IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 19, 2021, accepted November 29, 2021, date of publication December 22, 2021, date of current version January 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137585

Material Type Recognition of Indoor Scenes via Surface Reflectance Estimation

SEOKYEONG LEE[1], DONGJIN LEE[2], HYUN-CHEOL KIM[3], AND SEUNGKYU LEE[2]

[1]Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
[2]Department of Computer Engineering, Kyung Hee University, Yongin 17104, Republic of Korea
[3]Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

Corresponding author: Seungkyu Lee (seungkyu@khu.ac.kr)

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) Grant by the Korean Government (The research of the fundamental media contents technologies for hyper-realistic media space) under Grant 21ZH1200.

ABSTRACT There are fundamental difficulties in obtaining material type of an arbitrary object using traditional sensors. Existing material type recognition methods mostly focus on color based visual features and object-prior. Surface reflectance is another critical clue in the characterization of certain material type and can be observed by traditional sensors such as color camera and time-of-flight depth sensor. A material type is characterized well by relevant surface reflectance together with traditional visual appearance providing better description for material type recognition. In this work, we propose a material type recognition method based on both color and reflectance features using deep neural network. Proposed method is evaluated on both public and our own data sets showing promising material type recognition results.

INDEX TERMS Material type, surface reflectance.

I. INTRODUCTION

Objects are composed of multiple distinguishing materials. Material type is an essential characteristic in determining the class of object. Furthermore, materials composing an object reveal detailed sub-classes such as wooden vs iron table, fabric vs leather couch. Recognized material types enable realistic rendering of reconstructed three-dimensional model. Therefore, material type recognition has been importance research optic in computer vision and graphics. Earlier studies on material type recognition largely depend on visual features such as color patterns and texture shapes of each material type. Sharan *et al.* [1] investigate various combination of contemporary visual descriptors to suggest optimal color feature set for material recognition evaluated on FMD (Flickr Material Database). Bell *et al.* [2] prove that context information encoded from general convolutional neural network helps material recognition in indoor scenes. Materials IN Context(MINC) database proposed in this work has been widely used for the comparison of material recognition methods in the field. FVCNN [3] proposes a learning based texture descriptor using CNN filter banks and Fisher Vector pooling. Degol *et al.* [4] prove that using 3D geometry based features improves material recognition performance of a large-scale scene.

There have been many approaches for material type recognize using ToF camera [5]–[11]. Su *et al.* [7] have captured subsurface scattering that specifies the characteristics of material surface. However, it covers only small number of material-classes requiring additional light source. Tanaka *et al.* [8] propose an exemplar-based material classification method that uses the distortion of the observations of ToF camera. They claim that depth distortion caused by varying modulation frequencies and camera-object distance provide rich information on material class types. However, it relies on particular device-dependent characteristic observed during data acquisition process. Both prior work [7], [8] provide classification result only with controlled experimental environment. Kim *et al.* [9] propose a surface roughness estimation method with ToF camera. Off the shelve depth camera such as Microsoft Kinect is used to estimate surface roughness together with color features.

There are several attempts to use physically-motivated features like 3d geometry and surface reflectance for material recognition [12]–[14]. Due to the limited data acquisition environments and experimental costs, the performance of earlier methods is limited compared to color based approaches. Recent material recognition methods have addressed such

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

issues by properly designing deep neural networks for material type recognition. Zhang *et al.* [15] propose Deep TEN, a material recognition framework which uses generalized encoders like VLAD [16] and Fisher Vector [17] in a form of encoding layer. DEP(Deep Encoding Pooling) network [18] shows performance improvement over Deep TEN [15] by adding local spatial feature encoder. Zhai *et al.* [19] suggest MAP(Multiple Attribute Perceived)-Net which learns a bag of texture attributes representing texture-specific characteristics. Their improved work, called DSR(Deep Structure-Revealed)-Net [20], extracts an inherent spatial dependency of a texture which helps the textures to be recognized and deformed properly. However, both of the proposed methods work with only closed-up surface texture images. Various attempts to estimate or reconstruct surface reflectance have been conducted [12], [21]. Recent studies [22]–[26] have used multi-view images of material object to implicitly obtain physically-motivated features instead of using highly-restricted experimental setups [13], [27]–[30].

There has been large body of studies on capturing reflectance from a set of color images under varying experimental conditions [30]–[37]. Sengupta *et al.* [38] have suggested neural inverse rendering method which enables scene-attribute estimation from single indoor image. Despite several practical advantages of the method, it handles real image in a self-supervised manner showing dependency on training data. Li *et al.* [39] suggest indoor scene-wise SVBRDF(Spatially-Varying Bidirectional Reflectance Distribution Function) estimation framework from single synthetic image. But the method has limitation in that it cannot be trained on real data since it requires synthetic ground truth data for training. Murmann *et al.* [40] propose *Multi-illumination Images in the Wild(MIW)* dataset with per-image illumination condition and per-pixel material type label. Our color image based reflectance estimation network is trained on the data set.

Wang *et al.* [23] suggest light-field camera based approach for joint reconstruction of 3D shape and surface reflectance of an object. Xue *et al.* [25] suggest DAIN (Differential Angular Imaging Network), which uses small angular variance between closely captured image pairs as a reflectane-related material feature. However, these methods lack practicalities for its needs for customized devices or for conditioned environment. On the other hand, thanks to recent advance of deep learning, several attempts have been made for few-shot surface reflectance estimation [31]–[34], [37], [41], [42]. But these methods show limitation in that it can be applied in closed-up texture-like images or object-centered photos. Some of the recent works [38], [39] have achieved scene-wise reflectance estimation, but they show limitations in the application with real world data due to the dependency on *inverse rendering* based methods.

In this work, we propose surface reflectance estimation methods using either ToF (*time-of-flight*) depth camera (IR reflectance) or multiple color images (visible light reflectance) taken under varying practical conditions. Proposed methods are able to collect pixel-wise surface reflectance that enables dense material type classification of entire scene. Based on extensive experimental evaluation, we suggest an optimal network structure for a multi-modal material type recognition.

