

GINet: Graph Interaction Network for Scene Parsing

Tianyi Wu^{1,2*}, Yu Lu^{3*}, Yu Zhu^{1,2},
Chuang Zhang³, MingWu³, Zhanyu Ma³, and Guodong Guo^{1,2*}

¹ Institute of Deep Learning, Baidu Research, Beijing, China
{wutianyi01, zhuyu05, guogudong01}@baidu.com

² National Engineering Laboratory for Deep Learning Technology and Application,
Beijing, China

³ Beijing University of Posts and Telecommunications, Beijing, China
{aniki, zhangchuang, wuming, mazhanyu}@bupt.edu.cn

Abstract. Recently, context reasoning using image regions beyond local convolution has shown great potential for scene parsing. In this work, we explore how to incorporate the linguistic knowledge to promote context reasoning over image regions by proposing a Graph Interaction unit (GI unit) and a Semantic Context Loss (SC-loss). The GI unit is capable of enhancing feature representations of convolution networks over high-level semantics and learning the semantic coherency adaptively to each sample. Specifically, the dataset-based linguistic knowledge is first incorporated in the GI unit to promote context reasoning over the visual graph, then the evolved representations of the visual graph are mapped to each local representation to enhance the discriminated capability for scene parsing. GI unit is further improved by the SC-loss to enhance the semantic representations over the exemplar-based semantic graph. We perform full ablation studies to demonstrate the effectiveness of each component in our approach. Particularly, the proposed GINet outperforms the state-of-the-art approaches on the popular benchmarks, including Pascal-Context and COCO Stuff.

Keywords: Scene Parsing, Context Reasoning, Graph Interaction

1 Introduction

Scene parsing is a fundamental and challenging task with great potential values in various applications, such as robotic sensing and image editing. It aims at classifying each pixel in an image to a specified semantic category, including objects (*e.g.*, bicycle, car, people) and stuff (*e.g.*, road, bench, sky). Modeling context information is essential for scene understanding [2, 41, 3]. Since the work by Long [31] with fully convolutional networks (FCN), it has attracted more and more attention for context modeling in semantic segmentation or scene parsing.

* Equal contribution * Corresponding author

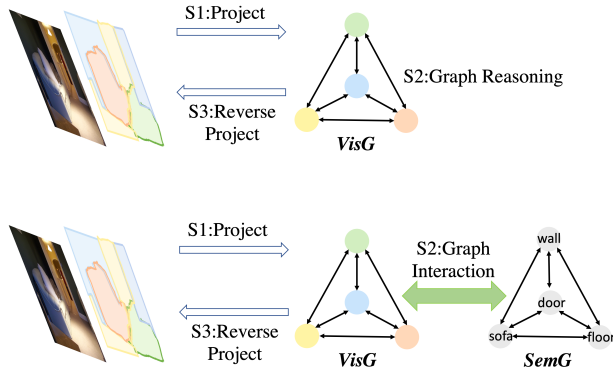


Fig. 1. Comparison of different contextual reasoning frameworks. *Top:* Contextual reasoning over the visual graph. *Bottom:* Our proposed Graph Interaction Network (GINet). Note: *VisG*: visual graph. *SemG*: semantic graph.

Early works are some approaches that lie in the stack of local convolutions to capture the context information. Several works employed dilation convolution [6, 51, 7–9, 48, 47], kronecker convolution [47] and pooling operations [30, 57] to obtain a wider context information. Recent works [53, 15, 55] introduced non-local operations [45] to integrate the local feature with their contextual dependencies adaptively to capture richer contextual information. Later, several approaches [21, 24, 61] were proposed to reduce the computation of non-local operations. More recently, using image regions for context reasoning [10, 25, 27, 56] has shown great potential for scene parsing. These methods were proposed to learn a graph representation from visual features, where the vertices in the graph define clusters of pixels (“region”), and edges indicate the similarity or relation between these regions in the feature space. In this way, contextual reasoning can be performed in the interaction graph space, then the evolved graph is projected back to the original space to enhance the local representations for scene parsing.

In this paper, instead of solely performing context reasoning over the visual graph representation for 2D input images or visual features (as shown in the top of Figure 1), we seek to incorporate linguistic knowledge, such as linguistic correlation and label dependency, to share the external semantic information across locations that can promote context reasoning over the visual graph. Specifically, we propose a Graph Interaction unit (GI unit), which first incorporates the dataset-based linguistic knowledge into feature representation over the visual graph, and re-projects the evolved representations of the visual graph back into each location representation for enhancing the discriminative capability (as shown in the bottom of Figure 1). Intuitively, the external knowledge is modeled as a semantic graph which is formed as vertices with linguistic entities (e.g., cup, table and desk) and edges with entity relationships (e.g., semantic hierarchy, concurrence and spatial interactions). GI unit shows the interaction between

the visual and semantic graph. Furthermore, we introduce a Semantic Context Loss, which aims at learning an exemplar-based semantic graph to better represents the sample adaptively, where the categories that appear in the scene are emphasized while those do not appear in the scene are suppressed. Details of the proposed method are presented in Section 3.

