Research Article Vol. 33, No. 20 / 6 Oct 2025 / Optics Express 41937
Optics EXPRESS

Non-line-of-sight multi-person pose sensing

YUSEN HOU,[1,†] XINGYU CUI,[1,†] SHIDA SUN,[2] YUE LI,[2] JING HUANG,[3] ZHI LU,[4] KUN LI,[3] ZHIWEI XIONG,[2] AND JINGYU YANG[1,*]

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[2]Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China
[3]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
[4]Department of Automation, Tsinghua University, Beijing 100084, China
[†]The authors contributed equally to this work.
[*]yjy@tju.edu.cn

Abstract: In the fields of anti-terrorism and emergency rescue, human pose sensing under non-line-of-sight (NLOS) conditions plays a critical role in enabling informed decision-making and effective response. Current research predominantly focuses on line-of-sight (LOS) scenarios but pays limited attention to NLOS conditions, and even these NLOS studies are confined to single-person pose detection, significantly restricting real-world multi-person applications. In this paper, we propose the first method that enables adaptive multi-person pose sensing in NLOS environments. Our approach reconstructs coarse 3D features from transients and refines these features through a neural network. The refined features are then utilized to predict a 3D body center heatmap and a mesh parameter map. Finally, using 3D body center-guided parameter sampling, the method enables adaptive multi-person 3D mesh regression. Comprehensive experiments on both simulated data and real-world data acquired using our self-built NLOS system demonstrate that our method effectively senses human poses in multi-person NLOS scenarios. Notably, the simulation pipeline we developed for constructing the multi-person NLOS dataset and the real-world data captured by our system provide valuable resources for advancing research in this field. We will publicly release our dataset to support the development of new methods and foster progress in NLOS human pose sensing.

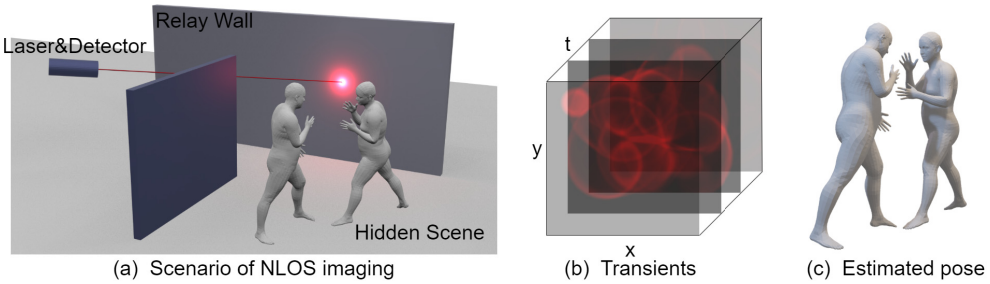
1. Introduction

Human pose estimation is a prevalent task in computer vision, supporting diverse applications such as motion analysis, human-computer interaction, and autonomous systems. Most existing methods in this domain are designed for general scenarios [1–7], where direct visual information is typically available to infer body poses accurately. However, in many practical scenarios, such as robot vision, autonomous driving, rescue operations, remote sensing, and medical imaging, the line of sight may be obstructed by walls, barriers, or other occlusions. Addressing these challenges makes non-line-of-sight (NLOS) human pose sensing critical.

NLOS imaging [8,9] is a technique that enables the detection of objects that are hidden or not directly visible by capturing light that has reflected off surfaces around the objects. Figure 1 shows a typical active confocal NLOS imaging system; a laser source and a detector are both focused on the same point on a relay surface. Pulses emitted by the laser reflect through the surface to illuminate the hidden scene. The detector captures photons that bounce back from the scene to the relay surface, represented as 3D spatiotemporal histograms of photons, known as transients. Using different imaging algorithms, such as Light Cone Transform(LCT [10]), Phasor Field (PF [11]), or F-K [12], the hidden scene can be reconstructed from the transients.

#570120 https://doi.org/10.1364/OE.570120
Journal © 2025 Received 17 Jun 2025; revised 31 Aug 2025; accepted 9 Sep 2025; published 24 Sep 2025

Optics EXPRESS



(a) Scenario of NLOS imaging (b) Transients (c) Estimated pose

Fig. 1. A typical active confocal NLOS imaging system. (a) A pulsed laser, combined with a picosecond detector (such as a SPAD), emits photons towards a relay surface. (b) 3D transient measurements from the optical NLOS system. The detector captures photons reflected from the hidden scene. (c) Estimated pose of the hidden scene.

In recent years, NLOS imaging technology has gained increasing attention. Physics-based direct reconstruction algorithms [10–16] have enabled the rapid reconstruction of hidden scenes. While model-based iterative algorithms [17–19] have achieved high-quality reconstructions, their iterative nature has long been a bottleneck, resulting in extended inference times. Recently, deep learning-based algorithms have made significant advancements in the NLOS field due to their ability to learn and perform rapid inference. Chopite et al. [20] were the first to introduce deep learning into the task of NLOS scene reconstruction. Chen et al. [21] introduced a method for learning feature embeddings tailored to NLOS reconstruction. They extract sparse hidden features from simulated transient images using learned feature extraction blocks and feature propagation units that leverage physical models, enabling effective scene reconstruction. Yu et al. [22] proposed a novel approach that enhances physics-based NLOS imaging methods by introducing a learnable inverse kernel in the Fourier domain and using an attention mechanism to improve the neural network to learn high-frequency information.

