Scene Parsing via Integrated Classification Model and Variance-Based Regularization

Hengcan Shi, Hongliang Li [*], Qingbo Wu, Zichen Song
University of Electronic Science and Technology of China
Chengdu, China
shihc@std.uestc.edu.cn, hlli@uestc.edu.cn, qbwu@uestc.edu.cn, szc.uestc@gmail.com

Abstract

Scene Parsing is a challenging task in computer vision, which can be formulated as a pixel-wise classification problem. Existing deep-learning-based methods usually use one general classifier to recognize all object categories. However, the general classifier easily makes some mistakes in dealing with some confusing categories that share similar appearances or semantics. In this paper, we propose an integrated classification model and a variance-based regularization to achieve more accurate classifications. On the one hand, the integrated classification model contains multiple classifiers, not only the general classifier but also a refinement classifier to distinguish the confusing categories. On the other hand, the variance-based regularization differentiates the scores of all categories as large as possible to reduce misclassifications. Specifically, the integrated classification model includes three steps. The first is to extract the features of each pixel. Based on the features, the second step is to classify each pixel across all categories to generate a preliminary classification result. In the third step, we leverage a refinement classifier to refine the classification result, focusing on differentiating the high-preliminary-score categories. An integrated loss with the variance-based regularization is used to train the model. Extensive experiments on three common scene parsing datasets demonstrate the effectiveness of the proposed method. [†]

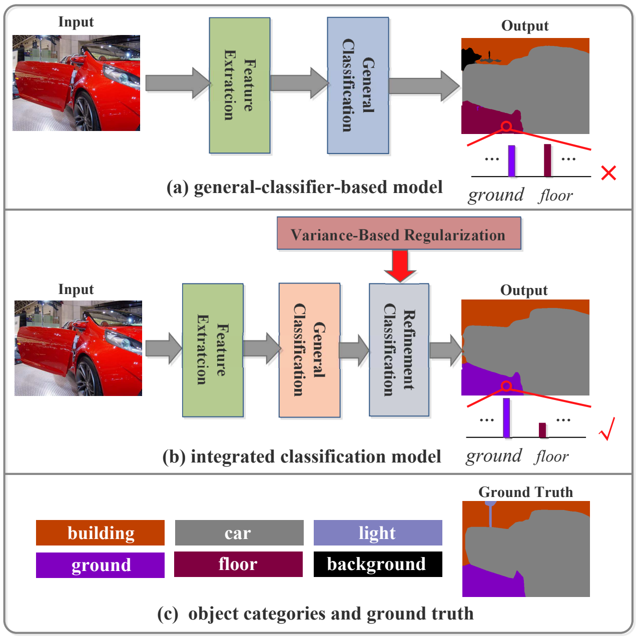


Figure 1. A comparison of general-classifier-based model and our proposed integrated classification model. (a) The general-classifier-based model. (b) The proposed integrated classification model. (c) The ground truth and object categories. The general-classifier-based model misclassifies the *ground* object as the *floor* object, whereas this misclassification is avoided by the integrated classification model. Moreover, the general-classifier-based model often predicts close scores for similar categories, while these scores are more different in the proposed method, benefiting from the variance-based regularization.

1. Introduction

Scene parsing expects to segment an entire image into multiple objects, which acts as a crucial component for many higher-level computer vision tasks, such as scene understanding [8, 20], object extraction [15, 26] and language-based vision analysis [11, 35]. The scene parsing task is usually formulated as a pixel-wise classification problem.

State-of-the-art scene parsing methods [1, 3, 5, 7, 10, 18, 23–25, 31, 36–38, 42, 44–46] mostly leverage deep neural network (DNN) to tackle this pixel-wise classification problem. These DNN-based methods encode the features of every pixel in an image and then classify these pixels by a general classifier, which focuses on classifying each pixel across all object categories by one step. However, as inves-

[*] Corresponding author.

[†] Code: https://github.com/shihengcan/ICM-matcaffe

tigated in [2], this strategy usually fails to distinguish the categories with similar appearances. For example, in Fig. 1, the general-classifier-based model mislabels the *ground* object as the *floor* object, which shares the similar shape and textures.

In this paper, we try to solve this problem from two aspects. Firstly, we propose an integrated classification model for scene parsing, which contains not only the general classification to recognize all object categories but also a refinement classification to distinguish confusing categories. Secondly, we observed that the scores of the confusing categories are usually close to each other within a classifier, which is also easy to result in misclassifications, such as the example in Fig. 1. Therefore, we propose a variance-based regularization to differentiate the scores of all categories as large as possible.

Specifically, the proposed scene parsing method can be divided into three steps. In the first step, we encode the features of all pixels by a deep learning network. Based on the extracted features, the second step is a general classification, which gives preliminary classification scores across all categories. In the preliminary scores, there may be more than one categories with the high score, which are confusing with respect to the general classifier. Therefore, in the third step, a refinement classifier is used to refine the scores, focusing on discriminating these confusing categories. To reduce the error accumulation between the two classifiers, we implement our general classifier with multiple binary classifiers rather than the commonly used multinomial classifier. An integrated classification loss with the variance-based regularization is used to train the integrated classification model to enhance its ability of differentiating similar categories. The proposed method is validated on three common scene parsing datasets, including the NYU Depth v2, Pascal-Context and SUN-RGBD datasets. The results show that our proposed method outperforms many state-of-the-art methods on these datasets.

