CONTEXTUAL GRAPH REASONING NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Reasoning has shown great potential recently in modeling long-range dependencies, which are crucial for various computer vision tasks. However, the graph representation learned by existing methods is not effective enough as the relation between feature and graph is under-explored. In this work, we propose a novel method named Contextual Graph Reasoning (CGR) that learns a context-aware relation between feature and graph. This is achieved by constructing the projection matrix based on a global set of descriptors during graph projection, and calibrating the evolved graph based on the self-attention of all nodes during graph reprojection. Therefore, contextual information is well explored in both graph projection and reprojection with our method. To verify the effectiveness of our method, we conduct extensive experiments on semantic segmentation, instance segmentation, and 2D human pose estimation. Our method consistently achieves remarkable improvements over state-of-the-art methods, demonstrating the effectiveness and generalization ability of our method.

1 INTRODUCTION

Over the past years, Convolutional Neural Networks (CNN) significantly boost the performance of computer vision tasks like image classification (Deng et al., 2009; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016), object detection (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015), semantic segmentation (Long et al., 2015; Zhao et al., 2017; Chen et al., 2017a), instance segmentation (He et al., 2017; Bolya et al., 2019) and human pose estimation (Wei et al., 2016; Chen et al., 2018b; Xiao et al., 2018; Sun et al., 2019) etc.

However, convolution is intrinsically limited in modelling long-range dependencies. Although stacking multiple convolution layers can enlarge the receptive field, it leads to higher computational cost and increases the over-fitting risk. Various solutions have been proposed to overcome the limitations of convolution, including but not limited to Conditional Random Fields (Chen et al., 2014; Gadde et al., 2016; Liu et al., 2015; Schwing & Urtasun, 2015; Wang et al., 2015; Zheng et al., 2015), multi-dimensional Long Short-Term Memory (Byeon et al., 2015; Liang et al., 2016), dilated convolution (Chen et al., 2014; 2017a;b; Yu & Koltun, 2015), pyramid pooling operation (Chen et al., 2018a; Zhao et al., 2017), non-local operation (Wang et al., 2018) and self-attention mechanism (Vaswani et al., 2017) to capture the long-range dependencies.

The above methods have shown the effectiveness of modeling long-range dependencies in coordinate space, however, they are computationally expensive. Recently, inspired by the graph reasoning mechanism, several works (Li & Gupta, 2018; Liang et al., 2018; Chen et al., 2019b; Zhang et al., 2019b) are proposed to resolve the problem in graph interaction space, where regions of the feature are defined as graph nodes and interaction between nodes are regarded as edges (Ladickỳ et al., 2009). Compared with coordinate space, graph interaction space is more efficient as the number of graph nodes is much smaller than the number of positions in the feature map. The pipeline of these works first projects the feature into a graph with the projection matrix. Then, reasoning based on graph convolution is performed to aggregate the interaction between nodes. Finally, the evolved graph is reprojected to coordinate space, delivering the feature for prediction. In general, methods based on graph reasoning consist of three steps: graph projection, graph reasoning and graph reprojection respectively.

Although existing methods have demonstrated great potential of graph reasoning, the learned graph representation is not effective enough as the relation between feature and graph is under-explored:

Existing methods adopt projection matrix to build the mapping between feature and graph. We carefully revisit two popular paradigms of projection matrix construction, however, neither of them explores the global contextual information. Apart from this, existing methods directly reproject the evolved graph to feature, assuming the graph convolution can adequately aggregate the interaction between nodes. Unfortunately, similar to standard convolutions, graph convolution also fails to exploit the global context, resulting in less effective feature for prediction. The above analysis motivates us to explore the context-aware relation for more effective graph representation.

In this work, we propose a novel method named Contextual Graph Reasoning (CGR), aiming to learn a context-aware relation between feature and graph. This is achieved by considering the contextual information during both graph projection and reprojection, leading to a new projection module called DGP and a new reprojection module called NCR. To be more specific, in DGP, the feature is convolved to deliver a global set of descriptors, which will be used to generate the projection matrix for graph construction. Compared with GloRe (Chen et al., 2019b), which directly predicts the projection matrix based on local feature, global descriptors can learn sufficient contextual information for more effective graph representation. In NCR, before reprojecting the evolved graph into coordinate space, the graph nodes are calibrated based on the self-attention of all nodes, improving their capability of capturing long-range dependencies.

To the best of our knowledge, our method is the first work to explore contextual information during graph projection and reprojection. The context-aware relation learned by CGR can further boost the performance of graph reasoning, which is evaluated on several computer vision tasks.

Our contributions are summarized as follows:

1. We propose the Descriptor Graph Projection (DGP) module, which constructs the projection matrix based on a global set of descriptors instead of local feature.

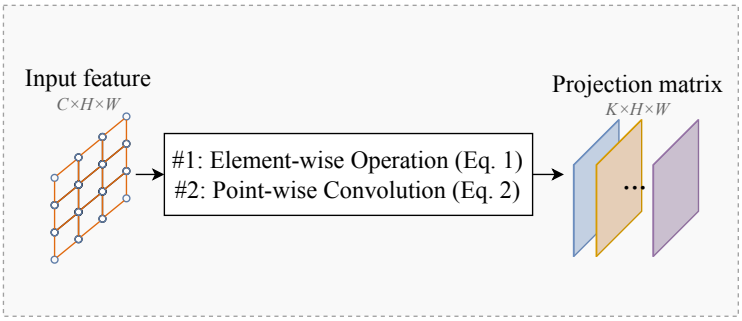
2. We present the Node Collaborative Reprojection (NCR) module, calibrating the graph nodes to improve its capability of long-range dependencies.

3. We conduct extensive experiments on semantic segmentation, instance segmentation and 2D human pose estimation, demonstrating the superiority of our approach by remarkable improvements.

