrid Transformer-CNN architecture (MB-TrAttUnet) for Multi-classes Covid-19 infection segmentation from CT-scans. The Encoder of TrAttUnet architecture combines CNN and Transformer blocks using Encoders Fusion Block (EFB). On the other hand, the Decoder of MB-TrAttUnet is AttUnet-like decoder.
- Furthermore, Multi Branches Skip Connection is introduced to extract more advanced features and to interpret and select the most significant ones from the encoder features and pass them to the decoder block.
- We propose a Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss function to cope with fuzzy boundaries, enhance the separability between classes and give more attention to the minority classes. The trained MB-TrAttUnet with MBA-CE is called EMB-TrAttUnet.
- The comparison between the proposed approach and both baseline and state-of-the-art segmentation architectures clearly demonstrates the superior performance of our method, particularly when faced with limited training data availability. The proposed EMB-TrAttUnet architecture is publicly available at: <https://github.com/faresboumourzi/EMB-TrAttUnet>.

The remainder of this paper is organized as follows: Sect. 2 summarizes the segmentation architectures for medical image segmentation and related work on Covid-19 infection segmentation. Section 3 explains the proposed approach. The datasets and evaluation metrics used are described in Sect. 4. Section 5 presents and discusses the experiments and results. Section 6 provides a visual analysis of the segmentation results of the proposed approach and the comparison methods. Finally, Section 7 concludes the paper.

2 Related work

This section is dedicated to describing the related works concerning state-of-the-art deep learning architectures for medical image segmentation, deep learning approaches for respiratory diseases, and deep learning approaches for COVID-19 infection segmentation.

2.1 DL architectures for medical imaging segmentation

In the last decade, Deep Learning approaches have proved their efficiency in many medical imaging tasks using different medical imaging modalities including but not limited to: Magnetic Resonance Imaging (MRI), X-rays, Computed Tomography (CT) and Ultrasound images (Litjens et al. 2017; Paladini et al. 2021). In particular, medical image segmentation is one of the most studied tasks in recent years. Since the development of Unet architecture in 2015, CNN-based architectures have become the dominant approaches in medical image segmentation. Consequently, plenty of the Unet Ronneberger et al. (2015) variants have been proposed such as Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018), ARU-GD Maji et al. (2022), FANet Tomar et al. (2022), and KiU-Net Valanarasu et al. (2021).

In short, Unet Ronneberger et al. (2015) has an Encoder-Decoder structure with a “U” shape. The encoder of the Unet architecture consists of successive convolutional blocks followed by max-pooling operators to extract high-level features from the input. On the other side, Unet’s decoder exploits the extracted features of the encoder to reconstruct the segmentation mask through successive deconvolution layers. Moreover, the skip connections incorporate the feature maps of the encoder layers into the decoder layers to prevent the loss of detailed features due to the shrinking of the encoder spatial dimensionality. In the Att-Unet architecture (Oktay et al. 2018), Attention Gates (AGs) are placed between the skip connections and the decoder layers to select the most important salient parts from the encoder features and pass them to the decoder.

In recent years, Transformers have shown great success in Natural Language Processing (NLP) domain (Vaswani et al. 2017; Khan et al. 2021). This inspired the computer vision community to exploit the transformer for vision tasks (Dosovitskiy et al. 2020; Liu et al. 2021; Touvron et al. 2021). In particular, transformers have shown promising performance for segmentation tasks in medical imaging (Li et al. 2023). Therefore, plenty of Transformer-based segmentation architectures have been proposed, such as UNETR Hatamizadeh et al. (2022), GT U-Net Li et al. (2021), and Swin-Unet Cao et al. (2021). In Li et al. (2021), a Group Transformer Unet (GT U-Net) architecture was proposed for tooth root segmentation. In more details, GT U-Net keeps the general structure of the Unet architecture, with the encoder and decoder consisting of Group Transformer blocks combining CNN and Transformer components. Also, SwinUnet Cao et al. (2021) architecture follows the same structure as Unet. However, the encoder and decoder of the Swin-Unet were purely constructed using Swin Transformer blocks (Liu et al. 2021) without any convolutional operations. The experimental results in Cao et al. (2021) proved the efficiency of the Swin-Unet architecture compared to CNN-based segmentation architectures for multi-organ and cardiac segmentation tasks.

In addition to pure Transformer architectures, hybrid Transformer-CNN architectures have been investigated in the last two years in the field of medical imaging segmentation (Wang et al. 2022; Huang et al. 2022; Wang et al. 2022). In Wang et al. (2022), the same structure as Unet was retained while proposing a Mixed Transformer Module (MTM), which aims to learn inter and intra-affinity features. MTM consists of a Local-Global Gaussian-Weighted Self-Attention (LGG-SA) and an External Attention (EA) block. In Huang et al. (2022), the aim of the MISSFormer architecture is to propose an Enhanced Transformer Block by using convolutional operations. Additionally, the authors introduced the Enhanced Transformer Context Bridge to extract long-range dependencies and local context from multi-scale features.

In contrast to these state-of-the-art hybrid approaches, our method involves two encoder paths. The first one is designed to extract global and long-range dependency features, while the second encoder aims to combine Transformer features at different levels with the local features extracted by the CNN blocks. Furthermore, a Multi-Branch Skip Connection (MBSC) mechanism is proposed to enhance the traditional skip connection or Att-Unet architecture. MBSC extracts higher-level features from different branches and concatenates their outputs to pass them to the second filtering stage through the attention gates.

2.2 DL approaches for respiratory diseases analysis

Over the last decade, deep learning approaches have been widely utilized for the analysis of respiratory diseases, including chronic obstructive pulmonary disease, pneumonia, asthma, tuberculosis, fibrosis, lung cancer, etc Bharati et al. (2020); Jasmine Pemeena Priyadarsini et al. (2023). Despite the wide variety of deep learning methods available, most of the research on respiratory diseases has predominantly employed Convolutional Neural Networks (CNNs) as the primary technique. These applications encompass numerous tasks, such as the recognition of multiple respiratory diseases from X-ray images (Bharati et al. 2020; Jasmine Pemeena Priyadarsini et al. 2023), the detection of tuberculosis from X-ray images (Duong et al. 2021), the identification of Covid-19 from X-ray images (Duong et al. 2023; Vantaggiato et al. 2021), and Covid-19 segmentation from CT scans (Müller et al. 2021; Ilhan et al. 2023; Saood and Hatem 2021; Fung et al. 2021). In Bharati et al. (2020), S. Bharati et al. introduced a hybrid deep learning framework that combines CNNs, data augmentation, and a spatial transformer network (STN) for the detection of lung diseases in X-ray images. This framework includes 15 different classes, one of which is “No findings,” while the others represent various diseases. In Fung et al. (2021), the authors proposed a two-stage approach with a self-supervised learning strategy for Covid-19 segmentation. They utilized the InfNet architecture as the backbone and integrated generative adversarial image inpainting to enhance the segmentation of lung lesions.

In addition to the CNN architectures, transformers have been investigated for respiratory diseases tasks (Duong et al. 2021), especially for Covid-19 analysis (Liang et al. 2021; Chaudhary et al. 2022). In Duong et al. (2021), L. D. Tuan et al. proposed a hybrid CNN-Transformer architecture for detecting tuberculosis from X-rays. First, they used a CNN backbone (EfficientNet) to extract 2D feature maps from the input X-ray, then they used an Encoder-Decoder Transformers. The first Transformer aims to contextualize features from the CNN features, and the second Transformer aims to auto-regressively generate one channel for each input feature at every time step. In Liang et al. (2021), S. Liang et al. proposed a hybrid CTNet framework that exploits CNN and Transformer approaches. In this approach, they used a CNN feature extractor with an SE attention Block and then passed the obtained results to a ViT module. Their approach demonstrated a good capability to learn discriminative features from 3D CT scans for Covid-19 infection recognition. S. Chaudhary et al. proposed Chaudhary et al. (2022) exploited the strength of the Swin Transformer to extract the feature from the CT slice and pass it through MLP layers for Covid-19 infection percentage estimation. To utilize the recent developments in Vision Transformers, we proposed a Hybrid Transformer-CNN approach for Covid-19 infection segmentation with multiple classes from limited training data.

2.3 DL approaches for Covid-19 infection segmentation

Since the spread of the Covid-19 pandemic in 2019, Covid-19 infection segmentation has been extensively studied by the machine learning and computer vision community. Most of the existing works have been concentrating on segmenting Covid-19 infections as a binary task (infection and background) Bougourzi et al. (2023); Pezzano et al. (2021); Yao et al. (2021); Ding et al. (2021); Paluru et al. (2021); Cong et al. (2022a, 2022b). Fewer works, however, have attempted to segment Covid-19 infection as Multi-classes (Jin et al. 2022; Yu et al. 2022). In Wu et al. (2021), G. Pezzano et al. proposed a multi-stages approach which consists of the following steps: (i) lung delineation, (ii) Covid-19 detection, and (iii) Covid-19 infection segmentation. First hand, they proposed a Unet variant with Multiple Convolutional Layers structure, that performs lung segmentation within a novel pipeline for direct Covid-19 detection and segmentation. They also proposed a new customized loss function for optimizing the performance.

Mu et al. (2021) proposed a progressive global perception and local polishing (PCPLP) approach for Covid-19 infection segmentation. In detail, the PCPLP encoder follows the VGG-19 (Simonyan and Zisserman 2014) architecture to extract multi-level low and middle features. On the other hand, the decoder of PCPLP integrates the multi-scale to learn high-level features for Covid-19 segmentation. To this end, the following blocks were proposed GPM, LPM, and BMS, which aim to localise the global feature of the infected regions, extract the local features of the infected regions contours and preserve the edge-related features, respectively. In Wang et al. (2022), X. Wang et al. proposed a new architecture called Spatial Self-Attention Network (SSA-Net), which exploits both the self-attention mechanism and spatial convolution to expand the receptive field, strengthen the learning process, and accelerate the training convergence. In addition to the proposed SSA-Net architecture, a semi-supervised few-shot iterative approach is proposed to cope with the limited labelled data available for training (Wang et al. 2022).

Since there is a paucity of work addressing the segmentation of Covid-19 infections as Multi-classes, this work focuses on this task, which is very important for identifying the stage of infection and predicting its progression and severity (Hefeda 2020; Salehi et al. 2020). This allows following-up the patient's state and taking the right actions to save the patient's life (Hefeda 2020; Salehi et al. 2020). Unlike the state of the art, our proposed approach efficiently combines CNN and Transformer encoders to segment Covid-19 infection into multiple classes. Moreover, a new skip-connection block is proposed to interpret and extract higher-level features from the encoder features and pass them to the decoder layers to preserve the finer details associated with the Multi-classes Covid-19 infection. Finally, transfer learning and freezing strategies are exploited to segment complicated task from very limited data.

3 The proposed approach

Our proposed approach consists of two main components. First, we propose a CNN-Transformer based architecture, which is described in detail in Sect. 3.1. Second, a Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss function is proposed. This is described in Sect. 3.2. Finally, the evaluation metrics are described in Sect. 3.3.

3.1 MB-TrAttUnet architecture

The proposed MB-TrAttUnet architecture is summarized in Fig. 1. Our proposed MB-TrAttUnet is an Encoder-Decoder architecture where the encoder consists of two components: Transformer path and Encoders Fusion Block (EFB). On the other hand, the decoder is an Att-Unet-like decoder consisting of Attention Gates and Deconvolution blocks. To connect the encoder and the decoder, an MBSC layer is proposed. Figure 2 shows the detailed structure of our proposed MB-TrAttUnet architecture.

