Global Context and Geometric Priors for Effective Non-Local Self-Attention

Sanghyun Woo[1]
shwoo93@kaist.ac.kr
Dahun KIm[1]
mcahny@kaist.ac.kr
Joon-Young Lee[2]
jolee@adobe.com
In So Kweon[1]
iskweon77@kaist.ac.kr

[1] KAIST
[2] Adobe Research

Abstract

Capturing relationships among local and global features in an image is crucial for proper visual understanding. However, the convolution operation is inherently limited at utilizing long-range information due to its small receptive field. Existing approaches thus heavily rely on non-local network strategies to make up for the locality of convolutional features. Despite their successful applications in various tasks, we propose there is still considerable room for improvement, by exploring the effectiveness of global image context and position-aware representations. Notably, the concept of the relative position is surprisingly under-explored in the vision domain, whereas it has proven to be useful for modeling dependencies in machine translation tasks. In this paper, we propose a new relational reasoning module, that incorporates a contextualized diagonal matrix and 2D relative position representations. While being simple and flexible, our module allows the relational representation of a feature point to encode the whole image context and its relative position information. We also explore multi-head and dropout strategies to improve the relation learning further. Extensive experiments show that our module shows consistent improvements over the state-of-the-art baselines on different vision tasks, including detection, instance segmentation, semantic segmentation, and panoptic segmentation.

1 Introduction

Relational reasoning is one of the core abilities for general intelligence and is beneficial for various vision tasks. However, it has proven difficult for convolutional neural networks (CNNs) to learn the concept of a relationship directly. By construction, the convolution is operated with a local receptive field, which can only model local information. Although stacking multiple convolution layers can enlarge the effective receptive field, it also brings several unfavorable issues in practice. First, stacking convolutional layers is not scalable because doing so will make the model excessively deep and inefficient, which increases the risk of over-fitting. Second, long-range relationships are mainly captured after the model

reaches deep layers with large receptive fields, and this may cause an information delay in shallow layers when performing higher-level reasoning. Third, delayed reasoning incurs an elongated computation graph, consisting of several layers for cross-layer communication, raising optimization difficulties during training.

Recent studies attempt to address these issues in many aspects. Yu *et al*. [59] and Dai *et al*. [12] augmented the sampling locations of convolution operation, using dilation or learnable offsets, respectively. Along the same line of reasoning, Peng *et al*. [39] enlarged the convolution kernel size with a decomposed structure. Though the receptive field of the single convolution layer was increased effectively, the result remained insufficient for global reasoning itself. Several studies [4, 23, 46] have shown that adopting recurrent neural networks (RNNs) along with a CNN can achieve long-range reasoning. However, these methods heavily rely on the memorization ability of the RNN, meaning that the global relationship is captured implicitly. Chen *et al*. [7, 8, 9] and Zhao *et al*. [65] attempt to aggregate multi-scale, pyramid feature representations using different dilated convolutions or pooling operations, respectively. While the global contextual information is well captured, in some challenging scenarios (e.g., occlusion, illumination), a rough context may not be enough to resolve ambiguities.

To go beyond the simple context and explore the fine-grained relational representations [2], more sophisticated approaches have been introduced. In particular, several studies[10, 27, 28] employ graph convolution [24]. This method initially transforms the coordinate space CNN features into the interaction space by projection. The graph convolution is then applied in the interaction space to perform graph reasoning. Finally, the resulting features are re-projected back into the original coordinate space. With semantic meanings stored in the node representations, graph convolution enables regional reasoning. However, the projection and re-projection steps significantly harm the existing geometric structures, losing meaningful spatial relationships in the image. Another line of studies [14, 15, 52, 60, 64, 66] adopts self-attention [51]. A non-local neural network [52] initially brings the Transformer formulation [51] to the vision field. It captures long-range relations by explicitly attending to all features in the image, which allows the model to build a direct relationship with another long-distance representation. Despite the success in incorporating long-range relational information, we find that two essential cues are missing in the current form. First, this approach treats each feature in the input image individually and performs attention over the whole input. Consequently, the contextual information is not taken into account in the computation of the relationships between elements. Second, representations of position information are absent, and thus cannot utilize well of the intrinsic spatial correlations in the images. We speculate that natural scenes exhibit certain motifs; particular object shape/size or object compositions repeatedly recur, hence, knowing their relative position distances makes it easier to model corresponding regularities. Despite the efficacy of relative position encoding, which has already been verified in machine translation tasks [13, 44, 58], we find it is surprisingly under-explored in the vision domain.

