Attention-based Fusion for Stroke Lesion Segmentation on Computed Tomography Perfusion Data

Attention-based Fusion for Stroke Lesion Segmentation on Computed Tomography Perfusion Data

CHINTHA SRI POTHU RAJU*, Speech and Image Processing Lab, Department of ECE, National Institute of Technology Silchar, India
RABUL HUSSAIN LASKAR, Speech and Image Processing Lab, Department of ECE, National Institute of Technology Silchar, India
ZULFIQAR ALI, School of Computer Science and Electronic Engineering (CSEE), University of Essex, Colchester, United Kingdom
GHULAM MUHAMMAD*, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

In recent times, stroke has emerged as a significant threat to humans, transforming affected brain tissue into core and penumbra regions. As the penumbra becomes irreversible over time, early core region segmentation is crucial. Automatic segmentation systems offer an efficient alternative to manual segmentation and aid radiologists in stroke lesion segmentation using Computed Tomography and Computed Tomography Perfusion (CTP) maps that comprise four parameter maps. This automatic segmentation is increasingly used in interactive, multimedia-based systems for diagnostic tools and AI-driven health applications. Top-performing models that follow the patch processing approach suffer from high inference times. To incorporate effective feature extraction at image-level inferences, which reduces the inference time, we present a hybrid fusion technique that combines early and bottleneck fusion, leveraging two separate encoders for effective feature extraction. Moreover, fusing the information from various fusion methods arbitrarily may not yield optimal results. Consequently, we have introduced two attention modules, i.e., cross-modal attention and cross-fusion attention modules, designed for the effective integration of features derived from diverse modalities and multiple fusion strategies, respectively. The findings highlight a considerable reduction in computational time alongside achieving a comparable dice score. Additionally, the incorporation of hybrid fusion and attention modules in the baseline notably increased the dice score from 0.482 to 0.521 in the validation dataset and achieved 0.48 in the test dataset of ISLES 2018. It also demonstrates competitive performance compared to existing models while maintaining efficient prediction times.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Image segmentation**.

Additional Key Words and Phrases: Hybrid Fusion, Stroke lesion Segmentation, CT Perfusion, Attention Modules, Multimodal data

Authors' Contact Information: Chintha Sri Pothu Raju, chintha_rs@ece.nits.ac.in, Speech and Image Processing Lab, Department of ECE, National Institute of Technology Silchar, Silchar, Assam, India; Rabul Hussain Laskar, Speech and Image Processing Lab, Department of ECE, National Institute of Technology Silchar, Silchar, India, rhlaskar@ece.nits.ac.in; Zulfiqar Ali, School of Computer Science and Electronic Engineering (CSEE), University of Essex, Colchester, Colchester, United Kingdom, z.ali@essex.ac.uk; Ghulam Muhammad, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Riyadh, Saudi Arabia, ghulam@ksu.edu.sa.

ACM 1557-735X/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Chintha Sri Pothu Raju, Rabul Hussain Laskar, Zulfiqar Ali, and Ghulam Muhammad. 2018. Attention-based Fusion for Stroke Lesion Segmentation on Computed Tomography Perfusion Data. *J. ACM* 37, 4, Article 111 (August 2018), 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Over the last two decades, stroke has emerged as a significant threat to human life. According to the World Stroke Organization (WSO), stroke is the second leading cause of death and the third leading cause of disability [16]. Ischemic stroke holds a portion of 87% of total cases. As the time lasts longer after stroke incidence, the penumbra will be converted into a core, which is irreversible. The necessity for the timely detection and segmentation of stroke lesions is to prevent the brain from being irreversibly damaged. The segmentation of stroke lesions requires brain imaging multimedia such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). CT Perfusion (CTP) imaging is a variant of CT in which multiple CT scans are acquired after fixed time intervals. It is called Perfusion Weighted Imaging (PWI). This CTP acquisition process is very similar to the acquisition of video frames in multimedia systems. The perfusion parameter maps are extracted from the raw PWI data. It does contain the penumbra information, which may be reversible if treated in time. Despite the challenges, various articles demonstrate the effectiveness of segmenting stroke lesions in CTP data [27, 39]. Manually delineating the stroke lesions from the brain image modalities is a gold standard practice. However, it is a time-consuming, tedious process facing issues like inter-observer variability. This motivated researchers to find alternatives to manually segmenting lesions on the brain imaging data.

The well-known alternatives are threshold-based software, and automatic algorithms developed using machine learning and deep learning concepts. Stroke lesion segmentation on CTP data is commonly performed using software tools such as RAPID [24] and Sphere. The problem with threshold-based softwares is that the lesion is segmented based on the threshold value for the parameter maps, utilizing fixed values for segmenting lesions [33]. The hemodynamics changes widely across the patients depending upon several factors. So, this software may not be the best alternative for segmenting the lesions. Deep learning has emerged as a solution to learn the patterns from the data to segment the lesions. The studies also suggest that the algorithms developed using deep learning models surpassed threshold-based techniques for segmenting lesions [18]. Despite the advancements in deep learning and medical technology, there is a lack of faster and more accurate algorithms for segmenting stroke lesions on the CTP data. This study aims to formulate and develop a better and faster algorithm for segmenting the lesions, improving precise segmentation, and achieving faster inference times. Accurate and efficient segmentation helps to reduce the overload on the radiologists and reduces the time required to delineate the lesion, eventually saving the brain from converting into the core.

Several algorithms have been designed in the existing literature to segment stroke lesions within CTP data [27, 39, 42]. Patch based processing techniques have also been employed for lesion segmentation on CTP data [11, 23]. However, these models often suffer from prolonged inference times. Furthermore, applying models designed for patch-based processing to whole-image inferences significantly degrades performance. An XGBoost-based machine learning approach has been introduced to segment core and penumbra regions directly on CTP data [45]. Another recent approach incorporated clinical data alongside CTP imaging to enhance stroke lesion segmentation results [2], although the study relied on proprietary datasets. Recent studies have focused on leveraging 4D CTP data directly to segment stroke lesions. To achieve superior performance in image-level inferences, models must integrate an efficient feature extraction mechanism.

The algorithms used for the stroke lesion segmentation can be broadly categorized into four groups based on their fusing multi-modal data approach. These categories are early, late, bottleneck, and multi-level fusion. Each strategy leverages the unique characteristics of stroke lesions captured within different modalities. However, arbitrarily combining the multimodal information may not lead to optimal results. While most research articles have proposed fusion strategies, each approach has its advantages and disadvantages. Additionally, the information obtained through these fusion methods may exhibit variations in certain aspects, such as integrated information and individual information of all modalities. Therefore, to ensure the precise integration of information produced by various fusion strategies, employing attention mechanisms, is imperative. Thus, integrating a cross-attention mechanism is vital for carefully fusing the features. The algorithms that achieve better performance were developed based on patch processing, which eventually takes much preprocessing and inference time. Image level prediction takes less time but requires information-rich features. These information-rich features may be provided by the hybrid fusion approach since it involves the integrated features and individual features from the multimodal data.

Multimodal or multimedia data carries richer information than single-modal data [7]. However, the method of fusing these data is a critical factor that decides the performance [9] since the multimodal data contains different information in each modality. Predominantly, algorithms have primarily concentrated on singular fusion strategies. However, investigating various fusion approaches presents potential solutions for addressing segmentation challenges, given the unique advantages inherent in each method. Consequently, a more sophisticated algorithm capable of precise multimodal data integration is to overcome the limitations associated with single fusion strategies and leverage the potential benefits of combining multiple approaches. To meet this requirement, we introduce a hybrid fusion technique for stroke lesion segmentation.

This paper introduces a hybrid fusion approach for stroke lesion segmentation in CTP data, equipped with attention modules for careful feature integration. The characteristics of these features depend on the input data provided to the model. In the context of multimodal data, feature characteristics vary based on whether they are derived from a single modality or multiple modalities. Single-modal data yields features specific to that modality, while features from multimodal data encapsulate the characteristics from all modalities. Combining these diverse feature types effectively may enhance the model's performance, yielding information-rich features. Moreover, the ensemble models that use the aggregation of features will surpass the single algorithm [46]. This motivated us to take the concept of hybrid fusion into consideration. First, the implementation of a hybrid fusion technique aimed at combining integrated features from all modalities with individual features from each modality. This integration of features with distinct characteristics poses a significant challenge, prompting the development of two distinct attention modules. One module is designed to fuse individual features, while the second is tasked with fusing features across different fusion strategies. To accomplish the fusion of individual features, a hierarchical attention mechanism is employed, initially fusing features from blood parameters (CBV: Cerebral Blood Volume and CBF: Cerebral Blood Flow) and time parameters (MTT: Mean Transit Time and Tmax: Time to Maximum) independently. Subsequently, these two sets of features are combined within and across the groups using non local block concept. The model generates integrated features along with the individual features in a different path. To effectively fuse these two sets of features generated in different ways, a cross-fusion attention module is introduced. The contributions of the paper are as follows:

(1) Development of a deep learning model for the precise delineation of stroke lesions with CT and CTP data with less computational time while capturing effective features from multimodal input data.

(2) This study introduces multimodal fusion techniques, including hybrid fusion, to effectively combine features derived from different modalities in stroke lesion segmentation. This contribution enables a more comprehensive analysis of the CTP data, taking into account the unique characteristics of individual modalities and all modalities together.
(3) The incorporation of attention mechanisms, including cross-modal and cross-fusion attention, enhances the fusion of features from different modalities and fusion strategies, respectively. This attention-based approach contributes to the model's capability to handle the fusion of diverse features effectively.

The rest of the paper is structured as follows: Section 2 summarizes the current state-of-the-art methods from the literature. Section 3 explains the proposed method incorporating hybrid fusion and cross attention modules . Section 4 presents the quantitative and qualitative results. Section 5 discusses the findings. Section 6 concludes the paper.