The most relevant works to our approach are [10, 27, 25]. Liang [27] proposed a Symbolic Graph Reasoning layer to perform reasoning over a group of symbolic nodes. The SGR explores how to harness various external human knowledge for endowing the networks with the capability of semantic global reasoning. In contrast, our method explores how to incorporate a dataset-based linguistic knowledge to promote context reasoning over image regions. Li [25] proposed a Graph Convolutional Unit to project a 2D feature map into a sample-dependent graph structure by assigning pixels to the vertices of the graph and learning a primitive grouping of scene components. Chen [10] introduced the Global Reasoning unit for reasoning globally, which projects information from the coordinate space to nodes in an interactive space graph to directly reason over globally-aware discriminative features. Different from these approaches, we propose to reason over the visual graph and the prior semantic graph. The semantic graph is employed to promote contextual reasoning and lead the generation of the exemplar-based semantic graph from the visual graph.

We conduct extensive experiments on different challenging datasets to validate the advantages of the proposed GI unit and SC-loss for scene parsing. Meanwhile, ablation studies are performed to demonstrate the effectiveness of each component in our approach. Experimental results are shown in Section 4.

The main contributions of this work include:

- A novel Graph Interaction unit (GI unit) is proposed for contextual modeling, which incorporates the dataset-based linguistic knowledge for promoting context reasoning over the visual graph. Moreover, it learns an exemplar-based semantic graph as well.
- A Semantic Context Loss (SC-loss) is proposed to regularize the training procedure in our approach, which emphasizes the categories that appear in the scene and suppresses those do not appear in the scene.
- A Graph Interaction Network (GINet) is developed, based on the proposed GI unit and SC-loss for scene parsing; It provides significant gains in performance over the state-of-the-art approaches on Pascal-Context [33] and COCO Stuff [4], and achieves a competitive performance on ADE20K dataset [59].

2 Related work

In this section, we briefly overview the recent progress in contextual modeling for scene parsing. They can be mainly divided into two categories based on whether graph reasoning is considered.

There are several model variants of FCN [31] proposed to exploit the contextual information. Some methods [14, 29, 44, 7, 8, 53, 43, 57, 42, 52] were proposed

to learn the multi-scale contextual information. DeepLabv2 [7] and DeepLabv3 [8] utilized an atrous spatial pyramid pooling to capture contextual information, which consists of parallel dilation convolutions with different dilation rates. TKCN [47] introduced a tree-structured feature aggregation module for encoding hierarchical contextual information. The pyramid pooling module is proposed by PSPNet [57] to collect the effective contextual prior, containing information on different scales. Moreover, the encoder-decoder structures [60, 5, 1, 36, 34] based on UNet [37] fuse the high-level and mid-level features to obtain context information. DeepLabV3+ [9] combines the properties of the above two methods that add a decoder upon DeepLabV3 to help model obtain multi-level contextual information and preserve spatial information. Differently, CGNet [49] proposed a Context Guided block for learning the joint representation of both local features and surrounding context. In addition, inspired by ParseNet [30], a global scene context was utilized in some methods [50, 58] by introducing a global context branch in the network. EncNet [54] introduced Encoding Module to capture the global semantic context and predict scaling factors to selectively highlight feature maps. Recently, there were many efforts [15, 53, 55] to break the local limitations of the convolution operators by introducing the Non-local block [45] into the feature representation learning to capture spatial context information. Furthermore, some methods [24, 61, 21] proposed to reduce the computational complexity of Non-Local operations. More recently, SPGNet [11] proposed a Semantic Prediction Guidance module which learns to re-weight the local features through the guidance from pixel-wise semantic prediction.

Some other methods introduced a graph propagation mechanism into the CNN network to capture a more extensive range of information. GCU [25] got inspiration from region-based recognition and presented a graph-based representation on semantic segmentation and object detection tasks. GloRe [10] and LatenGNN [56] performed a global relation reasoning by aggregating features with similar semantics to an interactive space. SGR [27] extracted the representation nodes of each category from the features and use external knowledge structures to reason about the relationship between categories. These methods have typical projection, reasoning, and back projection steps. Based on these steps, our approach further promotes graph reasoning by incorporating semantic knowledge. Finally, Graphonomy and GraphML [17, 18] proposed to propagate graph features of different datasets to unify the human parsing task. However, our approach explores the correlation between visual and semantic graph to facilitates the context modeling capabilities of the model.

3 Approach

In this section, we first introduce the framework of the proposed Graph Interaction Network (GINet). Then we present the design of the Graph Interaction unit (GI unit) in details. Finally, we give a detailed description of the proposed Semantic Context loss (SC-loss) for the GINet.

3.1 Framework of Graph Interaction Network (GINet)

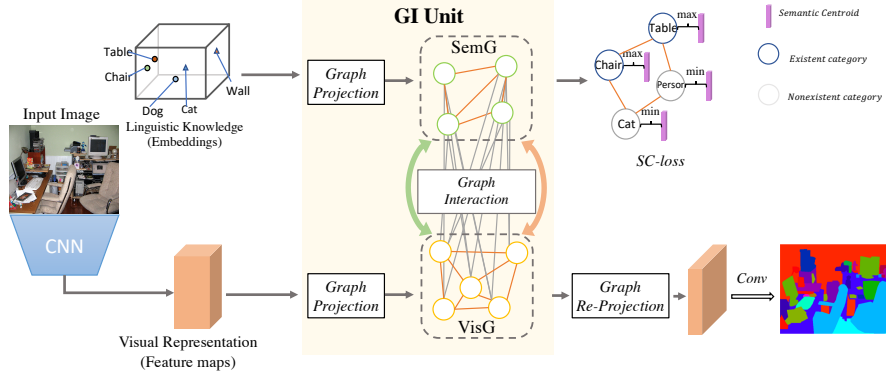


Fig. 2. The overall framework of the proposed Graph Interaction Network (GINet). (Best viewed in color.)