Motivated to tackle these limitations, we propose a novel relational reasoning module that is aware of the global context and relative position. First, we introduce a contextualized diagonal matrix that performs channel-wise attention on the relation computation. It effectively builds a connection between the global context and the individual input features. Next, we present 2D relative position representations. An apparent distinction from the previous studies [13, 44, 58] is that we consider pixels in the 2D image instead of words in sequence as our primitive elements. We thus decompose the target task into the two sub-tasks of 1D relative position embedding along the x-axis and y-axis, respectively. This facilitates ex-

ploitation of the spatial relationships within an image. Moreover, we investigate multi-head and dropout strategies for regularization. We finally incorporate all of the proposed components into the current non-local attention form [52]. In order to verify the effectiveness of our module, we experiment on the most common architecture designs in recent vision tasks. More specifically, we consider two important architecture groups: FCN-based and FPN-based, and present two derived instantiations. First, we apply our relational module to a fully convolutional network (FCN) [17, 33]. With other conditions set equally, we thoroughly compare the proposed method with other state-of-the-art approaches on semantic segmentation. Second, to validate our relational modeling ability on multi-scale hierarchical features, we integrate our module with a feature pyramid network (FPN) [29]. We evaluate the performance on three tasks: detection, instance segmentation, and panoptic segmentation. Extensive experiments show that our module can consistently boost the performance of state-of-the-art baselines with healthy margins, demonstrating the advantage of the proposed module in versatile scenarios.
We summarize our contributions as follows:

1. To our best knowledge, it is the first time that both global context and relative position representations are incorporated into a single non-local form. Our unified design is novel in that the previous works are limited to using either only the global context [6] or the relative position [44], which is complementary to each other. We also introduce multi-head and dropout techniques, improving the relation learning further.

2. A recent study [6] shows that the attention map of different query position is almost the same in the non-local block. We show that our formulation brings query-sensitivity into the non-local relation module, resolving the query-redundancy problem effectively.

3. We present two instantiations based on our new formulation. We consider two-important architecture groups. First, we append our module at the end of the FCN backbone, building a strong semantic segmentation model by aggregating global relational features. Second, we combine our module with FPN to further utilize multi-scale pyramidal features. We apply it to the various detection/instance segmentation frameworks, including one-stage, two-stage, and cascade.

4. We thoroughly investigate the effect of our proposals with extensive ablation studies. Finally, we show the superior performance of the proposed method over the state-of-the-art approaches on various vision tasks.

2 Related Works

Context Modeling Context matters [40, 45, 50]. Many efforts have been made to present an effective means by which to model and capture contextual information by, for example, enlarging the convolution sampling locations [12, 39, 59], using encoder-decoder designs [1, 33, 35, 42, 53, 56], combining global features with local patches [7, 8, 9, 32, 48, 57, 65], and incorporating image-level context with gating/attention modules [6, 21, 22, 37, 38, 55]. However, the context information is rarely considered with respect to relational embedding. Specifically, in the current non-local form [52], the calculation of the relationship between two individual features (i.e., a query and a key)

misses the opportunity to take advantage of a useful context. We therefore propose a simple yet effective way to combine the context with a non-local operation. In practice, we introduce a diagonal matrix that utilizes channel-wise attention during the relation computation. The implementation is inspired by SE [22], which gathers the image-level global information by spatial-pooling. As a result, our proposal effectively integrates two different algorithms, non-local [52] and SE [22], into a single formula.

