IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 14, 2020, accepted March 30, 2020, date of publication April 2, 2020, date of current version April 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985227

Overall Understanding of Indoor Scenes by Fusing Multiframe Local RGB-D Data Based on Conditional Random Fields

HAOTIAN CHEN[1], LONGFEI SU[2], BIAO ZHANG[1], FENGCHI SUN[2], (Member, IEEE), JING YUAN[3], (Member, IEEE), AND JIE LIU[3]

[1]College of Computer Science, Nankai University, Tianjin 300350, China
[2]College of Software, Nankai University, Tianjin 300350, China
[3]College of Artificial Intelligence, Nankai University, Tianjin 300350, China

Corresponding author: Fengchi Sun (fengchisun@nankai.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873327, and in part by the National Key Research and Development Plan under Grant 2018YFB0204304.

ABSTRACT Indoor mobile robots normally cannot capture the whole information of a scene by a single frame of perceptive data due to the limited sensor scope. The category of the current scene may be misjudged by robotics due to incomplete scene information, which leads to operation error. To address this problem, we propose an approach that leverages conditional random fields (CRFs) to fuse multiframe RGB and depth (RGB-D) visual data corresponding to the same scene. This method takes full advantage of prior knowledge that object categories significantly relate to the scene attributes. As a new image arrives, we incrementally integrate the current object detection results to update scene understanding by identifying duplicate objects between images, ranking available objects in terms of their relevance to the scene, and fusing new information with the existing CRF. With this approach, scene classification can be solved with higher precision based on multiview images than on single image frames sampled in the same places. Additionally, a configuration map of a scene is incrementally built into the above framework. The map includes identities of the recognized objects and various relations between them. This kind of map would not only benefit common robotic tasks but also offer a novel channel for human-robot interaction. We test the efficiency of our method on image sequences extracted from the NYU v2 dataset. The results show that our approach achieves the best performance against state-of-the-art baselines.

INDEX TERMS Conditional random fields, multiframe image fusion, scene configuration map, scene understanding.

I. INTRODUCTION

In recent years, indoor scene understanding has aroused immense interest in robotics research [9]. This technique enables service robots to understand their environment better, interact with humans and implement tasks more efficiently [3]. For example, when we ask a robot to take a book left in the bedroom, it can perform better if it knows where the bedroom is. In fact, the robot can learn the attributes of its surroundings by use of scene classification algorithms. Previous indoor scene classification systems are mainly based on local or global image features, i.e., local features scale-invariant feature transform (SIFT) [8] and GIST [11]. The performance of these systems is limited without using the high-level semantic knowledge of the scene. For instance, the category of a room is often determined by specific objects inside it. Recently, scene understanding algorithms using high-level semantic information have emerged. In particular, Lin *et al.* [7] proposed a conditional random fields (CRF) model to jointly implement scene classification and object recognition in a framework.

Nevertheless, indoor scene understanding remains a difficult task for robots. On the one hand, the composition of indoor scenes is more complicated than that of outdoor scenes. Due to the complexity and diversity of indoor scenes, the difference is large within the same class and small between different classes, which is difficult for scene classification algorithms. On the other hand, for spatial occlusion

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

and the limited detection scope of sensors, a single frame of perception data may not be sufficient to support a description fully covering the whole scene in many cases.

Furthermore, research on image sequences has been increasing. References [22], [23] used different methods to achieve good performance on optical flow estimation of image sequences. Reference [24] proposed a 3D geometry estimation method based on image sequences. Reference [25] presented a novel approach for unsupervised learning of depth and ego-motion from monocular image sequences. Reference [26] proposed a method combining convolutional neural networks (CNNs) and a state-of-the-art dense simultaneous localization and mapping (SLAM) system based on image sequences. However, there are still few studies on indoor scene understanding by using image sequences to obtain accurate classification results of the scenes.

For the above reasons, this paper improves the method proposed by Lin *et al.* [7] and presents a more powerful CRF model that fuses multiframe RGB and depth (RGB-D) images for scene understanding. First, we use two neural networks to perform object recognition and scene classification for each frame in the image sequence of the same scene. Based on this, to comprehensively infer the scene category, we build a compound semantic scene model from the image sequence. The model is completed by identifying and merging the duplicate objects in different images and, furthermore, removing the objects that have less impact on scene classification to increase the difference between scene classes. The above result is used as one of the object unary potentials in our multiframe fusion oriented CRF model. Additionally, we merge the scene classification results for images appearing in the sequence in terms of the 'importance index' of each frame and use the merged result as scene unary potential. Finally, we use the scene-object binary potential, which is estimated from the training set by counting the co-occurrence frequencies between certain scene-object pairs. This binary potential can fuse original object recognition and scene classification results and help to determine the scene category corresponding to the image sequence. Notably, this algorithm incrementally updates as a new image arrives by computing the latest unary potentials of objects and of the scene to obtain an overall recognition over the scene. We divide the NYU v2 [16] dataset into image sequences and test the efficiency of our approach. The experimental results show that our model achieves the best performance against state-of-the-art baselines.

To make scene understanding results more applicable, we build a scene configuration map while fusing images for scene understanding. After obtaining the results of object recognition and scene classification for the first image, we extract the relationship between objects on the single image and use this information to build the original configuration map. Incrementally, we fuse configuration maps of incoming images and, finally, obtain the configuration map of the whole scene. The configuration map can be used to help the robot conduct topology positioning, object searching, man-robot interaction, etc. Since the configuration map does not need specific positioning information, it is not sensitive to local changes such as object movement. Correspondingly, the configuration map has the advantages of better robustness and less time consumption for building than the existing compound semantic maps [4].

The rest of the article is organized as follows: Section II briefly reviews the existing work related to this paper. Section III introduces the multiframe fusion oriented CRF model, including the potential definition and incremental update of the model. In Section IV, we show how to build a configuration map and compare the proposed methods with existing representative methods on the NYU v2 dataset. Finally, Section V concludes and discusses future work.

II. RELATED WORK

The work involves research activities in three areas: object detection, scene classification, and semantic map building. With the rapid development of computer vision and robot SLAM, there have been many papers that address scene understanding. In this section, we discuss some of those that are closely related to this topic.

A. OBJECT DETECTION

In general, existing object detection methods can be divided into two categories: traditional methods and methods based on neural networks. In traditional methods, object detection results are obtained by the following steps [17], [18]. (1) Extract object rectangles from the pyramid of the image by using some prior knowledge. (2) Design features for object detection and extract the features from the object rectangles. (3) Design a classifier to determine the object category of the extracted rectangles.

