Research Article

ADVANCED
PHOTONICS
NEXUS

Object pose and surface material recognition using a single-time-of-flight camera

Dongzhao Yang,[a] Dong An,[b] Tianxu Xu,[c] Yiwen Zhang,[b] Qiang Wang,[d] Zhongqi Pan,[e] and Yang Yue[a,*]

[a]Xi'an Jiaotong University, School of Information and Communications Engineering, Xi'an, China
[b]Nankai University, Institute of Modern Optics, Tianjin, China
[c]Zhengzhou University, School of Electrical and Information Engineering, National Center for International Joint Research of Electronic Materials and Systems, Zhengzhou, China
[d]Angle AI (Tianjin) Technology Co. Ltd., Tianjin, China
[e]University of Louisiana at Lafayette, Department of Electrical and Computer Engineering, Lafayette, Louisiana, United States

Abstract. We propose an approach for recognizing the pose and surface material of diverse objects, leveraging diffuse reflection principles and data fusion. Through theoretical analysis and the derivation of factors influencing diffuse reflection on objects, the method concentrates on and exploits surface information. To validate the feasibility of our theoretical research, the depth and active infrared intensity data obtained from a single time-of-flight camera are initially combined. Subsequently, these data undergo processing using feature extraction and lightweight machine-learning techniques. In addition, an optimization method is introduced to enhance the fitting of intensity. The experimental results not only visually showcase the effectiveness of our proposed method in accurately detecting the positions and surface materials of targets with varying sizes and spatial locations but also reveal that the vast majority of the sample data can achieve a recognition accuracy of 94.8% or higher.

Keywords: data fusion; time-of-flight; object detection; diffuse reflection principles; machine learning.

Received Jan. 3, 2024; revised manuscript received Apr. 7, 2024; accepted for publication May 15, 2024; published online Jun. 4, 2024.

[DOI: 10.1117/1.APN.3.5.056001]

1 Introduction

Object detection and recognition constitute a pivotal aspect of 3D reconstruction techniques.[1] The primary objective is to ascertain the identity, position, and orientation of objects within a scene or image. Current methods are typically classified as either model- or context-based approaches.[2] Objects in the real world are composed of diverse materials with distinct physical, chemical, and optical properties. The discrimination of object materials in real-world scenarios holds significant importance in computer vision research.[3] Relying solely on the shape and orientation data may result in the omission of crucial surface material information. Moreover, objects in real-world scenarios often exhibit intricate surface geometrical postures. Existing methods face limitations in concurrently recognizing materials and surface poses. Many approaches tend to overlook surface information while concentrating on holistic object recognition. In our approach, to achieve improved recognition outcomes, the material, position, and orientation information of surface elements are incorporated. It does not solely rely on size and shape for classification, making it adaptable to various target sizes, shapes, and poses. By extracting and detecting the surface information, our method accurately determines the object's pose and material, thereby enhancing its applications in computer vision tasks, such as 3D reconstruction technology.

Image data play a pivotal role in object detection, yet their effectiveness is constrained in complex scenes when sourced from a single origin. The advent of image data fusion has addressed this limitation by integrating multisource information, resulting in enhanced detection capabilities. Common fusion methods encompass thermal infrared, visible, RGB, depth, and active infrared.[4] However, aligning fused data from diverse sensors proves intricate due to differences in resolution, focal length, and viewing angles. In the data acquisition setup, a single time-of-flight (ToF) camera concurrently captures the depth and active infrared data, presenting a solution that offers

*Address all correspondence to Yang Yue, yueyang@xjtu.edu.cn.

accurate depth information while mitigating the impact of ambient lighting. Notably, ToF sensor data remain independent of natural lighting, endowing the proposed method with robustness across various lighting conditions.[5]

Existing target detection methods often neglect the complex surface information on objects, including their material, optical properties, and geometry, thereby limiting the precision of object recognition. This research holds significant value in advancing high-precision object recognition. Moreover, there is a need for further investigation into fusion methods for target detection using image data from a single sensor. To address these challenges, we propose a recognition method that combines depth and active infrared intensity (D-AI) data from a single ToF camera. By analyzing the diffuse reflections of object materials in the near-infrared band, a framework for determining the pose and surface material of a detected object is established. Figure 1 provides an overview of the framework, involving the simultaneous acquisition of D-AI images, utilizing point clouds to ascertain the object's pose, and identifying the surface material by considering the reflective intensity and its influencing factors. This method finds diverse applications in industrial production, unmanned exploration, criminal investigation, monitoring, autonomous driving, etc., where object surface information is crucial, particularly in challenging lighting conditions.

The major contributions of this paper can be summarized as follows:

1. An innovative framework that integrates multiple image data to accurately process and recognize the objects' pose and surface materials focuses on acquiring and analyzing spatial positions and surface orientations, allowing recognition of objects of several materials with various shapes if the resolution conditions are met.
2. An approximate model for analyzing infrared intensity from conventional materials with diffuse reflection in the near-infrared wavelength band also identifies the factors affecting the infrared reflection intensity received by the ToF sensor from the object's surface.
3. The utilization of a single-ToF depth camera as the core acquisition device eliminates the need for complex alignment processes with multiple sensors. The fusion approach, combining D-AI information, demonstrates improved robustness in the face of ambient light and temperature changes.
4. The proposed method achieves high accuracy in pose calculation and recognition, with recognition accuracy surpassing 94.8% in 90.0% of the samples.

The subsequent sections of this paper are organized as follows: Sec. 2 reviews related works on the recognition of geometric pose and surface materials of objects, as well as image data fusion methods in object detection. Section 3 carries on the theoretical derivation of the diffuse reflection law of the object's surface and an analysis of the influencing factors. Section 4 provides a detailed description of the developed approach. Section 5 outlines the experimental setup and analyzes the experimental results. Finally, conclusions and avenues for future work are discussed in Sec. 6.

2 Related Works

2.1 Recognition of Object's Pose and Surface Materials

Diverse methods and applications have been developed for the recognition of object poses, utilizing 2D images or depth data. Pose estimation algorithms have reached maturity, finding applications in areas such as motion capture,[6] autonomous driving,[7] and robot grasping.[8] Deep learning, particularly in conjunction with convolutional neural networks (CNNs), has yielded impressive results in 2D image-based pose estimation. Pose convolutional neural network, proposed by Xiang et al.,[9] predicts the object's center coordinates and infers rotational information through regression. For 3D objects with depth information, traditional methods involve aligning the point clouds with known pose templates in a library and analyzing the pose relative to the template. Improved iterative closest point (ICP) algorithms, such as globally optimal-iterative closest point by Yang et al.[10] and GICP by Wang et al.,[11] address challenges such as variations in initial positions, missing points, noise, and different scale factors. In addition, attitude estimation can be achieved through deep networks, compensating for limitations in traditional ICP algorithms. Robust point matching-net by Yew and Lee[12] is a deep-learning-based method that is insensitive to initialization, while recurrent closest point by Gu et al.[13] overcomes the irregularity and inhomogeneity issues of point cloud data structures. Point cloud object retrieval and pose estimation method by Kadam et al.[14] is an unsupervised method

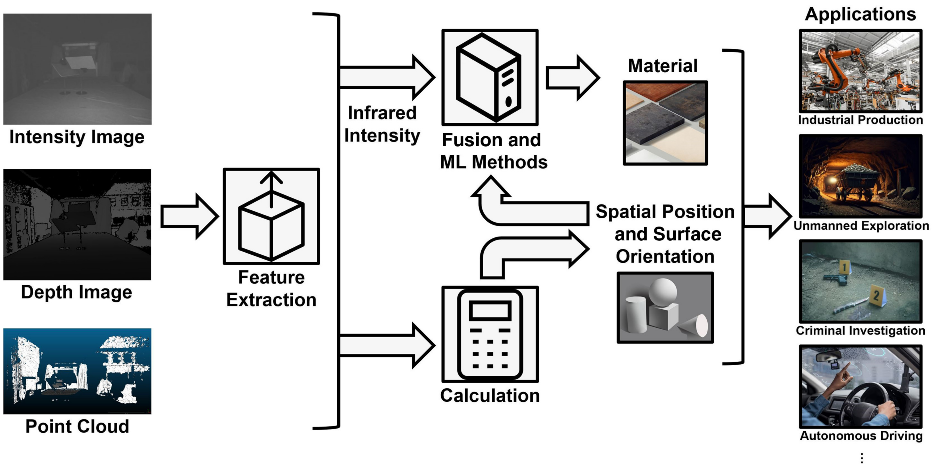

Fig. 1 Framework overview and applications.

that retrieves targets from point clouds and estimates their poses, surpassing traditional and learning-based approaches. Researchers have also proposed fusion methods that combine 2D data and depth information. Dense fusion (DF) by Wang et al.[15] separately processes the data sources and employs a DF network to extract pixel-level dense feature embeddings for pose estimation. Sparse DF (SDF) by Gao et al.[16] integrates sparse and DF modules using the Transformer architecture, enhancing semantic texture, and leveraging spatial structural information to improve pose estimation. Recent advancements took more account of the evaluation quality, indeterminacy, and target privacy for pose detection. Yang and Marco[17] proposed the first pose estimator that empowers the estimation with provable and computable worst-case error bounds. Gong et al.[18] explored a pose estimation framework, DiffPose, which treated 3D pose estimation as an inverse diffusion process, in response to the inherent ambiguity and occlusion. Aleksandr et al.[19] proposed an image-free human keypoint detection technique using a small number of coded illuminations and a single-pixel detector, addressing the data security and privacy issues faced by human pose estimation.

Research in material recognition encompasses various approaches aimed at perceiving and identifying materials across different application domains. Commonly framed as a texture classification problem in computer vision and image analysis, two primary categories of methods prevail: manual feature extraction coupled with classification techniques and deep-learning-based methods. One traditional method frequently employed is the bag-of-visual words approach,[20] which has demonstrated success in recognizing image materials. Feature-based frameworks for material recognition have also been proposed, utilizing techniques such as composite edge and color descriptors or the accelerated robust feature descriptor[21] to characterize localized image regions. Hierarchical material property representation mechanisms, exemplified by the GS-XGBoost model with a newly designed feature fusion algorithm and relative attribute model,[22] have been developed to enhance the material property classification. On the other hand, deep-learning-based approaches achieve material recognition through the construction of CNNs. For instance, the Materials in Context Database[23] has been leveraged to convert a pretrained CNN into an efficient fully convolutional framework and, when combined with a fully connected conditional random field, predicts materials within an image. Other deep-learning networks, such as the Material Attribute-Category CNN[24] and Deep CNNs,[25] along with feature encoding methods, have been employed to address the material recognition and visual attributes.

Besides focusing on texture, material perception can also rely on the fusion analysis of multimodal properties such as the optical properties of the target. Serrano et al.[26] proposed that the appearance of materials depends on the reflectivity, the surface geometry, and the illumination. In the acquisition of such optical properties, optical sensors with special capabilities, represented by depth cameras, have a certain degree of advantage. Mao et al.[27] proposed a method for surface material sensing based on structured laser points captured by a structured light camera and demonstrated the relationship between the active infrared image and the optical properties of different materials. There are also more related studies that selected ToF for this purpose. For example, on a pixel-by-pixel basis, Conde[28] utilized the band-limited character of material impulse response function for the material classification. Lee et al.[29,30] proposed a method for material type identification in terms of color and surface infrared reflectance features as well as the use of a single depth image for the estimation of reflectance and a segmentation method based on the similarity of reflectance. Mannan et al.,[31] on the other hand, determined the material of an object based on the smoothness of its surface by analyzing the reflectance pattern of infrared light.

