Hindawi
Journal of Sensors
Volume 2023, Article ID 9818617, 1 page
https://doi.org/10.1155/2023/9818617

Retraction

Retracted: Scene Classification Using Deep Networks Combined with Visual Attention

Journal of Sensors

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Shi, H. Zhu, Y. Li, Y. Li, and S. Du, "Scene Classification Using Deep Networks Combined with Visual Attention," *Journal of Sensors*, vol. 2022, Article ID 7191537, 9 pages, 2022.

Hindawi
Journal of Sensors
Volume 2022, Article ID 7191537, 9 pages
https://doi.org/10.1155/2022/7191537

Research Article

Scene Classification Using Deep Networks Combined with Visual Attention

Jing Shi,[1,2] **Hong Zhu**,[1] **Yuxing Li**,[1] **YangHui Li**,[1] **and Sen Du**[1]

[1]*School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China*
[2]*Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing, Xi'an University of Technology, Xi'an 710048, China*

Correspondence should be addressed to Jing Shi; shijing@xaut.edu.cn and Hong Zhu; zhuhong@xaut.edu.cn

Received 13 July 2022; Accepted 12 August 2022; Published 28 August 2022

Academic Editor: Yuan Li

In view of the scene's complexity and diversity in scene classification, this paper makes full use of the contextual semantic relationships between the objects to describe the visual attention regions of the scenes and combines with the deep convolution neural networks, so that a scene classification model using visual attention and deep networks is constructed. Firstly, the visual attention regions in the scene image are marked by using the context-based saliency detection algorithm. Then, the original image and the visual attention region detection image are superimposed to obtain a visual attention region enhancement image. Furthermore, the deep convolution features of the original image, the visual attention region detection image, and the visual attention region enhancement image are extracted by using the deep convolution neural networks pretrained on the large-scale scene image dataset Places. Finally, the deep visual attention features are constructed by using the multilayer deep convolution features of the deep convolution networks, and a classification model is constructed. In order to verify the effectiveness of the proposed model, the experiments are carried out on four standard scene datasets LabelMe, UIUC-Sports, Scene-15, and MIT67. The results show that the proposed model improves the performance of the classification well and has good adaptability.

1. Introduction

As a basic problem in the field of computer vision and image understanding, scene image classification has received extensive attention and research [1–7]. The most important problem to be solved in scene classification is to give proper expression to the content in the scene. In order to improve the accuracy of scene classification, the researchers constantly explore new ways, which has advantage of global features and local features as well as the middle to form visual word bag; the bag will represent a visual scene image word combination methods [8] and, by iteration and cross-validation, get the image block with degree of differentiation, as image middle expression method of [9]. The mean-shift algorithm is used to find the distinguishing mode in the image block distribution space, so as to create the image representation method of middle-level scene [10]. By establishing metric learning formulas and learning the best metric parameters, online metric learning and parallel optimization of large-scale and high-dimensional data can be solved [2]. Although these methods have achieved certain classification effects, the classification performance is reduced when there are many objects or complex contents in the scene image.

In recent years, the proposal of deep convolutional networks has made it possible to obtain richer high-level semantics of images [11–13]. Donahue et al. directly use convolutional neural networks (CNNs) [14], which are pretrained on ImageNet dataset, for scene classification. Zhou et al. constructed a large-scale dataset centered on the scene and trained convolutional neural network on this basis [15], which significantly improved the performance of scene classification. Bai proposed that through CNN transfer learning, deep features were used to express special scene targets for classification [16]. Zou et al. built a fusion method based on nonnegative matrix factorization, which can preserve feature nonnegative properties and improve their representation performance.

Furthermore, an adaptive feature fusion and boosting algorithm is developed to improve the efficiency of image features. There are two versions of the proposed feature fusion method for nonnegative single-feature fusion and multifeature fusion [1].

Zhang et al. proposed a spatially aware aggregation network for scene classification, which detects a set of visually semantically significant regions from each scene through a semisupervised and structurally reserved nonnegative matrix decomposition (NMF). Gaze shift path (GSP) was used to characterize the process of human perception of each scene image, and a spatial perception CNN called SA-NET was developed to describe each GSP in depth. Finally, the deep GSP function learned from the whole scene image is integrated into the image kernel, which is integrated into the kernel SVM to classify the scene [3]. Yee et al. propose a DeepScene model that leverages convolutional neural network as the base architecture, which converts grayscale scene images to RGB images. Spatial Pyramid Pooling is incorporated into the convolutional neural network [17]. These methods have greatly improved the effect of scene classification.