Although NLOS imaging technology has garnered increasing attention in the field of computer vision in recent years, most current research focuses on low-level visual tasks such as scene reconstruction, while work on higher-level semantic understanding remains relatively scarce. In particular, accurately estimating human pose in complex environments, where individuals are occluded or out of direct sight, is not only of great significance for applications such as security monitoring, counter-terrorism reconnaissance, rescue operations, and remote sensing, but also provides critical support for cutting-edge technologies such as autonomous driving and robotic vision navigation. Therefore, research on human pose sensing under NLOS conditions holds broad application potential and significant academic value. Isogawa et al. [23] employed reinforcement learning and the inverse point spread function (PSF) to achieve the first end-to-end method for 3D human pose sensing from optical NLOS transients. Liu et al. [24] used 3D networks combined with the LCT method to transform transients into 3D features, enabling single-person joint pose sensing. These methods typically assume that input transients are limited to single-person scenarios, but in complex hidden scenes in the real world, there are often multiple people involved, which greatly limits the applicability of these methods in practical environments.

To address the challenge of estimating multiple individuals in hidden scenes, we propose a unified and efficient method for NLOS multi-person 3D human pose sensing. Our approach adaptively processes diverse multi-person transients, achieving accurate 3D mesh regression for multiple individuals. We developed an automated pipeline that constructs multi-person 3D models along with their corresponding NLOS transient data. Using this pipeline, we created the first NLOS multi-person simulation dataset, which includes Skinned Multi-Person Linear Model

Optics EXPRESS

(SMPL [25]) parameters, voxels, and transients. By leveraging these simulated data, our method reconstructs 3D features from the transients and uses these features to predict a 3D body center heatmap and SMPL mesh parameter map, enabling robust multi-person pose sensing in complex NLOS scenarios. To validate its effectiveness, we deployed a self-built NLOS system, as shown in Fig. 2, to capture real-world data. Extensive experiments on both synthetic and real-world datasets demonstrate that our method effectively senses 3D human poses for multiple individuals in challenging NLOS environments. The simulation pipeline we developed and the real-world dataset captured by our system provide valuable resources for advancing research in this field.

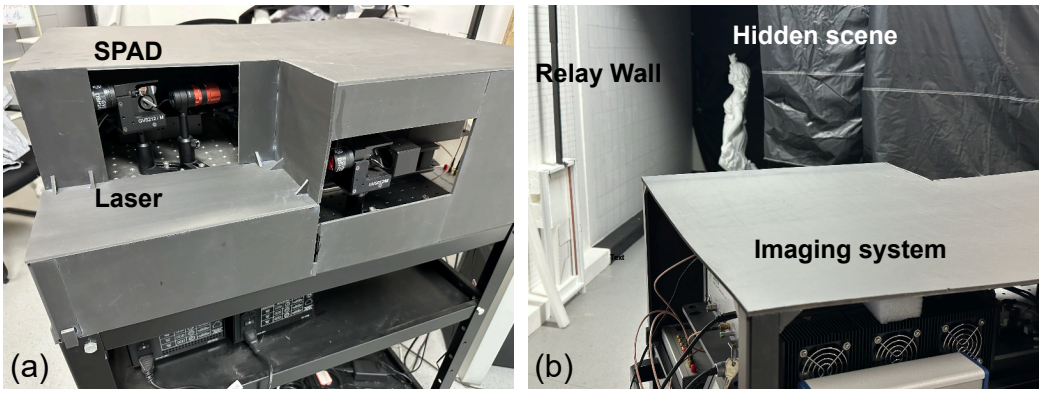


Fig. 2. (a) Our self-built NLOS imaging system, designed for capturing transient data in non-line-of-sight scenarios. (b) A schematic representation of the real-world setup.

2. Preliminary

2.1. Background: confocal NLOS imaging

Confocal NLOS imaging is a technique that utilizes a raster scanning process to collect transient data [10]. In the imaging system, both illumination and detection occur at the same point (x', y') on the reflective surface. When a pulse from a laser hits this point, the light scatters into occluded areas of the environment and later reflects back to the visible surface. A SPAD sensor records the time-resolved light intensity at this same location, which is captured as a histogram of photon arrival times. This process is repeated across a uniform grid of points on the surface, resulting in a 3D transient image $\tau(x', y', t)$, where each point on the grid has an associated temporal profile of the returning photons. The transient is modeled as:

$$\tau(x', y', t) = \int \frac{1}{r^4} \delta\left(t - \frac{r}{c}\right) \rho(x, y, z)\, \mathrm{d}V, \tag{1}$$

where $\rho(x, y, z)$ represents the albedo value at point (x, y, z) in the 3D hidden scene, r is the distance between the scene point (x, y, z) and the scanning point $(x', y', 0)$, c is the speed of light, and $\delta(\cdot)$ is the Dirac delta function.

The LCT [10] is based on a confocal NLOS imaging system, which can represent the NLOS imaging model as a translation invariant 3D convolution in the transformation domain, significantly improving computational efficiency, simplifying complex inverse problems, and achieving high-quality 3D reconstruction of hidden scenes.

2.2. Real-world data acquisition with self-built NLOS system

To capture real-world transients of human poses, we constructed a confocal NLOS imaging setup, as shown in Fig. 2, based on the forward model in Eq. 1. Using a pulsed laser and a single-pixel

Optics EXPRESS

SPAD via a beam splitter, we ensure that the illumination and detection points coincide on the relay wall, enabling the capture of multi-person NLOS data to verify the effectiveness of our method in real-world scenarios. Our system employs a 532nm pulsed laser with 750mW output power, providing optimal illumination for capturing high-quality transient measurements. The system uses a PDM photon counting detector module with high photon detection efficiency and superior timing resolution. A 2D galvanometer guides the laser light, scanning the relay wall, while the SPAD records returning photons via a TCSPC module to generate transient images.

We capture transients with real people wearing reflective clothing similar to those used in F-K [12] for our test set. Each hidden scene is located 0.8m to 1m away from the relay wall. To validate the method under more realistic conditions, we also captured data with subjects wearing ordinary clothing materials at distances of 1.2m to 1.4m away from the relay wall. The imaging system scans 128×128 points in the field of view (FOV) of 1.85 m $\times$ 1.85 m on the relay wall, with an acquisition time of 0.01s per scanning point and a bin width of 32 ps. The total scanning time for the entire scene is approximately 2.5 minutes.

