Part1: Highlight the advantages of our approach.

Let's reiterate the advantages of our method below, hoping to receive recognition from all the reviewers.

(a) This research aims to **address the lack of theoretical guidance** in existing methods for setting up attention networks. Often, attention network modules are considered as **black box designs**. To **bridge this gap**, our method utilizes NMF theory to guide the design of attention networks. By doing so, we hope to receive recognition from all the reviewers.

(b) The vanilla attention network is **initially designed in 2D**. Although the channel-spatial attention, such as CBAM, combines both channel attention and spatial attention either in series or in parallel, it first applies channel attention to obtain corrected features and then applies spatial attention to obtain the corresponding output that essentially belongs to 2D attention. However, this research takes a different approach by starting with **a 3D attention** design that combines channel and spatial attention and then **apply correction directly to the input features** and investigates the transformation from 3D attention to 3D self-attention mechanism. **Taking inspiration from GCNet**, a simple softmax can be directly exploited to achieve this transformation. The current methods have not yet explored this aspect, which is another key focus of this paper.

(c) We has conducted extensive experiments on medial and natural image datasets, namely **Synapse, Cityscapes, ImageNet, and COCO datasets**. The results of these experiments demonstrate that our DSNet performs well in various tasks, including semantic segmentation, image classification, and object detection, when compared to advanced network models.

Part2: In order to improve the quality of the paper, we have made the following **adjustments to the original manuscript**.

Revision1. We have added the following description **near line 61** of our main manuscript:

The extracted features for an image are denoted by $\mathrm{F} \in \mathbb{R}^{C\times H\times W}$. In a similar fashion to how NMF corrects the base matrix pixel-by-pixel, the attention mechanism should also be organized to perform pixel-wise feature correction, ultimately generating a 3D attention $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$. In order to capture long-distance dependencies, it is necessary to obtain a 3D shape of self-attention, referred to as 3D self-attention $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$, where the weights sum up to 1. The correction operation of 3D self-attention on the input involves two main steps: feature weighting and feature aggregation.

Revision2. We **merge** subsection "**Attention mechanism**" and subsection "**Self-attention mechanism**" in related work in Section 2.

Attention mechanism and Self-attention mechanism In general, attention modeling methods fall into three categories: channel attention, spatial attention, and hybrid attention (i.e., channel-spatial attention). Channel attention [5] captures information on feature maps at the channel level, while spatial attention [26, 27, 28] emphasizes critical spatial information in feature maps. Channel-spatial attention [6, 22, 29, 30] combines channel attention and spatial attention either in series or in parallel and thus possesses the inherent capability to encompass a broader context compared to channel-wise or spatial-wise attention. The self-attention mechanism is a type of attention mechanism that has earned considerable interest in the realm of computer vision [31]. To reduce high computational expense of the Non-Local Network (NLNet) [7], simplified variants of NLNet, such as EANet[32], GCNet[14], CCNet[33], Axial-DeepLab[34], are subsequently proposed. Among them, GCNet [14] finds that the global context modeled by NLNet is almost the same for various query positions within one feature map. Although the channel-spatial attention combines both channel attention and spatial attention either in series or in parallel, it first applies channel attention to obtain corrected features and then applies spatial attention to obtain the corresponding output that essentially belongs to 2D attention. This paper focuses on building 3D attention that combines channel and spatial attention and then applies correction directly to the input features. To reduce the computational load, the 3D attention is initially decoupled into spatial and channel attention and then recoupled back into its original 3D shape. Also, we exploit the observation that the self-attention for various query points is shared. And we convert the 3D attention to 3D self-attention by simply applying a softmax function to the original 3D shape, thereby further reducing computational workload.

Revision3. We have **clarified the difference** between "**channel-spatial attention**" and our proposed "**3D attention**" in related work in Section 2.
Although the channel-spatial attention, such as CBAM, combines both channel attention and spatial attention either in series or in parallel, it first applies channel attention to obtain corrected features and then applies spatial attention to obtain the corresponding output that essentially belongs to 2D attention. However, this research takes a different approach by starting with a 3D attention design that combines channel and spatial attention and then apply correction directly to the input features.

