2020 IEEE International Conference on
Autonomous Robot Systems and Competitions (ICARSC)
April 15-16, Ponta Delgada, Portugal

Deep-Learning based Global and Semantic Feature Fusion for Indoor Scene Classification

Ricardo Pereira[1], Nuno Gonçalves[1], Luís Garrote[1], Tiago Barros[1], Ana Lopes[1,2], Urbano J. Nunes[1]

***Abstract*—This paper focuses on the task of RGB indoor scene classification. A single scene may contain various configurations and points of view, but there are a small number of objects that can characterize the scene. In this paper we propose a deep-learning based Global and Semantic Feature Fusion Approach (GSF^{2}App) with two branches. In the first branch (top branch), a CNN model is trained to extract global features from RGB images, taking leverage from the ImageNet pre-trained model to initialize our CNN's weights. In the second branch (bottom branch), we develop a semantic feature vector that represents the objects in the image, which are detected and classified through the COCO dataset pre-trained YOLOv3 model. Then, both global and semantic features are combined in an intermediate feature fusion stage. The proposed approach was evaluated on the SUN RGB-D Dataset and NYU Depth Dataset V2 achieving state-of-the-art results on both datasets.**

***Index Terms*—Indoor Scene Classification, Deep Learning, RGB, Semantic Features**

I. Introduction

Indoor scene classification remains a challenging task in the machine learning and robotics communities. Due to the various configurations that a single scene may have, it becomes difficult to obtain a robust model for indoor scene classification tasks. On the other hand, a successful scene categorization could be very important for mobile robotics tasks [1][2], e.g. building maps, improve localization, and navigation.

Deep learning feature extraction layers have been achieving good performances in the object classification field. However, the results achieved in scene classification tasks need to be improved [3][4]. This weak performance may happen due to the lack of labeled datasets that comprise multiple scenes in different conditions (and points of view).

Several methods [5]–[7] have been proposed for classifying RGB-D scene images using local and global features. These works are based on CNN architectures to extract features from two distinct modalities, RGB and Depth, from which the relationships between local and global features were exploited. Their reported results showed that local features provide important information about the scene, making them very useful to improve the achieved results in scene classification. To recognize a scene, people focus on the objects present in the scene and also on correlations between objects [3]. As illustrated in Fig. 1, the *bedroom* class has various configurations and viewpoints. However, all images contain the same types of objects that allow to classify the scene as bedroom (e.g. bed object). In order to further exploit the global features and the correlations between objects present in the scene, we developed a fusion approach to learn how objects can be related to the indoor scene.

(a) Bedroom sample 1. (b) Bedroom sample 2. (c) Bedroom sample 3. (d) Bedroom sample 4.

Fig. 1. Different configurations from the bedroom class. Sample images are from the NYU Depth Dataset V2 [8].

In this work, a deep learning based Global and Semantic Feature Fusion Approach (GSF^{2}App) with two branches, shown in Fig. 2, is proposed for RGB indoor scene classification. Semantic features represent the objects recognized in the image along with the number of times that the same object appears in the image. The proposed architecture, consists of two branches operating on global and semantic features respectively, which are combined in an intermediate feature fusion model. The top branch uses the VGG16 [9] Convolutional Neural Network (CNN) to automatically learn and extract global features from the RGB modality. The bottom branch consists of recognizing objects present in the image and encode them into a Semantic Feature Vector (SFV) that feds two fully connected layers. Then, both branches output features undergo a fusion stage for a class prediction. To recognize objects in the RGB image, the COCO dataset pre-trained YOLOv3 model [10] is used. The proposed approach was evaluated on the SUN RGB-D Dataset [11] and the NYU Depth Dataset V2 [8].

[1]Authors are with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Portugal. Emails: {ricardo.pereira, nunogon, garrote, tiagobarros, anacris, urbano}@isr.uc.pt.

[2]Author is also with the Polytechnic Institute of Tomar, Portugal.