Geometric Modeling Only a few recent works [16, 19, 20, 31] aim to augment convolutional features with position information. Liu *et al*. [31] explicitly concatenate 2D absolute positional channels to the features. Hu *et al*. [19] learn the relative scale and position differences between the objects, Gu *et al*. [16] learn the relative position between objects and pixels, and Hu *et al*. [20] encode the relative positions of local convolution grids. Our formulation differs from the previous works in several aspects. First, unlike the absolute position [31], our relative position satisfies the translation-equivariance property [26]. This is crucial when dealing with images and helps the model generalize to unseen object positions during training. Second, in contrast to several earlier works [16, 19, 20], our form captures a global content-dependent positional bias. We find that this is essential for modeling complex object-dependent motifs in images. Third, instead of using distance information directly [16, 19, 20, 31], we use the sinusoid function [51] for distance encoding. This allows the model to attend to relative positions more easily. We thoroughly conduct ablation studies and demonstrate the effectiveness of our proposals.

Self-Attention Graph-based methods [24] and non-local models [52] are two dominant approaches for relational embedding. Due mainly to their powerful relationship modeling ability, these methods are widely adopted in various vision tasks [10, 19, 43, 52, 54, 62]. In this work, we propose an improved non-local form that incorporates the context and the relative position information, which are a surprisingly overlooked in previous studies. Recent works are limited to the use of either only the global context [6] or the relative position [3, 44], which is complementary to each other as we have empirically validated. Furthermore, we explore multi-head and dropout techniques and demonstrate these enhanced forms of relation learning further.

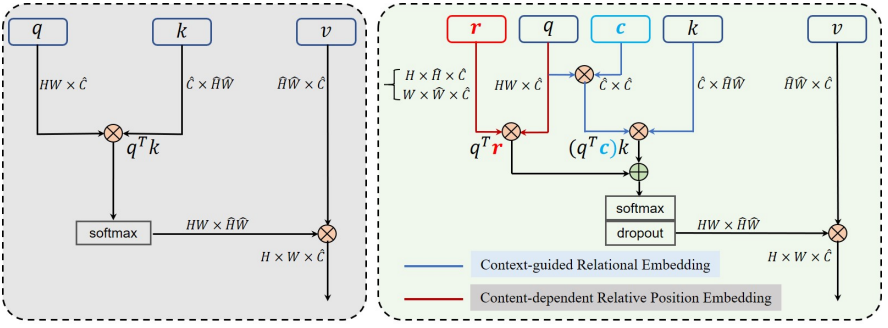
3 Method

In this section, we firstly revisit the definition of the non-local network [52] in Sec. 3.1, then detail the proposed formulation in Sec. 3.2. Finally, with the improved non-local formulation, we derive two practical instantiations.

3.1 Self-Attention in Non-local Network

Consider the input CNN feature map $X \in \mathcal{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, spatial width, and height, respectively. At first, three different 1×1 convolutions $W_q \in \mathcal{R}^{\hat{C}\times C}$, $W_k \in \mathcal{R}^{\hat{C}\times C}$, and $W_v \in \mathcal{R}^{\hat{C}\times C}$ are used to transform X into $q \in \mathcal{R}^{\hat{C}\times H\times W}$, $k \in \mathcal{R}^{\hat{C}\times H\times W}$, $v \in \mathcal{R}^{\hat{C}\times H\times W}$ embeddings as

$$q = W_q(X), \quad k = W_k(X), \quad v = W_v(X). \tag{1}$$



(a) Non-local Module (b) Context-guided Position-aware Non-local Module

Figure 1: Design of a standard non-local module (a) and the proposed approach (b).

$\hat{C}$ denotes the number of channels of the new embeddings. The three embeddings are then reshaped to the size of $\hat{C} \times N$, where N indicates the total number of the spatial locations (i.e., $N = H \cdot W$). A subsampling layer is often applied to k and v to reduce the overall computation. In this case, $k \in \mathcal{R}^{\hat{C} \times \hat{N}}$ and $v \in \mathcal{R}^{\hat{C} \times \hat{N}}$ have the total spatial locations of $\hat{N} = \hat{H} \cdot \hat{W}$ with $\hat{H} < H$ and $\hat{W} < W$. Next, the dense relationships are computed in the relation (i.e., affinity) matrix $A \in R^{N \times \hat{N}}$ as

$$A = f(q^T k). \tag{2}$$

Here, f is the normalizing function, which can take various forms (e.g., scaling, averaging, softmax). We choose the softmax function. The output, $O \in \mathcal{R}^{N \times C}$, is then computed based on the calculated relationships as

$$O = W_o(Av^T). \tag{3}$$

The 1×1 convolution, $W_o \in \mathcal{R}^{C \times \hat{C}}$, is used to recover the number of input channels C. In short, we initially encode the pixel-wise dense relations into a relation matrix (Eq. (2)). In the deep feature space, it basically determines the relevance in a patch-level. Afterward, we collect the necessary features from the entire input using attentional weighted-sum equation (Eq. (3)).

