Progressive Knowledge-Embedded Unified Perceptual Parsing for Scene Understanding

Wenbo Zheng [1,3] Lan Yan [3,4] Fei-Yue Wang [3,4] Chao Gou [2] *
[1] School of Software Engineering, Xi'an Jiaotong University
[2] School of Intelligent Systems Engineering, Sun Yat-sen University
[3] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
[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 [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 [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.

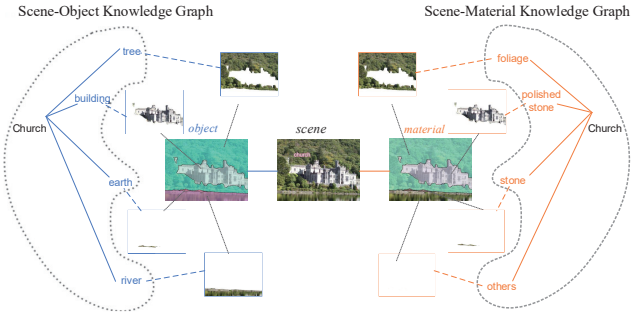

Figure 1. An example of knowledge graphs are able to to help with scene understanding. Our proposed framework is able to associate object-level attributes and material-level attributes with the results of image feature representations. Black color denotes "scene", blue denotes "object", orange denotes "material". The links in the graph correspond to object/material categories. The actual object/material pictures represent correspond category.

Up to now, mainstream algorithms of scene understanding are mainly divided into three categories: structure-based models [36, 37], visual attention network-based model [9, 50, 63], multi-task learning-based model [48]. These models all can locate discriminative regions/parts to distinguish

*Chao Gou is the corresponding author.

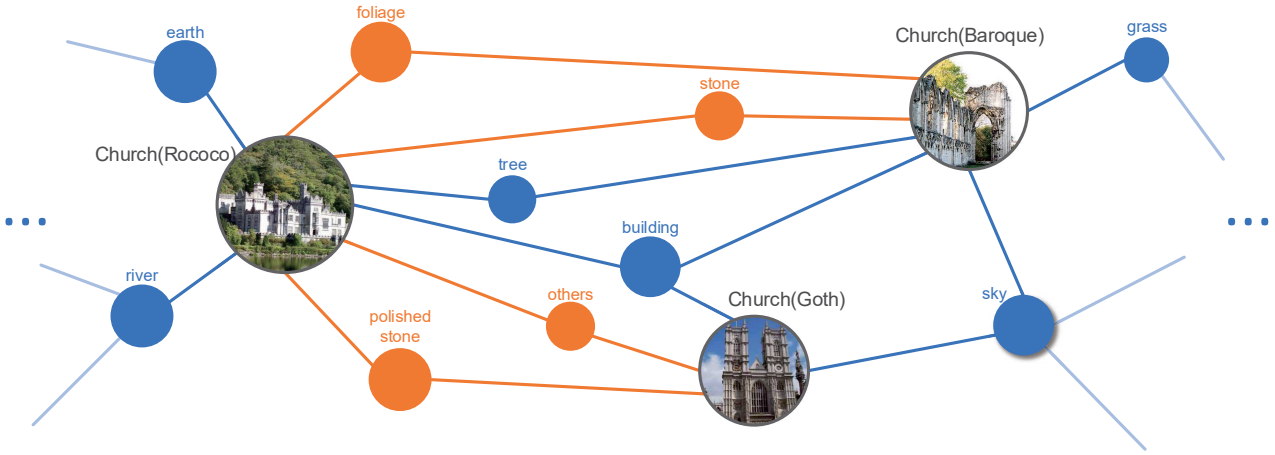

Figure 2. An example knowledge graph for modeling object-level attributes and material-level attributes with the results of image feature representations on the Broden+ dataset. Black color denotes "scene", blue denotes "object", orange denotes "material". The size of circles of "object" and "material" 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 $\mathcal{G} = \{\mathbf{V}, \mathbf{A}\}$, in which $\mathbf{V}$ is the node set and $\mathbf{A}$ 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 $\mathbf{x_v}$ that depends on the special task is initialized as the hidden state. Then, at time-step t, we define $\mathbf{h_v}^t$ as the hidden state. For each node $v \in \mathbf{V}$, the basic propagation recurrent