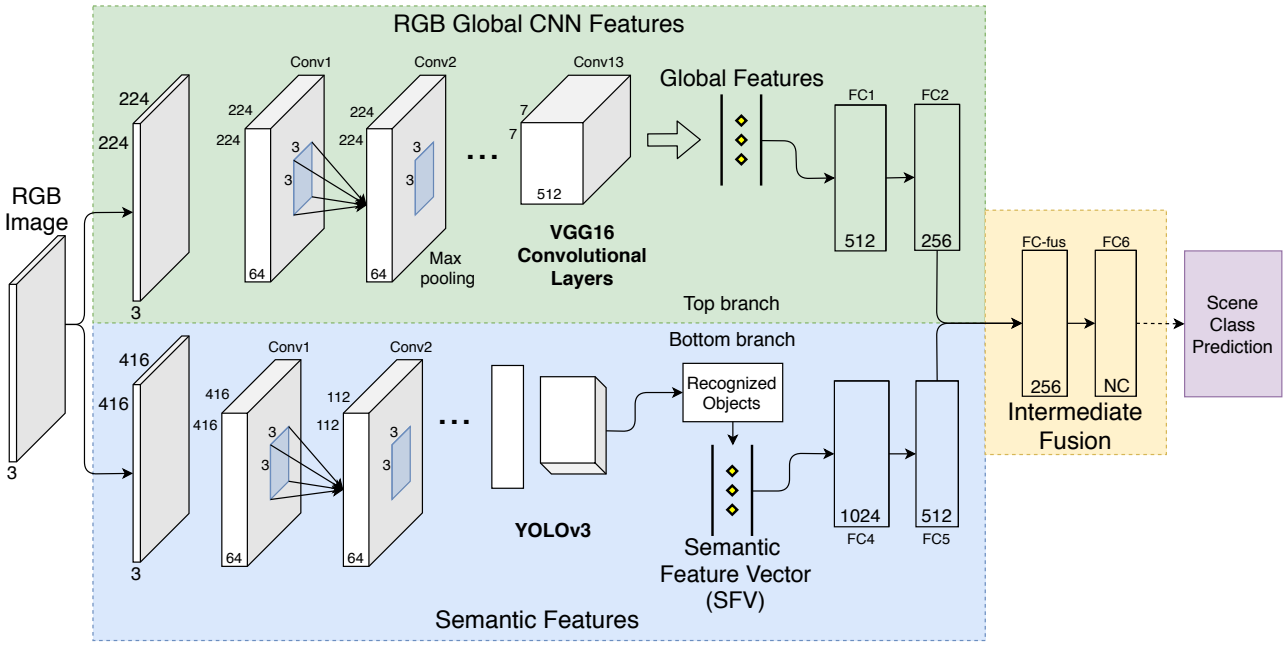


Fig. 2. An overview of the proposed GSF^{2}App (in FC6, NC represents the output number of classes). The top branch (green) uses the VGG16 network to extract global features. The bottom branch (blue) uses the YOLOv3 model to recognize the objects present in the scene, which are encoded into a Semantic Feature Vector. Both branches converge in two fully connected layers (yellow).

The main contributions of this work are:

- The development of an indoor scene recognition model based on the fusion of global and semantic features.
- Based on the achieved results, there is empirical evidence that semantics is able to improve indoor scene recognition models. The proposed GSF^{2}App achieves state-of-the-art results on both SUN RGB-D Dataset [11] and NYU Depth Dataset V2 [8].

II. Related Work

Scene Classification

Indoor scene classification methods have been researched over the past years. Gupta *et al.* [12] proposed an object-based pipeline to classify the indoor scene. They applied a semantic segmentation method to classify the objects present in a scene, which were encoded into a spatial pyramid (SPM), that contains the average presence of each semantic class, and is used as input feature of the SVM classifier. George *et al.* [13] argues that some objects are only available in one particular scene, while other objects can be identified in different indoor scenes. With this, they proposed a semantic scene descriptor based on the patterns of occurrence of objects in scenes to build small semantic clusters.

In the last years, due to the success of feature extraction layers, deep learning methods have been outperforming the state-of-the-art results in classification tasks. Inspired by the success of CNNs in image classification and object detection tasks, a large-scale scene recognition dataset, Places-CNN, was produced by Zhou et al. [4]. Places-CNN dataset has 365 scenes categories with at most 5000 images per category. As expected, [4] concludes that object-centric and scene-centric neural networks achieve different results. Therefore, they made available some pre-trained CNN models, which are ideal to use as starting point in small datasets.

Despite the CNNs success, full-image global CNN features are not enough to represent an indoor scene [6]. In an attempt to get more information about the scene, encoding methods such as Fisher Vector [5][14] or Vector of Locally Aggregated Descriptors [15] to encode local CNN features were proposed. Wang *et al.* [5] proposed a local and global CNN feature fusion method. They used an object proposal extractor method to generate Regions-of-Interest (RoI) from each RGB-D image, representing each RoI by local CNN features. For each modality, the Fisher Vector method was used to encode RoIs which were, in a final step, combined with full-image CNN features.