II. MATERIAL RECOGNITION WITH COLOR FEATURES

Materials in Context Database (MINC) is one of large material data set consists of 23 material classes with context information. Total 3 million sample patches are extracted from 435,749 images of the data set. Each material sample patch includes following attributes: material label, id of original photo, 2d coordinate information of center pixel used to extract the patch. There exists a subset of MINC with patch-wise material classification label called MINC-2500. The subset has balanced class distribution (2500 samples per class) while its original data set has biased distribution of material classes. MINC has been widely used for the color-based material classification evaluation due to its diversity and large number of samples. Zhao *et al.* [10] have pointed out the problem of MINC samples: non-values occurred by pixels out of image border and re-generated new material patches based on the extraction rules of MINC. They propose real-time 3D material segmentation framework trained with their newly sampled 917,839 patches. Jurado *et al.* [43] propose a semantic segmentation method of natural materials on a point cloud using multi-spectral features. Xue *et al.* [18] show improved performance on MINC-2500 (about 82.00% of classification) using deep encoding and pooling network called *DEP-Net*.

In order to perform cross-data set verification, we have generated two novel variants of MINC-2500 called MINC-NEW and OUR-NEW data set. In case of MINC-NEW, samples are extracted by the same extraction strategy with MINC-2500. Each patch is extracted from corresponding original image while none of the patches are extracted from identical image. Following the work, the size of patch is 32.9% of smaller one and resized to 362 × 362 resolution. The size of patch is decided to be 32.9% of smaller original image and then resized to 362 × 362. 500 patches per each material class (total 11,500 patches) are extracted. For OUR-NEW data set, original images are collected from two

TABLE 1. Performance comparison between the previous works and DenseNet-121 trained with MINC-2500.

Model	Accuracy	Epoch
AlexNet (Bell et al.)	76.0 ± 0.2%	N/A
Deep TEN * (*SOTA*)(Zhang et al.)	*multi : **81.3%** *single : 80.6%	>35
DenseNet-121	**81.13%**	35
DenseNet-169	78.64%	35

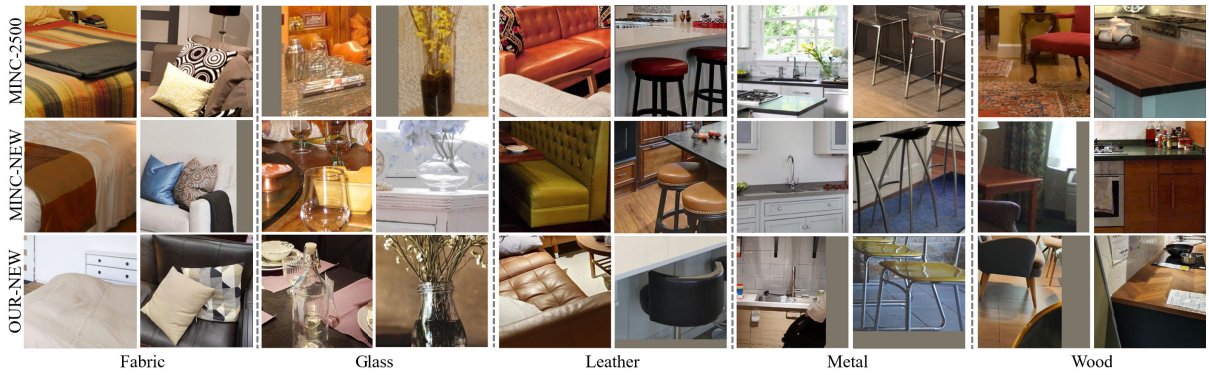


FIGURE 1. 3 different material data set used in our experiment and its corresponding reference images. Each sample shares same patch extraction strategy, containing single material label at the center of the image.

different image sources. First group consists of web-based images collected from internet. In order to keep the context of MINC data set, keywords '*interior*,' '*office*,' '*cafe*,' '*school*' are used for searching. A keyword '*Asia*' is also used to check if MINC has any bias on western environment context instead of material-specific context features. Second group includes images of material objects located in natural environment of furniture showroom, captured using mobile-phone camera. Each group consists of 300 to 400 original images and patches are obtained avoiding overlapping. Five commonly observed material classes including (*fabric, glass, leather, metal, wood*) are collected with 1,334 patches using identical patch extraction process with the first group. Figure 1 shows examples of 3 different material data set used in our experiment. Proposed two data sets have similar contextual characteristic compared to the original MINC-2500.

In this evaluation, we choose DenseNet-121 [44] trained on MINC-2500 aiming to verify whether typical material properties sufficiently appear in MINC. Proposed patch-wise material classification model is trained on the training set of MINC-2500 (48,875 instances). Table 2 shows material recognition performance of DenseNet-121 on our three different(but with similar context) material data sets. The model gives 81.13% of recognition accuracy on MINC-2500, which is compatible to Xue *et al.* [18] (82.00%). However, the model shows limited performance on our proposed data set: 69.27% on MINC-NEW and 49.78% on OUR-NEW. Considering the contextual similarities shown in Figure 1, the performance decrease indicates that trained model favors MINC-2500 images. Color-based features of an object can easily be contaminated by various real conditions such as lighting and viewing direction. Therefore, material recognition depending on only color feature has clear limitation in real-world applications.