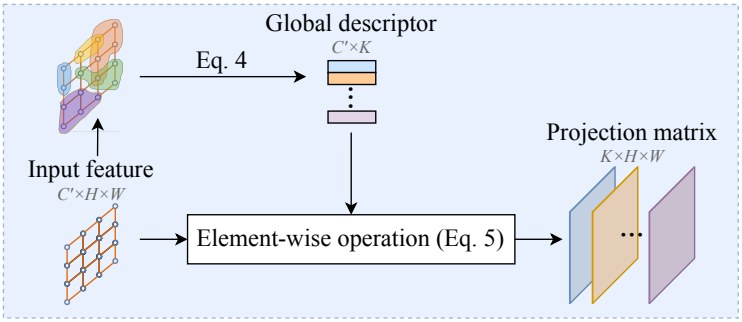
2 Related Work

The recent works on modeling global contextual information can be grouped into two categories. One is modeling contextual dependencies in coordinate space. For instance, (Zhao et al., 2017; Chen et al., 2017b) propose the novel pyramid sampling methods, improving network's capability of multi-scale information. (Wang et al., 2018) adaptively integrates local feature with their contextual dependencies. (Hu et al., 2018) calibrates each feature channel with the global contextual information. (Fu et al., 2019a) proposes the dual branch to calculate spatial-wise and channel-wise attention simultaneously. To reduce the computaional cost of (Wang et al., 2018), lastest works study more compact mechanisms. (Zhu et al., 2019) introduces the asymmetric attention map derived from pyramid sampling. (Huang et al., 2019) proposes a criss-cross attention mechanism. (Li et al., 2019) formulates the attention mechanism as the problem of expectation maximization. All of them try to reduce the computational cost and memory usage, while maintaining the network performance.

The other category is modeling the contextual dependencies in graph interaction space. Compared with coordinate space, graph interaction space is much more efficient since the number of graph nodes is significantly smaller than the activation of feature map. Graph reasoning (Li & Gupta, 2018; Liang et al., 2018; Chen et al., 2019b; Zhang et al., 2019b; Wu et al., 2020) has been proved to be an effective way of capturing global dependencies between distant regions. SGR (Liang et al., 2018) is proposed to construct a graph from local features by voting, and fuses human knowledge prior to graph reasoning. GCU (Li & Gupta, 2018) utilizes the global distribution to construct the projection matrix during graph projection, promoting the reasoning ability beyond regular grid. GloRe (Chen et al., 2019b) aggregates a set of global feature in coordinate space, and then projects them into graph interaction space via weighted global pooling and broadcasting. LatentGNN (Zhang et al., 2019b) introduces a mixture of learnable low-rank matrices to capture context between graph nodes. Our method explores the contextual information during graph projection and reprojection, achieving more effective representation for graph reasoning.



(a) Two popular paradigms



(b) Descriptor Graph Projection

Figure 1: Illustration of two popular paradigms and our approach for projection matrix construction. Note that $C' < C$ for the consideration of efficiency.

3 Approach

In this section, we first revisit two popular paradigms of constructing the projection matrix, which is important for graph projection and reprojection. Then, we propose a novel Contextual Graph Reasoning method (CGR), consisting of Descriptor Graph Projection (DGP), Graph Convolution (GC) and Node Collaborative Reprojection (NCR). Finally, we present the Contextual Graph Reasoning Network (CGRNet) built on CGR.

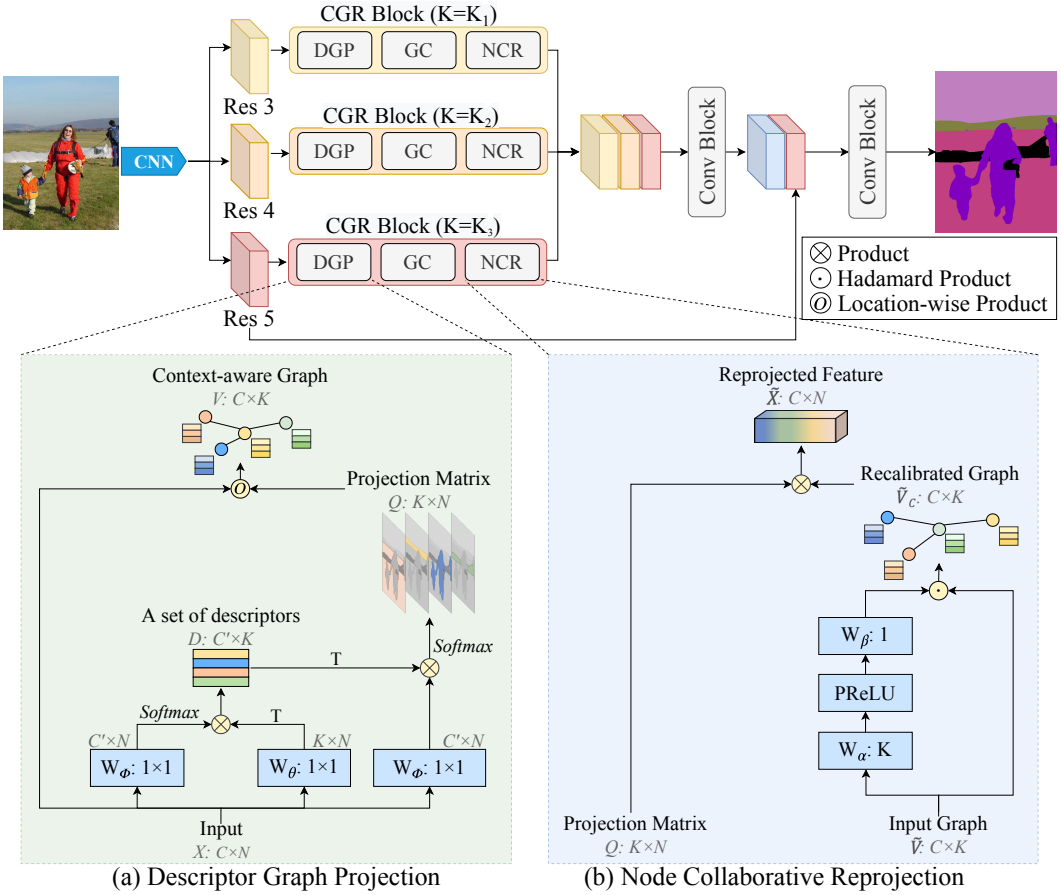
3.1 Projection Matrix Revisited

A standard graph can be represented as $G = \langle V, A \rangle$, where V and A denote graph nodes and adjacency matrix respectively. A nonparametric adjacency matrix is primarily adopted by recent graph reasoning works. Therefore, learning effective graph nodes from input feature is the key to improving the performance of graph reasoning. Projection matrix is adopted to project the feature into a graph and then reproject it back after reasoning. Recent works (Liang et al., 2018; Li & Gupta, 2018; Chen et al., 2019b; Zhang et al., 2019b; Wu et al., 2020) present two popular paradigms of constructing the projection matrix.