3.1.1 Transformer path

As shown in Fig. 2, the input image $x \in \mathbb{R}^{H \times W \times C}$, where H , W and C are the height, width, and input channels, respectively, is fed into two paths: Unet-like path and Transformer path. In the **Transformer path**, x is divided into uniform non-overlapping 2D patches $x_v = [x_v^1; x_v^2; \dots; x_v^N]$, where each patch is represented by a row vector $x_v^i \in \mathbb{R}^{S^2 \times C}$, where $(S \times S)$ is the spatial size of the patch and N is the number of the patches $N = (H \times W)/S^2$. These patches are projected into Embedding space z_0 using a linear transformation $E \in \mathbb{R}^{(S^2 \cdot C) \times K}$, where K is the dimension of the embedding space, which is fixed for all of the transformer layers. z_0 is defined by:

$$z_0 = [x_v^1 E; x_v^2 E; \dots; x_v^N E] \tag{1}$$

The embedded features $z_0 \in \mathbb{R}^{N \times K}$ are fed into Transformer layers similar to Dosovitskiy et al. (2020); Vaswani et al. (2017). As shown in Fig. 4c, the Transformer layer consists of two Layernorm (LN) blocks, a Multi-Head Self-Attention (MSA) block, a multilayer perceptron (MLP) block and residual connections. For the Transformer layer (l), the embedded input features z_{l-1} are fed into Layernorm (LN), followed by a Multi-Head Self-Attention block, which is then summed with z_{l-1} by a residual connection, as shown in equation (2):

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

The embedded features of z_{l-1} passed by the first LN are denoted by $s = LN(z_{l-1})$.

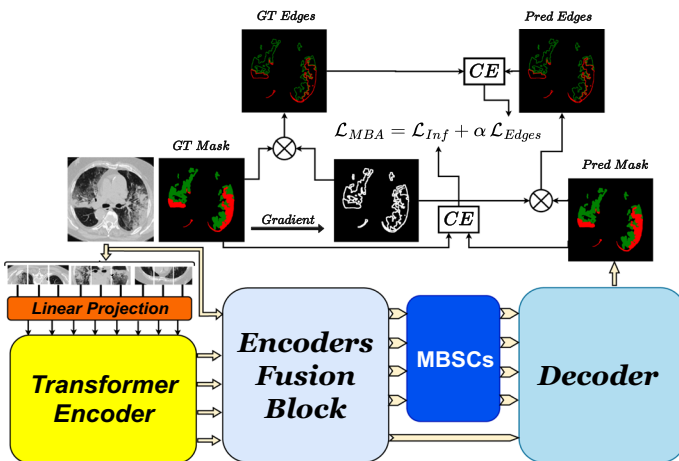


Fig. 1 The summary of our proposed EMB-TrAttUnet approach

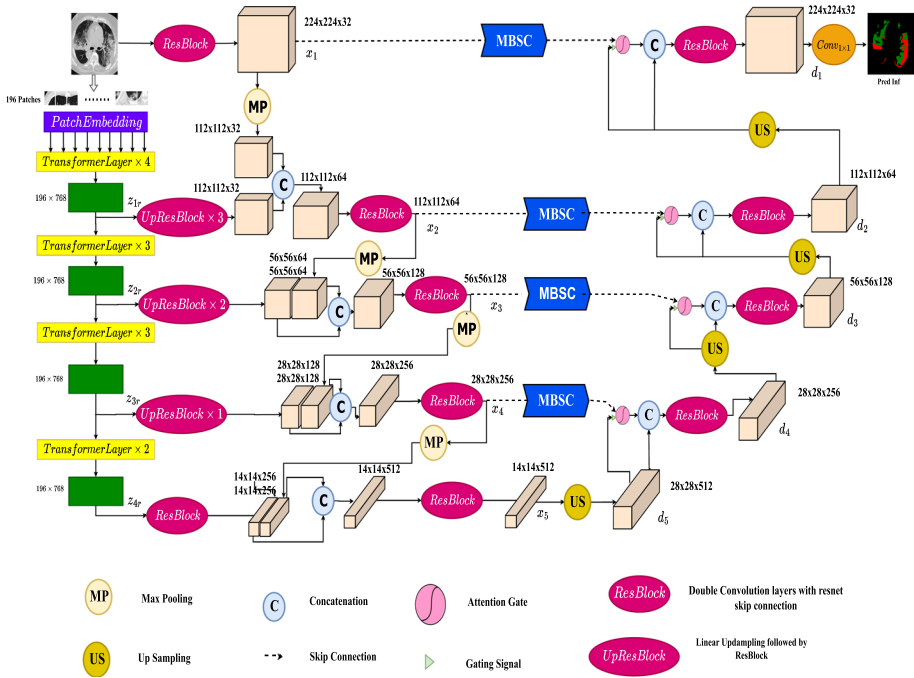


Fig. 2 Our proposed MB-TrAttUnet architecture

These features are processed by h heads. Multi-head attention allows the model to jointly attend to information from different representational subspaces. Each head generates attended features with a dimension of K/h . MSA is defined by:

$$MSA = [SA_1(s); SA_2(s); \dots ; SA_h(s)] \times U_{msa} \tag{3}$$

where SA_1, SA_2, \dots, SA_h are self-Attention results provided by the individual heads, and $U_{msa} \in \mathbb{R}^{K \times K}$ is the global projection matrix for the SA features.

The z'_i is fed into LayerNorm block followed by MLP block, and then summed by z_{i-1} through the residual connection, as depicted in equation (4):

$$z_i = MLP(LN(z'_i)) + z'_i \tag{4}$$

where MLP consists of two linear layers with a GELU nonlinearity. The first linear layer ($MLP_1 \in \mathbb{R}^{K \times K_{MLP}}$) projects $LN(z'_i)$ into K_{MLP} , then the second linear layer ($MLP_2 \in \mathbb{R}^{K_{MLP} \times K}$) projects it back onto K features.

In our approach, the number of layers of the Transformer, L , is set to 12, h to 12, $K = 786$ and $K_{MLP} = 3072$. The input image x size is $224 \times 224 \times 3$ ($W = H = 224$) and the transformer patch size is $S^2 = 16 \times 16$ pixels. Consequently, the number of patches is 196. To obtain diversity of learned features from different Transformer layers (levels), the embedded features of layers 4, 7, 10 and 12 are selected. These layers are denoted as Tr_1, Tr_2, Tr_3 and Tr_4 , respectively. Consequently, four layers from the Transformer path were injected into the Encoders Fusion Block (EFB), all of which are of shape 196×786 . To obtain a 3D tensors, z_i is reshaped to $14 \times 14 \times 786$, since $14 \times 14 = 196$. The reshaped

features of the embedded features of Tr_1, Tr_2, Tr_3 and Tr_4 are denoted by z_{1r}, z_{2r}, z_{3r} and z_{4r} , respectively.

3.1.2 Encoders fusion block (EFB)

Encoders Fusion Block combines the classical CNN features and Transformer features through ResBlocks in a hierarchical manner. Its main goal is to enrich the representations of the encoder at several stages. To inject the transformer features into different layers of the EFB and combine them with the CNN features, UpResBlock is introduced as depicted in Fig. 4-b. UpResblock consists of linear upsampling followed by ResBlock, which is depicted in Fig. 4-a. ResBlock consists of two consecutive 3 by 3 convolutional blocks, each followed by Batch Normalization and ReLU activation function. In addition, the input is summed with the output of the two convolutional layers using the residual connection, which consists of a 1 by 1 convolutional block, followed by Batch Normalization and ReLU activation function, as shown in equations (5) and (6):

$$x_{out_1} = ReLU(BN(Conv3 \times 3_1(x_{in}))) \tag{5}$$

$$x_{out} = ReLU(BN(Conv3 \times 3_2(x_{out_1}))) + ReLU(BN(Conv1 \times 1(x_{in}))) \tag{6}$$

where x_{out_1} and x_{out} are the outputs of the first 3 by 3 convolutional block and the whole ResBlock, respectively. $Conv3 \times 3_1, Conv3 \times 3_2$ and $Conv1 \times 1$ are convolutional kernels that transform the input number of channels into C_{out} .

The injected transformer features into the EFB are denoted by $z_{up_1}, z_{up_2}, z_{up_3}$ and z_{up_4} , respectively, which are defined by:

$$z_{up_1} = UpResBlock(UpResBlock(UpResBlock(z_{1r}))) \tag{7}$$

$$z_{up_2} = UpResBlock(UpResBlock(z_{2r})) \tag{8}$$

$$z_{up_3} = UpResBlock(z_{3r}) \tag{9}$$

$$z_{up_4} = ResBlock(z_{4r}) \tag{10}$$

Equations (7), (8) and (9) illustrate the number of UpResBlocks required to match the output of the transformer layers to the corresponding EFB level, using three, two, and one UpResBlock for Tr_1, Tr_2 and Tr_3 , respectively. For the Tr_4 layer, ResBlock is used instead of UpResBlock since no upsampling is required here to match the last EFB level, as depicted in equation (10).

On the other hand, the Encoders Fusion Block has five layers which will be denoted by Un_1, Un_2, Un_3, Un_4 and Un_5 , respectively. The first layer uses ResBlock on the input image $x \in \mathbb{R}^{H \times W \times C}$ to obtain the first EFB feature maps as shown in equation (11).

$$x_1 = ResBlock(x) \tag{11}$$

The second, third, fourth and fifth EFB layers combine the transformer features with the max-pooled features of the previous EFB layer as shown in equations (12), (13), (14) and (15):

$$x_2 = ResBlock([z_{up_1}; MP(x_1)]) \tag{12}$$

$$x_3 = ResBlock([z_{up_2}; MP(x_2)]) \tag{13}$$

$$x_4 = ResBlock([z_{up_3}; MP(x_3)]) \tag{14}$$

$$x_5 = ResBlock([z_{up_4}; MP(x_4)]) \tag{15}$$

where Max-Pooling layers (MP) reduces the spatial demensionality into half.

3.1.3 Multi-branches skip connection (MBSC)

Since the encoder of the proposed MB-TrAttUnet architecture generates composite and richer features by combining CNN and Transformer features, a more efficient skip connection is required to extract higher features and pass them to the decoder. As shown in Fig. 2, the encoder and decoder are connected thought Multi-Branches Skip Connections (MBSCs), where MBSC is depicted in Fig. 5. In detail, the proposed MBSC consists of four parallel paths, where each path performs different convolutional operations to learn the salient parts features from variety of feature sources (CNN and Transformer) x_i . The first path consists of a 1 by 1 convolutional block followed by a BN layer. The second path consists of an Average Pooling (AP) block, followed of a 1 by 1 convolutional Block. The third path consists of three consecutive convolutional kernels, which are 1 by 3, 3 by 1 and 3 by 3, each followed by BN layer. The fourth path consists of five consecutive convolutional kernels, which are 1 by 3, 3 by 1, 1 by 5, 5 by 1 and 5 by 5, each followed by BN layer. It should be noted that each path reduces the number of input channels into half in the last convolutional layer. The four paths are joined by concatenation, resulting a doubled number of feature maps of the input features (x_i). The concatenated features are passed into 3 by 3 convoutional kernel followed by BN layer, which reduces the number of feature maps into half (match the x_i number of feature maps). Finally, a skip residual is used as depicted in Fig. 5 and followed by a ReLU activation function. For an input x_i and four MBSC paths output $Skip_{i_1}$, $Skip_{i_2}$, $Skip_{i_3}$, and $Skip_{i_4}$, respectively. The skip connection on the i^{th} level $Skip_i$ is defined by:

$$Skip_{c_i} = [Skip_{i_1}; Skip_{i_2}; Skip_{i_3}; Skip_{i_4}] \tag{16}$$

$$Skip_i = ReLU(BN(Conv3 \times 3(Skip_{c_i})) + BN(Conv1 \times 1(x_i))) \tag{17}$$

where $Skip_{c_i}$ and $Skip_i$ are the concatenation of the four Skip paths output and the Skip features passed to the decoder, respectively. In the proposed MBSC the following properties are considered:

- The $n \times n$ convolution block is split into two consecutive convolution block of $1 \times n$ and $n \times 1$, to reduce the number of parameters.

- The last branch is implemented by increasing the kernel size to capture progressively higher features.
- Since the higher number of parameters block in the third and fourth branches ($n \times n$ convolutional block) are the last convolutional blocks, the number of channels of this block is reduced to the half to decrease the number of parameters.
- All branches features are concatenated, and then passed to a 3 by 3 convolutional block to summarize and capture the most important features from all of the four branches. Finally, the residual connection is added, and then passed to the ReLU activation function. As can be seen in Fig 5, the MBSC blocks do not change the dimension of the 3D input tensors.

