SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2025, Vol. 68, Iss. 4, 140407:1–140407:19
https://doi.org/10.1007/s11432-024-4256-3

Special Topic: Photonics Technology

Multispectral non-line-of-sight imaging via deep fusion photography

Hao LIU[1], Zhen XU[2], Yifan WEI[2], Kai HAN[1*] & Xin PENG[2*]

[1]*College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China*
[2]*School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Received 3 September 2024/Revised 25 October 2024/Accepted 17 December 2024/Published online 4 March 2025

Abstract Passive non-line-of-sight (NLOS) imaging is a promising technique that extends visual perception to hidden objects around the corner, offering advantages such as low-cost, portability, and real-time. However, the low quality of current passive NLOS images remains a significant barrier to field application of NLOS targets imaging at long standoffs. This study introduces a multispectral NLOS imaging approach utilizing a deep fusion framework to reconstruct images from visible, short-wavelength infrared, and long-wavelength infrared raw data captured by portable devices. The nonlinear representation capabilities and learnable activation function of the Kolmogorov-Arnold network (KAN) are particularly suited to the inverse light field transmission model in NLOS imaging, enhancing the interpretability of the deep neural network. Experimental results demonstrate that this deep fusion photography method provides satisfied performance to image the occluded individuals despite the polynomial attenuation of effective signals with increasing distance between hidden objects and the relay wall. Notably, the passive NLOS experiments reveal successful imaging of hidden people at distance >5 m from the relay wall. Remarkably, even at distances three times greater than those in previous studies, quantitative metrics validate the superior performance of the proposed method in the task of passive NLOS imaging.

Keywords non-line-of-sight imaging, multispectral deep fusion, Kolmogorov-Arnold network, VIS-SWIR-LWIR multispectral imaging, learning-based imaging method

Citation Liu H, Xu Z, Wei Y F, et al. Multispectral non-line-of-sight imaging via deep fusion photography. Sci China Inf Sci, 2025, 68(4): 140407, https://doi.org/10.1007/s11432-024-4256-3

1 Introduction

In real-world scenarios, the objects to be visually precepted are often occluded by obstacles. In 2008, the concept of non-line-of-sight (NLOS) imaging was introduced to recover hidden scenes around the corner [1]. As illustrated in Figure 1, detectors record either three-bounce (active NLOS imaging) or one-bounce (passive NLOS imaging) light field, enabling the reconstruction of NLOS objects using physics-based models and data-driven methods. These reconstruction results offer the capability to see the objects around the corner or behind the occluded wall. Over the past decade, NLOS imaging has attracted vast interests [2–18] because of its potential to overcome limitations in conventional optical imaging and enhance applications such as self-driving perception, survivor rescue, medical imaging, remote sensing, and public safety and security.

NLOS imaging can broadly be categorized into active and passive methods, depending on the presence of an additional light source during imaging. Active NLOS imaging typically employs ultra-short laser and ultrafast detectors, including intensified charge-coupled devices (iCCD) [5, 19], single-photon avalanche diodes (SPAD) [4, 9, 20–23], streak cameras [24], time-of-flight (ToF) [20, 25, 26] cameras, event cameras [27], and superconducting nanowire single-photon detectors (SNSPD) [28]. These devices measure the intensity or photon counts of scattered light from a relay wall and the return time after laser emission. Advanced methods, such as light-cone transform (LCT) [2], f-k [29], Fermat Paths [4], phasor-field virtual wave [30], prior knowledge [31], speckle correlation [32], point spread functions [21], and deep learning [13, 33,34], are utilized to obtain computational images from the blurry raw data measured in the experiments. Despite the relatively high quality of the imaging for the benefit of more modulable signals, the ultrafast laser and single-photon sensitive detectors often applied in the active techniques are prohibitively

* Corresponding author (email: hankai0071@nudt.edu.cn, pengxin_bupt@163.com)

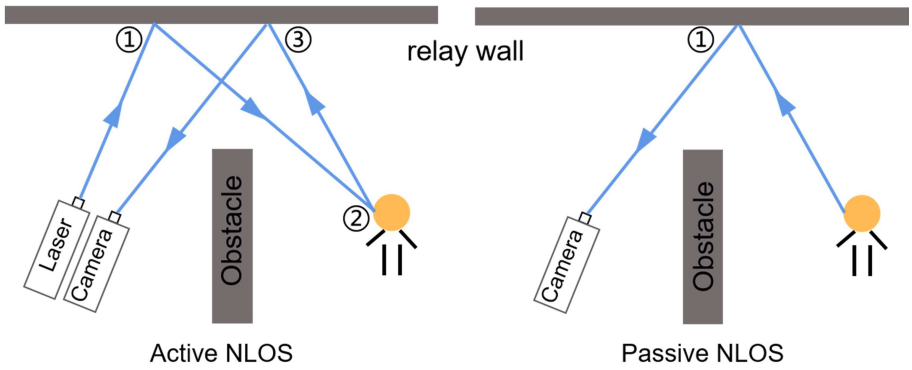

Figure 1 (Color online) Basic principle of NLOS imaging. Active NLOS imaging uses an additional light source to perform three reflections, while passive NLOS imaging only requires a single reflection and does not need an additional light source.

expensive. In addition, the combined illumination and computation time often span several minutes. In contrast, passive NLOS imaging is portable, cost-effective, and well-suited for practical real-world applications.

However, reconstructing hidden objects through passive NLOS imaging is an ill-posed problem. Recent studies have sought to enhance imaging quality using prior knowledge of the scene [5,18,31,35–38], physics-inspired modal ranging from coherence [39], polarization [40,41], long wavelength infrared [42–44], to color information [45], and deep learning methods [14, 46].

Related work via visible light. Bouman et al. [37] developed a corner camera based on a simplified light transmission matrix by employing the bidirectional reflectance distribution function (BRDF). By analyzing variations in ground shadows, they successfully inferred one-dimensional trajectories of individuals around the corner. Baradad et al. [47] explored how occluders code NLOS scenes, enhancing transformation matrix sparsity. Multi-angle NLOS scenes were reconstructed by utilizing shapes and positions of the partial occlusion. Saunders et al. [5] used the light cone matrix method to determine the positions of opaque obstacles and reconstruct images on screens behind them from single scattered photographs recorded by a conventional RGB camera. Hashemi et al. [45] utilized the color domain of the scattered light field and derived the optimization model requiring very few priors, performing accurate reconstruction. Czajkowski et al. [18] exploited two orthogonal edges for transverse resolution and scene representation for z-direction resolution to achieve 3D reconstruction of NLOS targets with an ordinary camera.

Related work via LWIR band. Reflection on relay walls includes specular reflection and diffuse components, with image reconstruction from specular reflection being much easier than the latter. For common materials with typical surface roughness, the specular reflection of long-wavelength infrared (LWIR) is stronger than that of visible light (VIS). Leveraging this, recent studies have utilized the blurred LWIR images scattered from relay wall surfaces to reconstruct NLOS objects. In 2019, Maeda et al. [42] proposed an LWIR NLOS imaging method to estimate hidden human poses in real-time. Subsequently, Kaga et al. [48] developed an empirical model based on the LWIR method to separate diffuse and specular reflection components, thereby estimating the position and temperature of hidden objects. Sasaki et al. [49] estimated the three-dimensional position of NLOS targets using the prior knowledge of BRDF of LWIR scattered by relay wall surface. Hashemi et al. [44] demonstrated an improved contrast-to-noise imaging around the corner using a parallax-driven denoising method.

Related work via learning-based methods. The existing passive NLOS imaging methods often rely on conventional image processing to extract information about hidden objects from reflected light signals. However, the light field transmission in real-world passive NLOS scenarios involves complex physical processes coupled with each other. The NLOS object reconstruction is always a high-dimensional inverse problem. Deep learning methods, with their powerful feature extraction and processing capabilities have found extensive applications in NLOS imaging. By learning mapping functions from a large amount of data, these methods effectively address complex scene information and nonlinear characteristics, excelling in noisy and highly complex environments. He et al. [50] proposed a deep learning framework that simultaneously achieves target reconstruction and dynamic tracking using a standard RGB camera. They first achieved sub-centimeter precision in 3D human pose estimation through a specifically designed neural network and successfully applied the system in real-world scenarios, recovering video signals for moving

targets and visually displaying the trajectories of their motion. Geng et al. [14] proposed the NLOS-OT model, which mapped from the projection image space to the hidden scene space by obtaining latent distribution information from the object space. To train the deep learning-based model, they captured the reflection data of images displayed on a screen after scattered by the relay wall, and constructed the first dataset for passive NLOS imaging. The NLOS-OT model significantly improved the generalization and robustness of reconstruction performance through manifold embedding and optimal transport separation methods.

It is noteworthy that while NLOS imaging studies, including recent advancements in the very latest articles, have achieved detectors-to-really wall distance of up to 1.43 km [9], the distance between hidden people and the relay wall typically remains less than <1.5 m. This limitation constrains practical applications. Given that the imaging light intensity diminishes polynomially with increased object-to-relay wall distance, the signal-to-noise ratio (SNR) declines sharply over distances of several meters, which is common in applications like autonomous driving perception and public security. Advancing NLOS imaging for real-world scenes requires innovative methods to reconstruct the NLOS objects from raw data effectively.

To address the challenges of NLOS imaging, on the basis of our previous physics-guided learning-based NLOS studies [17, 41], we propose a deep fusion photography by employing multispectral imaging data to reconstruct the scene around the corner based on Kolmogorov-Arnold network (KAN) embedded deep neural network (DNN) architecture. This method aims to extract and fuse both shallow and deep features of data from diverse spectral bands to improve the imaging quality of NLOS targets. To our best knowledge, this is the first NLOS imaging method that utilizes the unique advantages of VIS-short wavelength infrared (SWIR)-LWIR multispectral bands and their fusion. Experiment results, comparison with the latest state-of-the-art NLOS imaging techniques, and detailed discussions demonstrate that deep fusion photography outperforms prior passive NLOS imaging methods to see the hidden individuals around corners.

