# Deep Material Recognition in Light-Fields via Disentanglement of Spatial and Angular Information

Bichuan Guo[1][0000−0001−8475−427X], Jiangtao Wen[1], and Yuxing Han[2]

[1] Tsinghua University, Beijing, China
[2] Research Institute of Tsinghua University in Shenzhen, Shenzhen, China
gbc16@mails.tsinghua.edu.cn

**Abstract.** Light-field cameras capture sub-views from multiple perspectives simultaneously, with possibly reflectance variations that can be used to augment material recognition in remote sensing, autonomous driving, etc. Existing approaches for light-field based material recognition suffer from the entanglement between angular and spatial domains, leading to inefficient training which in turn limits their performances. In this paper, we propose an approach that achieves decoupling of angular and spatial information by establishing correspondences in the angular domain, then employs regularization to enforce a rotational invariance. As opposed to relying on the Lambertian surface assumption, we align the angular domain by estimating sub-pixel displacements using the Fourier transform. The network takes sparse inputs, i.e. sub-views along particular directions, to gain structural information about the angular domain. A novel regularization technique further improves generalization by weight sharing and max-pooling among different directions. The proposed approach outperforms any previously reported method on multiple datasets. The accuracy gain over 2D images is improved by a factor of 1.5. Ablation studies are conducted to demonstrate the significance of each component.

**Keywords:** light field, material recognition, angular registration

## 1 Introduction

Material recognition is a fundamental problem in vision and plays an important role in many industrial applications. For example, drones are used in topography [21] for recognizing ground surfaces, and mowers can operate autonomously by inspecting surrounding surface materials. As a result, material recognition has been an active area of research, which is inherently challenging because the appearances of materials depend on various factors such as shape, lighting and exhibit strong visual variations. A variety of clues, such as texture, reflectance and scene context, need be taken into account to achieve good performance.

The recent introduction of commercially available and compact light-field cameras (e.g. Lytro Illum) provides an efficient way to improve image based recognition, making it possible to significantly boost the performance of the

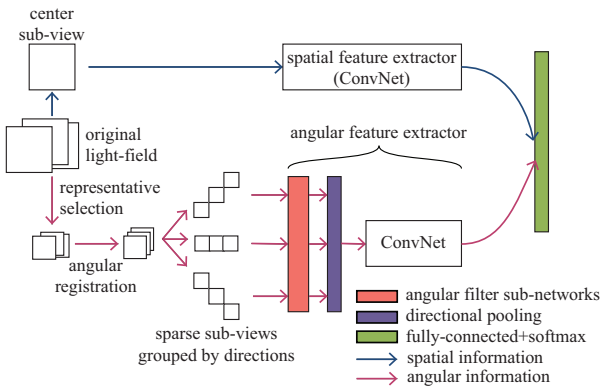


**Fig. 1.** Overview of the proposed framework. The light-field is represented by a spatial component (up), as well as an angular component (down) which is cropped to avoid occlusion and to facilitate alignment. Angular features are extracted from sparse sub-views grouped by perspective directions; all directions share filter weights and are aggregated via max-pooling. Spatial and angular features are combined at the end to produce class probabilities

above-mentioned applications with a simple optical upgrade. These cameras can capture sub-views from multiple viewpoints on a regular grid (i.e. the angular domain) in a single shot. Such rich information can be used to obtain critical cues such as depth information and intensity variation in different perspectives, which are not directly available from 2D images. Existing research have investigated using light-field cameras for recognizing not only materials [25], but also objects [14] and reflectance [13]. They confirmed that significant accuracy boost can be achieved by using light-field images over 2D images.

Several key aspects of light-field images were not addressed by existing methods. First, due to light-field camera optics, the angular domain is intertwined with the spatial domain to express reflectance variation. As a result, filters directly applied to each domain separately cannot disentangle reflectance variation efficiently. Moreover, occlusion in the scene can cause discontinuity of information distribution and misalignment, where same positions in sub-views correspond to different objects and materials. Third, the angular domain is highly redundant as sub-views exhibit strong similarities. Without regularization, it is hard to learn structural information and can easily lead to overfitting.

We propose a novel framework to tackle these problems. An overview of our methods is shown in Fig. 1. Our contributions are summarized below:

– Sub-views are aligned by estimating sub-pixel displacements using their Fourier transforms. A simple algorithm is deduced by leveraging angular domain geometry. This alignment procedure disentangles reflectance variation in angular domain from spatial domain.

- A representative region is selected based on inferred depth to avoid occlusion and to ease alignment. The spatial information is preserved by a separate, full sub-view. Two expert feature extractors (i.e. spatial and angular) are combined for joint classification.
- We employ regularization in angular domain by (1) using sparse sub-views grouped by perspective directions, so that the network is aware of the angular domain structure; (2) sharing parameters in all directions and enforcing rotational invariance with a novel directional pooling layer.

## 2 Related Work

**Material recognition.** Two distinct approaches have been taken in the literature for image based material recognition. One relies on information beyond object appearance, such as reflectance disks [33], scene depth [34], 3D surface geometry [7] and spatial thermal textures [5]. They incorporate special measurements and properties that are closely related to material characteristics. The other approach relies on 2D images alone, usually by efficient use of context information including object, texture and background. Schwartz and Nishino [17] proposed to separate local material from object/scene context and combine them later. Cimpoi et al. [6] introduced a new texture descriptor, which was used along with object descriptors. These 2D image based methods are not directly applicable to 4D light-field data studied in this paper; however, we are inspired by the idea of global-local decomposition in that we use separate expert models to extract spatial and angular features.

