Unsupervised Physics-Guided Deep Learning for Sparse-Data Compton Imaging*

Zhengtao Long[1] and Xiaofei Jiang[1,†]

[1]*College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China*[‡]

In response to the instability of Compton-camera reconstruction under ultra-low-count conditions, the substantial loss of usable information caused by conventional strict event screening, and the widespread reliance of existing deep-learning methods on image-label supervision, this study proposes SACN, a physics-constrained reconstruction framework for sparse Compton imaging. This method takes the set of encodable raw events as a unified input and achieves a unified representation of different interaction patterns through event-type identifiers and missing masks. During training, a differentiable Compton forward model is embedded into the optimization objective. Physical likelihood constraints are imposed on events for which Compton geometric consistency can be defined. Combined with geometric pre-localization and a conditional neural implicit field, the method reconstructs a continuous three-dimensional source distribution without image labels. Using a Geant4 simulation platform for a dual-layer GAGG:Ce Compton camera, we compared two data-usage strategies, namely strictly screened-event reconstruction and all-event-utilization reconstruction, under a unified physical simulation setting in single-source, double-source, and multi-energy three-source scenarios. Results show that, under the current simulation setting, SACN provides higher localization accuracy, better artifact suppression, and higher reconstruction efficiency under low-count conditions. Compared with the reference methods, the localization error is reduced by 64.2%–85.1%, the contrast-to-noise ratio is improved by 23.4%–189.1%, and the reconstruction time in the high-statistics single-source scenario is shortened by 99.96% relative to MLEM. In addition, the model is trained only with single-energy Cs-137 data, yet still maintains relatively stable performance for unseen isotopes and complex multi-source scenarios. This indicates a certain degree of cross-energy transferability under the current detector model and simulation conditions. This study provides a potentially useful physics-constrained and image-label-free reconstruction route for low-dose radionuclide imaging and weak-source detection.

Keywords: Compton camera, Deep learning, Image reconstruction, Physics-guided, Sparse data

I. INTRODUCTION

As a new gamma-ray detection device based on the principle of electronic collimation, the Compton camera has shown great application potential in medical radionuclide tracing, nuclear safety monitoring, and astrophysical observation because it can achieve three-dimensional imaging over a wide energy range without mechanical collimation [1–4]. In recent years, studies on online imaging in boron neutron capture therapy (BNCT), radioactive contamination imaging, dual-modality Compton–PET imaging, and X-ray fluorescence imaging have further shown that the Compton camera is moving from proof-of-concept toward more complex and more application-oriented imaging scenarios [5–9]. Its working principle is based on Compton scattering. An incident gamma photon first undergoes Compton scattering in the scatter layer and is then photoelectrically absorbed in the absorber layer. According to the Compton scattering formula, the scattering angle can be determined from the deposited energy:

$$\cos\theta = 1 - m_e c^2 \left(\frac{1}{E_{\mathrm{abs}}} - \frac{1}{E_{\mathrm{sca}} + E_{\mathrm{abs}}} \right), \qquad (1)$$

where $m_e c^2$ is the electron rest mass energy (511 keV), and E_{sca} and E_{abs} are the energy depositions in the scatter layer and the absorber layer, respectively. From a geometric point of view, a single valid event constrains the photon origin to a cone surface. Image reconstruction is essentially an ill-posed inverse problem that infers the source distribution from the spatial intersections of a finite number of cone surfaces [1, 2, 10].

Despite its clear advantages, the practical application of Compton imaging is still severely limited by reconstruction performance under low-count conditions. In particular, in low-dose radionuclide imaging, long-range weak-source detection, and nuclear emergency response, the number of valid Compton events available for reconstruction is usually very limited [1–4]. For such imaging tasks, conventional analytical methods such as simple back-projection (SBP) are computationally efficient, but they often produce obvious star-shaped artifacts. Statistical iterative methods such as maximum-likelihood expectation-maximization (MLEM) can improve image quality by exploiting measurement statistics, but under extremely sparse conditions they are still prone to noise amplification, unstable convergence, and local minima [1, 10, 11]. As the number of events decreases further, the ill-posedness of the reconstruction problem becomes much stronger. Conventional methods then struggle to balance localization accuracy, artifact suppression, and computational efficiency.