This paper is organized as follows. The related work is introduced in Section 2. In Section 3, we detail the integrated classification model and variance-based regularization. Experimental results are reported in Section 4 to demonstrate the effectiveness of our method. Finally, Section 5 concludes this paper.

2. Related Work

In this section, we review recent advances in the scene parsing task. The existing methods [1, 3–7, 9, 10, 12–14, 18, 21–25, 27, 28, 30–32, 36–38, 42–47] formulate the scene parsing task as a pixel-wise classification problem and tackle this problem with the deep neural network (DNN). Long *et al.* [28] proposed a fully convolutional network (FCN) [28]. They leveraged a DNN to directly encode features for every pixel and then used a general classifier to classify these pixels. However, since too many downsampling operations are involved in the DNN, the final predictions generated by the FCN [28] usually lose some details, such as the small objects and exact object edges.

Many works [3, 5, 30, 43] attempt to enhance the resolution of predictions to retain more detailed information. Chen *et al.* [5] and Yu *et al.* [43] replaced a part of downsampling layers with atrous convolutions and dilated convolutions, respectively. Noh *et al.* [30] trained a deconvolutional network to restore the details in the predictions, which is the mirror of a convolutional neural network. Bilinski *et al.* [3] changed the connections in the deconvolutional network into dense connections to enable a fusion between different output resolutions. To further segment an image on different resolutions, many methods [6, 10, 18, 23, 24, 31, 37, 42, 45] proposed to use multi-scale strategies, including multi-scale averaging [24, 25, 31, 42] and scale learning [6, 10, 18, 23, 37, 45]. These methods are able to provide more detailed scene parsing results and thus reduce the over- and under-segmentations. However, the highly diverse relationships among the objects in the scene are ignored, which is useful to constrain the semantic consistency among the scene and every object.

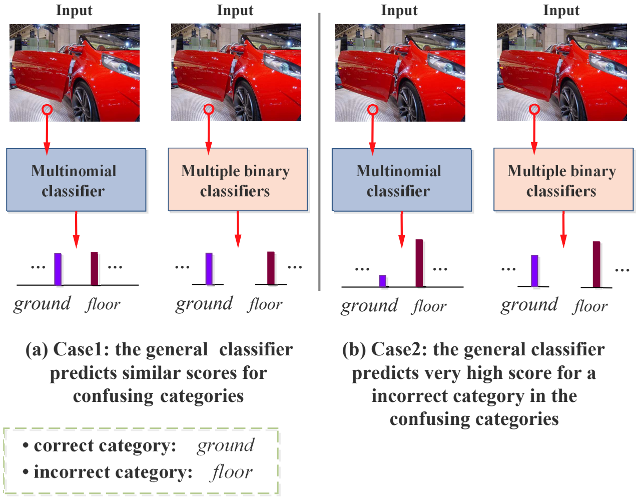
Some approaches [1, 7, 10, 21, 22, 24, 25, 27, 32, 36, 38, 44, 46, 47] model the relationships among the scene and objects by context models. Zheng *et al.* [47] and Lin *et al.* [25] leveraged the conditional random fields (CRFs) to model the relationships between each pair of pixels. In [7, 27, 46], convolutions with multi-size perspectives were used to model the hierarchical object relationships. Lin *et al.* [24] changed the size of input image instead of the size of convolution perspective to achieve the same goal as [7,27,46]. Zhang *et al.* [44] and Ding *et al.* [10] turned to use dictionary learning context and context contrasted local features to model these relationships, respectively. RNN-based context models were proposed by [21, 22, 32, 36, 38] to model the relationships including the relative positions of objects. Abdulnabi *et al.* [1] combined the RNN and attention model to learn more specific context. These methods also achieve remarkable performance for scene parsing.

However, these existing methods only adopt a general classifier, which is easy to confuse some categories with similar appearances or semantics. In order to solve this problem, this paper proposes an integrated classification model and a variance-based regularization to achieve more accurate classifications for the confusing categories.

3. Proposed Method

3.1. Integrated Classification Model for Scene Parsing

The scene parsing task can be formulated as a pixel-wise multinomial classification problem. Given an input image



(a) Case1: the general classifier predicts similar scores for confusing categories

(b) Case2: the general classifier predicts very high score for a incorrect category in the confusing categories

- correct category: *ground*
- incorrect category: *floor*

Figure 2. Comparison of the multinomial classifier and multiple binary classifiers in misclassification cases. (a) In the case of that the classifier generates similar scores for the correct and incorrect categories, the two types of classifiers show similar probability score distributions. (b) In the case of that the classifier predicts a very high score for the incorrect category, the multinomial classifier generate a lower score for the correct category than the multiple binary classifiers, due to the across-category competition.

