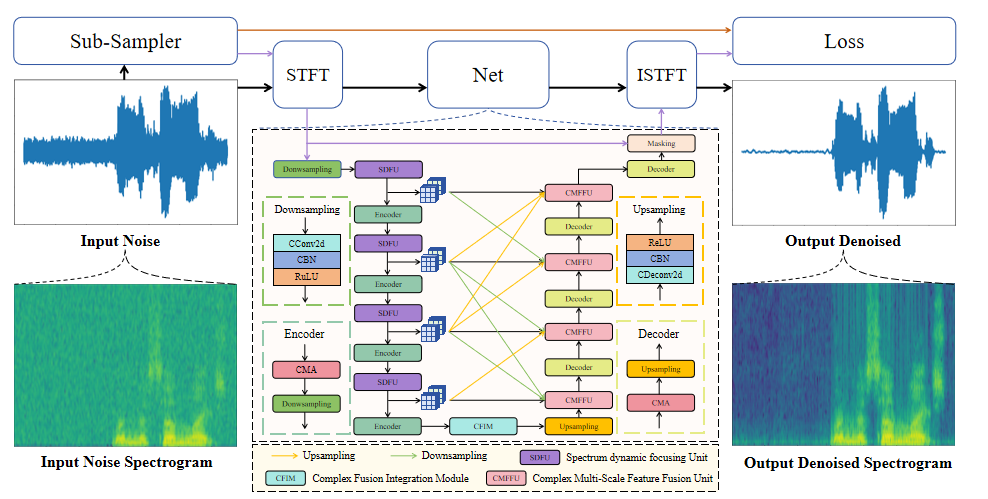
DMFNet: A self-supervised Dynamic Multi-Focusing Network for speech denoising without clean data

Given a noised speech through the sub-sampler to get two independent sub-audio and , where through the Short-Time Fourier Transform(STFT) to get as the input of the training network. And network outputs and are used as inputs to the loss function.

In this section, we first depict the overall architecture of the proposed Dynamic Multi-Focusing Network (DMFNet). Then, we introduce each module in the net in detail. Finally, we introduce related loss functions for self-supervised DMFNet training.



**Fig. 1.**

The presented speech denoising method does not rely on clean data, but focuses on self-supervision through the proposed Dynamic Multi-focusing Network (DMFNet). DMFNet is a U-shaped network divided into two main phases: encoding and decoding, connected by a Complex Fusion Integration Module (CFIM). In the encoding phase, the network dynamically extracts local and global features at different scales. The decoding phase is responsible for combining these features and reconstructing the spectral features.

1. **Overview**

As shown in Fig. 1, the proposed DMFNet is a U-shaped symmetrical hierarchical network

with two stages: encoding stage and decoding stage connected by Complex Fusion Integration Module (CFIM). In these two stages, the encoding stage contains 4 encoders designed to extract local and global features at different scales. The Spectrum Dynamic Focusing Unit (SDFU) was used before each encoder, to dynamically focus on the target spectral region of the input speech (the vocal spectral region). In another stage, the decoding stage contains 4 Decoders designed for feature fusion and spectrum reconstruction. Meanwhile, multi-scale connections are used before each encoder in the decoding stage to achieve dynamic focused full-feature aggregation.

1. Encoding Stage:

As we mentioned above, the encoding stage is designed for dynamical focus on the target spectral region and feature extraction at different scales. Initially, the network input , acquired through Short-Time Fourier Transform (STFT) from the sub-audio , undergoes downsampling. This process is essential for the efficient extraction of shallow features in both the real and imaginary components. Subsequently, the extracted features undergo processing via four encoders. Significantly, the deployment of Spectral Dynamic Focusing Units (SDFUs) preceding each encoder merits attention. This configuration facilitates dynamic and centralized acquisition of sophisticated data representations, specifically within acoustic spectral regions, under the framework of self-supervised learning. And each encoder comprises a uniquely designed Complex Multi-Scale Attention module (CMA) coupled with a downsampling block.