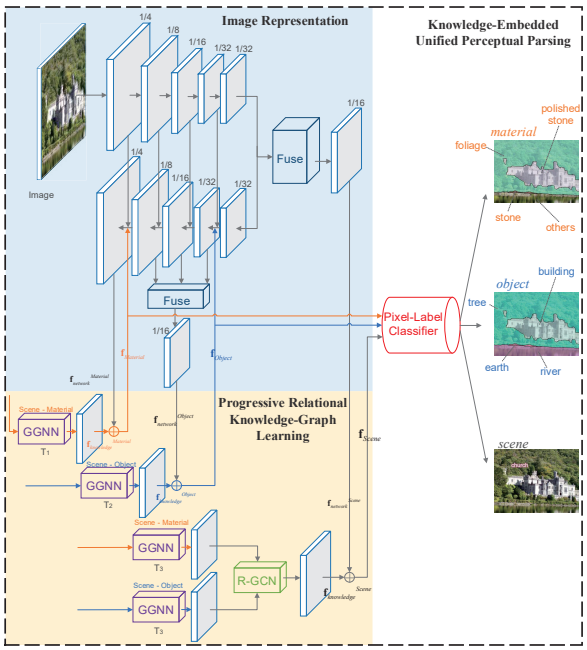

Figure 3. An overall pipeline of our proposed knowledge-embedded representation learning framework. The primary framework consists of a GGNN and R-GCN that takes the knowledge graph as input and propagates node information through the graph to learn knowledge representation under different time step, and a gated mechanism that embeds the representation into the image feature learning to learn attribute-aware features progressively. All components of the framework can be trained in an end-to-end fashion.

process is formulated as

$$\begin{aligned} \mathbf{h}_v{}^0 &= \mathbf{x}_v \\ \mathbf{a}_v{}^t &= \mathbf{A}_v{}^T[\mathbf{h}_1{}^{t-1} \cdots \mathbf{h}_{|\mathbf{V}|}{}^{t-1}]^T + \mathbf{b} \\ \mathbf{h}_v{}^t &= gate(\mathbf{a}_v{}^t, \mathbf{h}_v{}^{t-1}) \end{aligned} \tag{1}$$

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

$$\begin{aligned} \mathbf{z}_v{}^t &= \sigma(\mathbf{W}^z\mathbf{a}_v{}^t + \mathbf{U}^z\mathbf{h}_v{}^{t-1}) \\ \mathbf{r}_v{}^t &= \sigma(\mathbf{W}^r\mathbf{a}_v{}^t + \mathbf{U}^r\mathbf{h}_v{}^{t-1}) \\ \tilde{\mathbf{h}}_v^t &= \tanh(\mathbf{W}\mathbf{a}_v{}^t + \mathbf{U}(\mathbf{r}_v{}^t \odot \mathbf{h}_v{}^{t-1})) \\ \mathbf{h}_v{}^t &= (1 - \mathbf{z}_v{}^t) \odot \mathbf{h}_v{}^{t-1} + \mathbf{z}_v{}^t \odot \tilde{\mathbf{h}}_v^t \end{aligned} \tag{2}$$

where $\odot$, σ 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 $\{\mathbf{h}_1{}^T, \mathbf{h}_2{}^T, \ldots, \mathbf{h}_{|\mathbf{V}|}{}^T\}$. All in all, the computation process of equation (1) can be reduced to $\mathbf{h}_v{}^t = \text{GGNN}(\mathbf{h}_1{}^T, \mathbf{h}_2{}^T, \ldots, \mathbf{h}_{|\mathbf{V}|}{}^T; \mathbf{A}_v)$. Similar to [10], we employ an output network that is implemented by a fully-connected layer o, to compute node-level feature, expressed by

$$\mathbf{o}_v = o([\mathbf{h}_v{}^T, \mathbf{x}_v]), v = 1, 2, 3, \cdots |\mathbf{V}| \tag{3}$$

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 $\mathbf{V}$ with $J + L$ elements. Similar to [7], we define the $J \times L$ matrix $\mathbf{S}_{Scene-Object}$ that denotes the confidence that this category has the attribute and its value range is $[0, 1]$. Then, we can get the adjacency matrix $\mathbf{A}_{Scene-Object}$ can be expressed as

$$\mathbf{A}_{Scene-Object} = \begin{bmatrix} \mathbf{0}_{J\times J} & \mathbf{S}_{Scene-Object} \\ \mathbf{0}_{J\times L} & \mathbf{0}_{L\times L} \end{bmatrix} \tag{4}$$

where $\mathbf{0}_n$ is a zero vector with dimension n.