3.1.4 The decoder of MB-TrAttUnet architecture

The decoder of MB-TrAttUnet is an Att-Unet-like decoder, as shown in Fig. 2. In detail, the bottleneck feature maps x_5 of the encoder are fed into the first expansion layer of the decoder. First, x_5 is up-sampled using a linear transformation to obtain d_5 , and then passed to the decoder as shown in equation (18). On the other hand, the encoder feature maps x_1, x_2, x_3 and x_4 are forwarded to the decoder layers of MB-TrAttUnet via MBSC connections, as shown in Fig. 2. Following the Att-Unet architecture Oktay et al. (2018), three linear upsampling layers (US), four decoder layers, four Attention Gates (AG), and four ResBlocks are used, as shown in the following equations:

$$d_5 = US(x_5) \quad (18)$$

$$d_4 = ResBlock([AG(MBSC_4(x_4), US(x_5)); US(x_5)]) \quad (19)$$

$$d_3 = ResBlock([AG(MBSC_3(x_3), US(d_4)); US(d_4)]) \quad (20)$$

$$d_2 = ResBlock([AG(MBSC_2(x_2), US(d_3)); US(d_3)]) \quad (21)$$

$$d_1 = ResBlock([AG(MBSC_1(x_1), US(d_2)); US(d_2)]) \quad (22)$$

where ‘;’ denotes features concatenation.

Finally, a 1 by 1 convolutional block is used to match the feature map dimension of d_1 to the infection mask prediction, which consists of three channels for the Multi-classes segmentation (Background, GGO, and Consolidation).

The Attention Gate (AG) is depicted in Fig. 3 (Oktay et al. 2018), and is defined as follows:

$$M_{att_i} = \psi_i(ReLU(BN(W_x MBSC_i(x_i)) + BN(W_g g_i))) \quad (23)$$

where W_x and W_g are two 1 by 1 convolution kernels performing linear transformations of the number of channels (c_x and c_g of x_i and g_i , respectively) to c_{int} . ψ_i consists of $W_{\psi_i} \in \mathbb{R}^{1 \times 1}$ followed by Batch Normalization (BN) and sigmoid activation function to learn the spatial attention coefficient M_{att_i} for each pixel. The obtained spatial coefficients M_{att_i} are applied to the skip feature maps of the encoder x_i .

$$x_{att} = M_{att_i} \otimes MBSC_i(x_i) \quad (24)$$

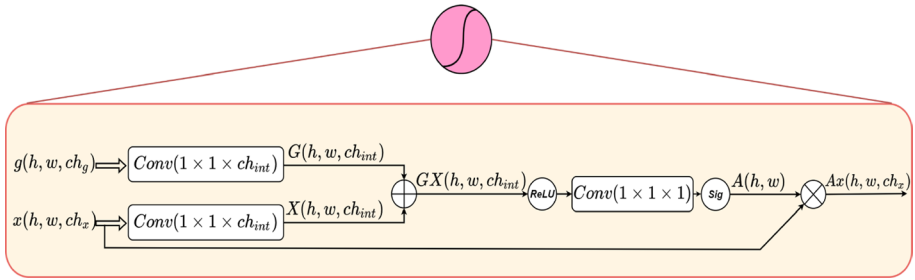


Fig. 3 Attention Gate block, where g is the gating signal and the x is the input feature maps. $A(h, w)$ is the obtained spatial attention, which is applied for all channels of the input feature maps (x)

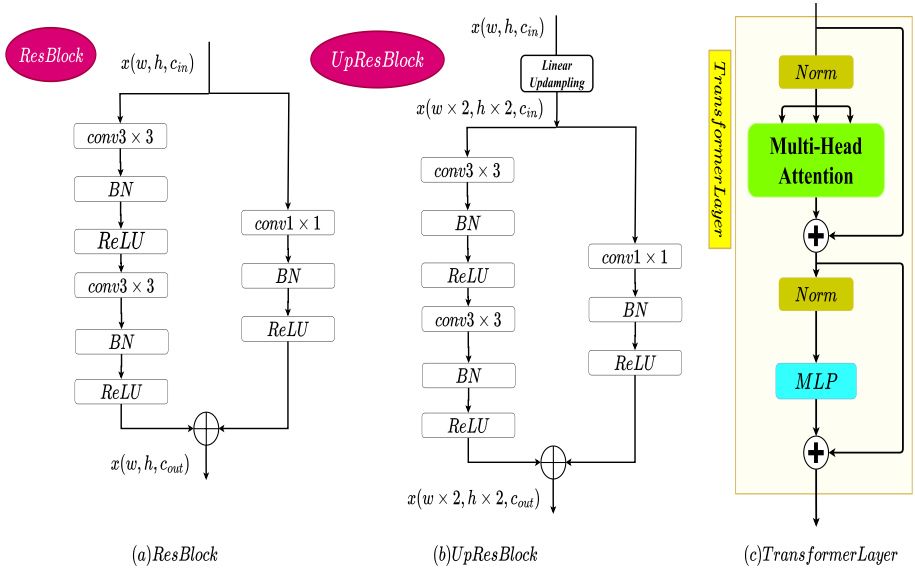


Fig. 4 Description of ResBlock, UpResBlock and TransformerLayer

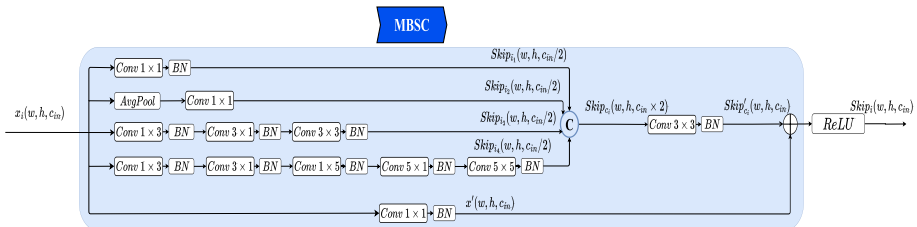


Fig. 5 Description of the proposed MBSC block

where \otimes is element-wise multiplication operator.

3.2 Multi-classes boundary aware cross-entropy (MBA-CE) loss function

In the field of medical imaging, objects can exhibit high variability in terms of their shapes, intensities, and positions (Bougourzi et al. 2023). This can make it challenging for traditional loss functions, such as Cross-Entropy, to accurately segment the target boundaries, resulting in fuzzy segmented object boundaries (Qin et al. 2019; Bougourzi et al. 2023). To address this issue, we propose a new Edge Aware Loss function for multi-class segmentation that specifically concentrates on segmenting infection boundaries, improving the separation between different classes, and giving greater attention to minority classes. Our proposed Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss Function is defined as following:

$$\mathcal{L}_{MBA} = \mathcal{L}_{Inf} + \alpha \mathcal{L}_{Edges} \tag{25}$$

Here, \mathcal{L}_{Inf} and \mathcal{L}_{Edges} are the Multi-classes infection and edges losses, respectively. Moreover, the value of the weight α is obtained experimentally and equals 2. By emphasizing the edges between infection types, the MBA-CE Loss Function encourages the model to focus on correctly predicting the boundaries between them, thereby improving its ability to distinguish between different types of infections. For the ground-truth mask width (W) and height (H), \mathcal{L}_{Inf} loss is defined by:

$$\mathcal{L}_{Inf} = - \sum_{i=1}^{W \cdot H} \sum_{c=1}^M G_{i_c} \log(p_{i_c}) \tag{26}$$

where G_i and p_i are the infection types ground-truth probability distribution and the model prediction probability distribution of pixel i , respectively. M is the total number of classes.

The \mathcal{L}_{Edges} is defined by:

$$\mathcal{L}_{Edges} = - \sum_{i=1}^{W \cdot H} \sum_{c=1}^M E_{i_c} \log(q_{i_c}) \tag{27}$$

where, E_i and q_i are the boundary ground-truth probability distribution and boundary prediction probability distribution of pixel i , respectively. As depicted in Fig. 6, the edge

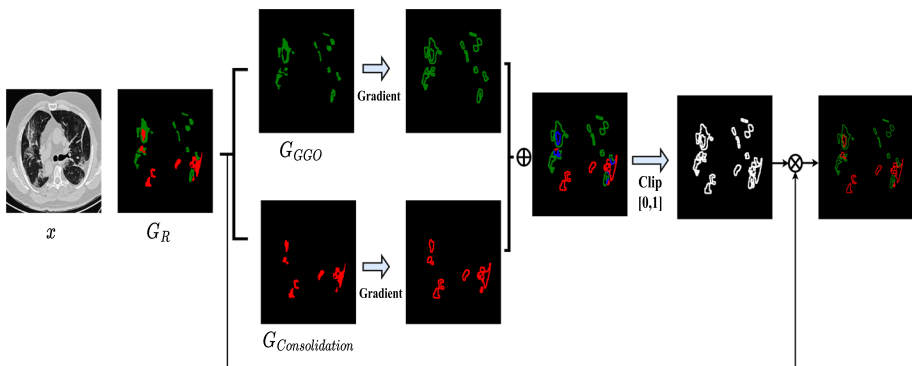


Fig. 6 The summary of Ground-truth Boundary edges process. Green and red colors represent GGO and Consolidation infection types, respectively

ground-truth E is obtained by multiplying the ground-truth mask G with the union of the morphological gradient applied to the ground-truth masks of GGO and Consolidation. In other words:

$$E = G \otimes (\text{Grad}(G_{GGO}) \cup \text{Grad}(G_{Con})) \quad (28)$$

Here, $\text{Grad}()$ denotes the morphological gradient operation.

Similarly, the edge prediction maps q is obtained by multiplying the model output map p with the union of the morphological gradient applied to the ground-truth masks of GGO and Consolidation. In other words:

$$q = p \otimes (\text{Grad}(G_{GGO}) \cup \text{Grad}(G_{Con})) \quad (29)$$

3.3 Evaluation measurements

The following evaluation metrics are used: F1-score (F1-S), Dice-score (D-S), Intersection over Union (IoU), and Hausdorff Distance (HD95) in pixels, to compare between different architectures and items. It should be noted that each infection type is evaluated using the four metrics.

It should be noted that F1-score and IoU (Jaccard Index) are micro metrics, where they are calculated for all images at one time using True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). However, the Dice-score is the macro version of the $F1 - score$. For N training or test images, it is defined by:

$$\text{Dice-score} = 100 \cdot \frac{1}{N} \sum_{i=1}^N 2 \cdot \frac{TP_i}{2 \cdot TP_i + FP_i + FN_i} \quad (30)$$

where TP_i , TN_i , FP_i and FN_i are True Positives, True Negatives, False Positives and False Negative for the i th image, respectively.

4 Experiments and results

4.1 Datasets and splitting configurations

Table 1 displays datasets labeled for multi-class COVID-19 infection segmentation. To replicate the urgency of pandemic situations, our approach is assessed in two scenarios: Scenario 1, depicting standard data availability, and Scenario 2, concentrating on the influence of limited training data on the segmentation approaches performance. Crafting an effective approach in the latter scenario is especially vital in pandemic situations.

In total, we performed four splitting configurations, in the first two, Dataset_1 RADIOLOGISTS (2019) and Dataset_2 RADIOLOGISTS (2019) are used. In the first splitting configuration, Dataset_1 and 50% of Dataset_2 are used as training data and the remaining 50% of Dataset_2 is used as test data. In the second splitting configuration, the training data are reduced to half to study the efficiency of the proposed approach in a more challenging few-shot learning strategy (less training data are available). More specifically, for the second splitting configuration experiment only 50% of CT-scans from Dataset_1 and

Table 1 The available Covid-19 Segmentation datasets with Multi-classes Covid-19 Infection Segmentation (GGO and Consolidation)

Name	Dataset	#CT-scans	#Slices	Classes	Settings
Dataset_1	Segmentation dataset nr. 2 RADIOLOGISTS (2019)	9	829	Background GGO Consolidation	Resolution: 512 × 512
Dataset_2	COVID-19 CT segmen- tation RADIOLOGISTS (2019)	40	100	Background GGO Consolidation	Resolution: 512 × 512
Dataset_3	CC-CCII Liu et al. (2020)	150	750	Background GGO Consolidation	Slicing: 1.0 mm Resolution: 512 × 512

25% of slices from Dataset_2 are used as training data. In the third and fourth configuration splittings, Dataset_3 is used. Where in the third one, 60% of Dataset_3 is used as training data and the remaining 40% is used as testing data. The fourth configuration corresponds to the second evaluation scenario where only 15% of Dataset_3 is used as training data.