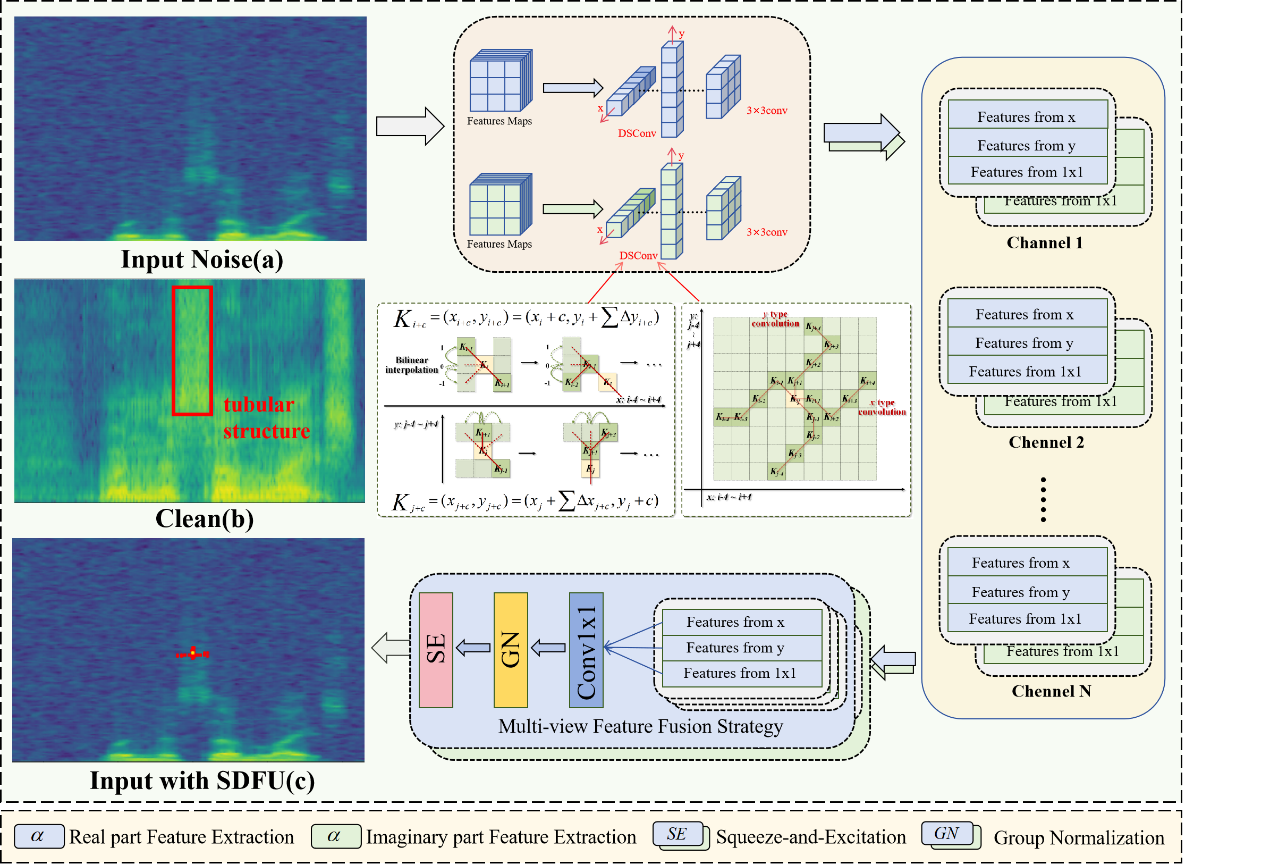
The downsampling block consists of Complex Convolution layers, Complex Batch normalization layer (CBN) and complex ReLU activation function (CReLU). The complex convolution layer employs a complex convolution filter . Given a complex vector , the complex convolution can be achievable through two independent real convolutions, expressed as . The architecture utilizes two real convolutions with a kernel size of 3×3 and a step size of 2, effectively reducing dimensionality while concurrently applying complex convolution for feature information extraction. The Complex Batch Normalization (CBN) and Complex Rectified Linear Unit (CReLU) have been effectively adapted for the complex domain. Specifically, the application of batch normalization (bn) and rectified linear unit (ReLU) operations on both the real and imaginary components has been demonstrated to yield superior outcomes among numerous proposed approaches[[1]](#endnote-1).

1. Decoding stage

In the decoding stage, we focus on feature utilization and aim to reconstruct local and global features on the spectrum. To achieve this, we introduce a new module called Complex Multiscale Feature Fusion Unit (CMFFU), which is applied before each decoder. Each decoder is composed of a Complex Mixture Attention (CMA) mechanism coupled with an upsampling block, optimizing feature integration and reconstruction. The upsampling module is composed of a complex inverse convolution layer, followed by a complex batch normalization layer and a complex ReLU activation function. Mirroring the downsampling approach, the complex transposed convolution layer is realized through two separate real inverse convolutions. These inverse convolutions employ 3×3 kernels with a stride of 2, facilitating both feature information extraction and simultaneous enlargement of the feature dimension.

