Research Article

Mining Sufficient Knowledge via Progressive Feature Fusion for Efficient Material Recognition

Hongbin Zhang¹, Qipeng Xiong¹, Haowei Shi¹, Weinan Liang¹, Xiang Zhong¹, Zhiliang Zhu¹, Xiong Li¹ and Donghong Ji²

¹School of Software, East China Jiaotong University, Nanchang 330013, China

²School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Correspondence should be addressed to Hongbin Zhang; zhanghongbin@whu.edu.cn

Received 18 June 2021; Revised 13 September 2021; Accepted 27 September 2021; Published 23 October 2021

Academic Editor: Sikandar Ali

Copyright © 2021 Hongbin Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Material images are susceptible to changes, depending on the light intensity, visual angle, shooting distance, and other conditions. Feature learning has shown great potential for addressing this issue. However, the knowledge achieved using a simple feature fusion method is insufficient to fully represent the material images. In this study, we aimed to exploit the diverse knowledge learned by a novel progressive feature fusion method to improve the recognition performance. To obtain implicit cross-modal knowledge, we perform early feature fusion and capture the cluster canonical correlations among the state-of-the-art (SOTA) heterogeneous squeeze-and-excitation network (SENet) features. A set of more discriminative deep-level visual semantics (DVSs) is obtained. We then perform gene selection-based middle feature fusion to thoroughly exploit the feature-shared knowledge among the generated DVSs. Finally, any type of general classifier can use the feature-shared knowledge to perform the final material recognition. Experimental results on two public datasets (Fabric and MattrSet) showed that our method outperformed other SOTA baseline methods in terms of accuracy and real-time efficiency. Even most traditional classifiers were able to obtain a satisfactory performance using our method, thus demonstrating its high practicality.

1. Introduction

Material recognition [1–5] has become a popular topic in the field of computer vision (CV). It has significant value in many practical scenarios, such as scene recognition [6], industrial inspection [7], medical image recognition [8], and robot vision [9]. For example, self-driving cars or mobile robots need to determine whether an incoming terrain is asphalt, gravel, or grass. Cleaning robots must distinguish between wood, tiles, and carpets. Clearly, material is an important visual cue that is widely present on the surface of natural objects and has abundant visual diversity. However, it is difficult for humans to distinguish these subtle differences. A computer-aided material recognition model can address this issue well and applies to the aforementioned practical fields; however, it is still a great challenge, as material recognition is normally affected by several external

factors, including light intensity, visual angle, shooting distance, and other conditions. Recently, feature learning has shown great potential for alleviating this problem.

Tan et al. [10] used a self-designed scatterometer to measure the bidirectional reflection distribution function (BRDF) of four different textured fabrics and analyzed the effects of surface texture and illuminating wavelengths. On this basis, to accomplish effective material recognition, they proposed a fast but simple BRDF model. Han et al. [11] transformed the original stimulus into a differential excitation domain according to Weber's law, and then they explored a local patch, called microtexton, in the transformed domain as the Weber local descriptor (WLD). Furthermore, the WLD space was characterized adaptively using a generative probability model, which can generate a more discriminative feature representation for material recognition. Tanaka et al. [12] used the depth distortion of

time-of-flight measurements as features and achieved the material classification of a scene. Their method can classify visually similar objects. Erickson et al. [13] proposed a multimodal sensing technique to capture spectral and close-range high-resolution texture information. They combined this high-resolution texture information with the output from a head-mounted camera on a robot to achieve accurate material classification. These methods [10–13] mainly focus on capturing the visual variations in material images. However, obtaining such spectral reflectance is time-consuming and error prone.

Recently, deep learning features have played an important role in material recognition. Shahriari [14] used deep neural networks with variable depths to learn the scales, orientations, and resolutions of texture filter banks for effective material recognition. The corresponding computational cost was highly reduced, and their method could extract very deep features through distributed computing architectures. Jiang et al. [1] also proposed a deep learning-based method that combines both traditional and dilated convolutional features to achieve pixel-level material recognition. In their method, heterogeneous convolutional features are used to remove artifacts. Schwartz Nishino [2] probed the human visual perception of materials. They compared pairs of image patches by asking some questions to complete the weak supervision required to build a set of attribute classifiers. Furthermore, they integrated their method into a convolutional neural network (CNN) that can simultaneously recognize materials and visual attributes. These deep learning methods [1, 2, 14] usually require massive amounts of data to train effective recognition models. Owing to the large cost of annotations, high-quality training samples are very scarce in the field of material recognition.

Moreover, some state-of-the-art (SOTA) feature fusion methods, such as multiset discriminant correlation analysis (MDCA) [15], gene selection eXtreme gradient boosting (GS-XGBoost) [16], gene selection adaptive boosting (GS-AdaBoost) [16], and hierarchical multifeature fusion (HMF²) [17], have been proposed to make full use of the limited training data. The MDCA model uses a single-layer early feature fusion algorithm for material recognition. However, the discriminative information is not balanced between the input and output interfaces of the model. To address this problem, the HMF² model implements multi-layer early feature fusion using residual connections. It modifies the traditional MDCA model [15] and fully mines cross-modal knowledge among the heterogeneous features. More importantly, inspired by the well-known residual structure [18], a hierarchical feature fusion strategy was proposed in HMF² to gather effective knowledge for material recognition. However, the HMF² model requires numerous features and multiple fusion operations, which may slow down its real-time efficiency. Unlike these early feature fusion methods [15, 17], GS-AdaBoost and GS-XGBoost perform middle feature fusion. They both use traditional features, including scale-invariant feature transform (SIFT) [19], Gist [20], local binary pattern (LBP) [21], and deep learning-based VGG16 [22], to complete the material

recognition. However, these two GS models ignore the beneficial early stage cross-modal knowledge among the heterogeneous features.

The abovementioned feature learning approaches [1, 2, 10–17] have shown remarkable progress in material recognition. However, the knowledge achieved by a simple feature fusion method is insufficient to completely represent material images, and few studies have exploited the implicit but valuable knowledge learned through progressive feature fusion. To address these issues, we propose a novel but effective progressive feature fusion method called SECGS that can seamlessly combine early and middle feature fusions together for material recognition. In our model, “SE” represents the well-known squeeze-and-excitation network (SENet) [23]. This means that we adopt only the SOTA SENet features for progressive feature fusion. “C” indicates that the SECGS model first mines the implicit cross-modal knowledge through cluster canonical correlation analysis [24]. It is a very important early feature fusion operation, as it not only creates a group of deep-level visual semantics (DVSS) but also serves as a basis for the subsequent middle feature fusion. “GS” indicates that we further capture the feature-shared knowledge among the implicit cross-modal knowledge using gene selection. Finally, deep-level but sufficient knowledge is mined for training material recognition model.

The main contributions of this study are summarized as follows:

- (1) We proposed a novel progressive feature fusion approach that can mine sufficient and effective knowledge among the heterogeneous SENet features for material recognition.
- (2) We conducted extensive experiments on two benchmark datasets. The results showed the superior performance of our model over other SOTA methods. With our approach, even traditional classifiers were able to achieve satisfactory performance. The code for our method is available at <https://github.com/Danicaghost/SECGS>.
- (3) We created multiple variants of the SECGS model. They are all effective for material recognition, which further validates the robustness of the proposed progressive feature fusion method.
- (4) We developed a real-time material recognition system based on our model and deployed it on a laptop. All we need is a mobile phone with a camera to complete the material recognition. This helps us to narrow the gap between theoretical research and practical applications. A demo video has also been uploaded to the above GitHub link.

2. Related Works

2.1. Material Recognition. Material recognition is an important research direction in the field of CV. Some researchers have used spectral reflectance from material surfaces to characterize materials [10, 25–29]. However, the

acquisition of spectral reflection is time-consuming and error prone, which may reduce the practicality of the material recognition model. Therefore, some researchers have focused on constructing local feature descriptors (such as WLD) [30], on using reflection-based edge features [31], or on adopting some CNN features. Feature learning helps in improving the accuracy of material recognition. For example, Cimpoi et al. [32] calculated the CNN-based Fisher vector (FV) [33] features and achieved the best performance on the well-known Flickr material database [34]. Bell et al. [35] combined a set of pretrained CNNs using a fully connected conditional random field model to predict the material label of each pixel. Schwartz and Nishino [2] introduced a material attribute-category CNN that considers the implicit attribute information for material recognition. Bruna Mallat [36] proposed scattering convolution networks (ScaNet) that take Gabor or Haar wavelets as filters for material recognition. Sifre Mallat [37] proposed a rotation- and scale-invariant feature extraction method for robust material classification. Inspired by ScaNet, Chan et al. [38] utilized local image patches as principal component analysis filters to complete the material recognition. The aforementioned studies [2, 32–38] mainly adopted a certain layer of the CNN model, without considering the implicit complementary information among the heterogeneous layers from a homologous neural network. This motivated us to mine the complementary knowledge among the heterogeneous SENet features, which helps improve the final recognition performance. Table 1 lists the recent major material recognition methods, which we roughly divided into four categories: “traditional feature”: traditional methods using mathematical calculations; “CNN-based feature”: methods using deep learning features; “ETE”: end-to-end methods; and “feature fusion”: feature fusion methods (see the next subsection to learn more about “feature fusion”).