I, our goal is to predict a conditional probability distribution $P(O|I,r)$ for each pixel r in the image. The random variable O can take the value in the set $\{o_c\}_{c=1,...,C}$, where C is the number of object categories in the dataset and $P(O=o_c|I,r)$ denotes the probability of that the r-th pixel belongs to the c-th category of objects. Once $P(O|I,r)$ is generated, the object category of the r-th pixel can be defined as follows:

$$\hat{c} = \arg\max_{c} P(O=o_c|I,r) \tag{1}$$

where $\hat{c}$ is the predicted object category of the r-th pixel in the image I.

In the proposed integrated classification model, we first leverage a general classifier to predict a preliminary probability distribution across all object categories and then use a refinement classifier to distinguish the high-score categories in the preliminary distribution.

To reduce the error accumulation between the general classifier and the refinement classifier, we adopt multiple binary classifiers as the general classifier instead of using a multinomial classifier. Since there are across-category competitions in the multinomial classifier, the misclassifications from the multinomial classifier result in more severe consequences than those from the binary classifiers. The reason is given as follows. Generally, there are two main cases of misclassifications. The first is that the classifier predicts a sightly higher score for the incorrect category than for the correct category, such as shown in Fig 2(a). In this case, misclassifications can be corrected by the refinement classifier. The second case can be found in Fig 2(b), where the classifier predicts a very high score for the incorrect category. In this case, the multinomial classifier would yield a very low score for the correct category because of the across-category competitions. This probability score distribution is hard to be corrected by the refinement classifier. In contrast, in the multiple binary classifiers, the scores of categories do not affect each other. Hence, the binary classifiers generate a relatively higher score for the correct category than the multinomial classifier, which is more favorable to the next refinement.

To leverage multiple binary classifiers as the general classifier, it is necessary to convert the multinomial classification problem into multiple binary classification problems. We rewrite each probability $P(O=o_c|I,r)$ in the distribution as an equivalent form $P(Y_1=0,...,Y_{c-1}=0,Y_c=1,Y_{c+1}=0,...,Y_C=0|I,r)$, where $Y_i\in\{0,1\}(i=1,...,C)$ denotes whether the pixel r belongs to the i-th object category. Based on this probability form, we decompose this multinomial classification problem into C binary classification problems, in which each binary classification problem determines the probabilities of whether the r-th pixel belongs to the c-th object category, namely $P(Y_c=1|I,r)$ and $P(Y_c=0|I,r)$. For simplicity, we use $p_{r,c}$, $p_{r,c}^{fg}$ and $p_{r,c}^{bg}$ to represent these probabilities, i.e.:

$$\begin{aligned} p_{r,c} =& P(Y_1=0,...,Y_{c-1}=0,Y_c=1, \\ & Y_{c+1}=0,...,Y_C=0|I,r) \end{aligned} \tag{2}$$

$$p_{r,c}^{fg} = P(Y_c=1|I,r) \tag{3}$$

$$p_{r,c}^{bg} = P(Y_c=0|I,r). \tag{4}$$

Then, the general classification can be formulated as follows:

$$\{p_{r,1}^{fg},...,p_{r,C}^{fg},p_{r,1}^{bg},...,p_{r,C}^{bg}\} = gcls(I,r) \tag{5}$$

where, $gcls(\cdot)$ denotes the general classifier.

After predicting general classification probabilities $p_{r,c}^{fg}$ and $p_{r,c}^{bg}$ for each category, a refinement classifier then generates the final probability distribution $P(O|I,r)$ as follows:

$$\begin{aligned} P(O|I,r) &= rcls(I,r,p_{r,1}^{fg},...,p_{r,C}^{fg},p_{r,1}^{bg},...,p_{r,C}^{bg}) \\ &= \{p_{r,1},p_{r,2},...,p_{r,C}\} \end{aligned} \tag{6}$$

where $rcls(\cdot)$ is the refinement classifier. We employ a multinomial classifier as the refinement classifier, which takes the features of the r-th pixel in the image I and the general classification probabilities $\{p_{r,c}^{fg},p_{r,c}^{bg}\}_{c=1,...,C}$ as inputs, and then outputs the final probability distribution. The refinement classifier focuses on differentiating the categories with high general classification scores (i.e., high