2 Related Works

The related works that employ machine learning or deep learning techniques for segmenting the acute infarct core are discussed. It encompasses methodologies focused on predicting the volume of the acute infarct core on the CTP data. The temporal acquisition process of CTP is similar to capturing video frames in multimedia systems, where sequences of images are sampled over time to provide meaningful visual information. However, despite these similarities in temporal sampling, the processing of PWI is different from traditional video segmentation methods. In video analysis, the focus is on segmenting objects or scenes across frames [10], while in PWI, the goal is to derive perfusion maps that quantify blood flow characteristics. In recent years, many studies on disease detection and classification utilized deep learning techniques [36]. Recent works utilized only the 4D CTP data without the perfusion maps to segment the lesions directly from the CTP [12]. Alternatively, these perfusion parameter maps are synthetically generated from CTP dynamic images using the LSTM-based neural network [37]. Recent works used patch processing techniques to segment the lesions on the CTP data [11, 23]. The models that involve patch processing suffer from high inference times. At the same time, these models applied to the image level inferences severely impact performance. A machine learning approach based on XGBoost is proposed for the segmentation of core and penumbra regions on the CTP [45]. Recent work included the clinical data along with the CTP data to improve the stroke lesion segmentation outcomes [2]. However, they have utilized the proprietary datasets. For better performance at image-level inferences, an effective feature extraction mechanism needs to be incorporated into the model.

CTP has multimodalities for the segmentation of the stroke lesions. These multimodal data provide complementary information to perform the segmentation task, which increases the performance compared to single modality [7]. The literature has several fusion methods involved in segmenting the lesions. Now, the question is how carefully the information is used in achieving better performance. There are several ways how these data from multimodalities have been combined or integrated from primitive to advanced methodologies [5]. In the work presented by [28], a multi-level fusion is developed by fusing the information at the image level, matrix level, and feature level for fusing the information between two modalities. It is very hard to decide the fusion technique that provides the information-rich features [48]. Many researchers designed the deep learning architecture to fuse the information from the multimodalities in different ways [25]. This overcomes the problem with the conventional fusion techniques by directly encoding the mapping between multimodalities. They are i) feeding the multi-modalities data as different channels to the model at the input itself, ii) processing these modalities individually in separate paths and combining these features at different levels, and iii) Processing these multiple data modalities

entirely using separate paths. All the basic methodologies of fusing multimodalities are broadly classified into the following strategies: i) early [3, 11], ii) late [32, 38], iii) bottleneck fusion [9], iv) multi-level fusion [26, 35].

The early fusion technique is the most commonly employed fusion method due to its simplicity in implementation and helps in focusing on the network architecture. However, despite incorporating early fusion in methodologies, there is a greater emphasis on diverse concepts. These include i) multiscale processing [26, 40], ii) adversarial learning [41, 47], iii) transfer learning [3, 9], iv) utilization of image synthesis [43], and various other innovative concepts. In the paper [38], the authors utilized the late fusion methodology to segment stroke lesions in CTP data. The various modalities were inputted into four separate U-Nets to carry out the lesion segmentation. They explored a voting classifier, weighted averaging, and logistic regression. Notably, the model employing logistic regression achieved better performance. In this scenario, all modalities were effectively exploited by the individual networks to enhance information quality. The late fusion approach proves advantageous in cases where multimodal information provides limited complementary insights. Chen et al [9] designed an encoder architecture designed to independently extract features from various modalities. These latent representations were subsequently integrated via a hyper-fusion module and fed to the decoder. Additionally, deep supervision was implemented to enhance convergence. In [49], a multi-encoder network was formulated specifically for segmenting brain lesions. This model processed distinct MRI sequences using separate encoders and these features from each modality were merged and fed to the decoder through a fusing block. Both channel and spatial attention mechanisms were utilized to capture informative features. Dolz et al. [13] developed a dense multipath U-Net architecture that incorporated distinct encoders for feature extraction from different modalities. This model employed dense connections within and across the encoders, enhancing the feature sharing. These individual features were fused at every stage within the encoder.

Selecting the fusion technique is still a predominant issue. Early fusion integrates the multimodalities at the input level but fails to establish further complex relationships between multimodalities. In the decision-level fusion (late) strategy, each modality employed a separate encoder to exploit each modality's independent feature representation. The disadvantage is that it requires a lot of memory as it involves the training of separate networks and also no cross-modal interaction between the modalities to establish the relation among the data. The bottleneck fusion strategy extracts the features from individual modalities. It is missing the integrated or joint representations of the multimodalities. Methodologically, each fusion strategy has its own advantages and disadvantages. To address the drawbacks of individual fusion strategies, we introduced a hybrid fusion strategy. This method combines early fusion and bottleneck fusion to effectively leverage the strengths of each fusion technique and efficiently derive features from multimodal data.

A simple or arbitrary fusion of information from multiple sources may not provide optimal results [9, 31]. From that perspective, careful integration of the information from the multimodalities and different fusion strategies necessitates the development of a fusion module with attention. In the literature, few research articles used attention concepts, particularly in the case of stroke lesion segmentation on the CTP data. Attention modules developed for other methodologies may not be helpful in this context as they have different paths in the encoder [1]. In [43], a simple attention block called a squeeze excitation module is used to enhance the encoder's feature extraction. Shi et al [35] proposed a cross-modal cross-attention module to process the different modalities and fuse the information properly from all the modalities. They used group convolutions to process the features in groups and group attention blocks based on the squeeze and excitation attention concept to enhance the important features. Two of such blocks have been used to process four modalities. A multiscale atrous convolution-based attention is used to acquire the multi-scale

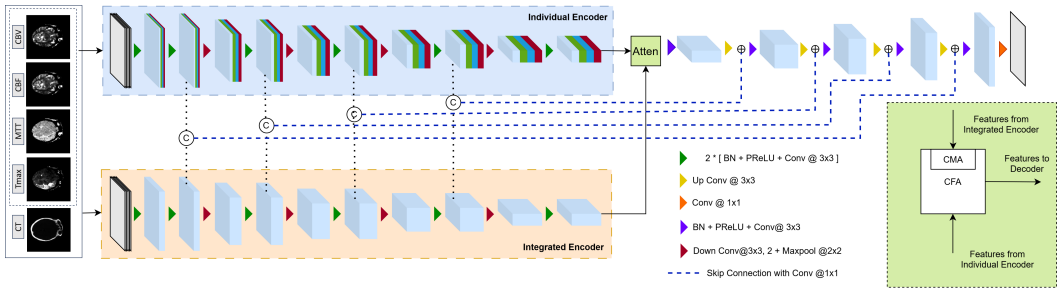

Fig. 1. Architecture of proposed model. The attention module is shown as a separate block in the diagram. The number of features in the individual and integrated encoders start from 32 to 512 and 8 to 128, respectively. The convolution is represented in Conv @ $k_x \times k_y, s_x$ format in which $k_x \times k_y$ represents kernel size and s_x represents stride of the kernel in convolution.

features for providing better features [50]. Few researchers also experimented with the concept of cross-attention in providing attention between the text and images [8]. While many research papers leverage attention mechanisms to improve the quality of features extracted from input images, our approach differs from the typical plug-and-play attention modules. The conventional attention modules are designed to handle a maximum of two inputs. These may not be appropriate for direct use. This is because we require input from four distinct modalities. We have developed an attention module based on non-local networks to address this challenge.

In summary, the literature review emphasizes the critical significance of employing fusion strategies to enhance the efficacy of stroke lesion segmentation by exploiting the advantages of various fusion strategies. This comprehensive review not only overviews the current state of algorithms utilized in this domain. Leveraging these observations, the present study seeks to advance the exploration of hybrid fusion techniques and attention modules to fuse the information from early and bottleneck fusions, capitalizing on the strengths inherent in diverse fusion strategies. The focus is on optimally integrating the information from various modalities and fusion strategies to enhance the segmentation performance.

3 Methodology

The U-Net architecture [34] has emerged as the gold standard for image segmentation and has found application in diverse medical image segmentation challenges, including brain abnormality segmentation [14, 21], coronary artery segmentation [15], polyp segmentation [22], skin lesion delineation [4], MRI reconstruction [29], and MRI translation [30]. Our proposed model is based on the U-Net architecture. This section provides a comprehensive architecture of the proposed methodology, including an in-depth explanation of the hybrid fusion technique and two attention modules used for stroke lesion segmentation.

Fig. 1 shows the proposed model, aiming to segment stroke lesions in CTP data with enhanced speed and accuracy. To achieve this, we have developed a multi-stage deep learning approach, including hybrid fusion and cross-attention techniques. This methodology uses a hybrid fusion approach to extract effective features with different characteristics from the input multimodal data. we utilize a cross-modal attention (CMA) module to fuse the multimodal information from the individual encoders that process the multimodal data separately. Along with this, a cross-fusion attention (CFA) module is employed to integrate information from the early and bottleneck fusion strategies.

3.1 Hybrid Fusion

The hybrid fusion approach is integrated into the encoder section of the U-Net architecture, while the attention modules are positioned at the end of the encoder, specifically at the bottleneck of the U-Net. The encoder consists of two distinct paths: i) The first path focuses on obtaining features from each modality individually and fusing the features at the bottleneck. The name bottleneck fusion originates from the fact that individual features are derived separately and combined at the bottleneck layer. This pathway produces individual features of multimodal data. So, it is referred to as the individual encoder. ii) The second path is responsible for aggregating features from all modalities together, known as early fusion. This configuration is denoted as the integrated encoder as it consolidates integrated features. At the bottleneck of the model, two sets of features are available. They are the integrated features and the individual features, derived from their respective paths.