3.2 Proposed formulation

We reformulate the original relation computation, $q^T k$, by introducing the ***context*** and ***relative position***. To improve the relation learning process further, we adopt *multi-head* and *dropout* strategies. An overview of the proposed method is shown in Fig. 1. These differences lead to significantly higher accuracy on various vision benchmarks. Below, we elaborate on each step of the process.

Incorporating Context Priors Even with the same content, pair-wise relations can vary in a different context. For example, imagine a human in the indoor scene and an outdoor scene. We can easily expect that humans may often co-occur with chairs or tables in the indoor scene. On the other hand, we expect traffic-lights and/or buildings instead in the outdoor scene. Thus, it is challenging to model accurate relationships only using their contents without considering the context. To offset the lack of contextual information and to maintain the flexibility of non-local operations, we propose the contextualization of the

input feature. Specifically, we present a diagonal context matrix, $\mathbf{c} \in \mathcal{R}^{\hat{C}\times\hat{C}}$, and reformulate the previous equation, $q^T k$, as

$$A = f((q^T \mathbf{c})k). \tag{4}$$

The context matrix is computed as $\mathbf{c} = diag(\sigma(W_c(AvgPool(X))))$. Here, $AvgPool$, $W_c \in \mathcal{R}^{\hat{C}\times C}$, and σ denote global average pooling, 1×1 convolution, and the sigmoid function, respectively. In fact, previous studies [22, 55] have shown that $AvgPool(X) \in \mathcal{R}^C$ softly encodes the global information of the features. We further embed and normalize this using W_c and the sigmoid function. To connect the summarized context with the input feature, we cast the vector to a diagonal matrix. Finally, the context matrix, $\mathbf{c}$, contextualizes the input feature, q, via channel-wise [22] attention. At a high-level, our new formulation enables the overall relation computation to be modulated by the given context.

Incorporating Geometric Priors Natural scenes generally exhibit certain motifs, which are highly correlated with the relative distance information. The relative distance can be defined either within the same object or between objects. The former is related to the particular object's shape/size. For example, paintings on a wall or street-lights along roads usually have regular shapes/sizes, which is information with which any pixels within the object could correlate. The latter is related to the object compositions. For example, when a human and a racket co-exist in the image, and they often do so at fairly short distances. Therefore, knowing the relative position distances is important and makes it easier to model these regularities. We therefore integrate the relative position information, $\mathbf{r}$, into the former formulation as

$$A = f((q^T \mathbf{c})k + q^T \mathbf{r}). \tag{5}$$

In contrast to earlier works [19, 20], note that position bias is established in a content-dependent manner (i.e., r vs q^Tr). Because we consider pixels in a 2D space instead of words in a 1D sequence [42], we consider the relative position embedding as a combination of the decomposed sub-tasks. Specifically, the relative position, $\mathbf{r}$, is computed as

$$\begin{aligned} \mathbf{r} &= r_x + r_y \\ r_x &= \text{Transpose}(W_{r_x}(R_X)) \in \mathcal{R}^{W\times\hat{C}\times\hat{W}} \\ r_y &= \text{Transpose}(W_{r_y}(R_Y)) \in \mathcal{R}^{H\times\hat{C}\times\hat{H}} \end{aligned} \tag{6}$$

where $R_X \in \mathcal{R}^{\frac{C}{2}\times W\times\hat{W}}$ and $R_Y \in \mathcal{R}^{\frac{C}{2}\times H\times\hat{H}}$ are 1D relative position encodings along the x-axis and y-axis, respectively. $W_{r_x} \in \mathcal{R}^{\hat{C}\times\frac{C}{2}}$ and $W_{r_y} \in \mathcal{R}^{\hat{C}\times\frac{C}{2}}$ are 1×1 convolutions. The Transpose$(\cdot)$ operation swaps the dimensions of the first and the second (e.g., $\mathrm{R}^{\hat{C}\times W\times\hat{W}} \rightarrow \mathrm{R}^{W\times\hat{C}\times\hat{W}}$). Here, $R_{X\,i,j} \in \mathcal{R}^{\frac{C}{2}}$ can be expressed as