However, most of the current algorithms regard the scene as a combination of multiple objects [18–20] and lack description of contextual semantic relations between objects, thus restricting the accuracy of scene classification [21]. For this purpose, the significance of detection algorithm based on context [22], annotation in the scene visual focus area, and the area contains the main target in the scene and can express the context of a part of the background region and at the same time, combined with the depth of the convolutional neural network, build a kind of fusion depth scene classification characteristic of visual attention model. It overcomes the limitation of using object and structure feature to classify effectively and obtains good scene classification performance.

2. The Construction of Scene Classification Model

In order to adapt to the diversity of images, this article will image the context of the significant characteristics as visual attention characteristics, superimposed onto the original image and into the depth of the convolution network; build scene classification model; make the model of images to express deep intrinsic characteristics at the same time; andalso can express the target in the scene context between semantic features.

2.1. Detection of Areas of Visual Attention. The area that has a major influence on visual judgment is called the area of visual attention. Here, the context-based saliency detection algorithm proposed by Goferman et al. is used to extract the visual areas of interest of the image [22]. The extracted saliency areas take full account of global and local features at different scales and mark saliency targets with their adjacent areas to varying degrees. It well reflects the contextual relationship between the objects in the scene and the surrounding scenery and filters out some repetitive texture information.

The image block is taken as the comparison unit and compared in Lab color space. The closer the distance is, the greater the difference is and the more significant it is. The difference degree of the two image blocks is expressed as

$$d\left(p_i, p_j\right) = \frac{d_c\left(p_i, p_j\right)}{1 + c \cdot d_p\left(p_i, p_j\right)}, \tag{1}$$

where p represents the image block, i is the central pixel point of p_i, and $d_c(p_i, p_j)$ and $d_p(p_i, p_j)$ represent the color distance and spatial distance between the two blocks, respectively.

For the difference degree value under a single scale, usually, only the difference degree value between the previous block and a certain block and the significance value of the pixel point under the scale need to be calculated as follows:

$$S_i^r = 1 - \exp\left\{-\frac{1}{M}\sum_{m=1}^{M} d(p_i^r, p_m^r)\right\}, \tag{2}$$

where M is the most similar front block taken, $\{p_m\}_{m=1}^{M}$, and r represents a certain scale.

In order to make the detected significance region significant in multiple scales, it is necessary to calculate multiple single-scale significance values and then take the average value, as follows:

$$\bar{S}_i = \frac{1}{K}\sum_{r\in R} S_i^r, \tag{3}$$

where K represents the number of scales and R represents the scale space.

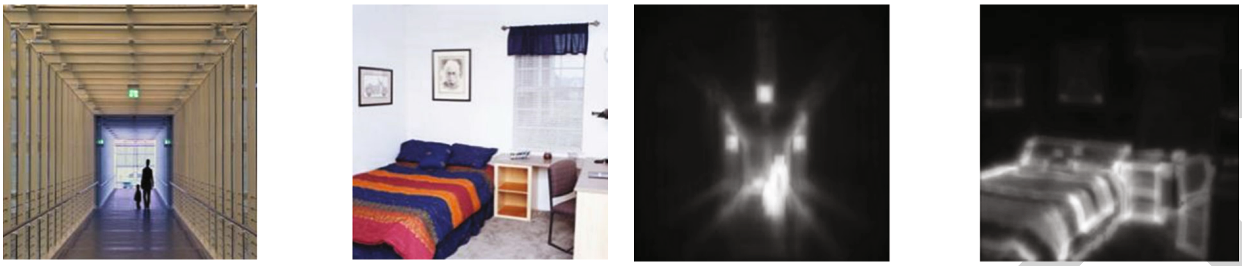
In addition, it is necessary to combine the context of the image to make the region with different distances from the saliency target have different saliency. The significance value of pixel point is finally defined as

$$\hat{S}_i = \frac{1}{K}\sum_{r\in R} S_i^r\left(1 - d_f^r(i)\right), \tag{4}$$

where $d_f^r(i)$represents the Euclidean distance between the pixel point i at the scale r and the nearest pixel point in the significant region. K represents the number of scales, and R represents the scale space.

Figure 1 is an example of detecting the area of visual attention. The brightness value in Figure 1(b) is the visual attention of this position. It can be seen that the attention of the background area varies with the change of closeness to the target.

2.2. Construction of Enhanced Images of Visual Areas of Concern. Although the scene information contained in the original image was comprehensive, it could not distinguish effective information from invalid information. In order to enhance the scene area containing different information to different degrees, the detection map of visual area of concern was superimposed on the original image to obtain the enhanced image of visual area of concern.



(a) Original images

(b) Visual attention area detection images

Figure 1: Example of visual area of attention detection.