III. SURFACE IR REFLECTANCE ESTIMATION

A. SURFACE IR REFLECTANCE FEATURE

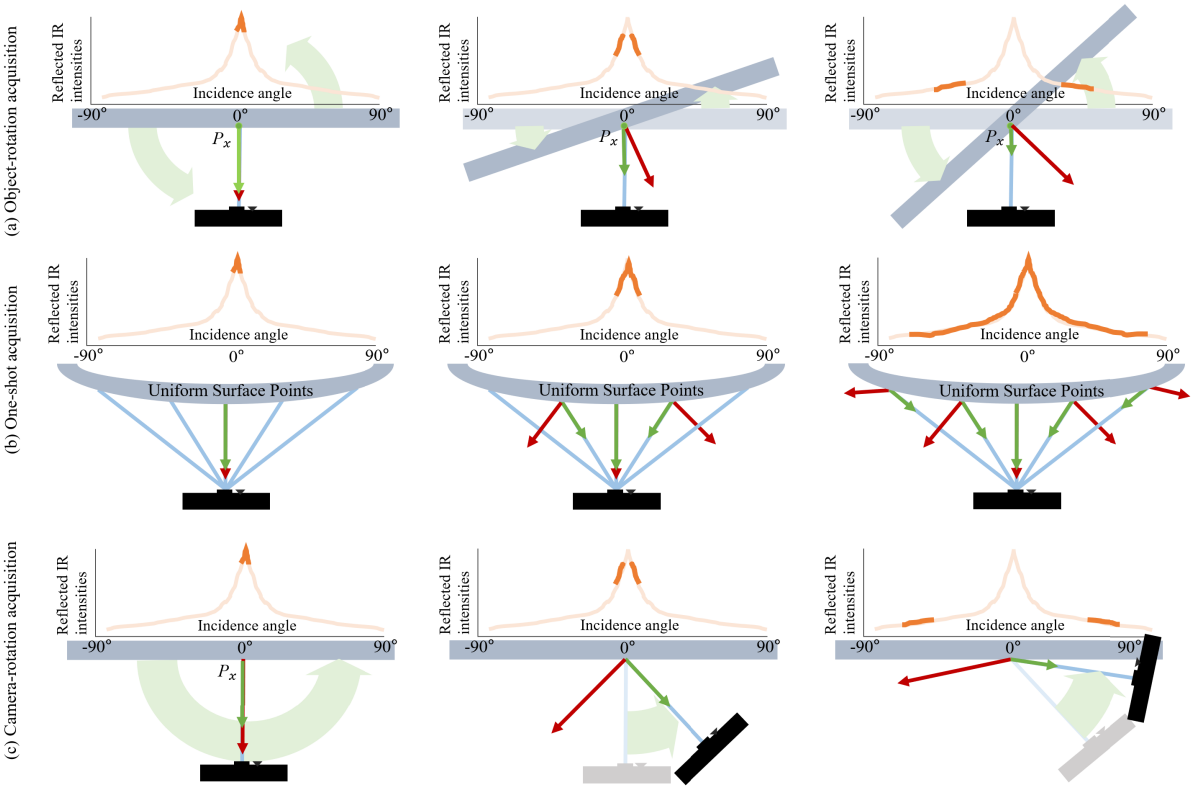
Surface reflectance of visible light has been obtained by exhaustive scanning from entire lighting or viewing directions. Anisotropic surface assumption based methods show limitations in real-world applications [28]. Kim *et al.* [9] propose surface roughness features for material recognition using single Kinect depth camera. Infrared reflectance obtained from time-of-flight depth camera is used for material feature extraction. However it still has limited application due to large angular condition (360°). By employing several practical assumptions, Lee *et al.* [11] propose a simple and practical acquisition setup. Isotropic surface assumption is reasonable since imaging setup with Kinect is limited to a single imaging device with light source. Based on Helmholtz reciprocity, they vary camera and light source positions and as a result vary incidence angle and reflectance angle. Our aim is not a precise reconstruction of BRDF of visible light. Infrared reflectance shows enough separation ability of diverse material types as verified in prior work [9], [11]. Furthermore, emitting visible light on to target objects or person for the acquision of reflectance is not practical. On the other hand, emitting invisible infrared light is free from such problem. Therefore, our proposed method employs IR Reflectance for material type classification.

B. ACQUISITION SETUPS

In order to enhance the practicality of surface IR reflectance feature acquisition process compared to the previous work [11], we use same assumptions explained above while collecting reflected IR of real-world samples. Note that some of our device setups are based on [11]. Figure 2 shows 3 different acquisition setups. Figure 2-(a) is object-rotation setup proposed by [11]. Reflected IR from the center of rotating

TABLE 2. DenseNet-121 accuracy on MINC-2500 test set and ours after trained with MINC-2500.

Train Data	Test Data	Accuracy	Test Patches
MINC-2500	MINC-2500	81.13%	5,750
	MINC-NEW	69.27%	11,500
	OUR-DATA	49.78%	1,334



	Isotropicity	Uniform Surface	Incidence Angle Range	Required Frame
(a)	isotropic	X	~35°	many
(b)	isotropic	O	~80°	1
(c)	anisotropic	X	~80°	many

(d) Experimental condition for each acquisition method

FIGURE 2. A 3 different acquisition setups of our experiments and corresponding experimental conditions.

target is acquired by camera fixed in front of the object. In order to obtain complete IR distribution, target surface has to be rotated for varying incidence angle. Figure 2-(b) shows one-shot acquisition with uniform surface material type assumption.

Consequently, it collects reflected IR from multiple surface points assuming that they share same reflectance characteristic. Especially, reflected IR intensities of a surface with sufficient surface normal variation(curved surface) can be acquired for one shot. Figure 2-(c) shows camera-rotation setup which requires point cloud registration. A newly acquired data from ToF camera are accumulated in dynamic voxel space. A camera-rotation acquisition setup is able to get pixel-wise reflectance while others get single reflectance per target object.

C. COLOR-IR MATERIAL DATA SET

We employ *Color-IR Material Data Set* [11]. Figure 3 shows examples of the data set of 7 common material types. Total 116 numbers of flat-surfaced samples are collected. Both color and proposed reflectance features are acquired using the object-rotation method described in Figure 2-(a).