These two paths have different input configurations. In the early fusion pathway, all four parametric maps and the CT image are provided as inputs. On the other hand, the bottleneck fusion pathway receives the four parametric maps for processing. The reason for excluding CT from the individual features is the complexity of fusing five distinct sets of features. However, recognizing the importance of CT information, we have integrated it into the early fusion pathway. As a result, the input for the early fusion pathway comprises a total of 10 channels. It consists of five channels from various modalities, and an additional five channels are symmetric images of modalities derived from the respective modalities. In contrast, the bottleneck pathway operates with eight input channels.

The basic building blocks for constructing both encoders share a similar structure except for the type of convolution layer employed. In the bottleneck fusion pathway, group convolutions are employed to extract features from various modalities, facilitating isolation among these distinct features. This method eliminates the need for multiple encoders to handle different modalities separately. Differently, in the early fusion pathway, conventional convolutions are utilized to jointly process the modalities. The building blocks found in both encoders include a dual residual block. This block consists of two sets of convolution layers, arranged consecutively with Batch normalization and ReLU layers. Importantly, a residual connection is incorporated within this block to facilitate feature propagation. Additionally, the reduction of feature map resolution is achieved through a combination of max pool operations and strided convolutions. The features from both paths are concatenated together. In the context of the early fusion encoder, the process is relatively straightforward. However, when dealing with the bottleneck fusion approach, certain adjustments are necessary to maintain feature isolation between different modalities. This is because group convolution operates differently from conventional convolution. This requires a special concatenation process to preserve the features within the same modality while ensuring the separation of features from other modalities. This process is clearly demonstrated in Fig. 2.

On the other hand, a single decoder is utilized to enhance the resolution of the feature maps derived from both fusion paths. Previous research studies have suggested using a lightweight decoder to reduce the computational complexity without compromising on performance. In our model, the decoder comprises a single residual block encompassing a set of convolution, batch normalization, and ReLU layers. The features on the decoder side starts from the 128 and ends with 8. The skip connections used in this model collect the features from both encoders at the same stage and perform a 1×1 convolution operation on the features to perform element-wise summation with features in the decoder.

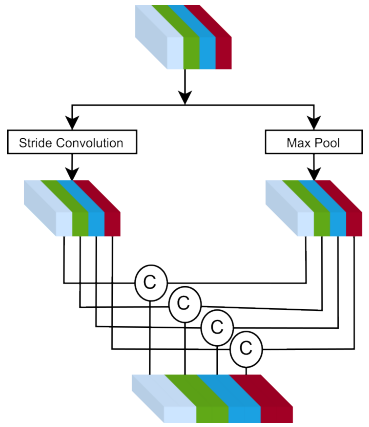

Fig. 2. Process of the concatenation of features from stridden convolution and max pool operations with feature isolation concept. Different colors indicate the features of different modalities. The kernel size for both the strided convolution and the max-pooling operation is 2×2.

3.2 Attention Stage

In the proposed model, the effective fusion of features requires the incorporation of two distinct attention mechanisms. The CMA mechanism is responsible for integrating the information derived from individual modalities, while the CFA mechanism focuses on fusing information generated by the two paths (integrated encoder and individual encoder). These attention modules draw inspiration from various cross-attention strategies found in the literature [17, 35, 44].

3.2.1 Cross Modal Attention (CMA) Module. While existing literature primarily focuses on attention modules designed for fusing two modalities, the proposed model requires the fusion of four distinct modalities. So, we developed a CMA module capable of receiving features from input modalities. The architecture of this CMA module is depicted in Fig. 3, where all features are treated with equal weightage. These modalities are categorized into two distinct groups, namely, the blood parameter group and the time parameter group. This categorization allows for a hierarchical fusion of information. Initially, it combines the information from the two blood parameters using a non-local block concept [44]. The same approach is applied to the time parameters. Subsequently, the features from these two distinct groups are merged using inter and intra-group attention concepts. The features in the CMA module are processed in several stages. As discussed above, the input of the CMA is from the individual encoder, which contains the features from the multimodal input images. Let us consider these features are represented with the $I_{cbv}, I_{cbf}, I_{mtt}, I_{tmax} \in R^{C \times H \times W}$ where C represents the number of channels, and H and W are the height and width of the feature maps. In addition to these, two more features (I_{bg} and I_{tg}), one from each group, are obtained by performing the elementwise summation operation ($\oplus$) between the features in each group. The shape of the generated features is also the same as the input features since elementwise summation is performed. These are obtained as shown in Eq. 1.

$$\begin{aligned} I_{bg} &= I_{cbv} \oplus I_{cbf} \\ I_{tg} &= I_{mtt} \oplus I_{tmax} \end{aligned} \tag{1}$$

In the next step, all the feature maps, along with two obtained feature maps from each group, undergo some operations such as convolution, flattening, and permutations before calculating the attention matrix. The Eq. 2 represents the convolution operation followed by the flattening

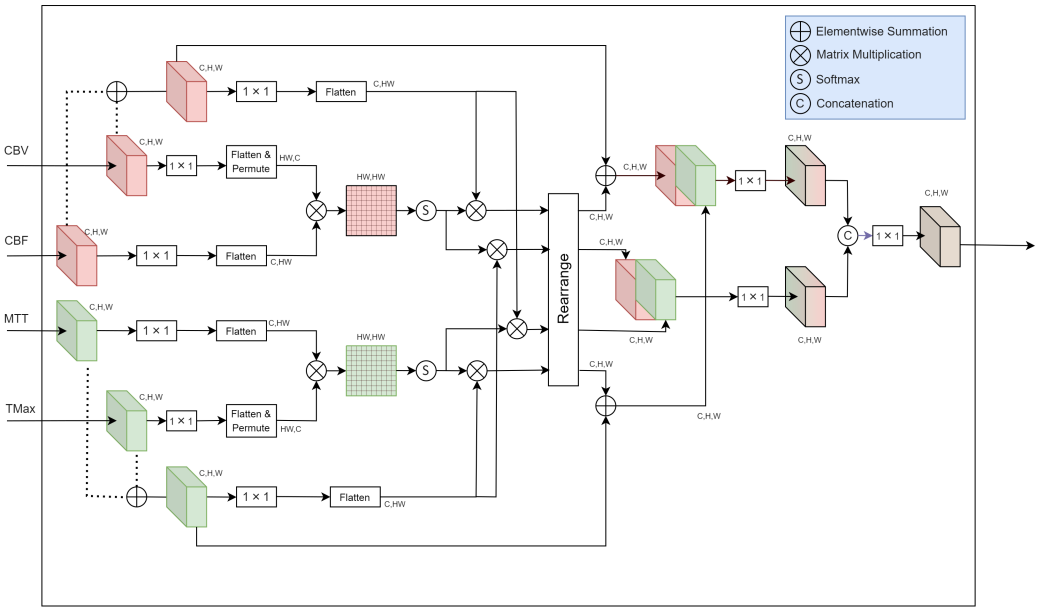

Fig. 3. The architecture of Cross-Modal Attention module

operation. This flattening operation is performed spatially.

$$F_x = f_{flatten}(f_{conv}(I_x)); x \in \{cbv, cbf, mtt, tmax, bg, tg\} \tag{2}$$

where $f_{flatten}(\cdot)$ flattens the features to $C \times HW$ shape, $f_{conv}(\cdot)$ performs a convolution operation with kernel size of 1×1. After these simple operations, the feature F_x has a dimension $R^{C \times HW}$. Further, as shown in Fig. 3, the features from CBV and Tmax undergo permutation operation to change the shape of the features as represented in Eq. 3.

$$F'_x = f_{permute}(F_x); x \in \{cbv, tmax\} \tag{3}$$

where $f_{permute}(\cdot)$ performs permutation operation which changes the shape of these features to $HW \times C$.

The next step is to calculate the attention mechanism between the features belonging to the same group as shown in Eq. 4. In this process, let us consider blood group parameters, a matrix multiplication operation ($\otimes$) is performed between the features F'_{cbv} and F_{cbf} followed by a softmax activation function. Similarly, in the time group, A_t is calculated from F'_{tmax} and F_{mtt}.

$$\begin{aligned} A_b &= S(F'_{cbv} \otimes F_{cbf}) \\ A_t &= S(F'_{tmax} \otimes F_{mtt}) \end{aligned} \tag{4}$$

where S is softmax activation used for normalization. A_b and A_t denote the attention matrices calculated from the features in blood and time parameter groups, respectively. Now, the shape of the attention matrix (A_b) and (A_t) is $R^{HW \times HW}$.

Now, attention features are calculated from these attention matrices. Moreover, we calculate two different types of features for these attention features. They are intra-group attention features and inter-group attention features. First, we discuss the generation of intra-group features. A matrix

multiplication operation is performed between the attention matrices and features (I_{bg}, I_{tg}).

$$FA_b = F_{bg} \otimes A_b$$
$$FA_t = F_{tg} \otimes A_t \tag{5}$$

Similarly, the inter-group attention features are also calculated, but the difference here is that instead of multiplying with the same group attention matrix, the other attention matrix is used.

$$FA_{bt} = F_{tg} \otimes A_b$$
$$FA_{tb} = F_{bg} \otimes A_t \tag{6}$$

The Eq. 5 and Eq. 6 represent the intra-group and inter-group attention features. The shape of these features is $R^{C \times HW}$. These features are subjected to the rearrangement of shape using permutation operation ($f_{permute}$) as in Eq. 7.

$$FA_x = f_{permute}(F_x); x \in \{b, t, bt, tb\} \tag{7}$$

Then, a residual connection is added from the features obtained by combining both features in a group, i.e., I_{bg}, I_{tg}, initially using Eq. 8.

$$FA'_b = FA_b \oplus I_{bg}$$
$$FA'_t = FA_t \oplus I_{tg} \tag{8}$$

where FA'_b and FA'_t represents the intra group features. Next, as shown in Eq. 9, the intra-group features are concatenated ($f_{concate}$), and a convolution operation (f_{conv}) is used to generate the final intra-group features.

$$F_{Intra} = f_{conv}(f_{concate}(FA'_b, FA'_t))$$
$$F_{Inter} = f_{conv}(f_{concate}(FA_{bt}, FA_{tb})) \tag{9}$$

the final attention features that are coming out of the CMA module are represented with F_{final} as shown in Eq. 10.

$$F_{final} = f_{conv}(f_{concate}(F_{Intra}, F_{Inter}) \tag{10}$$

3.2.2 Cross Fusion Attention (CFA) Module. While straightforward or simple information fusion of features from different sources may not yield improved performance, it has become evident that a more sophisticated approach to information fusion is necessary. Attention mechanisms are one of the techniques employed for this purpose. In this context, the non-local block concept has been selected for information fusion. However, a strategy is adopted where two individual paths are assigned to each type of fusion feature, as shown in Fig. 4. The early fused features are adapted to account for pixel-pixel relations and interactions, and a similar operation is applied to the bottleneck features path.