Qi et al. [15] explored the transform invariance of co-occurrence features, and introduced a pairwise rotation invariant feature for 2D images. We instead design a ConvNet architecture to enforce rotational invariance in angular domain for light-fields. Xue et al. [32] used a stereo pair to perform material recognition, by feeding their difference and one view to a neural network. Wang et al. [25] proposed multiple ConvNet architectures for recognizing materials using light-field images. Our work further explores this idea of using reflectance variation to improve material recognition, which requires algorithmic adaptations to several key characteristics of light-field images.

**Computer vision for light-fields.** Light-fields received increasing popularity in the vision community over the years. One direction is to enhance the quality of light-fields, as current acquisition techniques are still rudimentary. This includes super-resolution in spatial domain [27] or angular domain [31], denoising [4] and light-field video interpolation [26]. Another line of research is to tackle existing computer vision problems using light-field images. Many depth estimation algorithms [11, 12, 24, 28] that rely on efficient use of angular information were proposed. Similar methodologies were also adopted in other tasks. Lu et al. [14] used domain-interleaved filters to simultaneously extract spatial and angular features for object classification. Alperovich et al. [1] used an autoencoder to find compact representations for epipolar plane images [2] (EPI), which can be decoded to yield diffuse/specular intrinsic components. Chen et al. [3] used

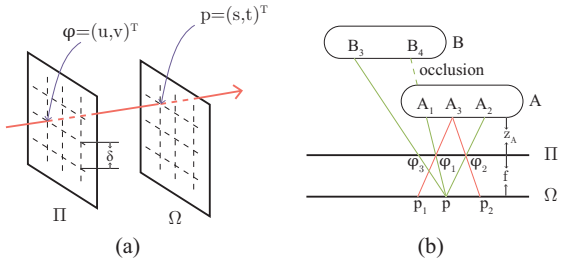


**Fig. 2.** Light-field geometry. (a) The 4D light-field parametrized by two parallel planes: the lens plane $\Pi$ (angular domain) and the sensor plane $\Omega$ (spatial domain). (b) Green lines: Pixels with same spatial coordinates $\boldsymbol{p}$ correspond to different objects. Red lines: Disparity $\boldsymbol{p}_1 - \boldsymbol{p}_2$ is linear to lens translation $\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2$

the surface camera to model reflectance variation of 3D points, leading to better stereo matching. We borrowed this idea in that our approach constructs the surface camera via alignment to disentangle angular and spatial variations.

**Neural networks for light-fields.** Many papers employed neural networks to analyze light-fields. To recognize materials or bidirectional reflectance distribution functions (BRDF), [13, 25] proposed to stack sub-views and take convolutions in spatial domain, or instead apply filters in angular domain. We extend the above ideas by enforcing sparsity in both domains in order to avoid occlusion and reduce overfitting. Heber et al. [9] introduced stacking in EPI domain for shape inference. [13, 25, 30] used alternating spatial and angular domain convolutions to mimic a full 4D filter while keeping computational costs low. Shin et al. [19] proposed to only use horizontal or crosshair sub-views to increase computational speed. Xue et al. [32] proposed the concept of angular gradients which enables the network to be aware of angular structure. In this work we further explore these ideas and find that, combined with weight sharing and pooling, such sparsity in angular domain improves generalization by enforcing rotational invariance, and improves the efficiency of extracting angular gradients.

## 3 Light-field Imaging Model

### 3.1 Light-field Geometry

The 4D light-field is usually parametrized using two parallel planes: a lens plane $\Pi$, representing the angular domain, and a sensor plane $\Omega$, representing the spatial domain, as shown in Fig. 2(a). A ray of light passes through a micro-lens on $\Pi$ and is captured by a pixel on $\Omega$. The light-field can be viewed as a function

$$L : \Pi \times \Omega \to \mathbb{R}^C, (\boldsymbol{\varphi}, \boldsymbol{p}) \mapsto L(\boldsymbol{\varphi}, \boldsymbol{p}), \tag{1}$$

where $\boldsymbol{\varphi}$ is a micro-lens on $\Pi$ with angular coordinates $(u, v)^\mathsf{T}$, $\boldsymbol{p}$ is a pixel on sensor $\Omega$ with spatial coordinates $(s, t)^\mathsf{T}$, and $C$ is the number of color channels.

The behavior of a micro-lens can be analyzed with a pinhole camera model. We first fix the pixel $\boldsymbol{p}$ in spatial domain and vary the lenses $\boldsymbol{\varphi}$ in angular domain. In Fig. 2(b), $\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2$ capture 3D points $A_1, A_2$, and $\boldsymbol{\varphi}_3$ captures a 3D point $B_3$. Here $B$ is occluded by $A$ since $\boldsymbol{\varphi}_1$ would capture $B_4$ without $A$. It is clear that variations in angular domain may correspond to multiple 3D points, or even different objects in case of occlusion. However, for material recognition, the key is to obtain reflectance variation of individual surface points in different perspectives. This can be done by fixing the 3D point and correcting for disparity in spatial domain. The light rays emitted from a 3D point $A_3$ with depth $z_A$ reach two lenses $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$, and are captured by sensors at $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. The disparity between $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ is related to lens translation linearly:

$$\boldsymbol{p}_1 - \boldsymbol{p}_2 = \left(1 + \frac{f}{z_A}\right)(\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2). \tag{2}$$

If scene depth $z_A$ and focal length $f$ are known, one can use (2) to group pixels in different sub-views with same originations. This is often called the angular sampling image (ASI) [18], describing the reflectance variation of individual 3D points. Operating on the ASI has the advantage that reflectance information is decoupled from spatial variations, which allows specialized expert models to be applied to each aspect, and later combined to form a joint decision.