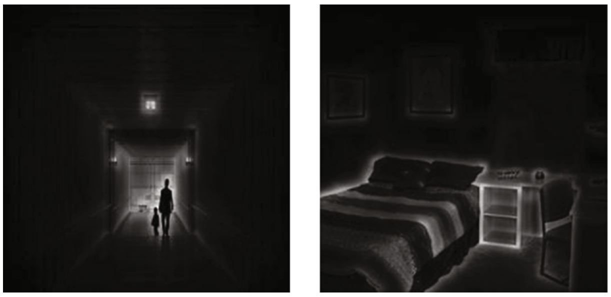


Figure 2: Enhanced images of areas of visual attention.

Assume that $f(i,j)$ is the original image, $f_s(i,j)$ is the detection map of visual area of concern, and $f_e(i,j)$ is the enhancement map of visual area of concern. Before stacking, the size of the original image and the detection map of visual area of concern are normalized to 256 × 256, and the significance value of the detection map of visual area of concern is normalized to [0,1].

$$f_e(i,j) = f_s(i,j).*f(i,j), \quad i \in M, j \in N. \tag{5}$$

Figure 2 shows the enhancement of visual attention area. It can be seen that different areas in the scene have different visual attention, and some repeated textures and less obvious regional information in the scene have been effectively suppressed, such as the passage in the airport scene and the decorative paintings in the bedroom. While preserving visual attention, the superimposed images supplement the information of some gray areas (transitional areas between the attention and unattention areas).

2.3. Deep Feature Fusion. In order to describe the content attributes of scene images effectively, AlexNet network model, which has been pretrained on large-scale scene dataset Places, is used to extract the deep convolution features of original image, visual area of concern detection image, and visual area of concern enhancement image. In addition, since different layers of the deep convolutional network have different abstract expressions of the original image data, the output of multiple fully connected layers of the deep convolutional network is used in this paper to form the deep fusion feature as the final expression of the scene image.

As shown in Figure 3, in the full connection layer of AlexNet, convolution feature is expressed in layer 6, and classification association feature is expressed in layer 7. Therefore, in this paper, the 4096-dimensional output feature of layer 7 and layer 6 is connected in series to generate the deep fusion feature of the image, and the calculation formula is as follows:

$$F_c = [F_{\mathrm{fc7}}, F_{\mathrm{fc6}}], \tag{6}$$

where F_c is the depth fusion feature of FC7 and FC6, F_{fc7} is the output feature of layer FC7, and F_{fc6} is the output feature of layer FC6.

Then, the deep fusion features of the original image, the detection image of visual focus area, and the enhanced image of visual focus area of the same image are spliced to generate the deep visual focus features. The calculation formula is as follows:

$$F_{\mathrm{VS}} = [F_c, F_{c\_s}, F_{c\_e}], \tag{7}$$

where F_{VS} are the features of deep visual attention, F_c, F_{c_s}, and F_{c_e} are the deep fusion features of the original image, the detection image of visual attention area, and the enhanced image of visual attention area, respectively.

Finally, the depth visual attention features of the training images in the target dataset are sent into the random forest to train the classifier, and the trained classifier is used for scene classification. Because the extracted features have both the contextual semantic relationship between objects in the image and the intrinsic characteristics of scene depth, the effectiveness of scene classification is greatly improved.

3. Experimental Results and Analysis

3.1. Datasets and Experimental Settings. This paper conducted tests on four common standard scene datasets, LabelMe (OT) [23], UIUC-Sports (SE) [24], Scene-15 (LS) [23, 25, 26], and MIT67 (IS) [27], respectively, partial images of each scene dataset are shown in Figure 4. In order to compare with similar algorithms, experiments were carried out according to the training and testing ratios of different datasets in the references, and the average classification accuracy of 10 experiments was taken as the final test result.

(i) The LabelMe (OT) dataset contains 2688 color images of 8 categories, all 256 × 256 in size. In each category, 200 images were randomly used for training, and the rest images were used as test images