2 Experiments setup

To achieve a more realistic scene, we configured NLOS devices to observe blurry but raw images projected onto the relay wall surface from the objects around the corners of a building corridor. Figure 2 showcased both the schematic diagram and the actual scene used for the measurements. The experimental setup comprised four components: the hidden objects (people and surroundings), a relay wall, the NLOS image acquisition unit, and the ground-truth (GT) image acquisition unit. We employed VIS, SWIR, and LWIR cameras to build the multispectral NLOS imaging devices to obtain three spectral bands of raw data. These three cameras in both NLOS and GT imaging units have been aligned and calibrated by the multi-vision algorithm.

The virtual GT position was symmetrically aligned with the NLOS imaging unit along the relay wall surface. However, since the relay wall was opaque, the GT was placed in the vicinity to capture direct sight-line images of the target.

Notably, images captured by the LWIR camera served as mask images for guiding the training process of the deep learning algorithm. Additionally, all cameras used in the NLOS and GT image acquisition units were of the same model to ensure consistency and comparability to build a dataset for deep learning training and validation.

During the measurements, the target-relay wall distance ranged from 5–6 m, and the distance between the NLOS imaging unit and the relay wall was 3.5 m. The spectral response band of the VIS, SWIR, and LWIR cameras is 400–900 nm, 900–1700 nm, and 8–14 μm, respectively. The final resolution of the raw images is 640×512 pixels. Each camera is coupled with a zoom lens with a suitable focal length for adjusting to the appropriate field of view during the measurements for subsequent image analysis. All cameras are externally triggered in synchronization with the same temporal series.

The material and surface roughness of the relay wall were critical considerations. While a mirror-like relay wall would dramatically reduce the difficulty of NLOS imaging, such smooth relay walls with high specular reflectivity are uncommon in real-world applications. Thus, we chose an unpolished wooden board with a surface roughness of $R_a \sim 12$ μm as the relay wall for our experiments. Its surface properties, including the predominance of diffuse reflection rather than specular reflection in its BRDF characteristics, present inherent challenges to high-quality NLOS imaging. Despite these challenges, we deliberately

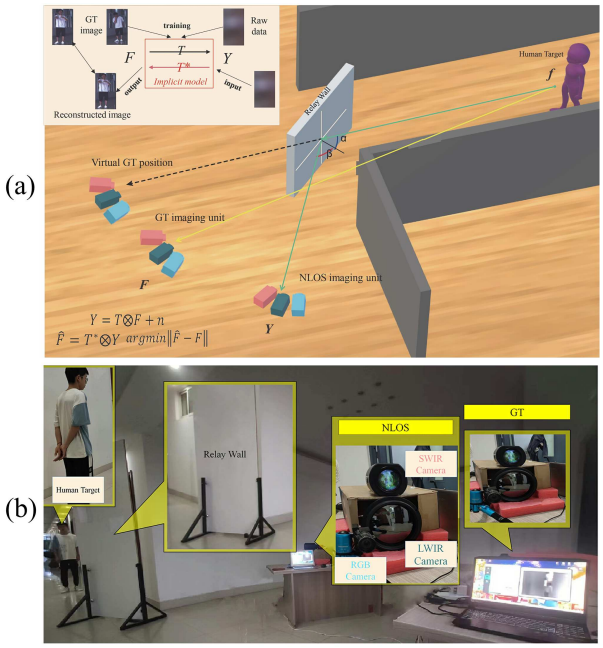

Figure 2 (Color online) (a) Schematic diagram and (b) actual scene of the NLOS imaging experiments. The data collection process was conducted in a corridor, comprising four components from left to right: the imaging target, a relay wall, the NLOS image acquisition unit, and the ground-truth image acquisition unit. Both units include the same VIS, SWIR, and LWIR cameras.

used the wooden board as the relay surface to demonstrate the superiority of the proposed deep fuse photography.

3 Deep fusion photography method

3.1 Passive NLOS imaging basic theory

Passive NLOS imaging involves capturing ambient light emitted or reflected by hidden human targets. The photons of this light are diffusely reflected by a relay wall before being received by the sensor, as shown in Figure 2. The light travels from the hidden target to the visible relay wall and finally reaches the LWIR, SWIR, and VIS cameras.

The light intensity of hidden target f was reflected (both specularly and diffusely) on the relay wall surface, and propagated to position y where the NLOS imaging unit captured the signal. Considering detector noise, the intensity measured by the NLOS imaging devices can be expressed as

$$I(y) = \iint_{f \in Q} T(f, y)\, I(f)\, \mathrm{d}f + n_b(y), \tag{1}$$

where $I(y)$ represents the signal intensity received by the sensor. Q is the set of hidden targets, where f denotes the position of an NLOS target. $T(f, y)$ represents the light transport matrix from the position f to the observation point y. $n_b(y)$ represents the noise at the observation point y.

Eq. (1) can be simplified as

$$Y = T(\mu, r, y, f)F + n_b, \tag{2}$$

where μ is the BRDF of the relay wall, r is the distance between the human target and the relay wall, and Y is the image obtained by the camera on the relay wall.

Passive NLOS imaging inherently represents an inverse problem in which we use the measured blurry raw image Y to reconstruct the target image F. The reconstructed image

$$\widehat{F} = T^*(\mu, r, y, f)Y. \tag{3}$$

However, due to the attenuation and diffuse reflection during propagation, the inverse problem is often highly ill-posed. Many prior studies utilized optimization [5, 40, 42] and deep learning [14] methods to solve the inverse problem.

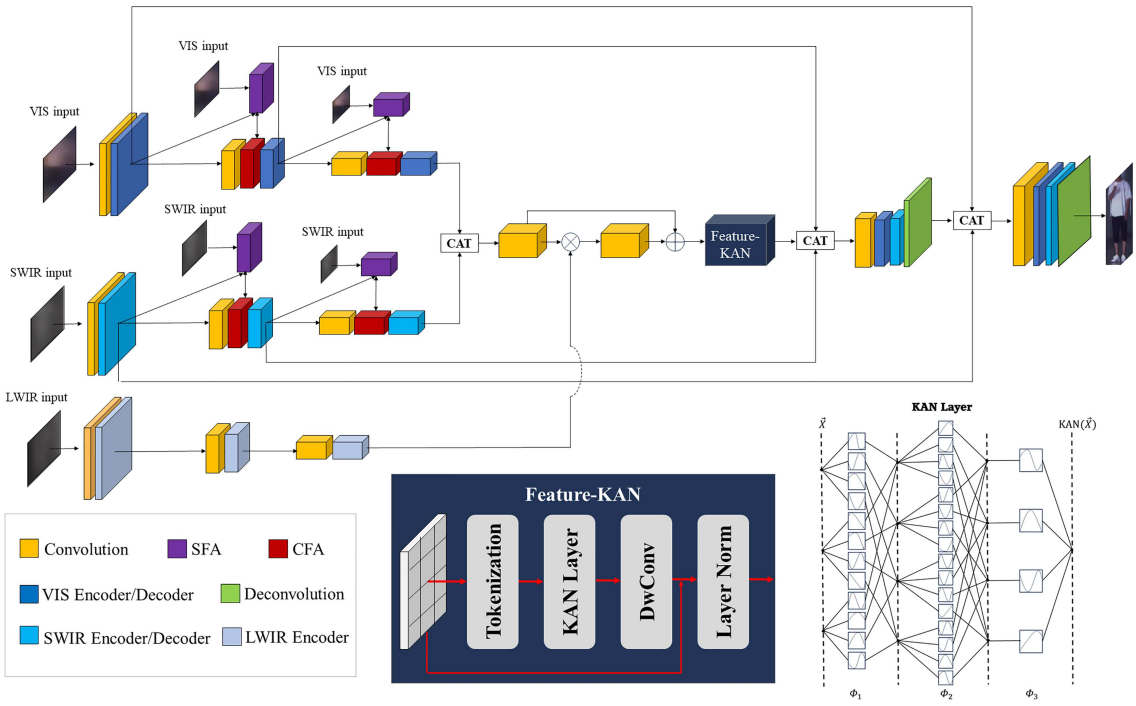

Figure 3 (Color online) Multi-band network architecture schematic. It details the feature processing and feature fusion processes for the three spectral bands.

3.2 KAN-embedding framework

NLOS imaging can be described as an inverse model of light transmission involving scattering (both specular and diffused reflection) on the relay wall surface, requiring the inferential calculations from the blurry raw data. The reconstruction based on the inverse process is a definite ill-posed problem. Thus, the primary objective of the proposed learning-based method is to infer the mapping from the vague raw images after scattering to the GT images. Hence, the photography architecture for better reconstruction of the occluded targets consists of two main steps: feature extraction and feature fusion.

Most existing data-driven passive NLOS imaging methods predominantly use single-band images, either in the visible or LWIR bands. In VIS-based reconstruction methods, the signal attenuates severely during propagation, resulting in a low SNR in an image captured by the camera on the relay wall. Conversely, while LWIR images often have higher SNR compared to VIS images, the VIS better aligns with human vision.

To address these challenges, this paper proposes a three-branch input deep fusion structure that extracts color information from VIS, overall contour information from LWIR images, and detailed texture information from SWIR images. The extracted information is then fused to obtain a higher quality reconstruction image that is more consistent with human vision.

Figure 3 shows the network architecture diagram proposed in this paper. In this task, the network takes LWIR images, SWIR images, and visible images as inputs.

During the encoding stage, downsampling is performed using the F.interpolate function. This method adjusts the feature maps using interpolation, and in this study, bilinear interpolation is employed. This downsampling operation helps gradually decrease the size of the feature maps in the encoding stage while retaining important feature information. By reducing the size of the feature maps, the computational burden is reduced, and higher-level features can be extracted. The interpolation adjustment maintains a certain level of smoothness, which helps preserve some feature information while reducing the size of the feature maps.

To extract more implicit features from the multi-band raw images after the NLOS scattering process, the KAN module is employed in the proposed framework. KAN [51], derived from the Kolmogorov-Arnold representation theorem [52], is a groundbreaking novel neural network architecture recently proposed with excellent approximation capacity and generalization. The Kolmogorov-Arnold representation theorem [52] has proven that multi-variate continuous function can be represented as the superposition of a series

of uni-variable continuous functions:

$$f(X) = f\left(x_1, x_2, \ldots, x_i, \ldots, x_n\right) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}\left(x_p\right) \right). \tag{4}$$

KAN applies adaptive, learnable activation functions, utilizing spline-like uni-variate functions along the edges, thereby significantly improving its capability to model nonlinear processes. This spline-like design achieves high accuracy for target functions, especially in feature extraction and reasoning within physics-informed deep learning [53]. Despite being a nascent development, KAN-based methods have already outperformed MLP-based architectures in tasks such as image classification, segmentation, and time-series prediction [51].