In recent years, deep-learning methods have been introduced into Compton-camera image reconstruction. These include image-supervised reconstruction enhancement, event-representation-based scattering information estimation, and hybrid reconstruction combined with physical models [12–19]. These studies suggest that data-driven methods may

* This work was supported by National Natural Science Foundation of China (No. 12205062) and Network Communication Signal Detection System (No. 1502195N).

† Corresponding author: 2188789329@qq.com

‡ 2291755892@qq.com

improve the reconstruction performance of conventional analytical and iterative methods under low-statistics conditions. However, current studies still have three main limitations. First, many methods still rely on strict event screening and use only valid events that satisfy the standard double-layer Compton criterion. Under low-count conditions, this procedure further compresses the already limited usable information. Second, existing studies mainly focus on single-source or relatively simple scenarios, and validation of robustness under multi-source, mixed-energy, and extremely sparse conditions remains insufficient. Third, many methods rely on image-level supervision. The model is then more likely to learn an input–output mapping coupled to the training-data distribution, rather than being directly constrained by the physics of Compton scattering. This may limit generalization under unseen energies, complex scenarios, and high-noise conditions.

It should be noted that the all-event-utilization strategy considered in this study is not proposed in isolation. Previous studies have preliminarily shown that, compared with the conventional strict-screening strategy, retaining raw encodable events as much as possible helps reduce information loss and improve reconstruction potential under sparse-data conditions [18, 19]. Building on those studies, the present work moves one step further. The reconstruction result is no longer restricted to a fixed-resolution three-dimensional image. Instead, it is represented as a continuous three-dimensional radioactive activity field conditioned on the event set. At the same time, geometric pre-localization and conditional neural implicit representation are introduced to unify event-level physical consistency constraints and continuous-space reconstruction within one optimization framework [20–25].

Based on the above considerations, this study reformulates Compton image reconstruction as an optimization problem centered on physical consistency. The basic idea is that the source distribution predicted by the network is not constrained by its pixel-wise difference from a labeled image. Instead, it is constrained by its ability to explain the observed events. For events that can define Compton-scattering geometric consistency, a physical likelihood constraint is constructed through a differentiable Compton forward model. For single-layer events, partial-energy-deposition events, and events with poor geometric quality but still encodable, event-type identifiers and missing masks are introduced into a unified input representation to reduce the information loss caused by conventional screening as much as possible. To alleviate gradient sparsity and the excessively large search space in continuous three-dimensional optimization under extremely sparse conditions, a geometric pre-localization mechanism is further introduced to provide coarse guidance for the potential source region. A conditional neural-field reconstruction network for low-count scenarios is then constructed to improve reconstruction stability and localization accuracy under multi-source and multi-energy conditions.

II. METHODS

A. Problem Formulation

The goal of Compton-camera image reconstruction is to recover the spatial distribution of the radiation source within the reconstruction region from a finite number of Compton scattering events. Different from the conventional practice of representing the reconstruction result as a discrete two-dimensional image, in this work the target to be solved is defined as a continuous three-dimensional radioactive activity field over the reconstruction region $\Omega \in \mathbb{R}^3$:

$$f_\theta(\mathbf{r} \mid \mathcal{E}), \qquad \mathbf{r} \in \Omega, \tag{2}$$

Here, $\mathcal{E} = \{e_i\}_{i=1}^{N}$ denotes the input event set, $\mathbf{r}$ denotes the spatial coordinate, and θ denotes the network parameters. Accordingly, the reconstruction task in this work is no longer to output an image with a fixed resolution, but to learn a continuous three-dimensional scalar field driven by the conditioning of the event set.