Table 2 provides an overview of the four evaluation configurations. It's important to note that not all slices necessarily exhibit infection, mirroring the real-world scenario where the prevalence of infected slices can vary across cases. Additionally, an infected slice may contain either a single infection type (GGO or Consolidation) or both, a critical aspect for tracking disease progression and severity. In the first two splitting configurations (as shown in Table 2), it's evident that the number of infected slices constitutes approximately half of the total slices in the training dataset. However, when considering all four splitting configurations, it becomes apparent that the number of slices infected with Consolidation is

Table 2 Data splits for the four splitting configurations

Splitting configuration	Phase	Involved dataset	Total # slices	# Infected slices	#Slices with GGO	#Slices with consolidation
First	Train	Dataset_1 50% of Dataset_2	879 (9+20 CT-scans)	422	345	272
	Test	50% of Dataset_2	50 (20 CT-scans)	50	50	40
Second	Train	50% of Dataset_1 25% of Dataset_2	304 (5+10 Ct-scans)	183	181	90
	Test	50% of Dataset_2	50 (20 CT-scans)	50	50	40
Third	Train	60% of Dataset_3	445 (90 CT-scans)	321	315	192
	Test	40% of Dataset_3	305 (60 CT-scans)	228	225	150
Fourth	Train	15% of Dataset_3	110 (942 CT-scans)	80	78	48
	Test	40% of Dataset_3	305 (60 CT-scans)	228	225	150

The details include the used dataset for each splitting configurations, number of slices (in brackets the number of CT-scans), infected slices among the total, slices with GGO infection, and slices with consolidation specified for each evaluated splitting configuration

notably lower, especially in the second evaluation scenario. This imbalance in the training data poses a significant challenge for multi-class Covid-19 infection segmentation.

4.2 Experimental setup

Paszke et al. (2019) Library with NVIDIA RTX A5000 GPU with 24 GB of memory is used to train and test our approach and the comparison ones. The used machine has 11th Gen Intel(R) Core(TM) i9-11900KF (3.50GHz) CPU and 64 of RAM. All models are trained for 60 epochs with batch size of 6 images and a learning rate schedule where the initial $lr = 0.01$ and then it decays by 0.1 twice after 30 and 50 epochs, respectively. In the training, two augmentation techniques are used: (i) Active Rotation using random angle between $[-35^\circ, 35^\circ]$, and (ii) random horizontal and vertical flipping. Multi-classes CE loss is used for all experiments as the loss function. To investigate the stability of the each architecture performance, each experiment is repeated five times and then the mean and standard deviation are declared for each evaluation metric.

4.3 Experimental results of the first splitting configuration

Table 3 summarizes the experimental results of the first splitting configuration. In order to compare our approach performance with state-of-the-art architectures, three CNN-Based Segmentation architectures (Unet Ronneberger et al. (2015), Attention-Unet Oktay et al. (2018) and Nested-Unet Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)), and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2022), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)) are evaluated as shown in Table 3. These results show that our proposed approach outperforms the comparison approaches by considerable margin for both GGO and Consolidation segmentation. For the GGO segmentation results, most of the comparison architectures achieved close performance (except for CopleNet and MTUnet, which achieved the lowest performance). The best comparison approach is MISSFormer for both GGO and Consolidation segmentation. Our approach outperformed the performance of MISSFormer by 1.41%, 6.75%, 1.6% and 3.97 for F-S, D-S, IoU and HD95 for GGO segmentation, respectively. By looking on the standard deviation values of the best comparison architecture, we find that our approach not only outperforms the best competing architecture, but also exhibits more stable performance at different running.

On the other hand, MISSFormer architecture performed the best in all metrics compared to the other state-of-the-art approaches for Consolidation segmentation. However, our approach outperformed its performance by 7%, 7.15%, 6.81% and 0.9 for F-S, D-S, IoU, and HD95 respectively. Compared with the results of GGO, it is noticed that the performance margin between our approach and the state-of-the-art architectures is larger. This proves that our approach can perform accurately even in unbalanced class segmentation scenarios, as is the case in most of Multi-class segmentation tasks in medical imaging domain. The results of GGO and Consolidation show that our approach has a high ability to segment different types of Covid-19 infections.

Table 3 The experimental results of the First Splitting Configuration for Multi-classes Covid-19 infection segmentation from CT-scans

Architecture	GGO				Consolidation			
	FI-S ↑	D-S ↑	IoU ↑	HD95 ↓	FI-S ↑	D-S ↑	IoU ↑	HD95 ↓
Unet (2015)	65.81 ± 1.26	50.13 ± 1.31	49.06 ± 1.41	33.58 ± 2.33	31.35 ± 12.96	15.45 ± 5.66	19.26 ± 8.76	37.81 ± 6.81
Att-Unet (2018)	64.81 ± 1.89	50.44 ± 1.35	47.97 ± 2.06	34.51 ± 3.64	39.04 ± 6.81	19.26 ± 3.55	24.48 ± 5.31	37.18 ± 5.14
Unet++ (2018)	65.69 ± 1.29	51.65 ± 4.12	48.92 ± 14.2	28.51 ± 2.94	31.31 ± 6.67	16.86 ± 4.48	18.75 ± 4.73	39.20 ± 7.12
CopleNet (2020)	60.44 ± 1.54	46.25 ± 3.13	43.33 ± 1.61	37.15 ± 7.12	29.70 ± 10.29	16.46 ± 4.76	17.90 ± 7.52	40.93 ± 5.46
AnamNet (2021)	65.10 ± 3.56	51.69 ± 4.81	48.36 ± 3.82	31.11 ± 4.95	31.97 ± 6.12	18.06 ± 4.61	19.18 ± 4.36	38.46 ± 6.17
SCOATNET (2021)	65.77 ± 3.28	50.80 ± 4.63	49.09 ± 3.56	34.16 ± 3.75	43.52 ± 1.67	23.32 ± 2.07	27.83 ± 1.38	36.47 ± 4.44
SwinUnet (2021)	60.99 ± 1.24	44.66 ± 3.16	43.89 ± 1.27	37.54 ± 4.91	29.51 ± 18.22	14.41 ± 7.96	18.65 ± 12.57	41.30 ± 7.17
MTUnet (2022)	59.99 ± 5.86	44.54 ± 7.39	41.01 ± 5.90	36.13 ± 7.35	27.25 ± 13.41	18.10 ± 8.60	16.41 ± 8.18	38.88 ± 7.01
UCTransNet (2022)	67.04 ± 1.48	51.77 ± 3.62	50.44 ± 1.66	31.65 ± 3.50	42.44 ± 7.97	22.83 ± 4.92	27.25 ± 6.30	37.69 ± 5.69
MISSFormer (2022)	68.65 ± 2.46	52.39 ± 3.09	52.32 ± 2.84	24.31 ± 1.33	53.26 ± 2.14	30.08 ± 1.75	36.32 ± 1.99	35.21 ± 4.06
EMB-TransUnet	70.06 ± 0.03	59.14 ± 0.87	53.92 ± 0.04	20.63 ± 2.34	60.26 ± 0.92	37.23 ± 0.97	43.13 ± 0.94	34.31 ± 2.49

Our proposed approach, three CNN-based segmentation architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)) and four and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2022), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)) are evaluated. The best performance and the second best performance are marked by *textit* and *textit*, respectively

The italic shows the best comparison approach

4.4 Experimental results of the second splitting configuration

Table 4 summarizes the experimental results of the second evaluation scenario, where half of the training data of first splitting configuration is used to investigate the efficiency of the proposed approach in more critical few shot-shot learning scenario. To compare the performance of our approach with state-of-the-art architectures, three CNN-based segmentation architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)) and four and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2022), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)), are evaluated as shown in Table 4. Similar to what noticed in the first evaluation scenario, our approach outperforms all of the comparison architectures on all of the four evaluation metrics for both GGO and Consolidation segmentation.

In more details, the three CNN baseline architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)) performed similar to the first scenario for GGO segmentation. In contrast, the three CNN baseline architectures failed to segment the Consolidation infection type. Similar behavior is observed for the CopleNet and AnamNet architectures. Table 4 shows that the SCOATNet architecture outperforms the other comparison architectures in segmenting GGO for F1-S, D-S and IoU metrics. Comparing the performance of SCOATNet in the first and second evaluation scenarios (Tables 3 and 4), it is noticed that SCOATNet achieves similar results for GGO segmentation. However, the segmentation results for Consolidation dropped to one-third by reducing the training data to half. For GGO segmentation in the second evaluation scenario, our approach outperforms SCOATNet architecture by 1.91%, 5.67%, and 1.65% for F-S, D-S, and IoU, respectively and outperforms MISSFormer by 7.12 for the HD95 metric. For the transformer based architecture, it is also noticed that their performance in segmentation GGO is decreased. From Consolidation results, it is noticed that the proposed approach surpasses MISSFormer architecture (the best comparison architecture) by a large margin: 8.68%, 3.7%, 7.26% and 1.8 for F-S, D-S, IoU and HD95, respectively. On the other hand, the comparison between the standard deviation values of our approach and the MISSFormer architecture in the second evaluation scenario shows that the performance of our approach is consistent across different runs.

The comparison between the results of the first and the second evaluation scenarios (from the first and second splitting configurations) shows that our approach performs consistently, although the training data in the second evaluation scenario are reduced to half. This proves the efficiency of the proposed components, which will be discussed in detail in the ablation study section.

4.5 Experimental results of the third splitting configuration

Table 5 summarizes the results of the third splitting configuration, where our proposed approach is compared with three baseline CNN architectures, three state-of-the-art approaches for Covid-19 segmentation, and four Transformer-based architectures similar to the first and second splitting configurations. From these results, it is noticed that SCOATNet and UCTransNet achieve similar performance in segmenting both classes (GGO and Consolidation), outperforming the other comparison approaches. However, our approach demonstrates its efficiency by outperforming the best competitor for each metric. For GGO

Table 4 The experimental results of the Second Splitting Configuration for Multi-classes Covid-19 infection segmentation from CT-scans.

Architecture	GGO				Consolidation			
	F1-S ↑	D-S ↑	IoU ↑	HD95 ↓	F1-S ↑	D-S ↑	IoU ↑	HD95 ↓
Unet (2015)	64.86 ± 1.59	51.38 ± 1.95	48.01 ± 1.73	38.58 ± 2.72	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	–
Att-Unet (2018)	64.83 ± 2.06	50.64 ± 2.20	47.99 ± 2.24	37.22 ± 3.51	0.65 ± 1.30	0.55 ± 1.11	0.33 ± 0.66	50.12 ± 3.31
Unet++ (2018)	64.80 ± 1.35	51.70 ± 2.34	47.94 ± 1.48	36.66 ± 5.14	3.62 ± 7.08	2.27 ± 4.45	1.98 ± 3.88	47.56 ± 4.58
CopleNet (2020)	58.23 ± 4.71	36.33 ± 6.35	41.23 ± 4.56	40.75 ± 6.18	1.79 ± 2.61	0.94 ± 1.28	0.92 ± 1.35	49.28 ± 6.12
AnamNet (2021)	58.40 ± 3.42	46.36 ± 5.66	41.33 ± 3.47	41.33 ± 5.47	0.07 ± 0.14	0.04 ± 0.07	0.03 ± 0.07	51.98 ± 7.75
SCOATNET (2021)	<i>67.31 ± 1.47</i>	<i>53.49 ± 1.74</i>	<i>50.74 ± 1.67</i>	<i>30.87 ± 3.79</i>	16.22 ± 12.68	7.74 ± 6.37	9.34 ± 7.42	44.79 ± 4.96
SwinUnet (2021)	56.24 ± 6.13	37.78 ± 6.03	39.36 ± 5.74	44.01 ± 2.12	23.41 ± 11.77	13.92 ± 6.50	13.77 ± 7.72	42.69 ± 2.43
MTUnet (2022)	54.32 ± 1.93	36.38 ± 1.94	37.31 ± 1.81	41.44 ± 1.72	27.58 ± 11.11	19.63 ± 7.95	16.45 ± 7.15	45.46 ± 3.83
UCTransNet (2022)	63.22 ± 2.99	46.70 ± 3.80	46.29 ± 3.21	36.52 ± 5.24	18.15 ± 7.46	11.366 ± 2.73	10.17 ± 4.59	42.08 ± 8.32
MISSFormer (2022)	66.86 ± 1.59	48.58 ± 1.20	50.24 ± 1.79	28.75 ± 1.63	<i>42.43 ± 5.44</i>	26.32 ± 2.54	<i>27.07 ± 4.21</i>	37.09 ± 3.89
EMB-TrAttUnet	69.22 ± 0.55	59.16 ± 1.33	52.39 ± 0.24	21.63 ± 2.56	51.11 ± 0.59	30.02 ± 1.55	34.33 ± 0.53	35.29 ± 1.68