### 3.2 Analysis of Baseline Methods

We now analyze several baseline methods (in *italic*) proposed in [25]. *2D-average* and *viewpool* feed each sub-view to a ConvNet and perform averaging or max-pooling for aggregation. They fail to exploit between-view correlation and perform poorly. *Stack* stacks all sub-views before a ConvNet, the first layer of which takes convolutions in spatial and angular domains simultaneously. We see previously that these two domains are intertwined; this complex coupling effect causes the model to perform only slightly better than 2D models. *EPI* is similar to *stack* as it stacks in EPI domain instead, achieving similar performance.

Two winning models from [25], namely *ang-filter* (referred to as *angular-filter* in the original paper) and *4D-filter*, provide significant performance boosts over 2D models. Our **first motivation** is due to the baseline method *ang-filter* outperforms *stack*. These two methods both apply convolutional filters in angular domain, the difference is that "ang-filter" has 1x1 spatial filters while "stack" has 3x3 spatial filters. This means *ang-filter* is only convolving in the angular domain while *stack* is convolving in spatial and angular domains at the same time, which suggests us to decouple these two domains. The **second motivation** can be considered as a step further from the first motivation. According to the analysis in Section 3.1, the angular domain itself is both affected by spatial and reflectance variations. This suggests us to align pixels into ASIs, which will decouple reflectance variations from spatial variations.

## 4 Methods

In our proposed framework, we first decompose the light-field into a spatial component and an angular component; features from both components are extracted separately and combined at the end to produce class probabilities, as shown in Fig. 1. The spatial component is the center sub-view, containing overall appearance, object and scene context. The angular component is then aligned into ASIs through *representative selection* and *angular registration*.

### 4.1 Representative Selection

ASIs can be formed by collecting pixels from sub-views according to (2). In practice, the dense depth field is usually not available and needs to be estimated. However, most depth estimation algorithms [10, 18, 24] rely on the assumption that objects are made of Lambertian materials with uniform reflectance across all viewing angles. This is clearly an oversimplification as our key to the problem is reflectance variation. Even for methods that are robust to non-Lambertian surfaces, dense depth estimation is still error-prone in case of occlusion [24].

We notice that by selecting a region with constant depth, the situation is greatly simplified. By the disparity linearity (2), the disparity between two sub-views becomes a constant for any 3D point in this region, which implies that we can form ASIs by simply translating sub-views. Also, a region with constant depth is mostly occlusion-free. Therefore, we crop the angular component to a representative region that has approximately constant depth. As the spatial component remains full resolution, we will not lose much spatial information by cropping the angular component.

The dense depth map $D(\Omega)$ estimated from [11] is used for region selection, which is less error-prone than using it to directly warp pixels (we verify this experimentally in Sec. 5.3). The spatial domain is partitioned into an $N \times N$ grid, and denote the depth values in each block by

$$\forall 1 \leq i, j \leq N, D_{ij} = \left\{ D(s,t) : \frac{(i-1)S}{N} < s \leq \frac{iS}{N}, \frac{(j-1)T}{N} < t \leq \frac{jT}{N} \right\}, \quad (3)$$

where $S \times T$ is the spatial resolution. The representative region is selected as the one that minimizes the following loss

$$(i^*, j^*) = \underset{i,j}{\operatorname{argmin}} \, \sigma[D_{ij}] + \lambda \big| \mu[D_{ij}] - \operatorname{med}[D(\Omega)] \big|, \quad (4)$$

where $\sigma[\cdot]$, $\mu[\cdot]$ and $\operatorname{med}[\cdot]$ denote the standard deviation, mean and median of a set. The first term penalizes the amount of depth variation as we wish it to be approximately constant, and the second term penalizes outlier regions that significantly deviate from the overall median depth, which usually correspond to backgrounds. The Lagrangian coefficient $\lambda$ is empirically set to 1.0 to provide a good trade-off between these two terms.

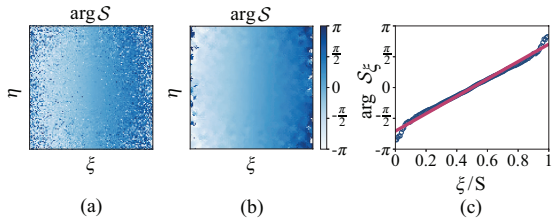


**Fig. 3.** Displacement estimation in Fourier domain. (a) If $\Delta t = 0$, $\arg \mathcal{S}$ is constant in $\eta$. There is a large amount of noise due to aliasing. (b) Median filtering on $\arg \mathcal{S}$ to remove isolated noise. (c) Median values of $\arg \mathcal{S}$ at each $\xi$ and its linear regression