For the i-th raw event considered in this work, it is no longer limited to the "valid event" that satisfies the dual-layer Compton criterion in the traditional sense. Instead, all detection records that can be encoded in a unified format are collectively referred to as raw events, and are denoted as

$$e_i = (\widetilde{q}_i, m_i, t_i), \tag{3}$$

Here, $\widetilde{q}_i$ denotes the basic observation completed by placeholder filling, including the obtainable interaction positions, energy depositions, and their derived physical quantities; m_i is the missing-value mask, which is used to indicate which observed components truly exist and which are placeholder-filled; and t_i is the event-type identifier, which is used to distinguish standard dual-layer events, partially incomplete events, compressed multiple-scattering events, and other event types that can be uniformly encoded. Therefore, the input of this work is no longer restricted to dual-layer Compton events in a fixed form, but is defined as the unified raw-event set:

$$\mathcal{E}_{\mathrm{raw}} = \{e_i\}_{i=1}^{M}, \tag{4}$$

where M denotes the total number of all encodable raw events retained under the same physical scene.

On this basis, this work further distinguishes two types of event subsets. One is the event subset for which a Compton geometric-consistency constraint can be defined according to the position and energy information, denoted as

$$\mathcal{E}_{\mathrm{cone}} \subseteq \mathcal{E}_{\mathrm{raw}}, \tag{5}$$

For events in this subset, the scattering angle can be calculated from the scattering point, the absorption point, and

the energy depositions, and the corresponding cone geometric constraint can be established. The other is the event subset that cannot fully define a Compton cone but still contains useful structural information, denoted as

$$\mathcal{E}_{\mathrm{aux}} = \mathcal{E}_{\mathrm{raw}} \setminus \mathcal{E}_{\mathrm{cone}}, \tag{6}$$

This subset includes single-layer interaction events, partial-energy-deposition events, and events with poor geometric quality but that are still encodable. These events do not directly provide a complete cone constraint, but can participate in overall representation learning through the event type and the missing pattern.

For any $e_i \in \mathcal{E}_{\mathrm{cone}}$, the observations that can be used for geometric modeling can be written as

$$q_i = (\mathbf{p}_i^s, \mathbf{p}_i^a, E_i^s, E_i^a), \tag{7}$$

Here, $\mathbf{p}_i^s$ and $\mathbf{p}_i^a$ denote the interaction positions in the scatter layer and the absorber layer, respectively, and E_i^s and E_i^a denote the energy depositions in the two layers, respectively. According to the Compton scattering relation, the scattering angle can be determined from the measured energy:

$$\cos\theta_i = 1 - m_e c^2 \left(\frac{1}{E_i^a} - \frac{1}{E_i^s + E_i^a} \right), \tag{8}$$

Here, $m_e c^2 = 511$ keV is the energy corresponding to the electron rest mass. This relation constrains the possible incident direction of the photon to the vicinity of a conical surface with $\mathbf{p}_i^s$ as the cone apex, the scattering direction as the axis, and θ_i as the half-apex angle. The essence of Compton image reconstruction is to stably invert the source distribution under noisy and incomplete raw-event observations by jointly using the events that can form cone constraints and the events that provide only weak auxiliary information.

Based on the above definitions, the reconstruction task in this work can be formulated as follows: under the condition of a given raw-event set $\mathcal{E}_{\mathrm{raw}}$, learn a continuous three-dimensional radioactive activity field driven by the conditioning of the event set, such that it can both utilize the auxiliary information in all encodable raw events and maintain high consistency, at the physical-loss level, with respect to the geometrically interpretable event subset $\mathcal{E}_{\mathrm{cone}}$.

B. Compton Camera System and Data Acquisition

1. Detector Configuration

In this work, a dual-layer Compton camera system based on GAGG:Ce scintillation crystals is used for validation. The system structure is shown in Fig. 1. GAGG:Ce crystals are widely used in gamma-ray detection because of their high light yield (about $54{,}000$ photons/MeV), moderate density (6.63 g/cm^3), and fast decay time (about 90 ns).

