A Novel Deep-learning Pipeline for Light Field Image Based Material Recognition

Yunlong Wang
CRIPAC, NLPR, CASIA
No.95 ZhongGuanCun East Street,
Beijing, China, 100190
Email: yunlong.wang@cripac.ia.ac.cn

Kunbo Zhang
CRIPAC, NLPR, CASIA
No.95 ZhongGuanCun East Street,
Beijing, China, 100190
Email: kunbo.zhang@ia.ac.cn

Zhenan Sun
CRIPAC, NLPR, CASIA
No.95 ZhongGuanCun East Street,
Beijing, China, 100190
Email: znsun@nlpr.ia.ac.cn

***Abstract*—The primitive basis of image based material recognition builds upon the fact that discrepancies in the reflectances of distinct materials lead to imaging differences under multiple viewpoints. LF cameras possess coherent abilities to capture multiple sub-aperture views (SAIs) within one exposure, which can provide appropriate multi-view sources for material recognition. In this paper, a unified "Factorize-Connect-Merge" (FCM) deep-learning pipeline is proposed to solve problems of light field image based material recognition. 4D light-field data as input is initially decomposed into consecutive 3D light-field slices. Shallow CNN is leveraged to extract low-level visual features of each view inside these slices. As to establish correspondences between these SAIs, Bidirectional Long-Short Term Memory (Bi-LSTM) network is built upon these low-level features to model the imaging differences. After feature selection including concatenation and dimension reduction, effective and robust feature representations for material recognition can be extracted from 4D light-field data. Experimental results indicate that the proposed pipeline can obtain remarkable performances on both tasks of single-pixel material classification and whole-image material segmentation. In addition, the proposed pipeline can potentially benefit and inspire other researchers who may also take LF images as input and need to extract 4D light-field representations for computer vision tasks such as object classification, semantic segmentation and edge detection.**

I. Introduction

Material recognition is crucial in scene understanding, which can be widely applicable in human-computer interaction, robotics, autonomous driving and so on. Since illuminations are typically unknown, it is quite intractable to distinguish the material type by measuring the surface reflectance or the bidirectional reflectance distribution function (BRDF). Hence, prior works on 2D material recognition [1]–[8] mostly depend on image appearance and shape cues, either classifying instance-level textures or exploiting category-level properties. Lack of multi-view analysis, these approaches usually fail to generate plausible predictions when only parts of the objects appear in the image or the materials look similar in colors and textures. For instance, flowers printed on a paper may be mistakenly categorized into the category of foliage.

An alternative way for material recognition is to consider the discrepancies in the reflectance of distinct materials under multiple viewpoints [9]. With additional optical components like the micro-lens array inserted between the main lens and the image sensor [10], light field (LF) cameras are capable of capturing both intensity and direction information of rays from real-world scenes. Hence, LF cameras possess coherent abilities to capture multiple sub-aperture images (SAIs) within one exposure [11], which can provide appropriate multi-view sources for material recognition.

Recently, Wang *et al.* [12] introduce a mid-size light-field image dataset captured by Lytro Illum camera, which contains 12 categories of materials, each with 100 images labeled with per pixel ground truth. In order to learn 4D representations that are compatible with 2D CNN models, they investigate several novel CNN architectures which take remapped light-field images as input. By aggregating multi-view information and reusing the spatial filters from previous 2D models, the best-performing Angular Filter structure in [12] boosts the accuracy of single-pixel material classification by 7% compared with single-view 2D prediction. The trained patch model is altered to a fully convolutional model, which performs material segmentation on an entire image. This sets an important baseline for the tasks of light-field image based material recognition.

Considering these CNN architectures, the basic idea is to construct *4D-to-2D adaptors*. In other words, the original light-field images are firstly passed through such adaptors which can aggregate the angular information and adapt input 4D data to 2D pre-trained networks. Take Angular Filter method in [12] for example, the kernel size of the angular filter itself is the same as angular resolution of remapped light-field images, and the stride is consistent with its kernel size. So after passing this layer, the 4D input reduces to the same spatial size as 2D single-view input. The aggregated angular information is then sent through the sequent 2D pre-trained network. Besides, this mechanism prefer to combine information from different views at a lower level. In brief, this series of methods with *4D-to-2D adaptors* transforms extracting 4D light-field representations to multi-view aggregation and 2D feature extraction.

However, we state that multi-view analysis for material recognition needs to act more like "differentiating" rather than "aggregating" in connecting the inconsistencies between sub-aperture images (SAIs). In this regard, we convert from approaches with 4D-to-2D adaptors to a novel "Factorize-Connect-Merge" (FCM) deep-learning pipeline to solve prob-

lems of light field image based material recognition. Specifically, 4D light-field data as input is initially decomposed into consecutive 3D light-field slices. Rather than extracting high-level semantic features with deep networks, shallow CNN without any pooling operations is leveraged to extract low-level visual features of each view inside these slices. Instead of combining multiple views to aggregate the angular information, we adopt Bidirectional long-short term memory (Bi-LSTM) network to differentiate the imaging differences and connect the inconsistencies between these SAIs. After feature selection including concatenation and dimension reduction, effective and robust feature representations for material recognition can be extracted from 4D light-field data. Experimental results indicate that the proposed framework can obtain remarkable performances on both tasks of single-pixel material classification and whole-image material segmentation. Compared with the best-performing Angular Filter method in [12], the accuracy of single-pixel material classification obtained by the proposed framework is 9% higher averaged on 12 categories of materials in the datasets. In addition, the proposed framework can potentially benefit and inspire other researchers who may also take LF images as input and need to extract 4D light-field representations for computer vision tasks such as object classification, semantic segmentation and edge detection.