1. Spectrum dynamic focusing Unit (SDFU)

The conventional 2D convolution kernel, characterized by its fixed geometry, encounters limitations in accurately discerning the intricate details inherent in complex spectral data. To augment its precision in detecting localized tubular structure features within the human voice spectrum, as illustrated in the unblemished depiction in Fig. 2, we introduce the Spectrum dynamic focusing Unit (SDFU). This unit fuses the traditional 2D convolution with DSConv to discern complex tubular structures. DSConv represents a novel convolutional approach, selectively concentrating on geometrically localized attributes of tubular structures, a technique has been demonstrated to amplify the discernment of intricate tubular formations[[2]](#endnote-2)[[3]](#endnote-3).



**Fig. 2.**

The architecture of the proposed Spectrum dynamic focusing Unit (SDFU). The figure (c) demonstrates the SDFU's capability to dynamically capture the spectrum of the human voice. The illustration includes yellow dots representing the convolution kernels and 360 red dots marking the locations of the learning samples.

Fig. 2 of input with SDFU reveals the convolution kernel's placement and configuration (the convolution kernel is indicated in yellow, while 360 red dots represent the learnt sampling locations). The visualization results show that the convolution kernel within SDFU is capable of dynamically conforming to the tubular structure of the human voice spectrum, ensuring precise alignment with the intended target. The specific implementation of SDFU is as follows:

For a 3×3 convolution kernel with its center coordinates, the standard 2D convolution kernel is denoted as:

Inspired by[[4]](#endnote-4), DSConv introduces a learnable deformation offset into the standard 2D convolutional kernel. in the direction of the x-axis becomes:

and in the direction of the y-axis becomes:

where since the learned offsets and are usually not integers, we consider the sampling bilinear interpolation method, denoted as:

where denotes the fractional position of equation ()(), represents all positions in integer space, is the bilinear interpolation kernel, which can be split into two one-dimensional linear interpolation kernels:

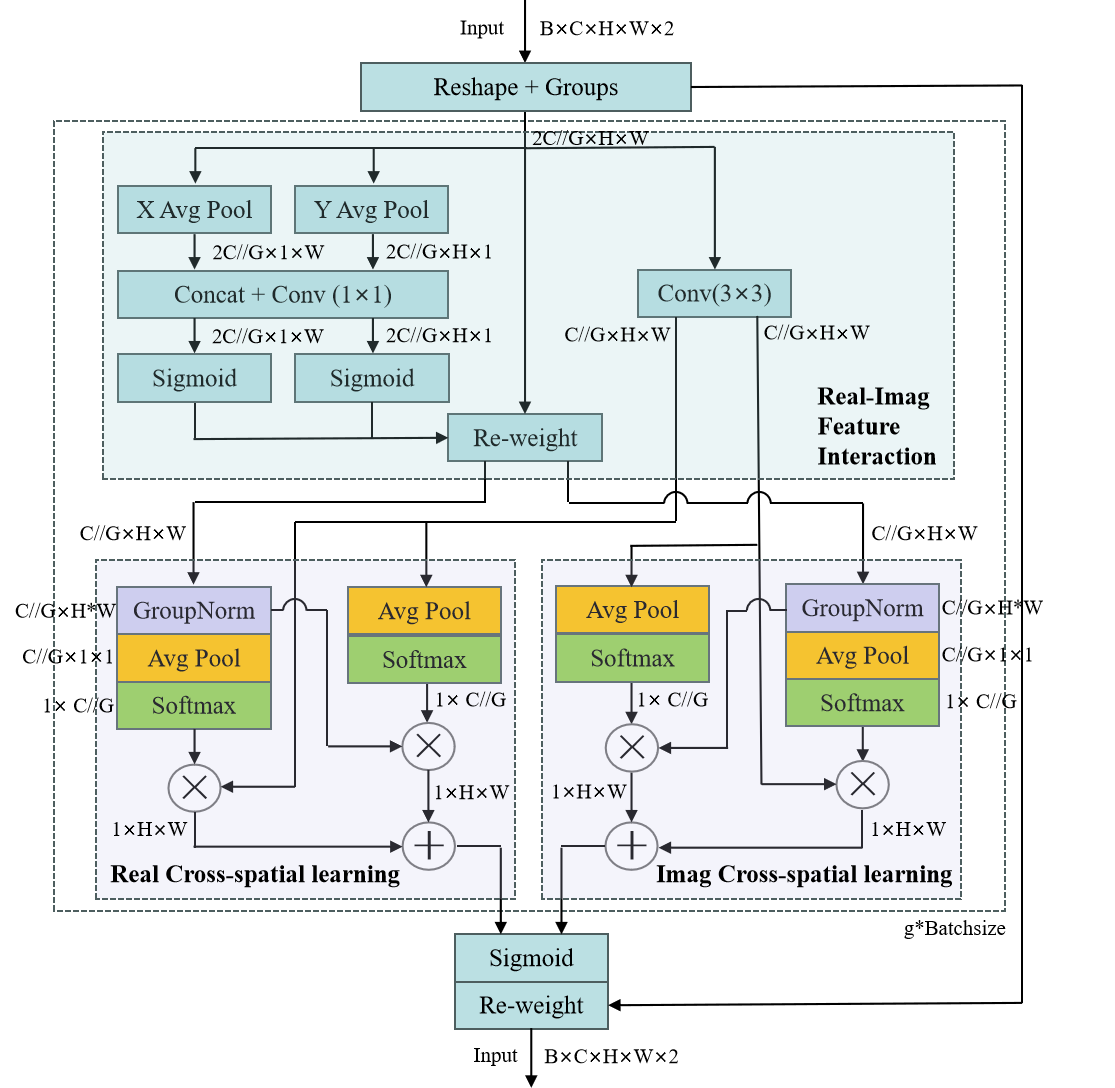
where b denotes the one-dimensional linear interpolation kernel. Following this mathematical exposition, the traditional 2D convolution is employed to apprehend the global spectral characteristics of the speech spectrum. Conversely, the DSConv can adaptively adjust the convolution kernel according to the shape of the feature map in the x-axis and y-axis directions. This ability enables it to more precisely discern the localized tubular structural features of the speech spectrum. The features extracted by DSConv and the conventional 2D convolution are concatenated along the channel dimensions , followed by the application of 1×1 convolution to yield initial fusion outcomes subsequent to a group normalization process. Finally , each channel is endowed with a channel direction attention weight via the Squeeze-and-Excitation mechanism[[5]](#endnote-5). The above steps can be expressed for input :

