Progressive Knowledge-Embedded Unified Perceptual Parsing for Scene Understanding

Wenbo Zheng\textsuperscript{1,3} Lan Yan\textsuperscript{3,4} Fei-Yue Wang\textsuperscript{3,4} Chao Gou\textsuperscript{2} *

\textsuperscript{1} School of Software Engineering, Xi’an Jiaotong University
\textsuperscript{2} School of Intelligent Systems Engineering, Sun Yat-sen University
\textsuperscript{3} The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{4} School of Artificial Intelligence, University of Chinese Academy of Sciences

zwb2017@stu.xjtu.edu.cn; yanlan2017@ia.ac.cn; feiyue.wang@ia.ac.cn; gouchao@mail.sysu.edu.cn

Abstract

Human can naturally understand scenes in depth with the help of various knowledge accumulated and by a comprehensive visual concept organization including category labels and different-level attributes. This inspires us to unify professional knowledge at different levels with deep neural network architectures progressively for scene understanding. Different from the general embedding approaches, we construct different knowledge graphs for different levels of vision tasks by organizing the rich visual concepts accordingly. We employ a gated graph neural network and relational graph convolutional networks to propagate node messages for different levels of tasks and generate progressively different levels of knowledge representation through the graph. Compared with existing methods, our framework has a main appealing property leading to a novel progressive knowledge-embedded representation learning framework that incorporates different level knowledge graphs into the learning of networks at corresponding level. Extensive experiments on the widely used Broden+ dataset demonstrate the superiority of the proposed framework over other existing state-of-the-art methods.

1. Introduction

Humans can not only extract a large amount of semantic information at a glance, but also acquire knowledge from daily lives or professions, thus completing the task of scene understanding\cite{3,8,27,44,61}. Specifically, human not only instantly segment and recognize the scene and objects contained within, but also identify the fine-grained attributes of a scene, such as objects and materials. Usually, our kind knowledge refers to a comprehensive visual concept organization including category labels and their attributes. It is incredibly beneficial to scene understanding as attributes are always crucial to distinguish different subordinate categories\cite{6,55,60,66}. For example, we might know from a book that certain kinds of churches are built from polished stones. With this knowledge, to recognize the scene category “church” given an image, we might first recall the knowledge, attend to the corresponding objects to see whether it possesses these attributes, and then perform reasoning. Figure 1 illustrates an example of how professional knowledge aids scene understanding.

Up to now, mainstream algorithms of scene understanding are mainly divided into three categories: structure-based models\cite{36,37}, visual attention network-based model\cite{9,50,63}, multi-task learning-based model\cite{48}. These models all can locate discriminative regions/parts to distinguish
Figure 2. An example knowledge graph for modeling object-level attributes and material-level attributes with the results of image feature representations on the Broden+ dataset. Black color denotes "scene", blue denotes "object", orange denotes "material". The size of circles of "object" and "material" mean frequency of corresponding category. The links in the graph correspond to object/material categories. The actual object/material pictures represent correspond category.

subtle differences among different subordinate categories. However, structure-based models involve heavy annotations or geometric annotations of objects, preventing them from application to large-scale data. Visual attention network-based model can only locate the parts/regions roughly due to the lack of supervision or guidance. Multi-task learning-based models can only locate parts of each attribute, and the identification label is inaccurate due to lack of supervision or guidance. Conventional approaches for scene understanding usually neglect this knowledge and merely rely on low-level image cues for parsing.

Recently, the theory [46] of neuroscience points out that one of the mechanisms of humans’ understanding of natural vision is obtained through stimulation of the brain. Further, the acquisition and loss of this neurological stimulus is a progressive process [5, 12]. When humans see a scene image (stimulation), humans always associate this image with humans’ knowledge to understand this scene. Inspired by aforementioned neuroscience, we organize knowledge about categories and different-level attributes in the form of the knowledge graph, and we proposed a progressive knowledge-embedded-representation-learning framework to incorporate knowledge graph into image feature learning to promote the process of scene understanding.