3. Methods

3.1. Overview

To handle transient inputs in diverse multi-person scenarios, we propose a framework named AMPE-NLOS, as illustrated in Fig. 3. Given transient images as input, we first extract and propagate features through a collection of physical models using the LCT method, producing coarse 3D features F_l . These features are then passed through a 3D U-Net network with voxel supervision to output refined 3D features F_r . Based on both the refined 3D features F_r and the coarse 3D features F_l, we generate the 3D body center heatmap C_m and SMPL parameter map S_m . The 3D body center heatmap C_m predicts the probability of each location in 3D space being a human body center, while the SMPL parameter map S_m predicts the SMPL parameters for a person centered at each spatial position.

Specifically, the input transient data is first downsampled from 128×128×512 to 128×128×128 as the input to the network. Following the settings in [24], feature extraction is performed using a 5-layer CNN with residual blocks to extract spatial-temporal features from the transient data. The generated 3D features and the transient input both have the same size of $128 \times 128 \times 128$. After the LCT transformation, a 7-layer 3D U-Net network is used to refine the $128 \times 128 \times 128$ voxel features F_r. Based on the refined features F_r, another 3D U-Net network is used to estimate the 3D body center heatmap $C_m \in \mathbb{R}^{1\times64\times64\times64}$. The fused feature map obtained by adding coarse features F_l and refined features F_r is used as input to output the predicted SMPL mesh parameter map $S_m \in \mathbb{R}^{79\times64\times64\times64}$ through a 3D CNN network with ResNet-50 [26] as the backbone.

3.2. Simulation pipeline for multi-person NLOS dataset

To train our AMPE-NLOS network, we require paired multi-person scenes and corresponding transients under NLOS conditions. However, existing NLOS datasets are scarce and insufficient for this task. Therefore, we utilized our own pipeline, as shown in Fig. 4, to construct a multi-person NLOS dataset. Our dataset includes 20K multi-person SMPL 3D models, their corresponding SMPL parameters, voxel data, rendered transients, and ground truth 3D body center maps.

Multi-person pose collection: VPoser [27] is a variational autoencoder-based model designed to represent human pose in a low-dimensional latent space. It efficiently captures and samples natural human pose variations by encoding complex pose configurations into a compact latent space, making it suitable for generative tasks and human mesh recovery. We first randomly sample latent variables from VPoser and pass them through the decoder to generate 20K single-person

Optics EXPRESS

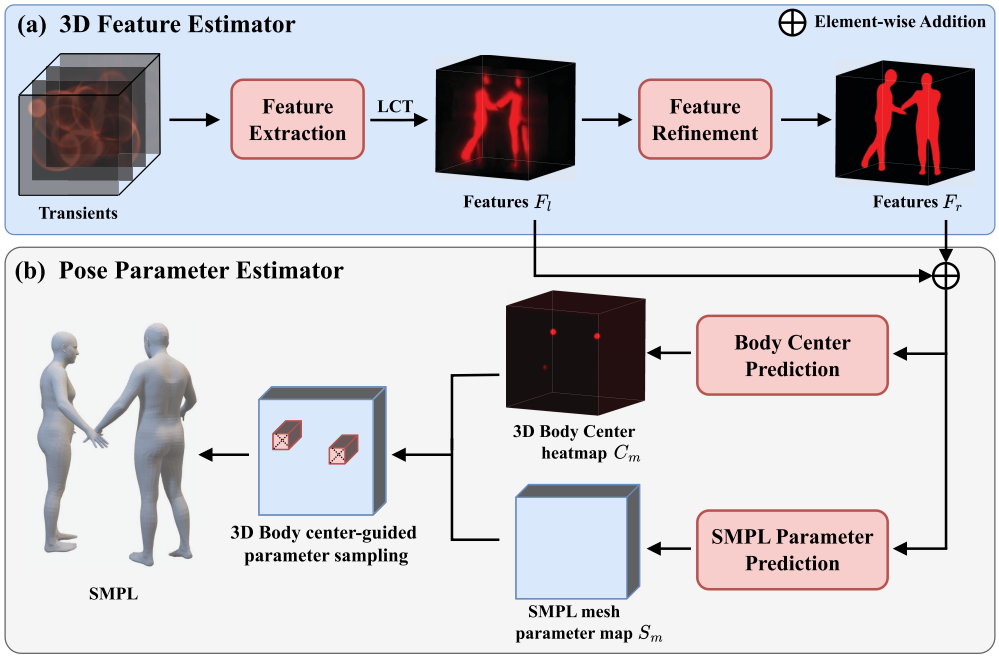


Fig. 3. AMPE-NLOS Framework Overview. Two-stage pipeline for adaptive multi-person NLOS pose estimation: (a) LCT+U-Net feature refinement; (b) SMPL + heatmap regression. Final poses regressed via body center-guided sampling — supports scenes with varying numbers of people.