II. Related Work

A. Image-based Material Recognition

In the field of classical 2D material recognition, several databases are published in the literature such as CUReT [1], KTH-TIPS [2], Flickr Material Database (FMD) [3] and Materials in Context Database (MINC) [8]. Among these databases, MINC [8] is a large-scale dataset with 3 million patches well-sampled across 23 categories of materials, where both tasks of classifying materials from cropped 2D patches and material segmentation in full images are tested on. Liu *et al.* [4] propose a Bayesian generative framework to fuse low- and mid-level features. For the sake of avoiding object-specific information, visual material traits is introduced by Schwartz and Nishino [13]. In subsequent work [14], they utilize partial supervision to discover locally-recognizable material attributes from crowdsourced perceptual material distances. Cimpoi *et al.* [7] exploit object-oriented and texture-oriented features which obtains superior results on FMD. However, these approaches working on 2D material recognition highly rely on image appearance and contextual information such as shape cues, and usually fail to generate plausible predictions when only parts of the objects appear in the image or the materials look similar in colors and textures.

LF cameras are able to capture a scene from multiple views in a single photographic exposure, which supplies the data source to measure the discrepancies in the reflectances of distinct materials by multi-view analysis. Recently, Wang *et al.* [12] capture a mid-size light field datasets using the Lytro Illum camera. This dataset contains 12 categories of materials, including *Fabric*, *Foliage*, *Fur*, *Glass*, *Leather*, *Metal*, *Plastic*, *Paper*, *Sky*, *Stone*, *Water* and *Wood*, along with extra pixels categorized into *Other*. Each class has 100 images, manually classified and labeled the images with single-pixel material category. Several novel CNN architectures are investigated so as to reuse the spatial filters from previous 2D models, consisting of *2D average*, *Viewpool*, *Stack*, *EPI*, *Angular filter on remap image*, *4D filter*. The Angular Filter method achieves the best performance among these architectures, which obtains about 7% gain compared to using 2D input on the task of single-pixel material classification. Similar to the training strategy on 2D images in [8], the Angular Filter model is firstly trained on image patches and then converted to a fully convolutional model, finally fine-tuned on entire images. The datasets and CNN architectures act as good baselines for light-field image based material recognition. However, these CNN structures are specially designed to learn 4D representations that are compatible with 2D CNN models. The *4D-to-2D adaptors* just as Angular Filter layer are constructed to aggregate the angular information from the SAIs of input light field images. Rather, we propose to differentiate the imaging differences under multiple viewpoints by connecting the inconsistencies between SAIs with Bi-LSTM. Meanwhile, low-level visual features of each view is extracted using shallow CNN without any pooling layers.

B. Deep-learning Approaches Working on LF Data

Unlike the pipelines dealing with 2D matrix *e.g.* images or 3D volume *e.g.* videos, it is quite a sticky problem to manipulate the high-dimensional LF data with plain CNN currently. To our best knowledge, Yoon *et al.* [15], [16] are the first to apply CNN framework to the research of light field super resolution (LFSR). They propose a new deep learning structure called LFCNN composed of a spatial SR network and an angular SR network to jointly increase the spatial and angular resolution. Wang *et al.* [17] propose LFNet and incorporate implicitly multi-scale fusion scheme into bidirectional recurrent convolutional neural network. Zhang *et al.* [18] propose a learning-based method using residual convolutional networks on stacked views to reconstruct light fields with higher spatial resolution. Yeung *et al.* [19] approximate 4D convolution with spatial-angular separable convolutions for extraction of spatial-angular joint features.

Kalantari *et al.* [20] propose the first deep learning system for light field reconstruction. Srinivasan *et al.* [21] build on the pipeline in [20] to synthesize a 4D RGBD LF from a single 2D RGB image. Jin *et al.* [22] focus on generating densely-sampled LFs with sparsely- and arbitrarily-sampled sub-aperture images (SAIs), which relieve the restriction on the regularity of the sampling pattern. Wu *et al.* [23] propose EPICNN and model view synthesis as learning-based angular detail restoration on 2D epipolar plane images (EPIs). Yeung *et al.* [24] adopt a coarse-to-fine strategy in a deep-learning framework to exploit the high dimensional spatial-angular clues inside 4D light field data. In [25], Wang *et al.* propose Pseudo 4DCNN which are assembled by 2D strided convolutions operated on stacked EPIs and detail-restoration