Our proposed approach, three CNN-based segmentation architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)) and four and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2021), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)) are evaluated. The best performance and the second best performance are marked by *textit* and *textit*, respectively

The italic shows the best comparison approach

Table 5 The experimental results of the Third Splitting Configuration for Multi-classes Covid-19 infection segmentation from CT-scans

Architecture	GGO				Consolidation			
	F1-S ↑	D-S ↑	IoU ↑	HD95 ↓	F1-S ↑	D-S ↑	IoU ↑	HD95 ↓
Unet (2015)	67.04 ± 1.25	32.79 ± 1.55	50.43 ± 1.40	14.11 ± 0.25	71.05 ± 1.24	24.16 ± 1.11	55.12 ± 1.50	8.71 ± 0.37
Att-Unet (2018)	65.34 ± 2.27	32.20 ± 1.49	48.57 ± 2.53	14.25 ± 0.93	71.12 ± 1.18	24.42 ± 0.78	55.20 ± 1.41	8.93 ± 0.54
Unet++ (2018)	66.73 ± 1.04	33.25 ± 0.87	50.08 ± 1.18	14.02 ± 0.44	71.83 ± 1.76	24.83 ± 1.22	56.07 ± 2.14	8.96 ± 0.63
CopleNet (2020)	55.57 ± 3.80	24.56 ± 2.96	38.57 ± 3.68	18.63 ± 0.84	67.22 ± 2.82	22.26 ± 1.60	50.69 ± 3.22	9.77 ± 0.64
AnamNet (2021)	66.22 ± 0.62	31.49 ± 0.63	49.51 ± 0.70	15.22 ± 0.84	70.09 ± 0.64	24.11 ± 0.55	53.96 ± 0.76	8.80 ± 0.27
SCOATNET (2021)	<i>69.63 ± 1.31</i>	<i>35.57 ± 0.47</i>	<i>53.42 ± 1.55</i>	14.08 ± 0.83	73.69 ± 0.62	26.32 ± 0.50	58.34 ± 0.78	8.41 ± 0.51
SwinUnet (2021)	55.65 ± 1.93	25.10 ± 0.75	38.58 ± 1.84	15.45 ± 1.00	64.34 ± 1.30	20.34 ± 0.77	4744 ± 141	9.28 ± 0.20
MTUnet (2022)	55.02 ± 3.59	28.10 ± 1.69	38.03 ± 3.38	19.44 ± 2.28	64.65 ± 2.95	22.54 ± 0.75	4783 ± 320	10.97 ± 1.36
UCTransNet (2022)	69.07 ± 0.91	35.28 ± 1.21	52.76 ± 1.05	<i>14.05 ± 0.54</i>	<i>73.79 ± 0.31</i>	<i>26.57 ± 0.55</i>	<i>58.47 ± 0.39</i>	<i>8.02 ± 0.23</i>
MISSFormer (2022)	52.98 ± 3.59	24.78 ± 2.58	36.12 ± 3.30	16.55 ± 1.70	62.79 ± 4.65	20.95 ± 2.33	45.92 ± 4.81	10.58 ± 0.72
EMB-TransUnet	69.74 ± 0.53	37.23 ± 0.53	53.54 ± 0.63	13.18 ± 1.25	75.93 ± 1.16	27.39 ± 0.32	61.22 ± 1.49	7.80 ± 0.67

Our proposed approach, three CNN-based segmentation architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)) and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2022), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)) are evaluated. The best performance and the second best performance are marked by textbf and textit, respectively

The italic shows the best comparison approach

segmentation results, our approach achieves better performance than the top competitor by 0.11%, 1.66%, 0.12%, and 0.87 for F-S, D-S, IoU, and HD95, respectively. Similarly, for Consolidation segmentation, our approach outperforms the best competitor by 2.14%, 1.66%, 0.12%, and 0.87 for F-S, D-S, IoU, and HD95, respectively. In this splitting configuration, with a substantial training dataset from nearly 100 CT scans, it is noticeable that some state-of-the-art approaches come close to our approach, but our approach still exhibits superiority. Furthermore, when compared to the previous two splitting configurations, UCTransNet emerges as one of the top competitors instead of the MISSFormer architecture, which showed lower performance than the best competitors. On the other hand, SCOATNET continues to perform as one of the best competitors. This demonstrates that the performance of certain architectures can vary depending on the scenario and the dataset used. In contrast, the proposed EMB-TrAttUnet consistently delivers strong performance across different scenarios.

4.6 Experimental results of the fourth splitting configuration

Table 6 summarizes the results obtained from the fourth splitting configuration. These results represent the second evaluation scenario of the third splitting configuration, in which the training dataset is reduced to a fourth. The comparison with state-of-the-art approaches reveals the superiority of our approach. When compared with the results from the first evaluation scenario in Table 5, it is noticed that the performance of all approaches has significantly decreased. For example, the F1-score of the top two competitors (SCOATNET and UCTransNet) has decreased by 12.34% and 11.11% for GGO segmentation, respectively. Similarly, their F1-score for Consolidation has decreased by 15.32% and 17.99%, respectively. In contrast, our approach demonstrates consistent performance, with the F1-score decreasing by only 4.2% and 3.86% for GGO and Consolidation, respectively.

4.7 Ablation study

In this section, the significance of each component of the proposed EMB-TrAttUnet approach is studied, which includes Transformer Encoder Path (TE), MBSC connection blocks (MBS), and Multi-classes Boundary Aware Loss Function (MBA). Tables 7 and 8 summarize the ablation study for the first and second evaluation scenarios correspond to the first and second splitting configurations, respectively. The Att-Unet architecture is considered as the baseline architecture for both evaluation scenarios. From Table 7, the comparison between the baseline architecture and the first ablation experiment shows that the Transformer path significantly improves the GGO and Consolidation segmentation results. This shows that the Transformer Path provides diversity of features in the encoding phase. In other words, the fusion of the global contextual and long-range dependencies features using the Transformer path and the local contextual features using the CNN Convolutional blocks provides finer details in the encoding phase, allowing to distinguish between infected, non-infected tissues and different infection types. In the second ablation experiment of Table 7, the results of the Att-Unet architecture with MBSC blocks are depicted. The results show that there is no improvement for GGO segmentation, but the results for Consolidation have dropped. The explanation of this result is that adding more Convolutional layers without rich and diverse features extraction tends to lead to more over-fitting. Experiment (3) shows that adding the Transformer path and MBSC connection blocks improves the results compared to both the baseline experiment and experiment (1) for both

Table 6 The experimental results of the Fourth Splitting Configuration for Multi-classes Covid-19 infection segmentation from CT-scans

Architecture	GGO				Consolidation			
	FI-S↑	D-S↓	IoU↑	HD95↓	FI-S↑	D-S↓	IoU↑	HD95↓
Unet (2015)	51.64 ± 3.35	21.75 ± 2.17	34.88 ± 3.01	19.65 ± 1.91	46.08 ± 7.61	12.98 ± 1.95	30.27 ± 6.60	11.88 ± 2.21
Att-Unet (2018)	54.41 ± 2.32	23.94 ± 1.53	37.40 ± 2.16	19.36 ± 1.41	46.86 ± 5.96	13.70 ± 1.82	30.80 ± 5.06	11.21 ± 0.96
Unet++ (2018)	54.46 ± 3.87	24.15 ± 2.26	37.52 ± 3.64	18.79 ± 0.88	51.13 ± 6.40	14.84 ± 2.18	34.60 ± 5.80	11.55 ± 1.01
CopleNet (2020)	45.18 ± 2.38	18.86 ± 1.48	29.21 ± 2.01	20.21 ± 1.34	39.13 ± 16.72	11.79 ± 5.12	25.52 ± 11.55	10.58 ± 0.28
AnamNet (2021)	51.44 ± 3.17	21.98 ± 1.92	34.69 ± 2.90	22.12 ± 1.41	46.39 ± 4.28	12.91 ± 1.78	30.30 ± 3.61	11.53 ± 0.67
SCOATNET (2021)	57.29 ± 1.89	26.73 ± 1.12	40.17 ± 1.87	19.33 ± 1.41	58.47 ± 4.97	18.13 ± 2.13	41.48 ± 4.90	10.56 ± 0.60
SwinUnet (2021)	15.15 ± 5.26	8.78 ± 4.45	11.25 ± 6.78	27.58 ± 4.65	8.18 ± 6.65	3.54 ± 4.45	6.56 ± 4.54	25.45 ± 5.78
MTUnet (2022)	45.11 ± 1.62	20.78 ± 0.85	29.14 ± 1.36	23.58 ± 0.50	48.18 ± 0.24	15.34 ± 1.04	31.76 ± 2.10	12.69 ± 0.91
UCTransNet (2022)	57.96 ± 3.86	28.31 ± 2.51	40.91 ± 3.84	20.17 ± 1.86	55.80 ± 2.50	17.72 ± 1.04	38.74 ± 2.39	12.64 ± 1.16
MISSFormer (2022)	32.76 ± 13.05	13.76 ± 5.75	20.26 ± 8.65	24.03 ± 2.19	16.64 ± 13.31	4.73 ± 3.84	9.67 ± 8.16	11.48 ± 4.78
EMB-TrAttUnet	65.54 ± 1.08	34.44 ± 0.91	48.76 ± 1.19	14.24 ± 0.75	72.07 ± 1.28	25.32 ± 0.93	56.36 ± 1.57	8.59 ± 0.98

Our proposed approach, three CNN-based segmentation architectures (Unet Ronneberger et al. (2015), Att-Unet Oktay et al. (2018), Unet++ Zhou et al. (2018)), three recent approaches that were designed to segment Covid-19 from the CT-scans (CopleNet Wang et al. (2020), AnamNet Paluru et al. (2021), and SCOATNet Zhao et al. (2021)) and four and four Transformer-Based architectures (SwinUnet Cao et al. (2021), MTUnet Wang et al. (2021), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022)) are evaluated. The best performance and the second best performance are marked by textbf and textit, respectively

The italic shows the best comparison approach

Table 7 Ablation study of the first evaluation scenario, in which the following components importance are investigated: Transformer Encoder path (TE), MBSC connection blocks (MBS), and Multi-classes Boundary Aware Loss Function (MBA)

Architecture	Ablation			GGO			Consolidation		
	TE	MBS	MBA	F1-S	D-S	IoU	F1-S	D-S	IoU
	Att-Unet (baseline)	✗	✗	✗	64.81 ± 1.89	50.44 ± 1.35	47.97 ± 2.06	39.04 ± 6.81	19.26 ± 3.55
(1) TrAttUnet	✓	✗	✗	67.83 ± 3.11	54.54 ± 0.38	50.65 ± 0.36	54.36 ± 0.71	30.15 ± 0.44	37.33 ± 0.66
(2) MB-AttUnet	✗	✓	✗	64.76 ± 2.06	48.42 ± 2.97	47.92 ± 2.26	30.35 ± 1.34	15.38 ± 5.74	18.61 ± 9.14
(3) MB-TrAttUnet	✓	✓	✗	69.04 ± 1.27	55.64 ± 1.86	52.73 ± 1.48	57.75 ± 1.75	31.77 ± 2.13	40.62 ± 1.72
(4) EMB-TrAttUnet	✓	✓	✓	70.06 ± 0.03	59.14 ± 0.87	53.92 ± 0.04	60.26 ± 0.92	37.23 ± 0.97	43.13 ± 0.94

Table 8 Ablation study of the second evaluation scenario, in which the following components importance are investigated: Transformer Encoder path (TE), MBSC connection blocks (MBS), and Multi-classes Boundary Aware Loss Function (MBA)

Architecture	Ablation			GGO			Consolidation		
	TE	MBS	MBA	F1-S	D-S	IoU	F1-S	D-S	IoU
Att-Unet (baseline)	✗	✗	✗	64.83 ± 2.06	50.64 ± 2.20	47.99 ± 2.24	0.65 ± 1.30	0.55 ± 1.11	0.33 ± 0.66
(1) TrAttUnet	✓	✗	✗	65.79 ± 0.50	57.45 ± 0.53	49.02 ± 0.55	9.64 ± 9.79	5.56 ± 5.04	5.36 ± 5.74
(2) MB-AttUnet	✗	✓	✗	64.03 ± 2.47	50.44 ± 2.67	47.14 ± 2.71	17.52 ± 12.30	8.61 ± 6.67	10.11 ± 7.53
(3) MB-TrAttUnet	✓	✓	✗	68.43 ± 0.68	57.46 ± 1.21	52.01 ± 0.79	43.11 ± 3.87	26.09 ± 4.03	27.64 ± 3.43
(4) EMB-TrAttUnet	✓	✓	✓	69.22 ± 0.55	59.16 ± 1.33	52.39 ± 0.24	51.11 ± 0.59	30.02 ± 1.55	34.33 ± 0.53

segmentation classes. Experiments (4) proves the effectiveness of the proposed Boundary Aware loss function in segmenting the two classes of Covid-19 infections. Compared with the results of Experiment (3), the GGO results are improved by 1%, 3.5%, and 1.19% for F-S, D-S, and IoU, respectively. Similarly, the Consolidation segmentation results are considerably improved by 2.51%, 5.46%, and 2.51% for F-S, D-S, and IoU, respectively.