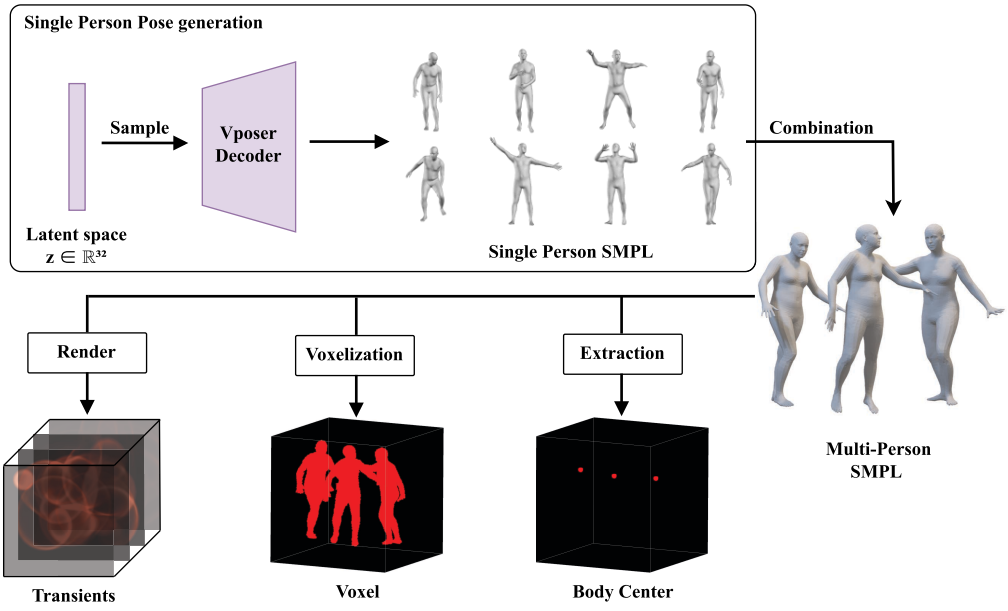


Fig. 4. NLOS Dataset Synthesis Pipeline — Combines VPoser pose sampling, multi-person composition, and noise-aware transient rendering to generate pose–voxel–measurement triplets for training NLOS pose estimators.

Optics EXPRESS

poses. These single-person SMPL models are then combined into two-person and three-person scenes through random translations and rotations around the z-axis.

To enhance the realism of the simulated data and the interactivity of multi-person scenes, we also introduce the InterHuman dataset [28] . This dataset is a comprehensive, large-scale collection of 3D human interaction motions, featuring various 3D movements between two interacting individuals. It includes interaction motions broadly classified into two categories: daily motions, such as passing objects, greeting, and communicating, and professional motions, like Taekwondo, Latin dance, and boxing. We extracted the SMPL parameters from three frames of each video and simultaneously constructed an NLOS version of this dataset.

Using the generated multi-person 3D SMPL mesh models, we employed the voxelization program provided by Binvox [29] to achieve a binary voxel representation of the 3D models. Using the representations introduced in Sec.3.3, we also constructed the ground truth for the 3D body center heatmap.

Transient simulation: Embed [21] implemented a CUDA-accelerated pipeline for rasterized rendering of transients. Using this approach, we simulated the transients for the SMPL 3D model. The transients have a spatial-temporal resolution of $128 \times 128 \times 512$ with a bin width of 33 ps.

However, in real NLOS imaging systems, due to the influence of the Single-photon avalanche diode (SPAD) sensor, there are various factors such as photon detection efficiency, time jitter, noise, dark counts, and afterpulsing that can affect the measurements. To reduce the gap between simulated and real data, we followed the SPAD simulation model presented in [30] to post-process the simulated transient data, simulating realistic human transients with noise. This model accounts for most of the SPAD's output response effects, including detection efficiency, time jitter, avalanche quenching, and both internal and external noise sources, providing a reliable computational sensor model.

3.3. Basic representations

We introduce the representation of the 3D body center heatmap and SMPL [25] mesh parameter map in this section. Each output map is of size $n \times D \times H \times W$, where n is the number of channels and $D = H = W = 64$.

3D body center heatmap: $C_m \in \mathbb{R}^{1\times D\times H\times W}$ is a heatmap representing the 3D human body center in space. We first determine the root node's spatial position using the translation parameter from the SMPL model and normalize it into a $D \times H \times W$ 3D spatial domain. Each human body center is modeled as a Gaussian distribution in the 3D body center heatmap. A 3D Gaussian kernel is applied to smooth the representation. The convolution of the body center space C with the Gaussian kernel G is expressed as:

$$C_{\mathrm{m}} = C * G, \tag{2}$$

where G is the 3D Gaussian kernel with a size of $4 \times 4 \times 4$ and a standard deviation of $\sigma = 1.0$, and $*$ denotes the convolution operation.

SMPL mesh parameter map: $S_m \in \mathbb{R}^{m\times D\times H\times W}$ includes the 79-dimensional parameters of the SMPL model, which describe the 3D human body's pose, shape, translation, and root orientation. We estimate a set of SMPL mesh parameters for each spatial position as the human body center, which collectively forms the SMPL map. The SMPL model defines an efficient unified parametric representation of the human body, mapping the pose θ and shape β parameters to a human 3D body mesh $M \in \mathbb{R}^{6890\times 3}$. The shape parameter $\beta \in \mathbb{R}^{10}$ is the top 10 principal components (PCA) coefficients of the SMPL shape space, representing the most significant variations in human body shape. The pose parameters $\theta \in \mathbb{R}^{3\times 22}$ define the 22 body joints, excluding the two hand joints, and describe the rotation of each joint relative to its parent using an axis-angle representation. The 3D rotation of the first joint denotes the body's root orientation

Optics EXPRESS

in the space. The translation parameter $T \in \mathbb{R}^{3\times1}$ represents the spatial displacement relative to the origin.

3.4. 3D body center-guided parameter sampling

In order to achieve adaptive person detection and human mesh recovery, we need to parse the 3D spatial coordinates $c \in \mathbb{R}^{n\times3}$ of multiple body centers from the 3D body center heatmap C_m, where n represents the number of detected individuals that meet the criteria. Based on the 3D spatial coordinates $c \in \mathbb{R}^{n\times3}$, we sample the SMPL parameters from the SMPL mesh parameter map S_m.

The 3D body center heatmap C_m represents the probability of each spatial location being the human body center. For the output heatmap, we first apply non-maximum suppression to extract the local maximum at different spatial positions. Through the max pooling operation, we determine c as the spatial coordinate of the local maximum. Based on the probability values of c being the body center in the heatmap, we sort them in descending order. By setting a confidence threshold t_c, the top n positions are selected as the final body centers. During the inference process, adaptive person detection is achieved based on the final selected c. Using the spatial positions of c, the corresponding SMPL parameters are sampled from the SMPL mesh parameter map S_m. During training, we sort the predicted c based on their spatial positions to match them with the ground truth body centers, enabling the calculation of the loss.