Given input feature $X = [x_0, ..., x_{N-1}] \in \mathbb{R}^{C \times N}$, both paradigms aim to obtain a projection matrix $Q = [q_0, ..., q_{K-1}] \in \mathbb{R}^{N \times K}$ for graph nodes $V = [v_0, ..., v_{K-1}] \in \mathbb{R}^{C \times K}$, where C, $N = W \times H$, and K denote the numbers of feature channel, activation and graph node individually. In the first paradigm, Q is calculated by element-wise operation with two parameters, namely visual code $W = [w_0, ..., w_{K-1}] \in \mathbb{R}^{C \times K}$ and scaling factor $\Sigma = [\sigma_0, ..., \sigma_{K-1}] \in \mathbb{R}^{C \times K}$. The formulation of the first paradigm can be illustrated as below:

$$q_i^k = \frac{exp(-\left\|(x_i - w_k)/\sigma_k\right\|^2/2)}{\sum_k exp(-\left\|(x_i - w_k)/\sigma_k\right\|^2/2)}, \tag{1}$$

where $q_i^k \in Q$ defines the mapping from i-th activation to k-th node. In the second paradigm, a point-wise convolution is employed to learn Q directly, reducing the computational cost of con-



(a) Descriptor Graph Projection (b) Node Collaborative Reprojection

Figure 2: Overview of the proposed Contextual Graph Reasoning Network (CGRNet). "DGP", "GC" and "NCR" denote Descriptor Graph Projection, Graph Convolution, and Node Collaborative Reprojection respectively. Best viewed in color.

structing the projection matrix:

$$Q = \theta(X; W_\theta), \tag{2}$$

where $\theta(\cdot)$ indicates the point-wise convolution layer, W_θ denotes the parameters of $\theta(\cdot)$. As can be seen from Figure 1, both paradigms derive the projection matrix with local feature, neglecting the global contextual information. Therefore, it is essential to explore the long-range dependencies for projection matrix construction, leading to context-aware relation between feature and graph.

3.2 Contextual Graph Reasoning (CGR)

In this section, we present the details of Descriptor Graph Projection, Reasoning by Graph Convolution and Node Collaborative Reprojection.

3.2.1 Descriptor Graph Projection (DGP)

Given input feature X after convolution layers, graph projection aims to transform feature vectors to a set of graph nodes. As mentioned above, existing methods fail to explore the holistic information, which is crucial for effective graph representation. Therefore, we propose a novel graph projection method based on a global set of descriptors $D = [d_0; ...; d_{K-1}] \in \mathbb{R}^{K \times C'}$, where $C' < C$ for the consideration of efficiency. We tactfully employ two maps to obtain the descriptors, which fully utilize the global contextual information. Specifically, we first employ a 1×1 convolution θ to compute the base map $B = [b_0, ..., b_{N-1}] \in \mathbb{R}^{C' \times N}$. Then we employ another 1×1 convolution

ϕ and normalization to compute the weight map $L = [l_0; ...; l_{K-1}] \in \mathbb{R}^{K \times N}$. This process can be represented as:

$$B = \theta(X; W_\theta), \qquad L' = \phi(X; W_\phi), \qquad l_{k,j} = \frac{e^{l'_{k,j}}}{\sum_j e^{l'_{k,j}}}, \tag{3}$$

where $l_{k,j}$ is the element of weight map L. Subsequently, we aggregate the feature of base map weighted by L, delivering the global descriptors:

$$d_k = l_k B^T = \sum_j l_{k,j} b_j^T. \tag{4}$$

In order to construct the projection matrix based on input feature and global descriptors, we first reduce the channels of input feature by a 1×1 convolution ψ, obtaining a compact feature $\hat{X} \in \mathbb{R}^{C' \times N}$ with the same channel dimension as D. This compact feature is used to generate the projection matrix as illustrated by

$$\hat{X} = \psi(X; W_\psi), \qquad q'_k = d_k \hat{X}^T = \sum_i d_{i,k} x_i^T, \qquad q_k = \frac{q'_k}{\|q'_k\|_2}, \tag{5}$$

where W_Ψ indicates the parameter of ψ. Finally, the graph representation V can be obtained by the input feature X and the projection matrix Q, which is formulated as

$$v'_k = \frac{1}{\sum_j q_k^j} \sum_j q_k^j x_j, \qquad v_k = \frac{v'_k}{\|v'_k\|_2}. \tag{6}$$

A graphical illustration of DGP is shown in Figure 2 (a).

3.2.2 Reasoning by Graph Convolution (GC)

We use graph convolution f_{gc} (Kipf & Welling, 2016) to perform global reasoning, aggregating contextual information in graph interaction space. For the purpose of efficiency, we only use one graph convolution layer to generate the evolved graph representation $\tilde{V}$:

$$\tilde{V} = \sigma(f_{gc}(V; W_{gc})) + V, \tag{7}$$

where $\sigma(\cdot)$ is the nonlinear activation function, W_{gc} is the parameters of graph convolution. Note that we also adopt Batch Normalization (Ioffe & Szegedy, 2015) after $\sigma(\cdot)$ to stabilize the training process.

3.2.3 Node Collaborative Reprojection (NCR)

Although graph convolution can aggregate the contextual information between nodes, it is still difficult for each node to exploit the global contextual information of the graph. Inspired by the self-attention mechanism (Vaswani et al., 2017), we propose a module named Node Collaborative Reprojection (NCR), which calibrates the evolved graph with channel-wise and spatial-wise attention. we first squeeze all nodes to obtain a C-dimensional vector by a convolution α:

$$\tilde{V}_\alpha = \sigma(\alpha(\tilde{V}; W_\alpha)), \tag{8}$$

where $\tilde{V}_\alpha \in \mathbb{R}^C$, $\sigma(\cdot)$ means PReLU, and W_α denotes the learnable convolutional kernel of α. This operation is similar to the squeeze operation in SENet (Hu et al., 2018) while we apply it to graph nodes. The squeezed feature of all nodes captures the global contextual information, which is essential to calibrate the evolved graph. Subsequently, we perform another convolution γ and normalization to expand the squeezed feature into scaling weights $S \in \mathbb{R}^{C \times K}$:

$$S' = \gamma(\tilde{V}_\alpha; W_\gamma), \qquad s_i^k = \frac{e^{s'_{i,k}}}{\sum_k e^{s'_{i,k}}}, \tag{9}$$

where s_i^k is the element of scaling weights S. Finally, S is used to calibrate the graph nodes and the calibrated graph is reprojected back to the coordinate space with Q^T as

$$\tilde{X} = (\tilde{V} \odot S) Q^T + X, \tag{10}$$

where $\odot$ denotes the hadamard product. Figure 2 (b) shows the detailed architecture of NCR.