With the emergence of low-cost RGB-D sensors (e.g. Microsoft's Kinect and Intel's RealSense) that are able to synchronously record both RGB and depth images, a new perspective for feature extraction from images was introduced. RGB data is only able to provide information about appearance and texture while depth data contain distances between the camera's position and its environment. With depth data, it is possible to extract additional information such as objects distances and their shape [16]. Methods operating on RGB-D data were proposed for object recognition [16][17], and scene recognition [6][18]. Gutpa *et al.* [19] proposed to encode depth data into three channels (Horizontal disparity, Height above ground, Angle with gravity) to extract depth features more efficiently. This kind of encoding methods (depth data to three channels) have become very popular and useful in

the machine learning community [16][19]. On the other hand, encoding depth data into three channels allows to apply pre-trained CNN models on the depth-images, which requires three channels as input. In the object classification field, Color Map and Surface Normals encoding methods were also proposed to extract depth features [17].

The majority of the aforementioned works simply concatenate local and global RGB-D features, expecting that the model can be able to learn the necessary correlations without any extra information. Recent methods [3][5][6][20] were able to improve the achieved results by proposing architectures that were not limited to exploiting local and global features, but were also able to incorporate additional information (features). Song *et al.* [20] proposed a multi-modal multi-feature scene recognition pipeline that is able to combine global and local RGB-D features with the spatial layout. Li *et al.* [3] aimed to learn correlative and distinctive features of each RGB-D modality. Xiong *et al.* [6] proposed an end-to-end multi-model feature learning framework, which is able to select important local region features from the high-semantic level CNN feature maps, concatenating, in a final step, both local and global features.

Object Recognition

Deep learning techniques have achieved cutting edge results in object recognition field. Some methods have been proposed, however, YOLO [10][21], Single Shot Detector (SSD) [22], and Faster R-CNN [23] remain the most popular end-to-end frameworks able to perform object recognition. Among these, Faster R-CNN and SSD achieved better results for the average precision metric on the VOC dataset, while YOLO is the fastest architecture. Joseph Redmon [21] argues that YOLO is less likely to predict false positives on background than any other state-of-the-art method.

III. Methodology

An overview of the proposed GSF^{2}App is presented in Fig. 2. Our approach consists of two branches processing RGB data (top branch) and semantic features (bottom branch) respectively, which are combined in an intermediate feature fusion model. This approach is also trained in two steps, the first learns global features from RGB images, while the second step, learns how to combine semantic features with global features.

A. RGB Global-CNN Features

The top branch of the proposed approach uses the VGG16 [9] CNN to process and extract full-image features. It consists of thirteen convolutional layers (with max-pooling after the second, fourth, seventh, tenth, and thirteenth convolutional layers) followed by two fully-connected layers and a softmax classification layer. Unless the final layer, all the others used Rectified Linear Units (ReLU) as activation function. The VGG16 weights' model is initialized by copying the parameters of the ImageNet pre-trained model. Then, we fine-tuned the parameters of the VGG16 model for classification of the target data. Although the model has been pre-trained for object classification instead of scene classification, preliminary results showed that this initialization has significant influence on the model's convergence.

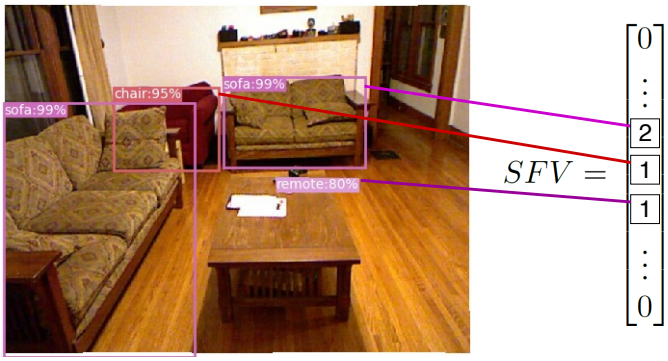


Fig. 3. Example of how YOLOv3's output object class predictions are encoded into a Semantic Feature Vector. Sample indoor scene image taken from the NYU Depth Dataset V2 [8].