### 4.2 Angular Registration

As the representative region has approximately constant depth, we can align (also called *register* in literature) any sub-view with the center sub-view via translation. By constraining disparity to be shared across spatial domain, the pixel level Lambertian surface assumption is relaxed to sub-view level translation, which holds as long as changing viewpoints does not cause drastic overall changes. Jeon et al. [11] used the Fourier domain to perform sub-pixel displacement. Here we instead use the Fourier domain to estimate the displacement itself. The Fourier shift theorem [16] states that if two images $I_1$ and $I_2$ of same size $S \times T$ are related by a translation $\Delta \boldsymbol{p} = (\Delta s, \Delta t)$, $I_2(s,t) = I_1(s + \Delta s, t + \Delta t)$, then their discrete Fourier transforms $\mathcal{F}\{I_1\}$, $\mathcal{F}\{I_2\}$ are related by

$$\mathcal{F}\{I_2\}(\xi,\eta) = \mathcal{F}\{I_1\}(\xi,\eta) \cdot e^{2\pi i(\xi \Delta s/S + \eta \Delta t/T)}, \tag{5}$$

therefore we can perform sub-pixel translation $\Delta \boldsymbol{p}$ to $I_1$ by

$$I_2 = \mathcal{F}^{-1}\{\mathcal{F}\{I_1\}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega} \cdot (\Delta s/S, \Delta t/T)\}. \tag{6}$$

To estimate the translation between $I_1$ and $I_2$, take the inverse Fourier transform of the cross-power spectrum

$$\mathcal{S} = \frac{\mathcal{F}\{I_1\}\mathcal{F}^*\{I_2\}}{|\mathcal{F}\{I_1\}\mathcal{F}\{I_2\}|} = \exp[-2\pi i(\frac{\xi}{S}\Delta s + \frac{\eta}{T}\Delta t)], \tag{7}$$

to arrive at an impulse function at $(\Delta s, \Delta t)$, where $F^*$ denotes the complex conjugate of $F$. However, this method can only measure integral translation, and its performance is not robust to aliasing, i.e. imperfect correspondence which is prevalent in case of non-Lambertian materials. Stone et al. [22] proposed a generic sub-pixel registration algorithm that deals with aliasing. Here we deduce a simplified approach by leveraging angular domain geometry.

By (2), $\Delta \boldsymbol{p}$ is linear to the lens translation $\Delta \boldsymbol{\varphi} = (\Delta u, \Delta v)$. If we choose two sub-views with $\Delta v = 0$, $\Delta t$ will also be zero, and the phase of the cross-power spectrum $\arg \mathcal{S} = -2\pi \Delta s \cdot \xi/S$ will be constant in $\eta$, as shown in Fig. 3(a).

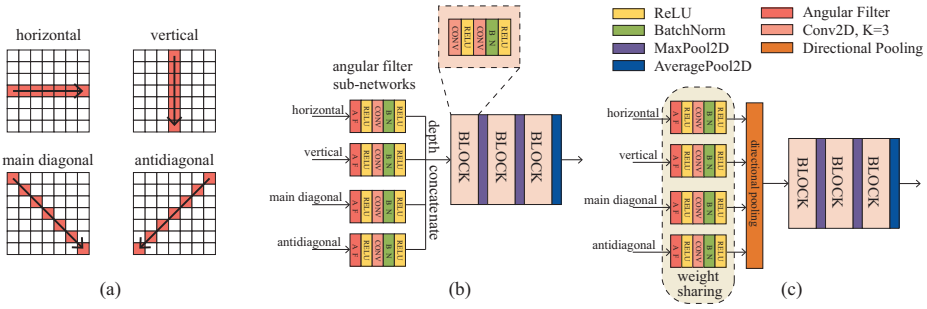


**Fig. 4.** Network architecture of the angular feature extractor. (a) Sub-views along particular directions are stacked and fed to the network. (b) *Angular-4*. Stacks of each direction have their own angular filter sub-networks. They are concatenated depth-wise before shared convolutional layers, which consist of three basic building blocks, each basic block is a sequence of "Conv-ReLU-Conv-BN-ReLU" layers. (c) *Angular-S*. Comparing to *angular-4*, the angular filter sub-networks have shared weights. The directional pooling layer is a MaxPool3D layer pooling among four directions. In (b) and (c), all "AF" layers are Conv2D layers with kernel size $K_{\text{af}} \in \{1, 3\}$, SAME padding, containing $\#\text{af} \in \{16, 32, 64\}$ different angular filters.

We first apply a median filter to $\arg \mathcal{S}$ to remove isolated noise due to aliasing, as shown in Fig. 3(b). Since we are dealing with periodic phase data, we take circular medians [23] instead of ordinary medians. To reduce clustered noise along the $\eta$ axis, the median value $\arg \mathcal{S}_\xi = \text{med}[\mathcal{S}(\xi, :)]$ is taken along $\eta$. Finally, we regress $\arg \mathcal{S}_\xi$ on $\xi/S$ using ordinary least squares to further reduce noise along the $\xi$ axis, as shown in Fig. 3(c). The estimated slope is then $-2\pi$ times the estimated displacement $\Delta\hat{s}$. Note that if $\Delta s$ is larger than one pixel, $\arg \mathcal{S}_\xi$ will wrap around $\pi/-\pi$. Therefore, we first locate the impulse $\mathcal{F}^{-1}\{\mathcal{S}\}$ to register $I_1$ and $I_2$ to the nearest integral pixel, then proceed with sub-pixel refinement.

The micro-lens array of a light-field camera can be modeled with a square grid, with constant translation $\delta$ between adjacent lenses [35], as shown in Fig. 2(a). Therefore, the disparity $\Delta$ between adjacent sub-views is also a constant. We estimate $\Delta s$ from two sub-views at both ends of the center row, and $\Delta t$ from two sub-views at both ends of the center column. $\Delta$ is estimated by averaging along two axes:

$$\Delta = \frac{1}{2}\left(\frac{\Delta s}{U-1} + \frac{\Delta t}{V-1}\right), \tag{8}$$

where $U$ and $V$ are the numbers of sub-views along axes $u$ and $v$. With $\Delta$, we register all sub-views to the center sub-view using (6), so that the angular domain of each spatial pixel forms an ASI. We provide some visual examples in the supplementary material.