Table 1: Comparison with the state-of-the-art approaches on PASCAL-Context, PASCAL-VOC 2012, ADE20K and COCO Stuff. "†" means the model has been pre-trained on COCO Stuff. "_" means no public results available. "⋆" means employing online hard example mining (Shrivastava et al., 2016). All methods employ ResNet-101 as the backbone.

Method	**mIoU**(%)			
	PASCAL-Context	PASCAL-VOC	ADE20K	COCO Stuff
PSPNet (Zhao et al., 2017)	47.8	82.6	43.29	-
EncNet (Zhang et al., 2018)	51.7	82.9	44.65	-
SGR† (Liang et al., 2018)	52.5	-	44.32	39.1
GCU (Li & Gupta, 2018)	-	-	44.81	-
DANet (Fu et al., 2019a)	52.6	82.6	-	39.7
CFNet (Zhang et al., 2019a)	54.0	84.2	44.89	-
APCNet (He et al., 2019)	54.7	84.2	45.38	-
ACNet⋆ (Fu et al., 2019b)	54.1	-	45.90	40.1
SPNet (Hou et al., 2020)	54.5	-	45.60	-
SpyGR (Li et al., 2020)	52.8	84.2	-	39.9
GINet (Wu et al., 2020)	54.9	-	45.54	40.6
CaC-Net (Liu et al., 2020)	55.4	85.1	46.12	-
OCRNet (Yuan et al., 2019)	54.8	-	45.28	39.5
CGRNet (Ours)	**56.5**	**85.7**	**47.19**	**41.1**

3.3 Framework of Contextual Graph Reasoning Network (CGRNet)

The CGR module can be taken as a plug-and-play component for convolutional neural networks. As an illustration of the usage, we design the Contextual Graph Reasoning Network (CGRNet). The whole framework of CGRNet is illustrated in Figure 2. We take ResNet (He et al., 2016) as our backbone following previous works (Fu et al., 2019a; Hou et al., 2020; Li et al., 2020; Liu et al., 2020). For the proposed CGR module, we deploy it on the last three stages, which can perform graph reasoning at multiple scales. After that, we concatenate the output features of CGR blocks and employ a 3×3 convolution layer to fuse them. The fused feature is concatenated with the the reduced feature, delivering the final feature for network prediction.

4 Experiments

The proposed Contextual Graph Reasoning module can be applied to various computer vision tasks. To demonstrate the effectiveness and generalization ability of our method, we conduct extensive experiments on four semantic segmentation benchmarks, including PASCAL-Context, PASCAL-VOC 2012, COCO Stuff and ADE20K (Mottaghi et al., 2014; Everingham et al., 2010; Caesar et al., 2018; Zhou et al., 2017), achieving the state-of-the-art performance. We also do a careful ablation study on semantic segmentation, giving a thorough analysis of our method. Besides, we evaluate our approach on instance segmentation and 2D human pose estimation. The results of 2D human pose estimation on COCO 2017 (Lin et al., 2014) and MPII (Andriluka et al., 2014) can be found in the Appendix.

4.1 Experiments on Semantic Segmentation

4.1.1 Implementation Details and Evaluation Metrics

During training, we use the poly learning rate scheduler $lr = base\_lr * (1 - \frac{iter}{total\_iter})^{0.9}$, where we set $base\_lr$ to 0.004 for ADE20K and 0.001 for others. SGD optimizer is adopted with weight decay 0.0001 and momentum 0.9. We also use the synchronized BN following existing methods (Fu et al., 2019a; Zhang et al., 2018) and train PASCAL-Context, PASCAL-VOC2012, COCO-Stuff, and ADE20K for 80, 50, 110, and 120 epochs respectively. The batch size for all datasets is set to 16 and input images are randomly cropped into 520×520. We use random flipping and scaling as the data augmentation to alleviate the problem of over-fitting (Zhang et al., 2018; Zhao et al., 2017).

Table 2: Ablation study of CGRNet. "CGP", "GC" and "NCR" denote Descriptor Graph Projection, Graph Convolution and Node Calibration Reprojection respectively. "MS" means Multi-Scale test.

Baseline	CGP	GC	NCR	MS	mIoU
✓					50.4
✓	✓				52.0
✓	✓	✓			52.2
✓	✓		✓		52.6
✓	✓	✓	✓		52.9
✓	✓	✓	✓	✓	**53.6**

Table 3: Ablation study of design choices. This study demonstrates the impact of node numbers and multi-stage features. $Res\#$ indicates the stage of backbone.

$Res\#3$	$Res\#4$	$Res\#5$	mIoU
16	8	4	51.8
32	16	8	**52.2**
64	32	16	52.1
✓	✓	✓	**52.2**
	✓	✓	51.7
		✓	51.3

During testing, We employ multi-scale test with input scaling factors [0.75, 1.0, 1.25, 1.5, 1.75, 2.0]. We adopt the mean Intersection-over-Union (mIoU) as the evaluation metric for semantic segmentation.

4.1.2 COMPARISON WITH STATE-OF-THE-ARTS

PASCAL-Context (Mottaghi et al., 2014) contains 4,998 and 5,105 images for training and validation, respectively. We report the mIoU on 60 categories (59 categories with the background) for evaluation, which is the same as previous works (He et al., 2019; Zhang et al., 2018). As shown in Table 1, Our result of 56.5% outperforms all previous methods based on graph reasoning (Liang et al., 2018; Wu et al., 2020) and non-local mechanism (Zhang et al., 2019a; Fu et al., 2019a; Yuan et al., 2019).

PASCAL-VOC2012 (Everingham et al., 2010) has 10,582 images for training, 1,449 images for validation, and 1,456 images for testing. We report results on 20 foreground object classes and one background class. We adopt ImageNet (Deng et al., 2009) pre-trained model and finetune our CGRNet on augmented training set for 80 epochs. Then, the model is finetuned on original trainval set for another 50 epochs. Finally, we evaluate our results on the official test server `http://host.robots.ox.ac.uk:8080`. As can been seen from Table 1, the result of 85.7% outperforms the start-of-the-art approaches on PASCAL-VOC 2012 test set.