Neural network-based methods can be divided into single-stage methods such as YOLO [14] and two-stage methods such as Faster R-CNN [15]. In two-stage methods, first obtain proposal regions where an object may exist, and then use a neural network to extract features and decide categories of proposal regions. In single-stage methods, the region proposal step is discarded, and the result of object detection is obtained in a single branchless convolutional network.

Methods based on neural networks in recent years have displayed better performance than traditional methods. Among them, the detection speed of single-stage methods is faster than two-stage methods, whereas two-stage methods work better when objects are relatively small in images. Considering both the speed and accuracy of the methods, we decide to use Fast R-CNN [2] as the method for acquiring the segmentation potential in our CRF model.

B. SCENE RECOGNITION

In terms of feature level, the methods of scene recognition can be classified into four major categories: methods based on underlying features, methods based on middle-layer features, methods based on high-level features, and methods based on multiple-level features.

Methods based on underlying features extract global or local appearance features such as color, edge, and texture, from the images and then use the extracted features to train a classifier such as KNN and SVM to classify the images. Many different feature configurations are proposed, and several remarkable configurations are widely used. For instance, the low-dimensional global feature description operator GIST in scale space [11], the local feature based on gradient histogram [1], and the local feature descriptor SIFT [8] are invariant to uniform scaling, orientation, and illumination changes. Since the release of the Places dataset, neural networks such as AlexNet [5] and ResNet [27] have been used to extract underlying features and predict scene labels [20].

Methods based on middle-layer features are proposed to bridge the gap between the underlying features and global context. The representative is the bag-of-visual-words (BoVW) [10]. In BoVW, images are treated as a collection of independent vocabularies and images are classified by transferring a text categorization method to scene classification.

Methods based on high-level features are consistent with the human cognitive principle. The objects in the scene are used as features to train a classifier. The typical method is the object bank method proposed by Li *et al.* [6], which first detects objects and then treats object classes as features to feed the scene classifier.

The underlying, middle, and high-level features of the scene are comprehensively utilized by methods based on multiple-level features. For instance, Lin *et al.* [7] uses CRF to combine the underlying and high-level features for determining the scene category of an image.

In general, by the existing scene recognition methods, different image features are utilized, and multiple-level features can offer a better overall description of the scene than other feature forms. Based on this consideration, we extend the method proposed in [7] and adapt it to the case with multiframe RGB-D image data for the same scene.

C. BUILDING SEMANTIC MAP

There are two different methods for building semantic maps in the robotics domain. The popular methods are based on simultaneous localization and mapping (SLAM). These methods use metric maps obtained by SLAM and add semantic information [12], [19]. The other methods do not involve low-level metric maps. For example, Ranganathan and Dellaert used the objects in the scene to build a semantic map [13].

Semantic maps based on SLAM can represent the scene in detail, including both space and category information of the objects. However, this kind of map is complicated and time-consuming to build and requires substantial memory. Thus, a lightweight semantic map without spatial details of the scene is adopted in this paper to serve practical robot tasks such as topology positioning, object searching, man-robot interaction, and even man-like navigation. Moreover, our method is simpler than [13] because it only saves categories of objects and relations between them.

III. METHODS

A. PRINCIPLE OF MULTIFRAME FUSION BASED ON CRF

Considering that a single image frame may be insufficient for describing and understanding an entire scene, we propose a novel framework that fuses multiple frames of RGB-D images by weighting the importance of an image to jointly determine the scene category. Furthermore, current research shows that the CRF-based model combining relations between scenes and objects achieves superior recognition performance over purely feature-based methods in complicated indoor scenes [7]. To improve the result, by integrating multiframe information and extending the model in [7], we formulate multiframe fusion based on the CRF model. Specifically, our model fuses scene appearance, object appearance, object geometry, the context of the scene and objects in multiframe images to improve scene classification accuracy. In this framework, the model incrementally updates the scene understanding result as a new RGB-D image arrives.

Formally, we define objects in a given image as $y_i \in \{0, 1, \ldots, C\}$ and the scene of an image in a sequence as $s \in \{1, 2, 3, \ldots, S\}$ where C represents the number of object classes and S represents the number of scene categories. Moreover, 0 represents the class ‘‘unknown’’. In addition, we rank recognized objects in terms of their relevance to certain scenes, deleting unimportant objects that may interfere with scene classification. For example, curtains may appear in various scenes, such as bedrooms, offices, and bathrooms, so they are not discriminative for scene classification and are regarded as unimportant objects. We use ‘importance’ for an object to denote its relevance to scene classification.

As illustrated in Fig. 1, we utilize the appearance and geometric properties of objects, appearance features of the scene and co-occurrence relationships between them. Whereas there are formal similarities to the CRF model in [7], our model has different definitions for the potential items, as depicted in (1):

$$p(y_i, s) = (1/z)\exp(\omega_s \psi_s(s) + \sum\nolimits_t \omega_t \sum\nolimits_i \psi_t(y_i) + \sum\nolimits_m \omega_m \sum\nolimits_i \varphi_m(s, y_i)) \tag{1}$$

There are three kinds of potentials in this model. $\psi_s(s)$ is a unary potential of scene s. $\psi_s(s)$ is obtained by computing the weighted average value with scene classification results of each frame in the image sequence. The weight of each frame is determined by the importance of the objects it contains. $\psi_t(y_i)$ is a unary potential of object y_i, where $i = 1, 2, 3, 4, 5, \ldots$ represents the different objects in an image sequence and $t = 1, 2, 3, \ldots$ represents the different object unary potentials. $\{\varphi_m(s, y_i), m = 1, 2, 3, \ldots\}$ is a set of binary potentials that capture the relationship between scene s and object y_i.With these three potentials, we can extract the most discriminative characteristic information in an image sequence of a scene.

B. POTENTIALS OF MULTIFRAME FUSION BASED CRF

As shown in (1), the potential configuration we adopt to depict and use CRF is similar to [7], but we redefine the unary scene potential to take advantage of the whole information from multiframe images. Moreover, we manage to cut object potentials by removing duplicate objects generated during multiframe integration and unimportant objects somewhat irrelevant to the scene category. This allows our model to outperform the work in [7] on datasets with multiview images for a single scene.