2.2. Feature Fusion. As shown in Table 1, feature fusion has shown great potential for material recognition. Currently, feature fusion methods can be divided into early, middle, and late fusions. Early feature fusion performs fusion at the raw feature level. For example, a group of heterogeneous features is first mapped into a semantic space. Then, early feature fusion is performed by feature summation or feature concatenation. Representative early feature fusion models include canonical correlation analysis [39], deep canonical correlation analysis [40], multiple canonical correlation analysis [41], cross-modal correlation learning [42], kernel discriminant correlation analysis [43], and consistent discriminant correlation analysis [44]. Early feature fusion is intended to mine low-level correlations among different features. Unlike early feature fusion, middle feature fusion builds a more powerful classification model using a group of linearly weighted features. Representative middle feature fusion models include multiple kernel learning [45], efficient range gene selection [16], and multiple kernel boosting [46]. Middle feature fusion attempts to mine the implicit feature-shared knowledge among the heterogeneous features. Finally, late feature fusion integrates the multiple predictions

of a set of weak classifiers to build a more powerful classification model. Representative late feature fusion methods include AdaBoost [47], XGBoost [48], categorical boosting (CatBoost) [49], and light gradient boosting machine (LightGBM) [50]. Unlike early and middle feature fusions, late feature fusion takes full advantage of the complementarity among different classifiers.

Recently, an increasing number of feature fusion methods have played an important role in material recognition (Table 1). The MDCA model uses a single-layer early feature fusion method for material recognition [15]. However, the effective discriminative information is not balanced between the input and output interfaces of the MDCA model. To resolve this problem, the HMF² model implements multilayer early feature fusion using residual connections [17]. It can fully mine the cross-modal knowledge among a group of heterogeneous image features. However, the HMF² model requires numerous features and multiple fusion operations, which may slow down its real-time efficiency. Moreover, it ignores the feature-shared knowledge and, thus, may miss some important material knowledge for effective material recognition. Unlike the early feature fusion approaches, the GS-AdaBoost and GS-XGBoost models perform middle feature fusion using SIFT, Gist, LBP, and VGG16 to complete the material recognition [16]. They can fully capture feature-shared knowledge to improve the performance of the final recognition. However, these two GS models ignore the early stage correlations among the heterogeneous features. In other words, the aforementioned feature fusion methods usually employ a single and independent feature fusion strategy. Moreover, only a few studies have explored the type of progressive relationship between the different feature fusion approaches for more effective material recognition.

In summary, a novel progressive feature fusion framework that can make full use of different fusion methods is required to resolve the abovementioned issue. The progressive feature fusion method can mine sufficient and effective knowledge among the heterogeneous SENet features for high-quality material recognition.

3. Our Approach

The overall scheme of the proposed progressive feature fusion approach is illustrated in Figure 1. In this section, we first present the heterogeneous SENet features. Five types of heterogeneous SENet features are used in our approach. Based on the SENet features, we then propose a progressive feature fusion approach. First, we mine the cross-modal knowledge among the heterogeneous SENet features using early feature fusion and generate a group of DVs with more discriminant ability. With the use of permutation and combination, a total of 10 different DVs are then generated. The 10 DVs mainly extract the more effective discriminative information from the heterogeneous SENet features. The DVs are more powerful than the original SENet features. Then, we mine the implicit feature-shared knowledge progressively among these DVs using middle feature

TABLE 1: Summary of recent material recognition models. Unit: %.

Method	Category	Dataset	Accuracy
WLD [30]	Traditional feature	Brodatz dataset	97.50
Alda [31]		KTH-TIPS2a dataset	64.70
PCANet [38]		Flickr material dataset	44.60
		MNIST	98.56
		CURET dataset	99.61
FC-CNN + FV-CNN [32]	CNN-based feature	Flickr material dataset	82.40
FC-CNN + FV-CNN [32]		Describable texture dataset	74.70
FV-CNN [32]		KTH-TIPS2b dataset	81.80
IFV + VGG [14]		KTH-TIPS2b dataset	81.50
IFV + DFB [14]		KTH-TIPS2a dataset	88.60
IFV + DFB [14]		Flickr material dataset	82.70
IFV + DFB [14]		Describable texture dataset	80.60
Ensemble 3 CNNs [35]	ETE	The materials in context database	83.50
ScaNet [36]		USPS dataset	97.70
SRP + multiscale train [37]		MNIST dataset	95.60
SRP + multiscale train [37]		KTH-TIPS dataset	99.40
SDC + CRF [1]		UMD dataset	99.70
		MINC dataset	74.27
GS-AdaBoost [16]	Feature fusion	Fabric dataset	79.52
		MattrSet dataset	66.11
GS-XGBoost [16]		Fabric dataset	84.20
		MattrSet dataset	67.67
HMF ² [17]		Fabric dataset	84.34
		MattrSet dataset	72.20
MDCA [17]		Fabric dataset	49.84
		MattrSet dataset	65.36

FIGURE 1: Overview of the proposed model.

fusion. The feature-shared knowledge is used to train our material recognition model. “Feature-shared” means that discriminative information is complementary between different features. Hence, in this stage, feature-shared knowledge among different DVSs is shared with all features to complete the middle feature fusion and obtain better recognition performance. Finally, the training procedure is described below. In addition, in this study, we used two benchmark datasets, namely, the Fabric and the MattrSet, to

complete all experiments. The Fabric dataset [51] is a fine-grained dataset and includes images of clothes made from any of these 9 materials: cotton, wool, terrycloth, fleece, nylon, silk, denim, viscose, and polyester. It contains 5064 images. The MattrSet dataset [16] is a coarse-grained dataset that includes images of bags (polyurethane (pu), canvas, nylon, and polyester) and shoes (pu and canvas). It contains 11,021 images (see Tables 2 and 3 for more details about these two datasets).

TABLE 2: The fabric dataset.

Category	Image	Number
Cotton		2352
Wool		360
Terrycloth		120
Fleece		132
Nylon		200
Silk		200
Denim		648
Viscose		148
Polyester		904

TABLE 3: The MattrSet dataset.

Category	Image	Number
Bag_pu		1982
Bag_canvas		1948
Bag_nylon		1764
Bag_polyester		1715
Shoes_pu		1757
Shoes_canvas		1855

3.1. Heterogeneous SENet Features. As discussed above, deep learning features [1, 2, 14] can better characterize images of materials. Current mainstream CNN models usually improve performance by using a spatial dimensional layer. Unlike the models in [52, 53], the SENet model [23] can adaptively recalibrate channel-wise feature responses by explicitly modeling the interdependencies between different channels. Moreover, it pays more attention to the features that contain the most effective information. Hence, the SENet features outperform the ResNet features in most CV tasks. Therefore, we chose SENet rather than ResNet as the basis for the subsequent progressive feature fusion (the results of extensive experiments also validated our selection, see Table 4). In this study, we used five types of heterogeneous layers of SENet, namely, SE-ResNet 50 (SR50), SE-ResNet 101 (SR101), SE-ResNet 152 (SR152), SE-ResNeXt 50 (SRXt50), and SE-

ResNeXt 101 (SRXt101). They contain abundant but complementary semantic information. Therefore, SENet features can achieve better recognition performance. This means that these heterogeneous features point to the same or similar material knowledge. Considering these factors, we attempt to make full use of the complementary knowledge among the SENet features using the progressive feature fusion approach.

3.2. Progressive Feature Fusion. Our progressive feature fusion approach is illustrated in Figure 2. The heterogeneous SENet features point to the same or similar material knowledge. Inspired by MDCA [15] and HMF² [17], we first attempt to obtain the implicit cross-modal knowledge from these SENet features using early feature fusion. We extract the average pooling layer (2048 dimensions) of the five heterogeneous SE modules introduced above. Then, we analyze the cluster canonical correlation among all the SENet features and generate 10 DVs. These DVs fully mine the effective discriminative information among the heterogeneous SENet features. This means that the semantic distance between the samples in the same class shrinks, whereas the corresponding semantic distance between the samples in different classes expands gradually. Hence, the classifier can better distinguish different material categories and obtain a better recognition performance (see Section 5.4).

The material dataset is defined as $D = \{(x_n, y_n), x_n \in R^k, y_n \in C, n = 1, 2, \dots, N\}$. Here, N is the number of samples. y_n is the material label of the sample x_n . $C = \{c_1, c_2, \dots, c_l\}$ is the set of all material labels, $c_r \in C$, and l is the number of material categories. X and Y are the two heterogeneous SENet features. We divide the two SENet features into “ X_1^* , Y_1^* ” (for training) and “ X_2^* , Y_2^* ” (for testing).

X and Y have classes c_l , $T_x = \{X_{c_1}, X_{c_2}, \dots, X_{c_l}\}$, and $T_y = \{Y_{c_1}, Y_{c_2}, \dots, Y_{c_l}\}$, where $X_{c_l} = \{X_{c_l}^{c_1}, X_{c_l}^{c_2}, \dots, X_{c_l}^{c_l}\}$ and $Y_{c_l} = \{Y_{c_l}^{c_1}, Y_{c_l}^{c_2}, \dots, Y_{c_l}^{c_l}\}$ represent the data of X_1^* and Y_1^* in class c_r , respectively. ω and ν are the projection vectors of X_1^* and Y_1^* , respectively. The correlation coefficient ρ between X and Y is computed as follows:

$$\rho = \max_{\omega, \nu} \frac{\omega' C_{XY} \nu}{\sqrt{\omega' C_{XX} \omega} \sqrt{\nu' C_{YY} \nu}}. \quad (1)$$

The covariance matrices C_{XY} , C_{XX} , and C_{YY} are, respectively, defined as follows:

$$C_{XY} = \frac{1}{S} \sum_{l=1}^l \sum_{i=1}^{|X_{c_l}|} \sum_{j=1}^{|Y_{c_l}|} x_i^{c_l} y_j^{c_l}, \quad (2)$$

$$C_{XX} = \frac{1}{S} \sum_{l=1}^l \sum_{i=1}^{|X_{c_l}|} |Y_{c_l}| x_i^{c_l} x_i^{c_l}, \quad (3)$$

$$C_{YY} = \frac{1}{S} \sum_{l=1}^l \sum_{j=1}^{|Y_{c_l}|} |X_{c_l}| y_j^{c_l} y_j^{c_l}, \quad (4)$$

TABLE 4: Comparisons of the different types of DVSs on each dataset. The best value of each column is shown in bold. Unit: %.