ADE20K (Zhou et al., 2017) is a large scale scene parsing dataset including 25K images annotated with 150 categories, which are split into 20K training images, 2K validation images, and 3K test images. In Table 1, our method shows an excellent mIoU of 47.19%, which again is better than the state-of-the-art methods.

COCO-Stuff (Caesar et al., 2018) is a very challenging dataset which contains 9000 training images and 1000 test images, annotated with 171 object and stuff categories. Our result of 41.1% also surpasses all the existing methods, demonstrating the effectiveness of our work.

4.1.3 ABLATION STUDY

In this section, we conduct experiments with different settings to evaluate the performance of our approach. For the consideration of efficiency, we use ResNet-50 (He et al., 2016) as the backbone in ablation study. We also give a detailed comparison with other graph reasoning methods in terms of accuracy and efficiency.

Effectiveness of DGP and NCR: We conduct experiments on several variants of our approach to evaluate each component. As illustrated in Table 2, we have the following observations: firstly, compared with the baseline, DGP blocks bring a significant improvement of 1.6% (52.0% vs. 50.4%), which demonstrates that DGP blocks can explore contextual relation between feature and graph adequately. Secondly, there is an additional improvement of 0.7% with NCR blocks, which indicates it can further improve the capability of assembling global contextual information.

Ablation for Design Choices: Table 3 shows the impact of node numbers and multi-stage features. The numbers of nodes in the last three stages are set to 32, 16, and 8, which can achieve the best result of 52.2%. Too many nodes may aggregate features with large differences, resulting in semantic

Table 4: Comparisons of efficiency and accuracy with other graph reasoning methods. "Params", "Mem" and "FPS" represent the Parameters (M), the extra memory (MB) and the inference speed respectively. We evaluate these blocks with 520×520 input size on the same server for a fair comparison.

Method	**Params**	**GFLOPs**	**Mem**	**FPS**	**mIoU**
Baseline	-	-	-	-	50.4
+ GCU (Li & Gupta, 2018)	11.2	93.5	366.7	95.7	51.5
+ GloRe (Chen et al., 2019b)	8.1	68.6	208.3	161.9	51.8
+ CGR	8.5	70.9	218.2	133.8	**52.9**

Table 5: Results of different graph reasoning methods on COCO 2017 validation set for object detection (AP^b) and instance segmentation (AP^m). All results are based on Mask R-CNN with ResNet-50 and ResNet-101 as the backbone.

Method (ResNet-50)	$\mathbf{AP}^b$	$\mathbf{AP}^m$	**Method** (ResNet-101)	$\mathbf{AP}^b$	$\mathbf{AP}^m$
Mask RCNN$*$	37.2	34.0	Mask RCNN	39.3	35.8
+ GCU (Li & Gupta, 2018)	37.2	33.9	+ GCU (Li & Gupta, 2018)	39.4	35.7
+ GloRe (Chen et al., 2019b)	37.8	34.6	+ GloRe (Chen et al., 2019b)	39.8	36.2
+ CGR	**38.2**	**34.8**	+ CGR	**40.3**	**36.6**

inconsistency. Too few nodes may not adequately capture the global contextual information. As shown in Table 3, it is also crucial to adopt multi-stage features for our method. Note that we do not fuse the features of $Res\#2$ since its spatial resolution is too large, resulting in substantial memory costs.

Comparisons with Related Graph Reasoning Methods: As shown in Table 4, we compare with two graph reasoning methods. Our CGR blocks achieve the best mIoU of 52.9% with relatively less memory, parameters, and GFLOPs. Compared with GCU and GloRe (Li & Gupta, 2018; Chen et al., 2019b), we tactfully explore the global contextual information for graph projection and reprojection, boosting the performance with negiligable additional storage and computational cost.

4.2 Experiments on Instance Segmentation

To further evaluate the generalization ability of our method, we also conduct experiments on instance segmentation using Mask R-CNN (He et al., 2017) as the baseline. We adopt ResNet-50/101 with FPN (He et al., 2016; Lin et al., 2017) as the backbone and add three CGR blocks in the last three stages. We implement our method based on mmdetection (Chen et al., 2019a) and report the results in terms of box AP and mask AP in Table 5. The results demonstrate that our method consistently improves the baselines in both metrics. Meanwhile, CGR also consistently outperforms existing graph reasoning methods like GCU and GloRe. Note that GCU even slightly declines the performance of baseline under the same training protocol.

5 Conclusion

In this paper, we propose a novel graph reasoning method named Contextual Graph Reasoning (CGR), exploring the long-range dependencies for effective graph representation. The core contributions are Descriptor Graph Projection (DGP) and Node Calibration Reprojection (NCR). DGP learns a global set of descriptors for projection matrix construction, capturing the contextual information for graph nodes. NCR calibrates the features of graph nodes with the self-attention mechanism, further exploiting the benefit of contextual information. We conduct extensive experiments on semantic segmentation, instance segmentation and 2D human pose estimation. Our method consistently achieves remarkable improvements over the state-of-the-art graph reasoning methods, demonstrating the effectiveness and generalization ability of our work.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 9157–9166, 2019.

Wonmin Byeon, Thomas M Breuel, Federico Raue, and Marcus Liwicki. Scene labeling with lstm recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3547–3555, 2015.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017a.

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

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018a.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7103–7112, 2018b.

Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 433–442, 2019b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, 2019a.

Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE international conference on computer vision*, pp. 6748–6757, 2019b.

Raghudeep Gadde, Varun Jampani, Martin Kiefel, Daniel Kappler, and Peter V Gehler. Superpixel convolutional networks using bilateral inceptions. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7519–7528, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. *arXiv preprint arXiv:2003.13328*, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 603–612, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

L'ubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 739–746. IEEE, 2009.

Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9167–9176, 2019.

Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. *arXiv preprint arXiv:2003.10211*, 2020.

Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems*, pp. 9225–9235, 2018.

Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *European Conference on Computer Vision*, pp. 125–143. Springer, 2016.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. In *Advances in Neural Information Processing Systems*, pp. 1853–1863, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Jianbo Liu, Junjun He, Jimmy S Ren, Yu Qiao, and Hongsheng Li. Learning to predict context-adaptive convolution for semantic segmentation. *arXiv preprint arXiv:2004.08222*, 2020.

Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE international conference on computer vision*, pp. 1377–1385, 2015.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1573–1581, 2015.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

Tianyi Wu, Yu Lu, Yu Zhu, Chuang Zhang, Ming Wu, ZhanYu Ma, and Guodong Guo. Ginet: Graph interaction network for scene parsing. In *Proceedings of the European conference on computer vision (ECCV)*, 2020.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 548–557, 2019a.

Songyang Zhang, Shipeng Yan, and Xuming He. Latentgnn: Learning efficient non-local relations for visual recognition. *arXiv preprint arXiv:1905.11634*, 2019b.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1529–1537, 2015.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 593–602, 2019.

A Appendix

A.1 Experiments on 2D Human Pose Estimation

To evaluate the generalization ability of our method, we conduct detailed experiments on 2D Human Pose Estimation. We compare our method with two state-of-the-art graph reasoning methods (Li & Gupta, 2018; Chen et al., 2019b) and all experiments are based on the same baseline (Xiao et al., 2018). We benchmark all the methods on COCO keypoints detection dataset (Lin et al., 2014) and MPII Human Pose dataset (Andriluka et al., 2014).

A.1.1 Dataset and Evaluation Metric

The COCO keypoints detection dataset (Lin et al., 2014) contains over 200k images and 250k person instances with manually annotated keypoints. We train the models on COCO *train2017* set, including more than 118k images and 150k personal instances. We evaluate the methods on COCO *val2017* set, which contains around 5k images. The Object Keypoint Similarity (OKS) is used for evaluation. We use the mean Average Precision (AP) and Average Recall (AR) over 10 OKS thresholds as the main metric. The OKS is calculated as the distance between predicted points and ground truth points, normalized by the scale of the person.

The MPII Human Pose dataset (Andriluka et al., 2014) consists of images obtained from a variety of real-world activities with full-body pose annotation. There are around 25K images and 40K subjects. We train the models on MPII training set, including around 28k subjects. we evaluate the methods on validation set. The data augmentation and training strategy are the same as COCO keypoints detection dataset, except that the input image is cropped to 256×256 for fair comparisons with other graph-based methods. We use PCKh score as the evaluation metric. A joint is correct if it falls within αl pixels of the ground-truth position, where α is a constant and l is the head size that corresponds to 60% of the diagonal length of the ground-truth head bounding box. The PCKh@0.5 (α = 0.5) score is reported.

A.1.2 IMPLEMENTATION DETAILS

We employ ResNet-50/101 as the backbone and add three CGR modules in the last three stages respectively. The numbers of nodes in three CGR blocks are set to 6, 4, and 2. We crop and resize the input images to two fixed resolutions, 256×192 for COCO dataset and 256×256 for MPII dataset. The optimizer, batch size, learning rate scheduler and data augmentations follow the baseline (Xiao et al., 2018). The backbone network is initialized by weights pre-trained on ImageNet. The initial learning rate is 0.001, and drops to 0.0001 at the 90-th epoch and 0.00001 at the 120-th epoch. The training process converges within 140 epochs.

Table 6: Comparisons of different graph reasoning methods on COCO *val2017* set for 2D pose estimation (without flip test).

Method	**Backbone**	**Input Size**	**AP**	**AR**
Simple baseline (Xiao et al., 2018)	ResNet-50	256×192	70.4	73.5
+ GloRe (Chen et al., 2019b)	ResNet-50	256×192	70.0	73.3
+ GCU (Li & Gupta, 2018)	ResNet-50	256×192	72.0	75.3
+ CGR	ResNet-50	256×192	**72.5**	**75.7**
Simple baseline (Xiao et al., 2018)	ResNet-101	256×192	72.0	75.3
+ GloRe (Chen et al., 2019b)	ResNet-101	256×192	71.1	74.6
+ GCU (Li & Gupta, 2018)	ResNet-101	256×192	72.8	76.0
+ CGR	ResNet-101	256×192	**73.3**	**76.3**

Table 7: Comparisons of different graph reasoning methods on MPII validation set for 2D pose estimation (without flip test).

Method	**Backbone**	**Input Size**	**Mean**
Simple baseline (Xiao et al., 2018)	ResNet-50	256×256	87.583
+ GloRe (Chen et al., 2019b)	ResNet-50	256×256	86.235
+ GCU (Li & Gupta, 2018)	ResNet-50	256×256	88.020
+ CGR	ResNet-50	256×256	**88.322**
Simple baseline (Xiao et al., 2018)	ResNet-101	256×256	88.374
+ GloRe (Chen et al., 2019b)	ResNet-101	256×256	86.656
+ GCU (Li & Gupta, 2018)	ResNet-101	256×256	88.660
+ CGR	ResNet-101	256×256	**88.798**

A.1.3 EXPERIMENTAL RESULTS

We compare three graph reasoning methods, namely GCU (Li & Gupta, 2018), GloRe (Chen et al., 2019b) and our CGR. As shown in Table 6, CGR outperforms GCU by 0.5 AP (72.5 vs. 72.0) and achieves 2.1 AP improvement over the baseline (72.5 vs. 70.4) with the backbone of ResNet-50 on COCO. GloRe declines the baseline under the same training protocol while our CGR consistently boosts the performance. The results on ResNet-101 and MPII show the same tendency, which demonstrates the effectiveness and generalization ability of our method.