Different from previous methods that only perform contextual reasoning over the visual graph built on visual features [25, 10], our GINet facilitates the graph reasoning by incorporating semantic knowledge to enhance the visual representations. The proposed framework is illustrated in Figure 2. Firstly, we adopted a pre-trained ResNet [20] as the backbone network, where visual features can be extracted given an input 2D image. Meanwhile, the dataset-based linguistic knowledge can be extracted in the form of categorical entities (classes), which is fed into word embedding (e.g., GloVe [35]) to achieve semantic representations. Secondly, visual features and semantic embedding representations are passed by the graph projection operations in the proposed GI unit to construct two graphs, respectively. A detailed definitions of graph projection operations are presented in Section 3.2. Accordingly, one graph that encodes dependencies between visual areas is built over visual features, where nodes indicate visual regions and edges represent the similarity or relation between those regions. The other graph is built over the dataset-dependent categories (represented by word embeddings), which encodes the linguistic correlation and label dependency. Next, a graph interaction operation is processed in the GI unit, where the semantic graph is employed to promote contextual reasoning over the visual graph and guide the generation of the exemplar-based semantic graph extracted from the visual graph. Then, the evolved visual graph generated by the GI unit is passed by Graph Re-projection operation for enhancing the discriminative ability for each local visual representation, while the semantic graph is updated and constrained by the Semantic Context loss during the training phase. Finally, we employ an 1×1 Conv followed by a simple bilinear upsampling to obtain the parsing results.

3.2 Graph Interaction Unit

The goal of the proposed GI unit is to incorporate dataset-based linguistic knowledge for promoting the local representations. First, the GI unit takes visual and semantic representations as inputs, conduct contextual reasoning by generating a visual graph and a semantic graph. Second, a graph interaction is performed between the two graphs to evolve node features by the guidance of similarity of visual nodes and semantic nodes.

Graph Construction:

The first step is to define the projection that maps the original visual and semantic features to an interaction space. Formally, given the visual feature maps $X \in \mathbb{R}^{L \times C}$, where $L = H \times W$, H and W indicate the height and width of the feature map, and C is the channel dimensions. We aim to construct a visual graph representation $P \in \mathbb{R}^{N \times D}$, where N is the number of nodes in the visual graph and D is the desired feature dimension for each node. Inspired by the works [26, 17], we introduce a transformation matrix $Z \in \mathbb{R}^{N \times L}$ that projects the local presentation X to a high-level graph representation P , which can be computed as follows:

$$P = ZXW, \quad (1)$$

where $W \in \mathbb{R}^{C \times D}$ is introduced as trainable parameters to convert the feature dimension from C to D , and Z adaptively aggregates local features to a node in the visual graph.

Next, we define a dataset-dependent semantic graph. Particularly, we aim to build a semantic graph presentation $S \in \mathbb{R}^{M \times D}$ over the object categories of a specific dataset, to encode the linguistic correlation and label dependency. M denotes the number of nodes, which is equal to the number of categorical entities (classes) in the dataset. D is the feature dimension for each node in the semantic graph. Specifically, we first use the off-the-shelf word vectors [35] to get semantic representation $l_i \in \mathbb{R}^K$ for each category $i \in \{0, 1, \dots, M-1\}$, $K = 300$. Then, a MLP layer is employed to adjust the linguistic embedding to suit the reasoning with the visual graph. This transformation process can be formulated as follow:

$$S_i = MLP(l_i), i \in \{0, 1, \dots, M-1\}, \quad (2)$$

where S_i represents node features for each category.

Graph Interaction: Next, we present how our GI unit incorporates dataset-based linguistic knowledge for promoting context reasoning and extracting the exemplar-based semantic graph from the visual graph. For simplicity, we abbreviate the visual graph and semantic graph as VisG and SemG, respectively. We first evolve both graphs separately and then perform the interaction between graphs. Then SemG and VisG attentively propagate information iteratively, including 1) Semantic to Visual (S2V), and 2) Visual to Semantic (V2S).

Specifically, we first perform graph convolution [23] on the VisG to get evolved graph representation \tilde{P} that is suitable for interacting with the SemG. This process can be formulated as follows:

$$\tilde{P} = f((A_v + I)PW_v), \quad (3)$$

where the adjacency matrix $A_v \in \mathbb{R}^{N \times N}$ is randomly initialized and updated by the gradient descent; I is an identity matrix; $W_v \in \mathbb{R}^{D \times D}$ are trainable parameters; and f is a nonlinear activation function. Through reasoning over the VisG, we can update node representation to capture more visual context information and to interact with SemG. Next we perform similar graph convolution [23] on the SemG for adjusting the representation of SemG, according to:

$$\tilde{S} = f((A_s + I)SW_s), \quad (4)$$

where \tilde{S} indicates an updated the graph representation of SemG; $A_s \in \mathbb{R}^{M \times M}$ is a learnable adjacency matrix or co-occurrence matrix that represents connections between semantic correlation or label dependency; $W_s \in \mathbb{R}^{D \times D}$ are trainable parameters. By performing a propagation of feature information from neighboring nodes, we can improve the representation of each semantic node.