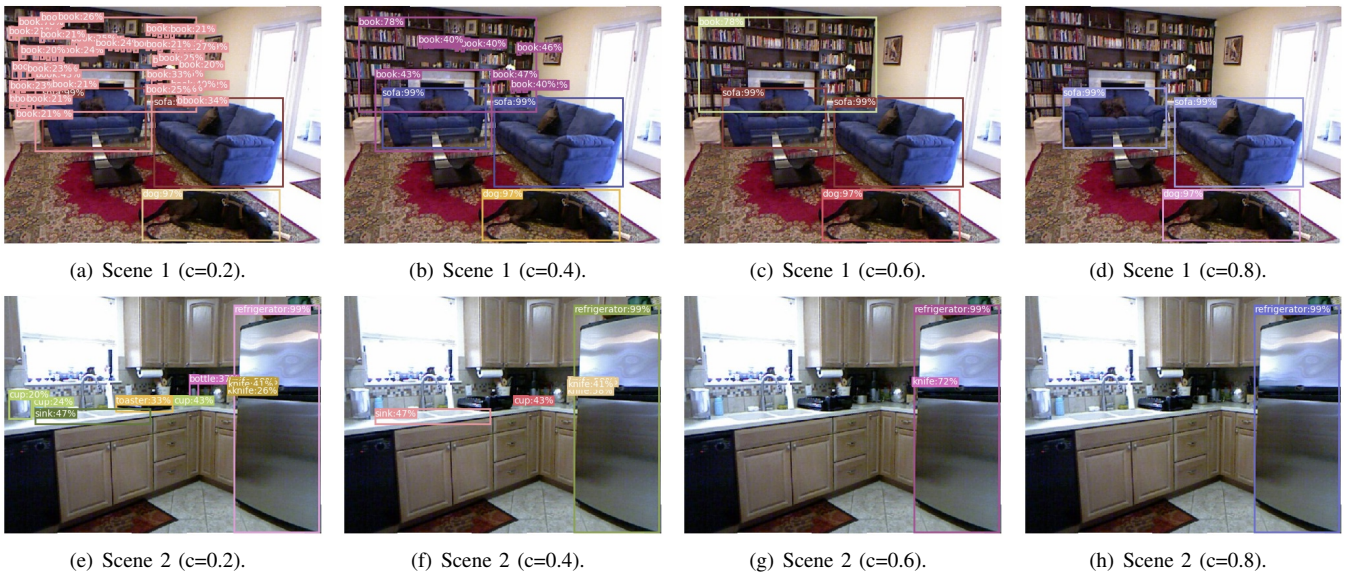
In order to use pre-trained models, the size of input images must be the same as that of the pre-trained model. So, all RGB images were resized to 224×224.

B. Semantic Features

In this work, semantic features represent the objects recognized in the RGB scene image along with the number of times that the same object appears in the same RGB scene image. More specifically, we use the publicly COCO dataset pre-trained YOLOv3 [10] model to recognize the objects present in RGB scene images. For each image, the YOLOv3's output object class predictions are encoded into a Semantic Feature Vector (SFV) (as shown in Fig. 3) which is processed by two fully connected layers in the bottom branch of the proposed approach. The particular COCO dataset pre-trained YOLOv3 model is able to recognize eighty different objects (e.g. sofa, oven, bed, and refrigerator), so the proposed SFV has 80 rows per column (image). It should be noted that, for any scene image, each object is always encoded in the same row as follows:

$$SFV = \begin{bmatrix} O_{1,1}(Person) \\ O_{2,1}(Bicycle) \\ \vdots \\ O_{80,1}(Toothbrush) \end{bmatrix}$$

To further exploit semantic features, for the same scene image, nine SFVs are created changing the YOLOv3's confidence threshold ($c = 0.1, c = 0.2, ..., c = 0.9$). Figure 4 presents the objects recognized using different YOLOv3's confidence thresholds in two distinguished indoor scenes. As expected, as the threshold decreases, more objects are recognized, leading to an increase in the number of semantic features that can characterize the indoor scene, however, the number of false positives also increased.



(a) Scene 1 (c=0.2). (b) Scene 1 (c=0.4). (c) Scene 1 (c=0.6). (d) Scene 1 (c=0.8).

(e) Scene 2 (c=0.2). (f) Scene 2 (c=0.4). (g) Scene 2 (c=0.6). (h) Scene 2 (c=0.8).

Fig. 4. Object recognition using different YOLOv3's confidence thresholds. Sample indoor scene images taken from the NYU Depth Dataset V2 [8].

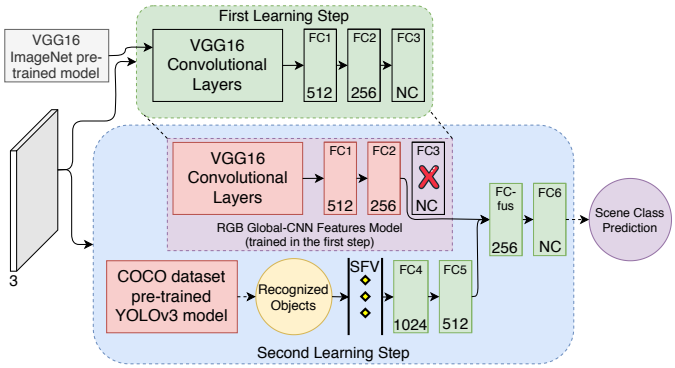


Fig. 5. Two-step learning of the GSF^{2}App.

