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. 1, 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 [[[1]](#endnote-1)][[[2]](#endnote-2)]. Figure 1 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[[3]](#endnote-3), 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 SE mechanism[[4]](#endnote-4). 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 kennelsize 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[[5]](#endnote-5).

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. [] Liu Q, Liu Y, Lin D. Revolutionizing Target Detection in Intelligent Traffic Systems: YOLOv8-SnakeVision[J]. Electronics, 2023, 12(24): 4970. [↑](#endnote-ref-1)
2. [] 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-2)
3. 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-3)
4. 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-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)