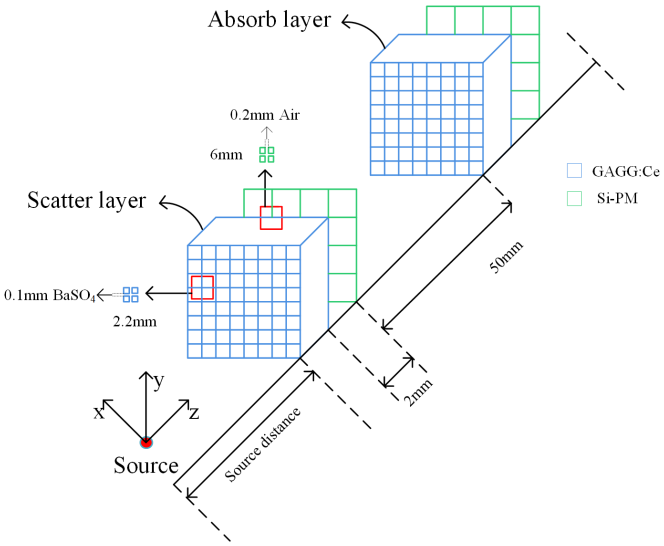

Fig. 1. System structure of the dual-layer Compton camera based on GAGG:Ce scintillation crystals.

Both the scatter layer and the absorber layer adopt a 23×23 crystal-array configuration, with a total of 529 crystal units. The crystal size of the scatter layer is $2.2 \times 2.2 \times 5$ mm^3. The thinner thickness is designed to maximize the Compton-scattering probability while minimizing multiple scattering. The crystal size of the absorber layer is increased to $2.2 \times 2.2 \times 10$ mm^3. The thicker design ensures efficient photoelectric absorption of the scattered gamma rays. The center-to-center distance between the two detector layers is set to 50 mm. This distance is optimized by Monte Carlo simulation to balance angular resolution and detection efficiency.

Adjacent crystals are separated by a 0.1 mm-thick $BaSO_4$ reflector layer to reduce optical crosstalk and improve position resolution. Photoelectric conversion is realized by an 8×8 silicon photomultiplier (SiPM) array. The effective area of a single SiPM is 6.0×6.0 mm^2, covering 9 crystal units.

2. Monte Carlo Simulation Platform

A high-fidelity Monte Carlo simulation platform is built based on Geant4 (version 11.1.1) to accurately simulate the detector geometry, material properties, and physical response process. The simulation physics processes adopt the G4EmStandardPhysics_option4 physics list, covering the major electromagnetic interaction processes such as Compton scattering, photoelectric absorption, Rayleigh scattering, and electron multiple scattering.

The detector-response modeling is calibrated based on actual experimental measurement data. The energy resolution is set to 5.4% FWHM, which is strictly set according to the actual experimental measurement results of our GAGG:Ce prototype system to ensure high fidelity. It should be explained here that, at the current simulation stage, the raw coordinates after energy-resolution broadening are directly

used, and the intrinsic interaction-depth (DOI) distortion effect of the 10 mm-thick GAGG:Ce crystal is not explicitly introduced for the time being. Although this represents an idealized geometric assumption to a certain extent, the core purpose of this work is to verify the feasibility of "all-event utilization" at the level of algorithm representation. Future work can be further improved by introducing physical DOI perturbations or by adopting a dual-ended readout detector model. Effects such as temporal response and electronic noise are also incorporated into the simulation model to ensure that the statistical characteristics of the simulation data are consistent with those of the experimental data.

3. *Dataset Construction*

To ensure the consistency of the spatial geometric distribution between the training samples and the test samples, this work keeps a unified spatial sampling strategy unchanged and performs stratified random sampling in the reconstruction region that is 10–40 cm away from the front surface of the detector. Source-position sampling is conducted by stratified random sampling in the three-dimensional space 10–40 cm away from the front surface of the detector. Specifically, the distance range is equally divided into three sampling layers (10–20 cm, 20–30 cm, and 30–40 cm). A lateral sampling plane of 24×24 cm^2 is defined in each layer. In each sampling layer, 144 nonrepeated source positions are randomly selected, yielding a total of 432 independent source configurations. The above sampling method can cover typical radiation scenes under different distances and lateral positions, and provides a unified spatial-distribution basis for model training and generalization testing.