1) UNARY POTENTIALS OF OBJECT

In our model, we define the segmentation potential and geometry potential to describe the properties of objects.

a: SEGMENTATION POTENTIAL

In [7], researchers first generate bounding boxes of objects in the image and then use six types of RGB-D kernel descriptors to train a classifier to obtain the segmentation potential. To obtain a more accurate segmentation potential, we use the Fast R-CNN proposed by Girshick [2] to detect the objects in the image and use the result of Fast R-CNN as our segmentation potential. The segmentation potential is defined in (2).

$$\psi_{seg}(y_i = v) = \Pr_{\text{Fast R-CNN}}(y_i = v) \tag{2}$$

where y_i is the i-th object in the image detected by Fast R-CNN, v is an object category and $\Pr_{\text{Fast R-CNN}}(y_i = v)$ is the probability that object y_i belongs to object category v.

b: GEOMETRY POTENTIAL

We first map the object bounding boxes detected by Fast R-CNN to the 3D coordinates and obtain their minimum circumscribed cuboids. Then, we capture ten properties of the cuboids, namely, *height*, *width*, *length*, *horizontal aspect ratio*, *vertical aspect ratio*, *area*, *volume*, *parallel-to-wall*, *close-to-wall*, and *close-to-ground*, to train a support vector machine (SVM) as described in [7]. The geometry potential is the result of an SVM and can be described as follows:

$$\psi_{geo}(y_i = l) = r_l \tag{3}$$

where r_l is the probability that the cuboid corresponds to the object category l obtained by the SVM.

c: DELETE DUPLICATE OBJECT INSTANCES

One novelty of this paper is to fuse multiframes of images for a better understanding of the environment instead of using a single image. This resembles human behavior of looking around when arriving in a large place. Since one object may be contained in different frames of images, duplicate object instances would affect the result of scene understanding and need to be deleted. The idea is that we first detect whether a new frame has the same object instances as the existing frames and then discard unary potentials of duplicate instances in the new frame. We adopt a method proposed by [21] for detecting duplicated instances.

d: DELETE UNIMPORTANT OBJECTS

As mentioned before, we delete objects that have little relevance to the scene category and retain the objects that tend to significantly impact scene classification results. We find that, statistically, the more likely an object appears in one or a few scenes than it appears in other scenes, the greater relevance to these scenes it has. Based on this rule, we propose a method for calculating the importance of an object to a scene, which can be defined as follows:

$$Imp_i = \sum\nolimits_{j=1}^{S} [(o_{ij}/s_j)/(\sum\nolimits_{j=1}^{S} o_{ij}/s_j) - 1/s_j]^2 \tag{4}$$

where Imp_i is the importance of the object class i, and S is the number of scene categories. o_{ij} represents the times that the i-th object occurs in the j-th scene, and s_j indicates the number of images contained in the j-th scene. The values of o_{ij} and s_j are obtained by counting the NYU v2 dataset. Imp_i is approximately equal to the variance in the probability of the i-th object appearing in each scene.

Based on the result of deleting duplicate objects, we retain the object whose Imp_i value is greater than the threshold α. Through extensive experiments, we found that our model performs stably when α is 0.35.

2) UNARY POTENTIALS OF SCENE

To obtain the unary potential of the scene on the image sequence, we fine-tune Place365-ResNet [20] to make it suitable for the NYU v2 dataset and then use it to obtain the scene category of each image. The output of Place365-ResNet can be seen as the probability of each scene category. Then, we obtain the weighted average of the scene category probabilities over each image in a sequence and use it as the unary potential of the scene for the image sequence. Specifically, the weight of each image is based on the importance of the objects detected in each image, and the unary potential of the scene is defined as follows:

$$\begin{aligned} \psi_s(s = u) &= \phi_s(s = u)/\sum\nolimits_{i=1}^{S} \phi_s(s = i), \\ \phi_s(s = u) &= \sum\nolimits_{p \in P} \mu_p \Pr_{\text{ResNet}}(p = u), \quad \text{and} \\ \mu_p &= \sum Imp_o \end{aligned} \tag{5}$$

In (5), $\psi_s(s = u)$ is the scene unary potential that the image sequence s corresponds to the scene category u, and it is the normalization of $\phi_s(s = u)$. $\phi_s(s = u)$ is the sum of $\Pr_{\text{ResNet}}(p = u)$, and we use it to incrementally update the unary potential of the scene as a new image is added. $\Pr_{\text{ResNet}}(p = u)$ is the probability that Place365-ResNet evaluates that the scene category of image p is u. μ_p is the importance of image p in the image sequence, equaling the sum of the importance value of the objects in image p. In addition, we retain the duplicate objects in the new image when calculating μ_p.

In practice, humans generally judge the scene category by sensing objects in the environment. It is obvious that images containing important objects carry more information about its scene attribution, and this kind of image is