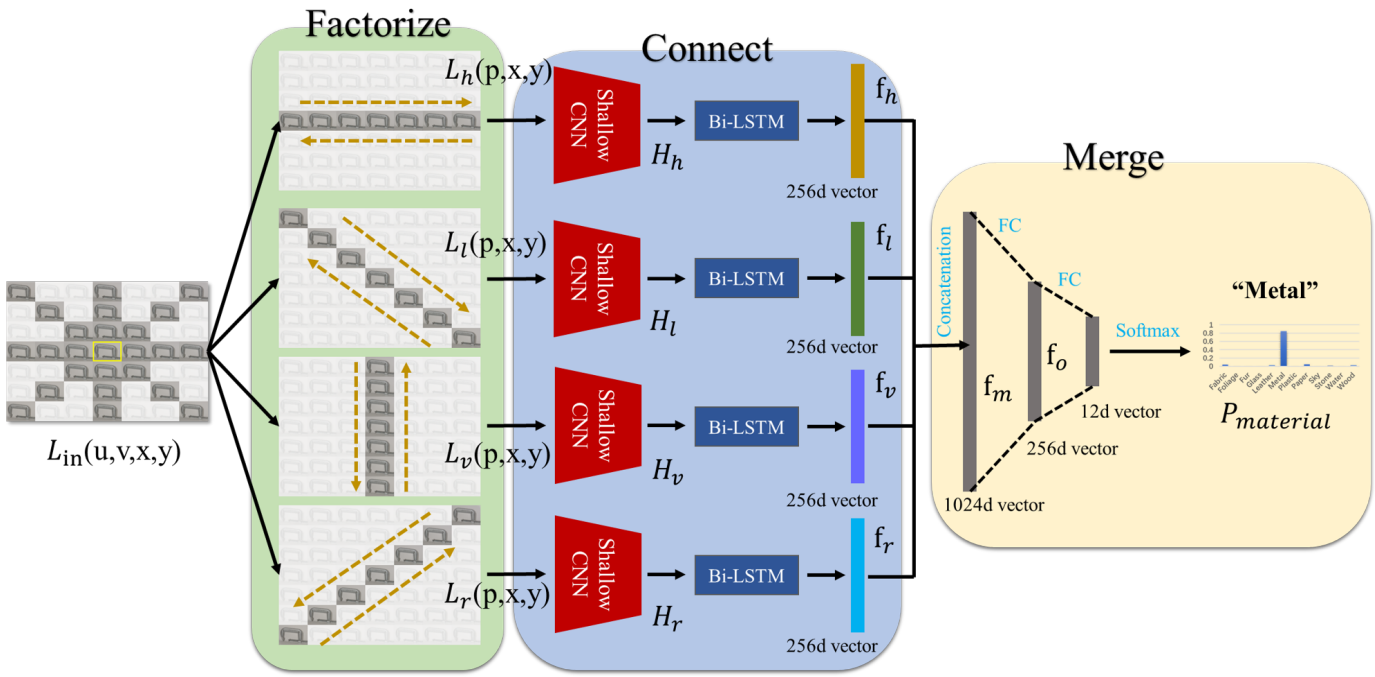

Fig. 1. Schematic of the novel pipeline (Factorize-Connect-Merge, FCM) for light field image based material recognition. Specifically, the process that a cropped 4D patch is categorized into the category of *Metal* is shown.

3D CNNs connected with angular conversion. Wu *et al.* [26] decompose the task of light field reconstruction as learning sheared EPI structure. Heber *et al.* [27] learns an end-to-end mapping between the 4D light field and a representation of the corresponding 4D depth field in terms of 2D hyperplane orientations. Further, Heber *et al.* [28] propose a U-shaped fully convolutional network that involves an encoding and a decoding part for shape from light field. Shin *et al.* [29] also adopts the structure of fully-convolutional neural network and introduce EPINET to estimate depth from light field images.

To sum up, the aforementioned deep-learning approaches working on LF data are proposed for the purpose of spatial super resolution [15]–[19], view synthesis [20]–[26] and depth estimation [27]–[29]. Except several CNN architectures proposed in [12], there are scarcely deep-learning frameworks tailored for material recognition currently. In this paper, we propose a novel "Factorize-Connect-Merge" (FCM) deep-learning pipeline to solve problems of light field image based material recognition.

III. Proposed Method

The details of the proposed pipeline, *i.e.* "Factorize-Connect-Merge" (FCM) deep-learning is described in this chapter. This pipeline is proposed to solve problems of light field image based material recognition in this paper. Specifically, 4D light-field data as input is initially decomposed into consecutive 3D light-field slices. Shallow CNN without any pooling layers is leveraged to extract low-level visual features of each view inside these slices. Next, we adopt Bidirectional long-short term memory (Bi-LSTM) network to differentiate the imaging differences and connect the inconsistencies between these SAIs. After feature selection including concatenation and dimension reduction, effective and robust feature representations for material recognition can be extracted from 4D light-field data. Inspired by the procedures in [8] and [12], we also train a patch model which classifies single-pixel categories of material firstly and then convert the patch model to a whole-image model which performs material segmentation.