To this end, our work focuses on a new task called Unified Perceptual Parsing [48]. Compared to conventional scene understanding task, this task emphasizes models parse various visual concepts at multiple perceptual levels such as scene, objects, and materials all at once. In other words, our framework is able to achieve different level vision task given one image. And our proposed framework contains three crucial components:

1. Gated graph neural networks (GGNN) [22] is employed to propagate node message through the graph to generate knowledge representation at the different level of the process of learning [4, 17, 21, 35, 45].
2. Relational graph convolutional networks (R-GCN) [38] is introduced to encode and combine different level node message through dealing with the different level data characteristic of the knowledge graph.
3. A novel progressive gated mechanism is introduced to learn the attribute-aware representation.

Specifically, we first construct a large-scale knowledge graph that associates category labels with their different-level attributes, as shown in Figure 2. As we can see, our framework initializes the constructed knowledge graph nodes with given image information for different levels of tasks implicitly. Thus, our framework associates these different levels of attributes with feature maps, and is able to reason about the discriminative attributes and categories for the image. In this way, our framework can learn feature maps with meaningful information that the parts/regions finely associate with the relevant different-level attributes in the graph. For example, the learned parts/regions of samples from category "church" always contains “building” and “stone” or “polished stone” in Figure 2. This category cannot contain other impossible attributes, such as “fabric”, because these scene related to attributes and these regions relate to attributes that are key to distinguish this category from others. This characteristic also provides insight into why the framework improves performance.

In summary, the contributions of our work can be concluded to three-fold:

1. Our work formulates a novel progressive knowledge-embedded representation learning framework that incorpo-
rates different level knowledge graph into the learning of network at the corresponding level.

(2) Through utilizing the gated graph neural networks, our work incorporates high-level knowledge graph as extra guidance into scene understanding. To the best of our knowledge, this is the first work to investigate this point.

(3) Extensive experimental results demonstrate the superiority of the proposed framework over existing state-of-the-art approaches.

2. Related Work

We review the related work about two research streams: unified perceptual parsing and knowledge representation. Then, we introduce the prerequisite knowledge about the brain of understanding the natural vision.

Unified Perceptual Parsing Humans recognize the visual world on multiple levels: we effortlessly classify scenes and detect internal objects, while also identifying the composition of the object’s material. Based on this, Xiao et al. [48] proposed a new task called unified perceptual parsing, which requires the machine vision systems to recognize as many visual concepts as possible from a given image. Obviously, this is a multitasking issue. Further, according to the principle of network dissection [2], Xiao et al. [48] proposed the unified perceptual parsing network. However, since the method does not notice the correlation between different levels of task attributes and the guidance of prior knowledge, this method is not very effective. It can be recognized that it is necessary to add knowledge guidance.

Knowledge Representation Representing prior knowledge in the form of graph structure [7, 18] and incorporating this structure for visual reasoning has received increasing attention [31, 33, 39, 49, 53]. For example, Malisiewicz et al. [29] build a large graph, with the nodes referring to object instances and the edge corresponding to associated types between nodes, to represent and reason about object identities and their mined relationships [9, 18, 26, 28, 34, 41, 43]. These methods usually involve hand-crafted features and manually-defined rules. Recently, more works are dedicated to exploring message propagation by learnable neural networks like [47] or neural network variants [51]. Relational graph convolutional networks (R-GCNs) [38] are encoder models which develop specifically to deal with the highly multi-relational data characteristic of realistic knowledge bases, and mine the implicit relationship between multi-relational data. Thus, we planned to use R-GCN to mine the relationship between the knowledge graphs of different levels of visual tasks. Gated graph neural network (GGNN) [22] is a fully differential recurrent neural network architecture for handling graph-structured data, which iteratively propagate node message through the graph to learn node-level or graph-level representation [32]. Several works have successfully developed GGNN variants for various vision tasks [45, 60]. Therefore, it is an excellent choice combining GGNN and R-GCN to handle different levels of visual tasks at once.

Prior Knowledge about Brain’s Understanding Natural Vision Our knowledge of brain processing has advanced dramatically in the last few decades, but this understanding remains far from complete, especially for stimuli with the broad dynamic range and strong temporal and spatial correlations characteristic of natural visual inputs. Maxwell et al. [46] highlight two broad strategies for approaching this problem: a stimulus-oriented framework and a goal-oriented one. In a stimulus-oriented framework, a common approach is to identify the transformations of sensory-input signals that optimize statistical and information-theoretic metrics. It’s worth noting that, after inputting stimulus, a message composed of an appropriately timed periodic train of pulse packets will be progressively amplified, and eventually will be strong enough to be propagated to the receiver neuronal network [5, 12]. Inspired by the theory of neuroscience, we consider using the different level of knowledge graphs to guide our neural network for corresponding levels of visual tasks, and further using the mined relationship to generate progressively different levels of knowledge representation to guide our neural network.