Dataset	Feature	AA _F					
		KNN	LR	SVM	Ada	XGB	MAA
MatrSet	DVS (re)	55.67	60.71	19.58	60.46	61.99	51.68
	DVS (re-se)	64.58	67.64	68.50	60.31	66.00	65.41
	DVS (se)	65.45	68.70	69.80	67.42	68.63	68.00
	$\Delta\text{Improve}_1$	7.53	4.19	-5.31	16.03	10.66	5.68
	$\Delta\text{Improve}_2$	0.32	0.05	-0.20	-2.72	0.27	-0.05
	$\Delta\text{Improve}_3$	1.19	1.11	1.10	4.39	2.90	2.28
Fabric	DVS (re)	42.36	56.12	46.25	48.19	56.78	57.45
	DVS (re-se)	86.21	87.01	78.42	70.68	74.11	79.29
	DVS (se)	88.97	90.97	85.36	87.71	87.57	88.12
	$\Delta\text{Improve}_1$	3.98	4.28	0.00	0.04	7.44	3.15
	$\Delta\text{Improve}_2$	0.03	-0.36	-2.33	-15.44	-11.85	-5.99
	$\Delta\text{Improve}_3$	2.79	3.60	4.61	1.59	1.61	2.84

FIGURE 2: Overview of the progressive feature fusion approach.

where $S = \sum_1^l |X_{c_i}| |Y_{c_i}|$ is the total logarithm of the paired relationship between X_1^* and Y_1^* .

According to ω and ν , X_1^* , X_2^* , Y_1^* , and Y_2^* are transformed into a set of mapped feature matrices $\tilde{X}_{\alpha_1}^* = \omega' X_1^*$, $X_{\alpha_2}^* = \omega' X_2^*$, $Y_{\alpha_1}^* = \nu' Y_1^*$, and $Y_{\alpha_2}^* = \nu' Y_2^*$, respectively. Then, U and V are constructed according to the mapped feature matrices, as follows:

$$U = \begin{pmatrix} X_{\alpha_1}^* \\ X_{\alpha_2}^* \end{pmatrix}, \quad (5)$$

$$V = \begin{pmatrix} Y_{\alpha_1}^* \\ Y_{\alpha_2}^* \end{pmatrix}.$$

Finally, we fuse U and V to generate 10 DVSs for performing middle feature fusion progressively. Unlike in HMF², we perform early feature fusion only once and only adopt the SENet features rather than numerous different features:

$$\text{DVS} = \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} X_{\alpha_1}^* & Y_{\alpha_1}^* \\ X_{\alpha_2}^* & Y_{\alpha_2}^* \end{pmatrix}, \quad (6)$$

$$\text{or DVS} = U + V$$

$$= \begin{pmatrix} X_{\alpha_1}^* + Y_{\alpha_1}^* \\ X_{\alpha_2}^* + Y_{\alpha_2}^* \end{pmatrix}.$$

As shown in Figure 2, to obtain the feature-shared knowledge from the DVSs, we perform middle feature fusion through gene selection. First, the DVSs are combined to form a group of DVS combinations. Second, we employ a general classifier to predict the estimated probability of each DVS in any DVS combination. Then, the corresponding gene selection weight of each DVS is calculated, and a sum function is used to fuse the weighted estimated probabilities. Finally, we use a max decision function to generate the final recognition result.

$DVS = \{DVS_1, DVS_2, \dots, DVS_z\}$ is the set of DVSs, $DVS_i \in DVS$. z is the number of DVSs. The effective range $ER_{dr} = [r_{dr}^-, r_{dr}^+]$ of DVS_i on the labeled samples c_r is defined as follows:

$$[r_{dr}^-, r_{dr}^+] = [\mu_{dr} - (1 - p_r)\gamma\sigma_{dr}, \mu_{dr} + (1 - p_r)\gamma\sigma_{dr}], \quad (7)$$

where r_{dr}^+ and r_{dr}^- represent the upper and lower boundaries, respectively, of the effective range of DVS_i for the samples c_r . μ_{dr} and σ_{dr} represent the mean value and standard deviation of DVS_i , respectively, for the samples c_r . p_r ($0 < p_r < 1$) is the estimated probability of the samples c_r , and $(1 - p_r)$ reduces the impact of the standard deviation σ_{dr} on the upper and lower boundaries of the effective range. γ is a constant derived from Chebyshev's inequality, which is equal to 1.732.

We compute the overlap area OL_d and the corresponding coefficient OC_d of DVS_i . Hence, the gene selection weight w_d of DVS_i is calculated according to OC_d . The probability p_{dk} is weighted by this weight:

$$OL_d = \sum_{k=1}^{l-1} \sum_{h=k+1}^l \varphi_d(k, r), \quad (8)$$

$$\varphi_d(k, r) = \begin{cases} r_{dr}^+ - r_{dr}^-, & r_{dr}^+ > r_{dr}^-, \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

$$OC_d = \frac{OL_d}{\max_r(r_{dr}^+) - \min_r(r_{dr}^-)}, \quad (10)$$

$$w_d = \frac{OC_d}{\sum\{OC_k: 1 \leq k \leq z\}}. \quad (11)$$

Lastly, we use a max decision function to generate the final recognition result. The material label is determined according to $\max(p_d) = \max\{\sum(p_{dk} * w_d)\}$. In summary, the implicit complementary knowledge among the heterogeneous SENet features is fully mined by the proposed progressive feature fusion method. The SECGS model is shown in Algorithm 1.

3.3. Training Procedure. Owing to its generality, the proposed SECGS model can use any type of general classifier to complete the final prediction. The proposed model was trained using four-fold cross-validation. In addition, the trained SECGS model was deployed on a laptop with the following configuration: Intel Core i7-7700 CPU (2.8 GHz), 16 GB of RAM, and GTX 1050 GPU. Hence, online real-time material recognition can be implemented easily on a laptop with a similar or higher configuration.

4. Experiments and Results

In this section, we evaluate the effectiveness, efficiency, and robustness of our approach on two benchmark datasets: (1) the Fabric and (2) the MattrSet datasets.

4.1. Datasets. The Fabric dataset is a fine-grained dataset constructed by Kampouris et al. [51] using photometric

stereoscopic sensors to collect clothing samples from physical stores and determine the final material according to the clothing label. It contains images of clothes made from any of these 9 materials: cotton, wool, terrycloth, fleece, nylon, silk, denim, viscose, and polyester. The Fabric dataset contains 1266 samples, according to different light intensities, visual angles, and other conditions. Each sample includes four images; thus, it has a total of 5064 images (Kampouris et al. cropped all the images to 400×400 pixels to avoid blurring the edges of the images). Table 2 provides more detailed information about the Fabric dataset.

The MattrSet dataset [16] is derived from real commodities on the web and constructed under the guidance of an experienced material expert. It is a coarse-grained dataset that includes images of bags (polyurethane (pu), canvas, nylon, and polyester) and shoes (pu and canvas). It contains 11,021 images (the corresponding image resolution ranges from 123×123 to 4300×2867 , which makes material recognition more challenging). Table 3 provides more detailed information about the MattrSet dataset.

4.2. Baselines. We used five general classifiers, namely, logistic regression (LR), k -nearest neighbor (KNN), support vector machine (SVM), AdaBoost (Ada), and XGBoost (XGB), to train the proposed SECGS model (e.g., "SECGS_{5_LR}" means that the LR classifier uses five DVSs to implement middle feature fusion). We compared our model with the baseline models listed below. The goal of this study was to obtain a better recognition performance using progressive feature fusion. We aimed to improve the recognition performance of our model on fine-grained and coarse-grained material images. Hence, our work can be regarded as a type of fine-grained image classification task. Therefore, we also compared the proposed model with some recently proposed fine-grained image classification models:

- (1) Some well-known late feature fusion models are as follows: AdaBoost [47], XGBoost [48], CatBoost [49], and LightGBM [50]. For a fairer comparison, we also used the DVSs generated by our method to train these models.
- (2) Recently proposed material recognition models are as follows: GS-XGB [16], GS-Ada [16], SVM-S [17], MDCA-XGB [17], HMF² [17], and Kampouris's Albe + Norm [53].
- (3) Recently proposed deep learning models are as follows: InceptionResNetV2 [54], DenseNet169 [55], SENet [23], and ResNeXt WSL [56].
- (4) Some well-known fine-grained classification models are as follows: DCL [57], NTS-Net [58], PMG [59], WILDCAT [60], WSCNet [61], and WS-DAN [62]. All these models are trained by using the four-fold cross-validation with an initial learning rate of $1e-3$, a batch size of 16 and trained for 40 epochs.
- (5) Multiple variants of the SECGS model are as follows: we removed the "GS" component from "SECGS" to generate a new early feature fusion model called SEC. We then added an ensemble learning strategy (a type

Input: image feature set: $S_1 = \{SR50, SR101, SR152, SRXt50, SRXt101\}$

Output: material recognition results

(1) **Repeat**

- (i) Select two features from S_1 .
- (ii) Divide the two features into the training and testing data.
- (iii) Use the training data to perform a cluster canonical correlation analysis and to obtain the projection vectors ω and ν .
- (iv) Use ω and ν to weight the training and testing data, respectively; therefore, U and V are constructed using equation (5).
- (v) Use equation (6) to obtain the corresponding DVSs according to the two features.

Until any pair of features in S_1 has been traversed.

(2) Obtain a DVS set: $S_2 = \{DVS_1, DVS_2, \dots, DVS_{10}\}$

(3) **For** $j = 2 : 10$

Build a DVS combination set S_3 by randomly choosing j DVSs from S_2 .