$$R_{X\,n,i,j} = \begin{cases} sin((i-j)/1000^{\frac{2n}{C}}) & \text{if } n \text{ is even, and} \\ cos((i-j)/1000^{\frac{2(n-1)}{C}}) & \text{if } n \text{ is odd,} \end{cases} \tag{7}$$

where $n \in [0, \frac{C}{2})$, $i \in [0, W)$, and $j \in [0, \hat{W})$. We use different wavelengths of sine and cosine functions for the encoding [51], which enables the model to better attend to the relative position. R_Y is defined similarly.

We also detail the computation of $q^T\mathbf{r}$ $(= q^T r_x + q^T r_y)$. When computing $q^T r_x$, we first reshape $q^T \in N\times C$ to $W\times H\times\hat{C}$. The reshaping operation moves W to the non-matrix

dimension. Next, we multiply it by $r_x \in \mathcal{R}^{W\times\hat{C}\times\hat{W}}$ as:

$$q^T r_x : \mathcal{R}^{W\times\overbrace{H\times\hat{C}}^{content}} \times \mathcal{R}^{W\times\overbrace{\hat{C}\times\hat{W}}^{rel-pos}} \rightarrow \mathcal{R}^{W\times H\times\hat{W}} \tag{8}$$

The operation gathers the interaction values between the content and the relative position features along the W dimension (i.e., x-axis). Similarly, when computing $q^T r_y$, we collect the interaction values over the y-axis as follows:

$$q^T r_y : \mathcal{R}^{H\times\overbrace{W\times\hat{C}}^{content}} \times \mathcal{R}^{H\times\overbrace{\hat{C}\times\hat{H}}^{rel-pos}} \rightarrow \mathcal{R}^{H\times W\times\hat{H}} \tag{9}$$

We finally element-wise sum the resulting matrices, $q^T r_x$ and $q^T r_y$, using the general broadcasting rules. This produces $q^T \mathbf{r} \in \mathcal{R}^{H\times W\times\hat{H}\times\hat{W}}(\mathcal{R}^{N\times\hat{N}})$, which is the global content-dependent positional bias. Finally, we transfer this positional representation using addition (Eq. (5)).

Ensemble & Regularization Two additional techniques that are disregarded in the existing non-local form [52] are considered here. We first adopt a multi-head strategy, which concatenates the module outputs using different model weights. That allows the module to learn diverse relation patterns, extracting relations densely. After conducting parameter analysis (see supplementary materials), we set the number of heads to 8. In the meantime, we try to prevent our module from adapting too much to the training data. During the training process, we thus apply the dropout technique [47] to the relation matrix (Eq. (5)). Partially masking out the relation patterns not only has a regularization effect but also helps our model to learn the general relationships that lead to performance improvements.

Two instantiations Based on the newly proposed non-local formulation, we present two derived instantiations. In particular, we consider two important architecture groups: the **FCN-based** and **FPN-based**. First, we append our module at the end of the FCN. For the FCN, we employ ImageNet-pretrained ResNet50 [17]. We remove the last two down-sampling operations and adopt multi-grid dilated convolutions [8]. We evaluate the network on a semantic segmentation task. Under the same setting, we show that our module outperforms other state-of-the-art approaches [7, 10, 14, 53, 54, 65].

Although our module is capable of capturing global relationships for any given feature representations, we find that our module can benefit further from in-network pyramidal representations, which contain different feature patterns in practice [51]. We therefore attempt to combine our module with FPN [29] as inspired by BFPN [36]. This is referred to as EBFPN. The detailed forward process of network is as follows: First, the input FPN [29] features $\{C_2, C_3, C_4, C_5\}$ are resized to an intermediate target size (*e.g.*, C_3). Second, the *balanced semantic features* are obtained by simple averaging as $C = \frac{1}{L}\sum_{l=l_2}^{l_5} C_l$, where L denotes the number of multi-level features. Third, the averaged feature, C, is refined with the proposed module. Finally, the enhanced feature is redistributed to the original features via residual summation. We verify the effectiveness of EBFPN on various vision tasks including detection, instance segmentation, and panoptic segmentation. We show that EBFPN consistently pushes the performance of state-of-the-art baseline models and outperforms BFPN [36] with large margins.