Inspired and driven by the excellent representation and outstanding interpretability of KAN in the area of nonlinear inverse problem solving, we proposed a KAN-embedding framework, as shown in Figure 3. The KAN module extracts more implicit features through its learnable activation functions and nonlinear representations. The framework employs three layers, with the output of the KAN layer performed by depth-wise separable convolution to reduce computational load.

In the decoder stage, transpose convolution layers upsample the feature maps, increasing their size to match the corresponding size in the encoder stage. This allows for the gradual ascent of features, enabling the recovery of high-resolution feature maps and providing richer detail information for subsequent feature fusion and prediction.

In the network, shallow features can usually be obtained with models that have fewer convolutional layers and pooling layers. These shallow features play a significant role in low-level feature extraction, such as color and texture. Deep features can usually be obtained with models that have more convolutional layers and pooling layers. These deep features play a significant role in high-level feature extraction, the spatial distribution of the color, contours, and shapes, and constraint relationships and mappings among them.

To fulfill the criteria of extracting pixel color information from visible images, detail texture information from SWIR images, and contour information from LWIR images, we used a deeper encoder for the visible branch, followed by the SWIR branch, and the shallowest for the LWIR branch. Finally, the features extracted from the three branches are fused and input into the decoder structure. Since the result of the image reconstruction task is to reconstruct an image with rich details and accurate contours, the decoder adopts the same structure as the encoder of the visible branch.

3.3 Feature extraction

In the feature extraction step, we processed VIS images, SWIR images, and LWIR images based on their unique characteristics, and obtained distinctive feature values for each type of image. As shown in Figure 4, y' represents the result of feature extraction from VIS images, E'_{s_obj} represents the result of feature extraction from SWIR images, and E'_{l_obj} represents the result of feature extraction from LWIR images. These features can express the color, texture, contour, and other information of the images in different wavelength bands, thereby better reflecting the true information of the occluded object.

Figure 4 shows that in y', almost no contour or pose information about the model can be captured, but y' has color information that the other two bands do not have, which is exactly what we need. In E'_{s_obj}, it can be observed that the SWIR features have clear texture features, but the distinction between the model and the background is not clear, such as the difference between the model's head and the background. In E'_{l_obj}, the overall contour information of the model is relatively pure, and there is a greater distinction from the background, but there is almost no texture detail of the model itself in the features. The differences in these feature information reflect the role and performance of different bands in NLOS imaging. Therefore, combining features from different bands can better reconstruct the image of the occluded object.

Taking the VIS branch input as an example, the structure of the VIS feature extraction branch is shown in the top-left of Figure 3. The branch mainly consists of three encoder structures. The input to this structure is a 256×256 VIS image. After the first layer convolution and encoder feature extraction, the feature map size is reduced to 128×128. After the second layer of convolution and encoder feature extraction, the feature map size is further reduced to 64×64. In the second and third layer structures, we input the corresponding size of the image into the spatial feature attention (SFA) module for processing.

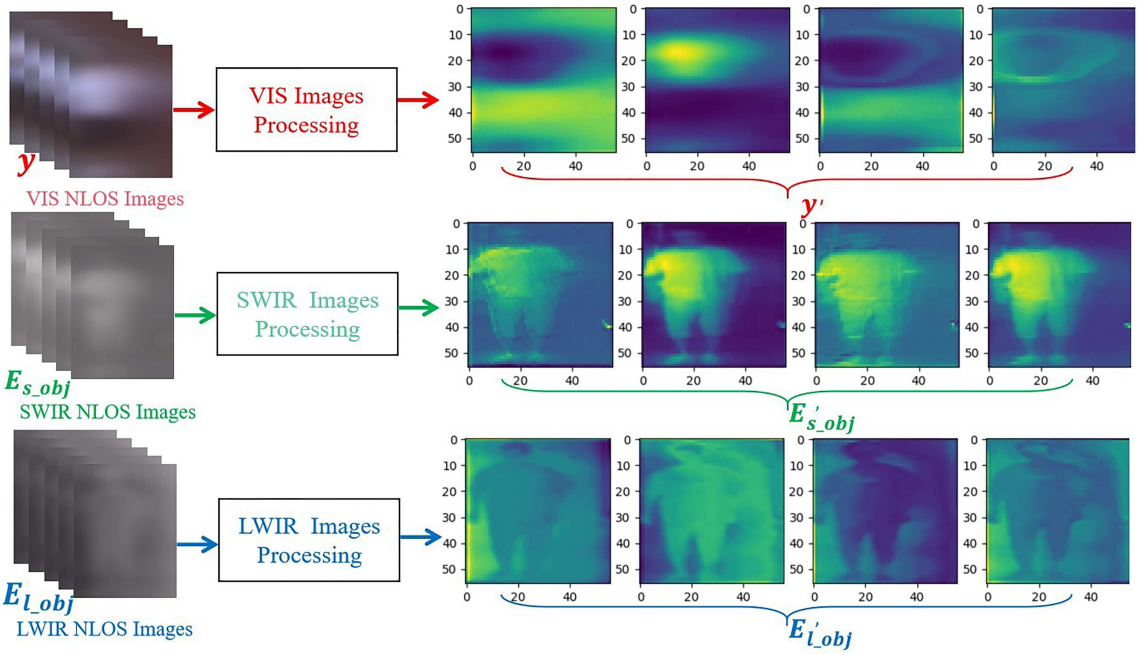

Figure 4 (Color online) Network's feature extraction processes and results for VIS images, SWIR images, and LWIR images.

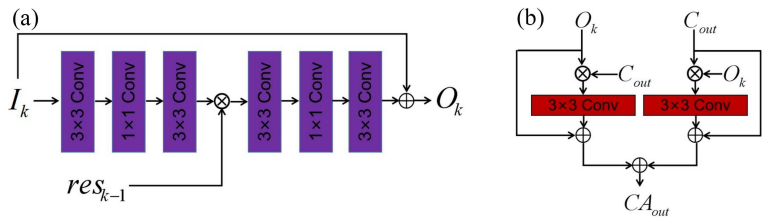

Figure 5 (Color online) (a) Spatial feature attention (SFA) module. It includes 6 convolutional layers. The input consists of the raw data of the NLOS image I_k, and the features res_{k-1} that are the output of the $k-1$ encoder. The output of the module is the weighted image shallow features O_k. This module uses res_{k-1} to weight I_k and then fits I_k to the feature map according to the residual idea, so that the output features are able to be enriched by this module. (b) Channel feature attention (CFA) module. The input of this module includes the output of the SFAM module and the output of the upper convolutional layer, and the output is the feature weighted by cross attention. This module uses O_k as C_{out} weighting, uses C_{out} as O_k weighting, performs convolution processing on the weighted features, and finally fits the features after convolution processing. The CA_{out} has the shallow detail information and deep feature information of the blurred image.

In both the visible and SWIR branches, image clarity improves progressively through multiscale input sub-networks via a coarse-to-fine approach. In this network, SWIR and visible channels use three sizes of images of the same band as inputs, and multiple asymmetric features are fused using the SFA and channel feature attention (CFA) attention modules.

The SFA structure is shown in Figure 5(a). The input I_k of the module represents the corresponding size of the VIS image, res_{k-1} represents the output of the previous layer encoder,k represents the number of times of convolution processing, and O_k represents the output of SFA. To ensure that the sizes of the matrices match, the size of res_{k-1} is compressed to half of itself in advance. In this module, first, downsampling is performed on I_k to extract features. In Figure 5, 3×3, 1×1, and 3×3 are used to stack features, and then the deep feature res_{k-1} is used to weight I_{k+3}. The specific process is as follows:

$$I_{k+4} = \left[\sum_{i=0}^{2} F_{3^\alpha}\left(x_{k+i}, \omega_{k+i}\right)\right]\left[\text{res}_{k-1}\right]. \tag{5}$$

In this context, $F_n(x,\omega)$ and ω represent n convolution processes, where $n = 3^\alpha$, and $\alpha = (i\%2)+1$. The weighted result I_{k+4} is then subjected to feature re-extraction using 3×3 and 1×1 convolutions, and the residual structure is used to merge the features of I_{k+7} and I_{k+1} for feature correction and

supplementation. The specific calculations are as follows:

$$I_{k+7} = \left[\sum_{i=4}^{5} F_{3^{\alpha}}\left(x_{k+i}, \omega_{k+i}\right)\right] + I_k. \tag{6}$$

The end of the SFA module is a 3×3 convolutional layer that sorts and outputs the fused features of I_{k+6} and I_k as O_k:

$$O_k = F_3\left(x_{k+7}, \omega_{k+7}\right). \tag{7}$$

In this module, the stride and padding of the convolutional layer are both set to 1.

For multi-input tasks, while input image sizes vary, it is assumed that the SNR of the images remains consistent during image size compression. Therefore, for original images of different sizes, their contribution to the network is also considered consistent.

Figure 5(b) shows the schematic diagram of CFA, where O_k and C_{out} are the inputs of the module, and CA_{out} is the output of the module. O_k is the output of the SFA module and C_{out} is the downsampled result in the encoder. In this module, the cross-attention module is used to fuse the deep semantic features of O_k and the shallow detail features of C_{out}. The mathematical expression of this module is shown as

$$\text{CA}_{\text{out}} = x_k + x'_k + F\left(x_{k+1}x'_{k+1}, \omega_{k+1}\omega'_{k+1}\right) + F'\left(x_{k+1}x'_{k+1}, \omega_{k+1}\omega'_{k+1}\right), \tag{8}$$

where $(x, \omega) \in O_k$, $(x', \omega') \in C_{\text{out}}$.