Finally, by this way, we can get the knowledge graph $\mathcal{G}_{Scene-Object} = \{\mathbf{V}_{Scene-Object}, \mathbf{A}_{Scene-Object}\}$. We are able to get the knowledge graph $\mathcal{G}_{Scene-Material} = \{\mathbf{V}_{Scene-Material}, \mathbf{A}_{Scene-Material}\}$ 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 → object → 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 $\mathbf{S} = \{s_0, s_1, \cdots, s_L\}$ 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. Thus, we can get the input feature for each node can be represented as

$$\mathbf{x}_v = \begin{cases} [s_i, \mathbf{0}_{n-1}] & \text{if node } v \text{ refers to scene category } i \\ [\mathbf{0}_n] & \text{if node } v \text{ refers to an attribute} \end{cases} \tag{5}$$

where $\mathbf{0}_n$ 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 ${\mathbf{o}_v}^{Object}$. Similarly, we also can get the the node-level feature ${\mathbf{o}_v}^{Material}$ for the material-level task by T_1 iteration. Finally, these features are concatenated to produce the final knowledge representation ${\mathbf{f}_{knowledge}}^{Object}$ and ${\mathbf{f}_{knowledge}}^{Material}$.

Final Scene Task According to the principle of the GGNN mentioned in Section 3.1 at T_3 iteration, we can get $\mathbf{h}_v^{T_3}{}_{Object}$ and $\mathbf{h}_v^{T_3}{}_{Material}$. We denote directed and labeled multi-graphs as $\mathcal{G}_{Scene} = \{\mathbf{V}, \mathbf{A}, \mathbf{R}_{Object-Material}\}$ with nodes $\mathbf{V}$ and labeled edges $\mathbf{A}$, where $\mathbf{R}_{Object-Material}$ 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:

$$h^{R-GCN} = \tanh\left(\frac{1}{\mathcal{N}}\mathbf{W}_r \mathbf{h}_v^{T_3}{}_{Object} + \mathbf{W}_0 \mathbf{h}_v^{T_3}{}_{Material}\right) \tag{6}$$

where $\tanh$ is hyperbolic tangent functions, N is a problem-specific normalization constant that can either be learned or chosen in advance. Each $\mathbf{W}_r$ is defined as follows:

$$\mathbf{W}_r = \sum_{b=1}^{B} C_b T_b \tag{7}$$

where T_b is a linear combination of basis transformations with coefficients C_b such that only the coefficients depend on $\mathbf{R}_{Object-Material}$, 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^{R-GCN}. Then, we can get the ${\mathbf{o}_v}^{Scene}$ computed by Eq.(3). Finally, these features are concatenated to produce the final knowledge representation ${\mathbf{f}_{knowledge}}^{Scene}$.

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 ${\mathbf{f}_{network}}^{Material}$, ${\mathbf{f}_{network}}^{Object}$ and ${\mathbf{f}_{network}}^{Scene}$, 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

$$\begin{aligned} \mathbf{f}_{Material} &= \sigma(g({\mathbf{f}_{network}}^{Material}, {\mathbf{f}_{knowledge}}^{Material})) \odot \\ &\quad {\mathbf{f}_{network}}^{Material} \\ \mathbf{f}_{Object} &= \sigma(g({\mathbf{f}_{network}}^{Object}, {\mathbf{f}_{knowledge}}^{Object})) \odot \\ &\quad {\mathbf{f}_{network}}^{Object} \\ \mathbf{f}_{Scene} &= \sigma(g({\mathbf{f}_{network}}^{Scene}, {\mathbf{f}_{knowledge}}^{Scene})) \odot \\ &\quad {\mathbf{f}_{network}}^{Scene} \end{aligned} \tag{8}$$

where σ 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

Figure 4. Predictions on the validation set using our framework. From left to right: original image, scene classification results, object parser results and material parser results.

perceptual parsing.

For the scene-level task, we choose the top-1 accuracy 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 [65] 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 effectively predicting the hierarchical output.

