### Complex Multi-Scale Feature Fusion Unit

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 superior audio signal processing

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 FigX. 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.

Considering the size of the feature maps, the fusion scheme can be categorized into four situations. The schematic diagram of how MFFU aggregates features from different scales is shown in Fig 5. To facilitate comprehension, only the formulation pertaining to Figure 5(b) is provided. The formulation for Figure 5(b) is delineated as follows:



 

where (k1\_k2 = 255\_127,127\_63,63\_31) represents the feature maps from the previous three encoding stages, each with dimensions of k1×k2.  represents the initial feature maps of the current decoder with the size of 127×63 .  indicates implies a size transformation of the feature maps from k1×k2 to 127×63.  and refer to the 3×3 convolution operation with the stride to be 2, whereas andrepresent the 3×3 transposed convolution operation with stride to be 2. denotes the concatenation process along the channel dimension. embodies the ultimate fusion outcome.