IV. MATERIAL RECOGNITION WITH ToF CAMERA

A. INDOOR SCENE-WISE SURFACE REFLECTANCE ESTIMATION

Proposed method shows reliable result both with/without color features in real-world environment. However, it has several problems. First, collected reflectance feature of one-shot acquisition method [Figure 2-(b)] has narrow incidence angle variation. Camera-rotation [Figure 2-(c)] has wider incidence angle but still it suffers from registration noises. To alleviate the problem, seven individual voxel spaces from seven different viewing directions are collected. Figure 4 shows comparison between this discrete observations and camera-rotation. In camera-rotation, a well-registered, normal-based segmented point clouds are first obtained to get object boundaries. Since the point-wise IR distribution is noisy, we set neighbor point clusters to get robust point-wise reflectance features. As the result of neighbor point clustering, points

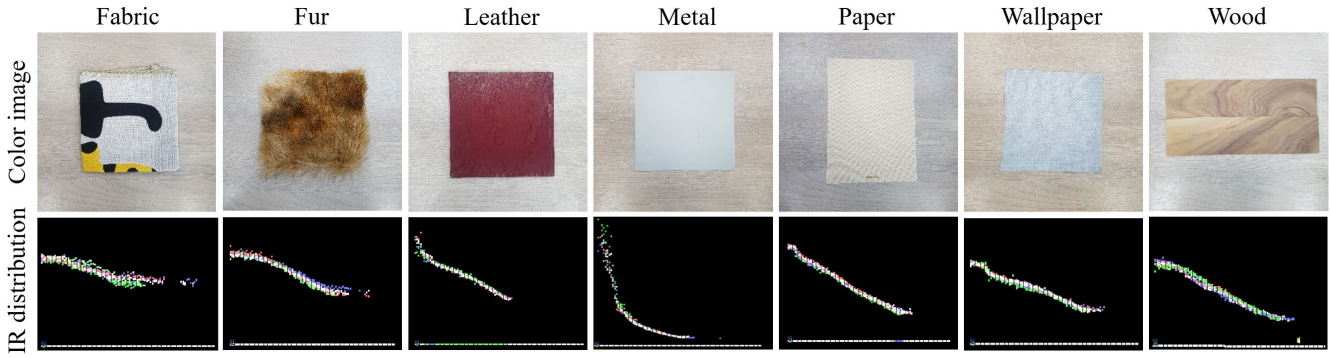


FIGURE 3. Sample of 7 material classes obtained using fixed Kinect v2 camera while rotating material object in front of it.

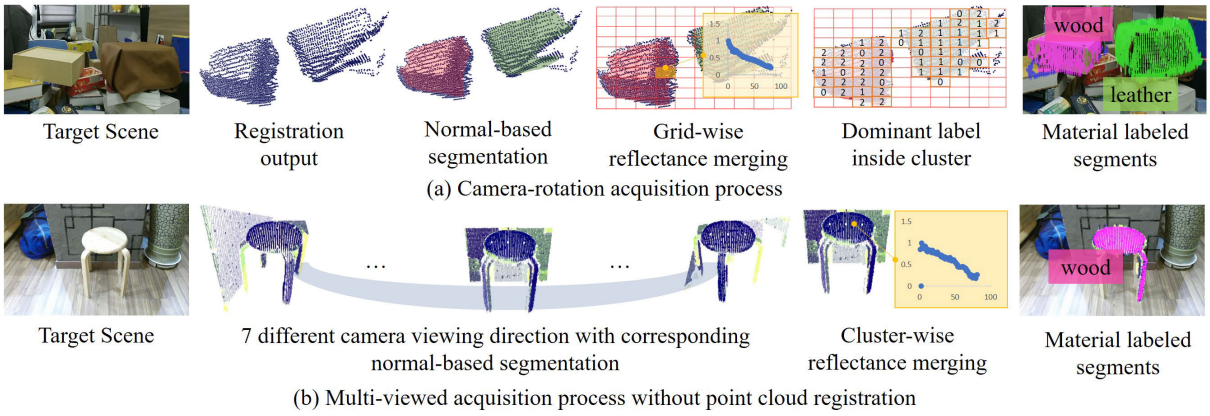


FIGURE 4. Comparison between camera-rotation acquisition process and multi-viewed, unregistered scene acquisition process. (a) Camera-rotation acquisition process. (b) Multi-viewed, unregistered acquisition process with enhanced practicality.

inside same clusters get identical reflectance of single material type. For further enhancement, each cluster finds and allocates dominant material type prediction result out of all pixels inside each segment.

B. MULTI-MODAL MATERIAL RECOGNITION NETWORK

Two-stream nets is widely used for the purpose of fusing multi-modal features. Xue *et al.* [25] have suggested a two-streamed convolutional neural network called DAIN which entangles spatial/angular gradients of material clues. Lee *et al.* [11] use matrix concatenation at the end of the network to fuse reflectance stream and color stream for material recognition. Similar to the previous work, we apply feature fusion at the final layer considering different level of the features. We adopt a SkipRNN [45] network. Lee *et al.* [11] define *partial gate*(p_t) which skips updates whenever noisy data is fed as input, modifying the structure to be noise-robust.

$$s_t = p_t \cdot S(s_{t-1}, x_t) + (1 - p_t) \cdot s_{t-1} \tag{1}$$

$$p_t = \begin{cases} u_t & \text{if } x_t \geq 0.05 \\ 0 & \text{if } x_t < 0.05 \end{cases} \tag{2}$$

Equations show activation conditions of partial gate used in partial skipRNN. The proposed network is trained on our proposed reflectance data set. We use gradients among adjacent sequence points $(x_t - x_{t-1}, x_{t+1} - x_t)$ to find noisy, missing value based on IR distribution continuity. If the obtained gradient is greater than defined threshold (0.5), the intensity value at the point is considered as noise and set to zero. Also, if a point located in small incidence angle has larger intensity value than the threshold, it will also be considered as noise.

Huang *et al.* [44] have suggested several modified versions of DenseNet, which share same dense connectivity strategy. Lee *et al.* [11] introduce fine-tuned model using MINC-2500 [2], enabling the network to be suitable for material recognition task. Along with their proposed two-stream network structure and concatenation based feature fusion, the model is used as a feature extractor of color feature stream. Since the feature of each stream is subject to be biased by the concatenation, it cannot fully be described by both reflectance and color characteristics in balance. Secondly, concatenation cannot guarantee sufficiently encoded correlation between the two features. To get better correlated feature fusion of multi-modal features, recent work [46], [47]

uses outer product fusion for two-stream network instead. By multiplying color and reflectance feature matrix, correlated multi-modal features are obtained.