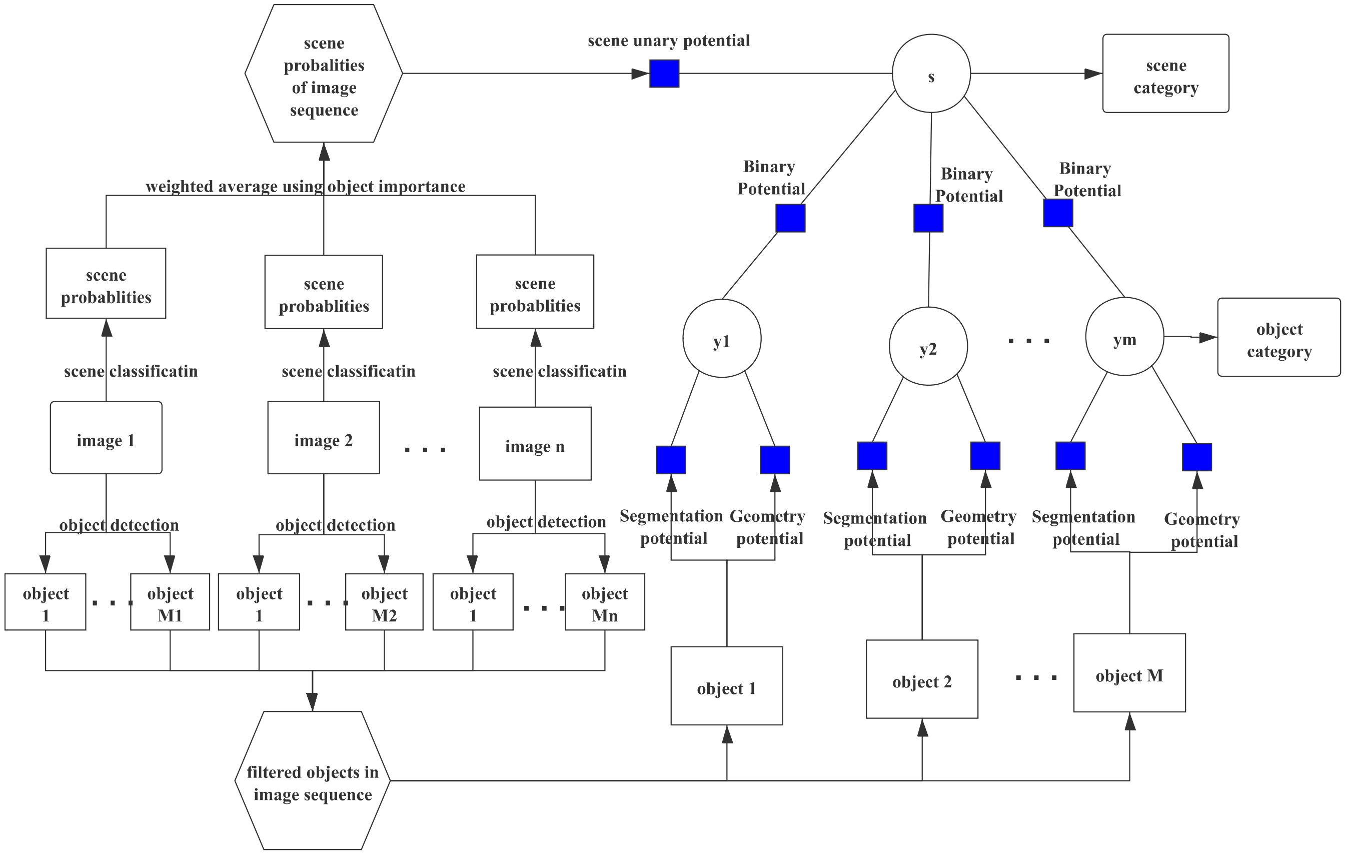


FIGURE 1. **The model of multiframe fusion, wherein $M_i, i = 1, 2, 3, \ldots$ represents the number of objects in image i.**

more important than the image without important objects. To determine whether the neural network would change the probability of the scene class corresponding to the change in the existing object categories, we perform a regression analysis on the probability of scene classes and object categories. For each scene category, we use the probability of Place365-ResNet output as the dependent variable and use the presence or absence of objects as the independent variables. In detail, the regression can be described as follows:

$$\Pr_{\text{ResNet}}(p = u) = \sum\nolimits_{i=1}^{C} \omega_i x_i \tag{6}$$

In (6), $\Pr_{\text{ResNet}}(p = u)$ is the probability that Place365-ResNet classifies the image p as scene u, and x_i is a Boolean variable such that $x_i = 1$ when object i appears in image p, and $x_i = 0$ otherwise. ω_i is the coefficient we want to obtain by regression. $\omega_i > 0$ indicates that the occurrence of object i increases the probability that Place365-ResNet classifies the image p as scene u, and $\omega_i < 0$ means that the occurrence of object i reduces the probability that the image's scene category is u.

According to regression analysis, we find that when there are objects typical to current scenes in the image, such as the sink in a possible ''kitchen'', the probability of the current scene is probably increased. When an image contains objects that also appear in other scenes, such as a sofa in the possible ''kitchen'', the probability of the current scene is likely to decrease. This shows that objects contained in images can influence the result of scene classification and proves that our weight setting method is correct and effective.

3) BINARY POTENTIAL BETWEEN SCENE AND OBJECTS

The binary potential between the scene and the objects can provide effective information for scene understanding. Through the interaction of the scene and objects, the accuracy of both scene recognition and object detection can be improved. According to the binary potential in [7], our binary potential is defined as follows:

$$\varphi_{so}(s = k, y_i = l) \triangleq \Big(\sum\nolimits_{j=1}^{N_{tr}} \sum\nolimits_{i=1}^{m_j} L(s_j = k, y_i^j = l)\Big)/N_{tr} \tag{7}$$

where y_i^j is the i-th object detected in the j-th image of training samples, and m_j is the number of objects in the j-th image. N_{tr} is the number of images in the training set. $L(\bullet)$ is an indication function such that its function value is 1 when the condition in parentheses is true, and 0 otherwise. The binary potential can be used to approximate the co-occurrence probability of the object and the scene.

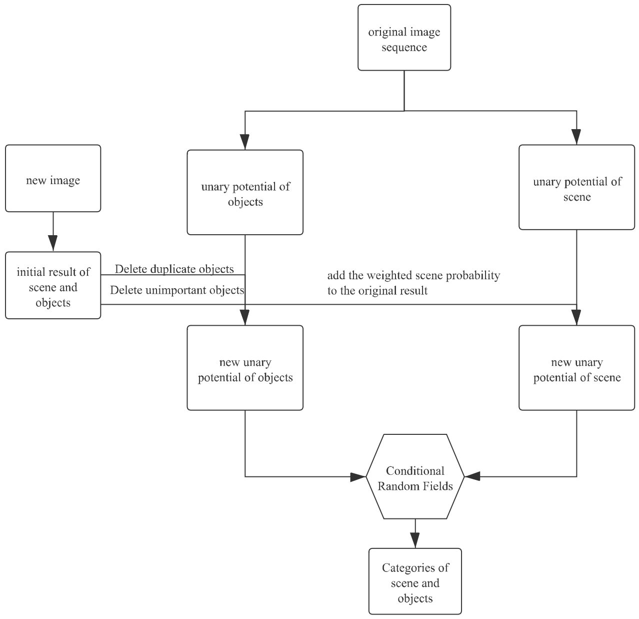

original image sequence
new image
unary potential of objects
unary potential of scene
initial result of scene and objects
Delete duplicate objects
Delete unimportant objects
add the weighted scene probability to the original result
new unary potential of objects
new unary potential of scene
Conditional Random Fields
Categories of scene and objects

FIGURE 2. Incremental update of multiframe fusion CRF.

C. INCREMENTAL UPDATE OF THE MULTIFRAME FUSION BASED CRF MODEL

When a new image frame arrives, the model is incrementally updated, as shown in Fig. 2.

We first use the duplicate object detection method mentioned above to delete duplicate objects appearing in the new image. Then, we select the objects with importance above the threshold and add their unary potentials to the original potential set. Additionally, we calculate the scene unary potential of the new image sequence by (8):

$$\begin{aligned}\psi_s(s=u) &= \phi_s^{new}(s=u)/\sum\nolimits_{i=1}^{S}\phi_s^{new}(s=i)\\ \phi_s^{new}(s=u) &= \phi_s^{old}(s=u)+\mu_{new}\Pr\nolimits_{\text{ResNet}}(p_{new}=u)\\ \mu_{new} &= \sum\nolimits_{o} Imp_o\end{aligned}\tag{8}$$

This specifies that we first calculate the importance of the new image μ_{new} and the probability $\Pr_{\text{ResNet}}(p_{new}=u)$ obtained by Place365-ResNet. Then, we obtain the final scene unary potential $\psi_s(s=u)$ by adding the product of the two to the original intermediate result $\phi_s^{old}(s=u)$ and normalizing the new intermediate result $\phi_s^{new}(s=u)$.