In S2V step: we utilize the evolved SemG to promote contextual reasoning over the VisG \tilde{P} . Specifically, to explore the relationship between two nodes from VisG and SemG, we compute their feature similarity as a guidance matrix $G^{s2v} \in \mathbb{R}^{N \times M}$. For one node $\tilde{p}_i \in \mathbb{R}^D$ in the VisG \tilde{P} and one node $\tilde{s}_j \in \mathbb{R}^D$ in the SemG, we can compute the guide information $G_{i,j}^{s2v}$ that represents the assignment weight of the node \tilde{s}_j in SemG to the node \tilde{p}_i in VisG as follows:

$$G_{i,j}^{s2v} = \frac{\exp(W_p \tilde{p}_i \cdot W_s \tilde{s}_j)}{\sum_{m=1}^M \exp(W_p \tilde{p}_i \cdot W_s \tilde{s}_m)}, \quad (5)$$

where $i \in \{1, \dots, N\}, j \in \{1, \dots, M\}$ and $W_p \in \mathbb{R}^{D/2 \times D}$ and $W_s \in \mathbb{R}^{D/2 \times D}$ are a learnable matrix to further reduce the feature dimension. After obtaining the guidance matrix G^{s2v} , we can distill information from SemG to enhance the representation of VisG, according to :

$$P_o = \tilde{P} + \beta_{s2v} G^{s2v} \tilde{S} W_{s2v}, \quad (6)$$

where $W_{s2v} \in \mathbb{R}^{D \times D}$ is a trainable weight matrix, $\beta_{s2v} \in \mathbb{R}^N$ is a learnable vector with zero initialization and can be updated by a standard gradient decent. We use a simple sum to melt information from graphs, which may be alternatively replaced by other commutative operators such as mean, max, or concatenate. With the help of the guidance matrix G_{s2v} , we effectively constructed the correlation between visual regions and semantic concepts, and incorporate corresponding semantic features into the visual node representation.

In V2S step: we adopt a similar method elaborated in Equation(5) to obtain the guidance matrix $G_{v2s} \in \mathbb{R}^{M \times N}$. Formally, the guide information $G_{i,j}^{v2s}$ that can be calculated as follows:

$$G_{i,j}^{v2s} = \frac{\exp(W_s \tilde{s}_i \cdot W_p \tilde{p}_j)}{\sum_{n=1}^N \exp(W_s \tilde{s}_i \cdot W_p \tilde{p}_n)}, \quad (7)$$

where $i \in \{1, \dots, M\}, j \in \{1, \dots, N\}$. After getting the guidance matrix G^{v2s} , we update the graph representation of the SemG for generating the exemplar-based SemG according to:

$$S_o = \beta_{v2s} \tilde{S} + G^{v2s} \tilde{P} W_{v2s}, \quad (8)$$

where $W_{v2s} \in \mathbb{R}^{D \times D}$ is a trainable weight matrix, $\beta_{v2s} \in \mathbb{R}^M$ is a learnable vector and initialized by zeros. We extract the exemplar-based semantic graph from VisG with the guidance matrix G_{v2s} . By combining the S2V and V2S steps, the proposed GI unit enables the whole model to learn more discriminative features for performing fine pixel-wise classification and generate a semantic graph for each input image.

Unit outputs: The GI unit has two outputs, one is the exemplar-based SemG, which are described in detail in section 3.3, and the other is the VisG enhanced by semantic information. The evolved node representation of VisG can be used to enhance the discriminative ability of each pixel feature further. As previous methods [26, 25], we reuse projection matrix Z to reverse project the P_o to 2D pixel features. Formally, Given node features $P_o \in \mathbb{R}^{N \times D}$ of the VisG, the reverse projection (or Graph Re-Projection) can be formulated as follows:

$$\tilde{X} = Z^T P_o W_o + X, \quad (9)$$

where $W_o \in \mathbb{R}^{D \times C}$ is a trainable weight matrix that transform the node representation from \mathbb{R}^D to \mathbb{R}^C , $Z^T \in \mathbb{R}^{L \times N}$ means the transposed matrix of Z , and we employ a residual connection [20] to promote the gradient propagation during training.