V. MATERIAL RECOGNITION WITH MULTIPLE COLOR IMAGES

A. MULTI-ILLUMINATION IMAGES

Multi-illumination images in the Wild (MIW) [40] data set consists of more than 1000 scenes with 25 illumination variations and per-pixel material labels. Using a customized camera setup with attached auto-controlled light source, MIW provides 25 images per scene with varying lighting conditions and shared (fixed) scene geometry. Per-pixel material labels consist of 41 material classes including wide range of materials such as *fabric, glass, metal, paper, plastic, marble, linoleum, wicker*. For better representation of real-world environment, photos are captured in 95 different rooms located in 12 different residential and office areas. Since all of the provided photos are taken as HDR images with varying and corresponding light probes, the data set is suitable for the applications like lighting prediction or relighting as they have tried. In our work, we focus on per-pixel material labels provided by the data set. Since MIW provides illumination varying scenes with static scenes, it reveals per-pixel reflectance characteristics of target objects.

B. TWO-STREAM NETS WITH ATTENTION MODULE

Our network consists of two neural networks encoding both color and brightness variation features as shown in Figure 5. In the color feature network (Figure 5 (b)), each patch is fed to Densenet121 fine-tuned with MIW data set. Since each patch has n different illumination conditions, the network encodes color features from all n patches. In the brightness variation network (Figure 5 (c)), we build a new patch set $\hat{P}$ by obtaining difference patch of two consecutive original patches in the patch set P as follows.

$$\hat{p_i} = \begin{cases} p_i - p_n, & i = 1 \\ p_i - p_{i-1}, & \text{otherwise} \end{cases} \tag{3}$$

The patch set consists of n patches with different brightness at the same viewpoint. So the new patch set $\hat{P}$ extracts brightness of original patch that is good for the extraction of brightness variation features. Extracted brightness variation reveals unique characteristic of surface reflectance of each material type. $\hat{P}$ is fed to Resnet34 fine-tuned with MIW data set extracting brightness variation features. We concatenate the color and brightness variation features along the sorted sequence. Each concatenated feature vector f_i has 192 channels.

The concatenated feature vectors $f_1 \sim f_n$ are fed to attention module (Figure 5 (d)) to emphasize features that contribute to class separation in the following Long Short-Term Memory (LSTM) network. Note that the LSTM extracts sequentially varying features from the illumination-varying inputs such as surface reflectance. Our attention module learns channel-wise significance assigning same attention to corresponding feature set ($f_1(j) \sim f_n(j)$) of j-th channel. And then, each feature vector f_i' obtained from corresponding input patches $p_i, \hat{p_i}$ is fed to each unit of the LSTM. Consequently, LSTM classifies material types from color and brightness variation features observed from the illumination-varying input patches.

C. PATCH EXTRACTION AND SORTING

Our network gets multiple images taken at a same location under illumination-varying condition. We hypothesize that the reflectance of material obtained from the observation helps to characterize material type thanks to unique meso-surface and BRDF (Bidirectional Reflectance Distribution Function) characteristics. We randomly extract m patches per scene for patch-wise classification. Multiple material classes exist inside a patch. We set the label of the center pixel of a patch as class label. Figure 5 (a) shows that a patch set P consists of n illumination conditioned patches as follows.

$$P = \{p_i | 1 \leq i \leq n\} \tag{4}$$

After the patches $p_1 \sim p_n$ are obtained, they are sorted along the average pixel-brightness. As a result, we extract m(number of patches) $\times$ n(illumination conditions) patches in a scene.

VI. EXPERIMENTAL RESULTS

A. MATERIAL RECOGNITION FROM COLOR AND DEPTH

Data set used for experimental evaluation includes reflected IR, RGB, vertex normal and 3D points. Incidence angle is calculated from vertex normal [9], [11]. Frame acquisition and calibration functions are implemented based on Kinect v2 SDKs. However, our proposed method can be applied to any other type of imaging device. Based on the acquisition setup in figure 4-(a), we construct multi-object scenes. With test objects, we minimize errors of poor registration and segmentation. Figure 6-(a) shows sample material segmentation results. Proposed feature successfully smooths out the noises of poor registration.

For evaluation comparison, combinations of different network structures are evaluated. One-dimensional *Gradual CNN* consists of 4 convolution filters which is gradually increasing within layers (1×3, 1×5, 1×7, 1×11). For fair comparison with [11] which uses 2-layered RNN based structure, total number of convolution layers of our network is fixed to 4. DenseNet-121 is first trained with MINC-2500 [11]. Table 3 shows comparison results among the combinations of different network structures. In the test with reflectance feature alone, Gradual CNN shows best performance (72.66%) compared to previous work (64.67%). Results indicate that the trained LSTM cells are affected by the noisy inputs coming from previous state. Proposed CNN structure isolates noisy inputs better within each kernel. Applying dilation to the network decreases performance

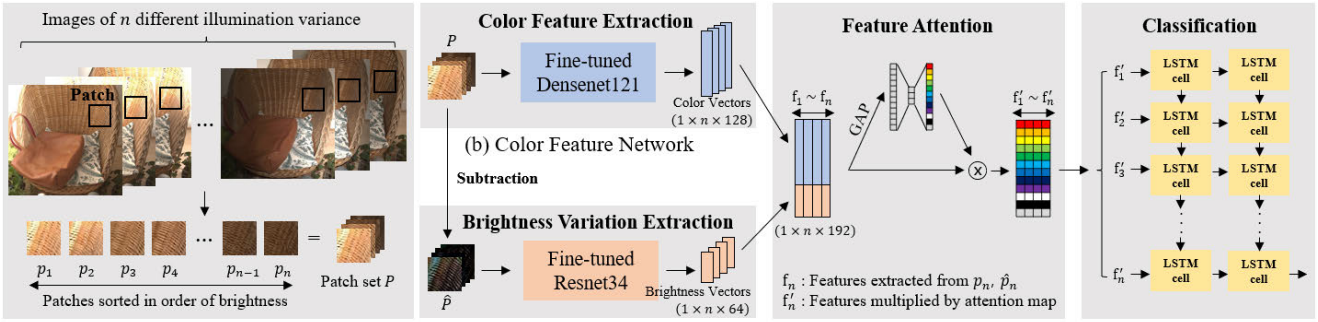


FIGURE 5. Proposed two-stream material type recognition framework. In the multi-illumination views, a patch set of *n* illumination conditions is extracted and is fed to our networks.