Finally, we input the new potentials into the CRF model to recalculate the categories of the scene and objects.

D. BUILDING SCENE CONFIGURATION MAP

Our scene configuration map for robots has a room-object hierarchy structure. There is a one-to-many relationship between the room and the objects, and one-to-one relationship between two objects. Fig. 3 is a sample scene configuration map of the living room including a sofa, a tea table and a wall-hanging TV.

We complete the construction of the scene configuration map by the following steps:

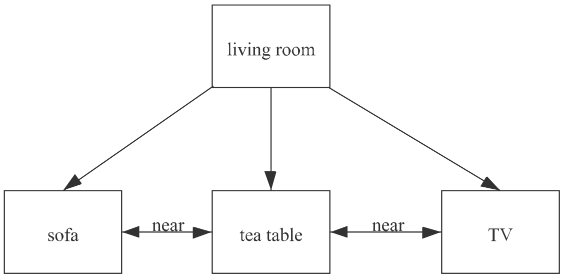


FIGURE 3. Sample of scene configuration map.

1. Object detection. Use the Fast R-CNN to detect the objects occurring in the current image.
2. Relationship detection. Use the object relationship detection method in [7] to extract the relationships between the objects in the current image. The relationships between objects include 'near' and 'above'.
3. Duplicate object detection. Use the method of duplicate object detection described in Section III.B to detect whether there are duplicate objects between the current image and the previous input image.
4. Scene classification. Use the model we proposed in Section III to classify the scene category of the current room.
5. Repeat steps 1 to 4 when a new image is input.
6. Integrate the results of single images and visualize the map. The scene category of the final scene configuration map is the result of the last scene classification, and the objects in the map are the result of merging duplicated objects.

IV. EXPERIMENTS

A. DATASETS

We evaluate the model in the NYU v2 dataset [16] and compare it with the method proposed in [7]. NYU v2 dataset is comprised of video sequences from a variety of indoor scenes [16]. Many images in NYU v2 dataset contain only a part of a scene. Incomplete scene information makes it difficult for existing image recognition methods to obtain excellent results on the NYU v2 dataset. Scene understanding on the NYU v2 dataset is still a challenging issue.

NYU v2 dataset contains 1,449 densely labeled aligned RGB-D images. The original annotation of the NYU v2 dataset assigns all the pixels to 894 object categories. Because there are too many object categories and some object categories have few object instances, it is difficult to manipulate the 894 categories in practice. To address the problem and to make it easy to compare with the method proposed in paper [7], we use 21 object classes and 13 scene classes, which are also used in [7]. The 13 scene classes are *kitchen*, *office*, *bathroom*, *living room*, *bedroom*, *bookstore*, *classroom*, *home office*, *playroom*, *reception room*, *study*, *dining room*, and *others*. The 21 object classes are *mantel*, *counter*, *toilet*, *sink*, *bathtub*, *bed*, *headboard*, *table*, *shelf*, *cabinet*, *sofa*, *chair*,

TABLE 1. The number of image sequences with different *m*.

m Dataset	1	2	3	4	5	6	7	8	14	27	33
NYUv2	248	118	130	69	30	9	2	1	1	1	1
test set of NYUv2	98	58	62	34	16	5	0	1	0	0	0

chest, *refrigerator*, *oven*, *microwave*, *blinds*, *curtain*, *board*, *monitor*, and *printer*.

We divide the 1,449 images of the NYU v2 dataset into different image sequences by manually putting the images obtained in the same room into a certain sequence. We partition the dataset into a training set containing 755 images and a test set containing 694 images using the same split as [7]. Then, we manually divide the test set into image sequences. Let m be the number of images in an image sequence. We present the number of image sequences with different m in Table 1.

B. VISUALIZATION OF EXPERIMENTAL RESULTS

The incremental result on the image sequences of our method is shown in Fig. 4.

If the initial input frame does not contain overall scene information, then the model makes an incorrect judgment on the scene category because the scene information is not sufficient. However, with the addition of new images, the information about the scene is gradually complemented, and the incorrect classification result based on the initial image is corrected.

Fig. 5 shows the construction process of the scene configuration map. In Fig. 5, (a) (b) (c) (d) are images in the image sequence, whereas (e) (f) (g) and (h) are configuration maps corresponding to each single frame image. (i) and (j) list the incremental results of scene configuration map building, where the map does not change when the third and fourth images are added without detecting new objects.

Qualitative experiments are given above and show that the proposed model yields meaningful results. Hereafter, we verify our model on an original test set of NYU v2 and on a combined test set of image sequences in Sections IV.C and IV.D. Each sequence in the combined test set contains several images misclassified by the model in [7]. We want to determine if the fused model can correct the scene classification fault by the original model. In Section IV.E, we perform a comparison against another multiframe fusion method. In Section IV.F, we conduct four-fold cross validation to select optimal parameters for the model. In Section IV.G, we compare our method with state-of-the-art scene classification methods to prove the effectiveness of our method. In Section IV.H we show the results of the ablation experiments performed on this method, and in Section IV.I we report the time efficiency of our method.

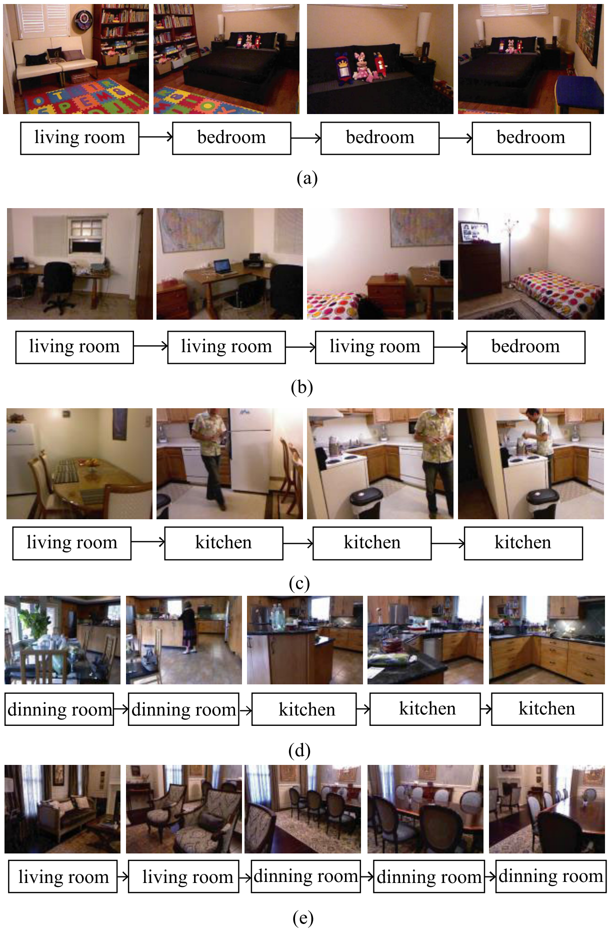


FIGURE 4. The incremental results of scene classification. (a) (b) (c) (d) (e) shows the scene classification results of five different image sequences. The images in image sequences are input into our model frame by frame, and the latest result of scene classification is displayed below the current image frame.