where and denote DSConv in the x and y directions, respectively. denotes a standard 2d convolution with a kennel size of 3. denotes a standard 2d convolution with a kennelsize of 1 and an outchannnel=inchannel //3 for the standard 2d convolution. denotes the group normalization process and SE denotes the attention mechanism of Squeeze-and-Excitation.

The Spectral Dynamic Focusing Unit (SDFU) synergistically combines Depth Separable Convolution (DSConv) and conventional 2D convolution. This fusion permits adaptive modification of the convolution kernel structure in response to varying spectral characteristics. Consequently, it enhances the precision in capturing localized tubular structures within the human voice spectrum, facilitating an efficient amalgamation of both global and local features.

1. **Complex multi-scale attention (CMA) module**

CMA is one of the most important modules in DMFNet, which is designed for local and global feature extraction and information interaction between real and imaginary parts. As shown in Fig. 3, CMA is mainly composed of complex feature interaction unit and cross-space learning unit, which are used for feature interaction and feature extraction respectively.



**Fig. 3.**

The architecture of Complex multi-scale attention(CMA) module.

1. Feature Grouping:

For any given spectral input  ,the last dimension of  represents the real and imaginary parts of the complex input, respectively. CMA initially splits  along the last dimension, followed by concatenation along the channel dimension. Subsequently, CMA divides  into G\*Batchsize sub-features along the channel dimension, facilitating the learning of diverse semantic representations, where the groups-style can be donated by:

，.

Without losing generality, we let  and assumed that the learnt attention weight descriptors will be utilized to strength the feature representation of interest region in each sub-feature.

2）Complex Feature Interaction Unit(CFIU):

In the field of speech enhancement (SE), a primary challenge lies in efficiently leveraging both the magnitude and phase information of noisy speech for clear signal reconstruction. To address this, we introduce the Complex Feature Interaction Unit. This unit is designed to maximize the extraction and interaction of valuable information from both the real and imaginary components of the speech signal. As shown in Fig. 3, the CFIU mainly consists of three parallel substructures, including two parallel instances of the shared components with 1x1 branch and a single 3x3 branch.

In the shared component of the 1x1 branch, we employ a structure similar to the 1D feature encoding vector found in the Coordinate Attention (CA) module[[[6]](#endnote-6)],which can capture position information along the horizontal and vertical dimensions of the real and imaginary parts. These vectors are obtained by global average pooling of the corresponding dimensions of the input tensor. These two 1D feature-encoded vectors are then concatenated and processed through a shared 1x1 convolutional layer with dimensionality reduction. This convolution is designed to capture local cross-channel interactions. The 1x1 convolution kernel has similarities to the channel convolution. After convolution the output passes through a nonlinear sigmoid function in each parallel path. Attentional weights learned from these parallel paths are used to reweight the original intermediate feature maps to produce the final output.

