high-level semantic features. Although the context fusion helps to capture different scales objects, it can not leverage the relationship between objects or stuff in a global view, which is also essential to scene segmentation.

Another type of methods employs recurrent neural networks to exploit long-range dependencies, thus improving scene segmentation accuracy. The method based on 2D LSTM networks [1] is proposed to capture complex spatial dependencies on labels. The work [18] builds a recurrent neural network with directed acyclic graph to capture the rich contextual dependencies over local features. However, these methods capture the global relationship implicitly with recurrent neural networks, whose effectiveness relies heavily on the learning outcome of the long-term memorization.

To address above problems, we propose a novel framework, called as Dual Attention Network (DANet), for natural scene image segmentation, which is illustrated in Figure. 2. It introduces a self-attention mechanism to capture features dependencies in the spatial and channel dimensions respectively. Specifically, we append two parallel attention modules on top of dilated FCN. One is a *position attention module* and the other is a *channel attention module*. For the position attention module, we introduce the self-attention mechanism to capture the spatial dependencies between any two positions of the feature maps. For the feature at a certain position, it is updated via aggregating features at all positions with weighted summation, where the weights are decided by the feature similarities between the corresponding two positions. That is, any two positions with similar features can contribute mutual improvement regardless of their distance in spatial dimension. For the channel attention module, we use the similar self-attention mechanism to capture the channel dependencies between any two channel maps, and update each channel map with a weighted sum of all channel maps. Finally, the outputs of these two attention modules are fused to further enhance the feature representations.

It should be noted that our method is more effective and flexible than previous methods [4, 29] when dealing with complex and diverse scenes. Take the street scene in Figure. 1 as an example. First, some 'person' and 'traffic light' in the first row are inconspicuous or incomplete objects due to lighting and view. If simple contextual embedding is explored, the context from dominated salient objects (e.g. car, building) would harm those inconspicuous object labeling. By contrast, our attention model selectively aggregates the similar features of inconspicuous objects to highlight their feature representations and avoid the influence of salient objects. Second, the scales of the 'car' and 'person' are diverse, and recognizing such diverse objects requires contextual information at different scales. That is, the features at different scale should be treated equally to represent the same semantics. Our model with attention mechanism just aims to adaptively integrate similar features at any scales from a global view, and this can solve the above problem to some extent. Third, we explicitly take spatial and channel relationships into consideration, so that scene understanding could benefit from long-range dependencies.

Our main contributions can be summarized as follows:

- We propose a novel Dual Attention Network (DANet) with self-attention mechanism to enhance the discriminant ability of feature representations for scene segmentation.
- A position attention module is proposed to learn the spatial interdependencies of features and a channel attention module is designed to model channel interdependencies. It significantly improves the segmentation results by modeling rich contextual dependencies over local features.
- We achieve new state-of-the-art results on three popular benchmarks including Cityscapes dataset [5], PASCAL Context dataset [14] and COCO Stuff dataset [2].

2. Related Work

Semantic Segmentation. Fully Convolutional Networks (FCNs) based methods have made great progress in semantic segmentation. There are several model variants proposed to enhance contextual aggregation. First, Deeplabv2 [3] and Deeplabv3 [4] adopt atrous spatial pyramid pooling to embed contextual information, which consist of parallel dilated convolutions with different dilated rates. PSPNet [29] designs a pyramid pooling module to collect the effective contextual prior, containing information of different scales. The encoder-decoder structures [**?**, 6, 8, 9] fuse mid-level and high-level semantic features to obtain different scale context. Second, learning contextual dependencies over local features also contribute to feature representations. DAG-RNN [18] models directed acyclic graph with recurrent neural network to capture the rich contextual dependencies. PSANet [30] captures pixel-wise relation by a convolution layer and relative position information in spatial dimension. In addition, EncNet [27] introduces a channel attention mechanism to capture global context.

Self-attention Modules. Attention modules can model long-range dependencies and have been widely applied in many tasks [11, 12, 17, 19–21]. In particular, the work [21] is the first to propose the self-attention mechanism to draw global dependencies of inputs and applies it in machine translation. Meanwhile, attention modules are increasingly applied in image vision flied. The work [28] introduces self-attention mechanism to learn a better image generator. The work [23], which is related to self-attention module, mainly