C. EXPERIMENTS ON THE ORIGINAL TEST SET

We use different object importance thresholds to test our model on the test set we split and compare the accuracy of our model with the model proposed in [7]. We take a threshold of every 0.05 in the range of 0 to 1 and retain the objects whose importance is greater than the threshold. The test results are shown in the second and fourth columns of Table 2. It can be seen from the table that the accuracy of the scene recognition of our method is higher than that of the method based on a single image proposed in [7]. The accuracy can be increased by 14% at most.

D. VERIFICATION OF THE ABILITY TO CORRECT THE MISCLASSIFICATION ON SINGLE IMAGES

We extract the misclassified images in [7] and use the image sequences containing these images as the test set. The images

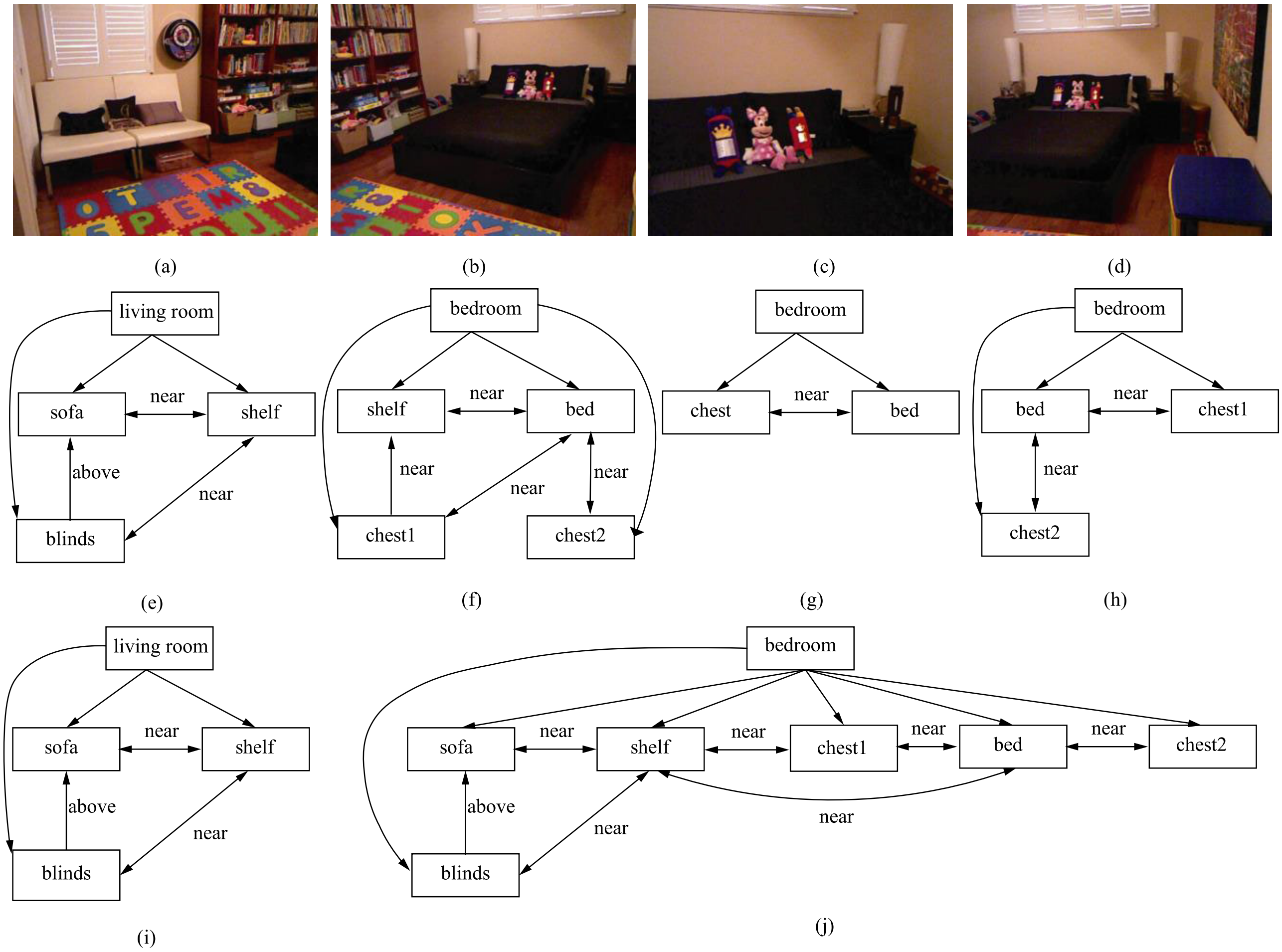


FIGURE 5. Incremental process of building a scene configuration map. (a) (b) (c) (d) are images in the image sequence, whereas (e) (f) (g) and (h) are configuration maps corresponding to each single frame image. (i) and (j) list the incremental results of scene configuration map building. (i) is the final configuration map corresponding to the first image, and (j) is the final configuration map corresponding to the second to fourth images.

in the remaining sequences are used as the training set. We use different object importance thresholds to test our model. The objects whose importance is greater than the threshold are retained. The experimental results are shown in the third and sixth columns of Table 2. It can be seen from the table that the accuracy of the scene recognition with our method is higher than that of the method based on a single image. This shows that our model can correct the misclassification on a single image by fusing information from multiview images within the same scene.

E. COMPARISON WITH OTHER MULTIFRAME FUSION METHODS

The above two experiments show that the multiframe information fusion method can improve the experimental accuracy of scene recognition compared with the single image-based method. Furthermore, we compare our model with the multiframe voting-based method, which may be considered the simplest multiframe fusion method. ‘‘Multiframe voting’’ refers to voting using the scene classification result for each image, and the scene category with the most votes is the final scene classification result for the image sequence. If there are two or more scene categories with the same number of votes, the number of objects contained in each image would be counted. The scene category including the most objects is taken as the final scene category of the image sequence. The last row of Table 2 shows the classification accuracy of the multiframe voting method. It can be seen from the table that the accuracy of our model is higher than that of the multiframe voting method. This proves that our model makes more efficient use of information in multiframe images.