In the training stage, only single-energy point-source data of Cs-137 are used. Both the training set and the validation set are composed of the simulated events corresponding to this isotope. Co-60, F-18, and the subsequent multi-source and multi-energy combined scenes do not participate in training, and are only used in the test stage to evaluate the generalization ability of the model under unseen energy conditions. To ensure comparability among different samples, the data of all isotopes are generated under the same detector structure, the same physical-process model, and the same spatial sampling rules. Only the incident-photon energy spectrum and the source-scene configuration are different.

Traditional Compton reconstruction methods usually rely on a strict event-screening procedure. Typical criteria include:

1. total-energy-window screening: for example, for the Cs-137 source, only events whose total deposited energy falls within the energy window near the target photon peak are retained;

2. interaction-pattern screening: only standard dual-layer events in which Compton scattering occurs in the scatter layer and absorption is completed in the absorber layer are retained;

3. spatial-position screening: the scatter and absorber interaction points are required to satisfy geometric conditions such as a minimum spatial interval so as to avoid nearest-neighbor mismatch and low-quality events.

The above screening procedure can improve the purity of the events used by traditional reconstruction methods, but it also significantly compresses the amount of usable data. Under the detector model and simulation conditions adopted in this work, the number of valid events finally retained after standard screening accounts for only a very small proportion of the raw detection records. This is one of the core challenges faced by low-count Compton reconstruction.

Different from the conventional idea of retaining only "high-confidence valid dual-layer events," this work adopts an "all-event utilization" strategy: raw detection events produced in the same physical simulation process and that can be uniformly encoded are retained as much as possible, so as to reduce the information loss caused by traditional screening. Specifically, the following types of events are retained in the unified raw-input set:

1. valid dual-layer events that satisfy the standard Compton-screening criteria;

2. events that do not fully satisfy the traditional energy-window or geometric-screening conditions, but can still form a unified dual-layer encoding;

3. partial multiple-scattering events that can still be mapped into a unified input format after regularized compression;

4. single-layer interaction events or information-incomplete events; such events are encoded jointly by missing-value placeholders and event-type masks so that they can be incorporated into the unified input-tensor representation.

It should be pointed out that this work does not indiscriminately include all raw detection records in training. Records that cannot be uniformly encoded, that have severely missing information, or that obviously do not satisfy basic physical interpretability are still removed. Therefore, the input of this work is "all encodable raw events," rather than all detection records in an unconstrained sense.

Based on the above definition, the data-usage protocol of different methods in this work is as follows: for the traditional methods SBP and MLEM, the input is N valid dual-layer events that satisfy the standard screening conditions; for the proposed SACN method and the 3D-UNet baseline, the input is all encodable raw events produced in the same physical simulation that generates these N valid events. Here, N is used only as a difficulty indicator that uniformly represents scene sparsity, and is not equivalent to the actual number of input events for the deep-learning methods.

The final dataset contains 432 independent source configurations. Each configuration corresponds to one set of Compton-event data and one ground-truth source-distribution image of 256×256 pixels (used for the training and evaluation

of U-Net). The dataset is randomly divided into the training set (346), validation set (43), and test set (43) at a ratio of $8:1:1$.

4. *Type Partition and Unified Encoding of Raw Events*

To reduce the information loss caused by traditional strict screening as much as possible, this work no longer retains only the "valid events" that satisfy the standard dual-layer Compton criterion. Instead, the detection records produced in the same physical simulation process and that can be represented in a unified format are constructed into a raw-event set. Let all encodable raw events be denoted as

$$\mathcal{E}_{\mathrm{raw}} = \{e_i\}_{i=1}^{M}, \tag{9}$$

Here, M is the total number of all encodable raw events retained under this scene.