The fields of object pose assessment and material recognition have historically been researched independently, resulting in disparate approaches. However, the integration of specific surface information effectively combines both aspects, enabling the simultaneous analysis of object pose and material judgment. This information fusion approach is not only efficient but also broadly applicable across various scenarios. In addition, relying solely on a single source of image data for material recognition often lacks comprehensive information. To overcome this limitation and enhance accuracy, data fusion techniques are employed.

2.2 Image Data Fusion Schemes in Object Detection

Image data fusion involves combining images from different sensors to generate robust and information-rich images. The key to successful fusion lies in effective information extraction and appropriate integration.[32] Thermal infrared and visible fusion schemes are well established.[4] Visible images offer high spatial resolution, detail, and contrast, while infrared images resist environmental interferences, providing more information, accuracy, and robustness than unimodal signals. However, combining images from different sensors can introduce resolution issues, and both types of images may be limited in extracting spatial information, impacting target detection and recognition in 3D space. RGB and depth image fusion is another widely adopted technique, proven successful in applications such as moving object recognition,[33] semantic segmentation,[34] scene reconstruction,[35] and medical image segmentation.[36] However, it necessitates better lighting conditions and tends to perform less effectively and accurately in dim or dark environments.

Depth and infrared fusion is particularly advantageous in acquiring data in dark environments and capturing the 3D stereo information of the target. This fusion category includes depth and thermal infrared (D-TI) as well as D-AI methods, with current research primarily concentrating on depth and thermal infrared fusion. Infrared or thermal imaging cameras gather emitted infrared information and integrate it with the depth data. Previous studies involved various sensors, necessitating specific algorithms to align focal lengths, resolutions, viewpoints, and other observation information, resulting in a complex process. Successful approaches have included the use of structure from motion[37] and multiview stereo techniques,[38] or combining lidar scanners with infrared cameras,[39] to reconstruct 3D models and temperature fields. However, these methods encounter challenges in detecting targets with poor infrared emission performance or similar thermal infrared (IR) intensities. By contrast, depth and active IR fusion achieves high-accuracy target detection using a single sensor, making it the chosen approach for our work.

3 Theoretical Model of the Diffuse-Reflection Law and Influencing Factors

In target detection and 3D reconstruction applications, we can approximate and analyze the optical properties and diffuse reflection laws of materials on the target object's surface using theoretical principles. This information is instrumental in understanding and leveraging factors that affect the active infrared intensity of camera acquisition in the subsequent fusion methods. Common materials such as paper, fabrics, leather, paint layers, and plastics exhibit similar diffuse reflection laws in the near-infrared band emitted by lidars (e.g., ToF cameras). Their behavior can be roughly approximated using Lambertian scattering[40] and Kubelka–Munk (K-M) theoretical models[41] for computational analysis, without necessitating precise reflection values. When considering diffuse reflection effects, the detected surface is treated as an infinitely wide sheet of material, neglecting edge effects and assuming that the object thickness is much larger than the size of particles interacting with absorbed and diffusely reflected light. In addition, when near-infrared light illuminates the detected surface, the model assumes isotropic energy reflection in all directions within the hemispherical space.

Since the size of the irradiation unit of the ToF camera used in this method is negligible with respect to the scale of the captured scene, the emitted infrared light is approximated as a point light source. According to the law of illumination, when illuminating with a point light source, the illuminance on the surface of an object perpendicular to the light is proportional to the luminous intensity of the light source and inversely proportional to the square of the distance from the illuminated surface to the light source. When the surface element d$\boldsymbol{A}$ is at a distance r from a point source with luminous intensity I, and the angle between it and the plane where the camera lens is located is θ, the illuminance E in the surface element d$\boldsymbol{A}$ can be expressed as

$$E|_{\mathrm{d}\boldsymbol{A}} = \frac{I}{r^2} \cos \theta \bigg|_{\mathrm{d}\boldsymbol{A}}. \tag{1}$$

According to Lambert's cosine law,[42] which is applicable to Lambertian scatterers, the reflected intensity Ir at the surface element d$\boldsymbol{A}$ of the object received by the ToF sensor is proportional not only to its own reflectivity R but also to the cosine of the angle γ between the direction of observation and the normal of its surface element d$\boldsymbol{A}$. The reflected intensity Ir can be expressed as

$$Ir|_{\mathrm{d}\boldsymbol{A}} = R \times \frac{I}{r^2} \cos \theta \times \cos \gamma \bigg|_{\mathrm{d}\boldsymbol{A}}. \tag{2}$$

The relative spatial positions and surface geometry of the surface element and the camera in Eq. (2) are illustrated in Fig. 2.

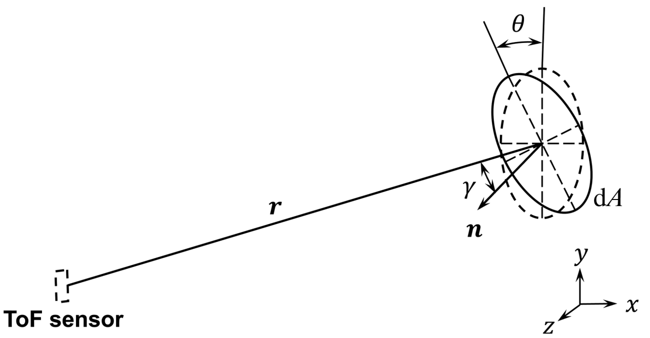

Fig. 2 Relative spatial position and geometric relationship between the ToF sensor and the surface element d***A***.

θ is related to the inclination of the surface element, while γ is related not only to the inclination of the object surface but also to the spatial position where the surface element is located. The reflectivity R of the surface element is connected to the specific optical properties of the object surface. As the object material is often irregular and inhomogeneous in the interior, the optical action within the scope of this paper is considered a semi-infinite sample optical model described by the K-M theory. In this context, the K-M diffuse modification equation can be expressed as the diffuse reflectivity of the surface of the object for a particular material.[43] The K-M function F follows

$$F(R) = \frac{K}{S} = \frac{(1-R_\infty)^2}{2R_\infty} \approx \frac{(1-R)^2}{2R}, \tag{3}$$

where K is the absorption coefficient, S is the scattering coefficient, and R_∞ is the reflectivity of the sample of the infinite thickness. Since the surface's optical impact within the range of scales compared with the macroscopic size of the object can be disregarded, R_∞ is approximately equal to the object surface reflectivity R.

After transforming and substituting Eq. (3) into Eq. (2), we obtain

$$Ir|_{\mathrm{d}\boldsymbol{A}} = \left[1 + \frac{K}{S} - \left(\frac{K^2}{S^2} + \frac{2K}{S}\right)^{\frac{1}{2}}\right] \times \frac{I}{r^2} \cos \theta \times \cos \gamma \bigg|_{\mathrm{d}\boldsymbol{A}}. \tag{4}$$

In the K-M theory, the absorption coefficient K and the scattering coefficient S are coefficients related to the particle size, refractive index,[44] and the composition of the surface material of the object and its microscopic properties. By analyzing Eq. (4), if we apply the above-mentioned model assumptions and approximations condition, the infrared intensity of a particular surface element d$\boldsymbol{A}$ in the acquisition scene received using the ToF sensor is mainly related to the material, the incident infrared intensity I, the distance r between d$\boldsymbol{A}$ and the camera lens, the angle θ between d$\boldsymbol{A}$ and the plane of the camera lens, and the angle γ between the line of the surface element and the camera lens.

The camera lens plane is denoted as the $z = 0$ plane, the equivalent luminous geometric point is labeled as (0, 0, 0), and the surface element d$\boldsymbol{A}$ is labeled as (x, y, z), then z = dep is the depth value of d$\boldsymbol{A}$ captured in the depth image. The horizontal inclination α and the vertical inclination β are introduced to characterize the inclination of the object's surface jointly, as shown in Fig. 3. Horizontal inclination α is defined as the angle between the projection of the surface normal vector in the horizontal plane and the line from the camera lens to the center of the field of view. When the projection of the surface normal vector of the object is located on the right side of the line from the camera lens to the center of the field of view, the horizontal tilt angle takes a positive value; conversely, the horizontal inclination takes a negative value. The vertical plane, defined as the plane that crosses the line from the camera lens to the center of the field of view and is perpendicular to the horizontal plane, is introduced. The vertical inclination β is defined as the angle between the projection of the object surface normal vector in the vertical plane and the line from the camera lens to the center of the field of view. It is situated on the surface of the object surface normal vector projection of the camera lens to the center of the field of view of the line of the upper. The vertical inclination is

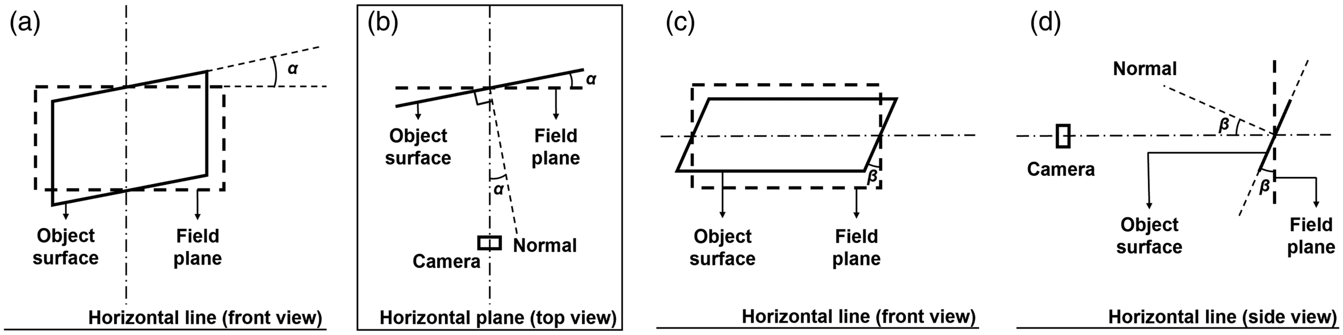

Fig. 3 Schematic diagram of horizontal inclination angle α and vertical inclination angle β. (a) is the front view of α, (b) is the top view of α, (c) is the front view of β, and (d) is the side view of β.

positive if it is located on the upper side of this line; otherwise, it is negative.