F. DETERMINING THE THRESHOLD BY CROSS-VALIDATION

To determine the threshold of retained objects that the model ultimately uses during fusing multiple frames, we divide the NYUv2 dataset into five equal parts into units of the image sequence. Then, we perform cross-validation on the first four parts and select the threshold with the highest accuracy as the candidate. Finally, we use the first four parts as the training set and the fifth part as the verification set to verify the validity of the candidate threshold.

TABLE 2. Comparison of scene classification accuracy on the NYU v2 dataset between other methods and our method under different thresholds.

Thresholds	Original test set	Misclassification-based test set	Thresholds	Original test set	Misclassification-based test set
0.95	66.67	52.02	0.9	66.67	52.02
0.85	70.03	57.72	0.8	70.03	57.72
0.75	70.03	57.72	0.7	72.02	59.86
0.65	72.02	59.86	0.6	72.32	60.57
0.55	72.32	61.52	0.5	71.87	61.52
0.45	71.87	61.52	0.4	72.32	61.52
0.35	72.02	60.57	0.3	70.34	57.01
0.25	69.27	56.06	0.2	69.11	55.82
0.15	70.49	57.72	0.1	70.49	57.72
0.05	70.49	57.72	0	70.49	57.72
Single image method in [7]	58.72	39.43	Single image method in [7]	58.72	39.43
Multiframe voting method	67.58	41.09	Multiframe voting method	67.58	41.09

TABLE 3. Cross-validation on NYUv2.

Threshold	Accuracy of first part	Accuracy of second part	Accuracy of third part	Accuracy of fourth part	Accuracy of verification
0.95	83.33	84.23	82.81	81.38	79.64
0.9	83.99	84.23	86.72	81.68	82.55
0.85	83.99	84.23	86.72	81.68	82.55
0.8	83.99	84.23	86.72	81.68	82.55
0.75	83.99	84.23	86.72	81.68	82.55
0.7	83.99	84.23	86.72	81.68	82.55
0.65	83.99	86.02	87.50	81.68	82.55
0.6	83.99	86.02	87.50	81.68	82.55
0.55	83.99	86.02	87.50	81.68	82.55
0.5	83.99	86.02	87.50	81.68	82.55
0.45	83.99	86.02	87.50	81.68	82.55
0.4	83.99	86.02	89.06	81.68	85.45
0.35	86.27	86.02	**89.84**	**81.68**	**85.45**
0.3	**90.20**	86.02	89.84	76.88	85.45
0.25	84.64	86.38	86.72	76.88	81.82
0.2	86.60	87.46	88.67	75.98	79.64
0.15	86.60	**87.46**	88.67	75.98	79.64
0.1	86.60	87.46	88.67	75.98	79.64
0.05	86.60	87.46	88.67	75.98	79.64
0	86.60	87.46	88.67	75.98	79.64
single image Method in [7]	83.66	84.23	82.42	81.38	78.91

The results of the cross-validation are shown in the first five columns of Table 3. The last column shows the accuracy of the verification. It can be seen that when the threshold is 0.35, the performance of our model is relatively stable. The performance of the multiframe fusion-based CRF model is better than that of the single-frame-based CRF method proposed in [7].

G. COMPARISON WITH STATE-OF-THE-ART SCENE CLASSIFICATION METHODS

Table 4 shows the performance comparison with state-of-the-art scene classification methods on the NYU Depth V2 test dataset. We select the stable experimental threshold obtained in the four-fold cross-validation and the threshold with the best experimental results in the original test set to compare with other methods. Since other methods use 10 scene categories, and our method uses 3 more scene categories, we map the extra 3 scene categories to 'others' and re-count the accuracy of scene classification. Our CRF with the stable threshold outperforms state-of-the-art methods, and our CRF with the best performance threshold is 0.2% more accurate than the former.

TABLE 4. Accuracy comparison with state-of-the-art methods on the NYU Depth V2 test set.

Method	Accuracy (%)
(Song et al. 2017 [28])	65.8%
(Song et al. 2017 [29])	66.7%
(Li et al. 2019 [30])	67.7%
(Song et al. 2019 [31])	67.5%
(Ali Ayub and Alan Wagner 2019 [32])	69.7%
Multiframe Fusion CRF ($\alpha = 0.35$, stable)	72.39%
Multiframe Fusion CRF ($\alpha = 0.40$, best)	**72.60%**

H. ABLATION STUDY

We performed ablation experiments on our method on the NYU Depth V2 dataset, as shown in Table 5. The baseline multiframe fusion model sends the image in the image sequence to Place365-ResNet for scene classification. Then,

TABLE 5. Ablation study for multiframe fusion CRF on the NYU Depth V2 Dataset.

Method	Accuracy (%)
baseline multiframe fusion	66.2%
MD (multiframe fusion + deleting unimportant objects)	71.55%
MC (multiframe fusion +CRF)	67.58%
MDC (multiframe fusion + deleting unimportant objects + CRF)	72.32%

it sums and normalizes the scene probabilities of the images as the scene classification result of the image sequence. It achieves 66.2% accuracy, which is already a high baseline compared with the state-of-the-art methods. The MD model designs weights for each frame of the image on the basis of the baseline. It first deletes unimportant objects in each frame and then uses the remaining objects' importance to calculate the image weight. The classification accuracy significantly improved after adding image weights to the baseline. The MC model takes the scene classification result of the baseline as the scene unary potential of the CRF and uses the relationship between objects and scenes to adjust the existing scene classification result. It can achieve better performance compared with the baseline model. The MDC model takes the classification result of MD as the scene unary potential of CRF and uses CRF to adjust the scene classification result. The accuracy of scene classification improves 1% after adjusting the MD model by CRF.

I. RUNTIME PERFORMANCE

We test the time efficiency of our algorithm on the NYU Depth V2 dataset. All tests are performed on an Intel Core i7-8700 CPU and an Nvidia GTX 1060 GPU. We obtain the scene unary potential and object unary potential for every frame. This process takes an average of 0.05 s/frame. Then, we use these potentials to form the CRF and use it to modify the results. This phase requires 0.01 s/frame on average.

The time for scene classification using Place365-ResNet is approximately 0.02 s/frame. Although the method in this paper is slightly slower than the neural network method, it still has better real-time performance. Furthermore, our method has a higher scene classification accuracy than Place365-ResNet. The time required for CRF iteration in [7] is approximately 0.01 s/frame, which is almost the same as the CRF in this paper. This shows that although the CRF in [7] is extended from a single frame to multiple frames in this paper, it does not increase the required processing time.