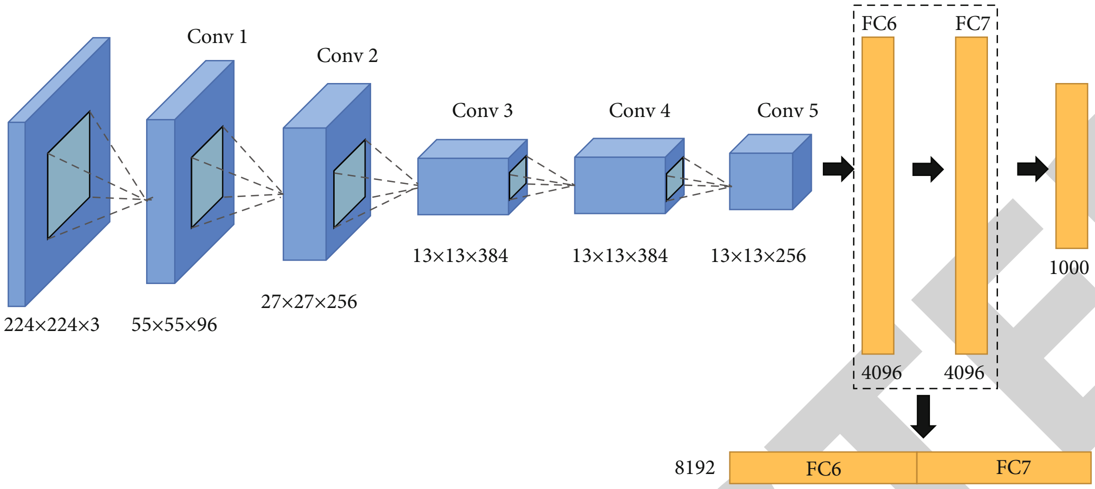


Figure 3: Schematic diagram of generating deep fusion features.

Figure 4: Partial images of each dataset.

(ii) UIUC-Sports (SE) dataset contains 1579 color sports scene images of different sizes in 8 categories. Each category was randomly assigned 70 images for training and 60 images for test

(iii) Scene-15 (LS) dataset contains a total of 4485 indoor and outdoor scene images of 15 categories, of which 8 categories are the same as the LabelMe dataset. 100 images were randomly used in each category for training, and the rest of the images were used as test images

(iv) MIT67 (IS) is a challenging indoor scene image dataset containing a total of 15,620 images in 67 categories. 80 images were randomly used in each category for training, and 20 images were used as test images

3.2. Classification Performance Evaluation. Figure 5 shows the comparison test results of classification on four datasets between deep learning features without visual attention region detection and features proposed in this paper by using the same classification method.

It can be seen that all the features proposed in this paper have certain effects on the test datasets, and the classification accuracy effect is most significantly improved in the LS dataset, mainly because the dataset contains indoor and outdoor scenes, which indicates that the algorithm is universal. In addition, the classification effect of simple outdoor scenes is also significantly improved. However, the effect of the SE and IS dataset is limited, mainly due to the fact that there are many objects in the scene, and the context relationship of prominent objects in the scene is complex. People in many scenes are not the main objects to distinguish scenes,

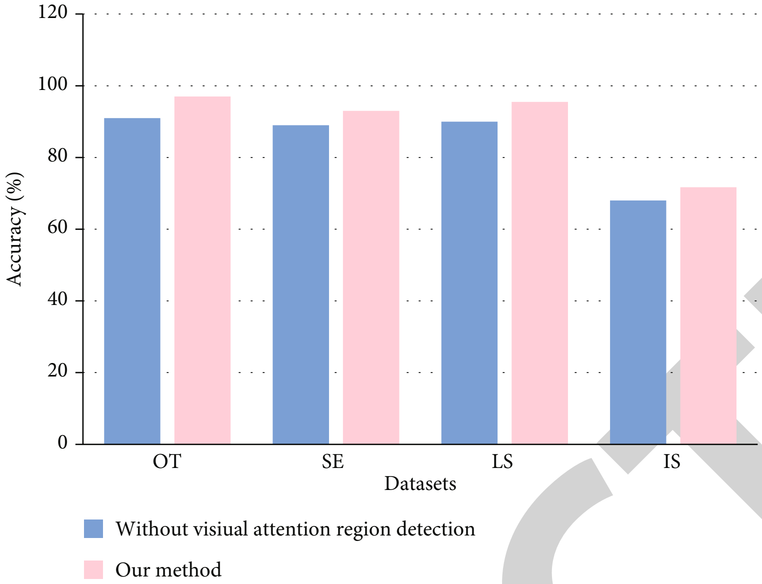


Figure 5: Comparison of experimental results using visual attention area detection.

Table 1: Fusion matrix of the OT dataset.

Category	1	2	3	4	5	6	7	8	Recall (%)
MITcoast (1)	154	2	—	—	—	4	—	—	96.3
MITforest (2)	—	127	—	—	1	—	—	—	99.2
MIThighway (3)	—	—	59	—	—	1	—	—	98.3
MITinsidecity (4)	—	—	—	108	—	—	—	—	100
MITmountain (5)	—	1	—	—	171	2	—	—	98.3
MITopencountry (6)	7	5	—	—	1	197	—	—	93.8
MITstreet (7)	—	—	—	—	—	—	91	1	98.9
MITtallbuilding (8)	—	—	—	—	—	—	—	156	100
Precision (%)	95.7	94.1	100	100	98.8	96.6	100	99.4	

but sometimes, they are enhanced as prominent objects, which interferes with the discrimination of scene content. In particular, scene discrimination in the SE dataset is mainly determined by the relationship between characters' actions and scenes, while characters' actions are sometimes very similar in multiple scenes. Therefore, the classification effect of visual area of concern detection algorithm on these datasets is limited.