Table 1. Results of Ablation study on the Broden+ dataset. O: Object. M: Material. S: Scene. mI.: mean IoU. P.A.: pixel accuracy. T-1: top-1 accuracy.

Training Data			Knowledge-Graph			Object		Material		Scene
+O	+M	+S	+O	+M	+S	mI.	P.A.	mI.	P.A.	T-1
✓						24.72	78.03	-	-	-
	✓					-	-	52.78	84.32	-
✓	✓	✓				*23.36*	*77.09*	*54.19*	*84.45*	*70.87*
✓			✓			31.45	84.27	-	-	-
	✓			✓		-	-	59.26	90.47	-
✓	✓	✓	✓	✓		30.99	83.86	59.26	90.47	75.29
✓	✓	✓	✓	✓	✓	**32.48**	**85.35**	**60.04**	**91.62**	**80.01**

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-

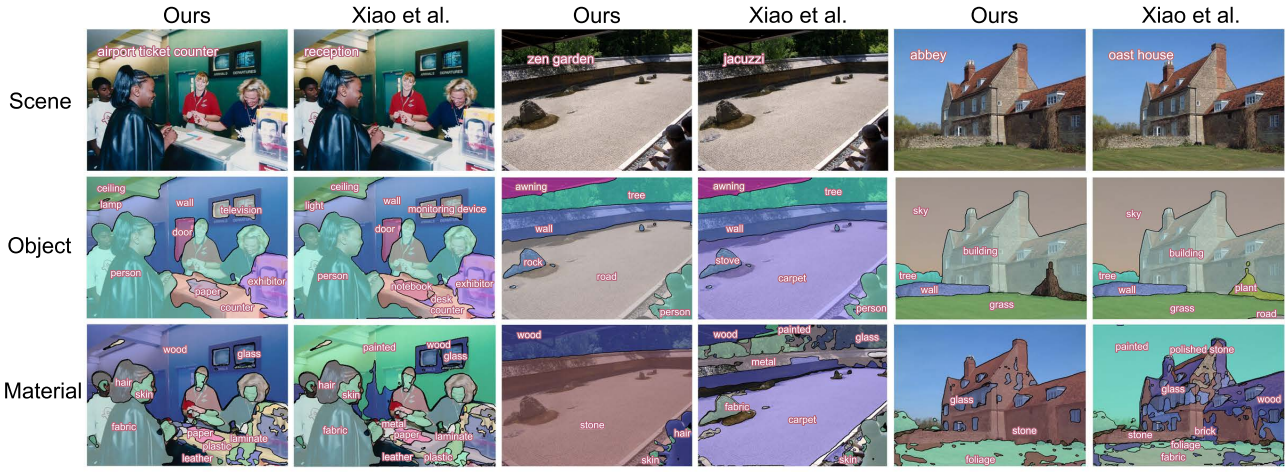

Figure 5. Comparison of ours and Xiao et al. proposed method on the task of unified perceptual parsing

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.

Table 2. Quantitative evaluations of the comparison experiment on the ADE20K dataset. mI.: mean IoU. P.A.: pixel accuracy.

Method	mI.	P.A.	Overall
Fei-Fei et al. [20]	21.34	69.44	40.23
FE-CCM [19]	22.37	69.75	43.24
FCN [40]	29.39	71.32	50.36
SegNet [1]	21.64	71	46.32
CascadeNet [65]	34.9	74.52	54.71
RefineNet [25]	40.7	-	-
DilatedNet [52]	34.28	76.35	55.32
PSPNet [62]	41.68	80.04	60.86
UPerNet [48]	41.22	79.98	60.6
PAC-multiple + CAB + MS [58]	45.21	82.14	-
GANet [56]	45.36	82.14	-
PDN [59]	41.89	80.81	-
EFCN-8s [42]	45	-	-
CaseNet [16]	45.28	-	-
SAC [57]	44.30	81.86	63.08
EncNet [54]	44.65	81.69	63.17
DSSPN [24]	43.68	81.13	62.41
PSANet [64]	43.77	81.51	62.64
CCNet [15]	45.22	-	-
APNB [67]	45.24	-	-
APCNet [13]	45.38	-	-
SPNet [14]	45.60	82.09	63.85
DPM [23]	44.86	81.55	
Ours	**46.34**	**83.76**	**64.98**