### 4.3 Network Architecture

We extract spatial features from the spatial component with a standard 2D image feature extractor (e.g. ResNet [8]). Angular features are extracted from the cropped and registered angular component using an angular feature extractor, shown in Fig. 4. We employ a multi-stream network [19] that reads sub-views in some certain directions, i.e. center horizontal, center vertical, main diagonal and antidiagonal, as shown in Fig. 4(a). The angular feature extractor *angular-4* (4 directions) is shown in Fig. 4(b). Sub-views in each direction are stacked and encoded separately by angular filter sub-networks at the beginning to produce meaningful representations. This not only explicitly supplies the network with structural information of angular domain, but also reduces angular redundancy. The first layer "AF" is a convolutional layer consisting of multiple filters that convolve across the angular domain. *Angular-4* is a modification of the multi-stream architecture in [19]. Convolutional kernel sizes and network depth are altered; the last convolutional layer is replaced by global average pooling to produce feature vectors. See more details in the caption of Fig. 4 and the supplementary material.

The multi-stream architecture was originally proposed for depth estimation. However, there is a major difference between depth estimation and material recognition: depth is a 2D function with axes orientations, but material is invariant to camera and object rotations. These actions cause angular domain axes to rotate. We can incorporate this prior knowledge by employing an invariance mechanism, which we call *angular rotational invariance*. Our final architecture *angular-S* (shared) augments *angular-4* with this invariance, as shown in Fig. 4(c). All four angular filter sub-networks have shared weights, which is implemented by passing all sub-view stacks to a single sub-network separately to produce four feature maps. These feature maps are concatenated along a new "direction" dimension, and a max-pooling layer is applied to this new dimension, which we call *directional pooling*. We can verify this architecture indeed enforces angular rotational invariance by observing that, if angular domain axes rotate by multiples of $\pi/4$, it results in a permutation of sub-view stacks, and the output of the directional pooling layer is invariant.

## 5 Experimental Results

### 5.1 Data and Training Procedure

We report results on multiple light-field material datasets to demonstrate the robustness of our methods. The first dataset is LFMR [25], which contains real light-field images captured by Lytro with spatial resolution $376\times541$ and angular resolution $7\times7$. The second dataset is the rendered light-field images with same resolutions from the BTF dataset [29], which was also used in [25] for evaluation. Our data preprocessing procedures follow [25]: square sample patches are extracted from whole light-field images; their centers are separated by at least half the patch size, and more than 50% of the pixels correspond to the target

material. Unless stated otherwise, most of our experiments use 128×128 patches. Each dataset is randomly split into a training set and a test set by 7:3, 1/7 of the training set is used as a validation set for hyper-parameter selection. Patches from the same light-field image only appear in one set to avoid strong correlation between the train/val/test sets.

**Table 1.** Classification accuracy (in percentage) comparison on test sets. "2D": use spatial features for classification. "model-4/S": *angular-4/S* is used as the angular feature extractor. "$\text{gain}_{2D}$": accuracy improvement over "2D"

| dataset | LFMR [25] | | BTF [29] | |
|---|---|---|---|---|
| method | accuracy | $\text{gain}_{2D}$ | accuracy | $\text{gain}_{2D}$ |
| 2D | $70.45_{\pm 0.23}$ | - | $67.35_{\pm 0.33}$ | - |
| StackNet [13] | $72.67_{\pm 2.39}$ | 2.22 | $70.19_{\pm 1.87}$ | 2.84 |
| AngConvNet [13] | $73.23_{\pm 2.29}$ | 2.78 | $72.16_{\pm 1.08}$ | 4.81 |
| Lu et al. [14] | $76.48_{\pm 1.23}$ | 6.03 | $\mathbf{73.84}_{\pm 1.31}$ | **6.49** |
| 4D-filter [25] | $77.29_{\pm 1.05}$ | 6.84 | $71.81_{\pm 1.93}$ | 4.46 |
| ang-filter [25] | $\mathbf{77.83}_{\pm 0.89}$ | **7.38** | $73.27_{\pm 0.85}$ | 5.92 |
| MDAIN [32] | $75.73_{\pm 1.88}$ | 5.28 | $69.43_{\pm 1.52}$ | 2.08 |
| model-4 (ours) | $80.38_{\pm 0.53}$ | 9.93 | $75.14_{\pm 0.66}$ | 7.79 |
| model-S (ours) | $\mathbf{81.75}_{\pm 0.61}$ | **11.30** | $\mathbf{77.52}_{\pm 0.58}$ | **10.17** |

Data augmentation is carried out in spatial domain, including random horizontal flipping, randomly cropping the spatial resolution to the largest factor of 224 (e.g. if the patch size is 128×128, we take 112×112 random crops), and then upsampling to 224×224. We also normalize all sub-views by subtracting half the max intensity (e.g. 128 for 8-bit images) uniformly in all color channels. At test time, we perform centered cropping instead of random cropping, followed by normalization. The angular feature extractor is trained from scratch, and the spatial feature extractor is ResNet-18 (except the last fully-connected layer), since it provides fast training and good generalization. All models are trained with stochastic gradient descent and cross entropy loss for 200 epochs, with a base learning rate of $10^{-4}$, momentum 0.9 and batch size 128. The angular feature extractor and the fully connected layer use 10× the base learning rate. All experiments use the top-1 accuracy as their performance measures.