The feature maps after feature extraction by the convolutional process are tokenized and further extracted by the KAN module. The tokens are sent into a three layers KAN, and the output of the KAN layers is

$$\text{KAN}(X) = \left(\Phi_3 \odot \Phi_2 \odot \Phi_1\right)(X), \tag{9}$$

where Φ_i $(i = 1, 2, 3)$ represents a KAN layer, and the depth of KAN is set as 3 that is the balance of more intricate patterns capture and risk of overfitting. Each layer transforms the input X through a series of learnable functions $\phi_{q,p}$, resulting in the learning-based method highly adaptable to the inverse light transmission of NLOS imaging.

The deep features then are dealt with depth-wise convolutional (DwConv) layer and a batch normalization layer. The output feature maps at the k-th layer Z_k can be written as follows:

$$Z_k = \text{LN}\left(Z_{k-1} + \text{DwConv}\left(\text{KAN}\left(Z_{k-1}\right)\right)\right). \tag{10}$$

3.4 Feature fusion

For the feature fusion step, we defined a multi-band fusion strategy, as shown in Figure 6. In this strategy, we first concatenated pixel-level features y' and E'_{s_obj} and reduced dimensionality through convolutional computation to obtain a feature $E'_{\text{texture\_obj}}$ with two bands. This method can combine pixel-level information of different bands into a unified feature representation, thus better expressing the information of the original image, such as color and texture. As y' and E'_{s_obj} do not have comprehensive information on posture and pose, we multiplied $E'_{\text{texture\_obj}}$ pixel by pixel with E'_{l_obj} and weighted the pixel values of $E'_{\text{texture\_obj}}$ to make the method more focused on reconstructing the model itself, thereby improving the algorithm's performance. Finally, we upsample and resize the fused feature $\text{Out}'_{\text{Features\_fusion}}$ to restore it to the original size of the image. During the upsampling process, we perform a "cat" operation between the color features $\text{Color}_{\text{features}}$ from X_rex2 and X_rex1 in Figure 6 and the feature $\text{Out}'_{\text{Features\_fusion}}$ to recover the color information of the reconstructed image.

3.5 Loss function

The loss function comprises three components. The first part is the mean absolute error (MAE) at the pixel level, which quantifies the discrepancy between predicted and true values. It is calculated as the average of the absolute differences between all pixel values. In image reconstruction tasks, the MAE loss function effectively penalizes errors in the pixel-level mapping, encouraging the model to learn more precise mapping. However, the MAE loss function does not consider the structure and texture information

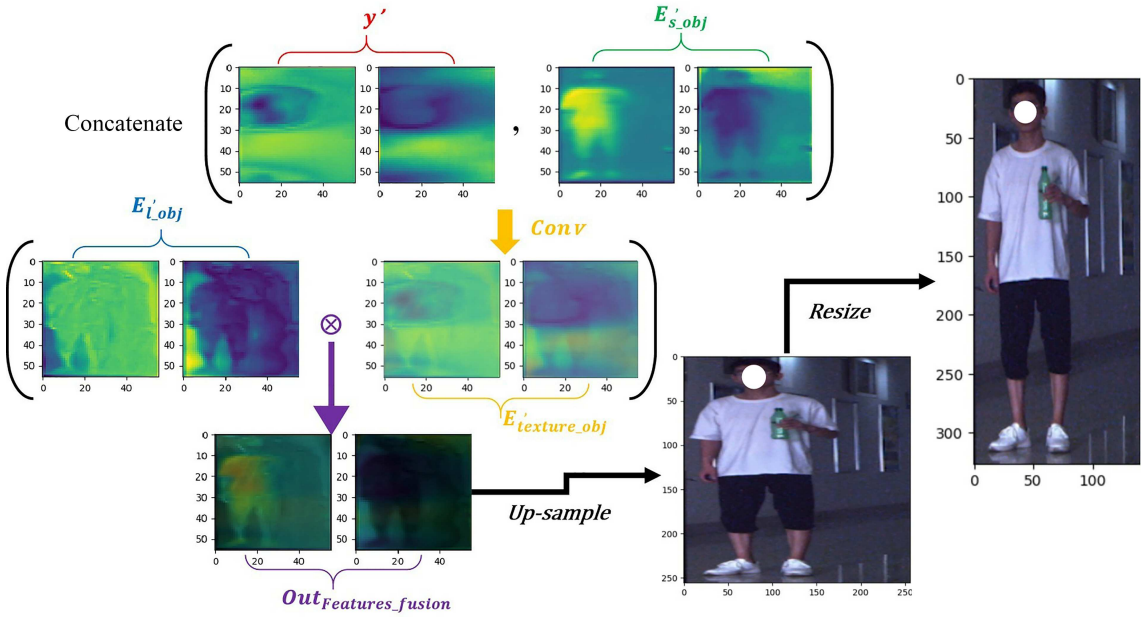

Figure 6 (Color online) Feature fusion process. After fusing the VIS features containing color and texture details with the SWIR features, the result is fused with the LWIR image features. The fused result is then upsampled to obtain the reconstructed image. To protect facial portraits, we artificially use white circles to shield the faces in the publicly published figures.

of the image, which may result in overly smooth and blurry generated images. The MAE loss is defined as follows:

$$L_{\mathrm{MAE}}\left(I_{\mathrm{vis}}, \hat{I}_{\mathrm{vis}}\right) = \frac{1}{H \times W} \sum_{i=1}^{H \times W} \left| I_{\mathrm{vis}}\left(x_i, y_i\right) - \hat{I}_{\mathrm{vis}}\left(x_i, y_i\right) \right|, \tag{11}$$

where $I_{\mathrm{vis}}(x, y)$ is the real VIS image, $\hat{I}_{\mathrm{vis}}(x, y)$ is the result of the network reconstruction, and H and W are the height and width of the image, respectively.

The second component is the structural similarity (SSIM) loss, which measures the similarity between predicted and real images by comparing their structural and texture similarities. The SSIM loss function can promote the generated images to be more realistic and clear while preserving image fine details. The SSIM loss is defined as follows:

$$\mathrm{SSIM}\left(y_i, \hat{y}_i\right) = \left[l\left(y_i, \hat{y}_i\right)^{\alpha} \cdot c\left(y_i, \hat{y}_i\right)^{\beta} \cdot s\left(y_i, \hat{y}_i\right)^{\gamma} \right], \tag{12}$$

where α, β, and γ are parameters used to balance the three factors, which are typically set to 1. $l(y_i, \hat{y}_i)^{\alpha}$, $c(y_i, \hat{y}_i)^{\beta}$, and $s(y_i, \hat{y}_i)^{\gamma}$ are the luminance factor, contrast factor, and structure factor, respectively. Therefore, the SSIM function can be rewritten as

$$\mathrm{SSIM}(y_i, \hat{y}_i) = \frac{\left(2\mu_{y_i}\mu_{\hat{y}_i} + C_1\right)\left(2\sigma_{y_i\hat{y}_i} + C_2\right)}{\left(\mu_{y_i}^2 + \mu_{\hat{y}_i}^2 + C_1\right)\left(\sigma_{y_i}^2 + \sigma_{\hat{y}_i}^2 + C_2\right)}, \tag{13}$$

where C_1 and C_2 are constants added to avoid division by zero, and their values can be adjusted based on the actual situation. In this task, the values of C_1 and C_2 are set to 1.

Due to the uniform experimental background during data acquisition, overfitting can occur in background reconstruction during network training. The loss value of non-region of interest (ROI) regions and their tendency to reach a minimum can reduce the performance of the network in reconstructing ROI regions. To mitigate this, a third component is incorporated into the loss function. The process begins with extracting a VIS mask for the ROI area using the human infrared mask, followed by applying the MAE loss to the network output. The detailed process is shown in Figure 7.

In this work, semantic segmentation of VIS GT images was not performed using a human body recognition network. This decision accounts for scenarios where models may include props or involve specular reflection, which dominate the scattering process on relay wall surfaces or outline targets due to thermal emission difference caused by temperature gradient or emissivity diversity [42, 44, 48]. Inspired by [44],

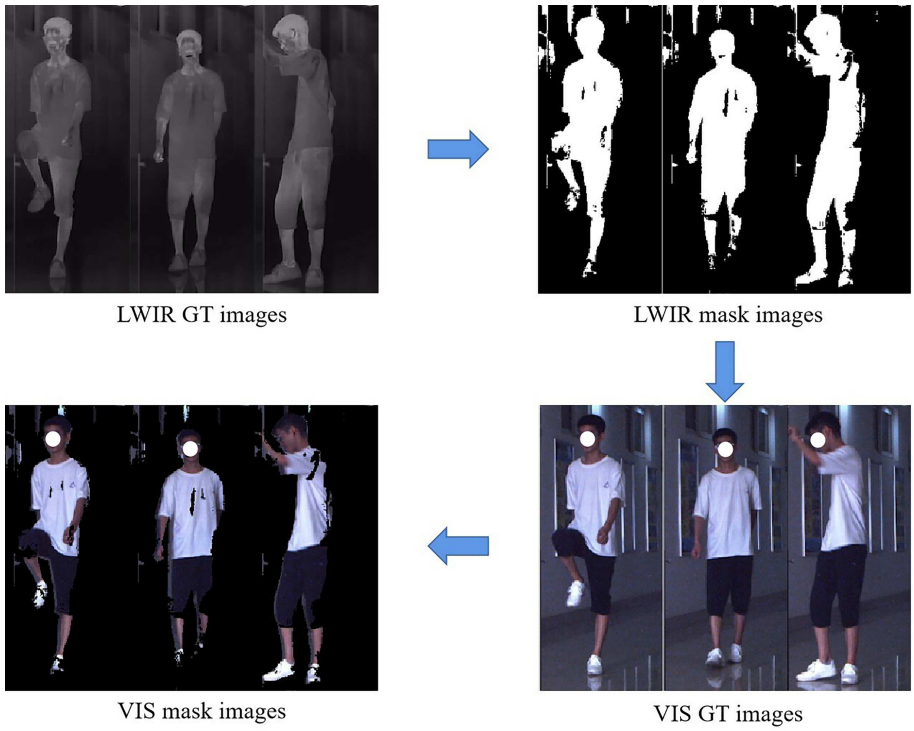

Figure 7 (Color online) Schematic diagram of the VIS mask image generation process. Firstly, the infrared mask image is calculated based on the LWIR ground-truth (GT) image. The normalized infrared mask image is then multiplied pixel by pixel with the VIS GT image to obtain the VIS mask image.

the human body and objects with significant temperature differences were annotated using a calculated infrared mask instead of employing machine learning algorithms. The process to calculate the infrared mask image involved analyzing pixel values in the infrared GT image. Pixels below a defined threshold were set to 0, while those above were set to 255. The number of non-zero pixels in the infrared mask was recorded. The threshold was determined as the average intensity of pixel points in each image.