Revision4. To provide a comprehensive **overview of previous work**, we have incorporated a new subsection at the end of the related work.
Attention mechanism in computer vision tasks The attention mechanism has become an integral component in computer vision tasks, with various variants being proposed. ECANet [41] introduces a local cross-channel interaction strategy that avoids dimensionality reduction by examining the adverse effects of the squeeze step in SENet. PSA [23] devises a polarized self-attention mechanism that effectively addresses the pixel-level regression task through polarized filtering and enhancement. FcaNet [46] utilizes frequency analysis to reparse channel attention and has demonstrated that Global Average Pooling (GAP) is a specific instance of feature decomposition in the frequency domain. Building upon this discovery, a multispectral channel attention approach is proposed. To extract the contextual information of the semantic segmentation model such as FCN [39], EncNet [40] introduces a new context encoding module that selectively enhances class-related feature maps and captures the semantic context. Additionally, UperNet [42] proposes a hierarchical network with a feature pyramid network to obtain an effective global prior representation. HamNet [17] utilizes strategies such as NMF [18] to decompose large matrices into low-rank

matrices and suggests that the low-rank embedding of input feature representations contains global information on self-attention. To promote the application of transformer [24] in semantic segmentation, Swin Transformer [45] introduces an inductive bias in CNN and proposes a shifted window-based design to gradually expand the receptive field. Twins-SVT [44] develops a spatially separable self-attention mechanism that can combine local and global attention. To reduce computational complexity, local attention is initially computed and grouped in the spatial dimension and then fused to obtain global attention. However, existing ADNN models currently lack theoretical guidance in the design of attention networks, which can be deemed a black box design. In contrast, this research aims to address this gap by integrating NMF into the design of attention mechanisms. Finally, we design the DSNet to conduct image classification, object detection, and semantic segmentation. And our proposed DSNet achieves comparable performance with almost incurring no additional computational costs.

Revision5. In order to justify our claims, we have **added** the following explanation in **section 3.1 on line 140**.

Ionescu *et al.* [12] rephrase the optimization process in DNNs as a recursive calculation of gradients for input layers and weight parameters. ADNN precisely adheres to this paradigm, and incorporating attention mechanisms to correct input is a more straightforward approach. Meanwhile, we find that the optimization process of NMF is similar to that of this ADNN, where NMF conducts the iterative correction of the base matrix and coefficient matrix [18]. As shown in Fig. 1, the optimization process of ADNN between any two consecutive layers is essentially a special form of solving NMF. Specifically, reweighting the input features through the attention in ADNN is equivalent to the correction of the base matrix in NMF.

Revision6. Furthermore, we have made **slight adjustments** to the references in the introduction section and the remaining subsections of the related work.

(a) Add reference [21] on line 23.

(b) Add references [44], [45] on line 44.

(c) Delete reference [21] on line 63.

(d) Add references PSA [23], ECANet [41], and FcaNet [46], etc. on line 65.

(e) Add references [41], [42] on line 68.

(f) Add references [43] [45] on line 81. “ConvNeXt [43] gradually transformed ResNet into a Transformer architecture, preserving the simplicity and efficiency of standard convolutions. This transformation ultimately resulted in surpassing the performance of the Swin Transformer[45].”

(g) Adjust the “**Theoretical foundation of the attention mechanism**” of the related work to: However, several works [9, 10, 11] have demonstrated a weak correlation between attention maps and features’ importance. Class activation mapping, such as Grad-CAM[35] is often employed to generate visual heatmaps that depict the underlying rationale for the network’s decision-making process. In addition, the strategy proposed by [36] utilizes deep Taylor decomposition to produce interpretability heatmaps. This paper argues that the attention mechanism primarily corrects input

features to obtain a more accurate feature representation, ultimately resulting in improved visualization of the class activation map. This paper points out that the attention mechanism essentially corrects its input features, which may be the fundamental reason for its effectiveness.
(h) The $\alpha_2 = \mathbf{H}/(\mathbf{N}^T\mathbf{HH})$ in L157 is modified to be $\alpha_2 = \mathbf{H}/(\mathbf{N}^T\mathbf{NH})$.
(i) L192 needs to be deleted, the **F** in L288 is modified to be **W**, and so on.

Revision7. We have reported the corresponding **experimental results** on the **synapse, Cityscapes, ImageNet, and COCO datasets**, as shown in the Section 5 "**Experiments**".
Note: Following the reviewer's comment, we focus on comparing the most advanced **FcaNet** (FcaNet: Frequency Channel Attention Networks). Due to the extensive training time of over twenty hours required for training a model on ImageNet, we prioritized evaluating the most advanced techniques. We will continue to supplement all experiments in the near future. Additionally, we plan to publish our latest codes.