C. Global and Semantic Feature Fusion

To explore the correlation between the RGB image and semantic features while simultaneously retain the distinctiveness in the RGB modality, we propose a feature fusion approach with two-step learning, as shown in Fig. 5. In the first step, only the VGG16 CNN is trained, originating a RGB-only base model. After that, the last layer of the VGG16 CNN is discarded, since its output is a scene class prediction and not a feature map. Then, in the second learning step, the model learns how to combine global and semantic features. For this, global features are extracted through the trained VGG16 model (in Fig. 5: the cascade composed by the VGG16 convolutional layers, FC1, and FC2), without any fine-tuning of its weights, while semantic features are processed in two fully connected layers (FC4 and FC5). Then, both global and semantic feature maps, FC2 and FC5 respectively, are concatenated feeding a fully connected layer, FC-fus (with 256 feature map output dimension). The last layer, FC6, outputs the scene class prediction. In summary, in the second learning step, only FC4, FC5, FC-fus, and FC6 weights are trained. After the conclusion of the second learning step, a single model for indoor scene classification tasks is obtained. Note that, the proposed GSF^{2}App expects an RGB image and its associated SFV as inputs. In the presented evaluation, a YOLOv3 model was used, however, other object recognition methods/pipelines could be employed.

IV. Experiments

The proposed approach was evaluated on two popular scene classification datasets: SUN RGB-D [11] and NYU Depth Dataset V2 [8]. In order to compare with state-of-the-art works [3][5][6][12][18][20], mean-class accuracy is used as the evaluation metric, which is calculated by averaging precision of all the categories as follows:

$$MeanAcc = \frac{1}{C}\sum_{c=1}^{C}\frac{Correct_c}{Total_c}$$

where $Correct_c$ is the number of correctly predicted samples of class c, $Total_c$ is the total number of samples of class c, and C is the total number of classes.

Additionally, an ablation study on both SUN RGB-D and NYU Depth datasets was conducted for more comprehensive evaluations of the proposed approach.

A. Datasets

SUN RGB-D Dataset: It contains 10,355 RGB and Depth image pairs captured from different cameras (Kinect v2, RealSense, Kinect v1, and Asus Xtion). It contains 10,335 RGB-D images distributed into 40 scene categories. Following the public split in [11], only 19 scene categories were selected for scene recognition evaluation, distributed in 4,845 images for training and 4,659 images for testing.

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE SUN RGB-D DATASET.

Method	Accuracy (%)	
	RGB	RGB-D
Li *et al.* [3]	46.3	54.6
Wang *et al.* [5]	-	48.1
Xiong *et al.* [6]	-	**55.9**
Du *et al.* [18]	42.6	53.3
Song *et al.* [20]	-	54.0
GSF^{2}App (ours)	55.3	-

NYU Depth Dataset v2: It has available 1449 images distributed into 27 scene categories. However, only a few of them are well represented. According to *Gupta et al.* [12], the original 27 scene categories must be reorganized into 10 scene categories (9 most common and "other"). Following [12], the dataset was split into 795 training and 654 test images.

B. Implementation Details

All experiments were performed using the publicly available PyTorch framework (version 1.0.1). In the first learning step, as mentioned, the VGG16 network was used as the base architecture and its parameters were initialized using the ImageNet pre-trained model. For the SUN RGB-D Dataset, in the first learning step, Stochastic Gradient Descent (SGD) optimizer method with learning rate of 0.001 was used over 75 iterations. In the second learning step, the ADAM optimizer method with learning rate of 0.0001 was used over 25 iterations. For the NYU Depth Dataset V2, during both learning steps, the ADAM optimizer method with learning rate of 0.0001 was used over 75 and 25 iterations respectively. A fixed momentum rate of 0.9, a weight decay rate of 0.0005, and a mini-batch size of 32 was also used in all learning steps. All experiences were performed using a NVIDIA RTX 2070 GPU, 32GB RAM, and an Intel i7-4790-@-3.60 GHz.

C. Results

We compare our achieved results with the recently state-of-the-art works [3][5][6][12][18][20]. Among them, Song *et al.* [20] and Wang *et al.* [5] introduced object detection based local feature learning methods. Li *et al.* [3] proposed a framework to learn distinctive and correlative features simultaneously. Xiong *et al.* [6] proposed a method that learns how to select important local region features. Among these works, [5] is the most related to ours. They achieved state-of-the-art performance by combining local with global features. In their work, local features represent the RoI detected in the image. It is important to highlight that most of these works used the AlexNet [24] as their CNN baseline and processed RGB and Depth data. In our work, VGG16 [9] network is used as the CNN baseline to process the RGB data.