Based on the aforementioned provisions and assuming a resolution of the collected scene as 640 × 480, the expressions for r, θ, and γ in terms of x, y, dep, α, and β can be formulated as follows:

$$r = \sqrt{(x-320)^2 + (y-240)^2 + \mathrm{dep}^2}, \tag{5}$$

$$\theta = \arcsin\left[\frac{\sqrt{\sin^2\alpha + \sin^2\beta}}{\sqrt{\sin^2\alpha + \sin^2\beta + (\cos\,\alpha + \cos\,\beta)^2}}\right], \tag{6}$$

$$\begin{aligned}\gamma &= \left(\frac{\pi}{2} - \theta\right) - \arcsin\left(\frac{\mathrm{dep}}{r}\right)\\ &= \left\{\frac{\pi}{2} - \arcsin\left[\frac{\sqrt{\sin^2\alpha + \sin^2\beta}}{\sqrt{\sin^2\alpha + \sin^2\beta + (\cos\,\alpha + \cos\,\beta)^2}}\right]\right\}\\ &\quad - \arcsin\left[\frac{\mathrm{dep}}{\sqrt{(x-320)^2 + (y-240)^2 + \mathrm{dep}^2}}\right],\end{aligned} \tag{7}$$

$$\begin{aligned}Ir|_{\mathrm{d}A} = {}& \left[1 + \frac{K}{S} - \left(\frac{K^2}{S^2} + \frac{2K}{S}\right)^{\frac{1}{2}}\right]\\ &\times \frac{I}{(x-320)^2 + (y-240)^2 + \mathrm{dep}^2}\\ &\times \frac{\cos\alpha + \cos\beta}{\sqrt{\sin^2\alpha + \sin^2\beta + (\cos\alpha + \cos\beta)^2}}\\ &\times \cos\left\{\left\{\frac{\pi}{2} - \arcsin\left[\frac{\sqrt{\sin^2\alpha + \sin^2\beta}}{\sqrt{\sin^2\alpha + \sin^2\beta + (\cos\alpha + \cos\beta)^2}}\right]\right\}\right.\\ &\left.- \arcsin\left[\frac{\mathrm{dep}}{\sqrt{(x-320)^2 + (y-240)^2 + \mathrm{dep}^2}}\right]\right\}\Bigg|_{\mathrm{d}A}.\end{aligned} \tag{8}$$

This represents the theoretical approximation of the infrared reflection intensity for a specific surface element on a particular material. It is important to highlight that employing professional simulation software for calculations can yield a more precise relationship between infrared reflective intensity and its influencing factors. However, it is crucial to recognize that simulations may vary for different surface materials and in diverse complex environments, leading to a scarcity of universally applicable outcomes. The proposed approximate calculation method outlined above offers a more intuitive visualization of the relationship between factors influencing infrared reflection intensity. Its significance lies in its general applicability and universality, providing insights into the broad understanding of the relationship between factors and intensity in different scenarios.

According to the analysis above, the relationship between the corresponding infrared reflection intensity of a specific surface element and the horizontal and vertical position of the surface element, along with the horizontal and vertical inclination angles, can be calculated separately, as illustrated in Fig. 4. Figure 4(a) depicts the theoretical approximation of the relationship between the infrared intensity and the landscape/portrait position of the surface element when both α and β are 0 at dep = 1.5 m. Figure 4(b) showcases the theoretical approximate relationship between the infrared reflection intensity and the horizontal/vertical inclination of the surface element when dep = 1.5 m, $x = 320$, and $y = 240$. In the calculations, the values of K, S, and I are all adjusted to ensure that the maximum value of the intensity aligns with the maximum value of the cardboard intensity obtained using the Intel RealSense LiDAR L515 under the same acquisition conditions.

The intensity value Ir corresponding to each surface element on the target object is associated with the material category, the depth value dep, its position (x, y) in the depth plane, as well as the horizontal inclination α and vertical inclination β. These factors satisfy a certain functional relationship and can be expressed by the following equation:

$$Ir|_{\mathrm{d}A} = f(\mathrm{Material}, x, y, \mathrm{dep}, \alpha, \beta)|_{\mathrm{d}A}. \tag{9}$$

Considering the aforementioned five feature values, excluding the material, as the spatial positions and surface orientations recognition information of each surface element on the object, if the acquired image data are processed to obtain the infrared intensity information of each pixel value (considered a surface element as described above), along with the spatial positions and surface orientations information, a specific recognition method can be employed to determine the pose and surface material of the target object.

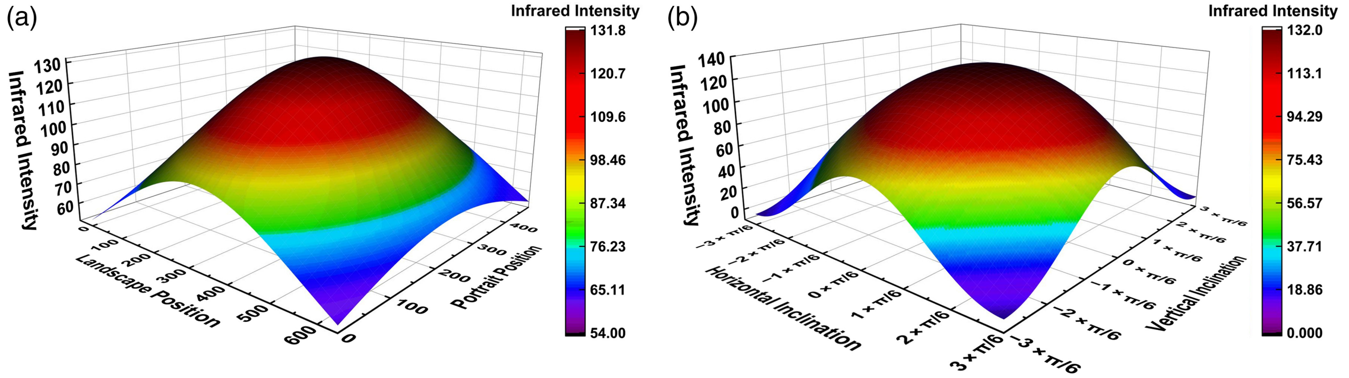

Fig. 4 Theoretical approximation of the infrared reflected intensity with respect to (a) the landscape and portrait positions and (b) horizontal and vertical inclinations of the surface element.

Moreover, another common material used in the target recognition tasks is highly reflective materials such as metals. The active reflective intensity of such materials is influenced by the energy distribution and polarization structure of the near-infrared light when it reflects on the metal surface. In this case, the diffuse reflection model assumed earlier is no longer applicable, and the reflected infrared intensity no longer follows a distribution law such as that shown in Fig. 4. However, despite the change in the reflective behavior of highly reflective materials, based on the derived geometric relationship and empirical qualitative judgment, this method still assumes that the intensity value Ir corresponding to each surface element on the target object is still related to the material, the depth value dep, its location (x, y) in the depth plane, as well as the horizontal inclination α and vertical inclination β, and the relationship described in the Eq. (9) is still adopted. It is essential to note that if a ToF camera is chosen to collect the data, and there is specular reflection in the shooting scene, it may lead to inaccuracies or deviations in the depth values according to its shooting principle. Therefore, this method is most suitable for shooting scenes that minimize the influence of specular reflection.

4 Detection Methods for Object Pose and Surface Material

Figure 5 illustrates the workflow of the proposed object recognition method based on the theoretical analysis above. The main sequence involves image preprocessing, contour extraction, and pixel value extraction on the depth-active infrared intensity image data set and the scene image to extract various influencing factors as analyzed in Sec. 3. Simultaneously, for the point cloud data, the normal vector acquisition is performed, and the surface orientations are calculated to comprehensively obtain the position of the object surface and pose information. Subsequently, an optimization neural network is employed for training and fitting, and the results obtained from the fitting are used for the material recognition of the object after optimizing the theoretical formulas.

4.1 Acquisition, Extraction, and Computation of Object's Spatial Positions and Surface Orientation Information and Infrared Intensity Information

As depicted in Fig. 6, the acquisition scene and equipment primarily consist of a computer, a ToF depth camera, objects (or scene) to be acquired, a camera tripod, and data cables, among other components. Throughout the acquisition process, it is crucial to ensure an unobstructed field of view in front of the target object. Simultaneously, the camera should be positioned stably and securely on a tripod, and data from the same scene should be acquired in multiple shots. The collected data encompass the infrared intensity image, the depth image, and the point cloud data that has been aligned with the 2D depth image, referred to as D-AI images.

In the image preprocessing stage, this method utilizes three noise reduction techniques: mean noise reduction, depth value filtering, and corrosion noise reduction. The objective is to enhance the quality of image data. Specific practices include eliminating irrelevant depth information and noise while minimizing errors such as multi-path interference (MPI), flying pixels, and intensity errors during image acquisition. MPI errors occur when the ToF sensor receives not only light reflected by the

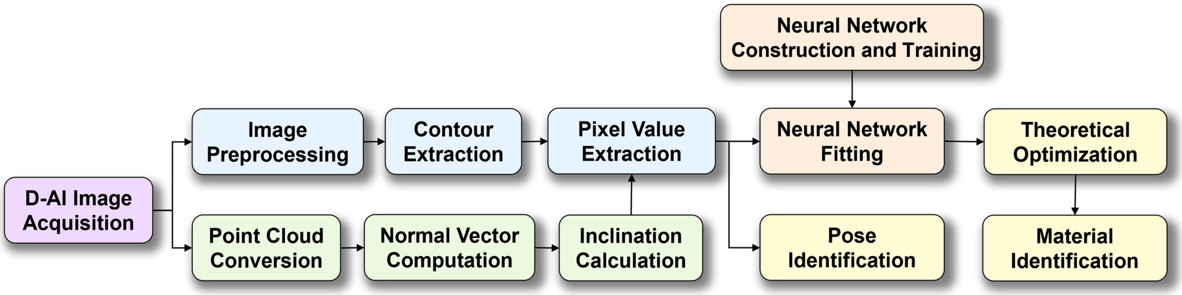

Fig. 5 Algorithm framework for recognizing object pose and surface material.

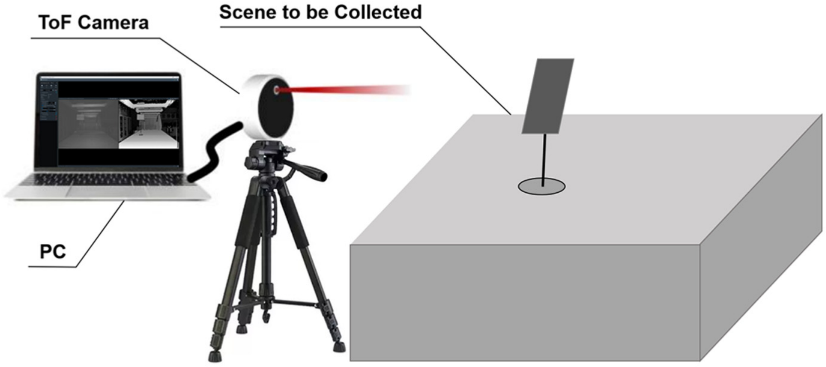

Fig. 6 Schematic diagram of acquisition scene and equipment arrangement.

target object but also complex diffuse or specular reflections, leading to measurement discrepancies. Flying pixel errors arise due to limited pixel values in the depth image, resulting in inaccurate depth values for pixels at object edges or regions with significant depth variations. Factors such as interpixel cross talk and lens scattering can contribute to flying pixel noise.[45,46] Intensity errors stem from non-flat object surfaces, distance, and integration time, causing fluctuations in values.