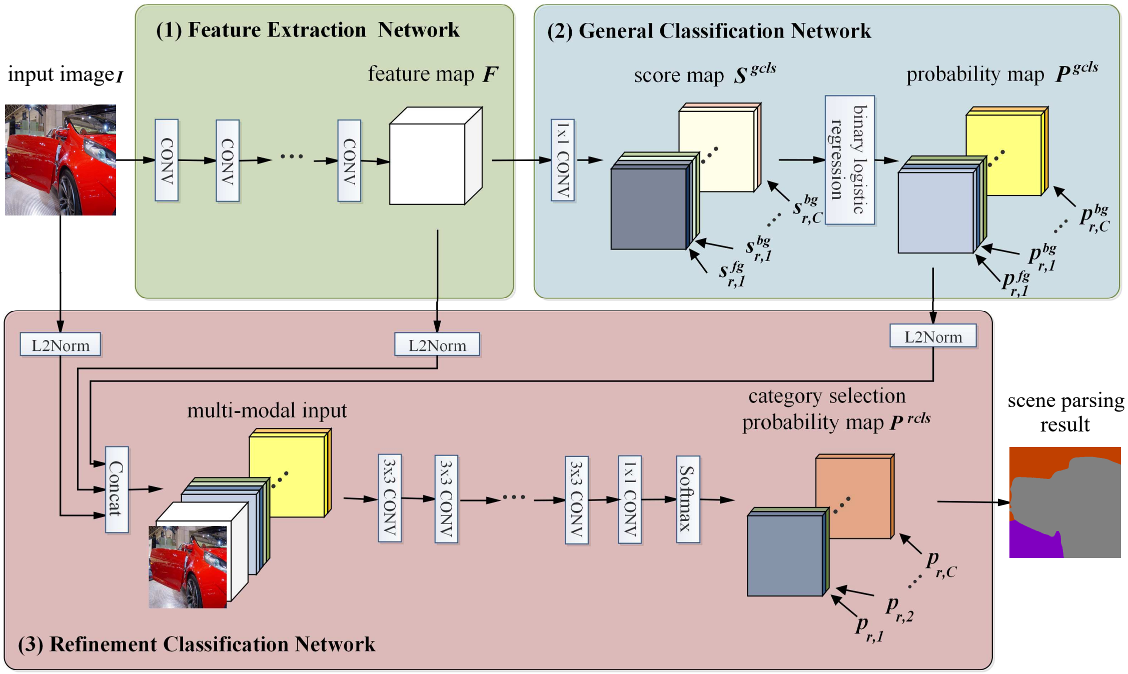


Figure 3. The proposed integrated classification model contains three parts: (1) a feature extraction network that encodes features of each pixel, (2) a general classification network that generates a preliminary probability distribution across all object categories, and (3) a refinement classification network that differentiates the high-preliminary-score categories and refines the probability distribution.

$p^{fg}_{r,c}$), but is not limited to this. It also has the ability of the second general classification to avoid the error accumulation caused by misclassifications from the general classifier.

We next illustrate how to implement the proposed integrated classification model with the deep neural network.

3.2. Integrated Classification Model Implemented with Deep Neural Network

The network structure of the proposed integrated classification model contains three parts, as illustrated in Fig. 3: (1) a feature extraction network that encodes features of the input image, (2) a general classification network that predicts a preliminary probability distribution, and (3) a refinement classification network that distinguishes the confusing categories and generates the final probability distribution.

3.2.1 Feature Extraction Network

Consider the input image $I \in R^{H \times W \times D}$, where H and W are the height and width of the image, respectively; and D is the number of channels. We first use the feature extraction deep neural network to encode features of this image as follows:

$$\begin{aligned} F &= DNN(I) \\ &= \{f_1, f_2, ..., f_{HW}\} \end{aligned} \tag{7}$$

where $F \in R^{H \times W \times d}$ is the encoded feature map, in which each feature vector $f_r \in R^d$ $(r = 1, ..., HW)$ encodes the appearance and semantic information of a region around the r-th pixel; and d is the dimensionality of each feature vector. The feature extraction deep network can be implemented with any network structure, such as the commonly used VGG [40] and ResNet [16].

3.2.2 General Classification Network

Based on the feature map F, we employ a series of binary classifiers as a general classifier to determine a preliminary probability distribution. Each binary classifier predicts a pair of foreground and background scores for each object category c as follows:

$$s^{fg}_{r,c} = (w^{fg}_c)^T f_r + b^{fg}_c \tag{8}$$

$$s^{bg}_{r,c} = (w^{bg}_c)^T f_r + b^{bg}_c \tag{9}$$

where $c = 1, .., C$. Here, C is the number of object categories in the dataset. For the c-th category, $s^{fg}_{r,c} \in R$ denotes the predicted score of the r-th pixel belonging to this category, while $s^{bg}_{r,c} \in R$ is the score of the opposite case. $w^{fg}_c, w^{bg}_c \in R^d$ and $b^{fg}_c, b^{bg}_c \in R$ are the parameters in the binary classifier. The scores for all pixels and all categories can be grouped into a score map $S^{gcls} \in R^{H \times W \times 2C}$, and the corresponding binary classifiers can be efficiently implemented with a convolutional layer, as shown in Fig. 3(a).

The foreground and background scores are then normalized into the form of probability by a binary logistic regression:

$$p^{fg}_{r,c} = \frac{exp(s^{fg}_{r,c})}{exp(s^{fg}_{r,c}) + exp(s^{bg}_{r,c})} \tag{10}$$

$$p^{bg}_{r,c} = \frac{exp(s^{bg}_{r,c})}{exp(s^{fg}_{r,c}) + exp(s^{bg}_{r,c})} \tag{11}$$

where $p^{fg}_{r,c}$ and $p^{bg}_{r,c}$ are the normalized foreground and background probabilities, respectively. Similar to the score map S, the probabilities can be grouped into a probability map $P^{gcls} \in R^{H\times W\times 2C}$. In the general classification network, the predictions of a category do not compete with ones of another category.