Table 8 summarizes the ablation study results of the second evaluation scenario. Similar to what were noticed in the first evaluation scenario (Table 7), the Transformer Path provides richer features in the encoding phase, leading to better segmentation performance, especially for the minor class (Consolidation). On the other hand, it is noticed from Experiment (2) that adding the MBSC connection blocks to the baseline architecture (Att-Unet) significantly improves the Consolidation segmentation results, while the GGO results slightly decrease. This shows that MBSC connection blocks can help hard class segmentation when less labelled data is available. However, this can affect the performance of the other classes in the case of only CNN encoder is used. Similar to the first evaluation scenario, Experiment (4) prove the effectiveness of the proposed loss function. By comparing with the results of Experiment (3), the GGO results are improved by 0.8%, 1.7%, and 0.38% for F-S, D-S, and IoU, respectively. Significantly, the Consolidation segmentation results are considerably improved by 8%, 3.93%, and 6.69% for F-S, D-S, and IoU, respectively. These results shows the efficiency of the proposed MBA-CE loss function in few-shot learning scenario, with imbalanced classes.

5 Discussion

5.1 Segmentation visualization analysis

Figures 7 and 8 depict visual comparison examples from the first and second evaluation scenarios which correspond to the first and second splitting configurations, respectively. The approaches that demonstrated the best segmentation performance in comparison to our approach are Att-Unet, SCOATNET, UCTransNet, and MISSFormer. All six examples in Fig. 7 demonstrate that our approach achieves superior segmentation performance for both GGO and Consolidation when compared to the other methods.

In more detail, the first example in Fig. 7 illustrates cases where the infection has spread in both lungs, manifesting as multiple small regions of GGO and Consolidation. In contrast to the comparison methods, our approach effectively segments most of the infected regions, including the smaller ones. This highlights the efficiency of combining CNN and Transformer encoders, where the Transformer path treats the input image as a collection of 2D patches, enabling the extraction of finer infection details, including small regions. Additionally, our proposed loss function places more emphasis on smaller regions.

The second example depicts a case where the infection has spread throughout most of the lung field as GGO, with Consolidation appearing in the lower region of the right lung. The segmentation masks generated by the comparison methods reveal their difficulty in accurately matching the ground truth masks for the GGO regions, particularly near the boundaries. In contrast, our approach demonstrates a strong ability to match the infection boundaries, thanks to the use of our proposed boundary-aware loss function. While some of the comparison approaches have identified Consolidation in the lower regions, many of them misalign with the ground truth mask. Notably, Att-Unet, SCOATNET, and UCTransNet show Consolidation in both lungs, which is inconsistent with the ground

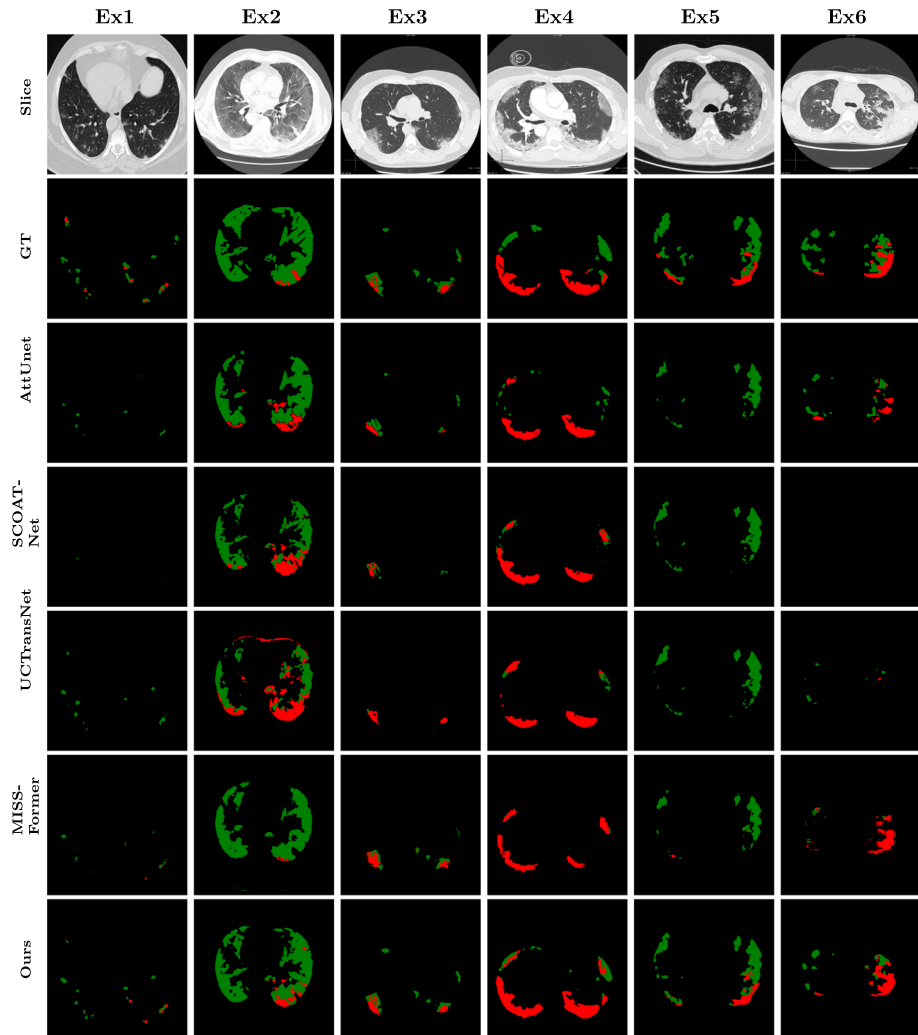


Fig. 7 Visual Comparison of COVID-19 Segmentation Results from the First Evaluation Scenario. The seven rows represents the input slice, ground truth, and the segmentation results of Att-Unet, SCOATNET, UCTransNet, MISSFormer and our proposed EMB-TrAttUnet, respectively. GGO is presented by the green color and Consolidation by the red color

truth. In contrast, MISSFormer and our approach accurately locate the region of Consolidation infection (the lower region of the right lung).

The third example presents a case where the infection appears in both lungs and is a combination of medium and small spots, with small Consolidation areas in the lower lung regions. The segmented masks demonstrate that most of the comparison methods struggle to segment the small regions and differentiate between the infection types. In contrast, our approach efficiently segments both medium and small infections and accurately distinguishes between the two classes (GGO and Consolidation), aligning with the ground

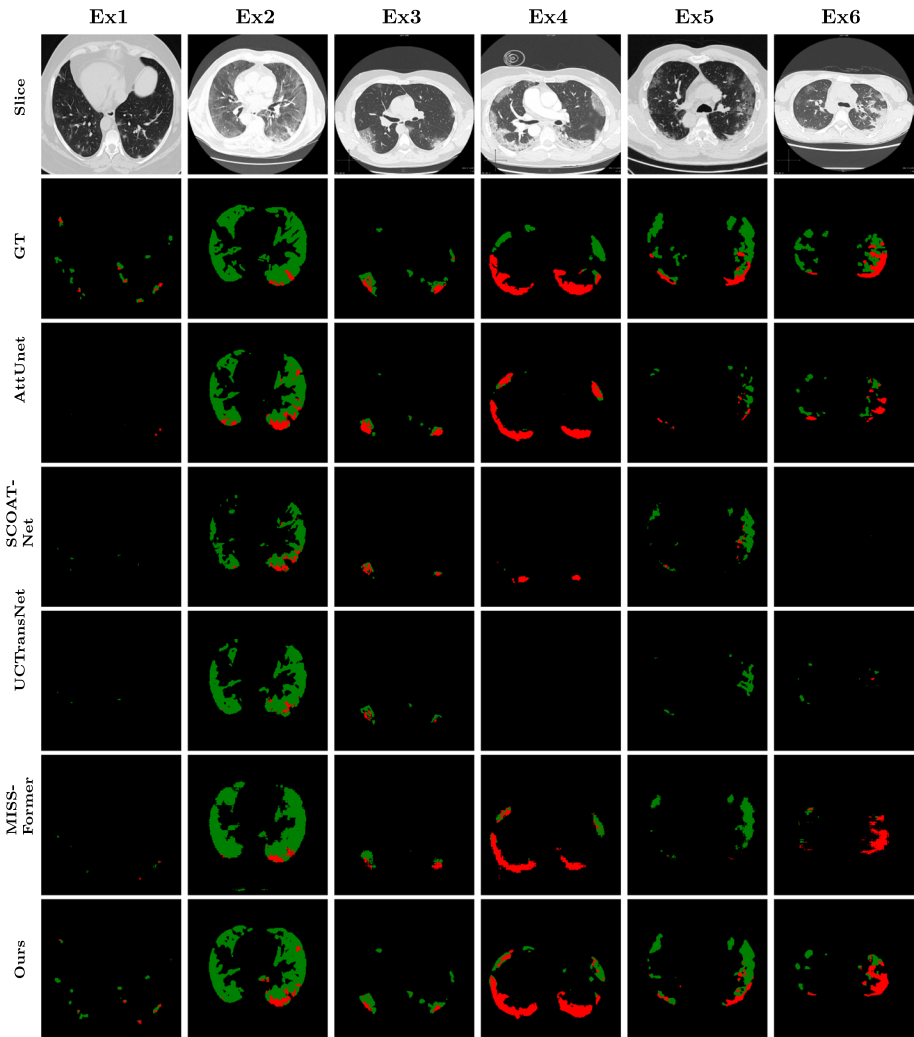


Fig. 8 Visual Comparison of COVID-19 Segmentation Results from the Second Evaluation Scenario. The seven rows represents the input slice, ground truth, and the segmentation results of Att-Unet, SCOATNET, UCTransNet, and MISSFormer and our proposed EMB-TrAttUnet, respectively. GGO is presented by the green color and Consolidation by the red color

truth. This underscores that our proposed approach has a better capability to segment small regions, as well as medium and large ones, compared to the state-of-the-art approaches.

The last three examples depict cases where the infection has a peripheral distribution with a mixture of GGO and Consolidation infection types, and Consolidation appears in the lower half of the lung regions. In the fourth example, the predicted masks of the segmentation methods struggle to distinguish between the infection classes and their sizes. For instance, the MISSFormer approach segments all infections as Consolidation, while other approaches consider most of the infected areas as Consolidation. Although our approach

identifies most of the infected regions, it shows limited distinguishability between the two infection types in this case.