In mean noise reduction, multiple images of the same scene in the acquired D-AI images are averaged. Depth value filtering involves selecting a threshold based on the spatial location of the target object, which can be selected according to the recommended working depth range of the ToF camera. This threshold produces a preliminary binarized image, effectively filtering out irrelevant depth information. The irrelevant depth value is set to 0, retaining only the depth information within the range of the target object. Image erosion is performed by convolving the depth-filtered image with a suitable kernel to derive the minimum value within the kernel's coverage area. For images with a resolution of 640 × 480, a convolution kernel of a size such as (5,5) can work well. This minimum value replaces the pixel values at reference points. The erosion operation further filters out irrelevant information and noise in the highlighted area, specifically removing flying point noise errors.

The contour and pixel value extraction process involves applying a Gaussian filter to the noise-reduced depth image. The filtered image is then binarized and saved. Gaussian filtering[47] smoothens the image by averaging the neighboring pixel values, facilitating the extraction of the object's edge contour. Since weighted averaging affects some pixels, a suitable gray threshold is chosen for the filtered contour image, which then undergoes secondary binarization. Once the contour is obtained, the pixel values of the target object in the D-AI image are extracted and outputted. The extraction is limited to the range within the contour edges. An example of this process is illustrated in Fig. 7.

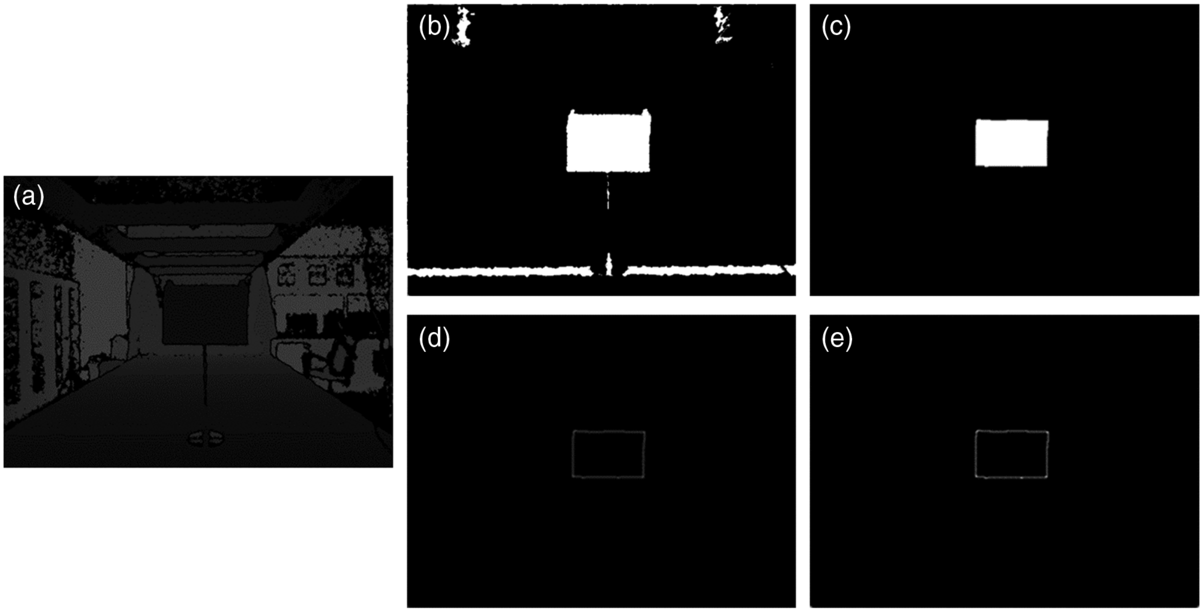

Fig. 7 Example of data preprocessing and contour extraction. (a) Depth map of the scene after mean noise reduction, (b) result of binarization after depth-valued filtering, (c) result after corrosion noise reduction, (d) result after contour extraction, and (e) result after the second binarization operation.

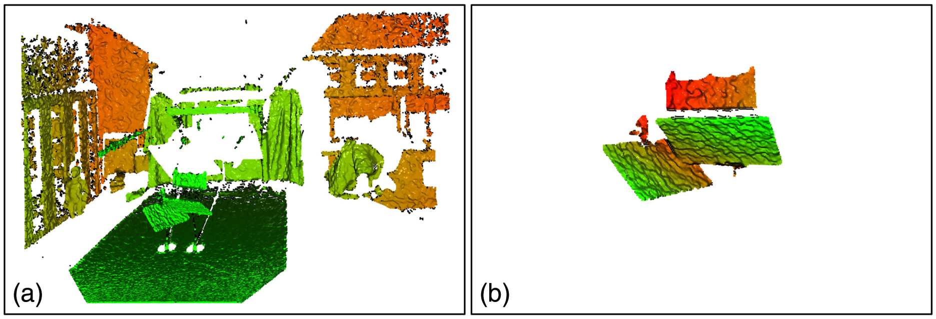

Fig. 8 Example of the point cloud information for (a) the scene and (b) the extracted result.

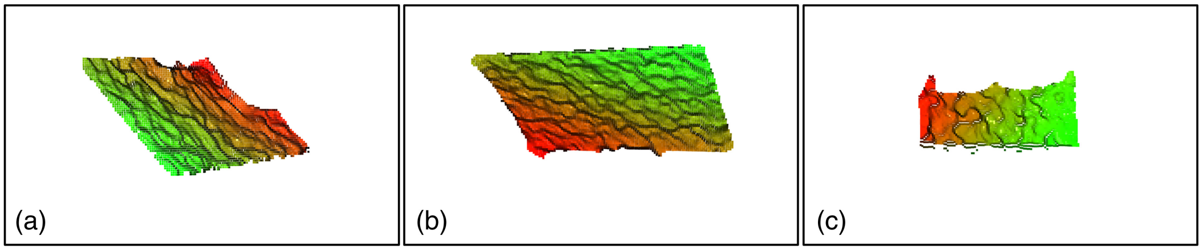

Fig. 9 Panels (a), (b), and (c) are the surfaces of different target objects.

The process encompasses parallel point cloud data processing, incorporating both preprocessing and inclination extraction. Background information is filtered out from the acquired point cloud, retaining only the foreground object data, as depicted in Fig. 8. Each point cloud is then individually extracted to obtain the normal vector information of each point, as demonstrated in Fig. 9. Parameters such as the local surface model and sphere neighborhood radius are utilized to adjust the normal direction. By employing a minimum spanning tree,[48] the normal vector value $\boldsymbol{N}$ of each point on the object's surface is calculated to obtain the normal components N_x, N_y, and N_z.

The direction of the line from the center of the field of view to the center of the camera lens is denoted as the z direction, and the direction parallel to the horizontal plane and perpendicular to the z axis is denoted as the y direction. Therefore, a 3D orthogonal coordinate system x-y-z can be established in the field of view. Within this coordinate system, the horizontal inclination α and vertical inclination β of each point in the point cloud can be expressed using the normal vectors of the three directions, N_x, N_y, and N_z

$$a = \arctan\left(\frac{N_x}{N_z}\right), \tag{10}$$

$$\beta = \arctan\left(\frac{N_y}{\sqrt{N_x^2 + N_z^2}}\right). \tag{11}$$

The inclination of each surface element is determined by calculating the horizontal inclination angle α and vertical inclination angle β from the normal vector of each point. By combining this information with the spatial position data from the point cloud, we obtain the surface orientations (α, β), depth value dep, and position (x, y) of each surface element. These factors with the actual diffuse reflection intensity Ir are analyzed to determine the material.

To achieve the target task, an artificial neural network is trained using a data set that includes the influencing factors and actual intensity values. The network learns to predict the fitted infrared intensity values for each surface element of the target object. The backpropagation (BP) neural network[49] is chosen due to its simple structure, low computational requirements, and strong nonlinear mapping capabilities. However, to address limitations such as slow convergence and the risk of local minima, the BP neural network is optimized using genetic algorithm (GA-BP)[50] and dung beetle optimizer (DBO-BP)[51] in subsequent calculations. Consideration can also be given to the other neural networks and advanced optimization algorithms for better fitting. Last, the fitted intensity of each surface element is combined with the actual diffuse reflection intensity to classify the materials of the object's surface elements in the next step.

4.2 Methods of Judging Surface Materials

For the application of material recognition on the object's surface in this study, the infrared reflection intensity, spatial positions, and surface orientations data of each surface element of the object's surface can be expressed in the following form in relation to the material, after the neural network is trained, fitted, and merged to obtain the results:

$$\text{Materials of } \mathrm{d}A_i \leftarrow g(x_i, y_i, \mathrm{dep}_i, \alpha_i, \beta_i, ir_{i,\text{actualvalue}}, ir_{i,\text{fittedvalue}}), \tag{12}$$

$$\text{Materials of the object} \leftarrow \sum_i g(x_i, y_i, \mathrm{dep}_i, \alpha_i, \beta_i, ir_{i,\text{actualvalue}}, ir_{i,\text{fittedvale}}). \tag{13}$$

The judgment method based on the above equation proposed in this study is shown in Fig. 10.

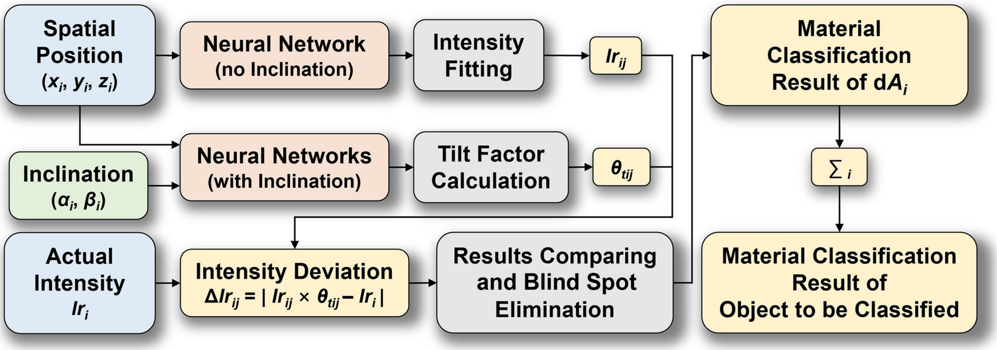
Spatial Position (x_i, y_i, z_i)
Neural Network (no Inclination)
Intensity Fitting
Ir_{ij}
Inclination (α_i, β_i)
Neural Networks (with Inclination)
Tilt Factor Calculation
θ_{tij}
Actual Intensity Ir_i
Intensity Deviation $\Delta Ir_{ij} = | Ir_{ij} \times \theta_{tij} - Ir_i |$
Results Comparing and Blind Spot Elimination
Material Classification Result of dA_i
$\sum_i$
Material Classification Result of Object to be Classified

Fig. 10 Flow chart of the judgment method based on the fitting results of the IR intensity values.

Two separate neural networks will be employed to simplify the training and improve the prediction: one network without considering inclination (marked as “no inclination” in Fig. 10) and another network that includes inclination, which are trained separately. These networks are designed for different material recognition scenarios. When the camera lens is parallel to a flat surface of the detected object or there is minimal surface curvature, the inclination has a smaller impact on material recognition compared with the spatial position. However, in cases with complex surface curvature or specific shooting viewpoints, surface element inclination becomes more important. Therefore, it is advantageous to employ two parallel neural networks that complement each other.

The first network solely considers the influence of spatial position and material on the reflected intensity when the normal vector of the surface element is perpendicular to the lens plane. By providing spatial position information to a neural network trained on various materials, the output represents the fitted intensity value corresponding to a specific material.