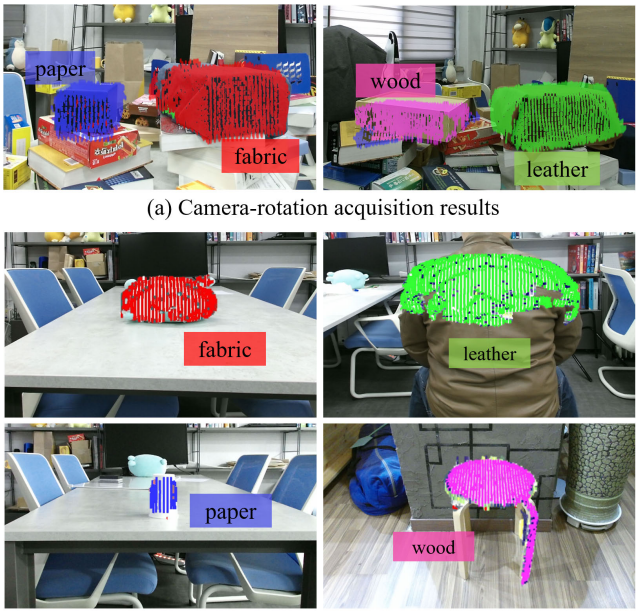


FIGURE 6. Example results of camera-rotation acquisition (figure 4-(a)) and multi-viewed acquisition without point cloud registration (figure 4-(b)).

from 72.66% to 68.67%, even though it is greater than previous work. Performance gain of outer product is 7.34% in two-stream network with partial skipRNN compared to concatenation based method. Outer product fusion encodes features of different modality better. Gradual CNN with outer product shows best performance of 86.00%.

B. 3D MATERIAL-AWARE SEGMENTATION

In this test, we perform 3D point cloud segmentation by material types. A neural network trained with 4 common material classes including (*fabric, leather, paper, wood*) are used. This test is performed in noisy, unconstrained real world conditions compared to the previous experiments. Seven point clouds from seven different viewing directions are acquired without registration. For each segment inside each angle-wise

TABLE 3. Material recognition from color and depth results.

Network	Fusion Method	Data	Top-5 Accuracy
SkipRNN [28]	-	Ref.	$62.67_{\pm 5.5}$
Partial SkipRNN [28]	-		$64.67_{\pm 1.8}$
Gradual 1D CNN	-		$\mathbf{72.66}_{\pm 1.5}$
Dilated 1D CNN	-		$68.67_{\pm 1.8}$
SkipRNN + DenseNet-121 [28]	Concat.	Ref. + Color	$74.67_{\pm 3.0}$
Partial SkipRNN + DenseNet-121 [28]			$76.00_{\pm 4.9}$
Partial SkipRNN + DenseNet-121	Outer product		$83.34_{\pm 4.7}$
Gradual 1D CNN + DenseNet-121			$\mathbf{86.00}_{\pm 4.3}$
Dilated 1D CNN + DenseNet-121			$83.33_{\pm 5.3}$

point clouds(from θ_1 to θ_7), total seven reflectance features are obtained and merged following the refinement process described in figure 4. As illustrated in figure 6, our framework shows meaningful classification performance despite the challenging environmental condition.

C. MATERIAL RECOGNITION FROM MULTIPLE COLORS

Implementation details and results of our experiments using multiple color images are as follows. Our model is implemented with Pytorch framework. ADAM optimizer is used for training and batch size is 32. Learning rate is initialized to 0.001 with learning rate scheduler ReduceLROnPlateau (patience = 10, factor = 0.95). Backbone network is

TABLE 4. Material recognition from multiple colors results.

Model	Input Patch ($1 \le i \le 25$)	Accuracy	
		mean	std
Single Stream (Densenet + LSTM)	$p_1 * 25$ (* 25)	61.58%	2.71
	p_i ( )	71.04%	2.90
Two Stream (Densenet/Resnet + LSTM)	p_i, p_i (,)	72.91%	3.06
	p_i, $\hat{p}_i$ (,)	75.64%	4.21
Two Stream + Attention	p_i, $\hat{p}_i$ (,)	**76.50%**	4.21

initialized with pre-trained models and training is finished after 300 epochs. We use MIW dataset for the evaluation. MIW dataset is randomly divided into 5 splits with 985 training scenes and 30 test scenes. We randomly extract 40(=m) patches per scene. There are 41 classes in the MIW dataset. We group the classes of similar material types into 8 superclasses(Fabric, Glass, Leather, Metal, Paper, Plastic, Stone, and Wood). We perform 5 different experiments (table 4) to verify the effect of multi-illumination conditions, two-stream network, adding new patch set $\hat{P}$, and attention module, respectively. Compared to single-illumination case, the accuracy of multi-illumination is around 9% higher. This shows that the difference in the surface appearance of materials along the illumination conditions improves material type classification. The result of two-stream nets using only patch set P is 1.87% higher than single-stream net. The result of using P for the color feature network and $\hat{P}$ for the brightness feature network is 75.64%, which is 2.73% higher than using only patch set P. This implies that brightness features from the new patch set $\hat{P}$ helps the networks to classify material types better encoding surface reflectance. Finally, attention map shows 0.86% higher accuracy than two-stream networks.