3. Our Framework

In this section, we first present the construction of our knowledge graphs that relate category labels with their different-level attributes. Then, we introduce our framework in detail, which consists of a GGNN for knowledge representation learning, a combining GGNN and R-GCN for progressive relational learning, and a gated mechanism to embed knowledge into discriminative image representation learning progressively. An overall pipeline of the framework is illustrated in Figure 3.

3.1. Knowledge-Graph Construction

Considering to parsing scene, objects, and materials all at once, we construct our knowledge graph that relates scene category labels with object-level attributes and material-level attributes. For the construction of our knowledge graph, we use the GGNN [22] method.

Principle of GGNN GGNN [22] is an end-to-end trainable network architecture that can learn features for arbitrary graph-structured data by iteratively updating node representation in a recurrent fashion. Formally, the input is a graph represented as $G = \{V, A\}$, in which $V$ is the node set and $A$ is the adjacency matrix denoting the connections among these nodes. We define $t$ is the time step of conducting the knowledge graph. At $t = 0$, input feature vectors $x_v$ that depends on the special task is initialized as the hidden state. Then, at time-step $t$, we define $h_v^t$ as the hidden state. For each node $v \in V$, the basic propagation recurrent
The propagation process is formulated as

\[ h_v^0 = x_v \]
\[ a_v^t = A_v^T [h_1^{t-1} \cdots h_{|V|}^{t-1}]^T + b \] (1)
\[ h_v = \text{gate}(a_v^t, h_v^{t-1}) \]

where \( A_v \) is a sub-matrix of \( A \) represents the connections of node \( v \) with its neighbors, and \( \text{gate} \) denotes gated update mechanism, which is defined as:

\[ z_v^t = \sigma(W^z a_v^t + U^z h_v^{t-1}) \]
\[ r_v^t = \sigma(W^r a_v^t + U^r h_v^{t-1}) \]
\[ \tilde{h}_v = \tanh(W^h a_v^t + U(h_v^{t-1} \odot r_v^t)) \]
\[ h_v^t = (1 - z_v^t) \odot h_v^{t-1} + z_v^t \odot \tilde{h}_v \] (2)

where \( \odot \), \( \sigma \) and \( \tanh \) are the element-wise multiplication operation, the logistic sigmoid and hyperbolic tangent functions, respectively.

The propagation process is repeated until our fixed iteration \( T \). During this process, we update the representation of each node based on its history state and the message sent by its neighbors. Thus, we can obtain the final hidden states \( \{h_1^T, h_2^T, \ldots, h_{|V|}^T\} \). All in all, the computation process of equation (1) can be reduced to \( h_v^T = \text{GGNN}(h_1^T, h_2^T, \ldots, h_{|V|}^T; A_v) \). Similar to [10], we employ an output network that is implemented by a fully-connected layer \( o \), to compute node-level feature, expressed by

\[ o_v = o([h_v^T, x_v]), v = 1, 2, 3, \ldots |V| \] (3)

**A Case of Constructing Scene-Object GGNN**

Distinctly, we need to construct two knowledge graph of which one relates scene category labels with object-level attributes and the other relates scene category labels with material-level attributes. We use the knowledge graph that relates scene category labels with object-level attributes as an example to illustrate the process of constructing GGNN. Given dataset that covers \( J \) scene categories and \( L \) object attributes, the graph has a node set \( V \) with \( J + L \) elements. Similar to [7], we define the \( J \times L \) matrix \( S_{\text{Scene-Object}} \) that denotes the confidence that this category has the attribute and its value range is [0, 1]. Then, we can get the adjacency matrix \( A_{\text{Scene-Object}} \) can be expressed as

\[
A_{\text{Scene-Object}} = \begin{bmatrix}
0_{J \times J} & S_{\text{Scene-Object}} \\
0_{J \times L} & 0_{L \times L}
\end{bmatrix}
\] (4)

where \( 0_n \) is a zero vector with dimension \( n \).