Precision and recall were used to evaluate and analyze the OT, SE, and LS dataset, respectively.

Table 1 shows the fusion matrix obtained by a test of the method in this paper on the OT dataset. It can be seen that this method can achieve 100% accuracy and recall rate in the "MITinsidecity" class and can also achieve good classification effect for other categories. It is easy to confuse the "MITopencountry" class with the "'MITcoast" class. Figure 6 shows partial misclassification images of the OT dataset. These two images misclassify images of "MITopencountry" into "MITcoast," because the context relationship between sky and land in "MITopencountry" is similar to that of sky and coast in "MITcoast." Lawns and deserts on the slopes have a similar texture to sea level.

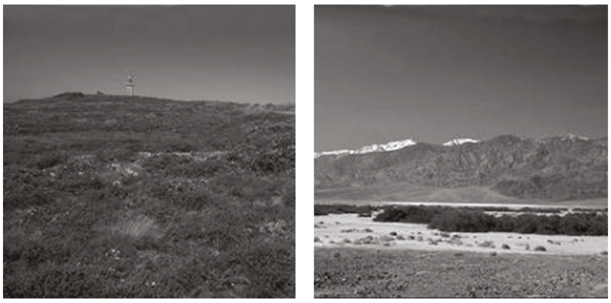


Figure 6: Misclassification images in OT.

Table 2 shows the fusion matrix obtained by the proposed method in a certain test on the LS dataset. The most confusing categories are "Bedroom" and "Livingroom" and "MITtallbuilding" and "Industrial." Figure 7 shows partial segmentation images of the LS dataset. In (a), the scene images of "Industrial" are misclassified to "MITtallbuilding," because the high-rise buildings in the image are very similar in appearance to tall buildings, and the context relationship with the surrounding environment is similar to that of

Table 2: The fusion matrix of the LS dataset.

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Recall (%)
Bedroom (1)	107	—	—	—	9	—	—	—	—	—	—	—	—	—	—	92.2
CALsuburb (2)	—	141	—	—	—	—	—	—	—	—	—	—	—	—	—	100
Industrial (3)	—	—	192	1	1	—	—	—	3	—	—	1	5	—	8	91.0
Kitchen (4)	—	—	—	103	2	—	—	—	—	—	—	—	—	4	1	93.6
Livingroom (5)	4	—	—	2	179	—	—	—	1	—	—	—	—	2	1	94.7
MITcoast (6)	—	—	—	—	—	252	2	—	—	1	5	—	—	—	—	96.9
MITforest (7)	—	—	—	—	—	—	221	—	—	6	1	—	—	—	—	96.9
MIThighway (8)	—	—	1	—	—	—	—	158	—	—	1	—	—	—	—	98.8
MITinsidecity (9)	—	2	5	3	—	—	—	—	194	—	—	3	1	—	—	93.3
MITmountain (10)	—	—	—	—	—	—	—	—	—	273	1	—	—	—	—	99.6
MITopencountry (11)	—	—	—	—	—	6	1	—	—	6	297	—	—	—	—	95.8
MITstreet (12)	—	—	3	—	—	—	—	—	2	—	—	186	—	—	1	96.9
MITtallbuilding (13)	—	—	9	—	—	—	2	—	1	—	—	—	243	—	1	94.9
PARoffice (14)	1	—	—	3	1	—	—	—	1	—	—	—	—	109	—	94.8
Store (15)	—	—	1	1	1	—	—	—	—	—	—	—	1	—	211	98.1
Precision (%)	95.5	98.6	91.0	91.2	92.7	97.7	97.8	100	96.0	95.5	97.4	97.9	97.2	94.8	94.6	

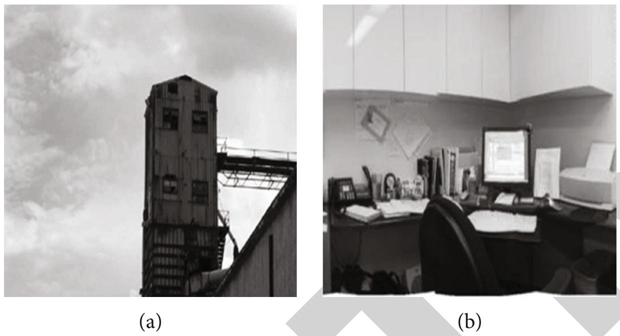

(a) (b)

Figure 7: Misclassification images in LS.

"MITtallbuilding." On the right, the scene image of "PARoffice" is misclassified into "Kitchen." The reason is that the cabinet in the upper part of the image has the same position and appearance with the cabinet, and the context relationship with the desktop is similar to that of "Kitchen," thus causing the misclassification.