A. Factorize 4D LF data

Throughout this chapter, 4D LF data as input are denoted as $L_{in}(u, v, x, y)$ decoded from the original LF image with the resolution $N \times N \times H \times W$. It should be noted that N is supposed to be odd here. For cropped 4D patches to train the patch model, the angular resolution is the same as full 4D LF data but the spatial resolution is $S \times S$, where S is the pre-set patch size. Based on the disparity-shift relations to central SAI, 4D light-field data is initially decomposed into consecutive 3D light-field slices. Firstly, horizontal 3D light-field data $L_h(p, x, y)$ is denoted as

$$L_h(v^*, x, y) = L_{in}(u^*, v^*, x, y) \tag{1}$$

where $u^* = floor(N/2) + 1$ and $v^* \in \{1, 2, ..., N\}$. Left-diagonal 3D light field data $L_l(p, x, y)$ is denoted as

$$L_l(u^*, x, y) = L_{in}(u^*, v^*, x, y) \tag{2}$$

where $(u^*, v^*) \in \{(1, 1), (2, 2), ..., (N, N)\}$. Vertical 3D light field data $L_v(p, x, y)$ is denoted as

$$L_v(u^*, x, y) = L_{in}(u^*, v^*, x, y) \tag{3}$$

where $u^* \in \{1, 2, ..., N\}$ and $u^* = floor(N/2) + 1$. Right-diagonal 3D light field data $L_r(p, x, y)$ is denoted as

$$L_r(u^*, x, y) = L_{in}(u^*, v^*, x, y) \tag{4}$$

where $(u^*, v^*) \in \{(1, N), (2, N-1), ..., (N, 1)\}$.

As shown in Fig.1, disparity shift between adjacent SAIs inside these 3D light-field slices are constant as they are seen through regularly-spaced portions of the main lens inside the LF camera despite that the affiliated epipolar planes are different from one another. From this point of view, these fractions of the original 4D LF data are more appropriate to discern the imaging differences and intensity variations under multiple views, which is also more efficient than feeding all the SAIs into the network. Moreover, 3D light-field slices can be bidirectionally analyzed and modelled, *i.e.* in forward and backward manners.

B. Connect multi-view inconsistency

It is mainly divided into two steps to connect the inconsistencies between SAIs inside 3D light-field slices. The first step is to extract low-level visual features of each view. The cropped 4D patches for single-pixel material classification only contain a portion of the contextual information around the central pixel to be classified in each SAI. Considering this observation, deep CNNs are not utilized to extract mid- or high-level semantic features, like object classes, shapes and so on.

The structure of shallow CNN adopted in the proposed framework is depicted as Fig.2, consisting of four convolutional layers and one flatten layers. The kernel size of all the convolutional layers is 5×5 with stride 2. Note that all the convolutional layers are activated by the rectified linear unit (ReLU), *i.e.* $\sigma(x) = \max(0, x)$, while the last layer flattens the feature map of the penultimate layer into a vector.

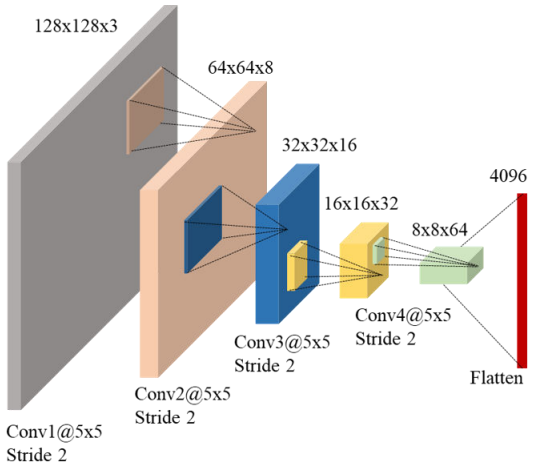

Fig. 2. The structure of shallow CNN.

Noticeably, there are no pooling layers inside the shallow CNN. As known, pooling operations can benefit deep CNNs in extracting semantic features invariant to color jittering, scales, translations and rotations, which is of vital importance in high-level vision tasks, like object recognition. Instead, the substantial features for material recognition such as imaging differences and intensity variations may be wiped away to large extent by max-pooling or average-pooling. To preserve the pixel-level inconsistencies, we substitute pooling operations with stride 2 to downsample the input feature maps.

After extracting low-level vision textures from each view, they are organized into sequence as Eq.5, where $F_{sc}(\cdot)$ and Θ denotes the operators and parameters of the shallow CNN, p_t denotes the patches or images from each view, $t \in \{1, 2, ..., N\}$.

$$H_{L(p,x,y)} = \{F_{sc}(p_1, \Theta), F_{sc}(p_2, \Theta), ..., F_{sc}(p_N, \Theta)\} \tag{5}$$

Next, $H_{L(p,x,y)}$ is sent through a Bi-LSTM [30] to connect the inconsistencies between SAIs as Eq.6.

$$f_{L(p,x,y)} = F_{bilstm}(H_{L(p,x,y)}, \Phi) \tag{6}$$

where $F_{bilstm}(\cdot)$ and Φ are the operators and parameters of the adopted Bi-LSTM, $f_{L(p,x,y)}$ is the feature vector representing one 3D light field slice.