The second network takes both spatial position and inclination into account. A tilt factor θ_t is introduced to quantify the intensity change for each surface element caused by inclination, and θ_t is defined as the ratio of the fitted intensity value considering horizontal inclination α and vertical inclination β to the fitted intensity value without inclination,

$$\theta_t = \frac{ir(\alpha,\beta)}{ir(0,0)}. \tag{14}$$

The fitted intensity value ir_{ij} corresponding to a surface element when it is a certain material j according to the neural network in the case without inclination is obtained, and then, the tilt factor θ_{tij} corresponding to the surface element through the neural network in the case with inclination when it is a certain material is obtained. The result obtained by multiplying ir_{ij} and θ_{tij} is used as the result of the fitted intensity value of the surface element corresponding to each material. Afterward, the method differs the true intensity value from the results of the fitted intensity values corresponding to all the materials in the data set, respectively, to obtain the difference Δir_{ij}, and the relative deviation err_{ij} can be obtained as well,

$$\Delta ir_{ij} = |ir_{ij} \times \theta_{tij} - ir_i|, \tag{15}$$

$$\mathrm{err}_{ij} = \frac{|ir_{ij} \times \theta_{tij} - ir_i|}{ir_i} \times 100\%. \tag{16}$$

It is worth noting that for some pixels at certain spatial locations, there are two or more materials corresponding to essentially the same intensity values, for example, at the spatial location corresponding to a given surface element k, there are

$$ir_{ka} \approx ir_{kb}. \tag{17}$$

The pixel points at these locations are recorded as recognition blind spots. If the data corresponding to the blind spots are used to make judgments, the specific material cannot be accurately determined, which will affect the overall judgment of the object material and cause a decrease in recognition accuracy. Therefore, the data corresponding to the blind spots should be eliminated according to certain criteria. For the example shown above, the quantitative criteria for eliminating blind spots are

$$\frac{|ir_{ka} - ir_{kb}|}{ir_{ka}} \times 100\% \le 1\%, \tag{18}$$

and

$$\frac{|ir_{ka} - ir_{kb}|}{ir_{kb}} \times 100\% \le 1\%. \tag{19}$$

The pixel corresponding to k is considered the blind spot for judging material a and material b. It is necessary to remove k from the set of pixels contained in the object to be measured, and after all the blind spots have been removed, judgment can be made to get the result.

After that, the Δir_{ij} of different materials is compared to find the smallest value of Δir_{imin} and its corresponding err_{imin}, then the material corresponding to Δir_{imin} can be regarded as the material corresponding to the surface element. Then, the material corresponding to each face element is totalized, and the material corresponding to the most surface elements is the material to which the object to be measured belongs as determined by this method. When a large number of Δir_{imin} at different locations of the object are different, it means that the target object is composed of different materials and should be re-segmented before repeating the above process.

4.3 Analysis of the Fitting Results of the Factors Affecting Infrared Intensity

Figures 11 and 12 display the fitting results of GA-BP and DBO-BP neural networks using data sets for cardboard, leather, and metal. These results depict infrared intensity distribution at a depth of 1.25 m, no inclination, full field of view, and resolution of 640 × 480. Data set acquisition and training details are in Sec. 5. The figures illustrate the relationship between the infrared intensity and the planar spatial position (x, y) for different materials. Both optimized neural networks accurately represent the consistent infrared intensity distribution of each material. The overall distribution of leather and cardboard aligns with theoretical laws in Sec. 3. However, changing position introduces bias and intensity fluctuations due to overfitting or missing data, necessitating further optimization.

In Fig. 13, the fitted infrared intensities are shown at different depths for the three materials. The center position is (320, 240), while the upper left third of the field of view is represented by the position (212, 80). At this position, leather has a higher intensity than cardboard for depths below 13.6 but lower for depths above 13.6. This depth, (212, 80, 13.6), creates a blind spot where leather and cardboard discrimination is inaccurate, hampering object recognition. Depths larger than 32 result in indistinguishable intensities for metal and leather, forming additional blind spots. Increased depth reduces recognition accuracy, which is influenced by the performance of the ToF depth camera used.

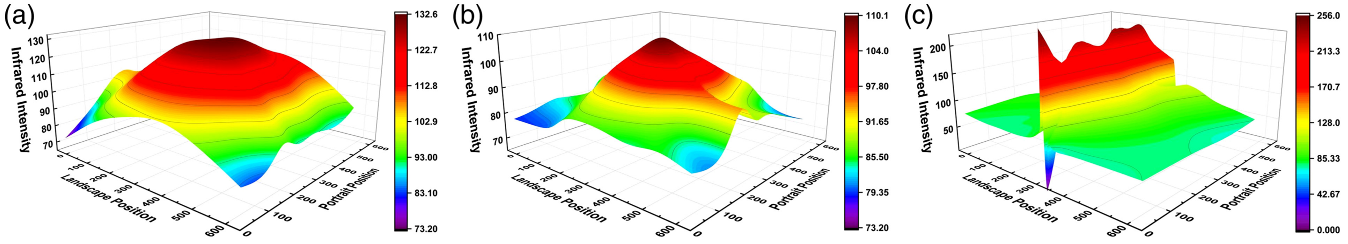

Fig. 11 Active infrared intensity distribution of different materials by GA-BP network fitting. (a) Cardboard, (b) leather, and (c) metal.

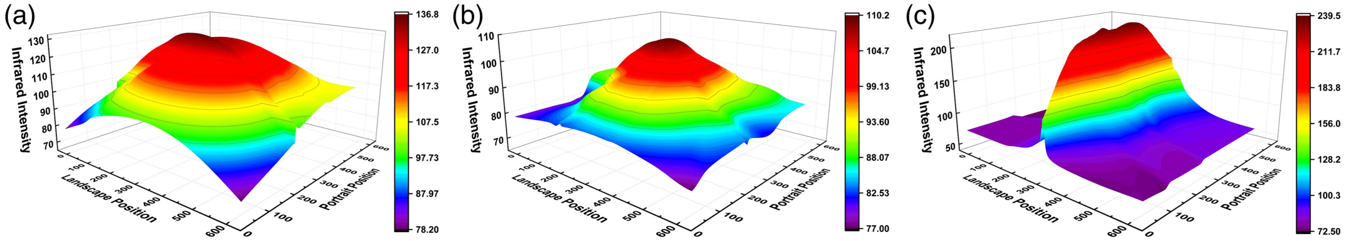

Fig. 12 Active infrared intensity distribution of different materials by DBO-BP network fitting. (a) Cardboard, (b) leather, and (c) metal.

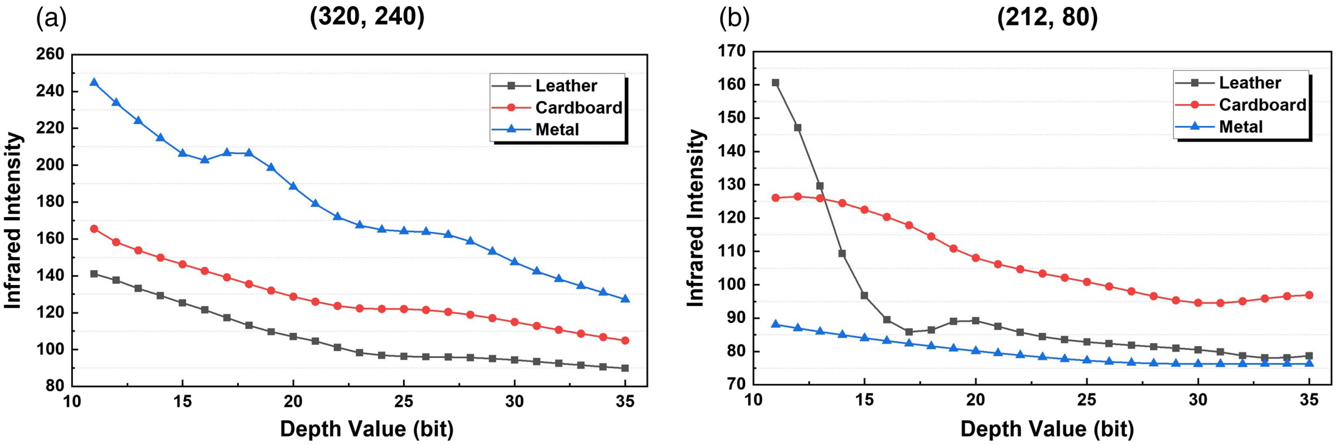

Fig. 13 Fits to the active IR intensity distribution of different materials as a function of the depth of the surface element (a) at position (320, 240) and (b) at position (212, 80).

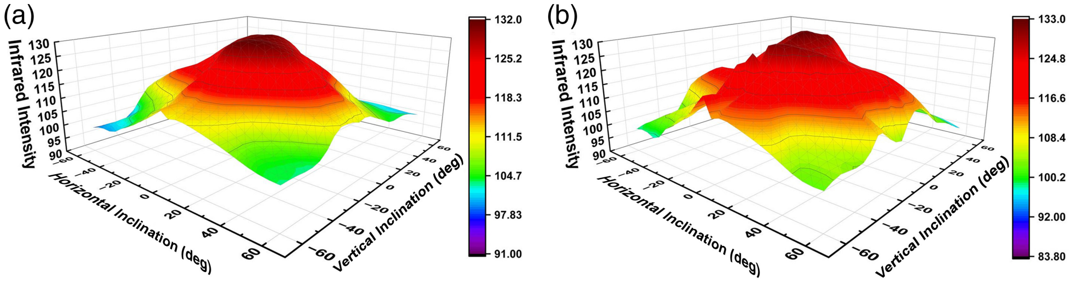

Fig. 14 Active infrared intensity fitting results at different inclination angles of (a) GA-BP network and (b) DBO-BP network.

Figure 14 exhibits the intensity fitting results for cardboard surface elements at the center point (320, 240) of the field of view. The depth is fixed at 1.25 m, and the inclination angles vary in increments of 5 deg. Both GA-BP and DBO-BP networks produce similar fitting results, which align with the theoretically derived laws in Sec. 3. However, there is a noticeable degree of deviation and fluctuation in intensity when the inclination angles change, regardless of the network used. This deviation could be due to factors such as overfitting or missing data, highlighting the need for further improvement.

4.4 Further Optimization of the Method According to the Diffuse Reflection Theory

The data set collected according to the above method is only a small number of samples in space, and the fitting ability of the neural network itself is always limited. It is important to note that while the overall distribution of the fitted results is generally consistent, there are still irregular fluctuations and deviations in intensity values when influencing factors change, as seen in the examples above. Although this overfitting can be made to be mitigated by increasing the amount of sample data and tuning the network hyperparameters, the inclusion of an optimization method for the neural outputs can be helpful in improving the accuracy due to the inherent intensity errors of the ToF sensors (attributed to the roughness of the object and the integration time). Thus, given the theoretical expectation of diffuse reflection, where intensity should be continuous and monotonous from the center to the surrounding areas, further optimization of the fitted intensity values is necessary before applying them to the judgment method. The optimization process employed in the method is illustrated in Fig. 15.