Repeat

- (i) Use a general classifier to calculate the estimated probabilities of the j DVSs.
- (ii) Use the j DVSs and the corresponding estimated probabilities to obtain the gene selection weight w_d .
- (iii) Use $\max(p_d) = \max\{\sum(p_{dk} * w_d)\}$ to generate the final material recognition result.

Until S_3 is null.

$j = j + 1$

(4) **End For**

(5) Output the final material recognition results.

ALGORITHM 1: The progressive feature fusion algorithm: SECGS.

of late feature fusion) to the SEC model and created two variants of our model: SEC(s) and SEC(h). SEC(s) uses soft voting-based late feature fusion after the early feature fusion, whereas SEC(h) uses hard voting-based late feature fusion. Hence, the two models also use the progressive feature fusion method. We wanted to validate the robustness of this method and to demonstrate the effectiveness of the “GS” component.

4.3. Evaluation Metrics. We employed accuracy (Acc.) (equation (12)) and precision (equation (13)) to evaluate our model. TP indicates true positive; FP, false positive; TN, true negative; and FN, false negative. Moreover, we designed a new metric called the average accuracy for each feature (AA_F) (equation (14)) to comprehensively evaluate each feature. Similarly, we propose a new metric called the average accuracy for each classifier (AA_C) (equation (15)) to comprehensively evaluate each classification model. Finally, we used the mean average accuracy (MAA) metric (equation (16)) to evaluate the overall performance of the proposed model:

$$\text{accuracy} = \frac{TP_{\text{all}} + TN_{\text{all}}}{TP_{\text{all}} + FP_{\text{all}} + TN_{\text{all}} + FN_{\text{all}}}, \quad (12)$$

$$\text{precision} = \frac{TP_{\text{all}}}{TP_{\text{all}} + FP_{\text{all}}}, \quad (13)$$

$$AA_F = \frac{\sum_{m=1}^{N_{\text{classifier}}} (\text{Accuracy}_m)}{N_{\text{classifier}}}, \quad (14)$$

$$AA_C = \frac{\sum_{n=1}^{N_{\text{feature}}} (\text{Accuracy}_n)}{N_{\text{feature}}}, \quad (15)$$

$$MAA = \frac{\sum_{n=1}^{N_{\text{classifier}}} \sum_{m=1}^{N_{\text{feature}}} (\text{Accuracy}_{(m)(n)})}{N_{\text{classifier}} \times N_{\text{feature}}}. \quad (16)$$

Moreover, as shown in equation (17), $\Delta\text{Accuracy}_1$ calculates the absolute performance improvement of the AA_F metric of the SENet features relative to that of the ResNet features. This metric can evaluate the effectiveness of the SENet features. As shown in equation (18), $\Delta\text{Accuracy}_2$ calculates the absolute performance improvement of the AA_F metric of the DVSs relative to that of the SENet features. This metric evaluates the effectiveness of the generated DVSs.

$$\Delta\text{Accuracy}_1 = AA_{F\text{-SENet}} - AA_{F\text{-ResNet}}, \quad (17)$$

$$\Delta\text{Accuracy}_2 = AA_{F\text{-DVS}} - AA_{F\text{-SENet}}. \quad (18)$$

4.4. Evaluation of the SEC Model. Table 5 shows the accuracy of the SEC model that uses different DVSs. Table 6 shows the AA_F and MAA values of the best ResNet features, SENet features, and DVSs. In this section, we discuss our experimental results in detail.

According to $\Delta\text{Accuracy}_1$, the maximum improvement of the SENet features relative to that of the ResNet features on the Fabric dataset is 47.79%, and the corresponding average improvement is 38.49%. These results confirm that the SE module can depict the key visual characteristics of different materials and that the SENet features from the SE

TABLE 5: Recognition performance of the proposed SEC models on each dataset. The best value of each column is shown in bold. Unit: %.

Dataset	DVS	Recognition performance of each SEC model					
		KNN	LR	SVM	Ada	XGB	Average
MatrSet	SR50-SR152	66.23	68.94	69.97	67.26	68.68	68.22
	SR50-SR101	65.48	68.72	69.70	67.49	68.40	67.96
	SR50-SRXt101	64.91	68.88	70.30	67.46	69.02	68.11
	SRXt101-SR152	65.15	68.73	69.61	67.79	68.89	68.03
	SR101-SR152	65.44	68.71	69.93	67.22	68.29	67.92
	SRXt50-SR152	65.47	68.22	69.29	67.30	68.51	67.76
	SR101-SRXt101	64.74	68.70	69.82	67.44	68.38	67.82
	SR50-SRXt50	64.25	68.44	69.63	67.59	68.82	67.75
	SRXt50-SRXt101	64.68	68.59	70.07	67.61	68.29	67.85
	SRXt50-SR101	64.49	68.63	69.47	67.52	68.24	67.67
	$\Delta\text{Accuracy}_1$	16.12	11.07	43.81	18.60	14.40	20.80
	$\Delta\text{Accuracy}_2$	0.83	1.07	1.08	4.44	2.82	2.05
Fabric	SR101-SRXt101	89.08	91.13	86.30	87.46	87.16	88.23
	SRXt50-SR101	89.00	90.98	86.02	87.32	88.11	88.29
	SRXt101-SR152	89.44	91.05	85.07	86.95	86.85	87.87
	SRXt50-SRXt101	88.74	90.60	85.68	87.91	87.56	88.10
	SRXt50-SR152	89.55	90.80	85.76	88.19	87.54	88.37
	SR50-SRXt101	87.99	91.23	83.31	88.45	88.19	87.83
	SR50-SRXt50	87.60	90.86	83.08	87.93	88.11	87.52
	SR101-SR152	88.55	90.28	84.56	86.65	86.04	87.22
	SR50-SR101	87.22	90.40	82.42	87.62	88.31	87.19
	SR50-SR152	87.97	89.93	81.38	87.72	87.97	86.99
	$\Delta\text{Accuracy}_1$	47.79	35.54	34.50	37.98	36.62	38.49
	$\Delta\text{Accuracy}_2$	2.34	3.35	3.61	1.50	1.62	2.48

TABLE 6: AA_F and MAA values of each feature on each dataset. In MAA values, the best performance of each column is shown in bold. In AA_F values, the best value of each row is shown in bold. Unit: %.

Dataset	Feature	AA_F					MAA
		KNN	LR	SVM	Ada	XGB	
MatrSet	ResNet	48.14	56.52	24.89	44.43	51.33	45.06
	SENet	64.26	67.59	68.70	63.03	65.73	65.86
	DVS	65.45	68.70	69.80	67.42	68.63	68.00
Fabric	ResNet	38.38	51.84	46.25	48.15	49.34	46.79
	SENet	86.18	87.37	80.75	86.12	85.96	85.28
	DVS	88.97	90.97	85.36	87.71	87.57	88.12

modules can be applied to different classification models. Similar results were also obtained on the MatrSet dataset, which confirms the robustness of the SENet features. Certainly, a larger performance improvement was obtained on the Fabric dataset owing to the fine-grained visual information contained in its samples.

According to $\Delta\text{Accuracy}_2$, the maximum improvement of the DVSs relative to that of the SENet features on the Fabric dataset is 3.35%, and the corresponding average improvement is 2.48%. Similar performance improvements were also obtained on the MatrSet dataset. These results indicate that the proposed DVSs are effective for material recognition on any dataset and that the proposed early feature fusion method is effective and robust. Moreover, the DVS (SR50-SRXt101) achieved the best accuracy on each dataset. This indicates that sufficient but effective cross-modal knowledge can be mined among the heterogeneous SR50 and SRXt101 features. Owing to this characteristic, even traditional classifiers, including LR and SVM, were able

to obtain satisfactory recognition performance. It is worth noting that the corresponding performance on the Fabric dataset was better than that on the MatrSet dataset because the former dataset contains much more valuable visual semantics, whereas the latter dataset contains certain noise (Tables 2 and 3).

According to AA_F and MAA in Table 6, we can obtain the overall recognition performance of different features from a statistical perspective. Therefore, many more meaningful conclusions can be drawn. According to AA_F , the proposed DVSs easily outperformed the SENet and ResNet features on each dataset. A larger performance improvement was obtained on the Fabric dataset. Hence, the proposed DVSs are robust and effective for material recognition. The MAA metric further validates this conclusion. For example, on the Fabric dataset, a performance margin of 41.33% was obtained between DVS and ResNet. These results clearly show that there is indeed a large amount of complementary information among the heterogeneous

SENet features and that the proposed early feature fusion method can capture such information to improve the recognition performance. In addition, even traditional models, such as LR, KNN, and SVM, were able to achieve better performance after using early feature fusion owing to the powerful discriminative ability of the proposed DVSSs.

In summary, the DVSSs generated by the proposed early feature fusion method are effective and robust for material recognition, providing a solid basis for the subsequent middle feature fusion. Sufficient cross-modal knowledge among the heterogeneous SENet features was mined to improve the recognition performance. Moreover, all the SENet features obtained satisfactory performance, which firmly supports our original selection.

4.5. Evaluation of the SEC(s) and SEC(h) Models. In this section, we evaluate two variants of the SECGS model, namely, SEC(h) and SEC(s). We want to determine whether the voting strategy contributes to improving recognition performance. More importantly, we intend to validate the robustness of the proposed progressive feature fusion method. In addition, we want to demonstrate the effectiveness of the “GS” component from another perspective. Figure 3 shows the AA_C values of the top three DVSSs combined with the corresponding voting strategy.

As shown in Figure 3(a), the SEC(h) model that uses SR101-SRXt101 with the top three classifiers achieved the best accuracy (91.51%) on the Fabric dataset. It outperformed the original DVS without voting by a large margin. Thus, we should carefully tune the number of classifiers for the final voting. Similarly, as shown in Figure 3(b), the SEC(s) model that uses SRXt101-SR152 with the top two classifiers obtained the best accuracy (93.19%) on the Fabric dataset. It also outperformed the original DVS without voting and the best SEC(h) model. Therefore, the soft voting strategy is more suited for the Fabric dataset.