A.2 QUALITATIVE RESULTS

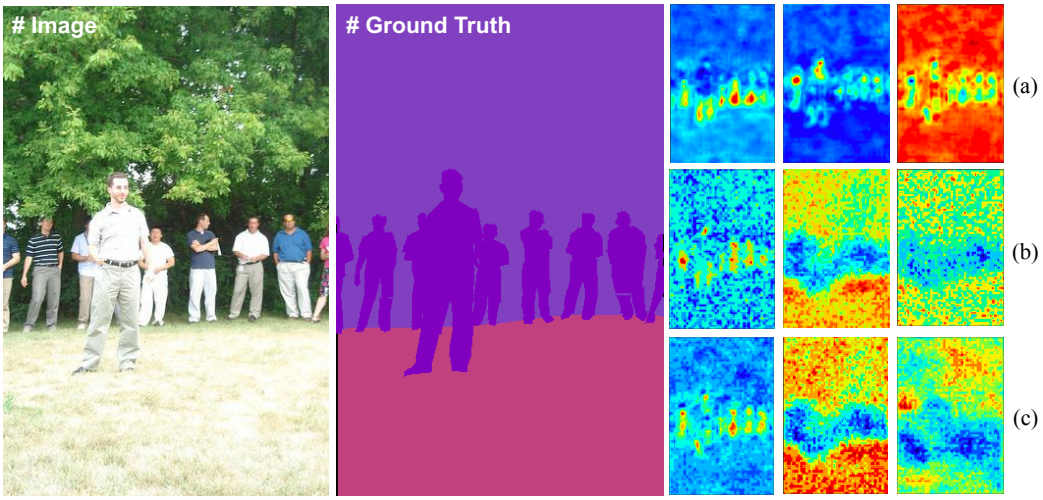


Figure 3: The visualization of projection matrices and feature maps (w/o. *vs* w/i. NCR module) on PASCAL-Context. (a) indicates projection matrices. (b) indicates feature maps without NCR module. (c) indicates feature maps within NCR module.

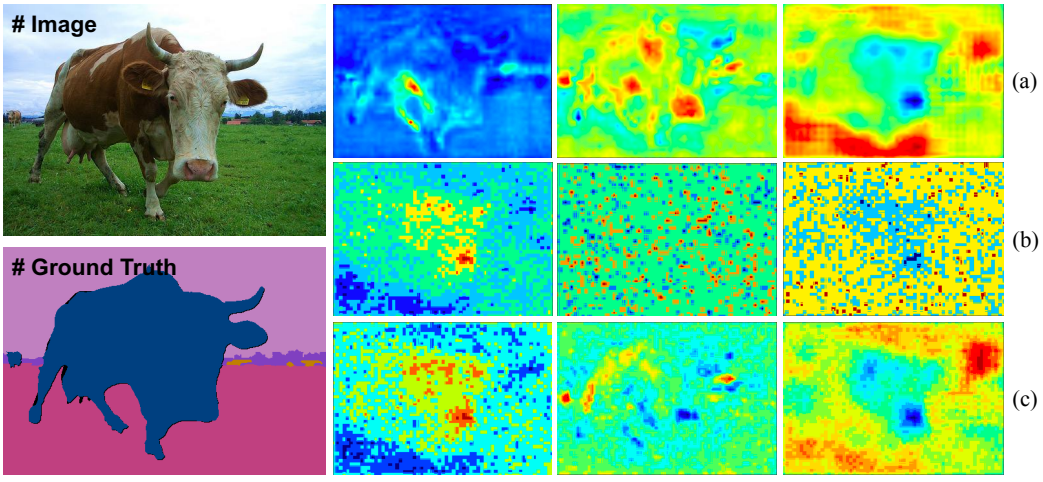


Figure 4: The visualization of projection matrices and feature maps (w/o. *vs* w/i. NCR module) on PASCAL-Context. (a) indicates projection matrices. (b) indicates feature maps without NCR module. (c) indicates feature maps within NCR module.

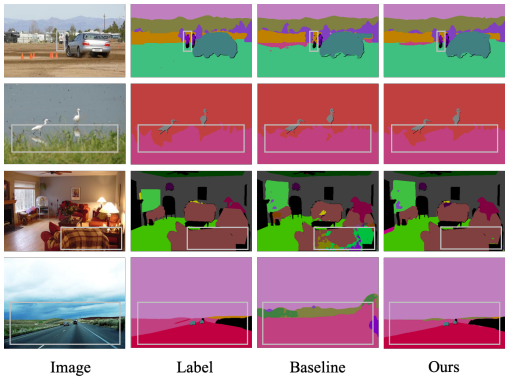


Figure 5: The visualization of PASCAL-Context validation set.

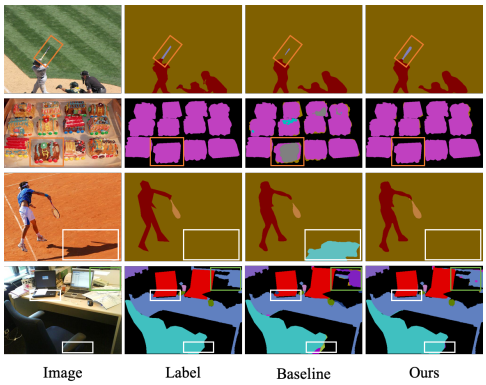


Figure 6: The visualization of COCO-Stuff test set.

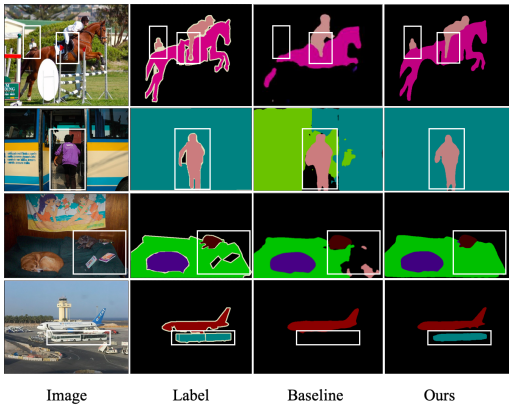


Figure 7: The visualization of PASCAL-VOC 2012 validation set.

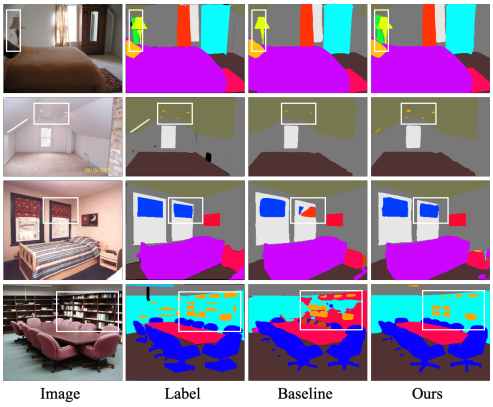


Figure 8: The visualization of ADE20K validation set.