Results on the SUN RGB-D Dataset

Table I shows the overall performance achieved in the SUN RGB-D Dataset. Our proposed framework achieves 55.3% of mean-class accuracy, getting close to the highest state-of-the art performance, which is 55.9% achieved by [6]. The confusion matrix can be seen in Fig. 7. An ablation study with intermediate evaluations of the proposed framework was conducted, whose results are summarized in Table II, which leads to the following observations:

TABLE II
ABLATION STUDY ON THE SUN RGB-D DATASET.

Proposed methods	Accuracy (%)
VGG16 (random initialization)	36.1
VGG16 (pre-trained model initialization)	54.5
SFV + SVM	39.2
SFV + FeedForward	42.0
VGG16 + SFV	55.3

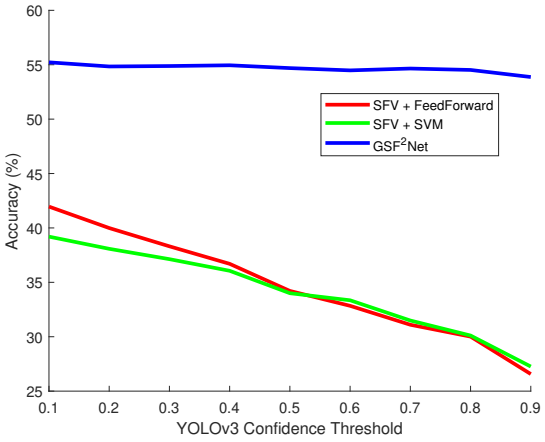


Fig. 6. Accuracy variation according with different YOLOv3's confidences thresholds on the SUN RGB-D Dataset.

CNN Baseline: It can be seen in Table II, that using the ImageNet pre-trained VGG16 model improves the model performance (from 36.1% to 54.5%). Note that during the training stage, both models (random/pre-trained model initialization) reached a very low loss value, however, as can be seen in the presented results, a better generalization of the model was achieved using the pre-trained model.

Semantic Features: In order to assess the potential that semantic features may have, a linear SVM and a FeedForward classifier were trained using the semantic feature vector as classifier inputs. Compared with the CNN baseline results, semantic feature learning had achieved acceptable results (39.2% with SVM and 42.0% with a FeedForward) showing that semantic features could be an asset to improve indoor scene classification tasks. Figure 6 shows the achieved results with YOLOv3's confidence threshold variations that directly affect the generated semantic feature vector. As expected, using lower thresholds leads to a higher number of semantic features which results in an improvement of the model's accuracy.

GSF^{2}App: It represents our final model, combining global and semantic features in a two-step learning approach (VGG16 + SFV). In overall, our approach combining the global CNN with the semantic features attained state of the art results,

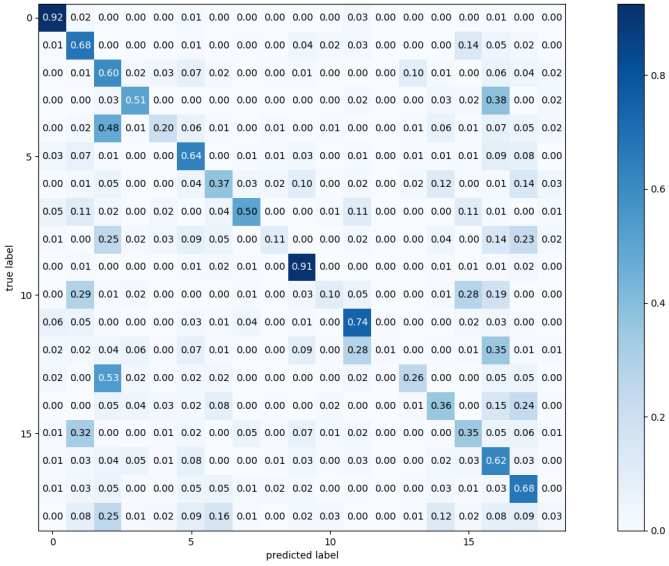


Fig. 7. Confusion matrix of GSF^{2}App on the SUN RGB-D dataset (Classes: 0 = bathroom, 1 = bedroom, 2 = classroom, 3 = computer room, 4 = conference room, 5 = corridor, 6 = dining area, 7 = dining room, 8 = discussion area, 9 = furniture store, 10 = home office, 11 = kitchen, 12 = lab, 13 = lecture theater , 14 = library, 15 = living room, 16 = office, 17 = rest space, 18 = study space).