To enhance accuracy, we utilize the Savitzky–Golay filter[52] to smooth the intensity data for materials following the diffuse reflection law. This filter fits a low-order polynomial to neighboring data points, preserving the overall trend and width of the signal. The collected infrared intensity exhibits a centrosymmetric distribution at the center point of the image and adheres to Lambert's cosine law. To achieve centrosymmetrization, we process the intensity distribution in both the planar position and inclination domains. Equidistant points from the center have their intensities averaged and assigned to all the data points on them. The selection interval of data points depends on the scene resolution in the planar position domain and can be customized in the inclination domain (e.g., using a 5 deg interval in this case). For materials with high reflectivity, such as metals, microstructures cause variations in their infrared intensity at different inclinations, which are not currently optimized. To address the intensity distribution within the same depth, we employ a specific localization method based on the intensity distribution. The final optimized intensity is calculated by averaging values from multiple optimized neural networks. Figure 16 illustrates the optimization results for the example in Figs. 11 and 12, and Fig. 17 shows the results for the example in Fig. 14. The optimized data significantly reduce distribution bias and improve overall accuracy.

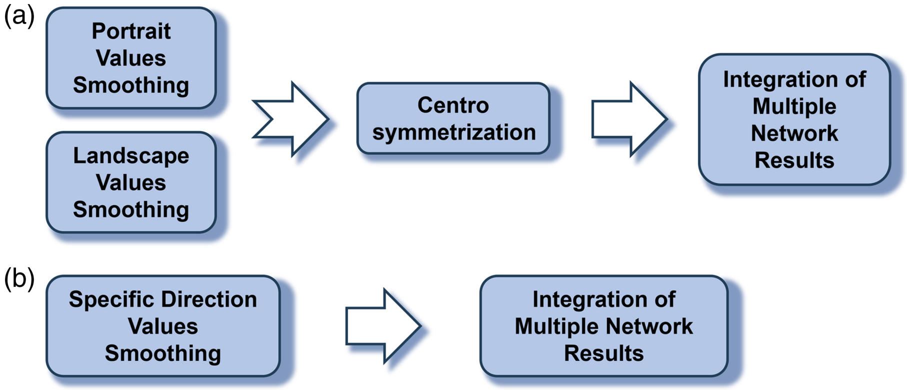
(a)
Portrait Values Smoothing
Landscape Values Smoothing
Centro symmetrization
Integration of Multiple Network Results
(b)
Specific Direction Values Smoothing
Integration of Multiple Network Results

Fig. 15 Optimization of the fitting intensity for (a) materials that follow the general diffuse reflectance law and (b) high reflectance materials.

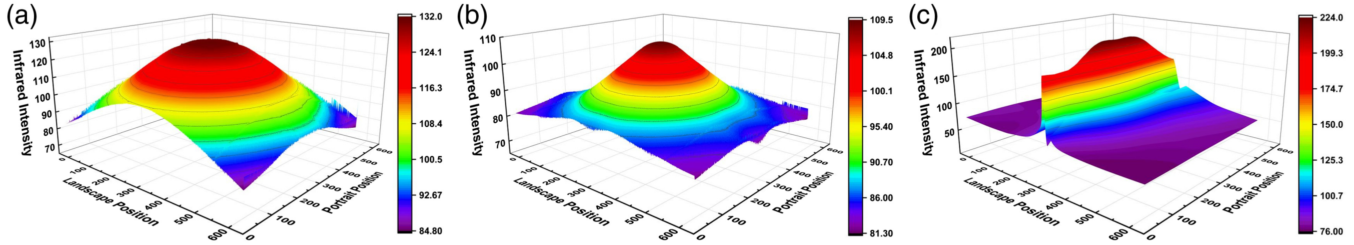

Fig. 16 Optimized fits of infrared intensity distributions for different materials based on the data in Figs. 11 and 12. (a) Cardboard, (b) leather, and (c) metal.

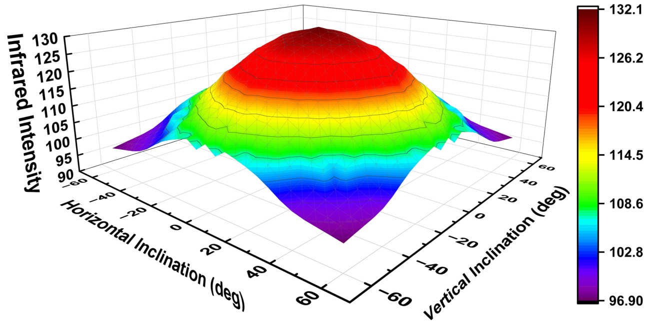

Fig. 17 Optimized fit of the infrared intensity of cardboard at different inclination angles based on the data in Fig. 14.

5 Experimental Setup and Results

5.1 System Building and Data Set Preparation

The data scene is arranged similarly to Fig. 6. Using the Intel RealSense Viewer software, a suitable shooting location is determined after conducting pre-experiments. Shots are captured with the Intel RealSense LiDAR L515 camera. D-AI data sets are collected from cardboard, leather, and metal materials at five depths (100, 125, 150, 175, and 200 cm) and nine spatial locations at each depth, as shown in Fig. 18. The surface orientation of each material is adjusted, and D-AI images are acquired with varying inclination angles (−60, −30, 0, 30, and 60 deg horizontally and vertically). Figure 19(a) illustrates an example set of intensity maps for cardboard material with different inclinations at the same location. Simultaneously, a point cloud map with matching resolution is obtained for each depth image. To reduce the flying point noise, three sets are collected for each position, material, and orientation. These sets form the basis of the adjustable D-AI data set, catering to specific application requirements. Our collected data set provides a viable, practical, and fundamental demonstration.

Following Fig. 5, the raw data undergo image preprocessing (wherein the depth threshold is set in accordance with a range of 100 to 200 cm), contour extraction [Fig. 19(b) shows examples after contour extraction], pixel value extraction, and inclination calculation. This provides the influencing factors for each extracted surface element, and these data can be used to train a practical neural network model. A total of 1245 images are collected according to the data set construction method described above. There are one to three objects of different materials distributed in each image and overlapped with each other. As the proposed method needs to ensure the accuracy and originality of the IR intensity and depth, neither of the two streams of data aligned need to undergo additional enhancement means. Our approach uses surface elements as the base magnitude for processing. A total of 107,492 leather data, 90,425 cardboard data, and 94,910 metal data are extracted for lightweight machine-learning training according to the methods provided in Sec. 4.1. Training and testing sets are divided according to the ratio of 7:3, and it is consistently ensured that the inputs in the testing set with and without tilting are in equal proportions. For the collected data set, the neural network without inclination has two hidden layers: the first hidden layer has eight neurons, and the second hidden layer has six neurons. Training parameters include a learning rate of 0.001, target error of 0.001, maximum

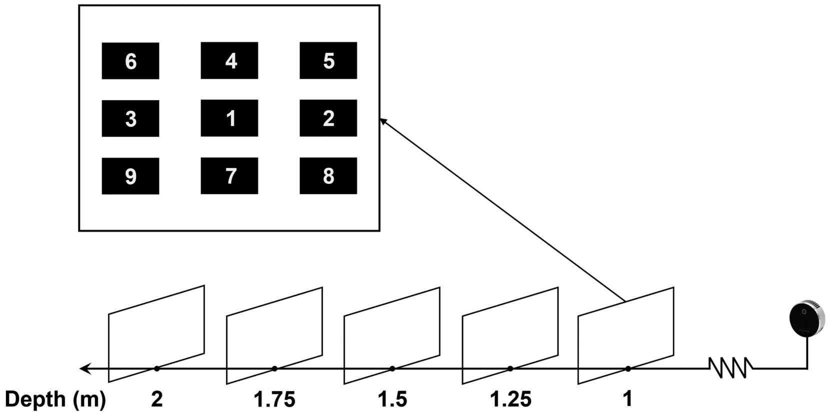

Fig. 18 Schematic illustration of D-AI image data acquisition locations for each material.

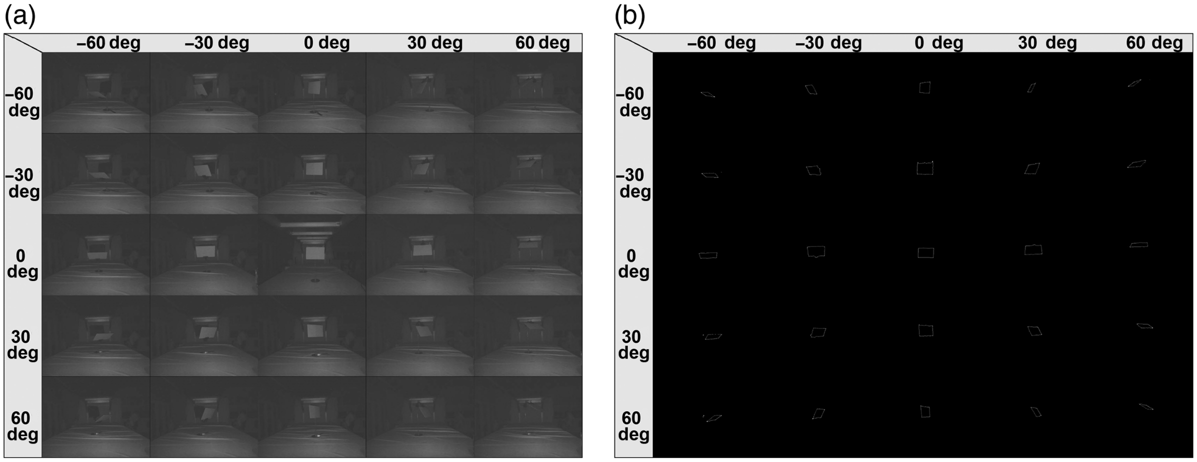

Fig. 19 Sets of (a) intensity images and (b) contour images obtained by depth image of the cardboard collected at the same location and different inclination angles.

of 4000 iterations, and 50 training sessions. In the GA-BP stage, there are 600 evolutionary iterations, a population size of 160, a crossover probability of 0.625, and a mutation probability of 0.05. In the optimization algorithm (DBO-BP) stage, there are 100 evolutionary iterations and a population size of 50. For the neural network with inclination, hidden layers are adjusted: the first hidden layer has 10 neurons and the second hidden layer has 8 neurons.

5.2 Test Results of Recognition

D-AI images of the objects without inclination are first randomly collected as a validation test of the recognition effect. For different materials, 10 areas of different sizes and locations are randomly selected as objects to be recognized in different sizes and locations for the recognition test. A total of 30 samples are obtained. Based on the recognition process in Sec. 4, the spatial positions and surface orientations of the samples are extracted, and the determination of the material to which each sample belonged is further obtained. The maximum deviation of each sample is defined as the maximum value in the err_{imin} corresponding to each surface element, in the determined material among all the surface elements of the sample, which can also be denoted as

$$\mathrm{dev}_{\max} = \frac{[\Delta ir_i]_{\max,\text{certain material}}}{ir_i} \times 100\%. \tag{20}$$

The average deviation is defined as the average value of err_{imin} corresponding to each surface element of the sample in the determined material, which can reflect the whole degree of proximity between the object and the predicted result, as

$$\mathrm{dev}_{\mathrm{avg}} = \frac{\sum_{i=1}^{n} \mathrm{err}_{imin}}{n}. \tag{21}$$

The total pixel accuracy is defined as the ratio of the number of surface elements contained in the material belonging to the one with the highest number of face elements, to the total number of surface elements (excluding blind spots) of the object to be measured, as

$$\mathrm{Acc} = \frac{[\text{surface elements of the same material}]_{\max}}{n} \times 100\%. \tag{22}$$

Then, the total pixel accuracy, maximum deviation, and average deviation of the test results for each set of samples are demonstrated in Fig. 20.