### 5.2 Overall Performance

Table 1 shows the results of our proposed framework on the test sets, and compares with winning baseline methods from [25]. We also compare with architectures designed for other closely related tasks. StackNet and AngConvNet were proposed in [13] for surface BRDF recognition using light-field images. MDAIN [32] uses multiple stereo pairs for material recognition. The authors tested on light-field datasets by selecting 4 sub-view pairs of entire light-field images. We replicate their test conditions except that patches are classified rather than the

whole light-field. Lu et al. [14] proposed a domain-interleaved architecture that resembles *4D-filter* for light-field object classification. Since it requires 8×8 angular resolution, we pad angular domain by replication and use it for material classification. The classification accuracy is reported by averaging 5 random train/test splits. More details are given in the table caption.

For fair comparison, all previous methods also use ResNet-18 as their backbone networks. Note that [25] originally used VGG-16 [20] as the backbone architecture and achieved 7% gain on both *ang-filter* and *4D-filter*. By switching to ResNet-18 we reproduce similar gains but with much less memory and shorter runtime. We see that comparing to the best baseline method, our *model-S* achieves 11.30%/7.38%=1.53 times gain on the LFMR dataset, and 10.17% / 6.49% = 1.57 times gain on the BTF dataset. This result shows that our proposed framework significantly outperforms previously reported methods. While previous methods achieve gains over 2D by *utilizing additional data* which is arguably expected, our method outperforms these methods by *more efficient usage of these additional data*.

### 5.3 Ablation Studies

We conduct extensive ablation experiments using the LFMR dataset, including hyper-parameter choices, contribution of each technical component and robustness tests.

**Hyper-parameters.** Table 2 (left) compares different angular filter kernel sizes

**Table 2.** Left: classification accuracy on validation set. Representative selection is enabled. Right: classification accuracy with different $N$ for representative selection. $N$=1 means no cropping. We report results using both angular and spatial features ("angular+spatial"), and only using angular features ("angular") for classification. In both tables, *angular-S* is the angular feature extractor, angular registration is enabled

| $K_{\text{af}}$ | #af | mean acc. (%) | std. (%) |
|---|---|---|---|
| 1 | 16 | 81.35 | 0.57 |
| 1 | 32 | **81.92** | 0.56 |
| 1 | 64 | 80.93 | 0.60 |
| 3 | 32 | 81.16 | 0.60 |

| feature | $N$ | mean acc. (%) | std. (%) |
|---|---|---|---|
| angular+spatial | 1 | 81.45 | 0.42 |
| angular+spatial | 2 | 81.56 | 0.69 |
| angular+spatial | 3 | **81.92** | 0.56 |
| angular+spatial | 4 | 81.28 | 0.53 |
| angular | 1 | **66.34** | 1.22 |
| angular | 2 | 60.15 | 1.04 |
| angular | 3 | 55.31 | 1.41 |
| angular | 4 | 51.93 | 1.53 |

$K_{\text{af}}$ and numbers of angular filters #af. The classification accuracy on the validation set is reported by averaging 5 random train/validation splits. This agrees with the observation in [25] that a medium #af offers the best performance, while a low #af reduces representation power, and a large #af leads to overfitting. We also observe that $K_{\text{af}} = 1$ outperforms $K_{\text{af}} = 3$. $K_{\text{af}} = 1$ corresponds to taking convolutions only in angular domain, i.e. aligned ASIs, while $K_{\text{af}} = 3$

corresponds to taking convolutions both in the ASI and spatial domain. This result confirms that decoupling angular and spatial information is beneficial.

Table 2 (right) compares different $N$ used for representative selection. $N = 1$ means the entire patch is used and selection is disabled. Besides the proposed framework ("angular+spatial"), we also report results where only angular features are used for classification ("angular"). For "angular+spatial", a medium sized region offers the best performance, as a large region may violate the constant depth assumption, leading to potential occlusion and bad registration; while a small region carries little information. In contrast, as the spatial component is missing in "angular", the best choice is to use the entire patch, which contains the most spatial information. This result shows that by decomposing light-fields into two components, we can conveniently represent local properties with a small region without worrying about losing much spatial information.

**Table 3.** Left: ablation study on the test set of LFMR. "model-4/S": *angular-4* / *angular-S* is used as the angular feature extractor. "select": use representative selection. "register": use angular registration. Right: classification accuracy (in percentage) comparison using different light-field patch sizes

| id | method | select | register | mean acc. | std. |
|---|---|---|---|---|---|
| 1 | ang-filter | - | - | 77.83 | 0.89 |
| 2 | model-4 | ✓ | ✓ | **80.38** | 0.53 |
| 3 | | ✓ | ✗ | 79.85 | 0.45 |
| 4 | | ✗ | ✓ | 79.36 | 1.05 |
| 5 | | ✗ | ✗ | 78.91 | 0.75 |
| 6 | model-S | ✓ | ✓ | **81.75** | 0.61 |
| 7 | | ✓ | ✗ | 81.00 | 0.56 |
| 8 | | ✗ | ✓ | 81.45 | 0.42 |
| 9 | | ✗ | ✗ | 80.74 | 0.52 |
| 10 | | random | ✓ | 81.41 | 0.73 |
| 11 | | ✗ | warp | 80.35 | 0.55 |

| data | 2D | ang-filter | model-S |
|---|---|---|---|
| size=32 | 49.18 | 58.79 | **63.29** |
| size=64 | 58.73 | 66.38 | **71.73** |
| size=128 | 70.45 | 77.83 | **81.75** |
| size=256 | 78.80 | 82.37 | **85.69** |