Typically, Bi-LSTM is applied in the fields of text generation [31], visual question answering [32] and video-based action recognition [33] for the sake of sequence prediction and spatio-temporal feature extraction. Here, we employ Bi-LSTM to selectively "remembering" and "forgetting" the discernable inconsistencies caused by reflectance differences between adjacent (short-term) and interval-spaced (long-term) SAIs.

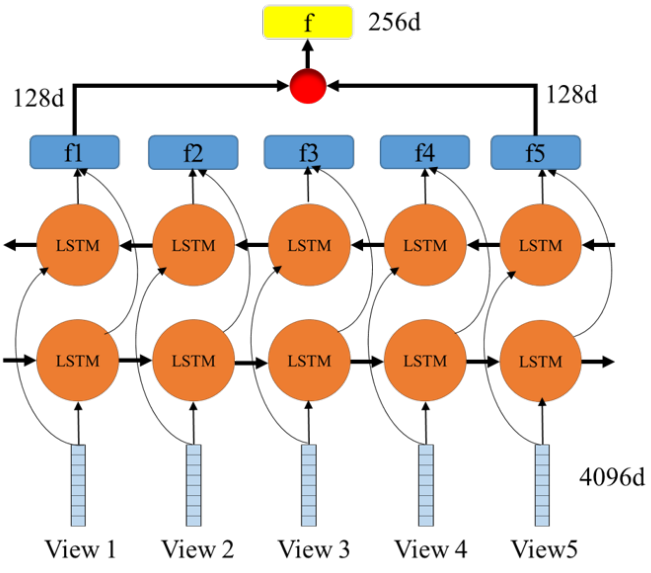

Fig. 3. The structure of specified Bi-LSTM in detail. Take 5 SAIs in 3D light field slices for example.

The structure of specified Bi-LSTM in detail is drawn in Fig.3. The low-level feature of each view is put into the LSTM unit in order as x_t of each time-step. The forward and backward path models the inconsistencies of imaging differences and outputs the hidden state as a 128d vector. The hidden state of the first and last time-step are chosen as the representation and concatenated together to yield a 256d vector $f_{L(p,x,y)}$, while the hidden states of other time-steps are neglected. $\{f_h, f_l, f_v, f_r\}$ represents feature vector for $L_h(p,x,y)$, $L_l(p,x,y)$, $L_v(p,x,y)$, $L_r(p,x,y)$ respectively.

It should be noted that the shallow CNN and Bi-LSTM in the proposed framework are weight-sharing and jointly-learning.

C. Merge 4D LF representations

In the last step of the proposed framework, we merge the feature vectors $\{f_h, f_l, f_v, f_r\}$. Firstly, they are concatenated together to form a 1024d vector f_m as Eq.7.

$$f_m = concat(f_h, f_l, f_v, f_r) \tag{7}$$

where the operator $concat(\cdot)$ means concatenation. Next, a Fully Connected (FC) is followed for feature selection and dimension reduction, which aims to compress 4 groups of feature vectors into one and finds the most discriminative feature. This FC layer outputs a 256d vector f_o which is the feature vector for material recognition. Finally, f_o is connected to another FC layer and Softmax layer, which yields a 12d vector $P_{material}$ with each dimension denoting the probability belonging to one category of material.

Inspired by the strategy in [8] and [12] when converting a trained patch model to a FCN model on full images, we remove the flatten layer in the shallow CNN. Besides, the Bi-LSTM is modified to similar ConvLSTM which takes sequences of feature maps as input rather than sequences of vectors in 4D patch model. The model for whole-image material segmentation is fine-tuned by reusing the weights of filters in the unchanged layers of pre-trained patch model.

D. Datasets and Training Details

To train the proposed framework, we utilize the light field material recognition datasets released by Wang *et al.* [12]. This dataset comprises 12 categories of materials and 100 LF images are captured with Lytro Illum camera for each category. Among these LF images, the material belonging to its category occupy large portions of the scene. The spatial resolution is 376×541 and angular resolution is 14×14. Every pixel in central view of these LF images is manually annotated with its material category.

For fair comparisons, we follow the same protocol in [12] to generate the training datasets. For single-pixel material classification, we crop 128×128 spatial region containing contextual information from the same position of 7×7 SAIs. In total, over 3×10^4 samples with the resolution $128 \times 128 \times 7 \times 7$ are extracted from original LF images. It should be noted that such 4D patch can be regarded as a valid sample only if more than 50% of the pixels belong to the same category of materials. We follow the same protocol in [12] to split the training and test datasets. To separate similar or duplicate samples, all 4D patches cropped from the same LF image only appear either in training or in test set.