Contextual relation in our material recognition task indicates how often one material type is observed with other material types within a patch. We hypothesize that material types of neighbor regions reveal critical information of the material type of target region. There is trade-off in the extraction of material type features between surface reflectance and color information along the variation of patch size. Each pixel of a patch has its own reflectance characteristic, within a patch, however, integrated reflectance of multiple pixels is obtained. Therefore, what we have obtained is a reflectance of all the pixels within the patch that is called a smoothed reflectance. If the patch size is large, the reflectance of a larger area is smoothed, so it is relatively difficult to define common reflectance. On the other hand, with bigger patch, color and contextual information are extracted better. On the contrary, if the patch size is small, relatively common reflectance label works better, but the extraction of other visual features may be limited. Based on the relationship, we tested increasing patch size. Patch is extracted only when the portion of pixel material label is higher than 65% within the patch. The 'Normal' column of Table 5 shows the test accuracy along the variation of patch size. Compared to the patch sizes 11 × 11 and 21 × 21, larger patch sizes get better performance. However the performance with patch size of 31 × 31 and bigger show saturated performance indicating that no better color and contextual information could be extracted with bigger patches lager than 31 × 31.

1) TRAINING DATA AUGMENTATION

We conduct training data augmentation. The key point is to keep contextual relation in a patch, when more than one class type exist in the patch. Since the class label of major region in a patch is material label of the patch, our augmentation keeps the region and replace remaining minor region by other patch of same minor class labels. To preserve contextual information, we create a new minor part of the same material class by randomly finding it from another training image. Figure 7 shows an example of applying the proposed augmentation. Figure 7-(a) is a patch whose class label is paper, and the minor part on the left top is bright-colored wood. Figure 7-(b) shows a patch that the minor part, wood, is replaced by a dark-colored wood from other training image. Table 5 shows that the best accuracy is obtained with 31 × 31 patch augmentation. When the patch size is small, there is little effect of augmentation because the contextual information contained within the patch is limited. When the patch size is increased, patch is extracted only when the pixel corresponding to the class label is 65% or more in the patch, so the surface diversity of the extracted patch is reduced. This condition, however, reduces the diversity of the data set.

TABLE 5. Experimental results along the variation of patch size and augmentation.

Patch Size	Test Accuracy				
	normal		augmentation		
	mean	std	mean	gain	std
11x11	59.28%	1.76	58.83%	-0.45	2.85
21x21	60.66%	1.23	61.55%	+0.89	1.51
31x31	63.64%	3.69	**67.39%**	**+3.75**	1.05
41x41	62.31%	4.47	63.42%	+1.11	1.44
51x51	63.43%	1.75	63.31%	-0.12	1.73

(a) Normal Patch (b) Patch applied augmentation

FIGURE 7. Example of patch with minor part replacement applied.

VII. CONCLUSION

In this work, we propose a material type recognition method of indoor scenes via surface reflectance estimation. Novel two-stream network extracting both reflectance and color features obtain pixel wise material type from objects of indoor real-world environment. Diverse experimental evaluations on public data set prove that conventional color features with reflectance feature outperforms prior approaches.

REFERENCES

[1] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson, "Recognizing materials using perceptually inspired features," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 348–371, Jul. 2013, doi: 10.1007/s11263-013-0609-0.

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3479–3487.