In the fifth example, the comparison approaches segment all infections as GGO. In contrast, our approach demonstrates a high capability to segment both classes accurately. Similarly, our approach performs well in the last example. However, some of the comparison approaches exhibit lower performance, such as UCTransNet and SCOATNet, as they struggle to segment the infection due to the low contrast between the infection and lung borders.

Figure 8 displays the segmented masks for the second evaluation scenario, where the training data was reduced to half compared to the first splitting configuration. From the visualized six examples, the following observations can be made: In the first example, almost all comparison approaches failed to segment any infection. However, our approach continues to demonstrate good performance in scenarios with more challenging data available. In the second example, all approaches are still capable of segmenting the widely spread infection and also successfully segment the Consolidation type in the lower part of the right lung. However, the accuracy varies from one approach to another. MISSFormer matches the Consolidation prediction, while our approach accurately captures the details of GGO infection in terms of shape and boundaries. This once again highlights the efficiency of the proposed methods, particularly the proposed loss function in this case. In the third example, our approach continues to perform well in segmenting small and medium infection regions with the correct infection type. The last three examples reveal that the comparison approaches miss a significant portion of the infection, especially UCTransNet and SCOATNet in examples 4 and 6. Additionally, it's worth noting that, as in the previous evaluation scenario, some approaches entirely predict only one infection class. However, our approach still demonstrates a high segmentation capability to accurately segment the infection and efficiently identify its type in scenarios with more data availability limitation regime.

5.2 Early stage infection localization and segmentation

One of the most important steps to stop the spreading of Covid-19 disease and save the infected person live is to detect the infection in the early stages. Where the golden standard RT-PCR test has a considerable false negative rate in this case (Jin et al. 2020). We propose to test the efficiency of our approach in detecting the early infection regions. To do so, we manually selected four CT-scans from Dataset_1 as the testing data, which shows infection in early stages. The remaining CT-scans of Dataset_1 and Dataset_2 are used as the training data. In order to compare with the state-of-the-art, the best performed approaches in the previous evaluated splitting configurations, which are Att-Unet (Oktay et al. 2018), and SCOATNet (Zhao et al. 2021), MISSFormer (Huang et al. 2022), and UCTransNet (Wang et al. 2022). Table 9 summarizes the obtained results, which show the superiority of our approach compared with the comparing ones for both infection types segmentation.

Figure 9 depicts three slices from two CT scans (each scan is represented by 3 slices) and the predicted masks of the four comparison approaches and our approach. The first three examples show that our approach correctly localizes the position and the number of tiny infected regions and classifies them as GGO. However, most of the comparison approaches failed to segment these infected regions. This once again demonstrates the efficiency of the proposed approach in dealing with even the tiniest infected regions in the early stage, allowing for the identification of infections in their very early stages. The second case (examples 4 to 6) depicts a situation where the infection is beginning to develop, and consolidation is starting to appear. In this case as well, it is evident that

Table 9 The experimental results to Evaluate the Performance of the proposed EMB-TrAttUnet, Att-Unet Oktay et al. (2018), SCOATNet Zhao et al. (2021), MISSFormer Huang et al. (2022), and UCTransNet Wang et al. (2022) for Early Stage Infection Localization and Segmentation

Architecture	GGO				Consolidation			
	F1-S	D-S	IoU	HD95	F1-S	D-S	IoU	HD95
Att-Unet	<i>50.67 ± 3.91</i>	<i>42.23 ± 4.20</i>	<i>34.02 ± 3.40</i>	<i>28.20 ± 3.69</i>	27.23 ± 18.10	6.09 ± 4.23	17.04 ± 12.22	5.02 ± 2.67
SCOATNET	49.14 ± 4.86	40.78 ± 6.08	32.71 ± 4.39	26.46 ± 2.39	<i>39.99 ± 7.39</i>	<i>7.87 ± 2.35</i>	<i>25.26 ± 5.80</i>	<i>7.17 ± 1.28</i>
UCTransNet	43.23 ± 5.12	29.79 ± 6.75	27.57 ± 5.22	23.95 ± 4.32	39.48 ± 5.12	6.73 ± 4.22	24.59 ± 6.21	4.79 ± 2.11
MISSFormer	39.48 ± 7.15	23.24 ± 5.29	26.30 ± 6.63	17.85 ± 4.19	26.97 ± 7.66	5.82 ± 4.03	16.77 ± 6.60	4.99 ± 2.64
EMB-TrAttUnet	62.45 ± 1.57	54.33 ± 3.55	45.42 ± 1.64	22.54 ± 2.91	42.84 ± 6.09	8.06 ± 1.76	27.45 ± 4.98	2.62 ± 0.94

The best performance and the second best performance are marked by textbf and textit, respectively
 The italic shows the best comparison approach

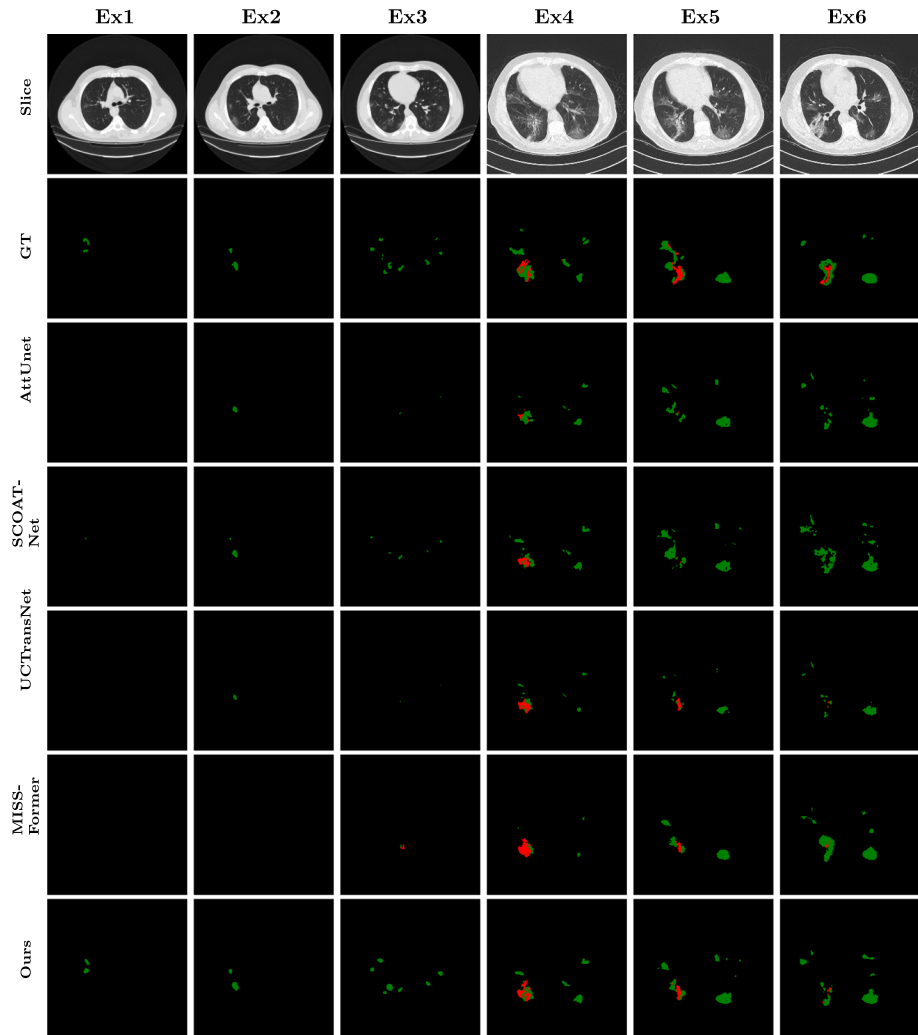


Fig. 9 Visual Comparison of COVID-19 Segmentation Results for Early Infection Segmentation Scenario. The seven rows represents the input slice, ground truth, and the segmentation results of The seven rows represents the input slice, ground truth, and the segmentation results of Att-Unet, SCOATNET, UCTransNet, and MISSFormer and our proposed EMB-TrAttUnet, respectively. GGO is presented by the green color and Consolidation by the red color

our approach has a high ability to detect small consolidation regions attached to the GGO infection, as well as small GGO infections spreading in both lungs. These two cases highlight our approach’s capability to localize and segment infections at different stages, including the early stage, which is crucial for detecting infected individuals, preventing the spread of the Covid-19 pandemic, and monitoring the patient’s condition.

5.3 Model size and inference time comparison

In this section, we investigate the number of parameters and inference times of our approach and the compared ones. Table 10 summarizes these comparisons. It is worth noting that our approach features a higher number of parameters compared to most of the compared architectures. This is due to the proposed Hybrid Encoder and the MBSC block, both of which have proven their efficiency in handling the complex task with very limited training data. Despite the larger parameter count, our approach still delivers competitive inference times. In fact, the inference time for a batch size of 50 slices is less than half a second, which is suitable for real-time scenarios.

6 Conclusion

In this paper, we proposed a Hybrid CNN-Transformer (MB-TrAttUnet) architecture and a Multi-classes Boundary Aware Cross-Entropy (MBA-CE) Loss function for Multi-classes Covid-19 infection segmentation. The Encoder of the proposed MB-TrAttUnet architecture fuses CNN and Transformer blocks to extract richer features, including local, global, and long-range dependencies features. On the other hand, the Decoder of the MB-TrAttUnet is similar to the Att-Unet decoder, which consists of AG and deconvolutional blocks. In addition, we proposed a new skip connection (MBSC) to integrate the encoder features into the decoder layers, where the MBSC blocks extract and combine higher level features related to the finer features of different infection types. On the other hand, MBA-CE Loss function is designed to enhance boundaries segmentation, the separability between classes and minority classes decision.

The performance of the proposed approach is evaluated and compared with three CNN-Based Segmentation architectures (Unet, Nested-Unet and Attention Unet), three proposed segmentation approaches for Covid-19 infection segmentation (CopleNet, AnamNet, and SCOATNet) and four recent Transformer based segmentation architectures (SwinUnet, MTUnet, MISSFormer, and UCTransNet) in two evaluation scenarios using two splitting configurations. From which, our approach outperformed the comparison methods for both Ground-glass opacity (GGO) and Consolidation segmentation. On the other hand, our approach showed consistent performance in the more

Table 10 Number of parameters of different architectures and Testing Time for a batch size of 50 slices

Architecture	Number of Parameters	Inference Time
Unet	7.85 M	89ms
AttUnet	7.98 M	102ms
Unet++	26.51 M	244ms
CopleNet	15.05 M	95ms
AnamNet	25.63 M	117ms
SCOATNET	40.21 M	407ms
SwinUnet	41.38 M	160ms
MTUnet	79.07 M	629ms
UCTransNet	66.43 M	423ms
MISSFormer	42.46 M	223ms
EMB-TrAttUnet	80.13 M	475ms

critical “few-shot” learning scenario (training data, in terms of scans, were reduced by half). In contrast, the performance of the comparison methods decreased in this scenario. Moreover, our approach showed a good ability to deal with imbalanced data and segmenting the infection in early stages. These advantages demonstrate the effectiveness and efficiency of the proposed EMB-TrAttUnet approach in a pandemic scenario where time is critical to save patient lives. As part of our future work, we intend to leverage recent advancements in vision transformer architectures to reduce the number of parameters in our model. Additionally, we plan to extend the application of our approach to various other medical imaging segmentation tasks.