Finally, by this way, we can get the knowledge graph \( G_{\text{Scene-Object}} = \{V_{\text{Scene-Object}}, A_{\text{Scene-Object}}\} \). We are able to get the knowledge graph \( G_{\text{Scene-Material}} = \{V_{\text{Scene-Material}}, A_{\text{Scene-Material}}\} \) that relates scene category labels with material-level attributes in a similar way.

### 3.2. Progressive Relational Knowledge-Graph Learning

Our work focuses on the unified perceptual parsing task. Based on the theory of the network dissection, Xiao et al. [48] find that unified perceptual parsing network can effectively parse the feature maps of different levels of visual tasks, such as the material-level feature maps, object-level feature maps and scene feature maps. Xiao et al. [48] also prove that this network parsed the features according to the following order: material \( \rightarrow \) object \( \rightarrow \) scene. However, due to the lack of the guidance of prior knowledge of this network, the output of this network might be irrational. For example, the material of the sky is painting; the material of the road is fabric; the position of the segmented object is biased, etc. In order to solve this problem, we plan to use the knowledge graph to guide the training of the network. Since this problem is three different levels of tasks, we create a corresponding knowledge graph for each task to guide corresponding tasks. For material-level and object-level tasks, we construct scene-material knowledge graph and scene-object knowledge graph, respectively. For the final task, we combine scene-material knowledge graph with
scene-object knowledge graph using R-GCN [38]. Besides, we design the progressive relational structure shown in Figure 3.

**Material-Level and Object-Level Tasks** We use the object-level task as an example to illustrate. For the material-level task, we can also use a similar method to learn under the guidance of knowledge graph. After building the scene-object knowledge graph, we employ the GGNN to propagate node message through the graph and compute a feature vector for each node. All the feature vectors are then concatenated to generate the final representation for the knowledge graph.

We define the score vector \( S = \{s_0, s_1, \cdots, s_L\} \) as the confidence of this category presented in a given image. We initialize the node refers to the category label \( i \) with \( s_i \), and the node refers to each attribute with a zero vector. Therefore, we can get the input feature for each node can be represented as

\[
x_v = \begin{cases} 
  [s_i, 0_{n-1}] & \text{if node } v \text{ refers to scene category } i \\
  [0_n] & \text{if node } v \text{ refers to an attribute}
\end{cases}
\]  

(5)

where \( 0_n \) is a zero vector with dimension \( n \). After \( T_2 \) iteration in Figure 3, according to the principle of the GGNN mentioned in Section 3.1, we can get the node-level feature \( o_v^{\text{Object}} \). Similarly, we can get the node-level feature \( o_v^{\text{Material}} \) for the material-level task by \( T_1 \) iteration. Finally, these features are concatenated to produce the final knowledge representation \( f_{\text{knowledge}}^{\text{Object}} \) and \( f_{\text{knowledge}}^{\text{Material}} \).

**Final Scene Task** According to the principle of the GGNN mentioned in Section 3.1 at \( T_3 \) iteration, we can get \( h_v^{T_3}^{\text{Object}} \) and \( h_v^{T_3}^{\text{Material}} \). We denote directed and labeled multi-graphs as \( G^{\text{Scene}} = \{V, A, R^{\text{Object-Material}}\} \) with nodes \( V \) and labeled edges \( A \), where \( R^{\text{Object-Material}} \) is relation between object and material.

According to the principle of R-GCNs [38], we define the following simple propagation model for calculating the forward-pass update in multi-graph:

\[
h_v^{R-GCN} = \tanh \left( \frac{1}{N} W_r h_v^{T_3} + W_0 h_v^{T_3} \right) \]  

(6)

where \( \tanh \) is hyperbolic tangent functions, \( N \) is a problem-specific normalization constant that can either be learned or chosen in advance. Each \( W_r \) is defined as follows:

\[
W_r = \sum_{b=1}^{B} C_b T_b
\]  

(7)

where \( T_b \) is a linear combination of basis transformations with coefficients \( C_b \) such that only the coefficients depend on \( R^{\text{Object-Material}} \), according to the principle of R-GCNs [38].