3.2.3 Refinement Classification Network

The refinement classification network refines probability distribution by differentiating the high-preliminary-score categories and the second general classification. We employ a multinomial classifier to achieve this goal.

The input of the refinement classification network is a concatenation of the binary classification probability map P^{gcls}, the image feature map F and the orignal image I, where the feature map F and the orignal image I are references to assist with the classification. Note that since the value ranges of P^{gcls}, F, and I may be different, we L2-normalize them before the concatenation.

The concatenated multi-modal input is then transformed by a series of convolutional layers. We use 3×3 convolutions in these layers to model contextual information among multiple pixels. Based on the transformed feature map, we employ an 1×1 convolution to generate the refined classification score map $S^{rcls} \in R^{H\times W\times C}$, where each element $s_{r,c}$ indicates the score of the r-th pixel belonging to the c-th category. The refined scores are normalized by a multinomial logistic regression (i.e., softmax) as follows:

$$p_{r,c} = \frac{exp(s_{r,c})}{\sum_{t=1}^{C} exp(s_{r,t})} \tag{12}$$

where $p_{r,c}$ is the final probability of the r-th pixel belonging to the c-th category. The probability set $\{p_{r,c}\}_{c=1,..,C}$ denotes the desired probability distribution $P(O|I,r)$.

3.3. Loss Function and Variance-Based Regularization

We use an integrated classification loss with the variance-based regularization to end-to-end train our full model:

$$L = L_{gcls} + \lambda_{rcls} L_{rcls} + \lambda_{vbr} L_{vbr} \tag{13}$$

where L_{gcls} and L_{rcls} denote the losses for the general classifier and the refinement classifier, respectively; L_{vbr} is the variance-based regularization to further reduce misclassifications; λ_{rcls} and λ_{vbr} are the factors controlling the relative importance among these losses and the regularization.

As the general classifier is composed of multiple binary classifiers, we train it in terms of an average loss of multiple binary cross entropy losses, in which each binary cross entropy loss corresponds to an object categories:

$$\begin{aligned} L_{gcls} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{HW}\sum_{r=1}^{HW}\frac{1}{C}\sum_{c=1}^{C}[y_{i,r,c} \times log(p^{fg}_{i,r,c}) \\ + (1 - y_{i,r,c}) \times log(p^{bg}_{i,r,c})] \end{aligned} \tag{14}$$

where N is the number of images in the training set; and $y_{i,r,c} \in \{0,1\}$ is the scene parsing label, which indicates whether the r-th pixel in the i-th image belongs to the c-th object category.

The loss L_{rcls} for the refinement classifier is formulated as a multinomial cross entropy loss as follows:

$$L_{rcls} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{HW}\sum_{r=1}^{HW}\sum_{c=1}^{C} y_{i,r,c} \times log(p_{i,r,c}). \tag{15}$$

Since similar probability scores of multiple categories may lead to misclassifications, we propose a variance-based regularization L_{vbr} to avoid this case. The variance-based regularization L_{vbr} constrains the scores of different categories to be as variant as possible. In this paper, inspired by the Herfindahl-Hirschman Index (HHI) [33] in economics, we adopt the second-order moment as the variance-based regularization L_{vbr}:

$$L_{vbr} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{HW}\sum_{r=1}^{HW}\sum_{c=1}^{C}(p_{i,r,c})^2 \tag{16}$$

where $L_{vbr} \in [0, 1 - 1/C]$ decreases with increasing variances among the probabilities $\{p_{i,r,c}\}_{c=1,...,C}$.

4. Experiments

In this section, we validated the proposed integrated classification model and variance-based regularization on multiple scene parsing datasets, including the NYU Depth v2 dataset [39], the PASCAL-Context dataset [29] and the SUN-RGBD dataset [41].

Datasets. The NYU Depth v2 dataset [39] contains 1449 pairs of RGB and depth images, where 795 pairs for training and 654 pairs for testing. We use 40 categories object labels as the same as [13]. Only RGB images and scene parsing labels are used to train the proposed model. The PASCAL-Context dataset [29] contains 10103 images. It is split into training and testing sets, including 4998 and 5105 images, respectively. We use 60 object category labels provided by [29]. The SUN-RGBD dataset [41] includes 37

Method	pAcc.(%)	mIoU(%)
FCN [28]	60.0	29.2
DilatedNet [43]	65.4	33.7
Context [25]	70.0	40.6
DeepLab v2 [5]	71.7	42.3
RefineNet [24]	73.6	46.5
PSPNet [46]	73.6	46.9
LoopNet [18]	72.1	44.5
Dense Decoder [3]	73.8	48.1
DeepLab v3+ [7]	73.8	47.4
Ours	**75.4**	**50.7**

Table 1. Comparison with state-of-the-art methods on the NYU Depth v2 dataset. LoopNet [18] is trained by both scene parsing and depth prediction labels, and all other methods are trained only by the scene parsing labels.

