Surface Material Perception Through Multimodal Learning

Shi Mao, Mengqi Ji, Bin Wang, Qionghai Dai, *Senior Member, IEEE*, and Lu Fang, *Senior Member, IEEE*

***Abstract*—Accurately perceiving object surface material is critical for scene understanding and robotic manipulation. However, it is ill-posed because the imaging process entangles material, lighting, and geometry in a complex way. Appearance-based methods cannot disentangle lighting and geometry variance and have difficulties in textureless regions. We propose a novel multimodal fusion method for surface material perception using the depth camera shooting structured laser dots. The captured active infrared image was decomposed into diffusive and dot modalities and their connection with different material optical properties (i.e. reflection and scattering) were revealed separately. The geometry modality, which helps to disentangle material properties from geometry variations, is derived from the rendering equation and calculated based on the depth image obtained from the structured light camera. Further, together with the texture feature learned from the gray modality, a multimodal learning method is proposed for material perception. Experiments on synthesized and captured datasets validate the orthogonality of learned features. The final fusion method achieves 92.5% material accuracy, superior to state-of-the-art appearance-based methods (78.4%).**

***Index Terms*—Material recognition, structured light camera, subsurface scattering, multimodal learning.**

I. Introduction

INFERRING objects' inherent material properties from captured images can provide a substantial understanding of the scene and benefit scientific areas including computer vision, computer graphics, and robotics. For example, although similar in appearance, a porcelain mug is much more fragile than a plastic one, requiring the robots to manipulate with carefulness. However, such image-based material recognition is ill-posed because the image perception process entangles material, geometry, and lighting in a complex way. With the ability of depth acquisition and active illumination, the structured light camera provides useful disentangling factors required for material recognition.

The key insight of using the structured light camera is that it provides an efficient probing function to detect the subsurface scattering effect, which describes how light penetrating the material surface is being scattered and exits the surface at a different point (see Fig. 1). Frequently observed in daily translucent surfaces like skins, plastics, marble, and wax, subsurface scattering acts as a discriminant feature for material classification. However, when illuminated by a diffusive light source, its characteristic point spreading function (PSF) is hard to be observed due to spatial integration. Therefore, a "probing function" is needed for the system diagnosis. Related works like [1], [2] modulate incident light temporally to implicitly reconstruct temporal PSF, and [3] modulates incident light spatially for binary material classification, showing the feasibility. Similar to [3], a speckle dot pattern projected by the structured light camera is used in this paper as a spatial "probing function". However, unlike their contrast-based method which is limited to fixed distance and flat surface, this paper aims to handle geometry and distance variance. In addition, the diffusive lighting, which is a by-product produced by the diffractive optical element (DOE) of the structured light projector, allows us to detect reflection features in the infrared (IR) spectrum.

Furthermore, a typical off-the-shelf structured light camera is not only equipped with an active IR sensor, but also an RGB sensor for visual acquisition. Such ambient light illuminated RGB image is more sensitive to material's texture, which is seriously damaged by the dot-pattern in IR image. Since the scattering and reflection feature learned from the IR image is orthogonal to the texture feature learned from the RGB image, the material discrimination power is increased by fusing these modalities. To validate the proposed method, both synthesized and captured datasets are collected for evaluation.

The main contribution of this paper is twofold. First, we show how the surface properties, including reflection and scattering, are related to the different components of the actively illuminated IR image. Second, a multimodal fusion method using

Manuscript received October 31, 2021; revised March 22, 2022; accepted April 20, 2022. Date of publication May 11, 2022; date of current version July 11, 2022. This work was supported in part by the Natural Science Foundation of China (NSFC) under Contracts 61860206003, 62088102, and 62171253, in part by Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20180507183706645, in part by the Beijing National Research Center for Information Science and Technology (BNRist) under Grant BNR2020RC01002, and in part by the China Postdoctoral Science Foundation under Grants 2020TQ0172, 2020M670338, and YJ20200109. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Miguel Rodrigues. *(Shi Mao and Mengqi Ji contributed equally to this work.) (Corresponding author: Lu Fang.)*

Shi Mao is with the Tsinghua-Berkeley Shenzhen Institute and the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: maos19@mails.tsinghua.edu.cn).

Mengqi Ji is with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China (e-mail: jimengqi@buaa.edu.cn).

Bin Wang is with the School of Electrical Engineering, Zhejiang University, Hangzhou 310051, China, and also with the Network and Information Security Laboratory of Hangzhou Hikvision Digital Technology Company, Ltd., Hangzhou, Zhejiang, China (e-mail: wangbin2@hikvision.com).

Qionghai Dai is with the Department of Automation, Tsinghua University, Beijing 100084, China, and also with the Institute for Brain and Cognitive Science, Tsinghua University (THUIBCS), Beijing 100084, China (e-mail: qhdai@tsinghua.edu.cn).

Lu Fang is with the Department of Electronic Engineering and the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: fanglu@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSTSP.2022.3171682

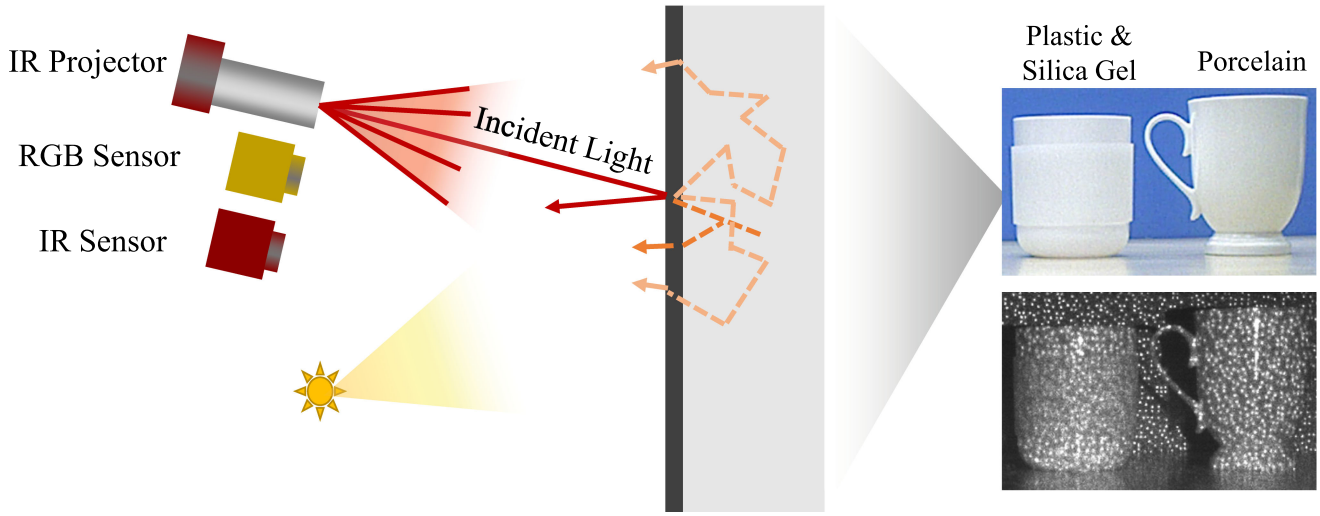

Fig. 1. Surface properties under different modalities. Although the difference between the porcelain mug and plastic one (with a silica gel protector) is limited in the RGB image, it's obvious in dot-pattern IR image with different point spreading effects. The underlying surface optical properties include reflecting (red arrow) and scattering (deep and light orange arrows).

all the images acquired from an off-the-shelf structured light camera is proposed based on the orthogonality of corresponding learned features. Both theoretical analysis and experimental validation are provided. Comparative experiments show that the multimodal fusion method is superior to state-of-the-art appearance-based methods (92.5% vs. 78.4%).