[3] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. Conf. CVPR*, Jun. 2015, pp. 3828–3836.
[4] J. Degol, M. Golparvar-Fard, and D. Hoiem, "Geometry-informed material recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1554–1562.
[5] S. Lee, J. Kim, H. Lim, and S. Ahn, "Surface reflectance estimation and segmentation from single depth image of ToF camera," *Signal Process. Image Commun.*, vol. 47, pp. 452–462, Sep. 2016.
[6] S. Lee and S. Lee, "Surface IR reflectance estimation and material recognition using ToF camera," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1554–1562.
[7] S. Su, F. Heide, R. Swanson, J. Klein, C. Callenberg, M. Hullin, and W. Heidrich, "Material classification using raw time-of-flight measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3503–3511.
[8] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 79–88.
[9] J. Kim, H. Lim, S. C. Ahn, and S. Lee, "RGBD camera based material recognition via surface roughness estimation," in *Proc. IEEE-WAVC*, Mar. 2018, pp. 1963–1971.
[10] C. Zhao, L. Sun, and R. Stolkin, "A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition," in *Proc. 18th Int. Conf. Adv. Robot. (ICAR)*, Jul. 2017, pp. 75–82.
[11] S. Lee, H. Lim, S. Ahn, and S. Lee, "IR surface reflectance estimation and material type recognition using two-stream net and Kinect camera," in *Proc. ACM SIGGRAPH Posters*, Jul. 2019, pp. 1–2.
[12] J. Gu and C. Liu, "Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral BRDF," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 797–804.
[13] S. Georgoulis, V. Vanweddingen, M. Proesmans, and L. V. Gool, "Material classification under natural illumination using reflectance maps," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 244–253.
[14] C. Liu, G. Yang, and J. Gu, "Learning discriminative illumination and filters for raw material classification with optimal projections of bidirectional texture functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1430–1437.
[15] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 708–717.
[16] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
[17] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 143–156.
[18] J. Xue, H. Zhang, and K. Dana, "Deep texture manifold for ground terrain recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 558–567.
[19] W. Zhai, Y. Cao, J. Zhang, and Z.-J. Zha, "Deep multiple-attribute-perceived network for real-world texture recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3613–3622.
[20] W. Zhai, Y. Cao, Z.-J. Zha, H. Xie, and F. Wu, "Deep structure-revealed network for texture recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11007–11016.
[21] O. G. Cula and K. J. Dana, "3D texture recognition using bidirectional feature histograms," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 33–60, Aug. 2004.
[22] M. Weinmann and R. Klein, "Material recognition for efficient acquisition of geometry and reflectance," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 321–333.
[23] T.-C. Wang, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "SVBRDF-invariant shape and reflectance estimation from light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5451–5459.
[24] T. Wang, J. Zhu, H. Ebi, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," *CoRR*, vol. abs/1608.06985, pp. 1–16, Aug. 2016.
[25] J. Xue, H. Zhang, K. J. Dana, and K. Nishino, "Differential angular imaging for material recognition," in *Proc. CVPR*, 2017, pp. 6940–6949.
[26] B. Guo, J. Wen, and Y. Han, "Deep material recognition in light-fields via disentanglement of spatial and angular information," in *Proc. ECCV*, vol. 12369, 2020, pp. 664–679, doi: 10.1007/978-3-030-58586-0_39.
[27] S. Lombardi and K. Nishino, "Reflectance and natural illumination from a single image," in *Proc. ECCV*, 2012, pp. 582–595.
[28] J. Filip, R. Vávra, M. Haindl, P. id, M. Krupika, and V. Havran, "BRDF slices: Accurate adaptive anisotropic appearance acquisition," in *Proc. CVPR*, Jun. 2013, pp. 1468–1473.
[29] H. Zhang, K. Dana, and K. Nishino, "Reflectance hashing for material recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3071–3080.
[30] C. Li, Y. Monno, H. Hidaka, and M. Okutomi, "Pro-cam SSfM: Projector-camera system for structure and spectral reflectance from motion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2414–2423.
[31] S. Bi, Z. Xu, K. Sunkavalli, D. Kriegman, and R. Ramamoorthi, "Deep 3D capture: Geometry and reflectance from sparse multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5960–5969.
[32] S. Bi, Z. Xu, K. Sunkavalli, M. Hasan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi, "Deep reflectance volumes: Relightable reconstructions from multi-view photometric images," in *Proc. ECCV*, 2020, pp. 294–311.
[33] C. Schmitt, S. Donne, G. Riegler, V. Koltun, and A. Geiger, "On joint estimation of pose, geometry and svBRDF from a handheld scanner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3493–3503.
[34] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim, "Practical SVBRDF acquisition of 3D objects with unstructured flash photography," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–12, 2018, doi: 10.1145/3272127.3275017.
[35] M. Innmann, J. Susmuth, and M. Stamminger, "BRDF-reconstruction in photogrammetry studio setups," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3346–3354.
[36] L.-P. Asselin, D. Laurendeau, and J.-F. Lalonde, "Deep SVBRDF estimation on real materials," in *Proc. Int. Conf. 3D Vis. (3DV)*, Nov. 2020, pp. 1157–1166.
[37] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan, "Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials," *IEEE Trans. Image Process.*, vol. 29, pp. 4159–4173, 2020.
[38] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz, "Neural inverse rendering of an indoor scene from a single image," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8598–8607.
[39] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2475–2484.
[40] L. Murmann, M. Gharbi, M. Aittala, and F. Durand, "A dataset of multi-illumination images in the wild," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4080–4089.
[41] Z. Li, K. Sunkavalli, and M. Chandraker, "Materials for masses: SVBRDF acquisition with a single mobile phone image," in *Proc. ECCV*, Sep. 2018, pp. 72–87.
[42] D. Gao, X. Li, Y. Dong, P. Peers, K. Xu, and X. Tong, "Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–15, Jul. 2019, doi: 10.1145/3306346.3323042.
[43] J. M. Jurado, J. L. Cárdenas, C. J. Ogayar, L. Ortega, and F. R. Feito, "Semantic segmentation of natural materials on a point cloud using spatial and multispectral features," *Sensors*, vol. 20, no. 8, p. 2244, Apr. 2020.
[44] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, pp. 1–9, Aug. 2016.
[45] V. Campos, B. Jou, X. I. Giró Nieto, J. Torres, and S. Chang, "Skip RNN: Learning to skip state updates in recurrent neural networks," *CoRR*, vol. abs/1708.06834, pp. 1–10, Aug. 2017.
[46] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.
[47] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proc. CVPP*, Jun. 2019, pp. 6639–6648.

SEOKYEONG LEE received the B.S. and M.S. degrees in computer science and engineering from Kyung Hee University, Yongin, South Korea, in 2018 and 2020, respectively. She is currently pursuing the Ph.D. degree in computer science with the Academy-Research Industry Collaboration Program, Korea University and the Korea Institute of Science and Technology, Seoul, South Korea. In Kyung Hee University, she has involved in the research on ToF camera-based surface reflectance estimation and material recognition as an Undergraduate Researcher and an M.S. Student. She is also participating in an advanced research related to plenoptic setup-based lighting and reflectance estimation with the Korea Institute of Science and Technology. Her research interests include physically motivated scene understanding, intrinsic decomposition, inverse rendering, and 3D-based image/view synthesis.

DONGJIN LEE received the B.S. degree in electrical engineering from Pusan National University, Busan, South Korea, in 2018. His research interests include material type classification, texture recognition, computer vision, and deep neural networks.

HYUN-CHEOL KIM received the B.S. and M.S. degrees in electronics engineering from Kyung Hee University and the Ph.D. degree from Chungnam National University, South Korea. In 2000, he joined ETRI, Daejeon, South Korea, where he is currently a Principal Researcher. He has been engaged in the development of immersive media systems and gaze tracking systems. His research interests include immersive media processing and UI/UX.

SEUNGKYU LEE received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1997 and 1999, respectively, and the Ph.D. degree in computer science and engineering from The Pennsylvania State University, in 2009. He has been a Research Engineer with the Technical Research Institute, Korean Broadcasting System, Seoul, South Korea, where he has been involved in the research on high-definition image processing, MPEG4 advanced video coding, and the standardization of terrestrial-digital mobile broadcasting. He has also been a Principal Research Scientist with the Advanced Media Laboratory, Samsung Advanced Institute of Technology, Yongin, South Korea. He is currently an Associate Professor with Kyung Hee University. His research interests include time-of-flight depth camera, color/depth image processing, symmetry-based computer vision, and 3-D modeling and reconstruction.