Table 3 shows the fusion matrix obtained by a certain test of the method in this paper on the SE dataset. It can be seen that the method can achieve 100% accuracy and recall rate in the "Sailing" class. However, "Bocce" has the lowest accuracy and recall rate, and it is most easily confused with "Croquet," mainly because the identification of these two scenes is mainly based on character movements and the relationship between character and scene, and the character movements of these two sports are very similar to the surrounding environment, so it is easy to misjudge.

3.3. Comparison of Experimental Results. The experimental results of the proposed method on four standard scene datasets were compared with those of the reference method.

The comparison test results on the OT dataset are shown in Table 4. It can be seen that the algorithm using deep convolutional network has obvious advantages over the traditional feature extraction algorithm. GECMCT method [28] adds the far neighborhood information to the nonparametric transformation calculation and spatial information. Gist feature and spatial correction census transform are combined to form a new image descriptor, but this method lacks the deep description of scene images. HGD algorithm [29] uses pLSA to train multichannel classifier on the topic distribution vector of each image, which is not only complex in modeling but has also limited classification effect. Compared with other algorithms using deep convolutional networks, the deep convolutional classification model constructed in this paper based on the visual area of interest of images has obviously better classification effect.

The comparative test results on the SE dataset are shown in Table 5. It can be seen that the classification performance of the model in this paper is significantly better than other classification algorithms. Among them, SKDES+Grad+color+shape method [8] embeds image and label information into block-level kernel descriptors to form supervised kernel descriptors and uses visual word bags to learn low-level block expressions. The implementation process of this method is relatively complex. And the representational ability of visual word bag is limited. LGF methods classified using global and local features of images, not the analysis on the contents of the image; using visual attention area detection algorithms on different areas of the image effectively; and combining the depth have been trained on the Places dataset convolution network and can better access the spatial structure of image information [30]. AdaNSFF-Color boost [1] proposes a novel fusion framework of adaptive nonnegative feature fusion (AdaNFF) for scene classification. The AdaNFF integrates nonnegative matrix factorization, adaptive feature fusion, and feature fusion

Table 3: The fusion matrix of the SE dataset.

Category	1	2	3	4	5	6	7	8	Recall (%)
Badminton (1)	57	3	—	—	—	—	—	—	95.0
Bocce (2)	3	45	9	—	2	—	—	1	75.0
Croquet (3)	—	5	54	—	—	1	—	—	90.0
Polo (4)	1	3	—	56	—	—	—	—	93.3
RockClimbing (5)	—	—	—	—	59	—	—	1	98.3
Rowing (6)	—	—	—	—	—	60	—	—	100
Sailing (7)	—	—	—	—	—	—	60	—	100
Snowboarding (8)	1	—	—	—	—	—	—	59	98.3
Precision (%)	91.9	80.4	85.7	100	96.7	98.4	100	96.7	

Table 4: Comparative test results on the OT dataset.

Methods	Accuracy (%)
CENTRIST [31]	76.49
CMCT [32]	79.91
GIST [23]	82.60
LPC [33]	83.40
GECMCT [28]	86.95
HGD [29]	87.80
ImageNet-CNN [14]	92.83
Places-CNN [15]	94.30
DSPMK+RVFs+BSRC [34]	95.11
Ours	97.70

Table 5: Comparative test results on the SE dataset.

Methods	Accuracy (%)
Object bank [19]	76.30
LRML-PCDM [2]	76.97
GECMCT [28]	77.96
GIST [23]	82.60
O2C kernels [18]	86.02
CENTRIST [31]	86.22
SC+LCSR [35]	87.23
LGF [41]	88.52
AdaNSFF-Color boost [1]	90.21
SKDES+grad+color+shape [8]	91.00
Ours	93.75

boosting into an end-to-end process. However, although this method fuses and enhances features, the training data lacks pertinence, which affects its generalization ability.

Table 6: Comparative test results on the LS dataset.

Methods	Accuracy (%)
Bow [25]	74.80
CMN [36]	77.20
SPMSM [37]	82.50
GECMCT [28]	82.96
EMFS [38]	85.70
DLGB (saliency) [5]	87.40
O2C kernels [18]	88.80
ISPR+IFV [39]	91.06
Hybrid-CNN [15]	91.59
DGSK [3]	92.30
MKL [30]	92.50
DDSFL+Caffe [11]	92.81
DeepScene [17]	95.60
SDO+ fc features [4]	95.88
Ours	96.01

Table 7: Comparative test results on the IS dataset.