3.3 Semantic Context Loss

We propose a Semantic Context Loss or simply SC-loss to constrain the generation of exemplar-based SemG. It emphasizes the categories that appear in the scene and suppresses those do not appear in the scene, which makes the GINet a capable of enhancing the external semantic knowledge adaptively to each sample. Specifically, we first define a learnable semantic centroid $c_i \in \mathbb{R}^D$ for each category. Then, for each semantic node $\tilde{s}_i \in \mathbb{R}^D$ in a SemG \tilde{S}_o , we compute a score v_i by performing a simple dot product with a sigmoid activation upon \tilde{s}_i and c_i . The v_i ranges from 0 to 1 and is trained with the BCE loss. The SC-loss minimizes the similarity between the node feature in the semantic graph and the semantic centroid of nonexistent categories, and maximizes the similarity to existent classes. If v_i is closer to 1, the corresponding category exists in the current sample; Otherwise, it does not exist. The SC-loss can be formulated as follows:

$$Loss_{sc} = -\frac{1}{M} \sum_{i=1}^M (y_i \cdot \log v_i) + (1 - y_i) \log(1 - v_i), \quad (10)$$

where $y_i \in \{0, 1\}$ represents the presence of each category in ground truth. The proposed SC-loss is different from SE-loss in EncNet [54]. It built an additional fully connected layer on top of the Encoding Layer [54] to make individual predictions and was employed to improve the parsing of small objects. However, our SC-loss is employed to improve the generation of exemplar-based SemG.

We also add a full convolution segmentation head attached to Res4 of the backbone to obtain the segmentation result. Therefore, the objective of the

GINet consists of a SC-loss, an auxiliary loss, and a cross-entropy loss, which can be formulated as:

$$Loss = \lambda Loss_{sc} + \alpha Loss_{aux} + Loss_{ce}, \quad (11)$$

where λ and α are hyper-parameters, the selection of λ is discussed in the experiment section, and α for auxiliary loss is set to 0.4 similar some previous methods [54, 55, 15, 46].

4 Experiments

In this section, we perform a series of experiment to evaluate the effectiveness of our proposed Graph Interaction unit and SC-loss. Firstly, we give an introduction of the datasets that are used for scene parsing, i.e., Pascal-Context [33], COCO Stuff [4], and ADE20K [59]. Next, we conduct extensive evaluations with ablation studies of our proposed method on these datasets.

4.1 Datasets

Pascal-Context [33] is a classic set of annotations for PASCAL VOC2010, which has 10,103 images. In the training set, there are 4,998 images. The remaining 5,105 images form the validation set. Following previous works [54, 55, 59], we use the same 59 most frequent categories along with one background category (60 in total) in our experiments.

COCO Stuff [4] has a total number of 10,000 images with 183 classes including an 'unlabeled' class, where 9,000 images are used for training while 1,000 images for validation. We follow the same settings as in [15, 24], the results are reported on the data contains 171 categories (80 objects and 91 stuff) annotated for each pixel.

ADE20K [59] is a large scale dataset for scene parsing with 25,000 images and 151 categories. The dataset is split into the training set, validation set and test set with 20,000, 2,000, and 3,000 images, respectively. Following the standard benchmark [59], we validate our method on 150 categories, where the background class is not included.

4.2 Implementation Details

During training, we use ResNet-101 [20] (pre-trained on ImageNet) as our backbone. For retaining the resolution of the feature map, we use the Joint Pyramid Upsampling Module [46] instead of the dilation convolution for saving the training time, resulting in stride-8 models. We empirically set the number of nodes in VisG as 64, and node dimensions as 256. Similar to prior works [8, 54], we employ a poly learning rate policy [7] where the initial learning rate is updated by $lr = base_lr * (1 - \frac{iter}{total_iter})^{0.9}$ after each iteration. The SGD [39] optimizer is applied with 0.9 momentum and 1e-4 weight decay. The input size for all

Table 1. Ablation study on PASCAL-Context dataset. “GI” indicates Graph Interaction Unit. “SC-loss” represents Semantic Context Loss. “VisG” means that context reasoning is only performed on visual graph.

Method	Backbone	GI	SC-loss	mIoU
baseline	Res50			48.5
+VisG	Res50			50.2
GINet	Res50	✓		51.0
GINet	Res50	✓	✓	51.7
baseline	Res101			51.4
+VisG	Res101			53.0
GINet	Res101	✓		53.9
GINet	Res101	✓	✓	54.6

Table 2. Comparisons of accuracy and efficiency with other methods. All experiments are based on ResNet50. “Para” represents the extra parameters relative to the backbone. “FPS” indicates the inference speed of models.

Methods	mIoU	Para (M)	FPS
baseline	48.5	9.5	48.6
+GCU [25]	50.4	11.9	40.3
+GloRe [26]	50.2	11.2	45.3
+PSP [57]	50.4	23.1	37.5
+ASPP [8]	49.4	16.1	28.2
GINet(ours)	51.7	10.2	46.0

datasets is set to 520×520 . For data augmentation, we apply random flip, random crop, and random scale (0.5 to 2) using the zero-padding if needed. The batch size is set to 16 for all datasets. We set the initial learning rate to 0.005 on ADE20K dataset and 0.001 for others. The networks are trained for 30k, 150k, 100k iterations on Pascal-Context [33], ADE20K [59], COCO Stuff dataset [4], respectively.

During the validation phase, we follow [54, 55, 46] to average the multi-scale $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$ predictions of network. The performance is measured by the standard mean intersection of union (mIoU) in all experiments.

4.3 Experiments on Pascal-Context

We first conduct experiments on the Pascal-Context dataset with different settings and compare our GINet with other popular context modeling methods. Then we show and analyze the visualization results of our GINet. Finally, we compare with state-of-the-art to validate the efficiency and effectiveness of our method.

Ablation Study First, we show the effectiveness of the proposed GI unit and SC-loss. Then, we compare our method with other popular context modeling methods and study the influence of SC-loss in terms of weight. Finally, different word embedding approaches are tested to show the robustness of our proposed method.