Method	pAcc.(%)	mIoU(%)
FCN [28]	65.9	35.1
DilatedNet [43]	66.4	37.0
Episodic CAMN [1]	72.1	41.2
Context [25]	71.5	43.3
DeepLab v2 [5]	73.6	44.5
RefineNet [24]	75.1	47.3
PSPNet [46]	75.1	47.0
Dense Decoder [3]	74.9	47.8
EncNet [44]	78.2	51.7
CCL&GMA [10]	78.4	51.6
DeepLab v3+ [7]	75.5	47.4
Ours	**80.5**	**52.6**

Table 2. Comparison with state-of-the-art methods on the PASCAL-Context dataset.

Method	pAcc.(%)	mIoU(%)
FCN [28]	68.2	27.4
Context [25]	78.4	42.3
DeepLab v2 [5]	71.9	32.1
RefineNet [24]	80.6	45.9
PSPNet [46]	79.7	46.2
CCL&GMA [10]	81.4	47.1
DeepLab v3+ [7]	80.5	46.7
Ours	**82.4**	**50.6**

Table 3. Comparison with state-of-the-art methods on the SUN-RGBD dataset.

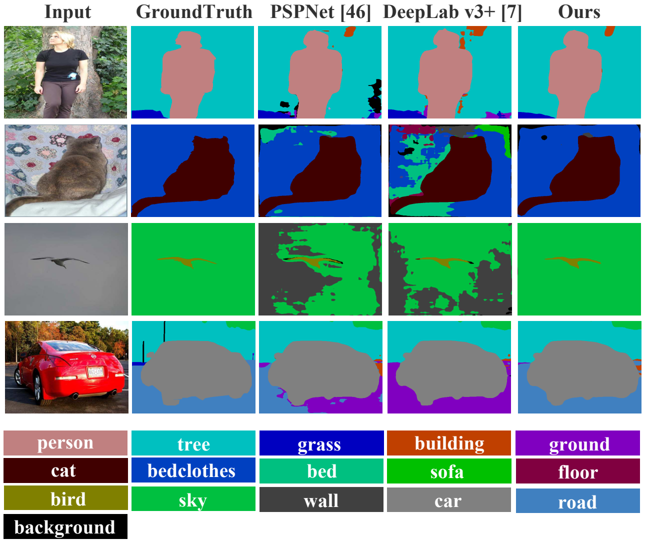


Figure 4. Visual comparison on the Pascal-Context dataset. Left to right: input images, ground truth, and results from PSPNet [46], DeepLab v3+ [7] and our method.

object categories. There are 10335 pairs of RGB and depth images in this dataset, 5285 pairs for training and 5050 pairs for testing. Here, we only use the RGB images in the experiments.

Evaluation Metrics. We adopted the pixel-wise accuracy (pAcc.) and the mean intersection-over-union (mIoU) metrics to evaluate the scene parsing performance. The pixel-wise accuracy is the percentage of correctly classified pixels in the entire dataset. The mean intersection-over-union is the average of the intersection-over-union between the predictions and ground-truths over all categories in the dataset.

Details of Implementation. We implement the proposed method with the Caffe [17] deep learning toolkit. We employed PSPNet [46] as the feature extraction deep network, which is implemented with the ResNet [16]. Meanwhile, a three-layer neural network is designed as the refinement classification network. Note that we used PSPNet [46] and the three-layer refinement classification network as a running example, which does not mean that the proposed model is limited to these networks. The feature extraction deep network is initialized from the weights pre-trained on ImageNet dataset [34], and other parts are initialized from random weights. We trained the proposed model end-to-end with the stochastic gradient descent (SGD) and the integrated classification loss including the variance-based regularization. The base learning rate was set to 0.00025, and the learning rates of the randomly initialized layers were 10 times higher than those of the pre-trained layers. The loss factors λ_{rcls} and λ_{vbr} were set to 1 and 0.2, respectively.

4.1. Comparison with State-of-the-art Methods

We compare the proposed method with nine state-of-the-art scene parsing methods on the NYU Depth v2 dataset. The results are shown in Table 1. All these methods only use a general classifier to classify each pixel, in which Dense Decoder [3] shows the best performance. Compared with Dense Decoder [3] , the proposed integrated classification model achieves improvements of 1.6% and 2.6% in terms of the pixel-wise accuracy and mean IoU, respectively. Even LoopNet [18] is trained by both scene parsing and depth prediction labels, the proposed model also outperforms it by 3.3% and 6.2% in terms of the pixel-wise

Method	Variance-based regularization	pAcc.(%)	mIoU(%)	Training Speed(Hz)	Testing Speed(Hz)	Params($\times 10^6$)
Baseline model [46]		73.6	46.9	4.3	9.6	65.7
Baseline model [46] + DSN [19]		73.8	46.5	4.2	9.7	65.7
Baseline model [46] + more layers		73.3	47.1	3.9	9.0	80.8
Integrated classification model		75.0	50.3	3.2	9.0	80.7
Baseline model [46]	✓	73.8	47.8	4.2	9.6	65.7
Integrated classification model	✓	**75.4**	**50.7**	3.2	9.0	80.7

Table 4. The effects of main components in the proposed method on the NYU Depth v2 dataset.