According to the completeness of the detection information and the degree of physical interpretability, this work divides the raw events into the following categories:

1. standard dual-layer valid events: these satisfy the traditional energy-window criterion, dual-layer interaction pattern, and basic geometric conditions, and can completely calculate the scattering angle and define the Compton-cone constraint;

2. weak-screened dual-layer events: although these do not fully satisfy the traditional strict energy-window or geometric-screening conditions, they still have the positions and energy information of the scatter layer and the absorber layer, and can form a unified dual-layer representation;

3. compressed multiple-scattering events: there are multiple interactions in the raw detection process, but after regularized compression they can still be mapped into a unified input format so as to preserve part of the additional structural information;

4. single-layer or information-incomplete events: these contain only single-layer energy deposition, incomplete positions, or partially missing observed quantities, and cannot completely define a Compton cone, but may still carry weak information related to the source position;

5. non-encodable events: records with severely missing information, records for which a unified input representation cannot be built, or records that obviously do not possess basic physical interpretability. Such events are directly removed in the preprocessing stage.

Based on the above partition, this work further defines two subsets. The first is the event subset for which a geometric-consistency constraint can be established:

$$\mathcal{E}_{\mathrm{cone}} \subseteq \mathcal{E}_{\mathrm{raw}}, \tag{10}$$

This subset includes standard dual-layer valid events and part of the weak-screened dual-layer events for which the scattering geometry can still be stably defined. The second is the event subset that only provides auxiliary structural information but cannot directly form a cone constraint:

$$\mathcal{E}_{\mathrm{aux}} = \mathcal{E}_{\mathrm{raw}} \setminus \mathcal{E}_{\mathrm{cone}}. \tag{11}$$

This subset mainly includes single-layer events, partially incomplete events, and weak-quality events for which the scattering angle cannot be stably defined.

To realize unified modeling of different types of events, this work represents each raw event as a triplet:

$$e_i = (\widetilde{q}_i, m_i, t_i). \tag{12}$$

Here, $\widetilde{q}_i$ is the basic observation vector completed by placeholder filling, m_i is the missing-value mask, and t_i is the event-type identifier. For missing observed quantities, a fixed placeholder value is used for filling, and the mask is used to explicitly indicate whether the corresponding component truly exists. For different event types, one-hot type coding or an equivalent type-embedding vector is used for distinction. The purpose of doing so is not to force all events to have the same physical meaning, but to enable the network to distinguish, within a unified input tensor, between "events that can form strong geometric constraints" and "events that provide only weak auxiliary information."

It should be emphasized that the unified encoding of raw events does not mean that all events are treated equivalently in the subsequent physical loss. For events with $e_i \in \mathcal{E}_{\mathrm{cone}}$, this work further imposes a physical-likelihood constraint based on a differentiable Compton forward model during training. For events with $e_i \in \mathcal{E}_{\mathrm{aux}}$, they mainly participate in the overall representation learning through the event type, the missing pattern, and their joint distribution with other events, rather than directly entering the cone forward-likelihood term.

The above design enables the proposed method to inherit the strong physical geometric constraints provided by traditional dual-layer events, while preserving as much auxiliary information as possible from the raw detection records under low-count conditions, thereby forming the unified data basis of the "all-event utilization" strategy.

C. Geometric Pre-localization

Within the conditional neural-field framework, the decoder needs to evaluate the source intensity in continuous three-dimensional space. If the query points are uniformly distributed over the whole reconstruction space, the vast majority of the sampled points are far away from the true source position, which leads to extremely sparse effective gradient

signals, low training efficiency, and aggravated optimization instability. To this end, this work introduces a lightweight geometric pre-localization module to make a coarse estimate of the potential source region, denoted by $\mathbf{r}_{\mathrm{geo}}$. This module does not directly produce the final reconstruction result. Instead, it serves as a geometric prior for the subsequent continuous-space optimization, and its main role is to narrow the effective search range.