The calculated infrared mask image was then overlaid on the VIS GT image. Based on the coordinates where the pixel values in the infrared mask image equaled 0, the corresponding pixel points in the VIS GT image were set. This process yielded the VIS GT image after infrared mask processing. During training, a similar procedure was applied to the network's output to generate a VIS-predicted image. The L1 loss was used to calculate the difference between these processed images, with the number of pixels determined by the non-zero pixels in the infrared mask. The process is mathematically described as follows:

$$\bar{G} = \frac{1}{N} \sum_{x=1}^{w} \sum_{y=1}^{\hbar} G(x, y), \tag{14}$$

where $\bar{G}$ represents the average value of pixel intensity, w and $\hbar$ are the width and height of the image, N is the total number of pixels, and $G(x, y)$ represents the intensity value of the image pixel point at coordinates (x, y). The infrared mask image is defined as

$$I_{\text{mask}}(x, y) = \begin{cases} 0, & G(x, y) < \bar{G}, \\ 1, & G(x, y) \geqslant \bar{G}, \end{cases} \tag{15}$$

where I_{mask} represents the calculated infrared mask, and the number of non-zero coordinates is defined as N_{mask}.

Therefore, Eq. (16) for calculating the infrared mask is as follows:

$$\text{VIS}_{\text{mask}}(x, y) = I_{\text{mask}}(x, y) \times I_{\text{vis}}(x, y), \tag{16}$$

where $I_{\text{vis}}(x, y)$ represents the VIS image and $\text{VIS}_{\text{mask}}(x, y)$ represents the VIS image processed by the infrared mask. The symbol "$\times$" represents element-wise multiplication of corresponding pixels. Through

this process, the pixel values in the overlapping areas of the VIS and infrared images are retained, while the pixel values in the non-overlapping areas are set to 0. Consequently, the infrared mask loss is defined as follows:

$$L_{\text{mask}(}y_i, \hat{y}_i) = \frac{1}{N_{\text{mask}}} \sum_{i=1}^{N_{\text{mask}}} \left| I_{\text{vis}}\left(y_i, \hat{y}_i\right) - \text{VIS}_{\text{mask}}\left(y_i, \hat{y}_i\right) \right|. \tag{17}$$

Therefore, the overall joint loss function is shown as follows:

$$L_{\text{joint}} = \lambda_1 L_{\text{MAE}}\left(I_{\text{vis}}, \hat{I}_{\text{vis}}\right) + \lambda_2\left(1 - \text{SSIM}\left(I_{\text{vis}}, \hat{I}_{\text{LWIR}}\right)\right) + \lambda_3 L_{\text{mask}}\left(I_{\text{vis}_{\text{mask}}}, \hat{I}_{\text{vis}_{\text{mask}}}\right), \tag{18}$$

where I_{vis} and I_{LWIR} represent VIS image and LWIR image, respectively. $\hat{I}_{\text{vis}}$ and $\hat{I}_{\text{LWIR}}$ represent the reconstructed VIS image and LWIR image, respectively. $I_{\text{vis}_{\text{mask}}}$ and $\hat{I}_{\text{vis}_{\text{mask}}}$ represent the VIS image mask and reconstructed VIS image mask, respectively. λ_1, λ_2, and λ_3 are hyperparameters used to balance different loss functions, where $\lambda_1 = 0.6$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.3$.

4 Results and discussion

The dataset collects data from two individuals in various human poses. For every measurement illustrated in Figure 2, we record raw images at different wavelength bands by the NLOS imaging unit and ground-truth images by the GT imaging unit. The dataset now has approximately 6000 pairs of data, including 5000 pairs of people with white shirts and 1000 pairs of people with deep red shirts. In order to create different scenes, the people act in diverse human poses at different target-relay wall distances, and hold distinct objects (e.g., bottles and Ping-Pong rackets).

We train our data-driven method with 4000 pairs of data of people with white shirts and 800 pairs of data of people with deep red shirts. The rest of the images are used to test and validate the learning-based model. We set the epoch to 2000 and the learning rate is initially set to 10^{-4}, decreasing by a factor of 0.5 for every 500 epochs. The algorithm experiments are conducted on an i7 Intel-(R) Xeno W-2275 CPU and an NVIDIA RTX A4000 GPU.

4.1 Imaging results via deep fusion photography

The reconstruction of NLOS images from VIS, SWIR, and LWIR bands is shown in Figure 8. From the qualitative perspective, the proposed data-driven method achieves high imaging quality for hidden subjects, including the object contours, human silhouettes, and texture details such as clothing patterns and color.

Three common-used metrics, which are structural similarity index measure (SSIM) [54], peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS) [55], are employed to evaluate the quantitative difference between the reconstructed images and the ground-truth.

SSIM can assess the structural similarity between two images. It has been widely used in the fields of image processing and computer vision, and can measure the degree of distortion of an image, thus evaluating its quality. The closer the value of SSIM is to 1, the higher the similarity between the two images. PSNR measures the logarithmic version of the SNR by calculating the ratio of the peak signal value to the root-mean-square error of the distortion.

The key idea of LPIPS is to use a convolutional neural network (CNN) to extract feature representations of images, and calculate the similarity between images by comparing these feature representations. Compared with conventional metrics such as MSE and PSNR, LPIPS is more in line with human visual perception and can better reflect the quality of images. The evaluation result of LPIPS is a value between 0 and 1, where 0 represents two identical images, and 1 represents two completely different images. That is, the smaller the LPIPS value, the better the network performance.

Table 1 shows high imaging quality with satisfied values of SSIM, PSNR, and LPIPS of both two targets. The white shirt data are dominant in the training set, so the deep neural networks (DNNs) learn more knowledge from the data distribution. The reconstructed images of people in deep red shirts are of lower quality, especially the arms of the target, and the color in the interface area.

According to our best knowledge, this is the first VIS-SWIR-LWIR multi-spectral NLOS imaging work. Consequently, there are no prior studies directly comparable on the exact same dataset or scene. Therefore, we present imaging results using our method alongside state-of-the-art passive NLOS imaging

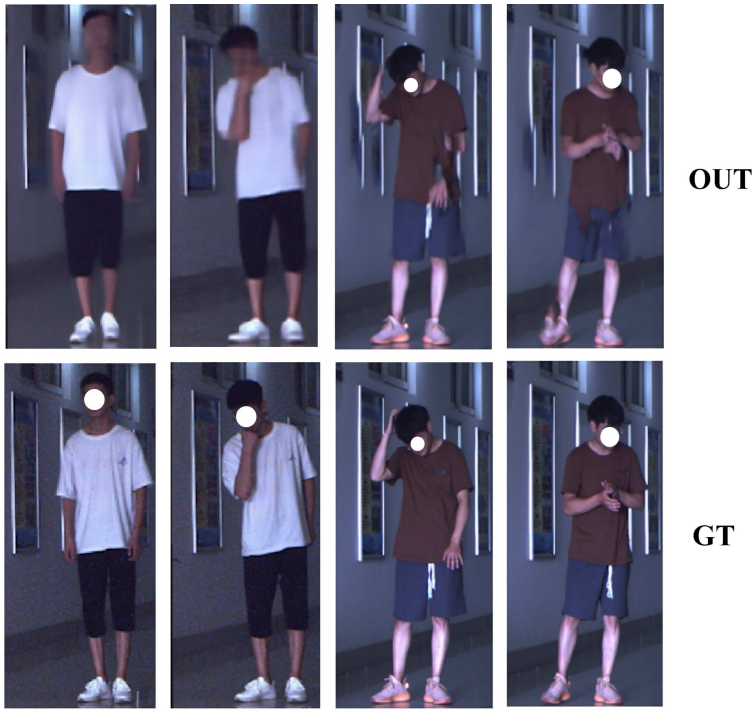

Figure 8 (Color online) Ground-truth images and reconstructed results output by multi-band deep fusion method. The "OUT" denotes the output image, and GT denotes the ground-truth image.

Table 1 Quantitative analysis results of reconstructed images for different targets.

Target	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
People with white shirt	0.88	23.14	0.133
People with red shirt	0.71	22.53	0.181

Table 2 Quantitative analysis results of reconstructed images via different methods.

Method	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
NLOS-OT [14]	0.65	17.96	0.277
MLP embedding method	0.81	21.94	0.152
KAN embedding method (our method)	0.88	23.14	0.133

methods. Figure 9 shows the qualitative comparison among physics-based models [5, 18, 45], learning-based NLOS-OT method [14], and the proposed approach here. It is important to note that each method and scene are diverse, making the results in Figure 9 unsuitable for direct comparison. However, they highlight notable advancements in their respective contexts.

Saunders et al. [5] employed partial occlusions to reduce the condition number, enabling strong generalization ability through their physics-based approach. Due to the neural networks' powerful feature extraction capabilities, learning-based methods, such as NLOS-OT [14] and the DNN model proposed here, achieve better reconstruction quality in specific scenes resembling those in the training dataset. However, both physics-based and learning-based methods have strengths and weaknesses. Passive NLOS imaging remains a trade-off between achieving high imaging quality and ensuring broad scene generalization.

Further comparison between our method and the most recent data-driven passive NLOS method [14] are performed on our multispectral dataset. The algorithm experimental conditions remain the same, and the results are presented in Figure 10 and Table 2. It is evident from the qualitative comparison that our reconstruction results outperform in terms of contour, color, and fine texture details. Furthermore, the quantitative results also confirm that the proposed DNN model here is superior from the perspective of SSIM, PSNR, or LPIPS metrics. The complementary features from cross-modalities of VIS, SWIR, and LWIR are extracted and then fused by the DNN shown in Figure 3, and our method is capable of achieving better passive NLOS quality.