Methods	Accuracy (%)
GECMCT [28]	36.57
Object bank [19]	37.60
Midlevel elements+IFV [10]	66.87
ImageNet-CNN [14]	56.79
MOP-CNN [12]	68.88
Places-CNN [15]	68.24
Hybrid-CNN [15]	70.80
DeepScene [17]	71.00
S2ICA [6]	71.20
DLGB(saliency) [5]	71.40
CNNaug+SVM [13]	71.90
RF-CNNs [40]	72.35
Ours	72.37

The comparison test results on the LS dataset are shown in Table 6. The algorithm in this paper still has a good classification effect, which is not only better than traditional classification methods but also better than many classification methods using deep learning. The features of SDO+fc algorithm of cooccurrence with all objects in the scene mode [4] and the correlation of different objects in the scene configuration to choose representative and distinguish between objects, thus, enhance the discriminability between the classes, with the emergence of identifying objects in the image block probability to represent the image descriptors and to eliminate the influence of the public target. Although the algorithm considers the correlation between objects in the scene, it is still limited to the simple object and does not consider the surrounding background area adjacent to the object, so the classification effect is limited. DeepScene [17] integrates Spatial Pyramid Pooling into the convolutional neural network to perform multilevel pooling on the scene image and implements the weighted average ensemble of convolutional neural networks to fuse the class scores thus improving the overall performance in scene classification. However, this method still takes the scene as a whole and does not enhance the information representing of scene, so the classification effect is limited.

The comparative test results on the IS dataset are shown in Table 7. It can be seen that, similar to the results of other datasets, the effect of using convolutional neural network is significantly better than that of traditional features in general, and the classification results in this paper still have obvious advantages. Among them, the Places-CNN algorithm [15] using the Places dataset pretraining network outperforms the ImageNet-CNNS algorithm [11] using the ImageNet dataset pretraining network by nearly 12%, mainly because the network using the scene training is more effective in judging the scene category. However, hybrid-CNN algorithm uses the pretrained network of ImageNet and Places datasets to extract deep convolution features of images [15]. Therefore, the classification effect is improved compared with the first two algorithms. However, the algorithm in this paper only uses the pretrained convolutional neural network of the Places dataset and combines the context information of the visual attention area of the image to achieve a better classification effect than hybrid-CNN algorithm.

4. Conclusions

This paper proposes a scene classification model based on the depth feature of visual focus area. Based on the context of significant regional detecting scene image visual interest areas in the image, with the original image overlay, enhance image visual interest area, then, will the three images into AlexNet, respectively, extraction depth visual focus features, in the end, will it into to the random forest classifier for training and classification.

Since the feature extraction model of deep visual attention in this paper also describes the target information in the image and the contextual semantic information between the target and the surrounding scene, different visual attention is constructed and the expression ability of scene characteristics is improved. Combined with the feature of multilayer deep convolution, the deep visual expression of scene image is constructed. The test results on four standard scene image sets verify the effectiveness of the proposed method, and it is better than several methods with good characteristics.

The method for the classification of the test datasets as a whole has a good effect, but when individual scene image content itself exists, ambiguity or visual attention content area does not fully express the scene; there is still a fault phenomenon, to this, and another step in the future research work will be digging deeper visual focusing on context information, in order to obtain better classification effect.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Natural Science Basic Research Program of Shaanxi Province, Grant Number 2021JQ-487.

References

[1] Z. Zou, W. Liu, and W. Xing, “AdaNFF: a new method for adaptive nonnegative multi-feature fusion to scene classification,” *Pattern Recognition*, vol. 123, article 108402, 2022.

[2] G. Sun, Y. Cong, Q. Wang, and X. Xu, “Online low-rank metric learning via parallel coordinate descent method,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 207–212, Beijing, China, 2018.

[3] A. Trouilloud, L. Kauffmann, A. Roux-Sibilon et al., “Rapid scene categorization: From coarse peripheral vision to fine central vision,” *Vision Research*, vol. 170, pp. 60–72, 2020.

[4] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, “Scene recognition with objectness,” *Pattern Recognition*, vol. 74, pp. 474–487, 2018.

[5] L. Zhang, R. Liang, J. Yin, D. Zhang, and L. Shao, “Scene categorization by deeply learning gaze behavior in a semisupervised context,” *IEEE Transactions on Cybernetics*, vol. 51, no. 8, pp. 4265–4276, 2021.

[6] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, “A spatial layout and scale invariant feature representation for indoor scene classification,” *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4829–4841, 2016.

[7] Y. Li, B. Geng, and S. Jiao, “Dispersion entropy-based Lempel-Ziv complexity: a new metric for signal analysis,” *Chaos, Solitons & Fractals*, vol. 161, article 112400, 2022.

[8] P. Wang, J. Wang, G. Zeng, W. Xu, H. Zha, and S. Li, “Supervised kernel descriptors for visual recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2858–2865, Portland, Oregon, USA, 2013.