The optimization of end-to-end training is conducted by the mini-batch Adam [34] stochastic optimization method with a batch size of 64. The learning rate is initially set to 1×10^{-3} and then decreased by a factor of 0.1 every 10 epochs until the validation loss converges, β_1 to 0.9, β_2 to 0.99. The filters of shallow CNNs are initialized from a zero-mean Gaussian distribution with standard deviation 0.01 and all the bias are initialized to zero. In the structure of Bi-LSTIM, rectified linear unit (ReLU) is employed as activation function for the convolutional layers and hard sigmoid is chosen as activation function for the recurrent step. The weights for the linear transformation of inputs and the recurrent state are initialized from Xavier [35] distribution. Except that the bias of inner forget gate is set to 1, other bias are set to 0. The proposed framework FCM is implemented using Keras package [36] and runs on a workstation with an Intel 3.6 GHz CPU and a TiTan X GPU. It takes less than 50 epochs to converge.

IV. Experimental Results

To validate the effectiveness of the proposed framework, we compare with MINC VGG model [8] fine-tuned on 2D central view of 4D patches in the datasets and the best-performing Angular Filter method in [12]. The source codes of these methods are searched online or supplied by the authors. Moreover, the released models are carefully re-trained, attempting to obtain the maximum performances for comparisons.

A. Single-pixel Material Classification

The quantitative results of single-pixel material classification on the 4D patch datasets is tabulated in Table.I. As can be seen, the Angular Filter method [12] increases the accuracy by 6% on average compared with MINC VGG model trained on 2D central view of these 4D patches. It is mainly due to the property that Angular Filter [12] discriminates multi-view imaging differences by convolving on macro-pixels grouped by a set of pixels from the same position of each SAI and then transferring to traditional 2D pixel-level material prediction.

TABLE I
ACCURACY ON SINGLE-PIXEL MATERIAL CLASSIFICATION.

Material Classes	MINC VGG [8]	Angular Filter [12]	Ours
Fabric	0.59	0.65	**0.88(+0.23)**
Foliage	0.88	0.92	**0.97(+0.05)**
Fur	0.7	0.78	**0.82(+0.04)**
Glass	0.69	0.65	**0.83(+0.18)**
Leather	0.8	0.91	**0.92(+0.01)**
Metal	0.66	0.73	**0.82(+0.09)**
Paper	0.49	0.6	**0.82(+0.22)**
Plastic	0.45	0.5	**0.71(+0.21)**
Sky	0.98	0.98	**0.95(-0.03)**
Stone	0.81	0.87	**0.92(+0.05)**
Water	0.87	0.92	**0.94(+0.02)**
Wood	0.62	0.73	**0.85(+0.08)**
Average	0.71	0.77	**0.86(+0.09)**

The proposed pipeline FCM is distinct to the former approaches that need adaptors like Angular Filter [12] to transform from 4D LF data to conventional 2D prediction. Rather, FCM adopts shallow CNNs to extract low-level visual features and Bi-LSTM to elaborately model the imaging differences between adjacent SAIs. Attributing to these properties, FCM gains significant improvements over Angular Filter [12] and boosts the accuracy of single-pixel material classification by 9% on average. Especially, the accuracy of classification results on some confusing materials rises by a large margin, such as *Fabric* (+23%), *Paper* (+22%) and *Plastic* (+21%). It is noticed that the accuracy on the category of *Sky* is a bit lower (-3%) than Angular Filter [12]. The underlying reasons are mainly because that the samples of *Sky* in the datasets are captured under various weathers like sunny or cloudy. This phenomenon from a side manifests that the proposed framework FCM is more sensitive to the changes of material itself in multi-view analysis.

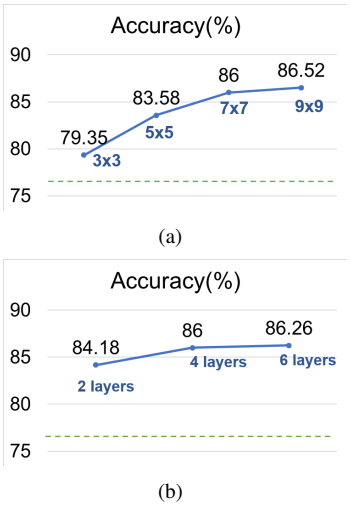

Fig. 5. Ablation Experiments. The green dashed line denotes the performance of Angular Filter in [12], *i.e.* 77% averaged on 12 categories of materials. (a) Number of SAIs in input LF data. (b) Number of layers in shallow CNNs.

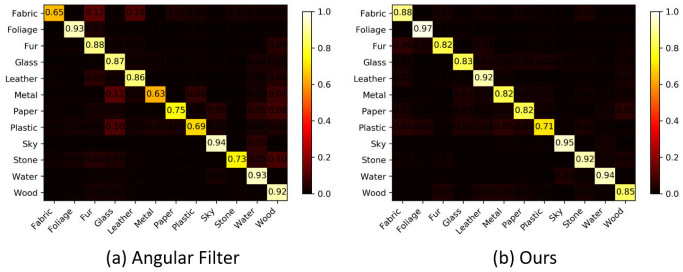

Fig. 4. The confusion matrix of single-pixel material classification. (a) Angular Filter [12]. (b) Ours.

Additionally, the confusion matrixes of Angular Filter method [12] and the propose framework FCM on Single-pixel Material Classification are shown in Fig.4. It is clear that FCM not only promotes the accuracy of material recognition on nearly all categories but also separates confusing materials such as *Fabric*, *Paper*, *Plastic*, *Glass* more precisely. In total, FCM can prominently boost the performance of single-pixel material classification.