As shown in Figure 3(c), the SEC(h) model that uses SR50-SRXt101 with all the classifiers obtained the best accuracy on the MattrSet dataset. It easily outperformed the original DVS without voting. Similarly, as shown in Figure 3(d), the SEC(s) model that uses SR50-SRXt101 with all the classifiers achieved the best accuracy. Unlike on the Fabric dataset, we require more classifiers to complete the voting-based late feature fusion on the MattrSet dataset owing to the much larger amount of noise information in its samples. More heterogeneous classifiers can make a more definite prediction. In addition, a larger performance improvement can be obtained on the MattrSet dataset if we use a soft voting strategy. This further indicates that the corresponding recognition of coarse-grained material images remains a significant challenge. In general, the soft voting strategy is also more suited for the MattrSet dataset.

In summary, both the SEC(h) and SEC(s) models achieved satisfactory performance on each material dataset. This also validates the robustness of the proposed progressive feature fusion method from another perspective. Hence, the voting-based late feature fusion method is another useful substitute for the proposed middle feature

fusion method. Unlike HMF² [17], the SEC model only adopts the SENet features, which helps reduce its complexity. This is an important step in narrowing the gap between theoretical research and practical applications.

4.6. Evaluation of the SECGS Model. In this section, we perform a more comprehensive evaluation of the proposed SECGS model. Table 7 shows the corresponding experimental results. We show the performance variations with the increase in the number of DVSSs on each dataset. For example, “2-SVM” means that the SVM classifier uses two DVSSs to implement the middle feature fusion. “Best Acc.” represents the best accuracy among all the DVS combinations. “ Δ_{max} ” represents the corresponding performance improvement of the SECGS model compared with the best SEC model described above. Some important conclusions can be drawn from Table 7.

- (1) On the MattrSet dataset, the corresponding performance rank of the different classifiers in descending order was “SVM > XGBoost > LR > AdaBoost > KNN.” On the Fabric dataset, it was “LR > KNN > XGBoost > AdaBoost > SVM.” These can help us with the selection of the optimal classifier for the SECGS model. The SECGS model prefers SVM on the coarse-grained MattrSet dataset, whereas it prefers LR on the fine-grained Fabric dataset. A similar result can be seen in Table 5. Thus, owing to the progressive feature fusion method, sufficient and effective knowledge was mined, enabling even the traditional classifiers to achieve satisfactory recognition performance.
- (2) We found that, regardless of the number of DVSSs, the SVM classifier always achieved the best performance on the MattrSet dataset and the LR classifier on the Fabric dataset. These findings are further illustrated in Figure 4. They indicate that the most suitable classifier should be chosen carefully for different material datasets to achieve the largest performance improvement.
- (3) According to Δ_{max} , the proposed middle feature fusion approach outperformed the voting-based late feature fusion approach. The feature-shared knowledge was fully mined using our progressive feature fusion model. Evidently, the proposed “GS” method is more effective than the voting-based late feature fusion method. Unlike the SEC model, our model adopts only one classifier. In addition, larger performance improvements were obtained on the fine-grained Fabric dataset. A similar conclusion can be drawn from the above subsection “Evaluation of the SEC Model.” Interestingly, owing to its different implementation mechanisms, the proposed SECGS model will be a powerful supplement to the well-known HMF² model (Table 8).
- (4) More DVSSs did not mean better recognition performance. For example, the proposed SECGS model that combines the SVM classifier and three DVSSs

FIGURE 3: Performance comparisons between the SEC(h) and SEC(s) models. (a) Fabric: SEC(h), (b) Fabric: SEC(s), (c) MatrrSet: SEC(h), and (d) MatrrSet: SEC(s).

TABLE 7: Recognition performance of the SEC GS model on each dataset. The best value of each column is shown in bold. Unit: %.

Dataset	SEC GS	BestAcc.	SEC GS	BestAcc.	SEC GS	BestAcc.	SEC GS	BestAcc.	SEC GS	BestAcc.
MatrrSet	2-KNN	68.89	2-LR	70.37	2-SVM	71.25	2-Ada	69.52	2-XGB	70.56
	3-KNN	69.56	3-LR	70.83	3-SVM	71.42	3-Ada	69.93	3-XGB	70.92
	4-KNN	69.98	4-LR	70.86	4-SVM	71.33	4-Ada	69.92	4-XGB	71.00
	5-KNN	69.89	5-LR	70.86	5-SVM	71.37	5-Ada	69.94	5-XGB	70.99
	6-KNN	69.88	6-LR	70.90	6-SVM	71.39	6-Ada	70.08	6-XGB	71.13
	7-KNN	69.90	7-LR	70.86	7-SVM	71.38	7-Ada	70.04	7-XGB	71.04
	8-KNN	69.95	8-LR	70.85	8-SVM	71.28	8-Ada	70.00	8-XGB	71.06
	9-KNN	69.86	9-LR	70.77	9-SVM	71.20	9-Ada	69.90	9-XGB	70.92
	10-KNN	69.61	10-LR	70.59	10-SVM	71.00	10-Ada	69.80	10-XGB	70.88
	$\Delta_{\max}$	3.75	$\Delta_{\max}$	1.96	$\Delta_{\max}$	1.12	$\Delta_{\max}$	2.29	$\Delta_{\max}$	2.11

TABLE 7: Continued.

Dataset	SECGS	BestAcc.	SECGS	BestAcc.	SECGS	BestAcc.	SECGS	BestAcc.	SECGS	BestAcc.
Fabric	2-KNN	92.48	2-LR	93.76	2-SVM	88.65	2-Ada	89.83	2-XGB	89.81
	3-KNN	92.91	3-LR	94.30	3-SVM	88.90	3-Ada	89.95	3-XGB	90.21
	4-KNN	93.23	4-LR	94.41	4-SVM	88.80	4-Ada	90.09	4-XGB	90.25
	5-KNN	93.42	5-LR	94.71	5-SVM	88.69	5-Ada	90.03	5-XGB	90.44
	6-KNN	93.42	6-LR	94.65	6-SVM	88.61	6-Ada	89.97	6-XGB	90.58
	7-KNN	93.37	7-LR	94.65	7-SVM	88.55	7-Ada	89.89	7-XGB	90.34
	8-KNN	93.31	8-LR	94.69	8-SVM	88.39	8-Ada	89.89	8-XGB	90.42
	9-KNN	93.21	9-LR	94.57	9-SVM	88.27	9-Ada	89.91	9-XGB	90.32
	10-KNN	93.15	10-LR	94.41	10-SVM	88.21	10-Ada	89.77	10-XGB	90.09
	$\Delta_{\max}$	3.87	$\Delta_{\max}$	3.48	$\Delta_{\max}$	2.60	$\Delta_{\max}$	1.64	$\Delta_{\max}$	2.27

achieved the best performance (71.42%) on the MattrSet dataset. Sufficient feature-shared knowledge among the three DVSs was better mined by our model, which improved the final recognition performance. A similar result can be seen in Figure 4. Hence, the proposed middle feature fusion strategy only requires careful tuning of the number of DVSs to capture sufficient feature-shared knowledge for more effective material recognition.

In conclusion, the proposed middle feature fusion approach outperformed the voting-based late feature fusion approach. Moreover, our progressive feature fusion-based SECGS model is effective and robust in recognizing both fine-grained and coarse-grained materials.

4.7. Comparisons with SOTA Baseline Models. In this section, we compare the SECGS model with a group of SOTA baseline models. Detailed accuracy comparisons are presented in Table 8. On the MattrSet dataset, the SECGS_{3-SVM} model achieved a very competitive performance compared with that of the best baseline (HMF²). Certainly, material recognition remains a challenge on this dataset. On the Fabric dataset, the SECGS_{5-LR} model outperformed all the baseline models, obtaining an approximately 10.37% performance improvement compared with that of the most competitive material recognition model recently proposed (HMF²). Hence, unlike the baseline models, the SECGS model achieved a more balanced recognition performance. This is mainly attributed to the diverse knowledge learned by the progressive feature fusion method. However, unlike GS-AdaBoost and GS-XGBoost, our model adopts only the SENet feature. Moreover, unlike HMF², our model performs early feature fusion only once. All these helps improve its real-time efficiency (Table 9).

In addition, on the Fabric dataset, the SECGS model achieved an approximately 1.11% performance improvement compared with that of the most competitive fine-grained classification baseline (DCL). This means that the SECGS model can mine sufficient knowledge to obtain better performance. On the MattrSet dataset, it performed comparatively well to the most competitive fine-grained classification model (PMG). This also demonstrated the effectiveness of the knowledge mined by our method.

We also found that our variants, SEC_{LR} and SEC(s), performed well (e.g., on the Fabric dataset, SEC(s) outperformed the HMF² model, with an improvement of 8.85%). The experimental results also illustrate an aspect of the ablation analysis; the corresponding accuracy in descending order was “SECGS > SEC(s) > SEC > general classifier.” This further confirms the effectiveness of our model, which can learn diverse valuable knowledge through progressive feature fusion. Notably, Figures 4 and 5 show another perspective of the ablation analysis.

The corresponding comparisons of the real-time efficiency are presented in Table 9. We performed our experiments using the recognition system we deployed on a laptop (Figure 6). For instance, on the MattrSet dataset, our model required only 13.26 s to test a material image, which is 5.43 s faster than the HMF² model. In summary, the results given in Table 9 indicate that, in addition to its effectiveness and robustness, the SECGS model is also more efficient than the SOTA baseline models.

In summary, the proposed SECGS model is effective, efficient, and robust for material recognition on both the Fabric and MattrSet datasets.