3.5. Loss functions

In this section, we introduce the configuration of the loss functions used throughout the network training process.

Voxel loss: Following [24], We use voxel loss to train the 3D U-Net by minimizing the difference between the ground truth voxels and the predicted refined features F_r . The voxel loss combines the binary cross-entropy loss and the dice loss [31] .

$$\mathcal{L}_{\text{voxel}} = \frac{1}{N}\sum_{n}\left(\sigma_n \log\hat{\sigma}_n + (1-\sigma_n)\log(1-\hat{\sigma}_n)\right) + \left(1 - \frac{2}{N}\sum_{n}\frac{\sigma_n\hat{\sigma}_n}{\sigma_n+\hat{\sigma}_n}\right), \tag{3}$$

where $\hat{\sigma}_n$ and σ_n represent the 3D features F_r predicted and the ground-truth volume, respectively, while N is the number of voxels.

3D body center loss: $\mathcal{L}_c$ is designed to ensure that the 3D human body center positions c have a high confidence probability, while in most other non-center spatial positions, the confidence in the heatmap C_m remains low. To address the significant imbalance and difference in the number of samples between the center and non-center points of the 3D body in C_m, we follow the focal loss in [32], 3D body center loss is defined as:

$$\mathcal{L}_c = -\frac{1}{\sum I_{\text{pos}}}\left(\sum_{\text{pos}}\log(C_m^p)(1-C_m^p)^2 I_{\text{pos}} + \sum_{\text{neg}}\log(1-C_m^p)(C_m^p)^2(1-C_m^{gt})^4 I_{\text{neg}}\right), \tag{4}$$

where C_m^p and C_m^{gt} represent the predicted and ground-truth 3D body center heatmaps. I_{pos} is a binary matrix with positive values at the 3D body center locations, defined as $I_{\text{pos}} = 1$ when C_m^{gt}>0.9, and I_{neg} is an indicator function for the negative samples, defined as $I_{\text{neg}} = 1$ when $C_m^{gt} \leq 0.9$.

Mesh parameter loss: As we introduced in Sec.3.4, we implemented the extraction and matching of multi-person SMPL parameters. The mesh parameter loss $\mathcal{L}_{\text{mesh}}$ is defined as:

$$L_m = w_{\text{pose}}L_{\text{pose}} + w_{\text{shape}}L_{\text{shape}} + w_{\text{root}}L_{\text{root}} + w_{\text{trans}}L_{\text{trans}} + w_{\text{j3d}}L_{\text{j3d}}, \tag{5}$$

where L_{pose} is the L2 loss computed by converting the pose parameters $\theta \in \mathbb{R}^{3\times21}$ into rotation matrices using the Rodrigues transformation, and calculating the pose parameter loss at the level

Optics EXPRESS

of the 3D rotation matrices. L_{shape} is the L2 loss of the shape parameters $\beta \in \mathbb{R}^{10}$. L_{root} is the L2 loss of the root orientation rotation matric, representing the overall orientation of the human body. L_{trans} is the L2 loss of the translation parameters. L_{j3d} is the L2 loss of the 3D joints J regressed from the body mesh M. $w(\cdot)$ denotes the corresponding loss weights.

4. Experimental results

4.1. Experimental settings

4.1.1. Implementation details

We divided our constructed simulation dataset into training, validation, and test sets in an 8:1:1 ratio, and trained our network accordingly. We employed the Adam optimizer with an initial learning rate of 0.001, applying a learning rate decay factor of 0.1 starting at the second epoch. We adopted a phased training strategy. In the first phase, we initially trained the feature extraction network, refined voxel features, and the 3D body center heatmap network. Once the 3D features and body centers were accurately predicting the number of people, we proceeded to the second phase. In this phase, we fixed the weights of the earlier networks and solely trained the SMPL mesh parameter sensing network. We trained the network for two days on two NVIDIA RTX 3090 GPUs with a batch size of 2.

4.1.2. Computational complexity and inference time analysis

The proposed network comprises several key components: a 5-layer CNN for feature extraction, an LCT module for spatiotemporal propagation, a multi-layer 3D U-Net for feature refinement and body center prediction, and a 3D CNN with a ResNet-50 backbone for SMPL mesh parameter regression. With a total of 88.43 million parameters, the network design incorporates input downsampling in the initial stage, a lightweight network architecture to reduce computational overhead, and parallel processing for body center prediction and SMPL parameter estimation. As a result, the model achieves an average inference time of 0.1 seconds per sample on an NVIDIA RTX 3090 GPU, making it well-suited for real-time imaging applications.

4.1.3. Comparison methods

Since we are the first to perform multi-person NLOS pose sensing, there is a lack of related work for comparison. Therefore, based on the single-person NLOS pose sensing work HiddenPose [24] and the reconstruction-based Embed [21], we implemented multi-person pose sensing versions to conduct comparative experiments. The output of HiddenPose consists of the 3D spatial coordinates of 24 joints for a single person. It is a regression probability model based on neural networks, and regardless of the number of people in the input transient scene, it can only output the spatial positions of 24 joints, making it suitable only for single-person scenarios. We manually modified the network structure of HiddenPose for each experiment and retrained different models for single-person, two-person, and three-person scenes to conduct comparative experiments. For Embed [21], we discard its reconstruction objective and reuse only its backbone architecture (e.g., feature extractor), training it end-to-end for pose regression under the same multi-person protocol. In contrast, our AMPE-NLOS network is a unified framework that requires only a single training process on mixed multi-person scenes and can adaptively handle transient scenes with varying numbers of people. For fairness and ease of comparison, we trained the SMPL version of the model using the same dataset.