Subsequently, a cross-operation is performed between these two feature types, aiming to extract effective information from both features. This process plays a pivotal role in extracting enhanced information from both integrated and individual features. The operation based on the non-local block concept proves its effectiveness in capturing long-range dependencies, a capability often lacking in standard convolution operations. These attention calculations are particularly effective for small-resolution images. Consequently, applying this attention module is confined to the bottleneck stage. In cases where the image resolution is high, the computation of pixel-pixel interactions among numerous pixels becomes challenging and considerably increases network complexity. To provide the fusion at different stages in the network, we followed a simple fusion process. A convolution block is introduced to facilitate the fusion of information from both fusion types at various levels or stages within the model. In the nonlocal attention calculations, a scaled dot product

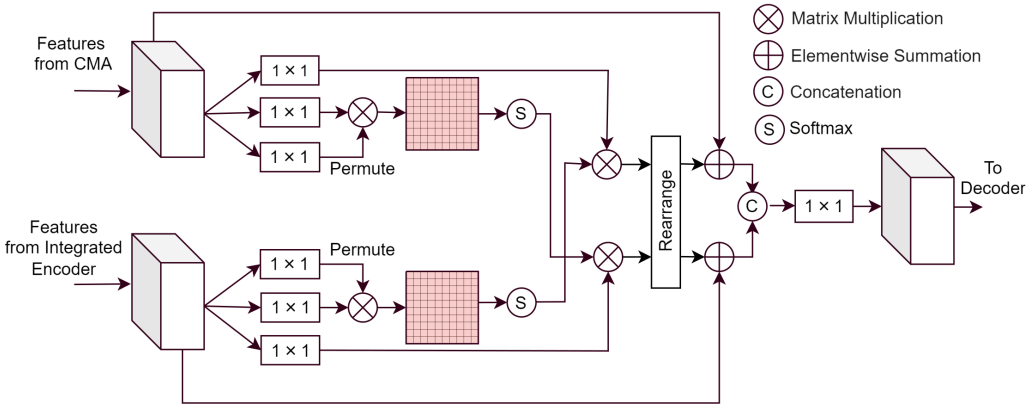

Fig. 4. The architecture of Cross Fusion Attention module

is employed to reduce the values generated after the matrix multiplication instead of a simple dot product. In contrast to the model presented by [44], we decided to continue with the same number of channels throughout the attention block since the number of channels at the deep stage is already kept at a minimum.

The internal architecture details of the CFA module are shown in Fig. 4. The CFA module receives two inputs. They are i) the features from the CMA module, expressed by $I_A \in R^{C \times H \times W}$ and ii) the features from the integrated encoder and denoted by $I_E \in R^{C \times H \times W}$. These features are processed in the CFA in several stages.

In the first stage, the feature maps from the CMA (I_A) are fed to the three parallel convolutional layers with 1×1 kernel size. Later, some features undergo simple operations that change their shape. The resultant features are represented with θ_A, ϕ_A, g_A and also shown in Eq. 11.

$$\begin{aligned} \theta_A &= f_{flatten}(f_{conv}(I_A)) \\ \phi_A &= f_{flatten}(f_{conv}(I_A)) \\ g_A &= f_{flatten}(f_{conv}(I_A)) \end{aligned} \tag{11}$$

Similarly, the features from the integrated encoder are processed and result in the generation of θ_E, ϕ_E, g_E. Later, the features ϕ_A and ϕ_E from both branches are permuted to change the shape of the features to $R^{HW \times C}$, and then these are represented with ϕ'_A and ϕ'_E.

In the next step, the attention matrix (A_A) is calculated by performing a scaled dot product between the features from the ϕ'_A branch with θ_A branch. Now, the attention matrix size is $R^{HW \times HW}$. Similarly, the attention matrix (A_E) is calculated using ϕ'_E and θ_E as shown in Eq. 12.

$$\begin{aligned} A_A &= \phi'_A \otimes \theta_A \\ A_E &= \phi'_E \otimes \theta_E \end{aligned} \tag{12}$$

Now, the attention features are generated using matrix multiplication operations from the attention matrices. The process is shown in Eq. 13.

$$\begin{aligned} F_{AE} &= g_A \otimes A_E \\ F_{EA} &= g_E \otimes A_A \end{aligned} \tag{13}$$

The features F_{AE} and F_{EA} represent the cross-fusion attention features obtained from both branches. Then, these features are permuted to change their shape back to normal. Then, these features are

denoted with F'_{AE} and F'_{EA}. Later, residual connections are added to F'_{AE} and F'_{EA} from the I_A and I_E branches, respectively, as shown in Eq. 14.

$$\begin{aligned} F''_{AE} &= F'_{AE} \oplus I_A \\ F''_{EA} &= F'_{EA} \oplus I_E \end{aligned} \tag{14}$$

After this residual connection, the features from both the branches are combined together using a concatenate operation followed by a convolution operation with kernel size of 1×1. The final features are the cross-attention features obtained from early and bottleneck fusions, as shown in Eq. 15.

$$F_{final} = f_{conv}(f_{concate}(F''_{AE}, F''_{EA})) \tag{15}$$

4 Experiments and Results

The proposed methodology undergoes rigorous testing on the ISLES 2018 dataset using a cross-validation approach to demonstrate its efficacy. The same methodology is applied to evaluate the test set of the ISLES 2018 dataset. The evaluation of the model's performance employs metrics widely recognized for assessing the quality of segmented results. Throughout the experiments, consistent hyperparameters are employed, ensuring uniformity across all evaluations. The quantitative and qualitative results are presented through a combination of tables and figures, providing a comprehensive assessment of the model's performance.

4.1 Experimental Procedure

4.1.1 Dataset. The Ischemic Stroke Lesion Segmentation Challenge (ISLES) 2018 dataset is particularly for the segmentation of CTP data. It provides train data consisting of 94 cases and test data consisting of 63 cases. The following data is provided: i) raw dynamic CTP data, ii) post-processed perfusion maps, i.e., CBF, CBV, MTT, Tmax, iii) CT without skull stripping, and iv) Ground Truth segmented on DWI. In these experiments, we used only CT and CTP maps. The data provided is 3D in nature, but the number of slices in the axial direction keeps varying from case to case, ranging from 2 to 22. The resolution of each scan is 256×256. This is also a reason for choosing the 2D processing techniques. The whole data is converted into the 2D slices through which our model is trained and tested.

4.1.2 Pre-processing and Augmentation. Two different preprocessing techniques are used: i) intensity normalization and ii) CT skull stripping. A simple intensity normalization technique is applied to all images to normalize the data to zero mean and unit variance. The CT images have the brightest pixel values on the skull region, which are not required for the stroke lesion estimation. Furthermore, removing those skull-related pixels enhances the visibility of the brain-related tissues. The method followed for skull stripping from the CT image is utilized in [11]. The parametric map image contains only the soft tissues of the brain. The intensity values in these four images are added to create a binary mask of the brain tissue. This mask is created by using non-zero values after summation. In the mask, all the pixels related to the skull will be zero. The CT is multiplied by the mask generated in the previous step. Now, the pixels related to the skull will be zero in the output image, i.e., the skull-stripped image.

The symmetric augmented modality is utilized in this experimentation for all the models, including baseline models, along with the other primitive augmentation techniques such as rotation, horizontal flip, and vertical flip during the training of the model. In detail, the symmetric version of each modality is created and fed to the network along with the original. The motive behind this augmentation is to extract the brain's quasi-symmetry property. Mainly, the stroke occurs on one side of the brain [44]. When a stroke occurs, the symmetrical nature of the brain is disturbed,

which is a clear indication of a stroke. This symmetric version of the brain is obtained by i) flipping the image and ii) applying the FLIRT [20] method on the flipped version of the original image for registration. Hence, feeding both original and symmetric versions enhances the visibility of the stroke. The primitive augmentation techniques are applied on-the-fly augmentation to avoid storing the images separately. The output for the preprocessing steps such as CT skull stripping and symmetric augmentation is shown in Fig. 5.