5. Parameter Tuning

5.1. Cross-Validation Analysis. Owing to the scarcity of material samples, cross-validation is required to objectively evaluate the proposed material recognition model. In this section, we evaluate two-fold and four-fold cross-validations to show the robustness of our method. Table 10 shows the corresponding performance comparisons. Here, we only compared three features, namely, ResNet, SENet, and DVS, which is sufficient to validate our choice.

As shown in Table 10, the SENet feature beats outperformed the ResNet feature, whereas DVS outperformed the SENet and ResNet features by a large margin. It is reasonable to mine the implicit cross-modal knowledge among the heterogeneous SENet features. In addition, the performance of the four-fold cross-validation was better than that of the two-fold cross-validation, which further indicates that the proposed DVSs are robust for material recognition and that they help enhance the practicability of the SECGS model. Certainly, material recognition remains a challenge on the MattrSet dataset. In conclusion, we used four-fold cross-validation to complete all experiments.

FIGURE 4: Evaluation results of all the DVS combinations in the middle feature fusion. “2 & 45” means that 2 DVSs were chosen to complete the middle feature fusion, and 45 DVS combinations were created for the material recognition. (a) MattrSet. (b) Fabric.

TABLE 8: Comparisons with other state-of-the-art baseline methods on each dataset. The best value of each dataset is shown in bold. Unit: %.

Dataset	Model	Acc.	Model	Acc.	Model	Acc.
MattrSet	CatBoost	68.68	Densenet169	59.77	NTS-Net	71.80
	LightGBM	68.72	ResNeXt WSL	67.87	PMG	71.79
	GS-Ada-SGL	66.11	SEC _{LR}	68.94	SEC(h)	70.81
	GS-XGB-SGL	67.67	SEC _{XGB}	69.02	SEC(s)	71.22
	SVM-S	47.78	SEC _{SVM}	70.30	SECGS _{4-Ada}	70.19
	HMF ²	72.20	DCL	70.97	SECGS _{4-KNN}	70.86
	MDCA	49.84	WS-DAN	71.67	SECGS _{6-LR}	70.90
	SENet	69.34	WILDCAT	71.04	SECGS _{6-XGB}	71.13
	InceptionResNetV2	52.09	WSCNet	52.12	SECGS _{3-SVM}	71.42
Fabric	CatBoost	88.47	ResNeXt WSL	59.66	NTS-Net	88.50
	LightGBM	88.03	Albe + Norm	79.60	PMG	93.30
	GS-XGB-SG	84.20	SEC _{Ada}	88.45	SEC(h)	91.61
	SVM-S	80.63	SEC _{KNN}	89.55	SEC(s)	93.19
	HMF ²	84.34	SEC _{LR}	91.23	SECGS _{3-SVM}	88.90
	MDCA	65.36	DCL	93.60	SECGS _{4-Ada}	90.09
	SENet	89.38	WS-DAN	91.86	SECGS _{6-XG}	90.58
	InceptionResNetV2	46.22	WILDCAT	81.13	SECGS _{5-KNN}	93.42
	Densenet169	46.29	WSCNet	62.50	SECGS _{5-LR}	94.71

TABLE 9: Comparisons of the real-time recognition efficiency on each dataset. The best value of each dataset is shown in bold. Unit: s.

Dataset	Model	Time (per image)
MattrSet	MDCA	19.85
	GS-XGBoost	16.82
	HMF ²	18.69
	SECGS _{3-SVM}	13.26
Fabric	MDCA	19.99
	GS-XGBoost	11.60
	HMF ²	16.86
	SECGS _{5-LR}	13.21

5.2. Comparisons of the Summation and Concatenation Modes in Early Feature Fusion. The proposed early feature fusion method uses two fusion modes, namely, summation (SUM) and concatenation (CON), to create the DVS. Ten DVSs were chosen to evaluate the effectiveness of these modes, and the corresponding results are shown in Figure 7. In the evaluation, we used five classifiers and four-fold cross-validation to perform material recognition, and the values of the AA_C metric were used to draw the line graphs.

As shown in Figure 7, the CON mode was superior to the SUM mode on each benchmark dataset. The SUM mode reduced the dimension of the fused features (DVSs), thus diluting

the effective cross-modal knowledge. In contrast, the dimension generated by the CON mode was more appropriate, as it retained the maximum clustering canonical correlations among the heterogeneous SENet features. Therefore, our model uses the CON mode to complete the early feature fusion.

5.3. Evaluation of the DVSs. In this section, we further make a comprehensive validation of the effectiveness of the proposed DVSs. Table 4 shows the performance of the corresponding DVS generated using different features. DVS (re) represents the DVS generated by the ResNet features, including ResNet 50, ResNet 101, ResNet 152, ResNeXt 50, and ResNeXt 101. DVS (re-se) represents the DVS generated by the ResNet and SENet features. DVS (se) represents the DVSs generated by the SENet features. $\Delta\text{Improve}_1$ represents the performance improvement of the DVS (re) compared with that of the corresponding ResNet feature. $\Delta\text{Improve}_2$ represents the performance improvement of the DVS (re-se) compared with that of the corresponding SENet feature. $\Delta\text{Improve}_3$ represents the performance improvement of the DVS (se) compared with that of the corresponding SENet feature. Several valuable conclusions can be drawn as follows:

- (1) According to MAA, there was an obvious upward trend with the adoption of the SENet features. This indicates that the SENet features, rather than the ResNet features, can provide more valuable knowledge for material recognition. This finding validates our choice of the SENet features as the basis for our progressive feature fusion method.
- (2) According to $\Delta\text{Improve}_1$, the corresponding DVSs achieved satisfactory performance even if we only used the ResNet features, which shows the robustness of the proposed early feature fusion method. Moreover, a larger MAA improvement was obtained on the MattrSet dataset, which confirms that a significant amount of cross-modal knowledge among the ResNet features was mined out from the MattrSet dataset.
- (3) According to $\Delta\text{Improve}_2$, performance improvements were not always positive when only one SENet feature was used to complete the early feature fusion. There must be some noise among the ResNet and SENet features, which significantly affected the quality of the acquired cross-modal knowledge (DVS) and led to a performance decline. This suggests that we should use the CNN-based features from a homogeneous network to complete the proposed early feature fusion, which helps obtain high-quality cross-modal knowledge for improving the performance of the final recognition.
- (4) According to $\Delta\text{Improve}_3$, the performance improvements were always positive when only the SENet features were used for the early feature fusion. High-quality cross-modal knowledge was fully mined to generate the proposed DVSs. This also

validates our original selection and serves as a basis for the subsequent middle feature fusion.

In conclusion, the proposed DVSs mined from the heterogeneous SENet features are effective and robust for high-quality material recognition.

5.4. Visualization of the Proposed DVSs. To further illustrate the effectiveness of the proposed DVSs, we employed the well-known t-SNE [63] algorithm to visualize some relevant features. Taking the Fabric dataset as an example, we chose several features, including ResNet 50 (R50) (similar to ResNeXt 101 (RXt101)), SR50, SRXt101, R50-RXt101 (DVS (re)), SR50-RXt101 (DVS (re-se)), and SR50-SRXt101 (DVS (se)), for the visualization. The corresponding experimental results are shown in Figure 5. Many valuable conclusions can be drawn as follows:

- (1) The original R50 feature represented very messy material categories. Hence, the traditional deep learning features are insufficient for describing different material categories (see Tables 6 and 10 for the quantitative results).
- (2) The original SR50 represented the material categories better than the R50 feature. However, confusion still occurred, as shown in Figure 5(b). A similar result can be seen in Figure 5(c). Therefore, the original deep learning-based features are insufficient for depicting material images.
- (3) DVS (se) can represent each material category more accurately than the single features (R50, SR50, and SRXt101). The same color blocks (samples in the same category) started to aggregate, whereas different color blocks (samples in different categories) began to separate from each other. This means that the semantic distance between the same samples shrank, whereas that between the different samples expanded gradually, which helped the classification model to train a more appropriate decision function (Table 5). When the proposed DVSs were used, even the traditional classification models, including SVM, KNN, and LR, were able to obtain satisfactory recognition performance.
- (4) Comparing the three types of DVS (DVS (re), DVS (re-se), and DVS (se)), we obtained the performance rank in the descending order as follows: "DVS (se) > DVS (re-se) > DVS (re)." According to Figures 5(d)–5(f), DVS (se) is better than DVS (re) in representing different material categories. The SENet features, rather than the ResNet features, provide more effective knowledge for early feature fusion. This is an intuitive explanation of the results presented in Table 4. This also validates our original selection. From Figures 5(a)–5(d), we can see that DVS (re) had a better ability to represent different material categories than that of the original R50, which shows the effectiveness of the proposed early feature fusion method from another perspective.

FIGURE 5: Continued.

FIGURE 5: Visualization results of different features on the Fabric dataset. (a) R50, (b) SR50, (c) SRXt101, (d) R50-RXt101 (DVS (re)), (e) SR50-RXt101 (DVS (re-se)), and (f) SR50-SRXt101 (DVS (se)).

FIGURE 6: Example demonstration of the SECGS-based real-time material recognition system.

TABLE 10: Comparisons of the two-fold and four-fold cross-validations. The best value of each column is shown in bold. Unit: %.

Dataset	Metric	Cross-validation	ResNet	SENet	DVS
MattrSet	MAA	Two-fold	39.78	61.76	64.19
		Four-fold	42.21	62.84	66.06
Two-fold		43.28	78.23	81.30	
Four-fold		44.05	81.60	84.64	

FIGURE 7: Comparison between the two modes of the early feature fusion method on each dataset. (a) MattrSet. (b) Fabric.