On the other hand, the 3x3 branch enhances the feature space by capturing cross-channel interactions of local real and imaginary data through 3x3 convolution. This approach allows the Enhanced Modular Architecture (EMA) to not only encode inter-channel information, thereby adjusting the importance of different channels post real-imaginary concatenation, but also to maintain precise spatial structural information within the channel.

3）Cross-spatial learning unit(CSLU)：

Channel-Spatial Learning unit (CSLU) exploits the interconnectedness between channels and spatial dimensions, a principle widely explored and applied in modern computer vision tasks[[[7]](#endnote-7)][[[8]](#endnote-8)].

Utilizing both short- and long-range dependencies among channels and spatial locations derived from outputs of 1x1 branchand 3x3 branchof the CFIU，CSLU demonstrates advanced proficiency in handling of complex spatial structures.

CSLU splits the input along the channel dimensions and directing it separately into the real Cross-spatial learning unit and imaginary Cross-spatial learning unit. The output from the 1x1 branch concentrates on localized interactions between the real and imaginary parts. Subsequently, we employ the 2D Global Average Pooling（GAP）to encode the global spatial information in the output of the 1x1 branch  and the 3x3 branch .The 2D global pooling operation is formulated as:



Which is designed for encoding real and imaginary part global information and modeling the long-range dependencies. And then CSLU applies nonlinear functions on the output of 2D GAP to fit the upon linear transformations.

In real Cross-spatial learning unit, we utilize the 2D GAP to encode global spatial information of the real part in the outputs of 1x1 branch, the outputs of the least branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e.,. Combining the outputs of the above parallel processing, CSLU generates spatial attention maps through matrix dot product operations, aiming to collect spatial information at different scales. Similarly, CSLU generates a second set of spatial attention maps，and aggregates the generated spatial attention weights by a Sigmoid function to produce the final output feature maps. This process ensures that the feature map incorporates cross-space dependencies between real and imaginary parts while maintaining the accuracy of spatial information.

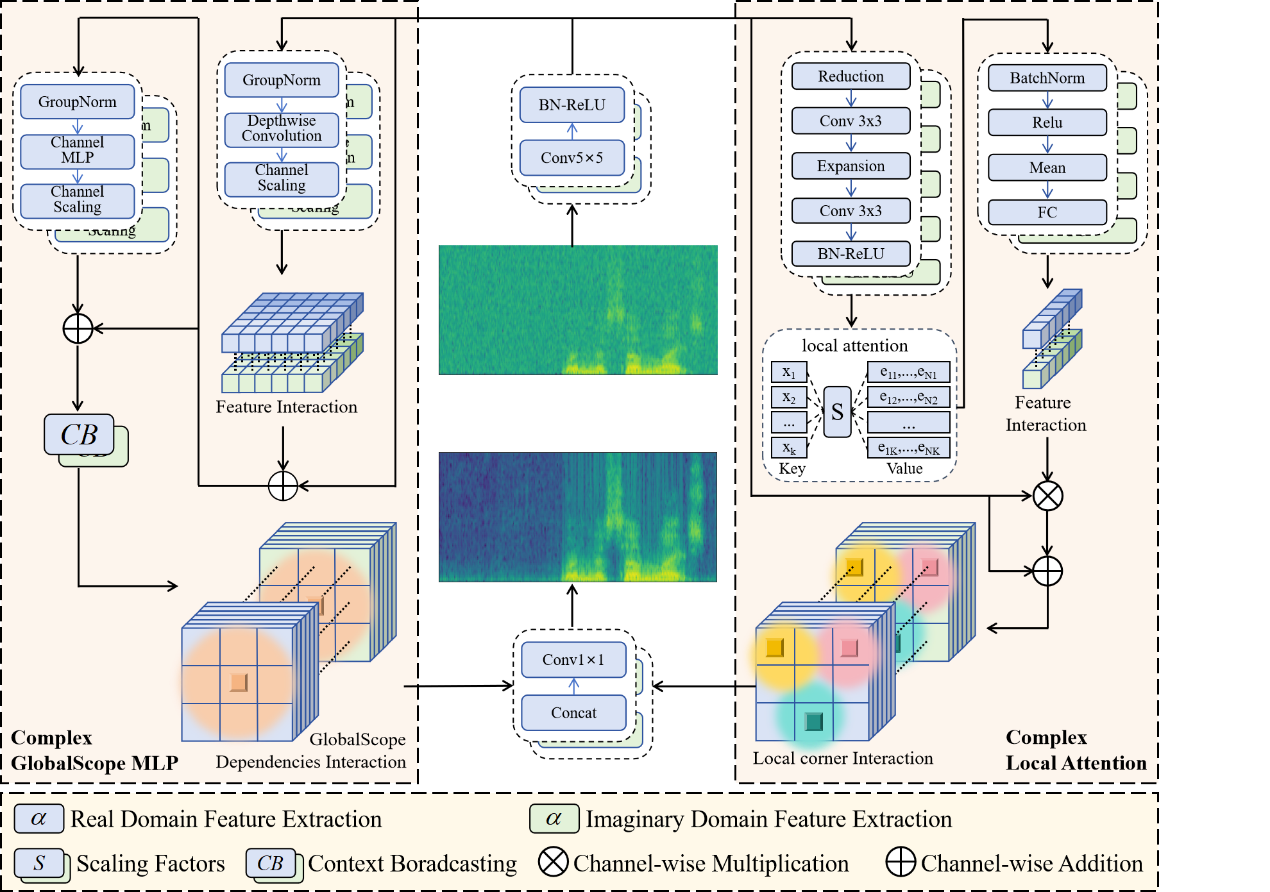
Via its cross-space information integration technique and parallel substructures, CSLU significantly enhances the network's proficiency ability to handle intricate spatial configurations, specially its ability to capture and interact with pixel-level details and global context in both the real and imaginary domains.