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

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, Dec 2017. 7

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017. 3

[3] M. Camplani, T. Mantecón, and L. Salgado. Depth-color fusion strategy for 3-d scene modeling with kinect. *IEEE Transactions on Cybernetics*, 43(6):1560–1571, Dec 2013. 1

[4] X. Cao, X. Wei, Y. Han, and X. Chen. An object-level high-order contextual descriptor based on semantic, spatial, and scale cues. *IEEE Transactions on Cybernetics*, 45(7):1327–1339, July 2015. 2

[5] Wei-Chih Chang, Jan Kudlacek, Jaroslav Hlinka, Jan Chvojka, Michal Hadrava, Vojtech Kumpost, Andrew D. Powell, Radek Janca, Matias I. Maturana, Philippa J. Karoly, Dean R. Freestone, Mark J. Cook, Milan Palus, Jakub Otahal, John G. R. Jefferys, and Premysl Jiruska. Loss of neuronal network resilience precedes seizures and determines the ictogenic nature of interictal synaptic perturbations. *Nature Neuroscience*, 21(12):1742–1752, 2018. 2, 3

[6] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[7] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 627–634. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 3, 4

[8] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, May 2019. 1

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[11] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, June 2016. 5, 6

[12] Gerald Hahn, Adrian Ponce-Alvarez, Gustavo Deco, Ad Aertsen, and Arvind Kumar. Portraits of communication in neuronal networks. *Nature Reviews Neuroscience*, 20(2):117–127, 2019. 2, 3

[13] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[14] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[15] 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/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Casenet: Content-adaptive scale interaction networks for scene parsing. *CoRR*, abs/1904.08170, 2019. 7

[17] V. J. Kok and C. S. Chan. Grcs: Granular computing-based crowd segmentation. *IEEE Transactions on Cybernetics*, 47(5):1157–1168, May 2017. 2

[18] C. Leng, H. Zhang, G. Cai, I. Cheng, and A. Basu. Graph regularized lp smooth non-negative matrix factorization for data representation. *IEEE/CAA Journal of Automatica Sinica*, 6(2):584–595, March 2019. 3

[19] C. Li, A. Kowdle, A. Saxena, and T. Chen. Toward holistic scene understanding: Feedback enabled cascaded classification models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1394–1408, July 2012. 7

[20] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036–2043, June 2009. 7

[21] X. Li, L. Mou, and X. Lu. Scene parsing from an map perspective. *IEEE Transactions on Cybernetics*, 45(9):1876–1886, Sep. 2015. 2

[22] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. In *Proceedings of ICLR'16*, April 2016. 2, 3

[23] Zhiheng Li, Wenxuan Bao, Jiayang Zheng, and Chenliang Xu. Deep grouping model for unified perceptual parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[24] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[25] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017. 7

[26] H. Liu, M. Zhou, and Q. Liu. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3):703–715, May 2019. 3

[27] X. Lu, X. Li, and L. Mou. Semi-supervised multitask learning for scene recognition. *IEEE Transactions on Cybernetics*, 45(9):1967–1976, Sep. 2015. 1

[28] L. Maccari. Detecting and mitigating points of failure in community networks: A graph-based approach. *IEEE Transactions on Computational Social Systems*, 6(1):103–116, Feb 2019. 3

[29] Tomasz Malisiewicz and Alexei A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, December 2009. 3

[30] K. Marino, R. Salakhutdinov, and A. Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–28, July 2017. 6

[31] Davide Mazzini and Raimondo Schettini. Spatial sampling network for fast scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[32] N. I. Nauata Junior, H. Hu, G. Zhou, Z. Deng, Z. Liao, and G. Mori. Structured label inference for visual understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 3

[33] L. Qi, Q. He, F. Chen, W. Dou, S. Wan, X. Zhang, and X. Xu. Finding all you need: Web apis recommendation in web of things through keywords search. *IEEE Transactions on Computational Social Systems*, pages 1–10, 2019. 3

[34] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Kegan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[35] M. J. Reale, P. Liu, L. Yin, and S. Canavan. Art critic: Multisignal vision and speech interaction system in a gaming context. *IEEE Transactions on Cybernetics*, 43(6):1546–1559, Dec 2013. 2