B. Whole-image Material Segmentation

To evaluate qualitative results of whole-image material segmentation, we compare with MINC VGG model [8], Angular Filter method in [12] and Ground Truth(GT) material label images. According to the number of materials appearing in the image, results are shown in Fig.6-7.

Firstly, when there exists only one category of materials in the scene, the proposed framework FCM can almost perfectly classify all the pixels into belonging categories as shown in Fig.6. At the same time, the predicted material segmentation by FCM is nearly the same as GT labeling image, which obviously outperforms MINC VGG model [8] and [12]. Some tiny errors only occur around the corners of the image, which may be caused by irrelevant contextual information absorbed from zero-padding effect during inference.

As can be seen from Fig.7, when there are two or more categories of materials in the scene, the proposed framework can yield satisfying material segmentation results. Especially, FCM still works even the objects are occluded each other, or the textures and colors of the materials are in similar pattern. FCM can not only obtain rather accurate pixel classifications inside the same material region, but also preserve sharp and clear boundaries of the material-segmented instances. Compared with 2D predictions in [8] and Angular Filter methods in [12], the proposed framework is more stable in complex scenes and obtains better segmentation results. These results demonstrate the superiority of our methods.

C. Ablation Study

To further demonstrate the effectiveness of the configurations in the proposed pipeline, ablation experiments of single-pixel material classifications are conducted on the number of layers of shallow CNN and number of SAIs in the consecutive 3D slices. The results are depicted in Fig.5. Note that the network structure of shallow CNN is quite simple, adding more layers and extracting deeper features will not obtain prominent improvements. On the other hand, less layers are not enough to capture sufficient receptive field which brings an obvious drop in the accuracy of single-pixel classifications. As we include 2 more layers, *i.e.* from 4 to 6 layers, the accuracy averaged on 12 categories of materials boosts just 0.26% (9% to 9.26%) than [12]. If shallow CNNs are reduced to 2 layers, *i.e.* from 4 layers to 2 layers, the accuracy drops 1.82% (9% to 7.18%) than [12]. Moreover, number of SAIs in consecutive 3D slices is a key factor. 4D LF with angular resolution 3x3, 5x5, 7x7, 9x9 gain 2.35%, 6.58%, 9%, 9.52% improvement on accuracy averaged on 12 categories of material types than [12] respectively. Taking balanced performance into account, although 9x9 gains 0.52% higher than 7x7 SAIs, but such LF input consumes big storage and the whole-image material

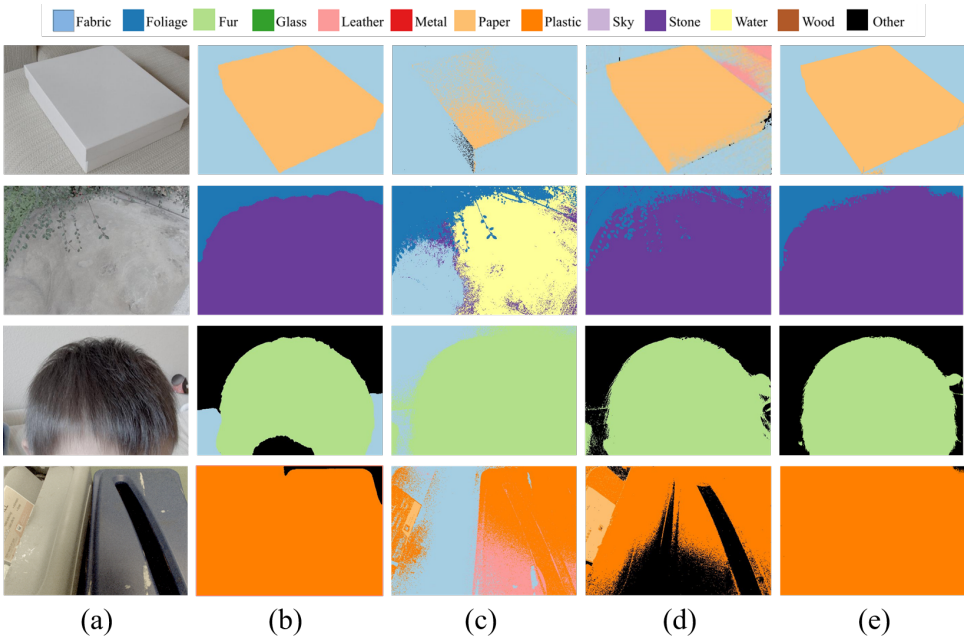

Fig. 6. The results of whole-image material segmentation under the circumstances that one category of materials occupies the majority of the scene. (a) Central SAI of input LF image. (b) Ground Truth. (c) 2D predictions by MINC VGG model [8]. (d) Angular Filter [12]. (e) Ours.