II. Related Work

Material recognition attracts a wide interest of researchers from areas of computer vision, computer graphics, and robotics. Related works can be grouped in the following categories:

Natural Appearance Methods. Typically, researchers using single image visual appearance classify materials by their characteristic features of texture, color, and context. The first group of works sought to use only local features (mainly textures and colors) to classify materials. Schwartz *et al.* [4] learned material attributes from local image patches for material representation. Zhang *et al.* [5] designed a special texture encoding network that utilized orderless representation for material and texture recognition. Xue *et al.* [6] extended the idea by encoding extra local spatial information in their DEP network. Such methods worked for materials with distinct textures, such as wood. But they struggled in recognizing texture-less yet geometrically deformable materials like paper and plastic.

Another group of works took non-local context information into account. Sharan *et al.* [7] introduced a pioneer material dataset FMD that contains images from Flickr. Bell *et al.* [8] introduced a larger dataset named MINC with materials in context. Methods including kernel-based nearest neighbor distance metric learning [9] and multiscale CNN [8] models are implemented. These approaches tended to rely on object-level information and took a shortcut by using object-level shape and context. However, such material-irrelevant information damaged the model's generalization performance. As pointed out by Sharan *et al.* [10], their material accuracy dropped when object-level features were removed. As our approach focuses on local material properties, we don't rely on object-level information, while addressing the texture-less problem by utilizing extra modalities provided by the structured light camera.

Reflection and Scattering Methods. To recognize material properties from the perspective of visual physical properties, methods based on reflection models were proposed. Bidirectional reflectance distribution function (BRDF), which describes the fraction of light coming out of the surface along a certain direction given its incident direction and surface normal, is a suitable representation for material recognition since it's invariant under different illumination and geometric structure. However, although simplified by Nielsen *et al.* [11], acquiring full measurement of BRDF is still complicated under control conditions [12]. Wang *et al.* [13] classified materials from BRDF slices collected by a hemispherical dome encircling the objects. Several works sought to optimize the feature collection effort for material classification [14], [15]. However, BRDF cannot model mesoscopic features like self-occlusions and inter-reflection [12]. To accommodate it, the Bidirectional texture function (BTF), an image-based representation that describes fine-scale appearance was suggested by Dana *et al.* [16]. A BTF dataset for material classification was introduced by Weinmann *et al.* [17], including both synthesized and measured data. However, the feature acquisition and storage were still complex although being optimized [18], making it unpractical in applications like robotics.

Another group of works gave up the complicated equipments and resorted to reflection variance detection under different viewing angles and wavelengths in the wild. Based on the difference in viewing angle, Wang *et al.* [19] proposed a 4D light-field dataset for material recognition, claiming that multiple sub-aperture views and view-dependent reflectance benefits material recognition. Xue *et al.* [20] captured images under a small angular variation to extract angular-gradient features to improve recognition. On the other hand, from the view of

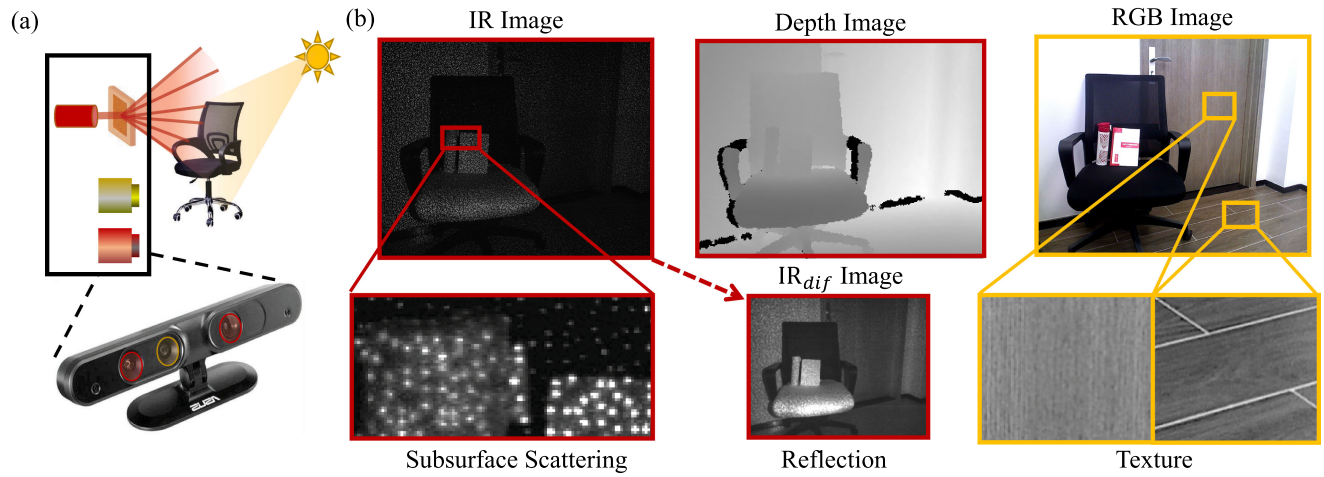

Fig. 2. (a) The physical imaging process. Infrared (IR) light is actively projected by the IR projector and received by the IR sensor after being reflected and scattered (red). While visible ambient light is received by the RGB sensor (yellow). (b) Different materials' properties are related to different images. On the top are raw images directly accessible from the structured light camera. Notice the reflection property is revealed in the IR$_{dif}$ image, which is recovered from the IR image by using the method described in Sec. IV-A2.

multi-spectral, Salamati *et al.* [21] classified materials by hand-craft feature obtained in NIR and RGB spectrum. Saragadam *et al.* [22] learned a hyper-spectral classifier by programmable spectral filters before it was recorded on CMOS, improving both acquisition speed and SNR ratio. However, without active illumination, their performance could be hindered by different ambient light.

Instead of describing materials by reflection properties, Su *et al.* [1] proposed to reconstruct the temporal point spread function (TPSF) from the raw data captured from the time of flight (TOF) camera to characterize the temporal scattering interaction of light and material. Similarly, Tanaka *et al.* [2] acknowledged the significance of TPSF and proposed to use the resultant depth distortion as a substitute for raw measurement, making the method more practical. However, owing to the multi-path effect of the TOF camera, these methods lack robustness in geometric variation. Steimle *et al.* [3] used the spatial point spreading function to distinguish hand and hard surfaces. However, they did not account for geometry and distance variation and only performed binary classification based on local image contrast. In this work, we not only learn material optical features (reflection and scattering) using a portable structured light camera but also disentangle geometry variation using depth images obtained from the camera.

Multimodal Methods. To take benefits from different modalities, multimodal fusion methods are proposed. Zheng *et al.* [23] proposed a deep learning method to use both contact (haptic) and non-contact (visual) information to classify materials. DeGol *et al.* [24] aligned 2D images with 3D geometry to proposed a geometry-informed material recognition method. Erickson *et al.* [25] mounted a high-resolution camera and near-infrared spectrometer on the PR2 robot's hand to classify materials based on texture and spectroscopy. By fusing different modalities, reported methods were more powerful by examining material properties from different aspects. Our approach validates such addable multi-dimensional discrimination power by fusing different modalities (dot, diffusion, and gray) accordingly.