According to the comparisons between Figures 5(e)–5(f), the material categories depicted by DVS (re-se) are messier than those depicted by DVS (se). This indicates that there is a certain amount of noise between the ResNet and SENet features, which might have affected the final quality of the acquired cross-modal knowledge. Moreover, this also reduced the ability of DVS (re-se) to characterize different material categories. All these findings also confirm the conclusions we mentioned in the above subsection “Evaluation of the DVSs.”

5.5. Evaluation of All the DVS Combinations. Figure 4 shows the corresponding recognition performance of all the DVS combinations in the proposed middle feature fusion procedure. As can be seen from the median line, the accuracy of each classifier tends to improve with the increase in the number of DVSs. This means that, as the number of DVSs increases, the SECGS model can acquire sufficient feature-shared knowledge to become more stable (similar results can also be seen in Table 7).

For the MattrSet dataset, an accuracy peak was observed when the SVM classifier used three DVSs. For the Fabrics dataset, an accuracy peak was observed when the LR classifier used five DVSs. These results are consistent with the best values given in Table 8. Moreover, the SVM classifier

achieved the best overall performance on the MattrSet dataset, whereas the LR classifier achieved it on the Fabric dataset. Hence, even the traditional classifiers were able to obtain satisfactory performance using the proposed progressive feature fusion method owing to the diverse knowledge obtained by our model. This finding is more evident on the MattrSet dataset.

In addition, the performance variations of our approach showed a positive effect on the number of DVSs, which indicates that more DVSs can obtain more stable performance (the lower edge (minimum) of each box increases with the number of DVSs, and the smaller space between the upper and lower edges indicates that the corresponding performance fluctuates less). These results show that feature-shared knowledge can be fully obtained using the proposed approach, thus improving the final recognition performance. In summary, like the cross-modal knowledge contained in the DVSs, the feature-shared knowledge among different DVSs can also be mined by the progressive feature fusion approach, and this knowledge is effective for high-quality material recognition.

5.6. Evaluation of the Contribution of Each Component of the SECGS Model. To evaluate the real contribution of each component of the SECGS model, we plotted the corresponding precision of SENet, DVS, SEC(s), and SECGS for

FIGURE 8: Precision of each component of SECGS for each material category. (a) MattrSet. (b) Fabric.

each material category. The corresponding results are shown in Figure 8. The evaluation was done as per the ablation analysis procedure, which helped us to objectively evaluate each component of the progressive feature fusion method.

First, gradual performance improvements were observed on each dataset using different components. This observation was more evident on the Fabric dataset (e.g., fleece, silk, and wool). Second, the proposed DVSs easily outperformed the original SENet feature, which indicates that they, unlike the single image feature (SENet), obtained sufficient cross-modal knowledge from the heterogeneous SENet features to better depict material images (Tables 6 and 10 and Figure 5). Third, the SECGS model outperformed the proposed DVSs, which means that implicit feature-shared knowledge was fully obtained through middle feature fusion to boost the final recognition performance.

Moreover, our model performed better on the Fabric dataset (similar results can be seen in Tables 7 and 8). Owing to its different implementation mechanisms, our model will be a good complement to the SOTA HMF² model, which did not do well on the MattrSet dataset. However, the recognition of materials, such as bag_pu and bag_polyester, remains challenging. We also found that better but more stable performance was achieved by using the proposed middle feature fusion method than that by using ensemble learning (Table 7). The proposed progressive feature fusion method (early feature fusion + middle feature fusion) is more effective and robust than the SEC(s) model (early feature fusion + late feature fusion). Although ensemble learning performed well on some materials, the proposed middle feature fusion method achieved a more balanced performance improvement on

most material categories. All these confirm the effectiveness and robustness of the proposed progressive feature fusion method from another perspective.

To further show the real contribution of each component of our approach, we performed another ablation analysis experiment and plotted the results in Figure 9. This figure shows the corresponding performance improvement after the ablation operation. “SE-SF” represents the largest performance improvement of the SENet feature relative to that of the best single feature. “DVS-SE” represents the largest performance improvement of the proposed DVSs relative to that of the SENet feature. “SECGS-DVS” represents the largest performance improvement of the proposed SECGS model relative to that of the generated DVS. “SEC(s)-DVS” represents the largest performance improvement of the proposed SEC(s) model relative to that of the generated DVS.

As shown in Figure 9, each component of the SECGS model tried its best to boost the final recognition performance. However, the contribution rank of the components in the descending order was “SENet > GS > DVS.” Hence, the SOTA SENet feature is the most important component of the proposed SECGS model, whereas the effect of the proposed DVSs is relatively small. This conclusion is more evident on the Fabric dataset. This also motivates us to design a more effective early feature fusion method to create novel DVSs in the future. We also found that our variant models (SEC(s)) obtained satisfactory recognition performance, which further confirms the robustness of the proposed progressive feature fusion method. Clearly, the proposed middle feature fusion method outperformed the voting-based late feature fusion method, which confirms the

FIGURE 9: Performance of each component of the SECGS model on the two datasets.

effectiveness of the implicit feature-shared knowledge mined by our GS-based middle feature fusion method.

6. Real-Time Material Recognition System

To verify the practicability of the proposed SECGS model, we designed a real-time material recognition system based on the SECGS model with two recognition modes that can be used through local uploading and instant shooting. The workflow of this system is shown on the left-hand side of Figure 6. First, users can upload local images or instantly shoot a photo using a mobile IP camera. Second, the system performs feature extraction and progressive feature fusion to complete the material recognition (see the right-hand side of Figure 6). The entire recognition operation is performed on the front page of the cloud application without any manual intervention. This can improve users' interaction experience and serve as an important basis for subsequent configurations on a cloud server.

As Table 9 shows, our SECGS model-based real-time material recognition system can achieve satisfactory efficiency compared with the SOTA baseline models. Hence, the SECGS model is efficient for real-time material recognition. If an advanced computer or server with a higher configuration is adopted, the corresponding real-time efficiency will be further improved. Unlike the HMF² and GS-XGBoost models, the SECGS model adopts only the SENet features to complete the progressive feature fusion. This can improve the final real-time efficiency from another perspective. Hence, the real-time material recognition system narrows the gap between theoretical research and practical applications (scene recognition, industrial detection, robot vision, instant shopping, etc.). For example, in the field of industrial detection, the system can assist textile enterprises in completing fabric defect classification more efficiently. Workers only need a mobile phone to perform this task more easily, thus improving work efficiency and reducing labor costs. For customers who are shopping, they only need a mobile phone to accurately identify the materials of the clothing goods they want to

purchase, to assist them in their final purchasing decisions. The system can also be combined with an intelligent robot such as NAO [64] to complete the real-time clothing recognition and grasp estimation, which can better assist the elderly. In summary, our SECGS model-based real-time material recognition system can be seamlessly used with many daily applications to meet practical demands.

7. Conclusions and Future Work

We presented a model called SECGS using progressive feature fusion for material recognition. We demonstrated its effectiveness, efficiency, and robustness on two benchmark datasets. By considering progressive feature fusion among the SOTA heterogeneous SENet features, the proposed SECGS model can gather sufficient material knowledge, including cross-modal and feature-shared knowledge, to improve the final recognition performance. Moreover, it can be deployed on a laptop for online real-time material recognition, thereby demonstrating its high practicality. In the future, we plan to advance our research in the following aspects:

- (1) We intend to use some SOTA data augmentation methods, such as AutoAugment [65] and Adversarial AutoAugment [66], to solve the problem of unbalanced material images and establish a firm data foundation for high-quality material recognition.
- (2) We plan to apply the SECGS model to other research fields, including tumor recognition [67], COVID-19 detection [68], and scene recognition [69]. This can help validate the versatility of our model.
- (3) We will try to develop a material recognition mobile application, which can really help narrow the gap between theoretical research and practical applications.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant numbers: 62161011, 61762038 and 61861016; the Natural Science Foundation of Jiangxi Provincial Department of Science and Technology, grant numbers: 20202BABL202044, 20202BABL212006, and 20192ACB21004; the Key Research and Development Plan of Jiangxi Provincial Science and Technology Department, grant numbers: 20171BBG70093, 20192BBE50071, and 20202BBEL53003; the Humanity and Social Science Fund of Ministry of Education of China, grant number: 20YJAZH142, the Science and Technology Projects of Jiangxi Provincial Department of Education, grant numbers: GJJ190323, GJJ200627, and GJJ200644; and the Humanity and Social Science Foundation of Jiangxi University, grant numbers: TQ19101 and TQ20108. The authors would like to sincerely thank Yi Yin who created the MatrrSet dataset; C Kampouris, S Zafeiriou, A Ghosh, and S Malassiotis who provided the Fabric dataset; Ziliang Jiang, Jinpeng Wu, Renzhong Wu, Chuanxiu Li, and Jianwu Zhuo who suggested many valuable pieces of advice about the SECGS model.