1. Complex Fusion Integration Module (CFIM)

Illustrated in Fig. 4, the Complex Feature Interaction Module (CFIM) primarily encompasses two units connected in parallel. The Complex Global Scope Multi-Layer Perceptron block (CGMLP) is designed to capture and integrate the Global Scope long-range dependencies of the real and imaginary parts in the uppermost feature . Simultaneously, for the aggregation of local area features within the real and imaginary layers, the complex local attention unit, deployed on , is introduced. The result feature maps from these two units are concatenate together along the channel dimension, forming CFIM's output for subsequent feature extraction. Between and the CFIM, a Stem block is utilized for feature smoothing, comprising two 5×5 convolution with 90 output channels for the real and imaginary part, followed by batch normalization layers and activation function layers. This process is encapsulated as follows:

Where denotes the output of CFIM, and represent the output features of the Complex GlobalScope Multi-Layer Perceptron and the Complex Local Attention mechanism, respectively. is the output of the Stem block, which is obtained by

Where and indicate a 5×5 convolution operation with stride 1 and the channel size is configured to 90 for both the real and imaginary parts. represents a batch normalization layer and denotes the ReLU activation function. Finally, the real and imaginary results are concatenated along the last dimension.



**Fig. 4.**

The architecture of Complex Fusion Integration Module (CFIM). The Complex GlobalScope Multi-Layer Perceptron Unit (CGMLP) primarily addresses interactions on a global scope, whereas the Complex Local Attention Unit(CLA) focuses on local corner interactions.

1. The Complex GlobalScope Multi-Layer Perceptron Unit(CGMLP)

Our proposed The Complex GlobalScope Multi-Layer Perceptron Unit(CGMLP) architecture primarily comprises three key components: a depthwise convolution residual module, a channel MLP residual module, and a Context Broadcasting (CB) module.

**Depthwise Convolution Residual Module.** Post-smoothing in the Stem block, the features output is directed into depthwise convolution for both real and imaginary parts, followed by group normalizations (i.e., the real and imaginary feature maps are grouped along the channel dimension). Contrasted with the conventional spatial convolution for s Spectral feature extraction, the depthwise convolution enhances the spectral feature representation ability while reducing the computational costs. Subsequently, channel scaling operation is employed to obolster feature generalization and robustness[[9]](#endnote-9).The above processes, performed independently in both the real and imaginary parts, can be formulated as:

Where represents the output of Depthwise Convolution Residual Module. is the group normalization and is a depthwise convolution with the kernel size of 1×1.The denotes the channel scaling operation.

**Feature Interaction Module.** Following the channel scaling operation, we concatenate the real features maps and the imaginary features maps along the channel dimension. Subsequently, a 1×1 convolution is employed to facilitate the learning of compressed latent space representations of parameter distributions, enhancing cross-channel interaction[[10]](#endnote-10)[[11]](#endnote-11). These processes can be formulated as:

Where denotes a concatenation along the channel dimension, and denotes a 1×1 convolution.