Method	Pascal Context		ADE20K		Cityscapes	
	pixAcc%	mIou%	pixAcc%	mIou%	pixAcc%	mIou%
FCN	75.57	45.78	78.09	37.44	95.28	73.32
FCN + ASPP [2]	77.72	48.79	78.92	38.76	95.32	73.49
FCN + PSP [63]	77.85	49.44	79.21	39.96	95.38	74.31
FCN + EncModule [63]	78.30	49.60	79.43	40.03	95.45	74.46
FCN + ACFModule [62]	78.53	50.19	79.56	40.15	95.47	74.85
FCN + GCModule [4]	78.26	50.01	79.32	40.07	95.41	74.57
FCN + DualAttention [14]	78.90	51.13	79.62	40.24	95.52	75.36
FCN + GloRe [10]	78.78	50.40	79.58	40.21	95.48	75.31
FCN + Ours	**79.14**	**51.27**	**79.72**	**40.41**	**95.65**	**75.55**

Table 1: Semantic segmentation results with single-scale testing. We use ResNet50 FCN backbone and test on Pascal Context, ADE20K, and Cityscapes.

Method	Params	Flops	mAP	m$AP_{.5}$	m$AP_{.75}$
RetinaNet [30]	37.74M	239.32G	35.5	55.4	37.7
RetinaNet + BFPN [36]	38.01M	240.37G	36.2	56.4	38.1
RetinaNet + EBFPN	38.14M	239.98G	**37.5**	**56.9**	**39.9**
FCOS [49]	32.02M	200.63G	36.5	55.8	38.6
FCOS + BFPN [36]	32.29M	201.68G	36.8	56.4	39.1
FCOS + EBFPN	32.42M	201.29G	**37.2**	**56.8**	**39.6**
Faster R-CNN [40]	41.53M	207.07G	37.1	59.3	40.1
Faster R-CNN + BFPN [36]	41.79M	208.12G	37.7	59.9	40.5
Faster R-CNN + EBFPN	41.93M	207.73G	**38.9**	**60.0**	**42.2**
Mask R-CNN [18]	44.18M	275.58G	37.2(34.1)	58.9(55.4)	40.3(36.2)
Mask R-CNN + BFPN [36]	44.44M	276.63G	38.1(34.8)	60.3(57.1)	41.5(37.1)
Mask R-CNN + EBFPN	44.57M	276.24G	**39.2(35.7)**	**61.7(57.7)**	**42.7(38.1)**
Cascade Mask R-CNN [6]	77.10M	440.23G	41.2(35.7)	59.3(56.4)	44.8(38.5)
Cascade Mask R-CNN + BFPN [36]	77.36M	441.28G	42.0(36.3)	60.8(57.7)	45.6(38.9)
Cascade Mask R-CNN + EBFPN	77.49M	440.89G	**42.7(36.9)**	**62.5(58.4)**	**46.4(39.4)**

Table 2: Detection/Instance segmentation on COCO *test-dev*. The numbers in the parentheses show instance segmentation scores.

4 Experiments

In this section, we evaluate our two instantiations on various vision tasks and, compared with the state-of-the-art baselines. we also visualize the learned relationships and show that diverse query-specific relationships can be modeled, which are rarely captured by existing non-local forms [6, 52]. Due to page limitations, we provide more experimental analysis including comprehensive ablation studies and qualitative results in the **supplementary materials**.