As we can see, Eq.(6) accumulates transformed feature vectors of neighboring nodes, and we obtain the \( h_v^{R-GCN} \). Then, we can get the \( o_v^{\text{Scene}} \) computed by Eq.(3). Finally, these features are concatenated to produce the final knowledge representation \( f_{\text{knowledge}}^{\text{Scene}} \).

3.3. Knowledge-Embedded Unified Perceptual Parsing

We introduce the gated mechanism that embeds the knowledge representation to enhance image representation learning.

**Image Representation** We start by introducing the image feature extraction. Based on the unified perceptual parsing network [48], we apply this model to extract image features. This model apply a pyramid pooling module from PSPNet [62] on the last layer of the backbone network before feeding it into the top-down branch in feature pyramid network. The down-sampling rates are \( \{4, 8, 16, 32, 32\} \), respectively. Specifically, given an image, we can extract material-level feature maps with \( \frac{1}{4} \) size of given image, object-level feature maps with \( \frac{1}{16} \) size of image, and scene-level feature maps with \( \frac{1}{16} \) size of image, by using the unified perceptual parsing network. Thus, we use the compact bilinear pooling method [11] shown in Figure 3 to produce feature maps \( f_{\text{network}}^{\text{Material}}, f_{\text{network}}^{\text{Object}}, f_{\text{network}}^{\text{Scene}} \), respectively.

**Unified Perceptual Parsing by Knowledge-Embedded Learning** Similar to [47], we embed this representation into image feature learning to learn feature corresponding to this attributes. Considering suppressing non-informative features and allowing informational features to pass under the guidance of different-level of knowledge graphs, we introduce a gated mechanism can be expressed as

\[
f_{\text{Material}} = \sigma(g(f_{\text{network}}^{\text{Material}}, f_{\text{knowledge}}^{\text{Material}}) \odot f_{\text{network}}^{\text{Material}})
\]

\[
f_{\text{Object}} = \sigma(g(f_{\text{network}}^{\text{Object}}, f_{\text{knowledge}}^{\text{Object}}) \odot f_{\text{network}}^{\text{Object}})
\]

\[
f_{\text{Scene}} = \sigma(g(f_{\text{network}}^{\text{Scene}}, f_{\text{knowledge}}^{\text{Scene}}) \odot f_{\text{network}}^{\text{Scene}})
\]  

(8)

where \( \sigma \) is the logistic sigmoid, \( \odot \) denotes the element-wise multiplication operation, \( g \) is a neural network that takes the concatenation of the feature of knowledge representation and the feature of extracting by using the unified perceptual parsing network.

4. Experiments

4.1. Experiment Settings and Experimental Results

4.1.1 Datasets

We evaluate our framework and the competing methods on the Broden+ dataset [48] that is specifically used for unified
For the scene-level task, we choose the top-1 accuracy as the performance evaluation index of the algorithm. For object-level task and material-level task, we choose the mIoU which indicates the intersection-over-union (IoU) between the predicted and ground truth pixels, and the pixel accuracy which indicates the proportion of correctly classified pixels, as the performance evaluation index of the algorithm. Also, to compare the effectiveness of our proposed architecture and the competing methods for semantic segmentation, we use the ADE20K dataset and choose the mIoU, pixel accuracy and overall which is averaged over all object classes as the performance evaluation index of the algorithm.

### 4.1.2 Implementation Details

For the GGNN, we utilize the compact bilinear model released by work [11] to produce the scores to initialize the hidden states. For R-GCNs, we build a 2-layer model with 16 hidden units by work [38], and trained for 50 epochs using a learning rate of 0.01. For fair comparisons, we set the epoch of our method as 40 similar to [48] and trained on the training part of the Broden+ dataset. The iteration time $T_1$, $T_2$, $T_3$ is set to 5, 10, 10, respectively. Our framework is jointly trained using the cross-entropy loss. All components of the framework are trained with SGD except GGNN and R-GCNs that are trained with ADAM following [30]. Our framework is able to uniformly parser visual knowledge while predicting the hierarchical output.


<table>
<thead>
<tr>
<th>Training Data</th>
<th>Knowledge-Graph</th>
<th>Object</th>
<th>Material</th>
<th>Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>M</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.1.3 Qualitative Results on ADE20K Dataset

We provide qualitative results of our framework, as visualized in Figure 4. Our framework is able to uniformly parser visual knowledge while effectively predicting the hierarchical output.