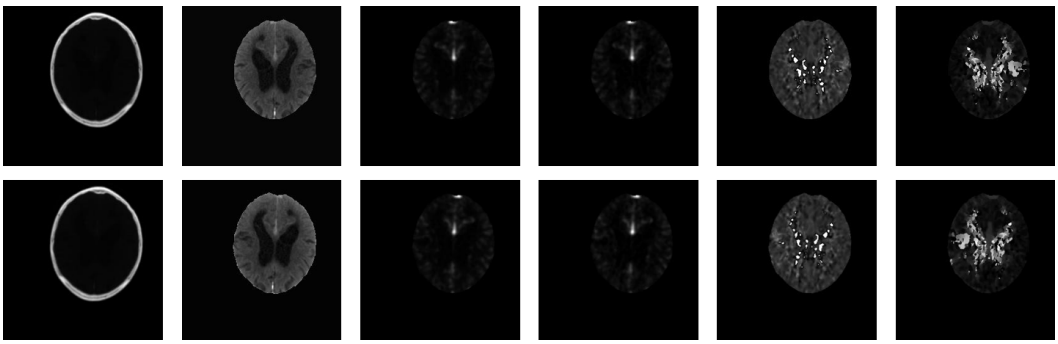

Fig. 5. The output of the preprocessing techniques. From left to right, CT input, skull stripped CT input, CBF, CBV, MTT, TMax. The first row represents the original input, and the second row represents its symmetric augmentation.

4.1.3 Experimental Configurations. The experiments are conducted on an Intel-based Xeon processor with 128GB of RAM and 16G of Tesla-T4 Graphic card memory from Nvidia. The models were developed using PyTorch Libraries and the Jupyter interface. The hyperparameters are configured as epoch: 200, optimizer: AdaDelta, batch size: 8. To combat the inherently presented class imbalance, a combination of focal loss and generalized dice loss is used, a common strategy to address the imbalances in medical datasets. For training and validation, the ISLES 2018 training dataset was split into an 80:20 ratio, respectively.

4.2 Experimental Results

4.2.1 Quantitaive Analysis. This experimentation aims to provide a robust and fast deep-learning model to perform the segmentation of lesions from the CTP data. In this section, we qualitatively assess the improvements from the proposed concepts of hybrid fusion, including both early and bottleneck fusions along with cross-attention methods. Two cross-attention modules, i) CMA is to fuse the features from the different modalities carefully, ii) CFA is to fuse the various types of features from the different fusion strategies. A step-by-step approach is followed to evaluate the proposed model. In this experimentation, two models are considered as baselines, i.e., early and bottleneck fusion-based deep learning models. Each module is added to the model and evaluated to show the performance of the added blocks. Each assessment entails utilizing a 5-fold cross-validation approach applied to the ISLES 2018 dataset comprising images from 94 labeled cases. The quantity of training and validation images differs in each fold due to variations in the number of slices for each case. Additionally, it is ensured that all slices belonging to one individual remain in the same set. The evaluation metrics used to calculate the performance of the models and additional blocks are Dice score, Hausdroff distance (HD), Precision, Recall, and Absolute volume difference (AVD). Table 1 presents evaluation metrics for model comparisons. It shows the highest Dice score achieved by the model consisting of hybrid fusion with CMA and CFA.

Table 1. Quantitative assessment of baseline models and addition modules incorporated into models

Model	Dice	HD	Precision	Recall	AVD
Early Fusion	0.482 ± 0.32	3.51 ± 1.85	0.4843 ± 0.29	0.5873 ± 0.30	2.37 ± 4.31
Bottleneck Fusion	0.491 ± 0.31	3.92 ± 1.97	0.4724 ± 0.26	0.6354 ± 0.28	2.18 ± 4.07
Hybrid Fusion	0.511 ± 0.31	4.07 ± 2.06	0.4664 ± 0.28	0.6971 ± 0.26	2.01 ± 3.89
Hybrid Fusion with CMA	0.516 ± 0.30	4.10 ± 2.08	0.4921 ± 0.26	0.5620 ± 0.28	1.99 ± 3.54
Proposed method	0.521 ± 0.29	3.04 ± 1.76	0.5341 ± 0.24	0.6235 ± 0.25	1.75 ± 3.12

Among the baselines, the early fusion model exhibited a lower dice coefficient in comparison to the bottleneck fusion model, primarily attributed to the superior adaptability in learning from multimodal data. Moreover, the bottleneck fusion approach, where multiple encoders are employed for various modalities, enables the more effective extraction of explicit information from the multi-modal data. While the bottleneck fusion approach did yield a higher Dice coefficient compared to early fusion, it is not without its drawbacks, notably in terms of potentially overlooking integrated or joint features. The proposed methodology introduces a hybrid fusion network that merges aspects of both early and bottleneck fusions, resulting in an improved performance level. As stated in [9], an arbitrary or simple fusion of features may not necessarily produce optimal results. With this consideration, we developed two attention mechanisms for effective feature fusion. The CMA module integrates data from various modalities into common data, distinct from the data processed by the encoder designed for early fusion. Following the incorporation of the CMA module, the Dice coefficient shows a more substantial increase, reaching 0.516, a testament to the effective fusion of multimodal features. Subsequently, an additional attention module known as CFA was introduced to integrate the distinct characteristic features from both encoders. This integration, encompassing information from both the individual and integrated paths, significantly boosted the Dice coefficient, reaching a value of 0.521.

The HD metric is used to measure the quality of the boundary. Ideally, HD should be less. The HD for hybrid fusion is a little higher compared to the early bottleneck. Instead of analyzing using HD alone, we correlate this with the Dice coefficient. The dice coefficient implies more correlation between prediction and ground truth. A small fraction of the increase in HD shows that the model struggles to predict the pixel along the boundary of the lesion. This may be due to an increase in a few false positives while a substantial decrease in the false negatives by the model. The incorporation of attention modules reduces the false positives and false negatives around the boundaries. So, the proposed model achieves the best HD among all the models.

Among the baselines, the highest precision score is achieved by the early fusion, indicating that it generates fewer false positives than the bottleneck fusion. The proposed model achieves the highest precision indicating that it minimizes false positives better than other models. The hybrid fusion achieves the highest recall, meaning it identifies fewer false negatives but has the highest number of false positives. Adding the CMA to hybrid fusion reduces the recall, providing a trade-off between precision and recall. Finally, adding CFA further increases both precision and recall. From the base model to the proposed model, the precision and recalls have increased by 10.2% and 6.1%, respectively. The mean AVD shows the difference in volumes between the predicted and original. The AVD is progressively decreasing from the base model to the proposed model indicating that the predicted volume is very much similar to the original volume for the proposed model.

We performed 5-fold cross-validation on the proposed model, and Table 2 shows the loss and dice coefficient for both training and validation sets. We also reported the number of images in train and validation sets for better interpretation of the scores. It also contains the number of images for

Table 2. 5 fold cross-validation of the proposed model on the train data

Fold #	Train			Validation		
	# of Images	Dice	Loss	# of Images	Dice	Loss
Fold 1	428	0.6613	0.3568	66	0.4709	0.3435
Fold 2	432	0.7308	0.3514	62	0.6011	0.2488
Fold 3	444	0.747	0.3514	50	0.4797	0.3294
Fold 4	452	0.7277	0.3388	42	0.6297	0.2744
Fold 5	224	0.7331	0.3566	270	0.4214	0.4627
Average		0.7199	0.351		0.5205	0.3317

training and testing. From Table 2, we can observe that training performance is stable across folds, with average Dice scores and loss values showing consistency. Validation performance varies more due to differences in the size and difficulty of validation sets. Fold 5's split highlights the impact of data imbalance, emphasizing the need to ensure equal distribution across folds in cross-validation. Moreover, this imbalance is caused because of variations in the number of slices in the dataset. The average metrics suggest the model has good training performance and moderate generalization to unseen data.

We also experimented to find the influence of the inclusion of CT. CT provides the structural information and whereas the CTP maps provide the blood dynamics in the brain. This combined input enables the model to leverage both anatomical and functional features, which enhances lesion detection and segmentation. The inclusion of CT in our model design leverages the full potential of the dataset by integrating the structural and perfusion information it provides. Along with that, we conducted an ablation study comparing model performance with and without CT as an input. The results demonstrated that the model with CT consistently outperformed the one without CT in terms of segmentation accuracy, further supporting its inclusion. These findings are shown in Table 3.

Table 3. Quantitative assessment for inclusion of CT as an input in the proposed model

Model	**Dice**	**HD**	**Precision**	**Recall**	**AVD**
Without CT	0.518 ± 0.29	3.56 ± 1.81	0.5221 ± 0.25	0.5920 ± 0.27	2.15 ± 3.98
With CT	0.521 ± 0.29	3.04 ± 1.76	0.5341 ± 0.24	0.6235 ± 0.25	1.75 ± 3.12

4.2.2 Qualitative Analysis. In the context of qualitative analysis, examining segmented images provides valuable insights into performance, any inconsistencies, errors, and improvements. It helps evaluate the model's adaptability and boundary definition by including various modules, enhancing performance. It allows us to evaluate the model's ability to capture intricate details, define boundaries, and adapt to changes within the images.