TABLE III
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE NYU DEPTH DATASET V2.

Method	Accuracy (%)	
	RGB	RGB-D
Gupta *et al.* [12]	58.0	-
Li *et al.* [3]	61.1	65.4
Wang *et al.* [5]	53.5	63.9
Xiong *et al.* [6]	53.5	67.8
Du *et al.* [18]	53.7	67.5
Song *et al.* [20]	57.3	66.9
GSF^{2}App (ours)	**70.6**	-

that were also similar to the achieved results by the VGG16 network with the ImageNet pre-trained model.

Results on the NYU Depth Dataset V2

We also obtained results on the NYU Depth Dataset V2, where new observations can be made. Table III shows the overall performance achieved in the NYU Depth Dataset V2. The proposed framework achieves 70.6% mean-class accuracy, which, to the best of our knowledge, outperforms the recently reported results for this indoor scene dataset [8] (outperforming the Xiong *et al.* [6] achieved result). The confusion matrix of the achieved results can be seen in Fig. 9. An ablation study was also conducted, attaining the results shown in Table IV, of which the following observations are taken:

CNN Baseline: As mentioned before, using the ImageNet pre-trained VGG16 model significantly improves the model performance. In the case of the NYU Depth Dataset V2, which contains much less data than the SUN RGB-D dataset, using the pre-trained model allowed to achieve promising results, leading to a 26.7% mean-class accuracy gap between the

TABLE IV
ABLATION STUDY ON THE NYU DEPTH DATASET V2.

Proposed methods	Accuracy (%)
VGG16 (random initialization)	42.4
VGG16 (pre-trained model initialization)	69.1
SFV + SVM	57.6
SFV + FeedForward	57.4
VGG16 + SFV	**70.6**

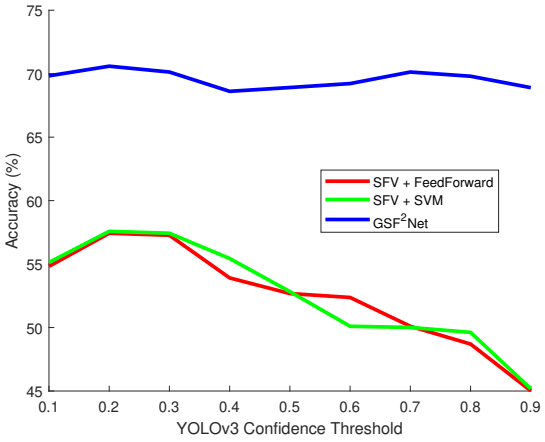


Fig. 8. Accuracy variation according with different YOLOv3's confidences thresholds on the NYU Depth Dataset v2.

random/pre-trained model initialization. The achieved results by the VGG16 network using the pre-trained model to initialize the parameters already outperforms the 67.8% of mean-class accuracy achieved by Xiong's work [6].

Semantic Features: The same individual semantic features evaluation was performed. Once again, promising results were achieved (≈57.7%), showing that semantic features can be very useful in this kind of applications. Specially, when compared with the random initialization CNN result that only achieved 42.4% of mean-class accuracy. As shown in Fig. 8, once again, with lower thresholds, the highest accuracy is achieved.

GSF^{2}App: Combining global and semantic features in a two-step learning approach, significantly improved the CNN baseline result by 1.5%. Our final performance gets 70.6% of mean-class accuracy on the NYU Depth Dataset V2, outperforming the recently state-of-the-art accuracy [6] by over 2.5%.

Notice that, the best final achieved results used semantic feature vectors that were extracted from the COCO dataset pre-trained YOLOv3 model with confidence thresholds of 0.1 and 0.2 on the SUN RGB-D and NYU Depth dataset respectively.

V. CONCLUSION

In this paper, a deep-learning based Global and Semantic Feature Fusion Approach (GSF^{2}App) with two branches for RGB indoor scene classification is proposed. Semantic features represent the objects recognized in the RGB image along with the number of times that the same object appears in the same image. Semantic features were extracted through the COCO