III. Structured Light Camera and Material Properties

In this section, we relate observed IR images captured by the structured light camera to different material optical properties. Specifically, subsurface scattering and direct reflection properties are revealed under the mixed illumination pattern projected by the structured light camera. As shown in Fig. 2, by observing the subsurface scattering effect, the plastic bottle cap is distinguishable from the upper-right part of the notebook in the raw IR image, although both are red in the RGB image. By observing the reflection effect, the chair's cushion is distinguishable from the chair's back in the 'smoothed' IR image (termed as IR$_{dif}$ in Sec. IV-A2), although both are black in the RGB image. However, when textures are considered as a material feature, the RGB image shows better discrimination power since it is illuminated by low-frequency ambient light. The physical imaging process of IR images (involving material optical properties, geometry, and lighting) is analyzed in this section.

A. Material Optical Properties

Following Jensen *et al.* [26], the general bidirectional surface scattering distribution function (BSSRDF) models the relationship between outgoing radiance and incoming radiant flux from different angles and locations, accounting for direct reflection and subsurface scattering. Such outgoing radiance L_o at point x from direction $\boldsymbol{w_o}$ can be split into two terms: the direct reflection term L_r and subsurface scattering term L_s [27].

$$L_o(x, w_o) = L_r(x, w_o) + L_s(x, w_o) \tag{1}$$

For subsurface scattering, Jensen *et al.* [26] decompose it into single and multiple (diffuse) terms. Since single-scattering for optically dense materials decreases much faster than multiple scattering as the distance between incident point x_i and outgoing point x_o increases, it attributes little to the overall outgoing radiance. The widely adopted dipole method models the outgoing radiance of multiple scattering as a spatial convolution of

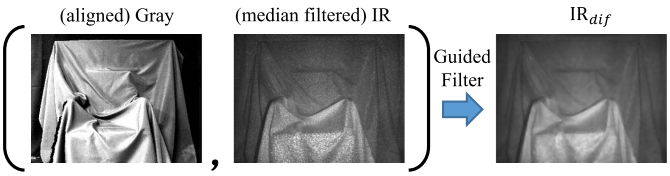

Fig. 3. Recovering IR_{dif} by filtering (median filtered) IR image guided by the (aligned) gray image. Notice both the median filtered IR image and IR_{dif} image was enhanced 4x in intensity for better visualization.

a PSF and incoming irradiance on the object surface. Where the material-related PSF $R_d(||x - x_i||; \sigma_a, \sigma_s, p)$ is a shift-invariant function of material's absorption coefficient σ_a, scattering coefficient σ_s and phase function p. The final subsurface scattering term can be integrated as:

$$\begin{aligned} L_s(x, w_o) &= L_1(x, w_o) + \int_A R_d(||x - x_i||)E(x_i)dA(x_i) \\ &\approx R_d(x) * E(x) \end{aligned} \tag{2}$$

where $E(x)$ is the incident irradiance at surface point x, and L_1 is the single-scattering term. Following [28], we ignored the Fresnel term in above equation because such Fresnel effect cannot be observed under our setting where incident and outgoing light directions are approximately the same, i.e. $w_i \approx w_o$, (see detailed analysis of this approximation in Sec. III-B).

The direct reflection is well studied by the bidirectional reflectance distribution function (BRDF), which describes the direct reflection without penetration, i.e. $x_i = x$. The famous physical-based Cook-Torrance model parameterizes BRDF with diffuse albedo k_d, specular albedo k_s and surface roughness r:

$$L_r(x, w_o) = k_d E(x) + k_s \int_{2\pi} S_p(n, w_i, w_o; r) L_i(x_i, w_i) dw_i \tag{3}$$

where S_p, parameterized by material roughness r, is the specular term describing the specular reflectance under surface normal n, and L_i is the incident radiance. Notice that since the pixels corresponding to strong specular reflection angles are corrupted by over-exposure, the structured light camera cannot calculate their valid depth. This effect seriously damages the local geometry information. Therefore, in this paper, we only focus on materials with relatively high roughness to avoid such depth incompleteness.

Finally, suppose the projection of a surface point x on the image is x^p. When enough resolution is provided, the intensity for this pixel, denoted as $I(x^p)$, is approximately proportional to the outgoing radiance.

$$I\left(x^p\right) \propto L_o\left(x, w_o\right) \tag{4}$$

B. Structured Light Camera

Structured light cameras are wildly used in object 3D sensing. By projecting a known pattern onto a scene and analyzing the reflected distorted patterns, a typical off-the-shelf structured light camera can calculate the depth of the scene being imaged. As shown in Fig. 2, such a camera is equipped with a projector emitting patterned infrared light and a sensor with a filter in the corresponding spectrum (usually the infrared spectrum, red-tinted in Fig. 2). Usually, an extra calibrated RGB sensor (yellow-tinted in Fig. 2) is equipped for general visual perception under ambient light.

While the projecting pattern varies among different cameras, the Primesense designed collimated dot pattern is the most commercially successful one. This pattern is widely adopted in Kinect-V1 (Microsoft), Xtion (Asus), and iPhone (Apple). This spatial multiplexing speckle dot pattern was created through diffracting the laser emitted light by a diffractive optical element (DOE). It provides an efficient probing function for spatial PSF. As a byproduct of the pseudorandom dot pattern, the diffuse lighting and 0-th order highlight are also created by DOE [29].

For subsurface scattering, the dot-pattern illumination provides multiple probing functions for the diffusive PSF kernel R_d. Specifically, the radiant of each single collimated laser beam can be modeled as a 'δ-function' with fixed intensity L_{dot}. Therefore, the irradiance on material surface point x being illuminated is only related to the incident direction:

$$E_{dot}(x) = L_{dot} M(x) \langle n, w_i \rangle \tag{5}$$

where $M(x)$ is a sparse mask indicating whether the surface point x is illuminated by the dot-pattern illumination, w_i is the direction from the surface point to the projector center, n is the normal of the surface at point x, and $\langle \cdot, \cdot \rangle$ denotes inner product operator. Since the designed energy of the outgoing light exceeds the dynamic range of the image sensor, it is unable to distinguish different reflection intensities.

On the contrary, diffusive isotropic illumination is suitable for direct reflection feature learning. Modeling such isotropic illumination as a point light source with fixed radiant intensity I_{dif}, the radiance at surface point x attenuates following an inverse square law of its distance to the light source, denoted as r:

$$E_{dif}(x) = \frac{I_{dif}}{r^2} \langle n, w_i \rangle \tag{6}$$

A simplified model considering only the diffuse subsurface scattering and diffuse direct reflection around a local region on the surface can be formulated as:

$$\begin{aligned} L_o(x, w_o) &\approx R_d(x) * E(x) + k_d E(x) \\ &= (R_d(x) + k_d \delta(x)) * (E_{dot}(x) + E_{dif}(x)) \\ &\approx (R_d(x) + k_d \delta(x)) * E_{dot}(x) \\ &\quad + (R_d * 1(x) + k_d) \frac{I_{dif}}{r^2} \langle n, w_i \rangle \\ &= R'_d(x) * M(x) L_{dot} \langle n, w_i \rangle + k'_d \frac{I_{dif}}{r^2} \langle n, w_i \rangle \end{aligned} \tag{7}$$

where $1(x)$ is a function with constant value one to approximate local diffusive illumination, $k'_d = R_d * 1(x) + k_d$ denotes augmented diffuse albedo, and $R'_d(x) = R_d(x) + k_d \delta(x)$ denotes augmented PSF. The first part of the formula involves a convolution of augmented local material PSF $R'_d(x)$ with dot-pattern illumination mask $M(x)$, while the second part is simply the

diffuse radiance amplified by a factor of augmented diffuse albedo k'_d.