Effectiveness of GI unit and SC-loss We design a detailed ablation study to verify the effect of our GI unit and SC-loss. Specifically, FCN with Joint Pyramid Upsampling Module [46] is chosen as our baseline model. As shown in Table 1, the baseline model achieves 48.5% mIoU. By performing reasoning over the VisG (row2), there is an improvement in performance by 1.7% (50.2 v.s. 48.5). Instead, by adopting our GI unit upon the baseline model to capture context

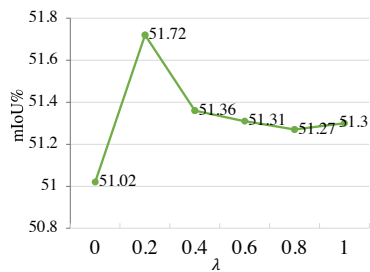


Fig. 3. Ablation study on the weight of SC-loss. GINet can achieve the best results with $\lambda = 0.2$.

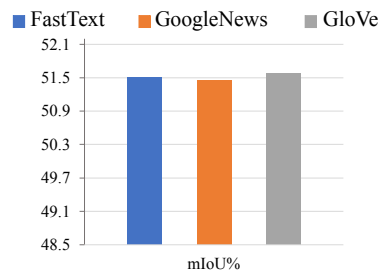


Fig. 4. Ablation study on different word embedding methods. (Best viewed in color.)

information from both visual regions and linguistic knowledge, one can see from Table 1 (row 3), there is a significant increase in performance by 2.5% (51.0 v.s. 48.5), which demonstrates the effectiveness of the GI unit and our context modeling method. Furthermore, by constraining the global information of semantic concepts, SC-loss can further improve the model performance to 51.7% mean IoU. Deeper pre-trained backbone provides better feature representations. GINet configured with ResNet-101 can obtain 54.6% mIoU, which outperforms the baseline mode by 3.2% in terms of mIoU.

Comparisons with context modeling methods Firstly, we compare the GINet with VisG-based context reasoning methods i.e., GCU [25], GloRe [10]. GloRe, GCU uses the typical projection, reasoning, and back-projection methods to model spatial context information. To ensure the fairness, we reproduce these methods. We report the model performances in terms of mIoU. As shown in Table 2, Compared to GCU’s 50.4% and GloRe’s 50.2%, our GINet achieves the highest score of 51.7% mIoU. This proves the effectiveness of introducing linguistic knowledge and label dependency upon the visual image region reasoning. PSPNet [57] and DeepLab[8] are classic methods for constructing visual context information, and their performance is lower than our GINet. To further analyze the efficiency of these context modeling methods, we list the inference speed (frame per second, denoted as FPS) of these models. FPS is measured on a Tesla-V100 GPU with input size 512×512 . As shown in Table 2, our model achieves 46.0 FPS, which outperforms all other context modeling methods.

Importance of the Weights of the SC-loss In order to study the necessity and effectiveness of the SC-loss, we train our GINet using different weights for the SC-loss, *e.g.*, $\lambda = \{0.2, 0.4, 0.6, 0.8, 1.0\}$. It is worth noting that $\lambda = 0$ means that the SC-loss is not applied. As shown in Figure 3, SC-loss can effectively improve the model performance when $\lambda = 0.2$. In our experiments, higher weights don’t bring more performance increase.

GINet with different word embedding By default, we use GloVe [35] as the initial representation of SemG. To verify the robustness of our method and see the influence of using different word embedding representations, we conduct experiments by applying three popular word embedding methods, *e.g.*, FastText

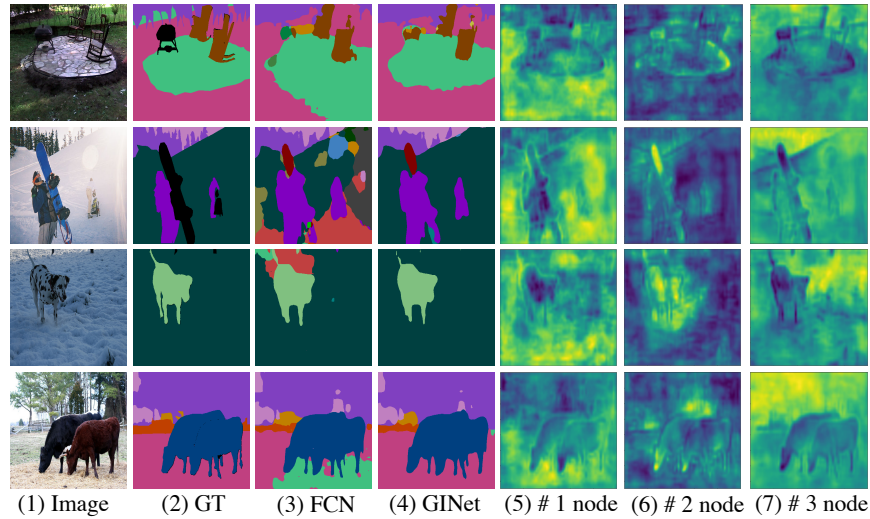


Fig. 5. Visualization of results and weights of projection matrices, all examples are from the Pascal-Context validation dataset. (*Best viewed in color.*)

[22], GoogleNews [32] and GloVe [35]. As shown in Figure 4, there is no significant performance fluctuation using different word embedding representations. This observation suggests that our method is quite general. No matter what word embedding methods are used, the proposed approach can capture the semantic information effectively.