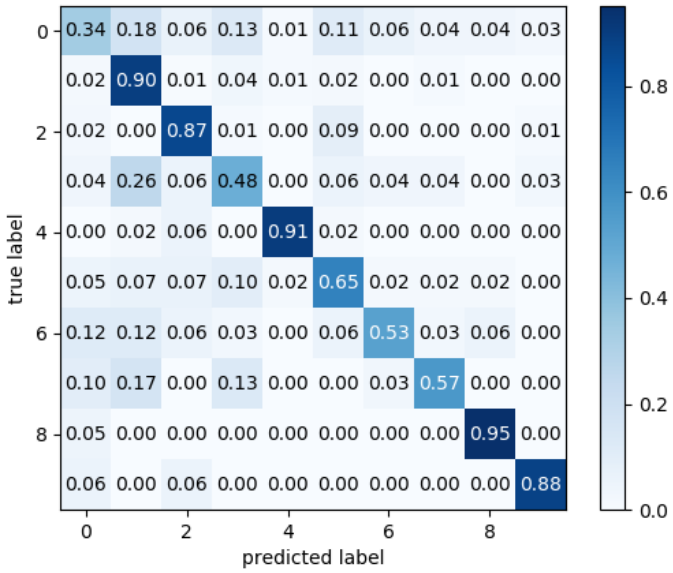


Fig. 9. Confusion matrix of GSF^{2}App on the NYU Depth Dataset V2 (Classes: 0 = others, 1 = bedroom, 2 = kitchen, 3= livingRoom, 4 = bathroom, 5 = diningRoom, 6 = office, 7 = homeOffice, 8 = classroom, 9= bookStorage).

dataset pre-trained YOLOv3 model. The proposed approach is trained in two steps. In the first one, the ImageNet pre-trained VGG16 CNN model is used to initialize our CNN model that is trained to extract global features from RGB image. In the second learning step, the model learns how to combine semantic features with Global-CNN features. Promising results were attained for RGB indoor scene classifications on both SUN RGB-D Dataset and NYU Depth Dataset V2, which validate the proposed learning approach, however, a good classification is dependent from the scene view. Note that we use an object-centric pre-trained model to initialize our model, however, promising results were achieved in indoor scene classification, showing that the model initialization is an important factor that must be taken into account. Despite the occurrences of errors in object recognition, reported results show that semantic features can be useful for indoor scene classifications tasks.

ACKNOWLEDGMENTS

This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the PhD grant with reference SFRH/BD/148779/2019. This work has been also supported by the projects B-RELIABLE with reference SAICT/30935/2017 (with FEDER/FNR/OE funding through programs CENTRO2020 and FCT) and MATIS-CENTRO-01-0145-FEDER-000014, Portugal. It was also partially supported by FCT through grant UID/EEA/00048/2019.

REFERENCES

[1] C. Premebida, D. Faria, and U. J. Nunes, "Dynamic bayesian network for semantic place classification in mobile robotics," in *Autonomous Robots*, vol. 41, 2017, pp. 1161–1172.

[2] C. Premebida, D. R. Faria, F. A. Souza, and U. J. Nunes, "Applying probabilistic mixture models to semantic place classification in mobile robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[3] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "DF2Net: Discriminative Feature Learning and Fusion Network for RGB-D Indoor Scene Classification," *AAAI Conference on Artificial Intelligence*, 2018.

[4] B. Zhou, g. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," *Advances in Neural Information Processing Systems*, 2015.

[5] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Modality and Component Aware Feature Fusion For RGB-D Scene Classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Z. Xiong, Y. Yuan, and Q. Wang, "RGB-D Scene Recognition via Spatial-Related Multi-Modal Feature Learning," *IEEE Access*, vol. 7, 2019.

[7] X. Song, S. Jiang, L. Herranz, and C. Chen, "Learning Effective RGB-D Representations for Scene Recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 980–993, 2019.

[8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGB-D Images," in *ECCV*, 2012.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICLR)*, 09 2014.

[10] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 2018.

[11] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 567–576.

[12] S. Gupta, P. Arbeláez, and J. Malik, "Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[13] M. George, M. Dixit, G. Zogg, and N. Vasconcelos, "Semantic Clustering for Robust Fine-Grained Scene Recognition," *CoRR*, 2016.

[14] M. Dixit, Si Chen, Dashan Gao, N. Rasiwasia, and N. Vasconcelos, "Scene classification with semantic Fisher vectors," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *CoRR*, 2015.

[16] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[17] M. M. Rahman, Y. Tan, J. Xue, and K. Lu, "RGB-D object recognition with multimodal deep convolutional neural networks," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[18] D. Du, X. Xu, T. Ren, and G. Wu, "Depth Images Could Tell us More: Enhancing Depth Discriminability for RGB-D Scene Recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[19] S. Gupta, R. B. Girshick, P. Arbelaez, and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," *CoRR*, 2014.

[20] X. Song, C. Chen, and S. Jiang, "RGB-D Scene Recognition with Object-to-Object Relation," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 600–608.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, 2016.

[23] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *CoRR*, 2015.

[24] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems*, vol. 25, 2012.