[9] S. Singh, A. Gupta, and A. A. Efros, “Unsupervised discovery of mid-level discriminative patches,” in *European Conference on Computer Vision*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7573 of Lecture Notes in Computer Science, pp. 73–86, Springer, Berlin, Heidelberg, 2012.

[10] C. Doersch, A. Gupta, and A. A. Efros, “Mid-level visual element discovery as discriminative mode seeking,” *Neural Information Processing Systems*, vol. 26, pp. 494–502, 2013.

[11] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang, “Exemplar based deep discriminative and shareable feature learning for

scene image classification," *Pattern Recognition*, vol. 48, no. 10, pp. 3004–3015, 2015.

[12] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European conference on computer vision*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8695 of Lecture Notes in Computer Science, , pp. 392–407, Springer, Cham, 2014.

[13] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, Columbus, OH, USA, 2014.

[14] J. Donahue, Y. Jia, O. Vinyals et al., "DeCAF: a deep convolutional activation feature for generic visual recognition," *International conference on machine learning*, vol. 32, pp. 647–655, 2014.

[15] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Neural Information Processing Systems*, vol. 27, pp. 487–495, 2014.

[16] S. Bai, "Scene categorization through using objects represented by deep features," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 9, article 1755013, 2017.

[17] P. S. Yee, K. M. Lim, and C. P. Lee, "DeepScene: scene classification via convolutional neural network with spatial pyramid pooling," *Expert Systems with Applications*, vol. 193, article 116382, 2022.

[18] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3241–3253, 2014.

[19] L. Li, H. Su, L. Feifei, and E. Xing, "Object Bank: a high-level image representation for scene classification & semantic feature sparsification," *Neural Information Processing Systems*, vol. 23, pp. 1378–1386, 2010.

[20] Y. Li, B. Tang, and Y. Yi, "A novel complexity-based mode feature representation for feature extraction of ship-radiated noise using VMD and slope entropy," *Applied Acoustics*, vol. 196, article 108899, 2022.

[21] Y. Li, B. Tang, X. Jiang, and Y. Yi, "Bearing fault feature extraction method based on GA-VMD and center frequency," *Mathematical Problems in Engineering*, vol. 2022, Article ID 2058258, p. 19, 2022.

[22] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[23] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[24] L. Li and L. Feifei, "What, where and who? Classifying events by scene and object recognition," in *2007 IEEE 11th international conference on computer vision*, pp. 1–8, Rio de Janeiro, Brazil, 2007.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pp. 2169–2178, New York, NY, USA, 2006.

[26] L. Feifei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 524–531, San Diego, CA, USA, 2005.

[27] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 413–420, Miami, FL, USA, 2009.

[28] G. K. A. de Souza and E. O. T. Salles, "Exploring neighborhood and spatial information for improving scene classification," *Pattern Recognition Letters*, vol. 46, pp. 83–88, 2014.

[29] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[30] X. Sun, L. Zhang, Z. Wang et al., "Scene categorization using deeply learned gaze shifting kernel," *IEEE Transactions on Cybernetics*, vol. 49, no. 6, pp. 2156–2167, 2019.

[31] J. Wu and J. M. Rehg, "CENTRIST: a visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[32] G. K. A. de Souza and E. O. T. Salles, "A contextual image descriptor for scene classification," in *Online Proceedings on Trends in Innovative Computing*, pp. 66–71, 2012.

[33] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *European conference on computer vision*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6311 of Lecture Notes in Computer Science, pp. 692–705, Springer, Berlin, Heidelberg, 2010.

[34] K. Sharma, S. Gupta, A. D. Dileep, and R. Rameshan, "Scene image classification using reduced virtual feature representation in sparse framework," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2701–2705, Calgary, AB, Canada, 2018.

[35] A. Shabou and H. Leborgne, "Locality-constrained and spatially regularized coding for scene categorization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625, Providence, RI, USA, 2012.

[36] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 902–917, 2012.

[37] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *European Conference on Computer Vision*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7575 of Lecture Notes in Computer Science, pp. 359–372, Springer, Berlin Heidelberg, 2012.

[38] H. O. Song, R. Girshick, S. Zickler, C. Geyer, P. Felzenszwalb, and T. Darrell, "Generalized sparselet models for real-time multiclass object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1001–1012, 2015.

[39] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3726–3733, Columbus, OH, USA, 2014.

[40] S. Bai, "Growing random forest on deep convolutional neural networks for scene categorization," *Expert Systems with Applications*, vol. 71, pp. 279–287, 2017.

[41] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Information Sciences*, vol. 348, Article ID 2058258, pp. 209–226, 2016.