The method successfully identifies all the 30 sample groups, achieving a total pixel accuracy above 92.5%. The maximum deviation from the 30 samples does not exceed 8%, and the average deviation for each sample is below 4%.

D-AI images of objects with inclination are collected for an additional validation test of the recognition method. A total of 30 sets of test samples are obtained in a manner similar to the previous approach, and the material recognition results are determined using the proposed recognition judgment method. The total pixel accuracy, maximum deviation, and average deviation for each set of samples are illustrated in Fig. 21.

The method can accurately identify all the 30 sample groups, except for three groups with fewer surface elements that had recognition accuracy below 90.0%. The remaining 27 groups achieve a recognition accuracy of above 91.3%. Accuracy may slightly deviate when fewer surface elements are used. For cardboard material, maximum deviations are below 5.1%, and the average deviation is below 1.8%. For leather material, 9 out of 10 samples had recognition accuracy above 95.0%. Although seven samples had larger maximum deviations, the overall impact on accuracy is minimal due to their small proportion. By contrast, the recognition accuracy for metal material is lower compared with the other materials. Error values are larger than those for cardboard, indicating that the neural network is less effective for highly reflective materials such as metal. There is still room for improvement in accurately identifying materials with similar infrared reflection characteristics. In addition, 10 sets of test samples with inclination are randomly selected from the test set, a total of 18,692 surface element intensity values and their influencing factor arrays are computed using the method of Sec. 4.1, and then, the training set is utilized for explicit nonlinear regression (NLR) fitting and K-nearest neighbors (KNN) classification to calculate the average accuracy. The obtained results are compared with the accuracy obtained by the proposed

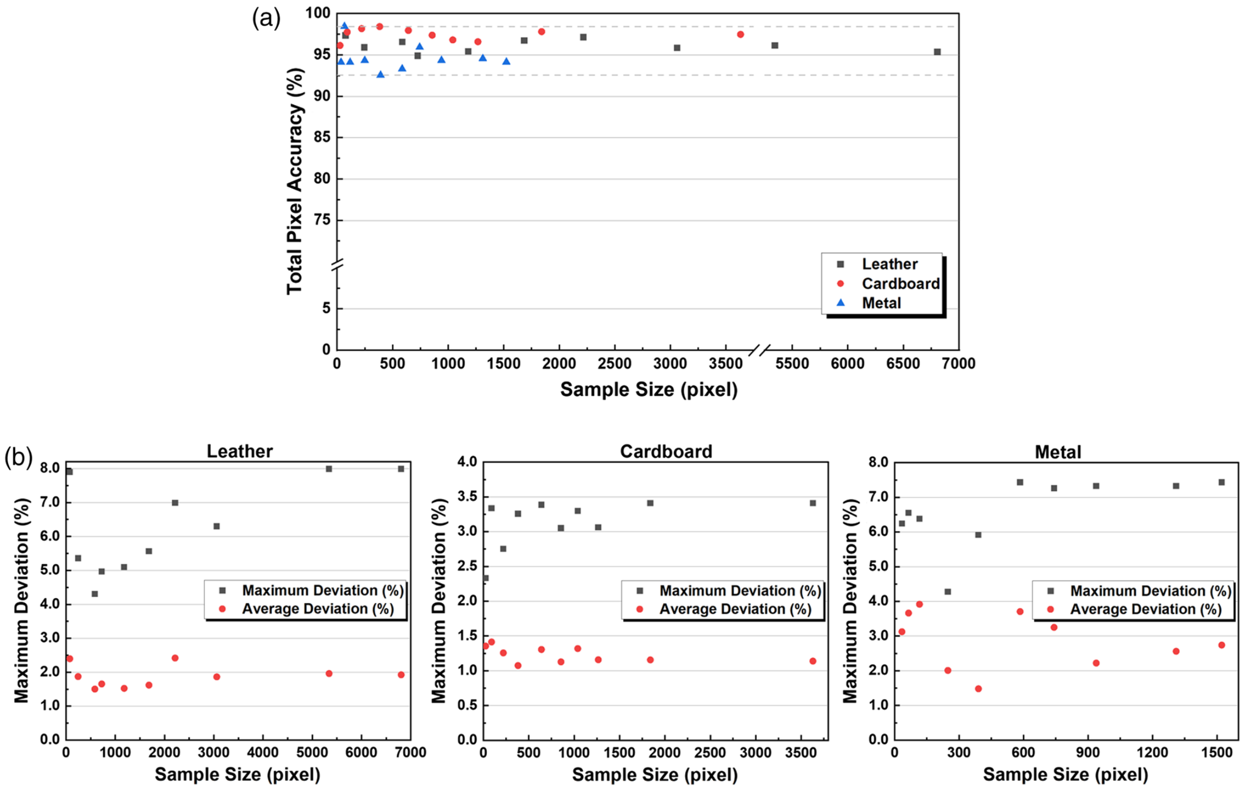

Fig. 20 (a) Total pixel accuracy and (b) maximum and average deviations of each set of test samples without inclination.

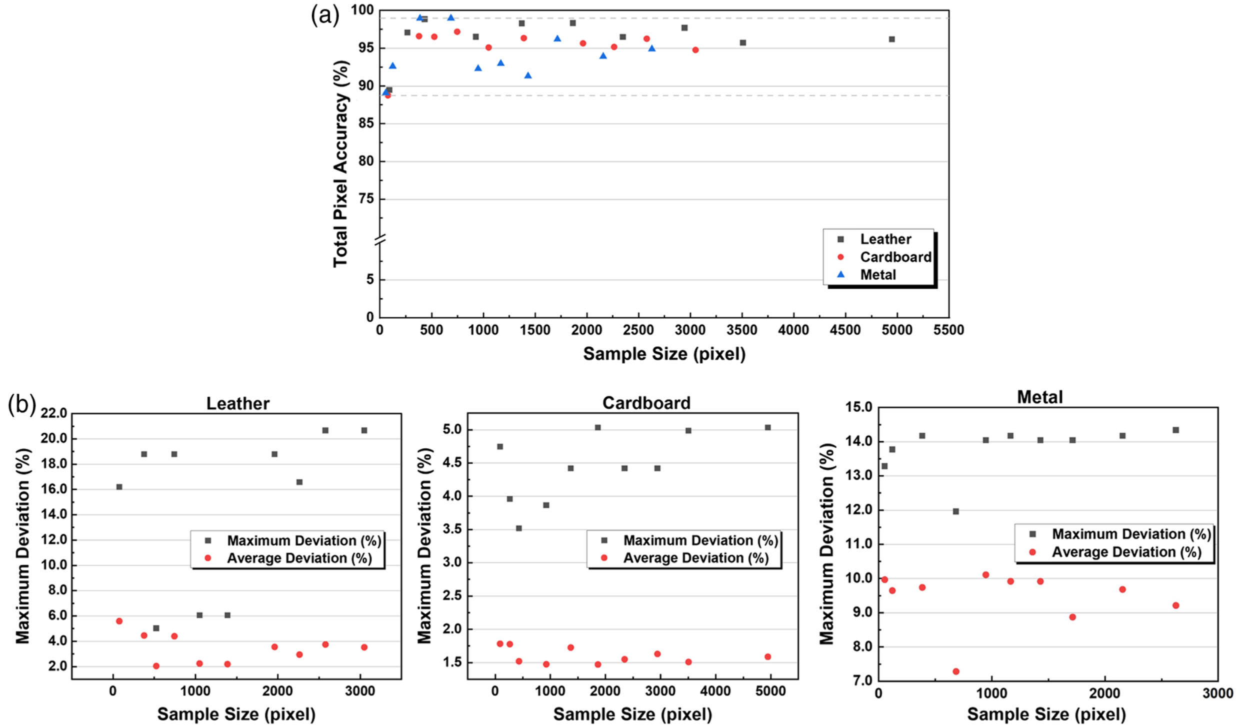

Fig. 21 (a) Total pixel accuracy and (b) maximum and average deviations for each set of test samples with inclination.

Table 1 Accuracies of material recognition by different methods.

Method	NLR	KNN	Our approach (unoptimized under Sec. 4.4 or by GA/DBO)	Our approach (optimized)
Acc. (%)	47.42	89.25	82.53	93.38

method and the results are displayed in Table 1. Upon comparison, the accuracy of our method has an advantage over the two conventional methods, and proper optimization steps are beneficial.

The proposed method utilizes two parallel independent neural networks for training, which can improve computational efficiency and save computational resources. When the inclination of the object surface is so slight that the effect on the reflection intensity is negligible, or when the surface of the target object is being captured directly, the efficiency can be greatly improved using only the first neural network. Splitting the neural network into two parallel ones based on the presence or absence of inclination can also reduce the training time and decrease the complexity of the network to improve accuracy. Conventional methods also include the use of an multilayer perceptron followed by softmax to classify a whole target object. However, the method of processing and classifying the whole target is not guaranteed for the classification results of small-sized object materials, whereas our method still has high accuracy when the size of the sample to be recognized is less than 500, as shown in Figs. 20 and 21. Moreover, this method cannot use the theoretical analysis in Sec. 3 to obtain an intuitive physical strength result for optimization in Sec. 4.4 to further improve the recognition accuracy. Furthermore, since the designed method effectively avoids the complex alignment process when fusing two different types of data streams, it reduces the complexity of the preprocessing algorithm and improves the efficiency to a large extent. After calculating or extracting the spatial position, pose, and infrared reflection intensity of the surface elements, the lightweight optimized neural network is used instead of conventional algorithms such as clustering or regression, which further saves computational time and resources, and improves the infrared intensity fitting at the resolution used for the experiment to the millisecond scale (even deployed on a personal laptop with a GPU of NVIDIA GTX 3060, 6 G of graphics memory, and 16 G of RAM).

6 Conclusion

A method that utilizes a single ToF camera and diffuse reflection principles is proposed to identify the pose and surface material of objects. Our approach involves a theoretical analysis and derivation of factors influencing the object's diffuse reflection. We have developed an image processing method to extract these factors, allowing for the calculation of pose positions and surface orientations. Data processing involves feature extraction and the application of lightweight machine-learning techniques. In addition, an optimization method is introduced for determining fitting values and surface materials based on intensity information and fitting results. To validate the feasibility of our method, we construct a data set containing diverse materials following theoretical laws and with varying inclined surfaces. Experimental results showcase the method's effectiveness in detecting the positions and surface materials of the targets with different sizes and spatial locations. The recognition accuracies for materials adhering to the theoretical laws without inclination are consistently above 94.9%. Even when there is an inclination, 90.0% of the samples can achieve a recognition accuracy of 94.8% or higher.