To take a deeper look at these equations, we can decompose images by different illumination conditions. Consider an infrared image IR_{dif} taken under diffuse illumination only, i.e. $E(x) = E_{dif}(x)$, it would be difficult to distinguish material's sub-surface scattering properties because in the second term such spatial PSF information is lost due to a convolution with similar local illumination, although being easy to distinguish augmented diffuse albedo k'_d (which is dominated by diffuse albedo k_d by value). On the contrary, for an infrared image IR_{dot} taken under dot-pattern illumination only, i.e. $E(x) = E_{dot}(x)$, the diffuse albedo contributes to the pixel's intensity only on those direct illuminated points, whose value is usually over-exposed under limited dynamic range. It would also be difficult to distinguish such albedo properties under dot-pattern illumination only, although being easy to capture spatial PSF by probing it with an illumination mask. Therefore, it would be beneficial to use both of them for better material recognition ability.

Notice that since the spatial convolution happens on the material surface instead of the image plane, the PSF kernel cannot be correctly recovered simply by deconvolution. Both the shape and distance of the surface deform the point spreading effect observed on the image, not to mention the anisotropic single-scattering term. Specifically, as the distance increase, the projected PSF kernel will shink inverse proportionally (i.e. approximately at rate $1/d$, where d is the depth from the surface point to the sensor's image plane), and the PSF will be tilted by different shapes.

IV. Method

Our method uses reflection, scattering, and texture features for material classification. A geometry term is derived from the lighting model and calculated from the depth image to disentangle the geometry-induced variation on the observed image.

A. Image Preprocessing

1) Geometry Term: As analyzed in Sec. III-B, under fixed lighting, the observed IR image deeply entangles both material and geometry of the objects. To classify materials regardless of their geometry, several factors are needed for disentanglement. Specifically, it includes the inverse of depth $1/d$, the inverse square between each point and the center of the projector $1/r^2$, and the inner product between surface normal and incident light direction $\langle n, w_i \rangle$. The last two factors need to calculate the incident light direction. Since the commercial structured light camera has a relatively small baseline (several centimeters), concerning the scale of the scene (several meters), it's reasonable to approximate the incident light direction w_i with the outgoing light direction w_o. The latter is easy to calculate given the camera's FOV and pixel location. We term the collection of such information as the geometry modality and use it as input for disentangling the geometry factors for material classification.

Specifically, we employ a conventional definition of camera space with the origin point locating in the camera's optical center, Z axis pointing towards the image center, and X, Y axes aligning with the camera's edges. The projection of material surface point $x(x, y, z)$ on IR image is $x^p(u, v)$, where (x, y, z) and (u, v) are their coordinates in camera space and image plane respectively. By stereo vision calculation, each pixel in IR image is assigned with a depth value $d = z - f$ representing the distance from its corresponding surface point x to the image plane, where f is the camera's focal length. When given the ratio between focal length and the pixel's width and height of IR sensor, termed as f_x, f_y respectively, the un-normalized surface normal direction of surface point x can be calculated from its projected location x^p:

$$n'(u, v) = \left[\frac{\mathrm{d}z}{\mathrm{d}x}, \frac{\mathrm{d}z}{\mathrm{d}y}, -1\right]^T = \left[\frac{f_x}{z}\frac{\mathrm{d}z}{\mathrm{d}u}, \frac{f_y}{z}\frac{\mathrm{d}z}{\mathrm{d}u}, -1\right]^T \tag{8}$$

Since d and z differ only by a constant value, their differential is the same. Therefore $\frac{\mathrm{d}z}{\mathrm{d}v}$ and $\frac{\mathrm{d}z}{\mathrm{d}u}$ can be calculated by applying differential operator on whole depth image. In our implementation, we use vertical and horizontal Sobel operators to calculate such differentials. The un-normalized out-going direction from surface point x to camera sensor can be calculated from its projected location x^p:

$$w'_o(u, v) = \left[\frac{x}{z}, \frac{y}{z}, 1\right]^T = \left[\frac{u}{f_x}, \frac{v}{f_y}, 1\right]^T \tag{9}$$

Since we approximates w_i with w_o, we calculate the desired $\langle n, w_i \rangle$ for every pixel x^p in image plane by first normalize the above un-normalized version by their L2 norm and then take inner product.

$$\langle n, w_i \rangle \approx \langle n, w_o \rangle = \left\langle \frac{n'}{||n'||_2}, \frac{w'_o}{||w'_o||_2} \right\rangle \tag{10}$$

Similarly, we approximate IR camera's origin point as the light source point to calculate the distance from surface point x to diffusive light source point as $r(u, v) \approx ||z w'_o(u, v)||_2$. By taking element-wise inverse square for all pixels in the image, we obtain the radiance attenuation factor $1/r^2$. Finally the PSF kernel shrinking factor $1/d$ is calculated by element-wise inversion of the depth image.

Notice every factor is calculated as a single channel image with the same size as IR image, we channel-wise stack them together to generate a three-channel geometry modality image. Specifically, if the IR image has shape (H, W) the stacked geometry modality image is a $(3, H, W)$ tensor, with $\langle n, w_i \rangle, 1/r^2$ and $1/d$ image being its channels sequentially.

2) IR Image Processing: As analyzed in Sec. III-B, the observed infrared image IR is a superposition of diffuse-illuminated image IR_{dif} (termed as diffusion modality) and dot-pattern-illuminated image IR_{dot} (termed as dot modality). Since the dot-pattern dominants the overall illumination, its corresponding image also dominate the observed image, i.e.

$$IR = IR_{dot} + IR_{dif} \approx IR_{dot} \tag{11}$$

To better distinguish material's albedo properties, the weak diffusion modality IR_{dif} needs to be explicitly recovered. A simple way to address the problem is to consider the observed infrared image as a distorted version of diffusion modality, being

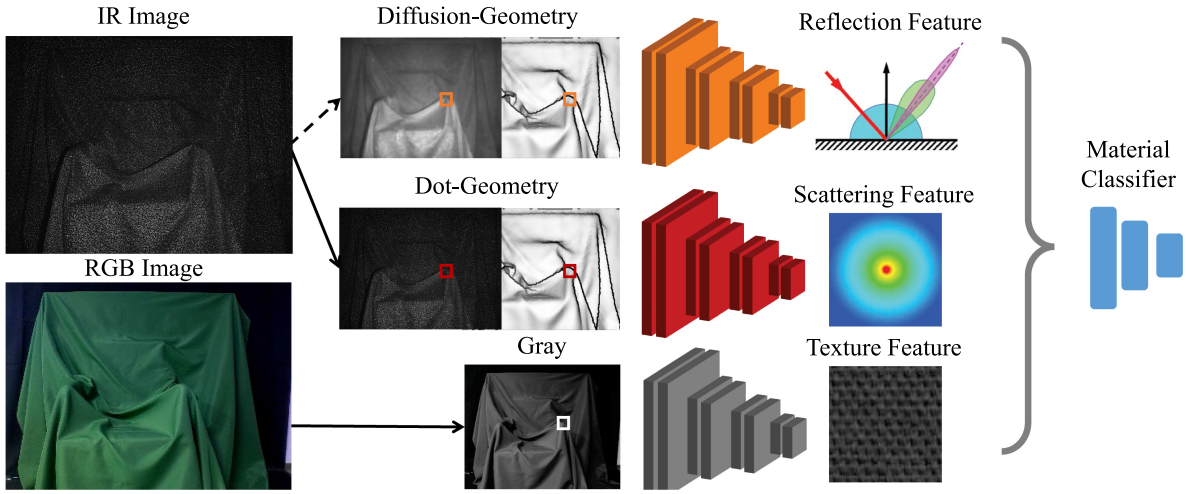

Fig. 4. An overview of method for material classfication.

distorted by strong "salt-and-pepper" noise. However, the noise cannot be fully removed simply by median filtering because it is not randomly distributed, but spatially related due to PSF. Here we use guided image filtering [30] to transfer edge clues from gray-scaled image (averaged from aligned RGB image described in Sec. IV-A3) to the median filtered IR image. This allows us to remove the dot-pattern while preserving the sharp edges. We directly use the observed IR image as dot modality since it is the dominant component, and validate this choice by experiments.