4.1.4. Evaluation metrics

We adopted the main evaluation metrics commonly used in the human pose sensing field: mean per joint position error (MPJPE), which calculates the average Euclidean distance between the

Optics EXPRESS

predicted joint coordinates and the ground truth coordinates. MPJPE is defined as:

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N}\|J_i - J_i^*\|_2, \tag{6}$$

where N represents the number of joints, J_i represents the predicted joint coordinates, and J_i^* represents the ground truth joint coordinates. PA-MPJPE calculates the MPJPE with the predicted joints rigidly aligned to the ground truth joints. This alignment removes the errors caused by translation, rotation, and scale differences.

Per-vertex error (PVE), which evaluates the 3D surface error by calculating the mean Euclidean distance between the predicted and ground truth mesh vertices. PVE is defined as:

$$PVE = \frac{1}{M}\sum_{i=1}^{M}\|V_i - V_i^*\|_2, \tag{7}$$

where M represents the number of mesh vertices, V_i represents the predicted vertex coordinates, and V_i^* represents the ground truth vertex coordinates.

Mean per joint angle error (MPJAE), which computes the mean error in joint angles, evaluates the accuracy of predicted joint rotations by calculating the mean angular distance between the predicted and ground truth joint angles. MPJAE is defined as:

$$\text{MPJAE} = \frac{1}{N}\sum_{i=1}^{N}|\Theta_i - \Theta_i^*|, \tag{8}$$

where N represents the number of joints, Θ_i represents the predicted joint angles, and Θ_i^* represents the ground truth joint angles.

4.2. Results on simulated data

To validate the effectiveness of our method, we conducted comparative experiments on the two simulation datasets we constructed. Figure 5 shows the experimental results on our dataset generated from InterHuman. Features F_r represent the refined features predicted by our network in the voxel domain. We present five different results of two-person pose sensing. As shown in the results, the SMPL models generated by our method are highly consistent with the ground truth. Compared to the implemented version of Embed [21] and HiddenPose [24], our method demonstrates significant advantages. By utilizing the 3D body center-guided parameter sampling method, we achieved effective pose sensing in multi-person NLOS scenarios. Even in challenging cases such as Scene 5, where significant front-back occlusions exist, our approach consistently delivers accurate pose sensing. In contrast, both Embed [21] and HiddenPose [24], which directly regress poses, fail to effectively handle multi-person tasks. The poses estimated by them tend to converge to a unified template, lacking the ability to distinguish between different individuals.

Table 1. Quantitative assessment for our method compared to Embed [21] and HiddenPose [24] on our dataset generated from InterHuman

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJAE ↓
Embed [21]	152.25	121.99	193.77	33.16
HiddenPose [24]	132.65	109.48	166.55	30.85
Ours	**43.40**	**35.50**	**54.99**	**12.35**

Table 1 shows the quantitative results of four evaluation metrics, which are calculated using the SMPL models predicted from the testing split of the Inter-NLOS dataset. The four evaluation

Optics EXPRESS

Fig. 5. Qualitative comparisons to Embed [21] and HiddenPose [24] on our dataset generated from InterHuman.

metrics of our method are significantly better than those of Embed [21] and HiddenPose [24], indicating improved accuracy in pose sensing and demonstrating the superiority of our approach.

Figure 6 shows the experimental results on our constructed dataset generated from VPoser. We evaluate our method on scenarios involving one, two, and three persons. Notably, our method employs a unified network architecture and achieves these results by training on a mixed dataset comprising all three scenarios. In contrast, Embed [21] and HiddenPose [24] utilize separate network architectures specifically designed for each scenario, and their results are obtained by training independently on datasets corresponding to each number of persons. As shown in Fig. 6, Embed [21] and HiddenPose [24] perform poorly in multi-person scenarios when directly regressing SMPL models. In contrast, our method uses the 3D body center-guided parameter sampling method to adapt effectively to multi-person scenarios. By leveraging the semantic information of body centers, our approach distinguishes the SMPL parameters and positions of different individuals with high accuracy, achieving adaptive non-line-of-sight multi-person pose sensing even under mixed training conditions.

Table 2 shows the quantitative results of four evaluation metrics, our method uses a unified network framework to sense multi-person poses in NLOS scenarios and achieves clear advantages.

4.3. Results on real data

Figure 7 shows the experimental results on real data captured using our self-built NLOS system. Our method can process transient data from scenarios with varying numbers of people using a unified network architecture and successfully accomplish the task of multi-person pose sensing. Regarding the orientation of the individuals in the second, third, and fifth columns, we made

Optics EXPRESS

Fig. 6. Qualitative comparisons to Embed [21] and HiddenPose [24] on our constructed dataset generated from VPoser.

Table 2. Quantitative assessment for our method compared to Embed [21] and HiddenPose [24] on our constructed dataset generated from VPoser

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJAE ↓
Embed [21]	123.77	85.23	164.46	28.54
HiddenPose [24]	108.40	75.37	144.75	27.74
Ours	**61.44**	**55.65**	**74.10**	**24.06**

the design choice to ensure that the human head effectively reflects light. As such, we consider the side of the hat as the front of the individual. This setup has led to an apparent reversal of orientation in the output, which might seem unusual from a visual perspective. Additionally, the distinction between a purely front-facing and purely back-facing position is relatively minimal in transient data due to the inherent properties of NLOS imaging. The reflected light from these positions is often quite similar, which can cause difficulties in differentiating the exact body orientation. The quality of the captured real data is influenced by the noise and time resolution limitations of the NLOS imaging system, resulting in a noticeable gap between real-world and simulated data. Although this issue can be present, it is important to note that it is a natural consequence of the limitations of the system's spatial and temporal resolution, and it does not significantly hinder the overall performance of the method.