We provide the qualitative results from the baseline to the proposed model, including various modules sequentially. Fig. 6 showcased different examples ranging from small to large lesions. These samples were chosen to represent the heterogeneity of the lesion size observed across the dataset. For instance, the baseline model is able to detect the lesion location. Still, the borders are unclear in the baselines, with more false positives and false negatives. Compared to the baseline models, the hybrid fusion model shows better precision and recall, representing a reduction in false positives and false negatives. The addition of CMA introduces a small number of false negatives while reducing a large number of false positives. The introduction of CFA further reduces the false

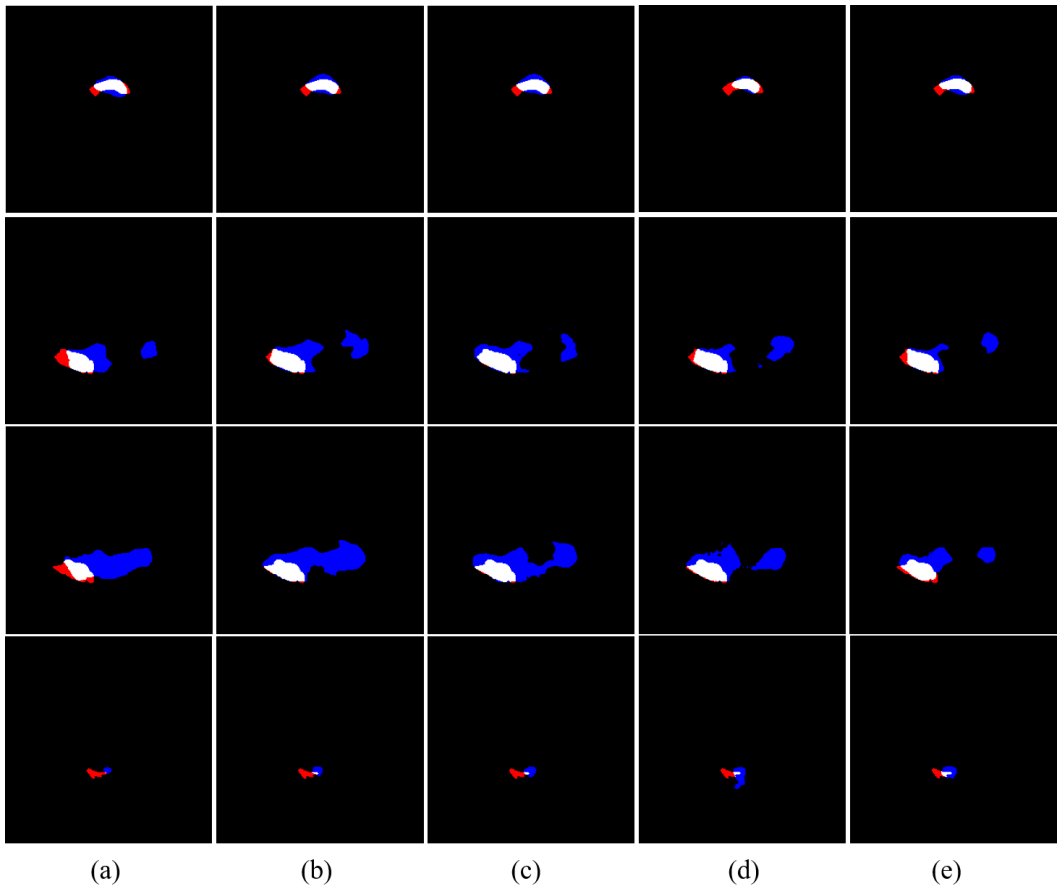

Fig. 6. Visualization of ischemic stroke core prediction of base models, hybrid fusion, and incremental developments for four subjects in the validation set. From Left to right, prediction by (a) early fusion, (b) bottleneck fusion, (c) hybrid fusion, (d) hybrid fusion with CMA, and (e) hybrid fusion with CMA and CFA. White denotes the true positives (TP), blue denotes false positives (FP), and false negatives (FN) are denoted in red. The TP and FN combined represent positive instances of ground truth. The TP and FP combined represent the predicted positive instances.

negatives and false positives by increasing the overlap between predicted and ground truth regions. Apart from these results shown here, there are a few instances where the baseline fusion strategies could not detect the input images without lesion, i.e., true negatives, whereas the hybrid fusion and its variants detected them correctly.

The scatter plots in Fig. 7 showcase the relationship between the original and predicted volumes under four different fusion strategies: (i) Early fusion, (ii) Bottleneck Fusion, (iii) Hybrid Fusion, and (iv) Hybrid Fusion with attention modules that offer valuable insights into the model's performance. To quantify these plots, R^2 value is employed as a metric to measure the correlation. However, volume difference alone is insufficient to conclude model performance, as it must be interpreted along with the Dice score. These plots coincide more with the AVD metric results. The R2 values for the model with early fusion are 0.69, the model with bottleneck fusion 0.71, the model with hybrid fusion is 0.75, and finally, the model with hybrid fusion along with attention modules is 0.79.

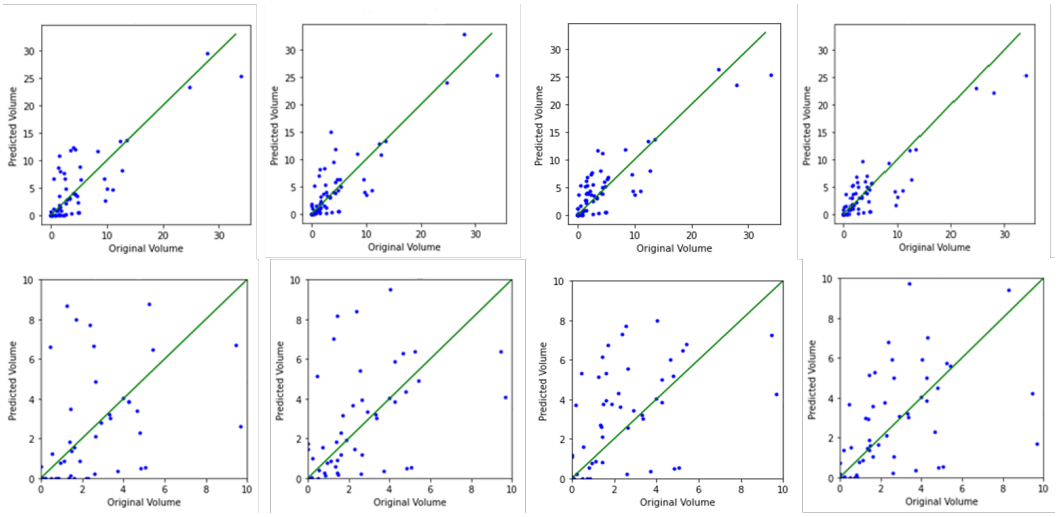

Fig. 7. The scatter plots between original and predicted volume. From left to right: i) Early fusion, ii) Bottleneck Fusion, iii) Hybrid Fusion, iv) Hybrid Fusion with Attention Modules. Additionally, the zoomed version (represented in the bottom row) of the scatter plot enhances the view of the small lesions.

These values show that the proposed additional modules, hybrid fusion, and attention modules, contribute towards better results in terms of predicting the stroke lesions approximately near to the original volume. The progression from early fusion to hybrid fusion with attention modules is marked by a notable enhancement in performance. This is evident from the scatter plots, where a clear improvement in volume prediction can be observed. It is important to note the Dice is also increasing in a similar pattern suggesting a better performance.

4.2.3 Analysis of the proposed architecture. Several experiments are carried out before settling on the final model. In this section, we explained the different variations of the model, a progressive approach to reach the final proposed model. These experiments are related to variations of the hyperparameters in the additional modules added to the network. For all the experiments, dice score and loss metrics are shown in Table 4. The first variation of the proposed model is the number of filers in the model. We developed a model with 8 and 16 as base filters. However, a model with 8 base filters shows better performance than the 16. This may be due to an increase in the number of parameters in the model with 16 base filters. Moreover, the variation of these two models is very small. Similarly, we also tested different variations in the attention module. In CMA, we first calculated only intra-group features. Later, we also implemented the inter-group features, which eventually enhanced the performance. In CFA, we decided not to reduce the number of features during the convolution layers since these features are already fewer. We performed two different experiments in relation to the decoder in the hybrid fusion model. They are the decoder with the same number of features as the encoder and a smaller number of features in the decoder than the encoder in the hybrid fusion model. Here, the decoder with fewer features worked better than the encoder with the same number of features.

Symmetric augmentation is used while training the model. Considering this perspective, a flipped version of the input might enhance the model's predictive abilities. An experimental procedure entails the model saving predictions derived from both the original and flipped versions of the input. These predictions are averaged to generate the final segmentation map.

Table 4. Performance metric for the parameters selected

Module in which variation involved	Variation	Training		Validation	
		Dice	Loss	Dice	Loss
Base filters in encoder	8	0.665	0.3157	0.511	0.3415
	16	0.657	0.3341	0.505	0.3618
CMA	Intra Group	0.655	0.3283	0.512	0.3398
	Intra & Inter Group	0.671	0.3194	0.516	0.3329

5 Discussion

We developed and tested a deep learning model using hybrid fusion, CMA, and CFA. The underlying concept of acquiring both integrated and individual features is to extract diverse feature characteristics from various modalities while preserving combined features from all modalities. This approach allows the model to acquire knowledge about both integrated and individual features, which is essential due to the distinct information carried by each modality. In this scenario, few modalities contain enriched information related to the core, and others may have sufficient information about the penumbra. However, the specific task is to segment core-related information from all modalities. Individual features enhance the ability of the model to understand the individual characteristics, while integrated features, representing characteristics across all modalities, facilitate joint representation learning that may not be disregarded. It can be observed that fusing features at a higher level yields superior results compared to early fusion strategies. Additionally, segmentation results show a noticeable visual enhancement by allowing the architecture the flexibility to determine the appropriate level of abstraction for combining various modalities. We modified the basic U-Net architecture to process the information present in multimodal data. Drawing inspiration from various studies in multimodal learning for medical image segmentation, the proposed approach involved processing each imaging modality separately within one encoding path and subsequently integrating them within another branch of the encoding path, which processes the multimodal data combined. This design provides the model with the flexibility to learn patterns both individually and in combination, thereby enhancing its ability to learn and represent complex information.

The attention modules, CMA and CFA, are inspired by non-local block networks and cross-modal attention networks. The primary focus is to effectively integrate the enriched contexts of individual features and also consider the global dependencies across the modalities. Hence, the combined feature response has a rich and joint representation that is more relevant. These attention modules have integrated diverse attention-learning mechanisms within their framework to enhance cross-modal feature extraction. Since the non-local operation [44] is an efficient attention process, it gives the pixel-to-pixel relations, i.e., long-range dependencies across all pixels in an image. Specifically, the non-local operation is designed to capture robust contextual information by means of spatial attention. Both attention modules are designed based on this concept. These experimental results demonstrate an increase in inter-class variations and highlight prominent features to suppress the irrelevant features. Few observations in this experimentation indicate that the cross-attention module architecture can more effectively capture the global interaction information, which is essential for semantic understanding and segmentation of complex multimodal information.

