Complex multi-scale attention(CMA) module

As one of the most important modules in xx net. CMA is designed for local and global feature extraction and interaction from real and imaginary parts. As shown in Fig.1 , CMA mainly consists of a Complex Feature Interaction unit and Cross-spatial learning unit ,which are used for feature interaction and feature extraction, respectively.

1. Feature Grouping:

For any given spectral input  ,the last dimension of  contains information about the real and imaginary parts, respectively. CMA will first split  from the last dimension and concatenate along the channel dimension. Then CMA will divide  into G\*Batchsize sub-features across the channel dimensions direction for learning different semantics, 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(CFI):

In speech enhancement(SE),the main challenge is how to effectively utilize the spectral magnitude and phase information of noisy speech while reconstructing the clear speech signal.To achieve this,we propose the Complex Feature Interaction Unit to make our model extract and interact as much as possible useful information from both the real and imaginary parts of the speech signal.As shown in Fig.X, the xxx unit mainly consists of three parallel substructures, including two parallel instances of the shared components with 1x1 branch and a singular 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[[[1]](#endnote-1)],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, 3x3 branch expands the feature space by capturing the cross-channel interaction of real and imaginary local information through 3x3 convolution. In this way, the EMA not only encodes the inter-channel information to adjust the importance of different channels after real-imag concating, but also preserves the precise spatial structure information into the channel.

1. Cross-spatial learning unit(CSL)：

Channel-Spatial Learning (CSL) capitalizes on the interdependence of channels and spatial locations, a concept extensively researched and implemented in contemporary computer vision tasks.[[[2]](#endnote-2)][[[3]](#endnote-3)]。Utilizing both short- and long-range dependencies among channels and spatial locations derived from outputs of 1x1 branchand 3x3 branchof the CFI，CSL demonstrates advanced proficiency in handling of complex spatial structures。

CSL splits the input along the channel dimensions and directing it separately into the real Cross-spatial learning unit and imag 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 CSL 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, CSL generates spatial attention maps through matrix dot product operations, aiming to collect spatial information at different scales.Similarly, CSL 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, CSL 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. [] 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-1)
2. [] 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-2)
3. [] 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-3)