**Components.** Table 3 (left) provides a breakdown of each component's contribution. Compare row 1 and 5, we see that using sparse inputs and explicit directional information provides 1.1% gain. Compare row 5 and 9, we see that enforcing angular rotational invariance provides 1.8% further gain. Within the *model-4* group, angular registration offers 0.5% gain, and representative selection offers 0.9% gain. When both are used, the combined gain 1.5% is nearly additive. This gain becomes less significant (1.0%) for *model-S*, implying that angular rotational invariance increases the network's robustness to misaligned and redundant data.

We analyze representative selection and angular registration in more detail with two more ablation test cases. Row 10 selects representative regions randomly rather than using (4) to select the best candidate. Its performance drops since it is susceptible to occlusion and background. Row 11 uses the estimated depth map to directly warp pixels, rather than to select representative regions. We observe a performance degradation comparing to row 8 as it imposes the


**Fig. 5.** Confusion matrices of *ang-filter* [25] and *model-S* on the LFMR dataset. Paper, glass, woods and plastic are the most improved categories

Lambertian assumption to all materials and alters ASIs. This further proves the necessity and significance of our proposed methods.

**Robustness.** We verify the robustness of our methods by varying the light-field patch sizes. Table 3 (right) compares performances of the 2D model using spatial features ("2D"), the best baseline method *ang-filter* and our best method *model-S*. It can be seen that under various input sizes, our method consistently outperforms *ang-filter* and significantly improves the gain over "2D".

### 5.4 Visualization

**Angular filter output.** Both *ang-filter* and our angular feature extractors use angular filters to perform convolutions in angular domain before subsequent layers. Fig. 6 (left) compares their activations in *ang-filter* and *angular-S*, both using 1×1 kernels. We observe that in (a3), the activations of different sub-view stacks differ, indicating the material (fabric) has strong reflection variation. Because all angular filter sub-networks have shared weights, difference of activations can only be caused by difference of inputs. Meanwhile in (b3), the material (plastic) has homogeneous reflection, and its activations are similar across directions. In contrast, this pattern is not present in (a2) and (b2), where directional information is hard to visualize.

**Angular v.s. spatial responses.** By combining angular and spatial features at the end, we can use the network to evaluate their relative strengths. For the neuron $i$ corresponding to the true class in the softmax layer, its pre-activation is a linear transform of angular and spatial feature vectors. Define the angular and spatial responses as the absolute values of $W_i^a \phi_a$ and $W_i^s \phi_s$, so that the pre-activation $y_i$ of the true class neuron $i$ has the decomposition $y_i = W_i^a \phi_a + W_i^s \phi_s$, where $W_i^s$, $W_i^a$ are weights of the fully-connected layer connected to neuron $i$, $\phi_a$ and $\phi_s$ are angular and spatial feature vectors, respectively. Fig. 6 (right) compares angular and spatial responses for two different patches in the same light-field image. If the patch contains object or shape information so that the

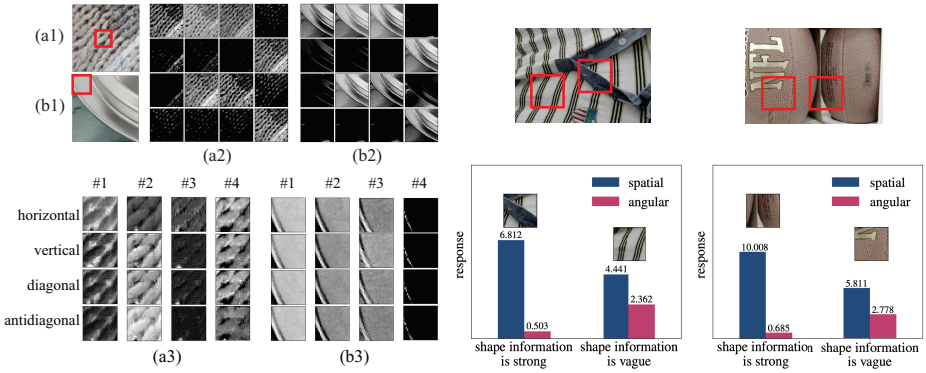


**Fig. 6.** Left: angular filter activations of *ang-filter* and *angular-S*. (a1) and (b1) are the original light-field patches, representative regions are outlined in red. (a2) and (b2) are the first 16 feature maps from *ang-filter*. Columns in (a3) and (b3) correspond to the first 4 feature maps from *angular-S*, each row corresponds to a stack of sub-views. Right: angular/spatial responses of different patches in the same light-field image. Top: original light-field images, selected patches are outlined in red. Bottom: corresponding patches are displayed above bar graphs

material can be easily inferred, the spatial response is much higher than angular response. Conversely, if context information is vague, then material has to be inferred from local properties such as texture and reflection, the angular response becomes more significant.

## 6 Conclusion

In this paper, we propose a novel framework for material recognition using light-fields that can potentially boost many industrial applications with a simple optical upgrade. The light-field is decomposed into the center sub-view and a representative crop, responsible for spatial and angular feature extraction, respectively. Thanks to the spatial-angular decomposition, we can keep most spatial information intact while cropping the angular component to avoid occlusion and for better registration. The angular feature extractor employs directional regularization by weight sharing in angular filter sub-networks and directional pooling; they together enforce rotational invariance in angular domain. Our methodology is verified by thorough ablation and robustness studies. It also casts light on how to efficiently learn from data with intertwined dimensions.