4.1 Results on Semantic Segmentation

We compare our method with existing state-of-the-art methods [6, 7, 10, 14, 63, 64, 65] using three different semantic segmentation datasets [11, 34, 67]. For a fair comparison, we reproduce all of the previous approaches in our Pytorch platform. Our focus was not on achieving state-of-the-art results, but on evaluating each module's pure long-term context modeling ability. Thus, we did not adopt any sophisticated backbones or heuristics (e.g., multi-scale testing, auxiliary loss), which can bring extra performance gains, in the experiments. We

Method	Backbone	Params	Flops	PQ	PQ^{Th}	PQ^{St}
Panoptic FPN [25]	ResNet50	45.82MB	275.58G	38.5	46.1	26.9
Panoptic FPN + BFPN [36]		46.08MB	276.63G	39.7	46.7	30.8
Panoptic FPN + EBFPN		46.21MB	276.24G	**41.3**	**47.3**	**32.2**
Panoptic FPN [25]	ResNet101	64.81MB	351.65G	40.5	47.8	29.5
Panoptic FPN + BFPN [36]		65.07MB	352.70G	41.8	48.5	32.9
Panoptic FPN + EBFPN		65.20MB	352.31G	**42.9**	**49.1**	**33.4**

Table 3: Panoptic segmentation results on COCO *val*.

train models end-to-end using synchronized multi-gpu batch normalization (SyncBN) [53]. For the testing, we adopt single-scale inference. The experimental results are summarized in Table 1. We can clearly see that our module outperforms the state-of-the-art approaches significantly. The proposed module outperforms DualAttention [14], a recent non-local attention based model without the context or relative position embeddings. Moreover, our method is superior to the GloRe [10], which is the latest graph convolution based approach. We note that the GC module [6], which only models the global context and misses relative position embedding, has a clear limitation with regard to aggregating rich contextual features compared to the proposed method.

4.2 Results on Detection & Instance Segmentation

At this point, we verify our second instantiation, EBFPN. We test the proposed method on various detection frameworks [6, 18, 30, 41, 49] and report the scores on COCO *test-dev*. We again reproduce all the previous methods in our Pytorch platform. We use the same ResNet50 + FPN backbone [29]. The experimental results are shown in Table 2. We observe that the proposed EBFPN greatly improves the baselines and outperforms BFPN [36] with healthy margins in all cases. The main design difference between EBFPN and BFPN lies in the refining step (i.e., Non-local [52] vs. Ours). This implies that our approach generates much stronger feature pyramid representations through its enhanced relational embedding ability. Also, our module feasibly maintains an appropriate level of parameter and computational overheads. Compared t one study [52], our module has slightly more parameters due to the additional position embedding layer, though ours also has less computational overhead due to its multi-head design.

4.3 Results on Panoptic Segmentation

In this experiment, we apply EBFPN to the Panpotic FPN [25], which is a strong baseline architecture for the panoptic segmentation (see Table 3). Surprisingly, we observe that EBFPN dramatically improves the baseline performances. We find that the significant improvement is mainly due to the non-locally aggregated global context and relative position information, which are crucial for both the instance and semantic segmentation tasks but are lacking in original model. Furthermore, an interesting point is that EBFPN with the ResNet50 backbone (41.3 PQ) significantly outperforms the baseline ResNet101 backbone (40.5 PQ) with far less parameter/computational overheads. In general, to obtain better accuracy, it is common to scale up a baseline network by employing a larger backbone. However, the results show that simply increasing the model capacity with larger backbones cannot capture necessary, fine-grained relational representations as ours does. We believe the provided experi-

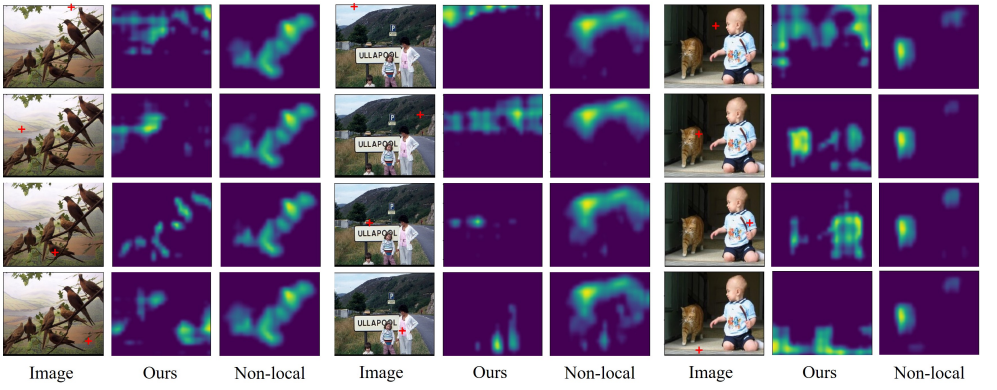


Figure 2: Visualization results of attention maps (i.e., relation) on Pascal Context. We compare our formulation with the non-local form [52]. We show the attention maps corresponding to the '+' sign marked in the input images. Our formulation enables the module to capture both intra- and inter-class relationships.

mental results provide new insight into future panoptic segmentation architecture designs.