Method	Type of general classifier	pAcc.(%)	mIoU(%)	Mean difference
Baseline model [46]	Multinomial Classifier	73.6	46.9	0.596
Integrated classification model	Multinomial Classifier	74.8	49.4	0.579
Integrated classification model	Multiple Binary Classifiers	**75.4**	**50.7**	0.414

Table 5. Comparison of different general classifiers on the NYU Depth v2 dataset. “Mean difference” is the the mean difference between the correct and incorrect probability scores when incorrect predictions are generated.

accuracy and mean IoU, respectively. This superior performance demonstrates the effectiveness of the proposed integrated classification model and variance-based regularization.

The results of comparative experiments conducted on the PASCAL-Context dataset and the SUN-RGBD dataset are shown in Table 2 and Table 3, respectively. On the PASCAL-Context dataset, compared with the previous state-of-the-art results, the proposed method achieves improvements of 2.1% and 0.9% in terms of the pixel-wise accuracy and mean IoU, respectively. On the SUN-RGBD dataset, our proposed method outperforms the previous state of the art by 1.0% pixel-wise accuracy and 3.5% mean IoU.

We depict some visualized scene parsing results in Fig. 4. It can be observed that the general-classifier-based methods mislabel some objects. For example, in the third image in Fig. 4, the *sky* object is mislabeled as the *wall* object; in the fourth image in Fig. 4, the *road* object is misclassified as the *ground* object. The proposed method successfully avoids such misclassifications, benefiting from the integrated classification and the variance-based regularization.

4.2. Ablation Study

In this section, we conduct a series of ablation experiments to further evaluate the effectiveness of our proposed method.

Effects of main components. We give the effects of our main components in Table 4. Without the variance-based regularization, compared with the baseline general-classifier-based model [46], the proposed integrated classification model improves pixel-wise accuracy and mean IoU by 1.4% and 3.4%, respectively. This demonstrates the effectiveness of the proposed integrated classification model.

DSN [19] adds multiple classifiers to different layers as supervision in the training stage. Our method outperforms “Baseline model [46] + DSN [19]”, because our method not only uses the classification results as supervision but also corrects these results in the test stage to promote the classification accuracy. In “Baseline model [46] + more layers”, we add more parameters to the baseline [46]. Compared with this method, our method achieves gains of 1.7% and 3.2%, respectively, in terms of pixel-wise accuracy and mean IoU. This result demonstrates that the proposed method improves parsing accuracy mainly through integrating multiple classifiers rather than adding more parameters.

Moreover, the baseline model [46] and proposed model trained with the variance-based regularization both achieve the better pixel-wise accuracy and mean IoU than them trained without the regularization. Ultimately, our full method (the integrated classification model and variance-based regularization) outperforms the baseline model [46] by 1.8% and 3.8% in terms of the pixel-wise accuracy and mean IoU, respectively.

Computation costs. Table 4 shows the computation cost of our method. It can be seen that the proposed method improves the scene parsing accuracy with acceptable computation overhead.

Effects of different general classifiers. The effects of different general classifiers are shown in Table 5. It can be observed that whether a multinomial classifier or multiple binary classifiers are used as our general classifier, the proposed integrated classification model provides better performance than the baseline model [46]. Using multiple binary classifiers achieves more improvements. The reason is that the multiple binary classifiers are easier to avoid error accumulation than the multinomial classifier, as explained in the Section 3.1. In Table 5, we depict the mean difference between the correct and incorrect scores when predictions are

The number of layers	Dimensionality of hidden layers	pAcc.(%)	mIoU(%)
1	/	74.1	49.0
2	256	73.9	48.9
2	512	74.1	49.3
2	4096	74.9	49.9
3	256	74.5	49.8
3	512	75.0	50.6
3	1024	75.4	50.7
3	2048	**75.5**	50.6
3	4096	75.2	**51.0**

Table 6. Comparison of different numbers of layers and different hidden layer dimensionalities of the refinement classification network on the NYU Depth v2 dataset.

General classification probability map	Image feature map	Original image	pAcc.(%)	mIoU(%)
✓			74.7	50.1
✓	✓		74.8	50.2
✓		✓	75.1	49.9
✓	✓	✓	**75.4**	**50.7**

Table 7. The effects of different inputs of the refinement classification network on the NYU Depth v2 dataset.

incorrect. It can be seen that the mean difference obviously decreases by using multiple binary classifiers.

Structures of the refinement classification network. Table 6 shows the effects of different structures of the refinement classification network. It can be seen that the best pixel-wise accuracy and mean IoU are achieved by the three-layer networks with 2048- and 4096-dimensional hidden layers, respectively. Compared with these structures, the three-layer network with 1024-dimensional hidden layers shows comparable performance but less computation costs. To balance the performance with the computation costs, we finally adopt the three-layer network with 1024-dimensional hidden layers in other experiments.

Inputs of the refinement classification network. The effects of different inputs of the refinement classification network are listed in Table 7. From Tables 7 and 4, it can be observed that when we only input the general classification probability map to the refinement classification network, the proposed model outperforms the baseline model [46] by 1.1% and 3.2% in terms of the pixel-wise accuracy and mean IoU, respectively. Inputting the image feature map and original image further improves the scene parsing accuracy, because they can be regard as references to assist with the classification. The best pixel-wise accuracy and mean IoU are achieved when the general classification probability map, image feature map and original image are input together.