3) RGB Image Processing: Being corrupted by dot-pattern, the recovered diffusion modality cannot capture the fine texture of materials, which is related to different self-occlusion of the surface micro-geometry. For example, the texture of fur and fabric are significant material features, but not fully captured by the optical properties described above. Here we use gray-scaled RGB image (termed as gray modality) for texture feature learning. We use gray modality instead of the raw RGB image to avoid color variance induced by ambient lighting and dyeing, which is invariant to materials.

Also, notice that the RGB image is not aligned with the IR image, because they are captured by different cameras. To avoid such misalignment, we need to align images from RGB image space to IR image space. We perform this alignment instead of the other way around because depth is calculated in IR image space. Specifically, for every pixel in IR image, we first calculate its corresponding coordinate in RGB image using their depth value and RGB-IR camera calibration parameters, and then assign its value by interpolation of neighboring RGB values. The final gray modality is averaged over the aligned RGB images.

B. Material Classification Model

The overall material classification process is depicted in Fig. 4. We first recover diffusion modality from IR image and directly use IR image as dot modality as described in Sec. IV-A2. As analyzed in Sec. III-B, both direct reflection and subsurface scattering entangles deeply with geometry to produce corresponding diffusion and dot modalities. We disentangle the geometry variance by taking it as an input to corresponding feature extractors. Specifically, we channel-wise stack diffusion and geometry modalities (denoted as diffusion-geometry modalities in Fig. 4), from which the material reflection features are extracted using ResNet-based convolutional neural networks (constitutes of four residual blocks followed by 2×2 max-pooling layer each). Similarly, the material scattering features are extracted from the channel-wise stacked image of dot and geometry modalities (denoted as dot-geometry modalities in Fig. 4). For texture feature learning, aligned gray modality is directly fed to the same feature learning networks described above.

The resulting reflection, scattering, and texture features are then stacked altogether to perform the final material classification through a fully connected network, which has two hidden layers with 512 and 128 neurons each. The final output layer has the dimension of materials types. Since materials with limited specularity are considered, we apply a local patch-based classification method for material recognition. Patches with size $P \times P$ on the same location of aligned modalities are cropped and fed into the networks.

V. Experiment

We collect both synthesized and captured datasets to validate the proposed material classification method.

A. Data Collection

1) Synthesized Dataset: Firstly, to validate that the reflection and scattering features are learned under known active illumination, a synthesized dataset is rendered using Blender, a physical-based rendering software. The illumination is modeled by assembling an isotropic point light and a dot-pattern projector that projects the inverse engineered Primsense designed pattern. Benefiting from the controllable lighting, we rendered two images under the isotropic and dot-pattern illumination respectively. As shown in Fig. 5 a, a camera with the same FOV and resolution as Xtion IR camera is coupled rigidly with the projector at a baseline of 6 cm. To align with the actual dynamic range of the Xtion sensor, we clipped the received intensity

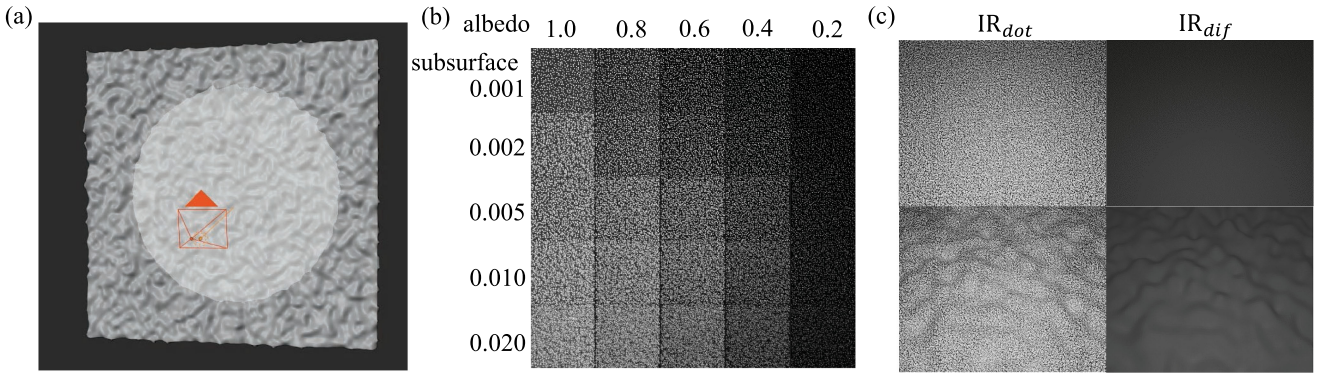

Fig. 5. Settings for synthesized dataset and different generated material properties. (a) Modeled IR camera, IR projector, and random bumpy planes in Blender. Both camera parameters and projected pattern are aligned with the Xtion sensor. (b) Materials with different properties are illuminated by mixed lighting. Different rows differ in subsurface scattering properties and different columns differ in reflection albedo properties. (c) IR_{dot} and IR_{dif} images rendered under their corresponding illumination, surfaces are either flat or random bumpy. Diffusion images are enhanced for better visualization.

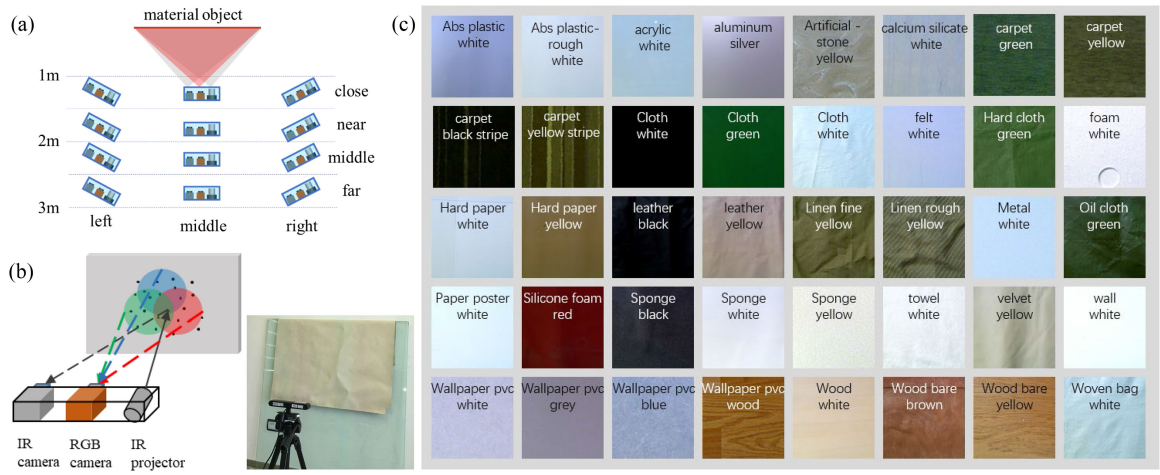

Fig. 6. Settings for captured dataset and selected materials. (a) Images are taken from 4 different depth ranges and 3 different viewing angles in the training set. Additionally, images taken from two random angles in all depth range constitute the testing set. (b) The Xtion sensor model and capturing setting. (c) Captured 40 different objects, with 30 distinct materials.

being over-exposed. With a physical-based rendering engine, Blender allows modification of object material properties like diffuse albedo, roughness, and optical depth for subsurface. In experiments, material varies in 5 different albedo values uniformly sampled in range 0.2-1 and 5 different optical depth values (subsurface scattering parameter) exponentially sampled in range 0.001-0.02 (Fig. 5 b). To disentangle geometry differences, the objects include both flat and random bumpy planes (Fig. 5 c). For each material in the training set, images are rendered at 4 different distances within the range of 1-4 m, which is a typical working range for the indoor structured light camera like Xtion. Also, 9 different angles are sampled uniformly from the range of $\pm 30^{\circ}$. In the testing set, two distances and angles are selected randomly within the described range.

