Appendix 1 The explanation of the attention mechanisms

The attention mechanism (36) is a technique in DL that simulates the human attention process, aiming to enable models to "focus" on specific important parts of the input, thereby enhancing the model's performance and efficiency. Initially, the attention mechanism achieved significant success in the field of natural language processing (NLP), and it has since been widely adopted in various fields, including computer vision.

The explanation of the self-attention mechanisms

The core concept of the self-attention mechanism (37) is to generate a weighted representation for each element in a sequence. Specifically, given an input sequence, each element undergoes 3 distinct linear transformations to generate query (Q), key (K), and value (V) vectors. Subsequently, the attention weights are computed by assessing the similarity between the Q vector and the K vector. These weights are then used to perform a weighted sum of the value vectors, resulting in the final output representation. Specifically, given an input sequence $X = [X_1, X_2, \ldots, X_n]$, where X_i represents the *i-th* input vector in the sequence, each input vector is first transformed into the query vector Q, key vector K, and value vector V through linear transformations:

$$Q = XW^Q \quad [1]$$

$$K = XW^K \quad [2]$$

$$V = XW^V \quad [3]$$

where, W^Q, W^K, and W^V represent weight matrix. Therefore, the computation of self-attention is represented as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [4]$$

where, d_k represents the dimension of the K vector.

Appendix 2 The explanation of the deformable convolution

Standard convolution performs excellently for lesions with regular shapes due to the fixed size and shape of the convolutional kernel. In contrast, deformable convolution (15) introduces additional offset parameters for each element of the convolutional kernel, allowing the sampling points of the convolutional kernel to shift within the feature map, thereby concentrating on feature extraction in regions of interest. The characteristics of deformable convolution are better suited for extracting boundary features of lesions with complex shapes, whereas standard convolution is more effective for capturing deep features within the lesion.

As shown in *Figure S1*, deformable convolution (15) can adaptively modify the sampling positions of the convolution kernel by learning additional offsets.

As shown in *Figure S2*, compared to standard convolution, deformable convolution (15) adjusts its sampling locations to better conform to the lesion's contour, thereby effectively eliminating interference from background noise. The standard and deformable convolutions can be represented by Eqs. [5,6]:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \bullet x(p_0 + p_n) \quad [5]$$

$$y(p_0) = \sum_{p_n \in R} w(p_n) \bullet x(p_0 + p_n + \Delta p_n) \quad [6]$$

where, $w(p_n)$ represents the weight of the convolution kernel at position p_n, x represents the input feature map, and Δp represents the offset. $R = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}$ (defines a 3×3 convolution kernel with dilation 1)

References

36. Bahdanau D. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 2014. doi: arxiv-1409.0473.
37. Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems 2017.