[36] Lukasz Romaszko, Christopher K. I. Williams, and John Winn. Learning Direct Optimization for Scene Understanding. *arXiv e-prints*, page arXiv:1812.07524, Dec. 2018. 1

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018. 1

[38] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, pages 593–607, Cham, 2018. Springer International Publishing. 2, 3, 5, 6

[39] Brigit Schroeder and Alexandre Alahi. Using a priori knowledge to improve scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[40] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, April 2017. 7

[41] Hengcan Shi, Hongliang Li, Qingbo Wu, and Zichen Song. Scene parsing via integrated classification model and variance-based regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[42] B. Shuai, H. Ding, T. Liu, G. Wang, and X. Jiang. Toward achieving robust low-level and high-level scene parsing. *IEEE Transactions on Image Processing*, 28(3):1378–1390, March 2019. 7

[43] M. Siddula, Y. Li, X. Cheng, Z. Tian, and Z. Cai. Anonymization in online social networks based on enhanced equi-cardinal clustering. *IEEE Transactions on Computational Social Systems*, pages 1–12, 2019. 3

[44] X. Sun, S. Shen, H. Cui, L. Hu, and Z. Hu. Geographic, geometrical and semantic reconstruction of urban scene from high resolution oblique aerial images. *IEEE/CAA Journal of Automatica Sinica*, 6(1):118–130, January 2019. 1

[45] X. Sun, L. Zhang, Z. Wang, J. Chang, Y. Yao, P. Li, and R. Zimmermann. Scene categorization using deeply learned gaze shifting kernel. *IEEE Transactions on Cybernetics*, 49(6):2156–2167, June 2019. 2, 3

[46] Maxwell H. Turner, Luis Gonzalo Sanchez Giraldo, Odelia Schwartz, and Fred Rieke. Stimulus- and goal-oriented frameworks for understanding natural vision. *Nature Neuroscience*, 22(1):15–24, 2019. 2, 3

[47] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1021–1028. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 3, 5

[48] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 432–448, Cham, 2018. Springer International Publishing. 1, 2, 3, 4, 5, 6, 7

[49] Y. Xiao, H. Yu, Q. Li, L. Liu, M. Xu, and H. Xiao. Mpurank: A social hotspot tracking scheme based on tripartite graph and multimessages iterative driven. *IEEE Transactions on Computational Social Systems*, pages 1–11, 2019. 3

[50] Y. Yao, W. Luo, L. Zhang, Y. Yang, P. Li, R. Zimmermann, and L. Shao. Learning latent stable patterns for image understanding with weak and noisy labels. *IEEE Transactions on Cybernetics*, pages 1–10, 2019. 1

[51] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[52] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[53] Oliver Zendel, Markus Murschitz, Marcel Zeilinger, Daniel Steininger, Sara Abbasi, and Csaba Beleznai. Railsem19: A dataset for semantic rail scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[54] 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 (CVPR)*, June 2018. 7

[55] L. Zhang, R. Liang, J. Yin, D. Zhang, and L. Shao. Scene categorization by deeply learning gaze behavior in a semisupervised context. *IEEE Transactions on Cybernetics*, pages 1–12, 2019. 1

[56] Pingping Zhang, Wei Liu, Hongyu Wang, Yinjie Lei, and Huchuan Lu. Deep gated attention networks for large-scale street-level scene segmentation. *Pattern Recognition*, 88:702 – 714, 2019. 7

[57] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[58] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan. Perspective-adaptive convolutions for scene parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 7

[59] Ruimao Zhang, Wei Yang, Zhanglin Peng, Pengxu Wei, Xiaogang Wang, and Liang Lin. Progressively diffused networks for semantic visual parsing. *Pattern Recognition*, 90:78 – 86, 2019. 7

[60] W. Zhang, W. Zhang, and J. Gu. Edge-semantic learning strategy for layout estimation in indoor environment. *IEEE Transactions on Cybernetics*, pages 1–10, 2019. 1, 3

[61] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for rgb-d scene understanding. *IEEE transactions on pattern analysis and machine intelligence*, July 2019. 1

[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5, 7

[63] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 1

[64] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 7

[65] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, July 2017. 6, 7

[66] W. Zhou, S. Lv, Q. Jiang, and L. Yu. Deep road scene understanding. *IEEE Signal Processing Letters*, 26(4):587–591, April 2019. 1

[67] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 7