Effects of different λ_{vbr}. Table 8 shows how the scene parsing performance is affected by the weight λ_{vbr} of the variance-based regularization in the integrated loss function. It can be observed that the best pixel-wise accuracy and mean IoU are achieved when λ_{vbr} is 0.2.

λ_{vbr}	0	0.1	0.2	0.5	1	2
pAcc.(%)	75.0	75.1	**75.4**	75.3	74.8	74.9
mIoU(%)	50.3	50.1	**50.7**	50.1	50.0	49.7

Table 8. The effects of different factor λ_{vbr} in the integrated loss function on the NYU Depth v2 dataset.

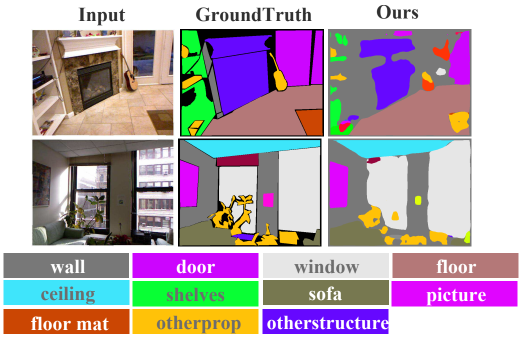


Figure 5. Parsing failures of the proposed method on the NYU Depth v2 dataset.

Analysis of failure cases. We display some parsing failures of the proposed method in Fig. 5. The first type of parsing errors is over-segmentation when objects contains various colors, such as the *otherstructure* (*fireplace*) object in the first image in Fig. 5. Another type of parsing errors is imprecise segmentation of delicate object edges, such as the *otherprop* (*plant*) object in the third image in Fig. 1. These problems may be alleviated by parsing the scene on different scales.

5. Conclusion

In this paper, we have presented an integrated classification model and a variance-based regularization for the scene parsing task. The integrated classification model first encodes features of each pixel. Then a series of binary classifiers are used to classify these pixels across all object categories. Based on the results of the general classification, we finally leverage a refinement classifier to discriminate the confusing categories. The variance-based regularization is used to train the proposed integrated classification model to differentiate the classification scores of all categories to be as large as possible. We have demonstrated the effectiveness of our method on three common scene parsing datasets. In the future, we hope to fuse multi-scale methods with our model to reduce over-segmentations and imprecise object edges.

Acknowledgement. This work was supported in part by National Natural Science Foundation of China (No.61831005, 61525102, 61601102 and 61871078).

References

[1] A. H. Abdulnabi, B. Shuai, S. Winkler, and G. Wang. Episodic camn: Contextual attention-based memory networks with iterative feedback for scene labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5561–5570, 2017.

[2] K. Ahmed, M. H. Baig, and L. Torresani. Network of experts for large-scale image categorization. In *Proceedings of the European Conference on Computer Vision*, pages 516–532. Springer, 2016.

[3] P. Bilinski and V. Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6596–6605, 2018.

[4] H. Caesar, J. Uijlings, and V. Ferrari. Region-based semantic segmentation with end-to-end training. In *Proceedings of the European Conference on Computer Vision*, pages 381–397. Springer, 2016.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, 2018.

[8] Y. Chen, W. Li, and L. Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[9] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3992–4000, 2015.

[10] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018.

[11] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1811–1820, 2017.

[12] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149, 2015.

[13] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[14] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[15] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang. Reinforcement cutting-agent learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9080–9089, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[18] S. Kong and C. C. Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Proceedings of the Artificial Intelligence and Statistics*, pages 562–570, 2015.

[20] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng. Foveanet: Perspective-aware urban scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 784–792, 2017.

[21] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph lstm. In *Proceedings of the European Conference on Computer Vision*, pages 125–143. Springer, 2016.

[22] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3185–3193, 2016.

[23] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1311–1319, 2017.

[24] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[25] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Exploring context with deep structured models for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1352–1366, 2018.

[26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[27] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015.

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[29] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[30] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[31] S.-J. Park, K.-S. Hong, and S. Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4980–4989, 2017.

[32] Z. Peng, R. Zhang, X. Liang, X. Liu, and L. Lin. Geometric scene parsing with hierarchical lstm. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3439–3445, 2016.

[33] S. A. Rhoades. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79:188, 1993.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[35] H. Shi, H. Li, F. Meng, and Q. Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the Proceedings of the European Conference on Computer Vision*, pages 38–54, 2018.

[36] H. Shi, H. Li, F. Meng, Q. Wu, L. Xu, and K. N. Ngan. Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Transactions on Multimedia*, 2018.

[37] H. Shi, H. Li, Q. Wu, F. Meng, and K. N. Ngan. Boosting scene parsing performance via reliable scale prediction. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 492–500. ACM, 2018.

[38] B. Shuai, Z. Zuo, B. Wang, and G. Wang. Scene segmentation with dag-recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1480–1493, 2018.

[39] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[41] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[42] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[43] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

[44] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2031–2039, 2017.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[47] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.