The dice scores achieved on the ISLES testing dataset by the proposed methodology, along with the state-of-the-art methodologies, are shown in Table 5. The inference time is shown in Table 6. Our proposed methodology involves an image-level prediction leading to quick inference time surpassing the time taken by the methods using the patch processing technique. The methods followed by [11, 23] involve patch processing, thus taking a lot of time in preprocessing and

Table 5. Comparison of other state-of-the-art models on test set computed by ISLES-2018 challenge

Reference	Dice	Precision	Recall	AVD
clera2 [11]	0.49	0.51	0.57	12.18
ghosp1 [23]	0.49	0.58	0.50	9.30
ours	**0.48**	**0.52**	**0.52**	**10.12**
tianayu [35]	0.48	0.48	0.59	-
baebj1	0.48	0.56	0.49	10.19
cheny11 [9]	0.48	0.59	0.46	9.7
xiaoh3 [19]	0.47	0.56	0.49	11.14

Table 6. Comparison with the state-of-the-art in terms of inference time

Reference	Time (Sec)
clera2 [11]	1.56 seconds
ghosp1 [23]	∼1 second
ours	0.3 second

prediction as well. In this paper [11], the authors also used uncertainty filtering, which causes three times more inference time. Moreover, each patch must be predicted separately, eventually leading to more inference time. In contrast, the proposed method involves an image-level prediction leading to a quick inference time. Moreover, it does not involve any complex preprocessing techniques.

Several factors contribute to this performance increase: i) The progression from early fusion to hybrid fusion, and further, the integration of attention modules, refines the precision of feature fusion. As a result, the model becomes more effective at capturing and exploiting the information from multiple modalities, leading to more accurate predictions. ii) Introducing attention mechanisms, such as cross-modal and cross-fusion, empowers the model to focus on critical regions or features, which may be especially important for small-volume lesions. This increased discriminative power allows the model to distinguish subtle differences more effectively iii) With improved feature integration and attention mechanisms, the model is better equipped to reduce variability in predictions, particularly for smaller lesions.

The main contributions of this article are the hybrid fusion and cross-attention modules. We performed experiments to prove the effectiveness of the hybrid fusion and attention modules. The input to the model is CT and CTP maps as whole images with direct and symmetric versions of the modality that help the model to get information about tissue in the other hemisphere region [6]. The proposed method is developed while using minimal pre-processing techniques such as intensity normalization and skull stripping. We have not also incorporated any post-processing techniques for the final prediction of the segmentation, which was followed in the literature. Including the multiple fusion strategies in the model helps overcome the issues associated with any single fusion strategy by acquiring various types of features (integrated and individual features) from different strategies. Developing the attention modules to fuse the multimodal data helps to enhance the performance significantly. The image level predictions allow the model to provide the segmentation results with less inference time. The limitations and future scope of this work are mentioned as follows. The ISLES 2018 dataset provides 4D CTP raw data in addition to the current inputs, including CT, CBF, CBV, MTT, and Tmax. 4D CTP data captures the temporal dynamics of blood flow in the brain, offering potentially richer information. However, our approach does not utilize the 4D CTP data, thereby avoiding additional computational complexity. Furthermore, while the

data itself is 3D, the proposed model operates on 2D image slices. This choice is motivated by the variability in the number of slices across datasets, which poses challenges for consistent 3D processing. In the future, we will include the 4D CTP data to explore the temporal dynamics for performance improvement.

6 Conclusion

Stroke is the second leading cause of death and permanent disability, so it is very necessary to detect in time the presence of stroke in the brain imaging modalities. While the manual approaches are more tedious, the automated methods can accomplish the task within very little time to reduce the burden on the radiologists. In this paper, we developed a hybrid fusion with attention modules based 2D deep learning model to segment the stroke lesions on the CTP data. We have assessed the advantages of employing the hybrid fusion strategy to collect the individual and integrated features from the multimodal images for the final prediction, leading to a decrease in the occurrences of false negatives and false positives. Along with that, the cross-modal attention module raises the dice score level further by properly fusing the features from the multimodalities. Moreover, the cross-fusion attention integrates the features from both encoders, which further increases performance. The proposed model achieves similar performance in terms of dice and outperforms most prior approaches in terms of faster inference times. The ablation study is also conducted to contrast the outcomes of various model variants. Finally, we evaluated the proposed method's performance against select top-performing approaches from the ISLES 2018 challenge, highlighting our model's state-of-the-art performance. It secured the 0.48 dice coefficient on the testing dataset. In the future, we plan to increase the dice score by incorporating more contextual information and using raw CT perfusion data and the perfusion parameter maps.

Acknowledgments

The authors acknowledge the Researchers Supporting Project number (RSP2025R34), King Saud University, Riyadh, Saudi Arabia.

References

[1] Hamdan Al Jowair, Mansour Alsulaiman, and Ghulam Muhammad. 2024. Multi-focal channel attention for medical image segmentation. *Expert Systems* February (2024), 1–21. https://doi.org/10.1111/exsy.13588

[2] Kimberly Amador, Alejandro Gutierrez, Anthony Winder, Jens Fiehler, Matthias Wilms, and Nils D Forkert. 2024. Providing clinical context to the spatio-temporal analysis of 4D CT perfusion to predict acute ischemic stroke lesion outcomes. *Journal of Biomedical Informatics* 149 (2024), 104567. https://doi.org/10.1016/j.jbi.2023.104567

[3] Vikas Kumar Anand, Mahendra Khened, Varghese Alex, and Ganapathy Krishnamurthi. 2019. Fully Automatic Segmentation for Ischemic Stroke Using CT Perfusion Maps. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer International Publishing, Cham, 328–334.

[4] Ginni Arora, Ashwani Kumar Dubey, Zainul Abdin Jaffery, and Alvaro Rocha. 2023. Architecture of an effective convolutional deep neural network for segmentation of skin lesion in dermoscopic images. *Expert Systems* 40, 6 (2023), 1–13. https://doi.org/10.1111/exsy.12689

[5] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16, 6 (2010), 345–379. https://doi.org/10.1007/s00530-010-0182-0

[6] Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens. 2019. Contra-Lateral Information CNN for Core Lesion Segmentation Based on Native CTP in Acute Stroke. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer International Publishing, Cham, 263–270.

[7] Liang Chen, Paul Bentley, and Daniel Rueckert. 2017. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage: Clinical* 15, June (2017), 633–643. https://doi.org/10.1016/j.nicl.2017.06.016

[8] Weidong Chen, Guorong Li, Xinfeng Zhang, Hongyang Yu, Shuhui Wang, and Qingming Huang. 2021. Cascade Cross-modal Attention Network for Video Actor and Action Segmentation from a Sentence. *MM 2021 - Proceedings of*

the 29th ACM International Conference on Multimedia (2021), 4053–4062. https://doi.org/10.1145/3474085.3475534

[9] Yu Chen, Jiawei Chen, Dong Wei, Yuexiang Li, and Yefeng Zheng. 2019. OctopusNet: A Deep Learning Segmentation Network for Multi-modal Medical Images. In *Multiscale Multimodal Medical Imaging: First International Workshop, MMMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings* (Shenzhen, China). Springer-Verlag, Berlin, Heidelberg, 17–25. https://doi.org/10.1007/978-3-030-37969-8_3

[10] Ying Chen, Rui Yao, Yong Zhou, Jiaqi Zhao, Bing Liu, and Abdulmotaleb El Saddik. 2023. Black-box Attack against Self-supervised Video Object Segmentation Models with Contrastive Loss. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023). https://doi.org/10.1145/3617502

[11] Albert Clèrigues, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, and Xavier Lladó. 2019. Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks. *Computers in Biology and Medicine* 115 (dec 2019). https://doi.org/10.1016/j.compbiomed.2019.103487

[12] Lucas de Vries, Bart J. Emmer, Charles B.L.M. Majoie, Henk A. Marquering, and Efstratios Gavves. 2023. PerfU-Net: Baseline infarct estimation from CT perfusion source data for acute ischemic stroke. *Medical Image Analysis* 85 (apr 2023). https://doi.org/10.1016/j.media.2023.102749

[13] Jose Dolz, Ismail Ben Ayed, and Christian Desrosiers. 2019. Dense Multi-path U-Net for Ischemic Stroke Lesion Segmentation in Multiple Image Modalities. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I* (Granada, Spain). Springer-Verlag, Berlin, Heidelberg, 271–282. https://doi.org/10.1007/978-3-030-11723-8_27

[14] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. 2017. Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. *Communications in Computer and Information Science* 723 (2017), 506–517. https://doi.org/10.1007/978-3-319-60964-5_44

[15] Jingfan Fan, Jian Yang, Yachen Wang, Siyuan Yang, Danni Ai, Yong Huang, Hong Song, Aimin Hao, and Yongtian Wang. 2018. Multichannel Fully Convolutional Network for Coronary Artery Segmentation in X-Ray Angiograms. *IEEE Access* 6 (2018), 44635–44643. https://doi.org/10.1109/ACCESS.2018.2864592

[16] Valery L. Feigin, Michael Brainin, Bo Norrving, Sheila Martins, Ralph L. Sacco, Werner Hacke, Marc Fisher, Jeyaraj Pandian, and Patrice Lindsay. 2022. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. *International Journal of Stroke* 17, 1 (2022), 18–29. https://doi.org/10.1177/17474930211065917

[17] Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, and Dacheng Tao. 2023. Hierarchical Multi-Attention Transfer for Knowledge Distillation. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2, Article 51 (Sept. 2023), 20 pages. https://doi.org/10.1145/3568679