2) Captured Dataset: To test the performance on real-world materials, we captured images by Xtion sensor. The IR camera of Xtion is equipped with a band-pass filter that is only sensitive to emitted 830 nm laser light while filtering out the visible spectrum. The horizontal and vertical fields of view are 58° and 45° respectively. The raw image resolution for RGB and IR cameras is 1280×1024, and the depth resolution is 640×480. Since the PSF spans limited space, the images under the highest resolution are captured in our experiment.

As shown in Fig. 6, We collected images from 40 different objects, with 30 distinct materials. several materials like cloth, carpet, and sponge are collected with different samples with varying colors and textures. The sampled materials cover a wide range of material variance including texture, transparency, reflectance, and subsurface scattering properties. To disentangle geometry difference, all materials are collected under 4 different depth ranges within 1-3 m, each with 3 different viewing directions (approximately left, middle and right) under flat and deformed shape (if deformable), resulting in 24 images for each material in the training set. The testing set is collected following the same protocol, except each material is sampled from 2 random angles.

B. Classification Experiments

1) Experiment on Synthesized Dataset: To validate the connections between different modalities and material properties in a controlled manner, we test the classification performance

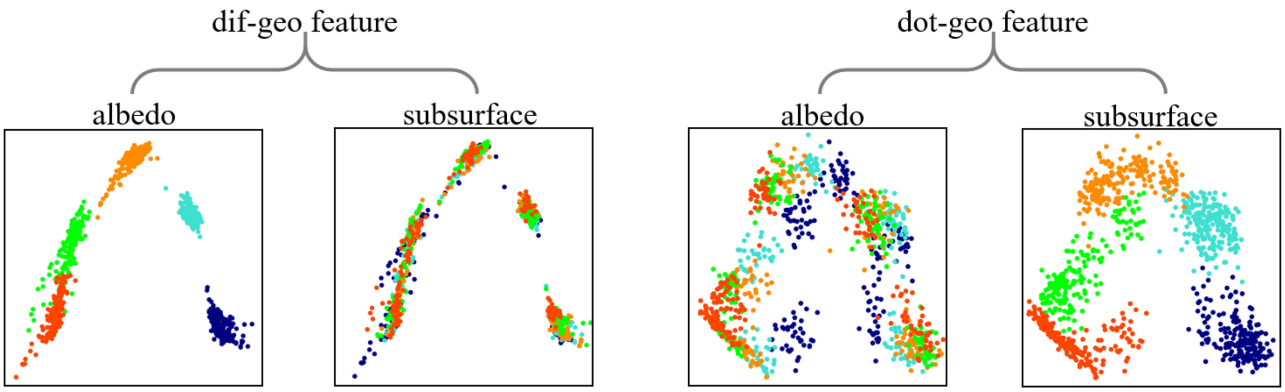

Fig. 7. PCA Visualizations of features learned from different material properties. Features learned from dif-geo modalities, denoted as the 'dif-geo features,' exhibit a clear clustering structure for material's albedo (but not for subsurface) on their major PCA directions. While the 'dot-geo features' behave just the opposite.

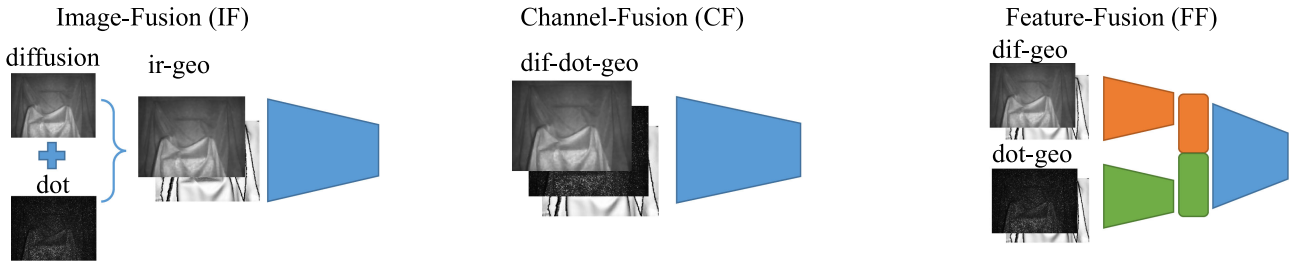

Fig. 8. Different fusion methods. Image-fusion method first element-wise adds diffusion and dot modalities to generate a synthesized 'ir' image, and then channel-wise concatenates ir modality (1 channel) with geometry modality (3 channels). After that, a CNN is attached for classification. Channel-Fusion uses 'early fusion' paradigm, which channel-wise concatenates all modalities before forwarding to a CNN classifier. Feature-Fusion method uses 'late fusion' paradigm, which channel-wise concatenates the features extracted from dif-geo modalities and dot-geo modalities before forwarding to a CNN classifier.

TABLE I
PERFORMANCE ON SYNTHESIS DATASET W/O GEOMETRY

Metric	dif-geo	dot-geo	diffusion	dot
Overall ac.(%)	22.8	**58.9**	4.3	34.0
subsurface ac.(%)	23.3	**96.9**	21.5	86.2
albedo ac.(%)	**97.7**	60.8	19.9	39.5

TABLE II
PERFORMANCE ON SYNTHESIS DATASET WITH DIFFERENT FUSION METHODS

Metric	IF	CF	FF
Overall ac.(%)	65.6	**96.0**	**96.0**
subsurface ac.(%)	98.3	**99.5**	99.0
albedo ac.(%)	66.7	96.5	**97.0**

on the synthesized dataset. We abbreviate diffusion and geometry modalities as 'dif' and 'geo' modalities respectively when modalities are channel-wise stacked together as input (see Sec.IV-B). As shown in Table I, Although the overall accuracy for the model using dif-geo is 22.8%, it achieves 97.7% on albedo accuracy, indicating that such modalities have good discrimination power on material's albedo property, while confused by its subsurface property. On the contrary, model taking dif-geo as inputs behaves just the opposite, showing a discrimination power in different subsurface properties.

This finding is further validated by visualizing the learned features using principal component analysis (PCA). Here the features refer to the flattened output vectors of the feature learning networks described in Sec.IV-B, before being fed to the downstream classifier. As shown in Fig. 7, in dif-geo feature space, similar albedo property are clustered together. Within each group, the subsurface property distributes almost uniformly. An exactly opposite behavior appears on dot-geo feature space. This validates the connection between learned features and material properties.