Acknowledgements This work is partially supported by Grant PID2021-126701OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amyar A, Modzelewski R, Li H, Ruan S (2020) Multi-task deep learning based CT imaging analysis for covid-19 pneumonia: classification and segmentation. *Comput Biol Med* 126:104037
- Bharati S, Podder P, Mondal MRH (2020) Hybrid deep learning for detecting lung diseases from x-ray images. *Inf Med Unlocked* 20:100391
- Bougourzi F, Distante C, Ouafi A, Dornaika F, Hadid A, Taleb-Ahmed A (2021) Per-COVID-19: a benchmark dataset for COVID-19 percentage estimation from CT-scans. *J Imaging* 7:189
- Bougourzi F, Contino R, Distante C, Taleb-Ahmed A (2021) Recognition of COVID-19 from CT scans using two-stage deep-learning-based approach: CNR-IEMN. *Sensors* 21:5878
- Bougourzi F, Distante C, Dornaika F, Taleb-Ahmed A, Hadid A (2022) Ilc-unet++ for covid-19 infection segmentation. In: International conference on image analysis and processing, pp 461–472. Springer, Berlin
- Bougourzi F, Dornaika F, Taleb-Ahmed A (2022) Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowl-Based Syst* 242:108246
- Bougourzi F, Distante C, Dornaika F, Taleb-Ahmed A (2023) Cnr-iemn-cd and cnr-iemn-csd approaches for covid-19 detection and covid-19 severity detection from 3d ct-scans. In: Computer vision—ECCV 2022 workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, pp 593–604. Springer, Berlin
- Bougourzi F, Distante C, Dornaika F, Taleb-Ahmed A (2023) Pdatt-unet: pyramid dual-decoder attention unet for covid-19 infection segmentation from ct-scans. *Med Image Anal* 886:102797
- Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M (2021) Swin-unet: unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*
- Chaudhary S, Yang W, Qiang Y (2022) Swin transformer for covid-19 infection percentage estimation from ct-scans. In: International conference on image analysis and processing, pp 520–528. Springer, Berlin
- Cong R, Yang H, Jiang Q, Gao W, Li H, Wang C, Zhao Y, Kwong S (2022) Bcs-net: boundary, context, and semantic for automatic covid-19 lung infection segmentation from CT images. *IEEE Trans Instrum Meas* 71:1–11
- Cong R, Zhang Y, Yang N, Li H, Zhang X, Li R, Chen Z, Zhao Y, Kwong S (2022) Boundary guided semantic learning for real-time covid-19 lung infection segmentation system. *IEEE Trans Consum Electron* 68(4):376–386
- Ding W, Abdel-Basset M, Hawash H, Elkomy OM (2021) Mt-ncov-net: a multitask deep-learning framework for efficient diagnosis of covid-19 using tomography scans. *IEEE Trans Cybern* 99:1

- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
- Duong LT, Le NH, Tran TB, Ngo VM, Nguyen PT (2021) Detection of tuberculosis from chest x-ray images: boosting the performance with vision transformer and transfer learning. *Expert Syst Appl* 184:115519
- Duong LT, Nguyen PT, Iovino L, Flammini M (2023) Automatic detection of covid-19 from chest x-ray and lung computed tomography images using deep neural networks and transfer learning. *Appl Soft Comput* 132:109851
- Fung DL, Liu Q, Zammit J, Leung CK-S, Hu P (2021) Self-supervised deep learning model for covid-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and covid-19. *J Transl Med* 19:1–18
- Ghosh S, Das N, Das I, Maulik U (2019) Understanding deep learning techniques for image segmentation. *ACM Comput Surv* 52(4):1–35
- Goceri E (2023) Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev* 20:1–45
- Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H. R, Xu D (2022) Unetr: transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 574–584
- Hefeda MM (2020) CT chest findings in patients infected with covid-19: review of literature. *Egypt J Radiol Nucl Med* 51(1):1–15
- Huang X, Deng Z, Li D, Yuan X, Fu Y (2022) Missformer: an effective transformer for 2d medical image segmentation. *IEEE Trans Med Imaging*
- Ilhan A, Alpan K, Sekeroglu B, Abiyev R (2023) Covid-19 lung CT image segmentation using localization and enhancement methods with u-net. *Proc Comput Sci* 218:1660–1667
- Jasmine Pemeena Priyadarsini M, Rajini G, Hariharan K, Utkarsh Raj K, Bhargav Ram K, Indragandhi V, Subramaniaswamy V, Pandya S et al (2023) Lung diseases detection using various deep learning algorithms. *J Healthcare Eng* 2023:1
- Jin Y-H, Cai L, Cheng Z-S (2020) A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). *Mil Med Res* 7:4
- Jin G, Liu C, Chen X (2022) An efficient deep neural network framework for covid-19 lung infection segmentation. *Inf Sci* 612:745–758
- Khan S, Naseer M, Hayat M, Zamir S. W, Khan F. S, Shah M (2021) Transformers in vision: a survey, *ACM Computing Surveys (CSUR)*. ACM, New York
- Li Z, Zhao S, Chen Y, Luo F, Kang Z, Cai S, Zhao W, Liu J, Zhao D, Li Y (2021) A deep-learning-based framework for severity assessment of covid-19 with CT images. *Expert Syst Appl* 185:115616
- Li Y, Wang S, Wang J, Zeng G, Liu W, Zhang Q, Jin Q, Wang Y (2021) Gt u-net: a u-net like group transformer network for tooth root segmentation. In: *International workshop on machine learning in medical imaging*, pp 386–395. Springer, Berlin
- Li J, Chen J, Tang Y, Wang C, Landman B. A, Zhou S. K (2023) Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Med Image Anal* p 102762
- Liang S, Zhang W, Gu Y (2021) A hybrid and fast deep learning framework for covid-19 detection via 3d chest CT images. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 508–512
- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88
- Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 10012–10022
- Liu X, Wang K, Wang K, Chen T, Zhang K, Wang G (2020) Kiseq: a three-stage segmentation framework for multi-level acceleration of chest CT scans from covid-19 patients. In: *International conference on medical image computing and computer-assisted intervention*, pp 25–34. Springer, Berlin
- Maji D, Sigedar P, Singh M (2022) Attention res-unet with guided decoder for semantic segmentation of brain tumors. *Biomed Signal Process Control* 71:103077
- Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D (2021) Image segmentation using deep learning: a survey. *IEEE transactions on pattern analysis and machine intelligence*
- Mu N, Wang H, Zhang Y, Jiang J, Tang J (2021) Progressive global perception and local polishing network for lung infection segmentation of covid-19 CT images. *Pattern Recogn* 120:108168

- Müller D, Soto-Rey I, Kramer F (2021) Robust chest CT image segmentation of covid-19 lung infection based on limited data. *Inf Med Unlocked* 25:100681
- Oktay O, Schlemper J, Folgoc L L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N. Y, Kainz B, et al.(2018) Attention u-net: learning where to look for the pancreas. arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999)
- Paladini E, Vantaggiato E, Bougourzi F, Distanto C, Hadid A, Taleb-Ahmed A (2021) Two ensemble-cnn approaches for colorectal cancer tissue type classification. *J Imaging* 7(3):51
- Paluru N, Dayal A, Jenssen HB, Sakinis T, Cenkeramaddi LR, Prakash J, Yalavarthy PK (2021) Anamnet: anamorphic depth embedding-based lightweight cnn for segmentation of anomalies in covid-19 chest CT images. *IEEE Trans Neural Netw Learn Syst* 32(3):932–946
- Parasher A (2021) COVID-19: current understanding of its pathophysiology. *Clin Present Treat Postgrad Med J* 97:312–320
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 32
- Pezzano G, Díaz O, Ripoll VR, Radeva P (2021) Cole-cnn+: context learning-convolutional neural network for covid-19-ground-glass-opacities detection and segmentation. *Comput Biol Med* 136:104689
- Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M (2019) Basnet: boundary-aware salient object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 7479–7489
- RADIOLOGISTS (2019) COVID-19 CT-scans segmentation datasets, <http://medicalsegmentation.com/covid19/>. Accessed 18 Aug 2021
- Rehman A, Khan A, Fatima G, Naz S, Razzak I (2023) Review on chest pathologies detection systems using deep learning techniques. *Artif Intell Rev* 1–47
- Ronneberger O, Fischer P, Brox T(2015) U-Net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention - MICCAI 2015, Lecture Notes in Computer Science*, pp 234–241. Springer, Cham
- Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, Huijben I, Chennakeshava N, Mento F, Sentelli A et al (2020) Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Trans Med Imaging* 39(8):2676–2687
- Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A et al (2020) Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients. *AJR Am J Roentgenol* 215(1):87–93
- Saood A, Hatem I (2021) Covid-19 lung CT image segmentation using deep learning methods: U-net versus segnet. *BMC Med Imaging* 21(1):1–10
- Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, Fan Y, Zheng C (2020) Radiological findings from 81 patients with covid-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis* 20(4):425–434
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
- Soomro TA, Zheng L, Afifi AJ, Ali A, Yin M, Gao J (2022) Artificial intelligence (ai) for medical imaging to combat coronavirus disease (covid-19): a detailed review with direction for future research. *Artif Intell Rev* 1–31
- Tomar N. K, Jha D, Riegler M. A, Johansen H. D, Johansen D, Rittscher J, Halvorsen P, Ali S (2022) Fanet: a feedback attention network for improved biomedical image segmentation. *IEEE Trans Neural Netw Learn Syst* 34:1
- Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention, in *International conference on machine learning*, pp 10347–10357, PMLR
- Valanarasu JMJ, Sindagi VA, Hacihaliloglu I, Patel VM (2021) Kiu-net: overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Trans Med Imaging* 41(4):965–976
- Van der Velden BH, Kuijff HJ, Gilhuijs KG, Viergever MA (2022) Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Med Image Anal* 79:102470
- Vantaggiato E, Paladini E, Bougourzi F, Distanto C, Hadid A, Taleb-Ahmed A (2021) Covid-19 recognition using ensemble-cnns in two new chest x-ray databases. *Sensors* 21(5):1742
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30:5998
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, kaiser, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30

- WHO (2023) World Health Organization. <https://covid19.who.int/>. Accessed 08 Nov 2023
- Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, Meng T, Li K, Huang N, Zhang S (2020) A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from CT images. *IEEE Trans Med Imaging* 39(8):2653–2663
- Wang X, Yuan Y, Guo D, Huang X, Cui Y, Xia M, Wang Z, Bai C, Chen S (2022) Ssa-net: spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning. *Med Image Anal* 79:102459
- Wang H, Cao P, Wang J, Zaiane OR (2022) Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc AAAI Conf Artif Intell* 36:2441–2449
- Wang H, Xie S, Lin L, Iwamoto Y, Han X.-H, Chen Y.-W, Tong R (2022) Mixed transformer u-net for medical image segmentation. In: *ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp 2390–2394, IEEE
- Wu Y-H, Gao S-H, Mei J, Xu J, Fan D-P, Zhang R-G, Cheng M-M (2021) Jcs: an explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Trans Image Process* 30:3113–3126
- Xie L, Huang J, Yu J, Zeng Q, Hu Q, Chen Z, Xie G, Feng Y (2023) Cntseg: a multimodal deep-learning-based network for cranial nerves tract segmentation. *Med Image Anal* 86:102766
- Yao Q, Xiao L, Liu P, Zhou SK (2021) Label-free segmentation of covid-19 lesions in lung ct. *IEEE Trans Med Imaging* 40(10):2808–2819
- Yu Q, Qi L, Gao Y, Wang W, Shi Y (2022) Crosslink-net: double-branch encoder network via fusing vertical and horizontal convolutions for medical image segmentation. *IEEE Trans Image Process* 31:5893–5908
- Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, Liang W, Wang C, Wang K et al (2020) Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell* 181(6):1423–1433
- Zhao S, Li Z, Chen Y, Zhao W, Xie X, Liu J, Zhao D, Li Y (2021) Scoat-net: a novel network for segmenting covid-19 lung opacification from CT images. *Pattern Recogn* 119:108109
- Zhou Z, Rahman Siddiquee M. M, Tajbakhsh N, Liang J (2018) UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov D, Taylor Z, Carneiro G, eds, *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Lecture Notes in Computer Science, pp 3–11. Springer, Cham

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Fares Bougourzi¹ · Fadi Dornaika^{2,3} · Amir Nakib¹ · Abdelmalik Taleb-Ahmed⁴

✉ Fares Bougourzi
faresbougourzi@gmail.com; fares.bougourzi@u-pec.fr

Fadi Dornaika
fadi.dornaika@ehu.eus

Amir Nakib
nakib@u-pec.fr

Abdelmalik Taleb-Ahmed
Abdelmalik.Taleb-Ahmed@uphf.fr

¹ University Paris-Est Creteil, Laboratoire LISSI, 122 Rue Paul Armangot, 94400 Vitry sur Seine, Paris, France

² University of the Basque Country UPV/EHU, San Sebastian 20018, Spain

³ IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

⁴ Université Polytechnique Hauts-de-France, Université de Lille, CNRS, 59313 Valenciennes, Hauts-de-France, France