Visualization and Analysis In this section, we provide a visualization of scene parsing results and projection matrix. Then we analyze the qualitative results delivered by the proposed method.

The scene parsing results are shown in Figure 5. Specifically, the first and second columns list the RGB input images and the ground truth scene parsing images, respectively. We compare baseline FCN [31] with our method in column 3 and column 4. One can see from the predicted parsing results images that our method shows considerable improvements. Particularly, it can be seen from rows 2 and 3, where the snow scene changes significantly in texture and color due to the illumination variation in the second and the fourth examples. By incorporating the semantic graph to promote the reasoning over the visual graph, our method successfully obtained more accurate parsing results. In the fourth row, it is fairly difficult to distinguish the green and yellow grass in the image only by the spatial context, while our method still identified the object correctly by incorporating semantic information, where the color changes mislead the FCN method.

Table 3. Comparison with the state-of-the-art approaches on PASCAL-Context dataset, COCO stuff test set, and ADE20K validation set. “†” means the model has been pre-trained on COCO Stuff. “-” means no public results available. “*” means employing online hard example mining(OHEM[38]).

Method	Backbone	mIoU%		
		PASCAL-Context	COCO Stuff	ADE20K
CCL [13]	ResNet-101	51.6	35.7	-
PSPNet [57]	ResNet-101	47.8	-	43.29
EncNet [54]	ResNet-101	51.7	-	44.65
TKCN [47]	ResNet-101	51.7	-	-
CFNet [48]	ResNet-101	52.4	36.6	-
DUpsampling [40]	Xception-71	52.5	-	-
SGR† [27]	ResNet-101	52.5	39.1	44.32
DSSPN [28]	ResNet-101	-	37.3	43.68
DANet [15]	ResNet-101	52.6	39.7	-
ANN* [61]	ResNet-101	52.8	-	45.24
FastFCN [46]	ResNet-101	53.1	-	44.34
GCU [25]	ResNet-101	-	-	44.81
EMANet [24]	ResNet-101	53.1	39.9	-
SVCNet [12]	ResNet-101	53.2	39.6	-
CCNet* [21]	ResNet-101	-	-	45.22
DMNet [19]	ResNet-101	54.4	-	45.50
ACNet* [16]	ResNet-101	54.1	40.1	45.90
GINet (Ours)	ResNet-101	54.9	40.6	45.54

Moreover, we can show that our method aggregates similar features from the visual feature map into a node in the visual graph. The graph nodes learn rich representations of different regions, and reasoning on these nodes can effectively capture relationships of image regions. We select three nodes (marked as #1, #2 and, #3) and show their corresponding projection weights in columns 5, 6, and 7, respectively. It can be observed that different nodes correspond to relevant regions in the image (the brighter areas in the image means high response). It can be seen from the 2nd row in Figure 5, node #1 aggregates and corresponds more with the background areas, while node #2 highlights the main objects in the images and node #3 shows more responses to the sky area in this example.

Comparisons with state-of-the-art methods We report 60 categories performance (including background) to compare with the state-of-the-art methods. As shown in Table 3, our GINet achieves the best performance and outperforms the SOTA DMNet[19] by 0.5%, which shows that our method is truly competitive. DMNet incorporates multiple Dynamic Convolutional Modules to adaptively exploit multi-scale filters to handle the scale variation of objects. In addition, the ACNet [16] obtains 54.1% mIoU, which captured the pixel-aware contexts by a competitive fusion of global context and local context according

to different per-pixel demands. However, our method extracts the semantic representation from the visual features under the guidance of the general semantic graph, and construct a semantic centroid to get the similarity score for each category.

4.4 Experiments on COCO Stuff

To further demonstrate the generalization of our GINet, we also conduct experiments on the COCO Stuff dataset [4]. Comparisons with state-of-the-art methods are shown in Table 3. Remarkably, the proposed model achieves 40.6% in terms of mIoU, which outperforms the best methods by a large margin. Among the current state-of-the-art methods, ACNet [16] introduced a data-driven gating mechanism to capture global context and local context according to pixel-aware context demand, DANet [15] deployed the self-attention module to capture long-range contextual information, and EMANet [24] proposed the EMA Unit to formulate the attention mechanism into an expectation-maximization manner. In contrast to these methods, our GINet considers the operation of capturing long-range dependencies as the way of graph reasoning, and additionally introduces the semantic context to enhance the discriminant property for the features.

4.5 Experiments on ADE20K

Finally, we compare our method and conduct experiments on the ADE20K dataset [59]. Table 3 compares the GINet performance against state-of-the-art methods. Our GINet outperforms the prior works and sets the new state-of-the-art mIoU to 45.54%. It is noting that our result is obtained by a regular training strategy in contrast to ANN [61], CCNet [21] and ACNet [16] where OHEM[38] is applied to help cope with difficult training cases. The ANN proposes an asymmetric fusion of non-local blocks to explore the long-range spatial relevance among features of different levels. The CCNet used a recurrent criss-cross attention module that aggregates contextual information from all pixels. We emphasize that achieving such an improvement on the ADE20K dataset is hard due to the complexity of this dataset.

