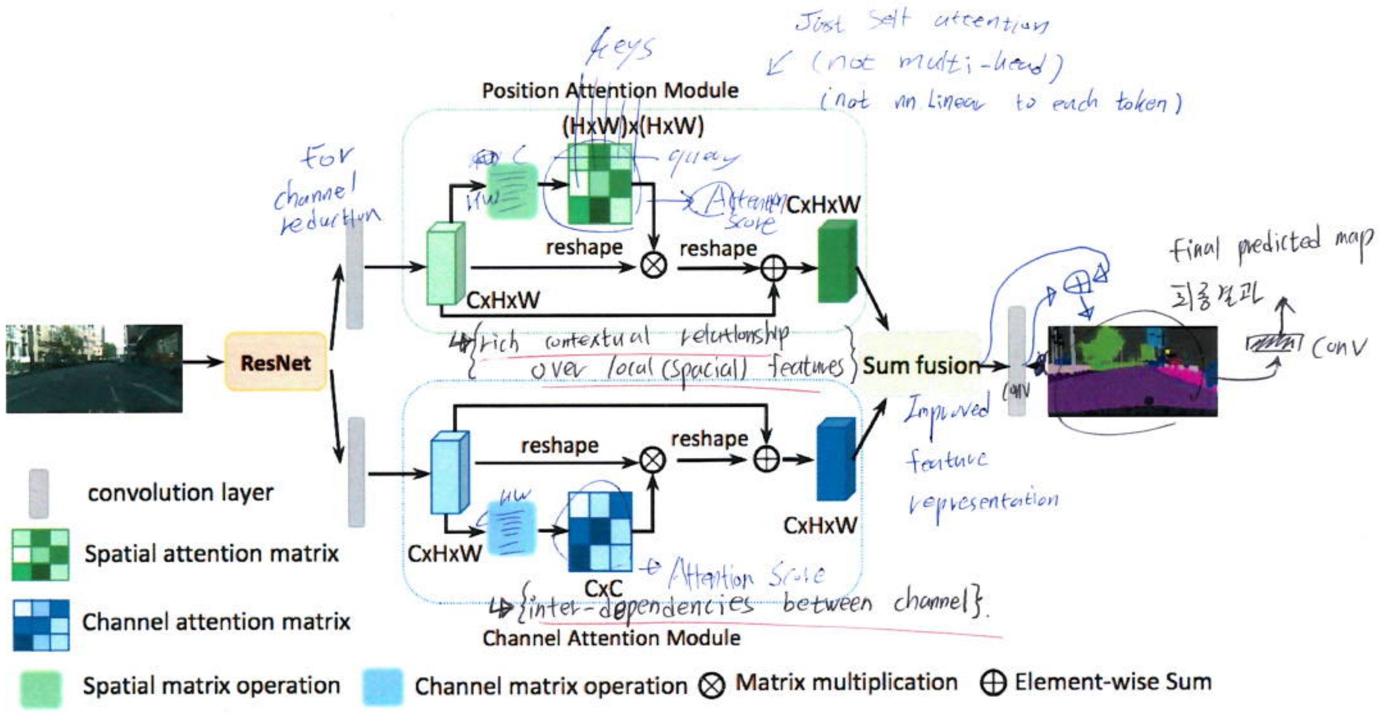

Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

exploring effectiveness of non-local operation in spacetime dimension for videos and images.

Different from previous works, we extend the self-attention mechanism in the task of scene segmentation, and carefully design two types of attention modules to capture rich contextual relationships for better feature representations with intra-class compactness. Comprehensive empirical results verify the effectiveness of our proposed method.

3. Dual Attention Network

In this section, we first present a general framework of our network and then introduce the two attention modules which capture long-range contextual information in spatial and channel dimension respectively. Finally we describe how to aggregate them together for further refinement.

3.1. Overview

Given a picture of scene segmentation, stuff or objects, are diverse on scales, lighting, and views. Since convolution operations would lead to a local receptive field, the features corresponding to the pixels with the same label may have some differences. These differences introduce intra-class inconsistency and affect the recognition accuracy. To address this issue, we explore global contextual information by building associations among features with the attention mechanism. Our method could adaptively aggregate long-range contextual information, thus improving feature representation for scene segmentation.

As illustrated in Figure. 2, we design two types of attention modules to draw global context over local features generated by a dilated residual network, thus obtaining better feature representations for pixel-level prediction. We employ a pretrained residual network with the dilated strategy [3] as the backbone. Noted that we remove the down-sampling operations and employ dilated convolutions in the last two ResNet blocks, thus enlarging the size of the final feature map size to 1/8 of the input image. It retains more details without adding extra parameters. Then the features from the dilated residual network would be fed into two parallel attention modules. Take the spatial attention modules in the upper part of the Figure. 2 as an example, we first apply a convolution layer to obtain the features of dimension reduction. Then we feed the features into the position attention module and generate new features of spatial long-range contextual information through the following three steps. The first step is to generate a spatial attention matrix which models the spatial relationship between any two pixels of the features. Next, we perform a matrix multiplication between the attention matrix and the original features. Third, we perform an element-wise sum operation on the above multiplied resulting matrix and original features to obtain the final representations reflecting long-range contexts. Meanwhile, long-range contextual information in channel dimension are captured by a channel attention module. The process of capturing the channel relationship is similar to the position attention module except for

backbone