### 4.2. Ablation Study

To verify the contribution of knowledge embedding and progressive knowledge embedding, we design the Ablation experiment. Note that our framework employs the unified perceptual parsing network [48] as the baseline. The quantitative evaluations of the unified perceptual parsing network and our framework are shown in Table 1. As shown in Table 1, it is evident that the results of our framework are better than the baseline.

To further clarify the contribution of knowledge guided, we analyze the following two aspects: comparison of tasks at the same level and comparison of the different level tasks under different-level knowledge graph.

**Comparison of The Same Level Tasks**

To better verify the benefit of embedding knowledge for feature learning at the same-level task, we conduct an experiment that removes the guidance of the knowledge graph and other components left unchanged. The comparison results are presented in Table 1. As shown in Table 1, for the object-level task, the value of mean IoU has increased by 6.73, and the value of pixel accuracy has increased by 6.24, after the guidance of the knowledge graph. It is obvious that the scene-level and the material-level task is improved by adding the guidance of the knowledge graph. This suggests that the design of embedding knowledge for feature learnings is reasonable and effective.

**Comparison of The Different Level Tasks**

Under the guidance of knowledge maps, we progressively add the guidance of different knowledge graphs to prove the importance and necessity of progressive knowledge-embedding learning. In this process, we cannot change other components. The comparison results are presented in Table 1. As shown in Table 1, the value of mean IoU and pixel accuracy under the guidance of object-level knowledge graph and material-level knowledge graph is lower than under only the guidance of object-level knowledge graph at the object-
level task. Due to object attributes and material attributes are context-sensitive, it does not incur additional information but only increasing the complexity of the model. This suggests that the design of the chosen R-GCN is reasonable and effective. The value of mean IoU and pixel accuracy under the guidance of all-level knowledge graph is higher than under the guidance of the corresponding level knowledge graph. This suggests that the design of progressive embedding knowledge for feature learnings is reasonable and effective.

4.3. Comparison with State-of-the-Art Methods

Qualitative Evaluation on Comparison Experiments.

It can be seen from the Figure 5 that, the results of our framework can effectively parse the object and material, and can effectively recognize the scene. For example, at the scene-level task, our framework is closer to the scene of a given image than the Xiao et al. proposed algorithms [48]. All in all, our algorithm in this paper has a higher overall effectiveness of understanding than Xiao et al. proposed algorithms [48].

Quantitative Evaluation on Comparison Experiments.

From Table 2, it can get the following these points: mean IoU of our method is 25, 23.97, 16.95, 24.7, 14.03, 11.44, 5.64, 12.06, 4.66, 5.12, 1.13, 0.98, 4.45, 1.34, 1.06, 2.04, 1.69, 2.66, 2.57, 1.12, 1.1, 0.96, 0.74, and 1.48 higher than Fei-Fei et al., FE-CCM, FCN, SegNet, DilatedNet, CascadeNet, RefineNet, DilatedNet, PSPNet, UPerNet, PAC-multiple + CAB + MS, GANet, PDN, EFCN-8s, CaseNet, SAC, EncNet, DSSPN, PSANet, CCNet, APNB, APCNet, SPNet, and DPM, respectively. In pixel accuracy and overall evaluation, there are also true of the case. This means our method has the best ability of understanding scene and outperforms state-of-the-art methods.

5. Conclusion and Future Work

In this paper, we have presented a novel framework for handling the problem of scene understanding. The key idea is that our work formulates a novel progressive knowledge-embedded representation learning framework that incorporates different level knowledge graph into the learning of network at the corresponding level. This not only helps to endow the deep model with learned relationships mined under the guidance of the knowledge graphs, but also provides a solution for scene understanding. Extensive experiments on the widely used Broden+ dataset demonstrate the superiority of our framework over existing state-of-the-art methods. Following the main idea of this work, future research can be expanded in various aspects, including the tasks of...
visual question answering and visual commonsense understanding.

Acknowledgments This work is supported in part by National Key R&D Program of China (2020YFB1600400), in part by Key Research and Development Program of Guangzhou (202007050002).

References