5 Conclusion

We have presented a graph interaction unit to promote contextual reasoning over the visual graph by incorporating the semantic knowledge. We have also developed a Semantic Context loss upon the semantic graph output of graph interaction unit to emphasize the categories that appear in the scene and suppress those do not appear in the scene. Based on the proposed graph interaction unit and Semantic Context loss, we have developed a novel framework called Graph Interaction Network (GINet). The proposed approach based on the new framework outperforms state-of-the-art methods by a significant gain in performance on two challenging scene parsing benchmarks, e.g., Pascal-Context and COCO Stuff, and achieves a competitive performance on ADE20K dataset.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* (2017)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2002)
3. Biederman, I., Mezzanotte, R.J., Rabinowitz, J.C.: Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology* (1982)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: *CVPR* (2018)
5. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: *2017 IEEE Visual Communications and Image Processing (VCIP)* (2017)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV* (2018)
10. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: *CVPR* (2019)
11. Cheng, B., Chen, L.C., Wei, Y., Zhu, Y., Huang, Z., Xiong, J., Huang, T.S., Hwu, W.M., Shi, H.: Spgnet: Semantic prediction guidance for scene parsing. In: *ICCV* (2019)
12. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic correlation promoted shape-variant context for segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8885–8894 (2019)
13. Ding, H., Jiang, X., Shuai, B., Qun Liu, A., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: *CVPR* (2018)
14. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *ICCV* (2015)
15. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: *CVPR* (2019)
16. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: *Proceedings of the IEEE international conference on computer vision*. pp. 6748–6757 (2019)
17. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: *CVPR* (2019)
18. He, H., Zhang, J., Zhang, Q., Tao, D.: Grapy-ml: Graph pyramid mutual learning for cross-dataset human parsing. *CoRR* **abs/1911.12053** (2019), <http://arxiv.org/abs/1911.12053>
19. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: *The IEEE International Conference on Computer Vision (ICCV)* (October 2019)

20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
21. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. arXiv preprint arXiv:1811.11721 (2018)
22. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fast-text. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016)
24. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9167–9176 (2019)
25. Li, Y., Gupta, A.: Beyond grids: Learning graph representations for visual recognition. In: Advances in Neural Information Processing Systems. pp. 9225–9235 (2018)
26. Li, Y., Gu, C., Dullien, T., Vinyals, O., Kohli, P.: Graph matching networks for learning the similarity of graph structured objects. arXiv preprint arXiv:1904.12787 (2019)
27. Liang, X., Hu, Z., Zhang, H., Lin, L., Xing, E.P.: Symbolic graph reasoning meets convolutions. In: Advances in Neural Information Processing Systems (2018)
28. Liang, X., Zhou, H., Xing, E.: Dynamic-structured semantic propagation network. In: CVPR (2018)
29. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR (2016)
30. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
32. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
33. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR (2014)
34. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters—improve semantic segmentation by global convolutional network. In: CVPR (2017)
35. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (2014)
36. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: CVPR (2017)
37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (2015)
38. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
39. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning (2013)
40. Tian, Z., He, T., Shen, C., Yan, Y.: Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: CVPR (2019)

41. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision* (2005)
42. Wang, Q., Guo, G.: Ls-cnn: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security* **15**, 1640–1653 (2019)
43. Wang, Q., Wu, T., Zheng, H., Guo, G.: Hierarchical pyramid diverse attention networks for face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
44. Wang, Q., Zheng, Y., Yang, G., Jin, W., Chen, X., Yin, Y.: Multiscale rotation-invariant convolutional neural networks for lung texture classification. *IEEE journal of biomedical and health informatics* **22**(1), 184–195 (2018)
45. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks (2017)
46. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816* (2019)
47. Wu, T., Tang, S., Zhang, R., Cao, J., Li, J.: Tree-structured kronecker convolutional network for semantic segmentation. In: *ICME* (2019)
48. Wu, T., Tang, S., Zhang, R., Guo, G., Zhang, Y.: Consensus feature network for scene parsing. *arXiv preprint arXiv:1907.12411* (2019)
49. Wu, T., Tang, S., Zhang, R., Zhang, Y.: Cgnet: A light-weight context guided network for semantic segmentation. *arXiv preprint arXiv:1811.08201* (2018)
50. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: *CVPR* (2018)
51. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015)
52. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065* (2019)
53. Yuan, Y., Wang, J.: Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916* (2018)
54. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: *CVPR* (2018)
55. Zhang, H., Zhang, H., Wang, C., Xie, J.: Co-occurrent features in semantic segmentation. In: *CVPR* (2019)
56. Zhang, S., Yan, S., He, X.: Latentgcn: Learning efficient non-local relations for visual recognition. *arXiv preprint arXiv:1905.11634* (2019)
57. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *CVPR* (2017)
58. Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D., Jia, J.: Psanet: Point-wise spatial attention network for scene parsing. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 267–283 (2018)
59. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: *CVPR* (2017)
60. Zhou, L., Zhang, C., Wu, M.: D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: *CVPR Workshops* (2018)
61. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 593–602 (2019)