https://dx.doi.org/10.21037/qims-24-1451

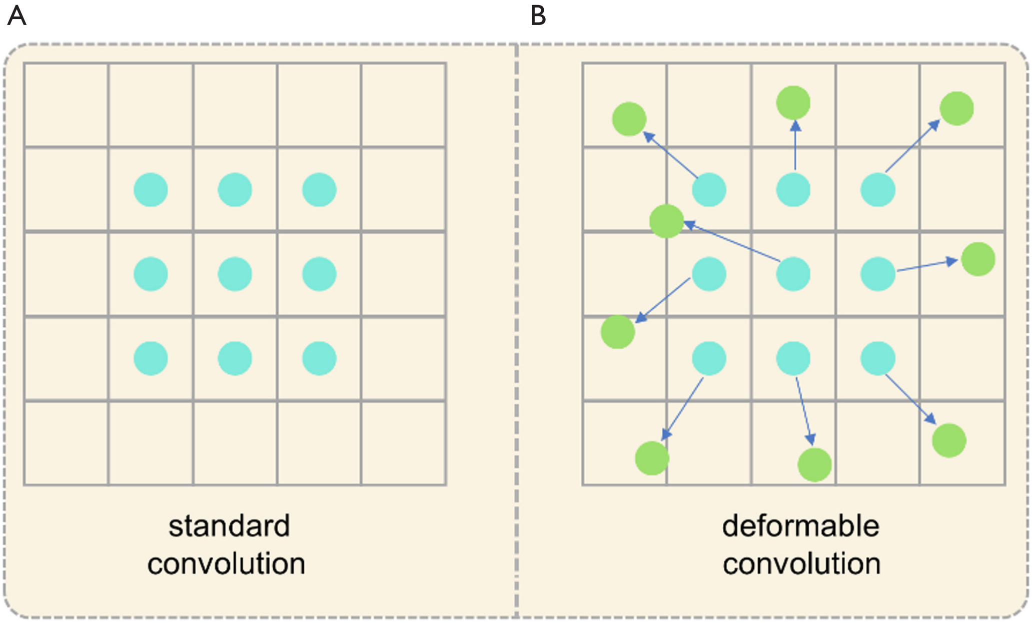

Figure S1 Illustration of the sampling locations in 3×3 standard and deformable convolutions. (A) Regular sampling points (blue points) for standard convolution. (B) Irregular sampling points (green points) with additional offsets (light blue arrow).

https://dx.doi.org/10.21037/qims-24-1451

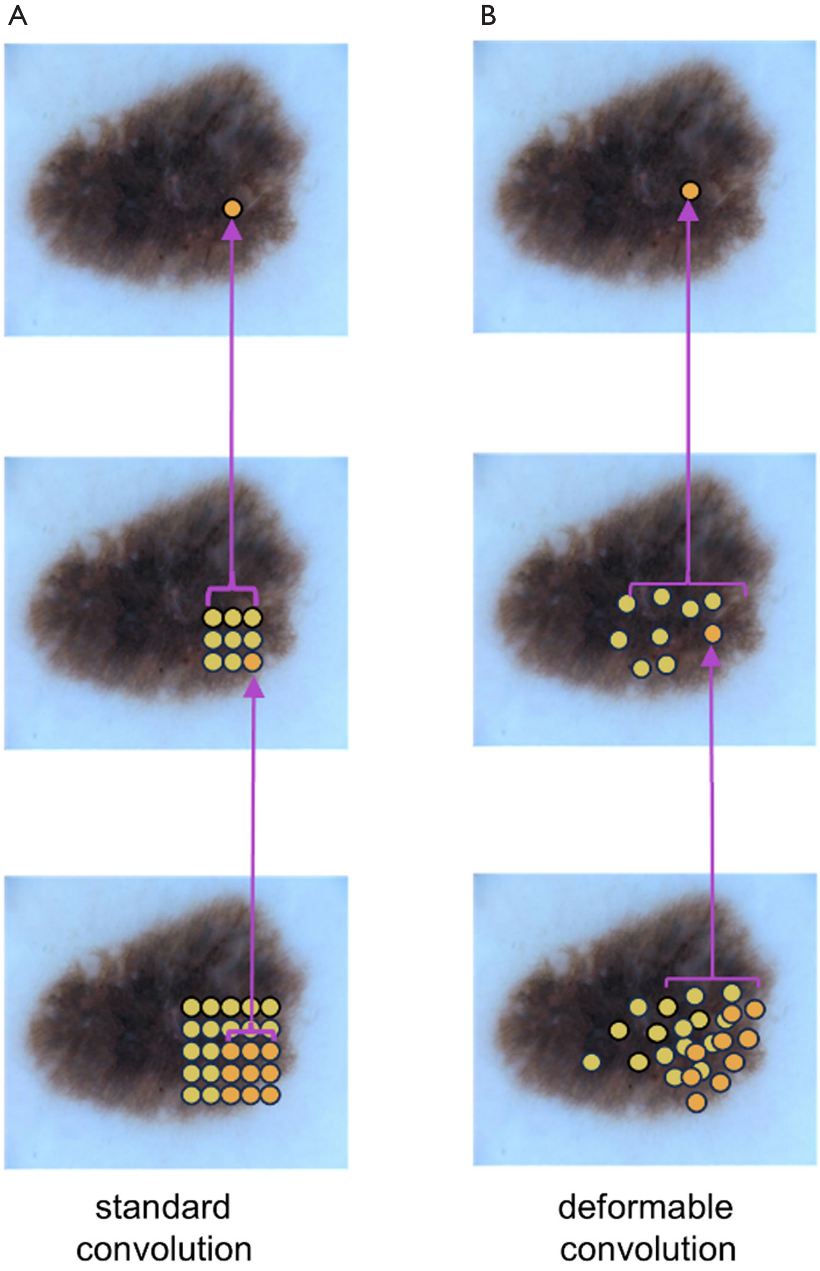

Figure S2 Illustration of the fixed receptive field of the 3×3 convolution kernel in standard convolution (A) and the adaptive receptive field of the 3×3 convolution kernel in deformable convolution (B).

https://dx.doi.org/10.21037/qims-24-1451