[18] Arsany Hakim, Søren Christensen, Stefan Winzeck, Maarten G. Lansberg, Mark W. Parsons, Christian Lucas, David Robben, Roland Wiest, Mauricio Reyes, and Greg Zaharchuk. 2021. Predicting Infarct Core from Computed Tomography Perfusion in Acute Ischemia with Machine Learning: Lessons from the ISLES Challenge. *Stroke* 52, July (2021), 2328–2337. https://doi.org/10.1161/STROKEAHA.120.030696

[19] Xiaojun Hu, Weijian Luo, Jiliang Hu, Sheng Guo, Weilin Huang, Matthew R. Scott, Roland Wiest, Michael Dahlweid, and Mauricio Reyes. 2020. Brain SegNet: 3D local refinement network for brain lesion segmentation. *BMC Medical Imaging* 20, 1 (2020), 1–10. https://doi.org/10.1186/s12880-020-0409-2

[20] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* 17, 2 (oct 2002), 825–841. https://doi.org/10.1006/NIMG.2002.1132

[21] Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, and Ben Glocker. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis* 36 (2017), 61–78. https://doi.org/10.1016/j.media.2016.10.004

[22] Taehun Kim, Hyemin Lee, and Daijin Kim. 2021. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 2167–2175. https://doi.org/10.1145/3474085.3475375

[23] Amish Kumar, Palash Ghosal, Soumya Snigdha Kundu, Amritendu Mukherjee, and Debashis Nandi. 2022. A lightweight asymmetric U-Net framework for acute ischemic stroke lesion segmentation in CT and CTP images. *Computer Methods and Programs in Biomedicine* 226 (2022), 107157. https://doi.org/10.1016/j.cmpb.2022.107157

[24] Brady Laughlin, Alex Chan, Waimei Amy Tai, and Parham Moftakhar. 2019. RAPID automated CT perfusion in clinical practice. *Practical Neurology* 2019 (2019), 41–55.

[25] Hui Liu, Shanshan Li, Jicheng Zhu, Kai Deng, Meng Liu, and Liqiang Nie. 2023. DDIFN: A Dual-discriminator Multi-modal Medical Image Fusion Network. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 4, Article 145 (Feb. 2023), 17 pages. https://doi.org/10.1145/3574136

[26] Liangliang Liu, Shuai Yang, Li Meng, Min Li, and Jianxin Wang. 2019. Multi-scale Deep Convolutional Neural Network for Stroke Lesions Segmentation on CT Images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic*

Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I (Granada, Spain). Springer-Verlag, Berlin, Heidelberg, 283–291. https://doi.org/10.1007/978-3-030-11723-8_28

[27] Pengbo Liu. 2019. Stroke Lesion Segmentation with 2D Novel CNN Pipeline and Novel Loss Function. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer International Publishing, Cham, 253–262.

[28] Wenbing Lv, Saeed Ashrafinia, Jianhua Ma, Lijun Lu, and Arman Rahmim. 2020. Multi-level multi-modality fusion radiomics: Application to PET and CT imaging for prognostication of head and neck cancer. *IEEE Journal of Biomedical and Health Informatics* 24, 8 (2020), 2268–2277. https://doi.org/10.1109/JBHI.2019.2956354

[29] Jun Lyu, Guangming Wang, and M. Shamim Hossain. 2024. Iterative Temporal-spatial Transformer-based Cardiac T1 Mapping MRI Reconstruction. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 6, Article 179 (March 2024), 18 pages. https://doi.org/10.1145/3643640

[30] Jun Lyu, Shouang Yan, and M. Shamim Hossain. 2024. DBGAN: Dual Branch Generative Adversarial Network for Multi-Modal MRI Translation. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 8, Article 235 (June 2024), 22 pages. https://doi.org/10.1145/3657298

[31] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H. Falk. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion* 76, July (2021), 355–375. https://doi.org/10.1016/j.inffus.2021.06.007

[32] Chintha Sri Pothu Raju, Anish Monsley Kirupakaran, Bala Chakravarthy Neelapu, and Rabul Hussain Laskar. 2023. Ischemic Stroke Lesion Segmentation in CT Perfusion Images Using U-Net with Group Convolutions. In *Computer Vision and Image Processing*, Deep Gupta, Kishor Bhurchandi, Subrahmanyam Murala, Balasubramanian Raman, and Sanjeev Kumar (Eds.). Springer Nature Switzerland, Cham, 276–288.

[33] Ryan A Rava, Alexander R Podgorsak, Mohammad Waqas, Kenneth V Snyder, Elad I Levy, Jason M Davies, Adnan H Siddiqui, and Ciprian N Ionita. 2021. Use of a convolutional neural network to identify infarct core using computed tomography perfusion parameters. In *Medical Imaging 2021: Image Processing*, Ivana Išgum and Bennett A Landman (Eds.), Vol. 11596. International Society for Optics and Photonics, SPIE, 1159611. https://doi.org/10.1117/12.2579753

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M Wells, and Alejandro F Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[35] Tianyu Shi, Huiyan Jiang, and Bin Zheng. 2022. C2MA-Net: Cross-Modal Cross-Attention Network for Acute Ischemic Stroke Lesion Segmentation Based on CT Perfusion Scans. *IEEE Transactions on Biomedical Engineering* 69, 1 (2022), 108–118. https://doi.org/10.1109/TBME.2021.3087612

[36] Mohammad Shorfuzzaman, M. Shamim Hossain, and Abdulmotaleb El Saddik. 2021. An Explainable Deep Learning Ensemble Model for Robust Diagnosis of Diabetic Retinopathy Grading. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 3s, Article 113 (Oct. 2021), 24 pages. https://doi.org/10.1145/3469841

[37] Mohsen Soltanpour, Pierre Boulanger, and Brian Buck. 2024. CT Perfusion Map Synthesis from CTP Dynamic Images Using a Learned LSTM Generative Adversarial Network for Acute Ischemic Stroke Assessment. *Journal of Medical Systems* 48, 1 (2024), 37. https://doi.org/10.1007/s10916-024-02054-2

[38] Mohsen Soltanpour, Russell Greiner, Pierre Boulanger, and Brian Buck. 2019. Ischemic Stroke Lesion Prediction in CT Perfusion Scans Using Multiple Parallel U-Nets Following by a Pixel-Level Classifier. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. 957–963. https://doi.org/10.1109/BIBE.2019.00179

[39] Mohsen Soltanpour, Russ Greiner, Pierre Boulanger, and Brian Buck. 2021. Improvement of automatic ischemic stroke lesion segmentation in CT perfusion maps using a learned deep neural network. *Computers in Biology and Medicine* 137, June (2021), 104849. https://doi.org/10.1016/j.compbiomed.2021.104849

[40] Tao Song. 2018. 3D Multi-scale U-Net with Atrous Convolution for Ischemic Stroke Lesion Segmentation. *International MICCAI Brainlesion Workshop* 40, 4 (2018), 4461.

[41] Tao Song. 2019. Generative Model-Based Ischemic Stroke Lesion Segmentation. arXiv:1906.02392 https://arxiv.org/abs/1906.02392

[42] Alzbeta Tureckova and Antonio J Rodríguez-Sánchez. 2019. ISLES Challenge: U-Shaped Convolution Neural Network with Dilated Convolution for 3D Stroke Lesion Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer International Publishing, Cham, 319–327.

[43] Guotai Wang, Tao Song, Qiang Dong, Mei Cui, Ning Huang, and Shaoting Zhang. 2020. Automatic ischemic stroke lesion segmentation from computed tomography perfusion images by image synthesis and attention-based deep neural networks. *Medical Image Analysis* 65 (2020), 101787. https://doi.org/10.1016/j.media.2020.101787

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7794–7803. https://doi.org/10.1109/CVPR.2018.00813

[45] Freda Werdiger, Mark W Parsons, Milanka Visser, Christopher Levi, Neil Spratt, Tim Kleinig, Longting Lin, and Andrew Bivard. 2023. Machine learning segmentation of core and penumbra from acute stroke CT perfusion data. *Frontiers in Neurology* 14 (2023). https://doi.org/10.3389/fneur.2023.1098562

[46] S. Winzeck, S. J.T. Mocking, R. Bezerra, M. J.R.J. Bouts, E. C. Mcintosh, I. Diwan, P. Garg, A. Chutinet, W. T. Kimberly, W. A. Copen, P. W. Schaefer, H. Ay, A. B. Singhal, K. Kamnitsas, B. Glocker, A. G. Sorensen, and O. Wu. 2019. Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. *American Journal of Neuroradiology* 40, 6 (2019), 938–945. https://doi.org/10.3174/ajnr.A6077

[47] Hao-Yu Yang. 2019. Volumetric Adversarial Training for Ischemic Stroke Lesion Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum (Eds.). Springer International Publishing, Cham, 343–351.

[48] Tongxue Zhou, Su Ruan, and Stéphane Canu. 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* 3-4 (2019), 100004. https://doi.org/10.1016/j.array.2019.100004

[49] Tongxue Zhou, Su Ruan, Yu Guo, and Stéphane Canu. 2020. A Multi-Modality Fusion Network Based on Attention Mechanism for Brain Tumor Segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 377–380. https://doi.org/10.1109/ISBI45749.2020.9098392

[50] Haichen Zhu, Yang Chen, Tianyu Tang, Gao Ma, Jiaying Zhou, Jiulou Zhang, Shanshan Lu, Feiyun Wu, Limin Luo, Sheng Liu, Shenghong Ju, and Haibin Shi. 2022. ISP-Net: Fusing features to predict ischemic stroke infarct core on CT perfusion maps. *Computer Methods and Programs in Biomedicine* 215 (2022), 106630. https://doi.org/10.1016/j.cmpb.2022.106630

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009