References

- [1] X. Jiang, J. Du, B. Sun, and X. Feng, "Deep dilated convolutional network for material recognition," in *Proceedings of the International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6, Xi'an, China, November.
- [2] G. Schwartz and K. Nishino, "Recognizing material properties from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1981–1995, 2020.
- [3] W. M. Bergmann Tiest and A. M. L. Kappers, "The influence of visual and haptic material information on early grasping force," *Royal Society Open Science*, vol. 6, no. 3, Article ID 181563, 2019.
- [4] Y. Quan, H. Teng, T. Liu, and Y. Huang, "Weakly-supervised sparse coding with geometric prior for interactive texture segmentation," *IEEE Signal Processing Letters*, vol. 27, pp. 116–120, 2020.
- [5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1785, Miami, FL, USA, June 2009.
- [6] C. Zheng, J. Chen, J. Kong et al., "Scene recognition via semi-supervised multi-feature regression," *IEEE Access*, vol. 7, Article ID 121612, 2019.
- [7] H. Ouanan and E. H. Abdelwahed, "Image processing and machine learning applications in mining industry: mine 4.0," in *Proceedings of the International Conference Intelligent Systems and Advanced Computing Sciences (ISACS)*, pp. 1–5, Taza, Morocco, December 2019.
- [8] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Medical image fusion via convolutional sparsity based morphological component analysis," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 485–489, 2019.
- [9] J. Bednarek, M. Bednarek, P. Kicki, and K. Walas, "Robotic touch: classification of materials for manipulation and walking," in *Proceedings of the International Conference Soft Robotics (RoboSoft)*, pp. 527–533, Seoul, Korea (South), April 2019.
- [10] J. Tan, Q. Lai, J. Zhao, Z. Zhao, and J. Tan, "BRDF characteristics of different textured fabrics in visible and near-infrared band," *Optics Express*, vol. 28, no. 3, pp. 3561–3575, 2020.
- [11] X. Xian-Hua Han, Y. Gang Xu, and G. Xu, "High-order statistics of weber Local descriptors for image representation," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1180–1193, 2015.
- [12] K. Tanaka, Y. Mukaigawa, T. Funatomi, H. Kubo, Y. Matsushita, and Y. Yagi, "Material classification using frequency-and depth-dependent time-of-flight distortion," in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 2740–2749, Honolulu, HI, USA, July 2017.
- [13] Z. Erickson, E. Xing, B. Srirangam, S. Chernova, and C. C. Kemp, "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, pp. 10452–10459, Las Vegas, NV, USA, Jan 2020.
- [14] A. Shahriari, "Learning deep filter banks in parallel for texture recognition," in *Proceedings of the International Conference on Image Processing (ICIP)*, pp. 1634–1638, Phoenix, AZ, USA, August 2016.
- [15] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis for feature level fusion with application to multimodal biometrics," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1866–1870, Shanghai, China, March 2016.
- [16] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, and T. Li, "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Applied Soft Computing*, vol. 80, pp. 57–79, 2019.
- [17] H. Zhang, Z. Jiang, Q. Xiong et al., "Gathering effective information for real-time material recognition," *IEEE Access*, vol. 8, pp. 159511–159529, 2020.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, Nevada, USA, Jun 2016.
- [19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [20] A. Oliva and A. Torralba, "Chapter 2 b," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.
- [21] M. Pietikainen, A. Hadid, G. Zhaomei, and T. Ahonen, *Computer Vision Using Local Binary Patterns*, Springer Berlin Heidelberg, London Ltd, Berlin, Germany, 2011.
- [22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for largescale image recognition," in *Proceedings of the International Conference on Learning Representation (ICLR)*, pp. 1–15, San Diego, CA, USA, April 2015.
- [23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.
- [24] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proceedings of the*

- International Conference Artificial Intelligence and Statistics (AISTATS)*, pp. 823–831, Reykjavic, Iceland, April 2014.
- [25] C. K. Gatebe and M. D. King, “Airborne spectral BRDF of various surface types (ocean, vegetation, snow, desert, wetlands, cloud decks, smoke layers) for remote sensing applications,” *Remote Sensing of Environment*, vol. 179, pp. 131–148, 2016.
  - [26] K. Z. Doctor, C. M. Bachmann, D. J. Gray, M. J. Montes, and R. A. Fusina, “Wavelength dependence of the bidirectional reflectance distribution function (BRDF) of beach sands,” *Applied Optics*, vol. 54, no. 31, pp. F243–F255, 2015.
  - [27] H. Wang, W. Zhang, and A. Dong, “Measurement and modeling of bidirectional reflectance distribution function (BRDF) on material surface,” *Measurement*, vol. 46, no. 9, pp. 3654–3661, 2013.
  - [28] W. Shi, J. Zheng, Y. Li, X. Li, and Q. An, “Measurement and modeling of bidirectional reflectance distribution function (BRDF) on cutting surface based on the coaxial optical microscopic imaging,” *Optik*, vol. 170, pp. 278–286, 2018.
  - [29] Y. Xie, J. Tan, L. Jing, W. Zhang, and Q. Lai, “Investigating directional reflection characteristics of anisotropic machined surfaces using a self-designed scatterometer,” *Applied Optics*, vol. 58, no. 29, pp. 7970–7979, 2019.
  - [30] J. Chen, S. Shan, C. He et al., “WLD: a robust local image descriptor,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.
  - [31] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, “Exploring features in a bayesian framework for material recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 239–246, San Francisco, CA, USA, June 2010.
  - [32] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, “Deep filter banks for texture recognition, description, and segmentation,” *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, 2016.
  - [33] F. Perronnin and D. Larlus, “Fisher vectors meet neural networks: a hybrid classification architecture,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3743–3752, Boston, MA, USA, June 2015.
  - [34] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson, “Recognizing materials using perceptually inspired features,” *International Journal of Computer Vision*, vol. 108, no. 3, pp. 348–371, 2013.
  - [35] S. Bell, P. Upchurch, N. Snaveley, and K. Bala, “Material recognition in the wild with the materials in context database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3479–3487, Boston, MA, USA, June 2015.
  - [36] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.
  - [37] L. Sifre and S. Mallat, “Rotation, scaling and deformation invariant scattering for texture discrimination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1233–1240, Portland, OR, USA, June 2013.
  - [38] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “PCANet: a simple deep learning baseline for image classification,” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.
  - [39] N. Rasiwasia, J. Costa Pereira, E. Coviello et al., “A new approach to cross-modal multimedia retrieval,” in *Proceedings of the ACM International Conference on Multimedia (ACM-MM)*, pp. 251–260, Firenze, Italy, October 2010.
  - [40] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1247–1255, Atlanta, USA, June 2013.
  - [41] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, “Joint blind source separation by multiset canonical correlation analysis,” *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.
  - [42] Y. Peng, J. Qi, X. Huang, and Y. Yuan, “CCL: cross-modal correlation learning with multigrained fusion by hierarchical network,” *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2018.
  - [43] Y. Bai, M. Haghighat, and M. Abdel-Mottaleb, “Kernel discriminant correlation analysis: feature level fusion for non-linear biometric recognition,” in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 3198–3203, Beijing, China, December 2018.
  - [44] E. Zhang, X. Chen, and L. Wang, “Consistent discriminant correlation analysis,” *Neural Processing Letters*, vol. 52, pp. 891–904, 2020.
  - [45] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, “Multiple kernels for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–613, Miami, FL, USA, June 2009.
  - [46] H. Xia, “SCH. Hoi. “MKBoost: a framework of multiple kernel boosting,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1574–1586, 2013.
  - [47] G. Rätsch, T. Onoda, and K. Müller, “Soft margins for adaBoost,” *Machine Learning*, vol. 44, pp. 287–320, 2001.
  - [48] T. Chen and G. Carlos, “XGBoost: a scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 785–794, New York, NY, USA, June 2016.
  - [49] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Proceedings of the Neural Information Processing Systems (NIPS)*, pp. 6638–6648, Montreal, Canada, January 2018.
  - [50] G. Ke, Q. Meng, T. Finley et al., “LightGBM: a highly efficient gradient boosting decision tree,” in *Proceedings of the Neural Information Processing Systems (NIPS)*, pp. 3146–3154, Long Beach, CA, USA, December 2017.
  - [51] C. Kampouris, S. Zafeiriou, A. Ghosh, and S. Malassiotis, “Fine-grained material classification using micro-geometry and reflectance,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 778–792, Amsterdam, Netherlands, September 2016.
  - [52] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, September 2015.
  - [53] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, “Inside-Outside net: detecting objects in context with skip pooling and recurrent neural networks,” in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 2874–2883, Las Vegas, NV, USA, June 2016.
  - [54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in *Proceedings of the International Conference Artificial Intelligence*, pp. 4278–4284, San Francisco, CA, USA, February 2017.

- [55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the International Conference Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, HI, USA, August 2017.
- [56] D. Mahajan, R. Girshick, V. Ramanathan, and K. He, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11206, pp. 185–201, October 2018.
- [57] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5157–5166, Long Beach, CA, USA, May 2019.
- [58] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference Computation Vision (ECCV)*, pp. 420–435, Munich, Germany, September 2018.
- [59] R. Du, D. Chang, J. Xie et al., "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Proceedings of the European Conference Computation Vision (ECCV)*, pp. 153–168, March 2020.
- [60] T. Durand, T. Mordan, N. Thome, and M. Cord, "Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5957–5966, Honolulu, Hawaii, July 2017.
- [61] J. Yang, D. She, Y. Lai, P. L. Rosin, and M. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7584–7592, Salt Lake City, UT, USA, June 2018.
- [62] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification," 2019, <https://arxiv.org/abs/1901.09891>.
- [63] V. D. M. Laurens and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [64] Q. Liu, C. Zhang, Y. Song, and B. Pang, "Real-time object recognition based on NAO humanoid robot," in *Proceedings of the 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 644–650, Halifax, Canada, August 2018.
- [65] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: learning augmentation strategies from data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, Long Beach, CA, USA, April 2019.
- [66] X. Zhang, Q. Wang, J. Zhang, and Z. Zhong, "Adversarial AutoAugment," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, December 2020.
- [67] W. Yao and S. Thomas, "Deep learning-based magnetic resonance imaging image feature analysis for pathological classification of brain glioma," *Scientific Programming*, vol. 2021, Article ID 6778009, 9 pages, 2021.
- [68] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine learning-based model to predict the disease severity and outcome in COVID-19 patients," *Scientific Programming*, vol. 2021, Article ID 5587188, 10 pages, 2021.
- [69] R. C. Joshi, S. Yadav, M. K. Dutta, and C. M. Travieso-Gonzalez, "Efficient multi-object detection and smart navigation using artificial intelligence for visually impaired people," *Entropy*, vol. 22, p. 941.