V. CONCLUSION

In this paper, we proposed a multiframe RGB-D image fusion model based on CRF to recognize the scene category in an overall view. The configuration map of a scene is incrementally built as an output of the scene understanding to be used for possible robot tasks. We divided the NYU v2 dataset manually into image sequences corresponding to different places and compared our model with the typical single image-based method. The experimental results show that our model has a better performance in scene understanding against state-of-the-art baselines.

Future work can be summarized as two-fold. First, we plan to optimize the methods for deleting duplicate objects. We can use the context information in an image, for example, the relationship between objects, to improve the accuracy for deleting duplicate object instances with smarter methods. Moreover, we plan to perform 3D reconstruction of image sequences and use the 3D reconstruction result to classify the surrounding scene.

ACKNOWLEDGMENT

(Haotian Chen and Longfei Su contributed equally to this work.)

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, San Diego, CA, USA, 2005, pp. 886–893.

[2] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 1440–1448.

[3] R. Li, D. Gu, Q. Liu, Z. Long, and H. Hu, "Semantic scene mapping with spatio-temporal deep neural network for robotic applications," *Cognit. Comput.*, vol. 10, no. 2, pp. 260–271, Apr. 2018, doi: 10.1007/s12559-017-9526-9.

[4] I. Kostavelis and A. Gasteratos, "Learning spatially semantic representations for cognitive robot navigation," *Robot. Auto. Syst.*, vol. 61, no. 12, pp. 1460–1475, Dec. 2013.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, NV, USA, 2012, pp. 1106–1114.

[6] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014, doi: 10.1007/s11263-013-0660-x.

[7] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1417–1424.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, 1999, pp. 1150–1157.

[9] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3D for autonomous agents: A survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019, doi: 10.1109/ACCESS.2018.2886133.

[10] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," 2011, *arXiv:1101.3354*. [Online]. Available: http://arxiv.org/abs/1101.3354

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001, doi: 10.1023/A:1011139631724.

[12] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *Proc. IEEE Int. Conf. Robot. Autom.*, Paul, MN, USA, May 2012, pp. 3515–3522.

[13] A. Ranganathan and F. Dellaert, "Semantic modeling of places using objects," in *Robotics: Science and Systems III*. Atlanta, GA, USA, 2007, pp. 1–8.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[16] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, Florence, Italy, 2012, pp. 746–760.

[17] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013, doi: 10.1007/s11263-013-0620-5.

[18] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Dec. 2001, pp. 511–518, doi: 10.1109/CVPR.2001.990517.

[19] H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J.-M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robot. Auto. Syst.*, vol. 56, no. 6, pp. 493–502, Jun. 2008, doi: 10.1016/j.robot.2008.03.007.

[20] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018, doi: 10.1109/tpami.2017.2723009.

[21] X. Li, "Object detection and scene understanding based on fusing multiview RGB-D frames," M.S. thesis, Dept. College Softw., Nankai Univ., Tianjin, China, 2019.

[22] M. Zhai, X. Xiang, R. Zhang, N. Lv, and A. El Saddik, "Optical flow estimation using dual self-attention pyramid networks," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2019, doi: 10.1109/TCSVT.2019.2943140.

[23] M. Zhai, X. Xiang, R. Zhang, N. Lv, and A. E. Saddik, "Optical flow estimation using channel attention mechanism and dilated convolutional neural networks," *Neurocomputing*, vol. 368, pp. 124–132, Nov. 2019, doi: 10.1016/j.neucom.2019.08.040.

[24] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "LEGO: Learning edge with geometry all at once by watching videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 225–234.

[25] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5667–5675.

[26] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 4628–4635.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] X. Song, L. Herranz, and S. Jiang, "Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs," in *Proc. AAAI Conf. Artif. Intell.*, pp. 4271–4277, 2017.

[29] X. Song, S. Jiang, and L. Herranz, "Combining models from multiple sources for RGB-D scene recognition," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4523–4529.

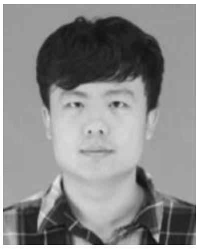
[30] Y. Li, Z. Zhang, Y. Cheng, L. Wang, and T. Tan, "MAPNet: Multimodal attentive pooling network for RGB-D indoor scene classification," *Pattern Recognit.*, vol. 90, pp. 436–449, Jun. 2019, doi: 10.1016/j.patcog.2019.02.005.

[31] X. Song, S. Jiang, L. Herranz, and C. Chen, "Learning effective RGB-D representations for scene recognition," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 980–993, Feb. 2019.

[32] A. Ayub and A. Wagner, "CBCL: Brain-inspired model for RGB-D indoor scene classification," 2019, *arXiv:1911.00155*. [Online]. Available: http://arxiv.org/abs/1911.00155

HAOTIAN CHEN was born in Tianjin, China, in 1988. He received the B.S. and M.S. degrees in computer software from Nankai University, Tianjin, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the Intelligent Information Processing Lab.

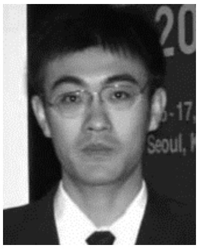
His research interests include mobile intelligent robot systems, scene understanding, and object recognition.



LONGFEI SU born in Dezhou, China, in 1996. He is currently pursuing the M.S. degree in software engineering with Nankai University, Tianjin, China. His current research interests include computer vision and scene understanding.

BIAO ZHANG born in Shanxi, China, in 1996. He received the B.S. degree in software engineering from Nankai University, Tianjin, China, in 2017, where he is currently pursuing the M.S. degree in software engineering. His main research interests include computer vision and scene understanding.

FENGCHI SUN (Member, IEEE) born in Shandong, China, in 1973. He received the Ph.D. degree in control theory and control engineering from Nankai University. He is an Associate Professor with the College of Software, Nankai University. His main research interests include artificial intelligence, robotics, and embedded systems.



JING YUAN (Member, IEEE) received the B.S. and Ph.D. degrees in control theory and control engineering from Nankai University. He is currently a Professor with the College of Artificial Intelligence, Nankai University. His main research interests include robot control, mobile robot navigation, SLAM, multirobot systems, and target tracking.

JIE LIU is currently a Professor with the College of Artificial Intelligence, Nankai University. His main research interests include machine learning, data mining, and natural language processing. He is also a member of the CCF and ACM.