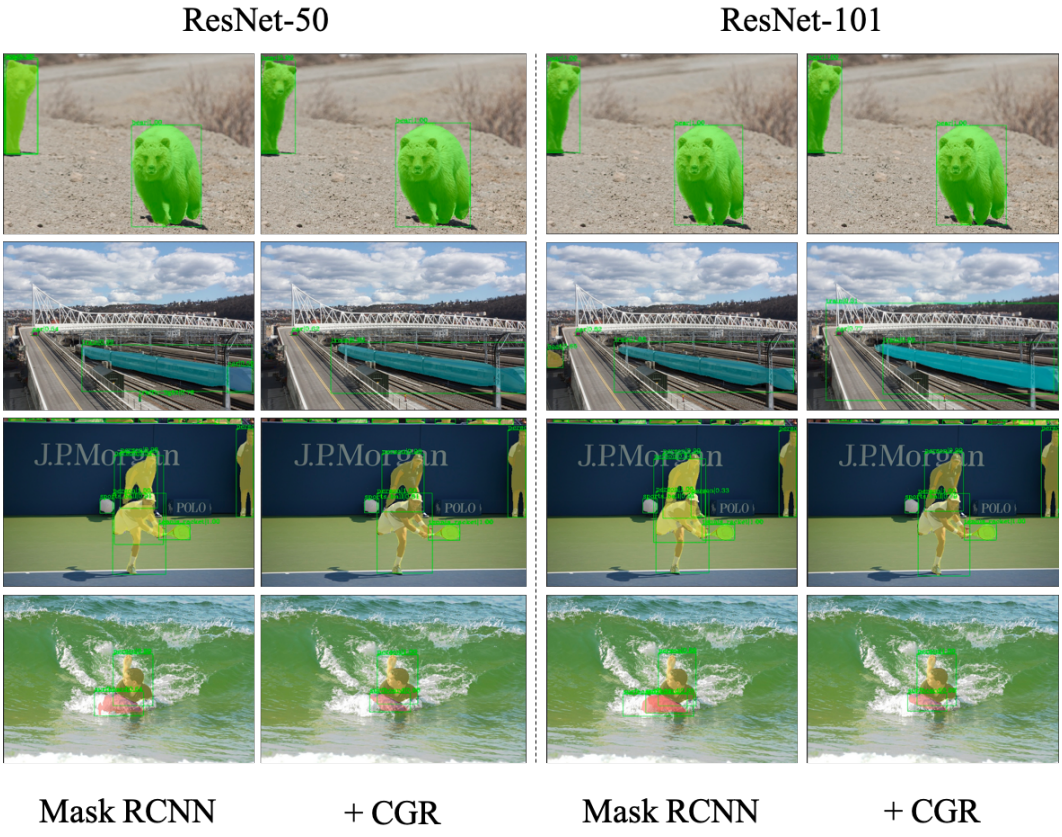


Figure 9: The visualization of object detection and instance segmentation with ResNet-50/101 as the backbone.

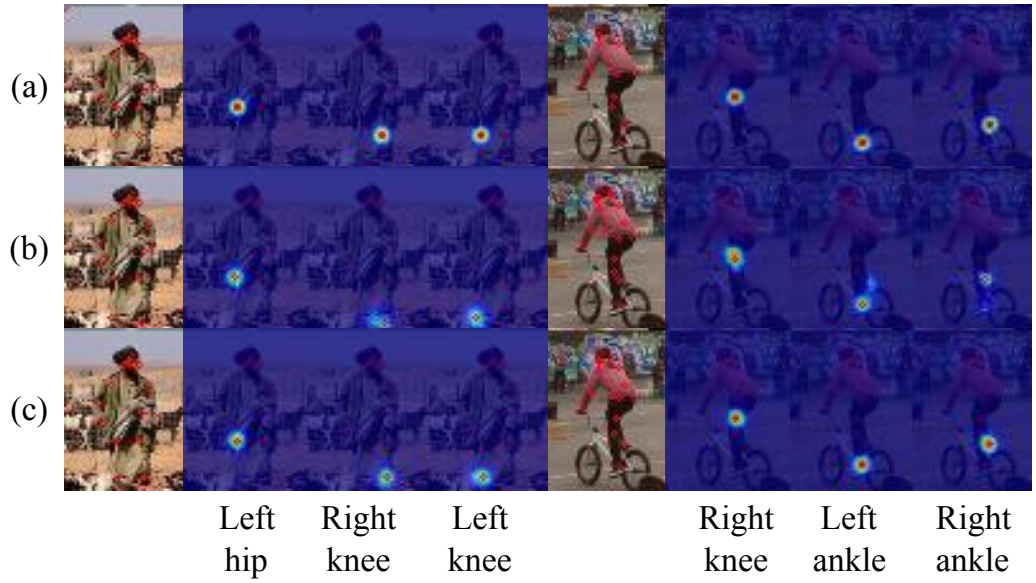


Figure 10: The visualization of 2D Human Pose Estimation on COCO 2017 validation set. (a) denotes "Ground Truth". (b) denotes "Baseline". (c) denotes "Baseline + CGR".

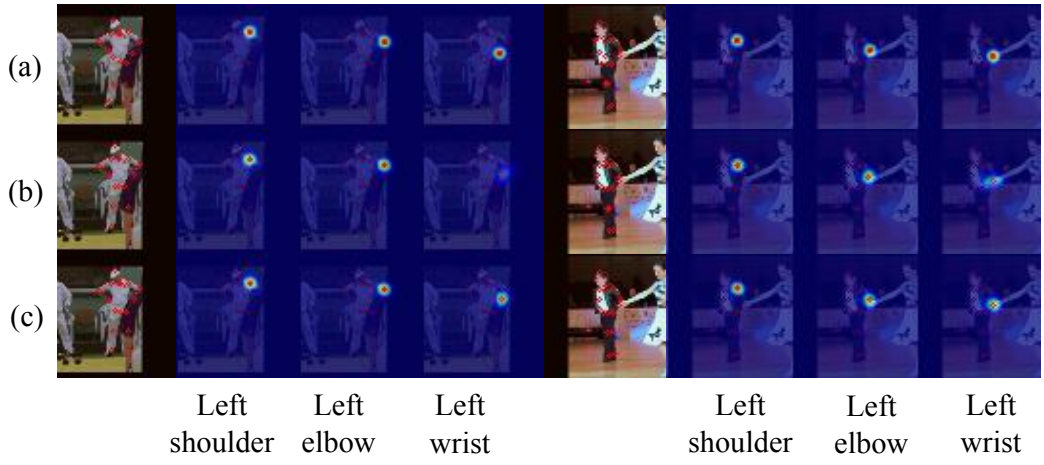


Figure 11: The visualization of 2D Human Pose Estimation on MPII validation set. (a) denotes "Ground Truth". (b) denotes "Baseline". (c) denotes "Baseline + CGR".