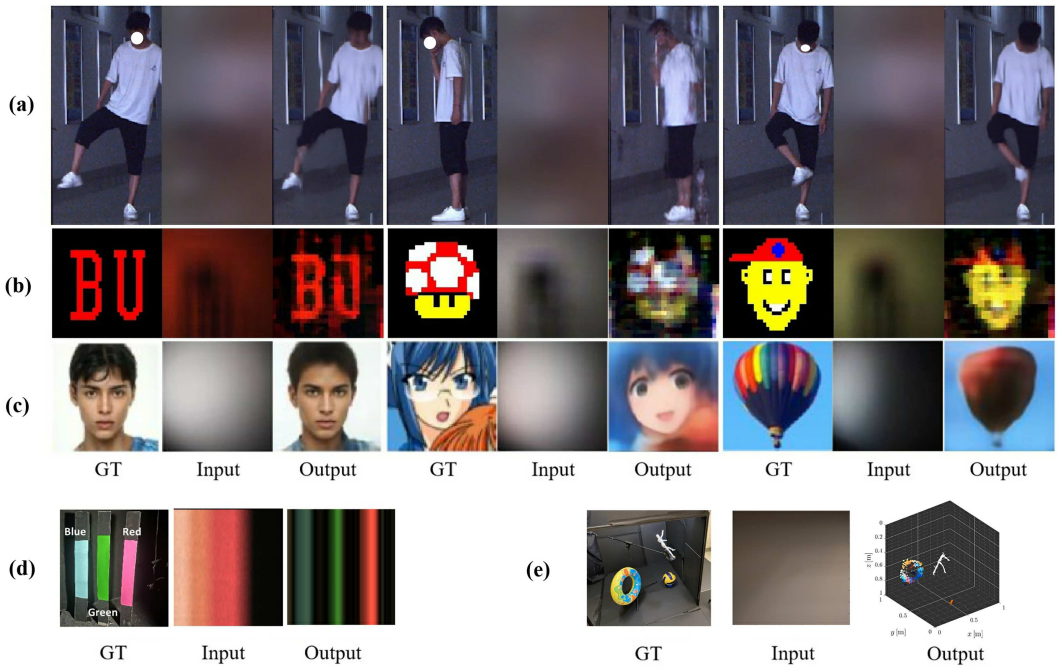

Figure 9 (Color online) Qualitative comparison results with recent state-of-the-art studies by various passive NLOS methods. (a) Our reconstruction result; (b) the result using scene priors [5] Copyright 2019 Nature; (c) the result of NLOS-OT [14] Copyright 2021 IEEE TIP; (d) the visible color NLOS imaging result [45] Copyright 2023 IEEE PAMI; (e) the 3D passive NLOS imaging result [18] Copyright 2024 Nature Communications.

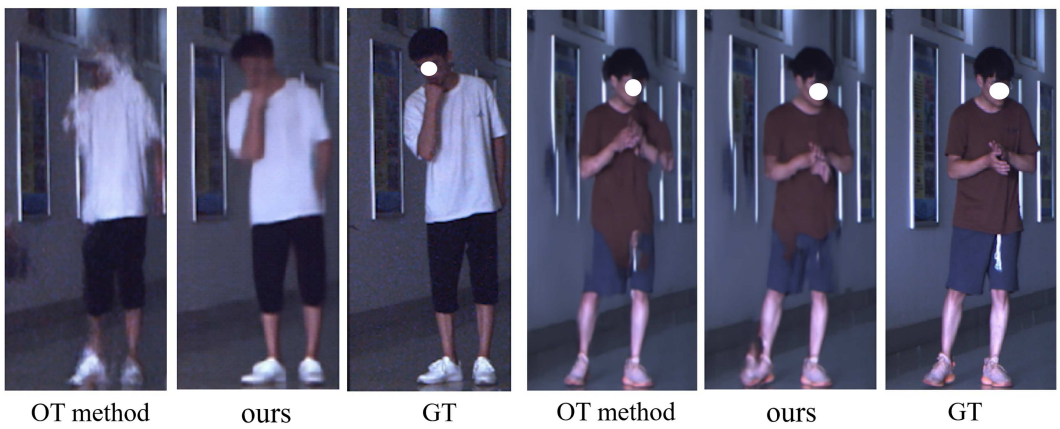

Figure 10 (Color online) Qualitative comparison results of data-driven methods on the same dataset. OT method here is based on prior work named NLOS-OT [14] listed in Table 2.

4.2 Ablation study on spectral bands

An ablation study of spectral bands is conducted to demonstrate the advantages of the multi-spectral fusion method and assess the individual strengths of SWIR and LWIR bands in image reconstruction. We reconstruct the images in the absence of SWIR, LWIR, and both infrared bands. Since the network ultimately outputs visual images, in the band ablation experiments, the three-channel network structure remains unchanged, and only changes the input to the network. For example, in the SWIR ablation experiment with LWIR and VIS as inputs, the SWIR channel input changes to a VIS image, and the corresponding encoder is changed to a structure of 8 ResBlocks which remains the same.

The ablation study results are illustrated in Figure 11. The learning-based method is trained to learn the representation of high-dimensional mapping from the raw images to GT images. For the imaging task of the hidden people around the corner, the proposed method needs to learn two main categories of knowledge: explicit information that is decided by the inverse process of the light field transmission, and the implicit features involved in the data pattern. For the LWIR band, a relay wall surface with a

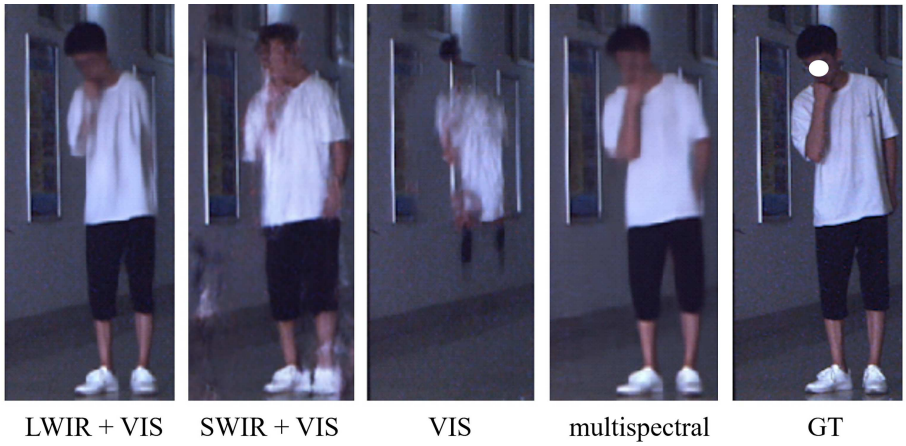

Figure 11 (Color online) Spectral band ablation experiment results. Without changing the shape of the network model backbone (replacing the channel of the ablated band with the channel of the VIS band), NLOS reconstruction is performed using the LWIR and VIS images, SWIR and VIS images, and VIS image alone. Because the VIS absence imaging will loss the color information, the VIS ablation experiment is not included here.

Table 3 Quantitative results of spectral band ablation experiments.

Evaluation standard	LWIR+VIS	SWIR+VIS	VIS	Multispectral
SSIM ↑	0.69	0.78	0.45	0.88
PSNR (dB) ↑	21.59	20.31	14.32	23.14
LPIPS ↓	0.20	0.17	0.42	0.133

roughness of approximately 12 μm acted more mirror-like, enhancing the contours in reconstructed images through specular reflection. Consequently, the LWIR modality contributed more to learning inverse light transport processes than the VIS modality. The quantitative evaluation in Table 3 indicates a decline in metrics when infrared bands are omitted.

However, similar to all thermal imaging techniques, NLOS imaging in the infrared band fails to offer more detailed information, including color and texture, and thus it is adverse to neither human visual perception nor artificial intelligence (AI) perception for subsequent high-level vision tasks (e.g., recognition, segmentation, and semantic cognition). While the VIS modality theoretically can provide ideal complement features, the diffused reflection on the same relay wall introduces challenges in achieving high-quality reconstruction. Ref. [14] has demonstrated the capability of deep learning-based methods to represent prior, data effectively and enhance VIS passive NLOS imaging. With a well-trained deep learning model, the VIS modality can learn data distribution, enhancing the prediction and reconstruction of images through learned mappings between inputs and outputs.

The SWIR band positioned between the LWIR and VIS bands, combines the advantages and limitations of both LWIR and VIS. For instance, SWIR raw images exhibit clearer contour information than VIS but remain blurrier compared to LWIR raw images. While SWIR does not capture color, it retains more texture information than LWIR. Comparing the quantitative metrics of VIS and SWIR+VIS in Table 3, as well as the differences between LWIR+VIS in Table 3 and all multi-spectral bands in Table 1, underscores that incorporating SWIR enhances imaging quality.

According to Table 3, the metrics value of both overall output images and ROI demonstrated strong results. Although the values of ROI slightly reduced, they still indicate satisfactory quality. These minor deviations in overall and ROI values highlight the effectiveness of the proposed data-driven model in the reconstruction tasks for both people and surroundings. This also validates the utility of the human mask loss incorporated in the joint loss function design, as shown in Figure 11 and (18).

4.3 Generalization of the deep fusion photography

Methods that address the inverse problem in passive NLOS imaging generally fall into two categories: (1) interpretable models based on physics-driven approaches, signal processing, optimization, sparsity, and conventional machine learning; (2) data-driven models that, while capable of high-quality imaging, require improvements in interpretability due to their "black box" nature.