Our method presents a more efficient alternative to deep CNNs and point cloud neural networks. It leverages a single ToF sensor and integrates spatial information with active infrared intensity data. Through the analysis of diffuse reflection of near-infrared light from common material surfaces, we extract comprehensive influencing factors. This approach finds applications in various fields such as industrial production, unmanned exploration, criminal investigation, security surveillance, and unmanned vehicle driving. It successfully addresses the challenges posed by the complex ambient illumination and delivers valuable surface information through the use of active optical detection techniques.

However, certain limitations must be taken into account. The method's generalization ability requires improvement to accommodate a broader range of surface materials. The accuracy is constrained when dealing with highly reflective materials, primarily due to the operating principle of the ToF camera and potential specular reflections. Currently, the method does not encompass materials such as glass, water, and other liquids, as well as those with complex surface textures such as wood. As for conventional materials, there are limitations in the proposed method when the surface of the object consists of different materials and the proportion of one is too small. Future efforts will extend the scope to include these materials and develop targeted approaches. Our goal is to integrate more comprehensive optical analysis techniques and optimize feature utilization to effectively address these challenges.

Disclosures

The authors have no relevant financial interests in this paper and no other potential conflicts of interest to disclose.

Code and Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the Shaanxi Province Innovation Talent Promotion Program-Science and Technology Innovation Team (Grant No. 2023-CX-TD-03).

References

1. Z. Wang et al., "Object as query: equipping any 2D object detector with 3D detection ability," arXiv:2301.02364 (2023).
2. S. Rani, K. Lakhwani, and S. Kumar, "Three dimensional objects recognition & pattern recognition technique; related challenges: a review," *Multimedia Tools Appl.* **81**(12), 17303–17346 (2022).
3. D. Hu, L. Bo, and X. Ren, "Toward robust material recognition for everyday objects," in *Br. Mach. Vis. Conf.*, Vol. 2, p. 6 (2011).
4. J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: a survey," *Inf. Fusion* **45**, 153–178 (2019).
5. S. Su et al., "Material classification using raw time-of-flight measurements," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 3503–3511 (2016).
6. D. Mehta et al., "XNect: real-time multi-person 3D motion capture with a single RGB camera," *ACM Trans. Graph.* **39**(4), 82:1–82:17 (2020).

7. R. Wang et al., "A robust registration method for autonomous driving pose estimation in urban dynamic environment using LiDAR," *Electronics* **8**(1), 43 (2019).
8. G. Du et al., "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artif. Intell. Rev.* **54**(3), 1677–1734 (2021).
9. Y. Xiang et al., "PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes," arXiv:1711.00199 (2017).
10. J. Yang et al., "Go-ICP: a globally optimal solution to 3D ICP point-set registration," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(11), 2241–2254 (2015).
11. X. Wang et al., "A coarse-to-fine generalized-ICP algorithm with trimmed strategy," *IEEE Access* **8**, 40692–40703 (2020).
12. Z. J. Yew and G. H. Lee, "RPM-Net: robust point matching using learned features," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 11824–11833 (2020).
13. X. Gu et al., "RCP: recurrent closest point for point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 8216–8226 (2022).
14. P. Kadam et al., "PCRP: unsupervised point cloud object retrieval and pose estimation," in *IEEE Int. Conf. Image Process. (ICIP)*, IEEE, pp. 1596–1600 (2022).
15. C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 3343–3352 (2019).
16. Y. Gao et al., "Sparse dense fusion for 3D object detection," arXiv: 2304.04179 (2023).
17. H. Yang and P. Marco, "Object pose estimation with statistical guarantees: conformal keypoint detection and geometric uncertainty propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.* (2023).
18. J. Gong et al., "Diffpose: toward more reliable 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.* (2023).
19. T. Aleksandr et al., "Image-free single-pixel keypoint detection for privacy preserving human pose estimation," *Opt. Lett.* **49**, 546–549 (2024).
20. J. M. Dos Santos et al., "Color and texture applied to a signature-based bag of visual words method for image retrieval," *Multimedia Tools Appl.* **76**, 16855–16872 (2017).
21. E. R. Vimina and K. P. Jacob, "Feature fusion method using BoVW framework for enhancing image retrieval," *IET Image Process.* **13**(11), 1979–1985 (2019).
22. H. Zhang et al., "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Appl. Soft Comput.* **80**, 57–79 (2019).
23. S. Bell et al., "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 3479–3487 (2015).
24. G. Schwartz and K. Nishino, "Recognizing material properties from images," *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(8), 1981–1995 (2019).
25. M. Cimpoi, S. Maji, and I. Kokinos, "Deep filter banks for texture recognition, description, and segmentation," arXiv:1507.02620 (2015).
26. A. Serrano et al., "The effect of shape and illumination on material perception: model and applications," *ACM Trans. Graph.* **40**(4), 1–16 (2021).
27. S. Mao et al., "Surface material perception through multimodal learning," *IEEE J. Sel. Top. Signal Process.* **16**(4), 843–853 (2022).
28. M. Conde, "A material-sensing time-of-flight camera," *IEEE Sens. Lett.* **4**(7), 1–4 (2020).
29. S. Lee et al., "Surface reflectance estimation and segmentation from single depth image of ToF camera," *Signal Process. Image Commun.* **47**, 452–462 (2016).
30. S. Lee and S. Lee, "Surface IR reflectance estimation and material recognition using ToF camera," in *25th Int. Conf. Pattern Recognit. (ICPR)* (2021).
31. M. A. Mannan et al., "Material information acquisition using a ToF range sensor for interactive object recognition," in *Adv. in Visual Comput.: 7th Int. Symp. (ISVC)* (2011).
32. G. Piella, "A general framework for multiresolution image fusion: from pixels to regions," *Inf. Fusion* **4**(4), 259–280 (2003).
33. X. Bi, S. Yang, and P. Tong, "Moving object detection based on fusion of depth information and RGB features," *Sensors* **22**(13), 4702 (2022).
34. D. Seichter et al., "Efficient RGB-d semantic segmentation for indoor scene analysis," in *IEEE Int. Conf. Rob. and Autom. (ICRA)*, IEEE, pp. 13525–13531 (2021).
35. B. Kuang, J. Yuan, and Q. Liu, "A robust RGB-D SLAM based on multiple geometric features and semantic segmentation in dynamic environments," *Meas. Sci. Technol.* **34**(1), 015402 (2022).
36. Q. Wu et al., "Human 3D pose estimation in a lying position by RGB-D images for medical diagnosis and rehabilitation," in *42nd Annu. Int. Conf. of the IEEE Eng. in Med. & Biol. Soc. (EMBC)*, IEEE, pp. 5802–5805 (2020).
37. J. Wardlaw et al., "A new approach to thermal imaging visualisation—thermal imaging in 3D," Tech. Rep., Rep. no, 2010, University of College London, London, UK (2010).
38. Y. Ham and M. Golparvar-Fard, "An automated vision-based method for rapid 3D energy performance modeling of existing buildings using thermal and digital imagery," *Adv. Eng. Inf.* **27**(3), 395–409 (2013).
39. M. I. Alba et al., "Mapping infrared data on terrestrial laser scanning 3D models of buildings," *Remote Sens.* **3**(9), 1847–1870 (2011).
40. L. Kocsis, P. Herman, and A. Eke, "The modified Beer–Lambert law revisited," *Phys. Med. Biol.* **51**(5), N91 (2006).
41. W. E. Vargas and G. A. Niklasson, "Applicability conditions of the Kubelka–Munk theory," *Appl. Opt.* **36**(22), 5580–5586 (1997).
42. A. S. Marathay and S. Prasad, "Rayleigh-Sommerfeld diffraction theory and Lambert's law," *Pramana* **14**, 103–111 (1980).
43. R. Alcaraz de la Osa et al., "The extended Kubelka–Munk theory and its application to spectroscopy," *ChemTexts* **6**, 1–14 (2020).
44. K. T. Mehta and H. S. Shah, "Simplified method of calculating legendre coefficients for computing optical properties of colorants," *Color Res. Appl.* **10**(2), 98–105 (1985).
45. A. Sabov and J. Krüger, "Identification and correction of flying pixels in range camera data," in *Proc. 24th Spring Conf. Comput. Graph.*, pp. 135–142 (2008).
46. Y. He and S. Chen, "Recent advances in 3D data acquisition and processing by time-of-flight camera," *IEEE Access* **7**, 12495–12510 (2019).
47. A. Kumar and S. S. Sodhi, "Comparative analysis of Gaussian filter, median filter and denoise autoenocoder," in *7th Int. Conf. Comput. for Sustain. Glob. Dev. (INDIACom)*, IEEE, pp. 45–51 (2020).
48. O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum spanning tree based clustering algorithms," in *18th IEEE Int. Conf. Tools with Artif. Intell. (ICTAI'06)* (2006).
49. R. Hecht-Nielsen, *Theory of the Backpropagation Neural Network, Neural Networks for Perception*, pp. 65–93, Academic Press (1992).
50. D. E. Golberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, p. 372, Addison Wesley (1989).
51. J. Xue and B. Shen, "Dung beetle optimizer: a new meta-heuristic algorithm for global optimization," *J. Supercomput.* **79**, 7305–7336 (2022).
52. M. Schmid, D. Rath, and U. Diebold, "Why and how Savitzky–Golay filters should be replaced," *ACS Meas. Sci. Au.* **2**(2), 185–196 (2022).

Dongzhao Yang received his BS degree from Xi'an Jiaotong University, China, in 2023. He is currently pursuing an MS degree in information and communication engineering at the School of Information and Communications Engineering, Xi'an Jiaotong University, China.

Dong An received his BS degree in optical information science and technology from Anhui University, China, in 2019. He is currently pursuing a PhD in optical engineering at the Institute of Modern Optics, Nankai University, China.

Tianxu Xu received her BS and MS degrees from Zhengzhou University, China, in 2015 and 2018, respectively, and her PhD from the Institute of Modern Optics, Nankai University, China, in 2021. She is currently a lecturer at the National Center for International Joint Research of Electronic Materials and Systems, School of Electrical and Information Engineering, Zhengzhou University, China. She serves as a review editor for *Frontiers in Physics*, and a reviewer for several journals.

Yiwen Zhang received her BS degree in optical information science and engineering from Dalian University of Technology, China, in 2018. She is currently pursuing a PhD in optical engineering at the Institute of Modern Optics, Nankai University, China.

Zhongqi Pan is a professor in the Department of Electrical and Computer Engineering, University of Louisiana at Lafayette, United States. He also holds a BORSF Endowed Professorship in Electrical Engineering II and a BellSouth/BORSF Endowed Professorship in Telecommunications. His research is in the area of photonics, including photonic devices, fiber communications, wavelength-division-multiplexing (WDM) technologies, optical performance monitoring, coherent optical communications, space-division-multiplexing (SDM) technologies, and fiber sensor technologies. He has authored/co-authored over 200 papers.

Yang Yue is a professor at the School of Information and Communications Engineering, Xi'an Jiaotong University, China. He is the founder and current PI of Intelligent Photonics Applied Technology Lab (iPatLab). His current research interest is intelligent photonics, including optical communications, optical perception, and optical chips. He is a fellow of SPIE and a senior member of IEEE and Optica. He has published approximately 300 journal papers (including *Science*) and conference papers with more than 12,000 citations.