To further validate generalization under practical conditions, we conducted additional experiments with subjects wearing ordinary clothing at distances of 1.2m to 1.4m away from the relay wall. Results from this realistic setup, shown in Fig. 8, confirm that our method remains

Optics EXPRESS

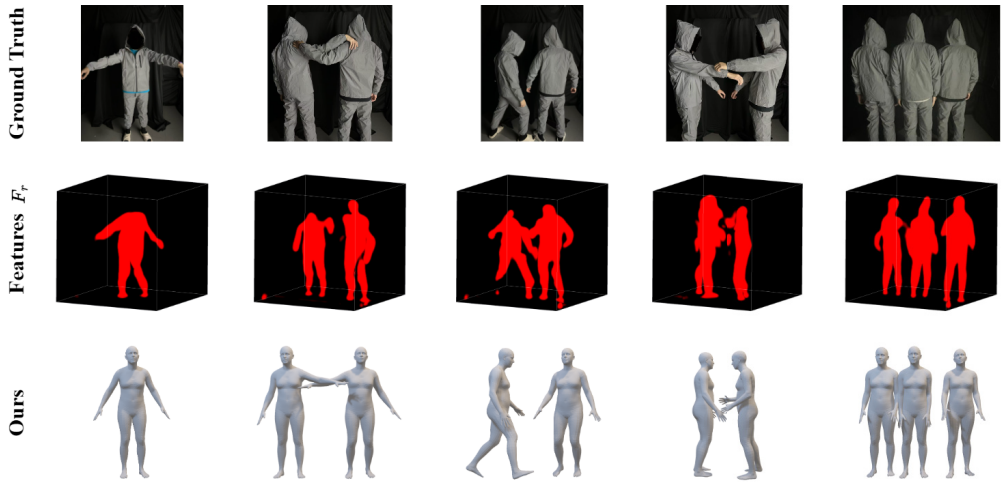


Fig. 7. Qualitative results of multi-person 3D pose estimation on real data captured with subjects wearing reflective clothing.

functional with ordinary clothing. The inherently weaker signal from low-reflectivity clothing, combined with the system's temporal resolution constraints, results in noisier transient data. This noise propagates into the subsequent feature reconstruction, leading to blurred and incomplete structural information, particularly in peripheral regions like the arms and extremities. As our pose estimation network is highly dependent on the quality of these input features, the final accuracy in estimating fine-grained poses, especially arm orientations, is consequently diminished under these challenging conditions. This also represents one of the limitations of our method.

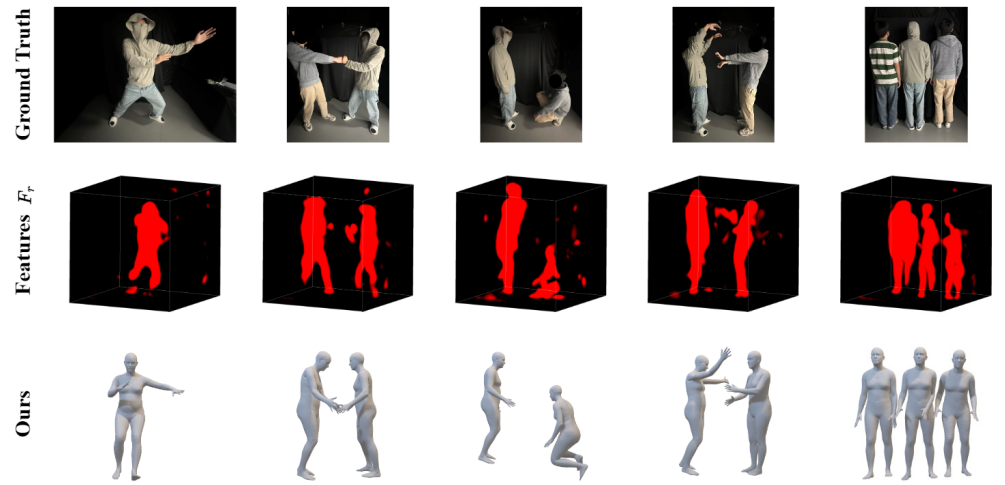


Fig. 8. Qualitative results of multi-person 3D pose estimation on real data captured with subjects wearing ordinary clothing.

4.4. Ablation study

To validate the contribution of each module in our proposed framework, we conduct systematic ablation studies focusing on the feature extraction and refinement components. The complete

Optics EXPRESS

network architecture consists of four key stages: feature extraction, feature refinement, body center prediction, and SMPL parameter estimation. The latter two components constitute the core function of implementing adaptive multi person pose perception and are indispensable, so we conducted ablation experiments on the first two modules. We implement the feature extraction and refinement modules in different combinations, and show the quantitative results of MPJPE, PA-MPJPE, PVE, and MPJAE in Table 3. The results clearly indicate that bypassing the feature processing stages and attempting to directly estimate poses from transient data is not feasible, as this approach fails to provide the necessary semantic information for effective body center detection and multi-person counting. The feature extraction and refinement modules play a critical role in reconstructing meaningful representations, with the complete configuration achieving the best performance across all metrics.

Table 3. Ablation study for the feature extraction and refinement modules on our constructed dataset generated from VPoser

Feature Extraction	Feature Refinement	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJAE ↓
×	×	219.76	123.71	283.97	39.09
✓	×	81.78	73.17	101.17	26.91
×	✓	73.10	66.95	88.17	26.10
✓	✓	**61.44**	**55.65**	**74.10**	**24.06**

5. Conclusion

This work presents the first unified and efficient method for NLOS multi-person 3D human pose sensing, overcoming the limitations of previous works confined to single-person scenarios. Our approach adaptively handles diverse multi-person NLOS conditions through a novel framework that reconstructs coarse 3D features from transients, refines them via neural networks, and leverages 3D body center-guided parameter sampling for adaptive mesh regression.The method adapts to diverse multi-person NLOS scenarios and includes the first simulated dataset for multi-person pose sensing in NLOS conditions. Experimental results on both simulated and real-world data demonstrate the effectiveness of our approach. Notably, the simulation pipeline we developed for constructing the multi-person NLOS dataset and the real-world data captured by our system provide valuable resources for advancing research in this field.