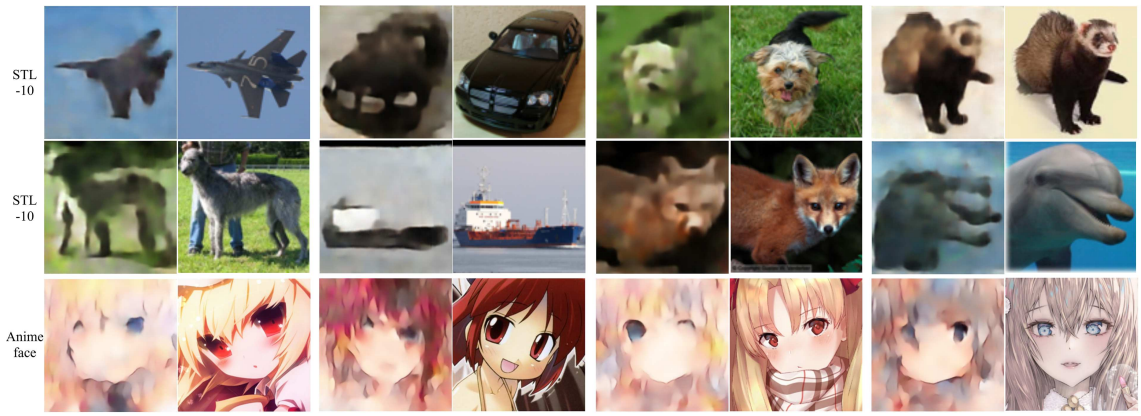

Figure 12 (Color online) Generalization of deep fusion model. The data-driven method is trained on the NLOS-OT dataset [14]. The training set is the supermodel sub-dataset, and the test sets are STL-10 [56] and anime face [14] sub-dataset. The images on the left side are the reconstruction, and the images on the right side are the ground-truth. The figures are sourced from publicly available open-source datasets and used only for academic research purposes.

Table 4 Quantitative metrics of reconstruction results output by our deep fusion method on the NLOS-OT dataset [14]. The training set is supermodel sub-dataset, and the test sets are STL-10 and anime face sub-dataset.

Test set	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
STL-10	0.51	16.37	0.44
Anime face	0.42	14.48	0.53

The former category offers broad generalization across diverse scenarios. For instance, several brilliant previous studies have utilized partial occlusions [5], thermal emission [42], polarization [40], and color [45] to infer the inverse process of light field transport, thereby demonstrating strong generalization potential.

In contrast to the former, the latter method excels in imaging quality, especially as the NLOS target-to-relay wall distance increases, providing that they are trained on sufficient data and applied to similar scenarios.

To assess the generalization capability of the deep fusion method, we validate it using the NLOS-OT dataset [14], which includes diverse scenes and various types of data distribution. Our data-driven model is trained on the supermodel sub-dataset, and tested on the STL-10 sub-dataset and anime face sub-dataset. The results are presented in Figure 12 [56] and Table 4.

The outputs from the never-seen test set show acceptable reconstruction from both qualitative and quantitative perspectives. This indicates that the data-driven deep fusion method here achieves satisfactory generalization performance. Theoretically, generalization depends on the diversity of the training set scenes. With sufficient data, the deep learning method can be trained on broader data distributions and varied scenes, enabling the DNN models to learn more general features, patterns, and knowledge, thereby enhancing generalization. However, increasing the complexity of training scenes often results in a decline in reconstruction quality for specific scenes. As mentioned above, learning-based methods in the field of NLOS imaging and, more broadly, in AI-related research consistently face a trade-off between generalization ability and reconstruction quality [57–59].

For applications involving similar scenes within a specific scope, the data-driven method achieves outstanding reconstruction results when trained and tested on the dataset with appropriate data distribution. In this study, we would like to emphasize the potential of the learning-based deep fusion method in using VIS, SWIR, and LWIR multi-spectral bands for imaging NLOS targets around the corner. Beyond the generalization assessment conducted on the NLOS-OT dataset [14], we further assess its performance on our dataset.

Figure 13 illustrates an example of the deep fusion method's generalization capabilities in this work. The model was trained on our training set, which excluded objects such as bottles and Ping-Pong rackets. When the never-seen data were input into the trained DNN model, it produced reconstructed images of commendable quality. As shown in Figure 13, the LWIR modality offers contour information, the SWIR channel captures texture patterns, and the VIS channel conveys color features. The encoder of the DNN extracts these features from multi-spectral bands, and the network fuses them for reconstruction tasks. The carefully designed deep fusion network can learn the implicit mapping from the features to the GT,

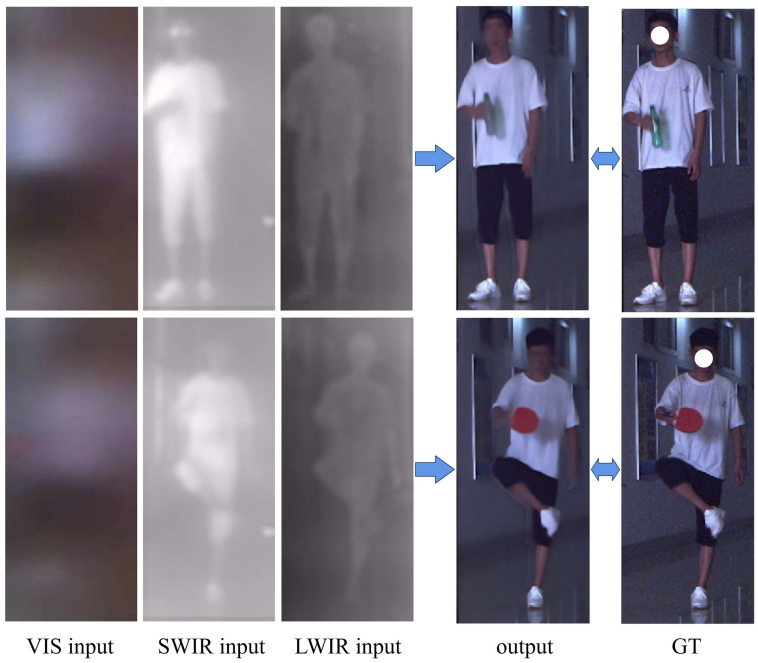

Figure 13 (Color online) Generalization of deep fusion model. The data-driven method is trained on our dataset. There are no bottles nor Ping-Pong rackets in the training set, and the people holding green bottles and red Ping-Pong rackets are set as NLOS imaging targets to test the generalization.

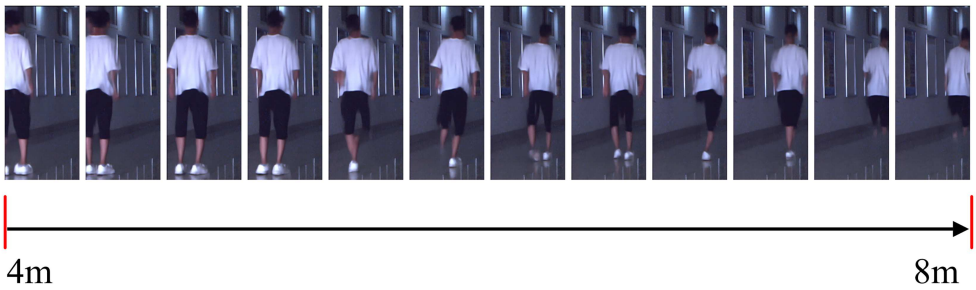

Figure 14 (Color online) Adaptability of deep fusion method to the target-relay wall distance. The target slowly walks away from the relay wall from a distance of 4–8 m.

assimilate the data priors as the distribution of the training set, and achieve high-quality reconstruction. For the never-seen objects, the contour information from the LWIR band and the detailed textures from the SWIR band significantly contribute to the deep fusion method's ability to infer the inverse process of light field transport, thereby improving the reconstruction quality for the never-seen objects. Although the predicted images in Figure 13 show slight inaccuracies in the details of hands, face, and color compared to the GT images, the reconstruction quality of our method still exceeds previous approaches. These results underscore that the method also demonstrates an acceptable level of generalization ability.

Finally, we demonstrate the adaptability of the deep fusion method to variations in distance between the target and the relay wall, which can be interpreted as another aspect of its generalization ability. Figure 14 shows the experimental results when the target-relay wall distance ranges from 4 to 8 m. Due to the polynomial attenuation of effective signals with increasing distance, the reconstruction task becomes more challenging at greater distances. As the target moves further from the relay wall, imaging quality decreases. Furthermore, since multi-camera alignment is optimized for a fixed distance of approximately 5 m, the data captured at varying distances may lack precise alignment across cameras. For instance, as shown in Figure 14, images captured at greater distances exhibit vagueness in reconstruction details, such as when one leg occludes the other, or when legs are positioned on different planes. Nevertheless, the overall imaging quality remains acceptable.

These results collectively demonstrate the proposed method's balance between high-quality reconstruction and its reasonable generalization ability.

5 Conclusion

In this study, we proposed a deep fusion method for imaging hidden individuals beyond the line-of-sight. Unlike prior research, where the target-relay wall distance typically remains <1.5 m, our work focuses on distance >5 m. While this experimental setup aligns more closely with real-world scenarios, it introduces challenges due to the polynomial attenuation of the signal-to-noise ratio at increased distances. To address this, we developed a multispectral NLOS imaging system employing VIS, SWIR, and LWIR cameras, supported by a deep learning framework. This framework solves the inverse problem of passive NLOS imaging by representing implicit mapping relationships between the blurry raw images and the target.

To implement the learning strategy, we introduce the KAN-embedded feature extraction module and a multi-band feature fusion network that processes NLOS images from different wavelength bands, enabling high-quality passive NLOS imaging tasks. The encoder structure within the feature extraction channel was tailored to accommodate the depth required for feature processing across different spectral bands. Unlike the multi-layer perception (MLP)-based DNN methods, the KAN can leverage learnable activation functions to model nonlinear mapping in the inverse problem, making it particularly effective for physics-based model challenges. Experimental results showed that this optimization enhances the reconstruction quality without impacting the network's operational efficiency. In the proposed neural network, we leverage GT images from LWIR and VIS to define a sub-loss function called the VIS mask loss, which directs the model to focus more on reconstructing hidden people targets during training. In addition, multiple sets of spectral band ablation experiments demonstrated that using different spectral band images as input in the network can provide richer information, thus improving the quality and accuracy of NLOS imaging. For spectral band combinations, VIS images contribute color information, SWIR images provide detailed information, and LWIR images capture contour information. The integration of these spectral bands enhances image reconstruction using their distinct physical characteristics.