**Channel MLP Residual Module.** Features from the Feature Interaction Unit, denoted as , are initially subjected to group normalization, followed by the application of the Channel Multi-Layer Perceptron (MLP)[[12]](#endnote-12). The Channel MLP can effectively learn the complex features of data through its multi-layer structure, concurrently reducing computational complexity[[13]](#endnote-13)[[14]](#endnote-14). Following this, a channel scaling operation is executed to enhance the generalization and robustness of features. The above processes are expressed as:

Where represents the output of Channel MLP Residual Module. is the group normalization and is the channel MLP. The denotes the channel scaling operation, and denotes the output of Feature Interaction Module.

**Context Broadcasting Module (CB).** Fig. 4 illustrates our Context Broadcasting Module (CB) module. The CB module is positioned at the end of the Complex Long-range MLP block. See Fig. 4 for the overall CFIM architectures with our CB module. The CFIM is inclined towards dense interactions encompassing both real and imaginary parts, factoring in both long-range and short-range dependencies. However, due to the steep gradient of the softmax function, learning dense attention becomes more challenging. In contrast, the integration of the CB module mitigates the density in the original attention graph, thereby enhancing the capacity and generalization potential of the CFIM. Specifically, given a input ,our CB module supplies the average-pooled operation onto the channel features as follows:

Where is output of the Complex Long-range MLP block.denotes a average pooling operation on channel dimensions.

1. The Complex Local Attention Unit(CLA)

The Complex Local Attention Unit (CLA) operates as an encoder with an inherent dictionary mechanism. This encoder is composed of two primary elements: Local attention unit and Feature Interaction Unit.

**Local attention unit**. As shown in Fig. 4, the input feature initially undergoes processing in a combinatorial module, which encompasses two 1x1 convolutions and 3x3 convolutions. This module serves multiple purposes: compression of features, capturing spatial relationships, and expansion of channel information. The resultant transformed features are further refined through a CBR (Convolution-Batch Normalization-ReLU) module, which integrates a 3x3 convolutions, batch normalization, and ReLU activation. This integration is pivotal in enhancing feature characterization while preserving nonlinearity. Subsequently, the encoded features, denoted as are fed into the local attention unit. A set of scaling factor is utilized to align and map the corresponding position information. The comprehensive information of the entire features in relation to the k-th codeword is computed as follows:

Where is the update to the kth codebook vector. is a set of scaling factor. and are the real and imaginary parts of the i-th input vector, respectively. and are the real and imaginary parts of the k-th codebook vector, respectively. We use features fuse unit to fuse all and highlight key classes, which encompasses a batch Normalization (BN) layer, ReLU activation function, a mean layer and a fully connection layer. Building upon this, the full information of the whole *K* codewords is calculated as follows.

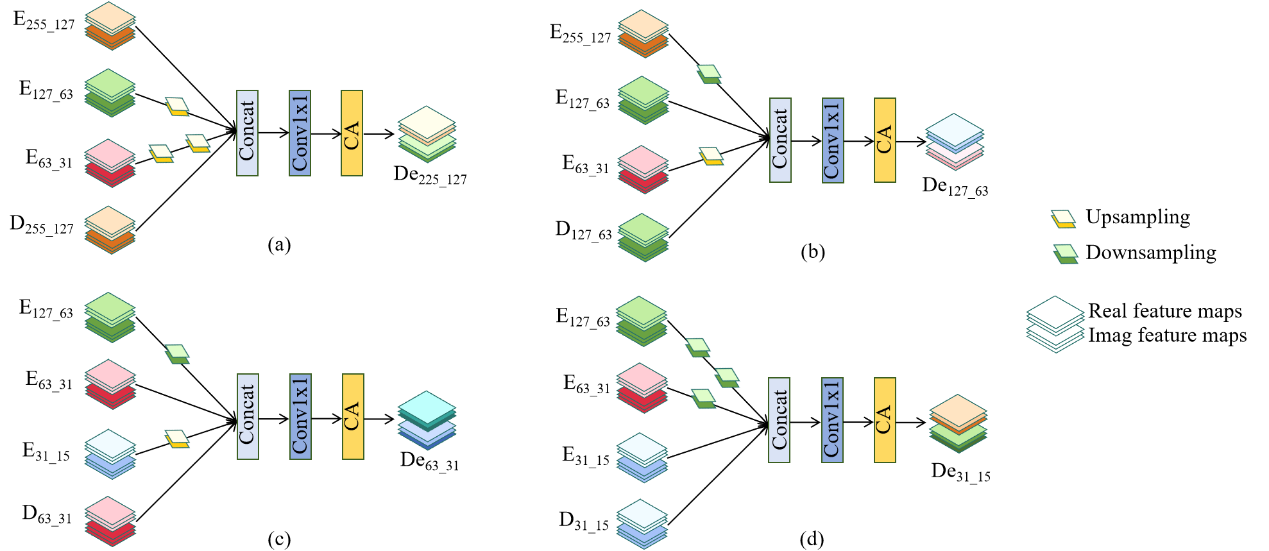
Where denotes BN layer, ReLU activation function and mean layer, and represents the fully connection layer.