To validate the necessity of introducing the geometry term in dot and diffusion modalities, an ablation study is performed. The overall accuracy for models using only dot and dif modalities are 34.0% and 4.3% respectively, both suffer from a significant performance drop, compared to the geometry-stacking version. This performance drop validates the necessary to accounts for geometry variance. In the following experiments, we use geometry-stacked versions for diffusion and dot modalities by default.

To validate the necessity of explicitly recovering diffusion modality, we test different fusion methods. As shown in Fig. 8, fusion methods includes image-fusion (IF) that image-wise add diffusion and dot modalies before stacking with geo modality, channel-fusion (CF) that channel-wise concatenate diffusion, dot and geo modalities, and feature-fusion (FF) that channel-wise concatenate dif-geo and dot-geo features. Notice in image-fusion, we do not explicitly recover diffusion modality by using a synthesized 'raw IR' image. While for the last two fusion methods, an explicitly recovered diffusion modality is required. As shown in Table II, since image-fusion is dominated by

TABLE III
PERFORMANCE ON CAPTURED DATASET W/O GEOMETRY MODALITY. ACCURACIES ARE CALCULATED FOR 40 SAMPLES AND 30 MATERIALS SEPARATELY, THE INNER-MATERIAL SIMILARITY GAIN MEASURES THE DIFFERENCE BETWEEN THESE TWO ACCURACIES

Metric	without geo			with geo			fusion (with geo)			
	dot	dif	gray	dot	dif	gray	dot-dif	dot-gray	dif-gray	all
Sample ac. (%)	51.9	28.0	**64.9**	62.4	35.8	**67.8**	64.2	90.7	81.7	**90.8**
Material ac. (%)	59.3	33.4	**66.9**	**70.0**	41.4	69.9	71.70	92.2	83.8	**92.5**
inner-mat. gain (%)	**7.4**	5.3	2.0	**7.6**	5.7	2.2	**7.5**	1.5	2.1	1.7

dot-pattern, it has a similar performance of dot-geo modalities. On the contrary, both channel-fusion and feature-fusion benefit from the explicit recovered diffusion modality, leading to a better performance on albedo accuracy.

2) Experiment on Captured Dataset: Once the physical connection between different modalities and corresponding properties is validated on the synthesized dataset. We examine the proposed method on the captured dataset for real-world application. Since we sampled 40 objects with 30 different materials in the captured dataset, we train our model only on the objects sample's label, and test the accuracy on their corresponding object samples and material labels. This training method tends to 'over-classifying' the materials, but provides a method to check whether the learned models classify material based on material-invariant features, e.g. color and geometry bias. We denoted the difference between material accuracy and sample accuracy as inner-material similarity gain. Given similar sample accuracy, the larger the gain is, the better the samples from the same materials are clustered together. Indicating that the learned model tends to perform classification based on the joint feature of the same material instead of the inner-material difference.

As shown in Table III, when being trained using single modalities without stacking with geometry, model using gray modality have the highest sample accuracy, we attribute it to the fact that texture difference is significant in the captured dataset. However, models using dot modality have the highest inner-material similarity. Indicating that the model using dot modality is more robust in inner-material variance, capturing good material-related properties. When stacking with geometry, models using dot-geo and dif-geo modalities exhibits significant inner-material gain, compared to gray-geo modalities. These results aligned with the synthesized experiments and theoretical analysis, where the dot and diffusion modalities derived from the IR image is illuminated by known pattern, and the material's optical properties can therefore be faithfully inferred by disentangling the geometry factors involved in the imaging process. On the other hand, the gray image is illuminated by uncontrolled ambient light, making it difficult to disentangle the material's optical properties even given geometry factors. We attribute the insignificant performance gain by stacking geometry term on gray modality to the fact that it helps the model to explain the depth-related texture variation.

To further validate that IR modalities can handle difficult textureless cases while gray modality cannot, we select 10 textureless white materials from the captured materials, including materials like white plastic, paper, sponge, wall, cloth, and foam. The model using gray modality achieves 63.5% material accuracy, compared with 92.4% for the model using dot-dif-geo. It seems the model using gray modal struggles to find useful texture differences for these materials, while the optical subsurface and reflection differences among sampled materials are more significant.

TABLE IV
COMPARING WITH STAT-OF-THE-ART METHODS

Methods	Sample ac.(%)	Material ac. (%)
DEP w/o pretrain	75.0 / 67.3	78.4 / 71.1
DeepTEN w/o pretrain	72.9 / 67.1	75.1 / 70.5
Gray (ours)	64.9	66.9
Fusion-all (ours)	90.8	92.5

Since the features learned from the three modalities lie in different dimensions, we fuse them in a channel-fusion fashion to evaluate the classification performance. As shown in Table III, on average, the pair-wise fusion exhibits performance gain compared with single modal. The method fusing all modalities achieves the highest 92.5% on material accuracy.

3) Comparing With State-of-the-Art Methods: We compare the proposed method with state-of-the-art material classification methods like DeepTEN [5] and DEP [6], [20]. Since we are the first to use the structured light camera and none of these methods incorporate dot and diffusion modalities described above, we compare the performance using gray images only. Both of these state-of-the-art models take pretrained ResNet18/50 as the backbone, we validate their performance with and without using the pretrained weights to have a fair comparison with our method. The comparisons are summarized in Table IV. With a dedicated design on texture feature recognition and pretrained backbone, DEP and DeepTEN outperform our method with better texture recognition power. However, the performance using only texture information (78.4%) is not comparable to our fusion method which incorporates both textures, reflection, and scattering information (92.5%), indicating the significance of extra modalities. Further ablation study indicates the superior performance of DEP and DeepTEN relies largely on pretraining for better texture feature extraction rather than dedicated network design. It is worth pointing out that the purpose of this work is not to fully exploit the texture feature but to utilize extra optical material features, whose performance is shown by the fusion-all model. Also, the performance gain by using pretrained weights suggests that a larger real-world dataset that incorporates both RGB and IR images is beneficial for multi-modal feature learning.

4) Segmentation Experiment: Beyond the patch-based CNN classification, one can easily apply a sliding window to the full image to achieve material segmentation. In this experiment, we instead train our classifier in a fully convolutional way and run

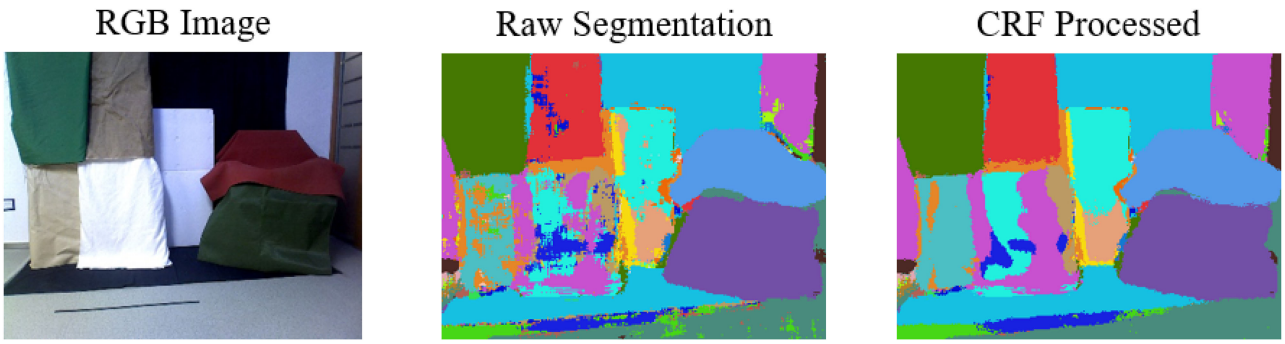

Fig. 9. Segmentation result of the proposed method. The raw segmentation (middle) lacks local consistency, which is mitigated by CRF backend (right).

a single fully convolutional network (FCN) at test time. Since such patch-based training methods lack local consistency on the output, a conditional random field (CRF) was attached as a backend to provide the final output. The Result is shown in Fig. 9.