4.4 Visualization of Learned Feature Relations

A recent study [6] shows that the attention maps at different query position are nearly identical in non-local blocks [52]. In fact, this means query-specific diverse relationships are barely modeled with the current form. We also observe a similar phenomenon from the visualization task (see Fig. 2). The current non-local form tends to capture only the salient information, and it is quite redundant. On the other hand, the proposed method learn diverse relationships; For example, the first two images demonstrate intra-class feature aggregation, and the last image shows that inter-class feature aggregation, which is related to co-occurrence modeling, is also possible (e.g., a baby and a cat). A more quantitative cosine distance analysis of the feature maps is provided in the **supplementary materials**.

5 Conclusion

In this paper, we aim to enrich local convolutional features using long-range, contextual relationships. We propose an improved non-local form that incorporates context and geometric priors. During the relation computation, our method enables the model to be aware of both the image-level context and relative distance information effectively. We further improve relation learning by introducing multi-head and dropout strategies. We show our proposals consistently boost the performance of state-of-the-art baselines on various vision tasks.

6 Acknowledgements

This work was supported in part by Samsung Electronics Co., Ltd (G01200447), and National Research Foundation of Korea (NRF2020M3H8A1115028, FY2021).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, volume 39, pages 2481–2495. IEEE, 2017.

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[4] Wonmin Byeon, Thomas M Breuel, Federico Raue, and Marcus Liwicki. Scene labeling with lstm recurrent neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3555, 2015.

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. In *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*. IEEE, 2019.

[6] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 40(4):834–848, 2017.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 801–818, 2018.

[10] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 433–442, 2019.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 764–773, 2017.

[13] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[15] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 6748–6757, 2019.

[16] Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei, and Jifeng Dai. Learning region features for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 381–395, 2018.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[20] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, October 2019.

[21] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 9401–9411, 2018.

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[23] Xiaojie Jin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Multi-path feedback recurrent neural networks for scene parsing. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2017.

[25] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, 2019.

[26] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 991–999, 2015.

[27] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 9225–9235, 2018.

[28] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 1853–1863, 2018.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[31] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 9605–9616, 2018.

[32] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014.

[35] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[36] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 821–830, 2019.

[37] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. 2018.

[38] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. A simple and light-weight attention module for convolutional neural networks. *International Journal of Computer Vision*, 128(4):783–798, 2020.

[39] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4353–4361, 2017.

[40] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICAI)*, pages 234–241. Springer, 2015.

[43] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 4967–4976, 2017.

[44] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[45] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. In *Int'l Journal of Computer Vision (IJCV)*, volume 81, pages 2–23. Springer, 2009.

[46] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Dag-recurrent neural networks for scene labeling. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3620–3629, 2016.

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 9627–9636, 2019.

[50] Antonio Torralba. Contextual priming for object detection. In *Int'l Journal of Computer Vision (IJCV)*, volume 53, pages 169–191. Springer, 2003.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[53] Sanghyun Woo, Soonmin Hwang, and In So Kweon. Stairnet: Top-down semantic aggregation for accurate one shot detection. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1093–1102. IEEE, 2018.

[54] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 560–570, 2018.

[55] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 3–19, 2018.

[56] Sanghyun Woo, Soonmin Hwang, Ho-Deok Jang, and In So Kweon. Gated bidirectional feature pyramid network for accurate one-shot detection. *machine vision and applications*, 30(4):543–555, 2019.

[57] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3684–3692, 2018.

[58] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of Neural Information Processing Systems (NeurIPS)*, pages 5754–5764, 2019.

[59] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2016.

[60] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[61] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2019.

[63] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2018.

[64] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2019.

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[66] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 267–283, 2018.

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017.