**Feature Interaction Unit.** Upon acquiring the output from the features fuse unit, we introduce a feature interaction unit designed to facilitate the interaction between real and imaginary local features. Subsequently, a channel-wise multiplication is executed between the input features from the Stem block, denoted as and the scaling factor coefficient δ (·). Finally, we perform a channel-wise addition between features from the Stem block and the local attention features. The above processes are expressed as:

Where denotes the sigmoid function. is channel-wise multiplication and is the channel-wise addition.

1. Complex Multi-Scale Feature Fusion Unit (CMFFU)

To make full use of the multi-scale Properties of both the real and imaginary components, which are elicited from the spectrogram of an audio signal after short-time Fourier transform (STFT), our approach employs a multi-scale feature fusion strategy within the decoding stage of the network architecture. This strategy is specifically designed to enhance the network's ability in feature propagation and representation, capitalizing on the critical information extracted from the real and imaginary parts during the encoding stage. This comprehensive approach guarantees a thorough exploitation of the spectral data, essential for self-supervised speech denoising.



**Fig. 5.**

The architecture of Complex Multi-Scale Feature Fusion Unit (CMFFU)

Specifically, our main goal is to explore and exploit the real and imaginary features from the encoding stage during the decoding process. The disparate sizes of these features present a significant integration challenge. For instance, with an input image size of 512×256×2,the feature map sizes in the decoding stage are 255×127×2, 127×63×2, 63×31×2, and 31×15×2, respectively. In contrast, the feature map sizes in the decoding stage follow the sequence of 31×15×2, 63×31×2, 127×63×2, and 255×127×2.To address this, we introduce the Complex Multi-scale Feature Fusion Unit (CMFFU), detailed in Fig. 5. As illustrated, each feature map in the decoding stage is merged with the three feature maps that are most similar in size. This process involves up-sampling and down-sampling to adjust the feature maps to comparable sizes, notably using distinct sampling layers for the real and imaginary components. Following the standardization of feature map sizes, we concatenate these maps along the channel dimension. Subsequently, a 1 × 1 convolutional layer is applied to achieve initial fusion outcomes. Ultimately, we allocate a channel-specific attention weight to each channel through the CA mechanism.

1. Loss function

1. Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep

   complex networks. In International Conference on Learning Representations, 2018. [↑](#endnote-ref-1)
2. Liu Q, Liu Y, Lin D. Revolutionizing Target Detection in Intelligent Traffic Systems: YOLOv8-SnakeVision[J]. Electronics, 2023, 12(24): 4970. [↑](#endnote-ref-2)
3. Qi Y, He Y, Qi X, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6070-6079. [↑](#endnote-ref-3)
4. Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773. [↑](#endnote-ref-4)
5. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141. [↑](#endnote-ref-5)
6. [] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, "Image super-resolution using very deep residual channel attention networks", Proc. Eur. Conf. Comput. Vis., pp. 286-301, 2018. [↑](#endnote-ref-6)
7. [] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, “Aˆ 2-nets: Double attention networks,”Advances in neural information processing systems, vol. 31,

   2018. [↑](#endnote-ref-7)
8. [] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, “Non-local neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803. [↑](#endnote-ref-8)
9. W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan,“Metaformer is actually what you need for vision,” arXiv, 2021. [↑](#endnote-ref-9)
10. Maksoud S, Zhao K, Peng C, et al. Scalable Bayesian Deep Learning with Kernel Seed Networks[J]. arXiv preprint arXiv:2104.09005, 2021. [↑](#endnote-ref-10)
11. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11531-11539, doi: 10.1109/CVPR42600.2020.01155. [↑](#endnote-ref-11)
12. Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in neural information processing systems, 2021, 34: 24261-24272. [↑](#endnote-ref-12)
13. L. Nosrati, M. S. Fazel and M. Ghavami, "Improving Indoor Localization Using Mobile UWB Sensor and Deep Neural Networks," in IEEE Access, vol. 10, pp. 20420-20431, 2022, doi: 10.1109/ACCESS.2022.3151436. [↑](#endnote-ref-13)
14. Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series

    in 2021,” arXiv, 2021. [↑](#endnote-ref-14)