VI. Discussion and Future Direction

We propose a multimodal fusion method for material recognition using an off-the-shelf dot-pattern projecting structured light camera. Leveraging the illumination pattern that consists of both isotropic and dot-pattern ones, we decompose the captured IR image and relate different material optical properties (i.e. direct reflection and subsurface scattering) to them separately. A geometry modality, that helps disentangle material properties form geometry variations, is derived from the rendering equation and calculated based on the depth image obtained from a structured light camera. Further, together with the texture feature learned from gray modality, a multimodal fusion model is proposed for material classification. We collected data by both physical based rendering and real-world capturing. Using the controlled synthesized dataset, we reveals the connections between IR modalities and their corresponding materials properties being derived from the theoretical analyzing. The necessity of introducing geometry term and explicitly recovering diffusion modality is validated using the synthesized dataset. And the multimodal fusion performance is evaluated using the captured real-world dataset. The final fusion-all methods achieves best performance in our dataset, and an extended segmentation methods is implemented.

Comparing with the state-of-the-art methods, the key priorities of the proposed method are as follows: (1) Superior in textureless surface recogintion. By using active dot pattern illumination, materials with different subsurface scattering effects are distinguishable, even for texture-less surfaces. Experiments on a white subset of captured dataset validate this priority by showing that model using gray modality only achieves 63.5% material accuracy, while model using dot-dif-geo achievese 92.4%. (2) All-Day availability. Since IR image is taken under active illumination emitted by the structured light camera itself, the fusion model using dot and diffusion modalities would keep working at night without ambient light. While it is not possible for methods using RGB image only (e.g. Deep-TEN and DEP). (3) Multi-dimensional descrimination power. By showing connections among different imaging modalities (dot, diffusion, and gray) and different material properties (subsurface scattering, reflecting, and texture), the discrimination power of the fusion model is addable. The full multimodal model outperforms its components and other state-of-the-art methods using texture properties only.

A major limitation is that materials with significant specularity can not be recognized properly, due to overexposure. An interesting future direction would be to move from material classification to inverse rendering. Instead of learning the discriminate feature for direct reflecting, multiple scattering and texture for classification, a descriptive representation that enables scene generation under different lighting and viewpoints are more powerful. And we believes such descriptive representation that only relates to object's intrinsic material properties can promote not only the scene understanding in computer vision domain, but also applications in robotics and computer graphics domains.

References

[1] S. Su *et al.*, "Material classification using raw time-of-flight measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3503–3511.
[2] 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.*, 2017, pp. 79–88.
[3] J. Steimle, A. Jordt, and P. Maes, "Flexpad: Highly flexible bending interactions for projected handheld displays," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2013, pp. 237–246.
[4] G. Schwartz and K. Nishino, "Recognizing material properties from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1981–1995, Aug. 2020.
[5] H. Zhang, J. Xue, and K. Dana, "Deep TEN: Texture encoding network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 708–717.
[6] J. Xue, H. Zhang, and K. Dana, "Deep texture manifold for ground terrain recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 558–567.
[7] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?," *J. Vis.*, vol. 9, no. 8, pp. 784–784, 2009.
[8] 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.*, 2015, pp. 3479–3487.
[9] D. Hu, L. Bo, and X. Ren, "Toward robust material recognition for everyday objects," in *Proc. Brit. Mach. Vis. Conf.*, 2011, Art. no. 48.
[10] 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, 2013.
[11] J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi, "On optimal, minimal BRDF sampling for reflectance acquisition," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 186.

[12] D. Guarnera, G. C. Guarnera, A. Ghosh, C. Denk, and M. Glencross, "BRDF representation and acquisition," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 625–650, 2016.

[13] O. Wang, P. Gunawardane, S. Scher, and J. Davis, "Material classification using BRDF slices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2805–2811.

[14] M. Jehle, C. Sommer, and B. Jähne, "Learning of optimal illumination for material classification," in *Proc. Joint Pattern Recognit. Symp.*, 2010, pp. 563–572.

[15] C. Liu and J. Gu, "Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral BRDF," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 86–98, Jan. 2014.

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

[17] M. Weinmann, J. Gall, and R. Klein, "Material classification based on training data synthesized using a btf database," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 156–171.

[18] J. Filip and M. Haindl, "Bidirectional texture function modeling: A state of the art survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1921–1940, Nov. 2009.

[19] 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 *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 121–138.

[20] J. Xue, H. Zhang, K. Nishino, and K. Dana, "Differential viewpoints for ground terrain material recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1205–1218, Mar. 2022.

[21] N. Salamati, C. Fredembach, and S. Süsstrunk, "Material classification using color and NIR images," in *Proc. Color Imag. Conf.*, 2009, no. 1, pp. 216–222.

[22] V. Saragadam and A. C. Sankaranarayanan, "Programmable spectrometry: Per-pixel material classification using learned spectral filters," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2020, pp. 1–10.

[23] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2407–2416, Dec. 2016.

[24] J. DeGol, M. Golparvar-Fard, and D. Hoiem, "Geometry-informed material recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1554–1562.

[25] Z. Erickson, E. Xing, B. Srirangam, S. Chernova, and C. C. Kemp, "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10 452–10 459.

[26] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in *Proc. 28th Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 511–518.

[27] J. R. Frisvad, T. Hachisuka, and T. K. Kjeldsen, "Directional dipole model for subsurface scattering," *ACM Trans. Graph.*, vol. 34, no. 1, 2014, Art. no. 5.

[28] C. Schmitt, S. Donné, 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.*, 2020, pp. 3493–3503.

[29] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, *Time-of-Flight and Structured Light Depth Cameras: Technology and Applications*. New York, NY, USA: Springer, 2016.

[30] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

Shi Mao received the B.E. degree from the South China University of Technology, Guangzhou, China, in 2019. He is currently working toward the M.S. degree with Shenzhen International Graduate School, Tsinghua University, Beijing, China. His research interests includes 3D vision and visual intelligence.

Mengqi Ji received the B.E. degree from the University of Science and Technology Beijing, Beijing, China, in 2012, and the M.Sc. and Ph.D. degrees from the Hong Kong University of Science and Technology, Hong Kong, in 2013 and 2019, respectively. He is currently an Associate Professor with the Institute of Artificial Intelligence, Beihang University, Beijing, China. Before that he was a Postdoc with Tsinghua University, Beijing, China. His research interests include 3D vision and computational photography.

Bin Wang received the master's degree and the Ph.D. degree from the China National Digital Switching System Engineering and Technological Research and Development Center. He is currently a Professor with the school of Electrical Engineering, Zhejiang University, Hangzhou, China. His research interests mainly include new computer networks, artificial intelligence security, and Internet of Things security.

Qionghai Dai (Senior Member, IEEE) is currently a Professor with the Department of Automation, and an Adjunct Professor with the School of Life Science, Tsinghua University, Beijing, China. He is the Academician of the Chinese Academy of Engineering. His research interests include computational photography, brain science, and artificial intelligence.

Lu Fang (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2007 and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2011. She is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include computational imaging and visual intelligence. She is currently an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON MULTIMEDIA.