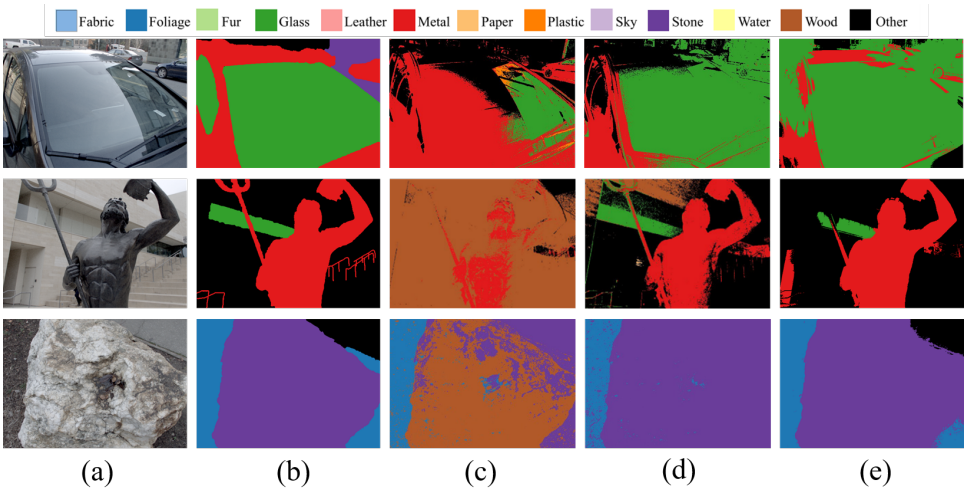

Fig. 7. The results of whole-image material segmentation under the circumstances that there exist two or more categories of materials in the scene. (a) Central SAI of input LF image. (b) Ground Truth. (c) 2D predictions by MINC VGG model [8]. (d) Angular Filter [12]. (e) Ours.

segmentation model is too large to be trained or tested on typical GPU cards so that 7x7 LF input can be considered as better options for practical applications.

V. Conclusion

Rather than following the approaches with 4D-to-2D adaptors, a novel "Factorize-Connect-Merge" (FCM) deep-learning pipeline is proposed to solve problems of light field image based material recognition in this paper. 4D light-field data as input are factorized into 3D slices. Shallow CNN and Bi-LSTM are leveraged to differentiate the imaging differences and connect the inconsistencies between SAIs caused by reflectances of distinct materials. The features from separated 3D slices are then merged together by feature selection and dimension reduction. The proposed framework can extract effective and robust feature for 4D light filed representations. It is experimentally verified that the proposed framework can obtain remarkable performances on both tasks of single-pixel material classification and whole-image material segmentation, which set a higher baseline for LF image based material recognition. As this pipeline is compatible with other vision tasks, researchers who may also take LF images as input and need to extract 4D light-field representations will benefit from this work.

REFERENCES

[1] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Transactions on Graphics*, vol. 18, no. 1, pp. 1–34, 1999.

[2] B. Caputo, E. Hayman, and P. Mallikarjuna, "Class-specific material categorisation," in *International Conference on Computer Vision*, vol. 2, 2005, pp. 1597–1604.

[3] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Material perception: What can you see in a brief glance?" *Journal of Vision*, vol. 9, no. 8, pp. 784–784, 2010.

[4] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a bayesian framework for material recognition," in *Proceedings ofthe IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 239–246.

[5] D. Hu, L. Bo, and X. Ren, "Toward robust material recognition for everyday objects," in *British Machine Vision Conference (BMVC)*, 2011, pp. 1–11.

[6] X. Qi, R. Xiao, C. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise rotation invariant co-occurrence local binary pattern," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2199–2213, 2014.

[7] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3828–3836.

[8] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3479–3487.

[9] H. Zhang, K. J. Dana, and K. Nishino, "Reflectance hashing for material recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3071–3080.

[10] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[11] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.

[12] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4d light-field dataset and cnn architectures for material recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 121–138.

[13] G. Schwartz and K. Nishino, "Visual material traits: Recognizing per-pixel material context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Dec 2013, pp. 883–890.

[14] G. Schwartz and K. Nishino, "Automatically discovering local visual material attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3565–3573.

[15] Y. Yoon, H. Jeon, D. Yoo, J. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 57–65.

[16] ——, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, June 2017.

[17] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.

[18] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 046–11 055, 2019.

[19] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2319–2330, May 2019.

[20] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 193, 2016.

[21] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4d rgbd light field from a single image," *International Conference on Computer Vision (ICCV)*, pp. 2262–2270, 2017.

[22] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, "Flexible, fast and accurate densely-sampled light field reconstruction network," 2019. [Online]. Available: https://arxiv.org/abs/1909.01341v1

[23] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on epi," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1638–1646.

[24] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 138–154.

[25] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with pseudo 4dcnn," *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 340–355, 2018.

[26] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared epi structure for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3261–3273, July 2019.

[27] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3746–3754.

[28] S. Heber, W. Yu, and T. Pock, "U-shaped networks for shape from light field," in *British Machine Vision Conference (BMVC)*, 2016.

[29] C. Shin, H. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 4748–4757.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] J. Li, T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," *International Joint Conference on Natural Language Processing (IJCNLP)*, vol. 1, pp. 1106–1115, 2015.

[32] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[33] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 12, pp. 3007–3021, 2018.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.

[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[36] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.