The qualitative and quantitative analyses of the experiments demonstrate high-quality NLOS imaging results. Three standard quantitative metrics, SSIM, PSNR, and LPIPS, are employed to evaluate the reconstructed images. This study first highlights the strengths and weaknesses of the proposed method and compares it to other state-of-the-art methods in specific scenes using the same dataset. The proposed multispectral NLOS method outperforms competing approaches across all quantitative metrics. The spectral band ablation study further underscores the utility of infrared bands, demonstrating that each wavelength contributes valuable cross-modal features. This enables the deep fusion method to deliver superior outputs. Addressing a common concern in all deep learning methods, the generalization capability of the data-driven method here was rigorously discussed. Validations using never-seen test data from both the self-build dataset and the public dataset with broader data distribution confirmed the generalization capability of the proposed method. Furthermore, adaptability to varying target-relay wall distances further supports its generalization. In summary, deep fusion photography demonstrates exceptional imaging capability as well as reasonable generalization, making it significant for potential real-world applications, including self-driving perception, medical imaging, and public security.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62272421, 62375283) and Science Fund Program for Outstanding Young Scholars of Hunan Province (Grant No. 2024JJ4044). The authors thank Dr. Xin HE from Zhengzhou University for useful help on the code discussions, and appreciate Dr. Xin HE and Mr. Pengfei WANG for the dataset acquisition.

References

1 Faccio D, Velten A, Wetzstein G. Non-line-of-sight imaging. Nat Rev Phys, 2020, 2: 318–327
2 O'Toole M, Lindell D B, Wetzstein G. Confocal non-line-of-sight imaging based on the light-cone transform. Nature, 2018, 555: 338–341
3 Batarseh M, Sukhov S, Shen Z, et al. Passive sensing around the corner using spatial coherence. Nat Commun, 2018, 9: 3629
4 Xin S, Nousias S, Kutulakos K N, et al. A theory of Fermat paths for non-line-of-sight shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 6800–6809
5 Saunders C, Murray-Bruce J, Goyal V K. Computational periscopy with an ordinary digital camera. Nature, 2019, 565: 472–475
6 Rapp J, Saunders C, Tachella J, et al. Seeing around corners with edge-resolved transient imaging. Nat Commun, 2020, 11: 1–10
7 Metzler C A, Heide F, Rangarajan P, et al. Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging. Optica, 2020, 7: 63–71
8 Feng X, Gao L. Ultrafast light field tomography for snapshot transient and non-line-of-sight imaging. Nat Commun, 2021, 12: 2179

9 Wu C, Liu J, Huang X, et al. Non-line-of-sight imaging over 1.43 km. Proc Natl Acad Sci USA, 2021, 118: e2024468118
10 Shen S, Wang Z, Liu P, et al. Non-line-of-sight imaging via neural transient fields. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 2257–2268
11 Liu X, Wang J, Li Z, et al. Non-line-of-sight reconstruction with signal-object collaborative regularization. Light Sci Appl, 2021, 10: 198
12 Cao R, de Goumoens F, Blochet B, et al. High-resolution non-line-of-sight imaging employing active focusing. Nat Photon, 2022, 16: 462–468
13 Mu F, Mo S, Peng J, et al. Physics to the rescue: deep non-line-of-sight reconstruction for high-speed imaging. IEEE Trans Pattern Anal Mach Intell, 2024. doi: 10.1109/TPAMI.2022.3203383
14 Geng R, Hu Y, Lu Z, et al. Passive non-line-of-sight imaging using optimal transport. IEEE Trans Image Process, 2021, 31: 110–124
15 Seidel S, Rueda-Chacón H, Cusini I, et al. Non-line-of-sight snapshots and background mapping with an active corner camera. Nat Commun, 2023, 14: 3677
16 Huang X, Ye R, Li W, et al. Non-line-of-sight imaging and vibrometry using a comb-calibrated coherent sensor. Phys Rev Lett, 2024, 132: 233802
17 Jin S, Xu Z, Xu M, et al. Time-gated imaging through dense fog via physics-driven Swin transformer. Opt Express, 2024, 32: 18812–18830
18 Czajkowski R, Murray-Bruce J. Two-edge-resolved three-dimensional non-line-of-sight imaging with an ordinary camera. Nat Commun, 2024, 15: 1162
19 Laurenzis M, Velten A. Nonline-of-sight laser gated viewing of scattered photons. Opt Eng, 2014, 53: 023102
20 Nam J H, Brandt E, Bauer S, et al. Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. Nat Commun, 2021, 12: 6526
21 Pei C, Zhang A, Deng Y, et al. Dynamic non-line-of-sight imaging system based on the optimization of point spread functions. Opt Express, 2021, 29: 32349–32364
22 Ahn B, Dave A, Veeraraghavan A, et al. Convolutional approximations to the general non-line-of-sight imaging operator. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 7889–7899
23 Tsai C Y, Sankaranarayanan A C, Gkioulekas I. Beyond volumetric albedo—a surface optimization framework for non-line-of-sight imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 1545–1555
24 Velten A, Willwacher T, Gupta O, et al. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. Nat Commun, 2012, 3: 745
25 Kadambi A, Zhao H, Shi B, et al. Occluded imaging with time-of-flight sensors. ACM Trans Graph, 2016, 35: 1–12
26 Boger-Lombard J, Katz O. Passive optical time-of-flight for non line-of-sight localization. Nat Commun, 2019, 10: 3343
27 Wang C, He Y, Wang X, et al. Passive non-line-of-sight imaging for moving targets with an event camera. Chin Opt Lett, 2023, 21: 061103
28 Feng Y, Cui X, Meng Y, et al. Non-line-of-sight imaging at infrared wavelengths using a superconducting nanowire single-photon detector. Opt Express, 2023, 31: 42240–42254
29 Isogawa M, Chan D, Yuan Y, et al. Efficient non-line-of-sight imaging from transient sinograms. In: Proceedings of the 16th European Conference on Computer Vision, Glasgow, 2020. 193–208
30 Liu X, Guillén I, Manna M L, et al. Non-line-of-sight imaging using phasor-field virtual wave optics. Nature, 2019, 572: 620–623
31 Heide F, O'Toole M, Zang K, et al. Non-line-of-sight imaging with partial occluders and surface normals. ACM Trans Graph, 2019, 38: 1–10
32 Chopite J G, Hullin M B, Wand M, et al. Deep non-line-of-sight reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 960–969
33 Li Y, Peng J, Ye J, et al. NLOST: non-line-of-sight imaging with transformer. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 13313–13322
34 Su X, Hong Y, Ye J, et al. Model-guided iterative diffusion sampling for NLOS reconstruction. IEEE J Sel Top Quantum Electron, 2024, 30: 1–11
35 Yedidia A B, Baradad M, Thrampoulidis C, et al. Using unknown occluders to recover hidden scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 12231–12239
36 Wang Y, Zhang Y, Huang M, et al. Accurate but fragile passive non-line-of-sight recognition. Commun Phys, 2021, 4: 88
37 Bouman K L, Ye V, Yedidia A B, et al. Turning corners into cameras: principles and methods. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2270–2278
38 Lin D, Hashemi C, Leger J R. Passive non-line-of-sight imaging using plenoptic information. J Opt Soc Am A, 2020, 37: 540–551
39 Beckus A, Tamasan A, Atia G K. Multi-modal non-line-of-sight passive imaging. IEEE Trans Image Process, 2019, 28: 3372–3382
40 Tanaka K, Mukaigawa Y, Kadambi A. Polarized non-line-of-sight imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 2136–2145
41 Liu H, Wang P, He X, et al. PI-NLOS: polarized infrared non-line-of-sight imaging. Opt Express, 2023, 31: 44113–44126
42 Maeda T, Wang Y, Raskar R, et al. Thermal non-line-of-sight imaging. In: Proceedings of IEEE International Conference on Computational Photography (ICCP), 2019. 1–11
43 Divitt S, Gardner D F, Watnik A T. Passive, thermal, reference-free, non-line-of-sight imaging. In: Proceedings of Conference on Lasers and Electro-Optics (CLEO), 2020
44 Hashemi C, Sasaki T, Leger J. Parallax-driven denoising of passive non-line-of-sight thermal imagery. In: Proceedings of IEEE International Conference on Computational Photography (ICCP), 2023. 1–12

45 Hashemi C, Avelar R, Leger J. Isolating signals in passive non-line-of-sight imaging using spectral content. IEEE Trans Pattern Anal Mach Intell, 2024. doi: 10.1109/TPAMI.2023.3301336

46 Chen M, Liu H, Jin S, et al. Hyper-NLOS: hyperspectral passive non-line-of-sight imaging. Opt Express, 2024, 32: 34807–34824

47 Baradad M, Ye V, Yedidia A B, et al. Inferring light fields from shadows. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6267–6275

48 Kaga M, Kushida T, Takatani T, et al. Thermal non-line-of-sight imaging from specular and diffuse reflections. IPSJ T Comput Vis Appl, 2019, 11: 1–6

49 Sasaki T, Hashemi C, Leger J R. Passive 3D location estimation of non-line-of-sight objects from a scattered thermal infrared light field. Opt Express, 2021, 29: 43642–43661

50 He J H, Wu S K, Wei R, et al. Non-line-of-sight imaging and tracking of moving objects based on deep learning. Opt Express, 2022, 30: 16758–16772

51 Li C, Liu X, Li W, et al. U-KAN makes strong backbone for medical image segmentation and generation. 2024. ArXiv:2406.02918

52 Givental A B, Khesin B A, Marsden J E, et al. On the representation of functions of several variables as a superposition of functions of a smaller number of variables. In: Collected Works. Berlin: Springer, 2009. 25–46

53 Shukla K, Toscano J D, Wang Z, et al. A comprehensive and FAIR comparison between MLP and KAN representations for differential equations and operator networks. Comput Methods Appl Mech Eng, 2024, 431: 117290

54 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612

55 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 586–595

56 Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning. J Mach Learn Res Proc Track, 2011, 15: 215–223

57 Genkin M, Engel T A. Moving beyond generalization to accurate interpretation of flexible models. Nat Mach Intell, 2020, 2: 674–683

58 Karniadakis G E, Kevrekidis I G, Lu L, et al. Physics-informed machine learning. Nat Rev Phys, 2021, 3: 422–440

59 Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence. Nature, 2023, 620: 47–60