Although our method achieves robust performance, current results exhibit minor inaccuracies in arm pose estimation, particularly in challenging cases with significant occlusions or complex limb configurations. These subtle imperfections highlight the inherent difficulties of inferring poses from transient NLOS data. Future research is needed to reduce the cost of capturing real data with NLOS imaging systems and bridge the gap between real-world and simulated data. Ultimately, we believe this work lays the foundation for broader applications of NLOS imaging in real-world multi-person environments and inspires further advancements in this emerging field.

Funding. National Natural Science Foundation of China (62231018. 62171317); Fundamental Research Funds for the Central Universities (WK2100000059).

Disclosures. The authors declare no conflicts of interest.

Data availability. The data that support the findings of this study are openly available in github repository [33].

References

1. A. Kanazawa, M. J. Black, D. W. Jacobs, *et al.*, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp.7122–7131.
2. Y. Sun, Q. Bao, W. Liu, *et al.*, "Monocular, one-stage, regression of multiple 3d people," in *Proceedings of the IEEE/CVF international conference on computer vision*, (2021), pp.11179–11188.
3. M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2020), pp.5253–5263.

Optics EXPRESS

4. N. Kolotouros, G. Pavlakos, M. J. Black, *et al.*, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE/CVF international conference on computer vision*, (2019), pp.2252–2261.
5. J. Wang, S. Tan, X. Zhen, *et al.*, "Deep 3d human pose estimation: A review," Computer Vision and Image Understanding **210**, 103225 (2021).
6. C. Zheng, W. Wu, C. Chen, *et al.*, "Deep learning-based human pose estimation: A survey," ACM Comput. Surv. **56**(1), 1–37 (2024).
7. H. Ge, Q. Feng, H. Jia, *et al.*, "Lpsnet: End-to-end human pose and shape estimation with lensless imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), pp.1471–1480.
8. D. Faccio, A. Velten, and G. Wetzstein, "Non-line-of-sight imaging," Nat. Rev. Phys. **2**(6), 318–327 (2020).
9. R. Geng, Y. Hu, Y. Chen, *et al.*, "Recent advances on non-line-of-sight imaging: Conventional physical models, deep learning, and new scenes," APSIPA Transactions on Signal and Information Processing 11 (2021).
10. M. O'Toole, D. B. Lindell, G. Wetzstein, *et al.*, "Confocal non-line-of-sight imaging based on the light-cone transform," Nature **555**(7696), 338–341 (2018).
11. X. Liu, I. Guillén, M. La Manna, *et al.*, "Non-line-of-sight imaging using phasor-field virtual wave optics," Nature **572**(7771), 620–623 (2019).
12. D. B. Lindell, G. Wetzstein, and M. O'Toole, "Wave-based non-line-of-sight imaging using fast fk migration," ACM Trans. Graph. **38**(4), 1–13 (2019).
13. S. Xin, S. Nousias, K. N. Kutulakos, *et al.*, "A theory of fermat paths for non-line-of-sight shape reconstruction," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), pp.6800–6809.
14. S. Shen, Z. Wang, P. Liu, *et al.*, "Non-line-of-sight imaging via neural transient fields," IEEE Trans. Pattern Anal. Mach. Intell. **43**(7), 2257–2268 (2021).
15. F. Mu, S. Mo, J. Peng, *et al.*, "Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging," IEEE Trans. on Pattern Anal. Mach. Intell. (2022).
16. Y. Li, J. Peng, J. Ye, *et al.*, "Nlost: Non-line-of-sight imaging with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), pp.13313–13322.
17. M. La Manna, F. Kine, E. Breitbach, *et al.*, "Error backprojection algorithms for non-line-of-sight imaging," IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1615–1626 (2019).
18. B. Ahn, A. Dave, A. Veeraraghavan, *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), pp.7889–7899.
19. X. Liu, J. Wang, Z. Li, *et al.*, "Non-line-of-sight reconstruction with signal–object collaborative regularization," Light:Sci. Appl. **10**(1), 198 (2021).
20. J. Grau Chopite, M. B. Hullin, M. Wand, *et al.*, "Deep non-line-of-sight reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp.960–969.
21. W. Chen, F. Wei, K. N. Kutulakos, *et al.*, "Learned feature embeddings for non-line-of-sight imaging and recognition," ACM Trans. Graph. **39**(6), 1–18 (2020).
22. Y. Yu, S. Shen, Z. Wang, *et al.*, "Enhancing non-line-of-sight imaging via learnable inverse kernel and attention mechanisms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), pp.10563–10573.
23. M. Isogawa, Y. Yuan, M. O'Toole, *et al.*, "Optical non-line-of-sight physics-based 3d human pose estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp.7013–7022.
24. P. Liu, Y. Yu, Z. Pan, *et al.*, "Hiddenpose: Non-line-of-sight 3d human pose estimation," in *2022 IEEE International Conference on Computational Photography (ICCP)*, (IEEE, 2022), pp.1–12.
25. M. Loper, N. Mahmood, J. Romero, *et al.*, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, (2023), pp.851–866.
26. K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp.770–778.
27. G. Pavlakos, V. Choutas, N. Ghorbani, *et al.*, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2019), pp.10975–10985.
28. H. Liang, W. Zhang, W. Li, *et al.*, "Intergen: Diffusion-based multi-human motion generation under complex interactions," International Journal of Computer Vision pp. 1–21 (2024).
29. P. Min, "binvox," (2019). Accessed: 2022-01-05.
30. Q. Hernandez, D. Gutierrez, and A. Jarabo, "A computational model of a single-photon avalanche diode sensor for transient imaging," arXiv (2017).
31. S. A. Taghanaki, Y. Zheng, S. K. Zhou, *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," Computerized Medical Imaging and Graphics **75**, 24–33 (2019).
32. T.-Y. Lin, P. Goyal, R. Girshick, *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp.2980–2988.
33. Y. Hou, X. Cui, S. Sun, *et al.*, AMPE-NLOS Dataset, Github, 2024 [Accessed date: Sept. 2024] https://github.com/Syh-20/AMPE-NLOS.