## 7 Acknowledgement

Yuxing Han is the corresponding author. This work was supported by the Natural Science Foundation of China (Project Number 61521002) and Shenzhen International Collaborative Research Project (Grant GJHZ20180929151604875).

## References


1. Alperovich, A., Johannsen, O., Strecke, M., Goldluecke, B.: Light field intrinsics with a deep encoder-decoder network. In: CVPR (2018)
2. Bolles, R.C., Baker, H.H., Marimont, D.H.: Epipolar-plane image analysis: An approach to determining structure from motion. IJCV **1**(1) (1987)
3. Chen, C., Lin, H., Yu, Z., Bing Kang, S., Yu, J.: Light field stereo matching using bilateral statistics of surface cameras. In: CVPR (2014)
4. Chen, J., Hou, J., Chau, L.P.: Light field denoising via anisotropic parallax analysis in a CNN framework. IEEE Signal Processing Letters **25**(9) (2018)
5. Cho, Y., Bianchi-Berthouze, N., Marquardt, N., Julier, S.J.: Deep thermal imaging: proximate material type recognition in the wild through deep learning of spatial surface temperature patterns. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018)
6. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: CVPR (2015)
7. DeGol, J., Golparvar-Fard, M., Hoiem, D.: Geometry-informed material recognition. In: CVPR (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Heber, S., Yu, W., Pock, T.: Neural EPI-volume networks for shape from light field. In: ICCV (2017)
10. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4D light fields. In: Asian Conference on Computer Vision (2016)
11. Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., So Kweon, I.: Accurate depth map estimation from a lenslet light field camera. In: CVPR (2015)
12. Johannsen, O., Sulc, A., Goldluecke, B.: What sparse light field coding reveals about scene structure. In: CVPR (2016)
13. Lu, F., He, L., You, S., Chen, X., Hao, Z.: Identifying surface BRDF from a single 4-D light field image via deep neural network. IEEE Journal of Selected Topics in Signal Processing **11**(7) (2017)
14. Lu, Z., Yeung, H.W., Qu, Q., Chung, Y.Y., Chen, X., Chen, Z.: Improved image classification with 4D light-field and interleaved convolutional neural network. Multimedia Tools and Applications (2018)
15. Qi, X., Xiao, R., Li, C.G., Qiao, Y., Guo, J., Tang, X.: Pairwise rotation invariant co-occurrence local binary pattern. IEEE TPAMI **36**(11) (2014)
16. Reddy, B.S., Chatterji, B.N.: An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE TIP **5**(8) (1996)
17. Schwartz, G., Nishino, K.: Material recognition from local appearance in global context. arXiv preprint arXiv:1611.09394 (2016)
18. Sheng, H., Zhang, S., Cao, X., Fang, Y., Xiong, Z.: Geometric occlusion analysis in depth estimation using integral guided filter for light-field image. IEEE TIP **26**(12) (2017)
19. Shin, C., Jeon, H.G., Yoon, Y., So Kweon, I., Joo Kim, S.: EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images. In: CVPR (2018)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)

21. Sonnemann, T., Ulloa Hung, J., Hofman, C.: Mapping indigenous settlement topography in the Caribbean using drones. Remote Sensing **8**(10) (2016)
22. Stone, H.S., Orchard, M.T., Chang, E.C., Martucci, S.A.: A fast direct Fourier-based algorithm for subpixel registration of images. IEEE Transactions on Geoscience and Remote Sensing **39**(10) (2001)
23. Storath, M., Weinmann, A.: Fast median filtering for phase or orientation data. IEEE TPAMI **40**(3) (2018)
24. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: ICCV (2015)
25. Wang, T.C., Zhu, J.Y., Hiroaki, E., Chandraker, M., Efros, A.A., Ramamoorthi, R.: A 4D light-field dataset and CNN architectures for material recognition. In: ECCV (2016)
26. Wang, T.C., Zhu, J.Y., Kalantari, N.K., Efros, A.A., Ramamoorthi, R.: Light field video capture using a learning-based hybrid imaging system. ACM TOG **36**(4) (2017)
27. Wang, Y., Liu, F., Zhang, K., Hou, G., Sun, Z., Tan, T.: LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. IEEE TIP **27**(9) (2018)
28. Wanner, S., Goldluecke, B.: Reconstructing reflective and transparent surfaces from epipolar plane images. In: German Conference on Pattern Recognition (2013)
29. Weinmann, M., Gall, J., Klein, R.: Material classification based on training data synthesized using a BTF database. In: ECCV (2014)
30. Wing Fung Yeung, H., Hou, J., Chen, J., Ying Chung, Y., Chen, X.: Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In: ECCV (2018)
31. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on EPI. In: CVPR (2017)
32. Xue, J., Zhang, H., Dana, K., Nishino, K.: Differential angular imaging for material recognition. In: CVPR (2017)
33. Zhang, H., Dana, K., Nishino, K.: Reflectance hashing for material recognition. In: CVPR (2015)
34. Zhao, C., Sun, L., Stolkin, R.: A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. In: 2017 18th International Conference on Advanced Robotics (2017)
35. Zhao, S., Chen, Z.: Light field image coding via linear approximation prior. In: ICIP (2017)