1. Cone-Consistency Objective Function

The core idea of geometric pre-localization is as follows: if the candidate point $\mathbf{r}^*$ is close to the true source position, then for any Compton event e_i, the angle between the incident direction from $\mathbf{r}^*$ to the scattering point $\mathbf{S}_i$ and the scattering direction $\hat{\mathbf{u}}_i$ should be consistent with the measured scattering angle θ_i. Accordingly, the cone-consistency residual is defined as

$$\delta_i(\mathbf{r}^*) = \frac{\mathbf{r}^* - \mathbf{S}_i}{\|\mathbf{r}^* - \mathbf{S}_i\|} \cdot \hat{\mathbf{u}}_i - \cos\theta_i. \tag{13}$$

This residual characterizes the geometric deviation between the candidate position and the cone constraint of the i-th event. A coarse estimate of the source position can be obtained by jointly optimizing over all events.

2. Three-Stage Coarse-to-Fine Solution

Because the above objective function is highly nonconvex with respect to $\mathbf{r}^*$, and is significantly affected by noisy events and outlier events, this work adopts a three-stage coarse-to-fine strategy for solution.

a. Stage 1 (coarse grid search). A uniform three-dimensional grid of $25 \times 25 \times 15$ is constructed in the reconstruction space Ω. For each grid node, the median absolute deviation (MAD) of the residual is calculated as

$$\mathcal{J}_{\mathrm{MAD}}(\mathbf{r}^*) = \mathrm{median}_{i=1}^{N}\Big(|\delta_i(\mathbf{r}^*)|\Big). \tag{14}$$

MAD rather than the mean squared error is used as the evaluation criterion because the median statistic is naturally robust to outlier events, such as noisy events produced by multiple scattering. The grid node with the minimum $\mathcal{J}_{\mathrm{MAD}}$ is selected as the initial estimate $\mathbf{r}^{(1)}$.

b. Stage 2 (refined grid search). Centered at $\mathbf{r}^{(1)}$, a refined grid of $20 \times 20 \times 15$ is constructed within a local region of ± 60 mm (lateral) and ± 40 mm (axial), and the optimal node $\mathbf{r}^{(2)}$ is selected according to the same MAD criterion. The asymmetric design of the lateral and axial search ranges reflects the physical characteristic that the Compton camera has lower resolution in the axial (depth) direction than in the lateral direction.

c. Stage 3 (gradient refinement). Using $\mathbf{r}^{(2)}$ as the initial value, the Adam optimizer (learning rate 0.5) is used to perform 300 steps of continuous optimization on the source position. In this stage, the Huber loss is used in place of MAD to obtain smooth gradients:

$$\mathcal{L}_{\mathrm{Huber}}(\mathbf{r}^*) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}\delta_i^2, & |\delta_i| < 0.1,\\ 0.1\,(|\delta_i| - 0.05)\,, & |\delta_i| \geq 0.1.\end{cases} \tag{15}$$

The Huber loss provides a quadratic gradient when the residual is small so as to accelerate convergence, and degenerates to a linear form when the residual is large so as to suppress the excessive influence of outlier events. During optimization, coordinate clipping is imposed on $\mathbf{r}^*$ to ensure that it always stays within the reconstruction space Ω.

It should be emphasized that geometric pre-localization is an offline preprocessing step. It is executed only once for each sample before training, and does not participate in network gradient propagation. Therefore, it does not increase the back-propagation overhead during training. Its accuracy also does not need to reach the level required by the final three-dimensional reconstruction. It only needs to provide a geometric center with a correct directional tendency.

D. Network Architecture

This section describes in detail the architectural design of the proposed network. Unlike conventional methods that represent the reconstruction result as a discrete pixel image, this paper models the radioactive source distribution as a continuous three-dimensional neural implicit field, $f_\theta : \mathbb{R}^3 \to \mathbb{R}^+$, which is driven by the global feature conditioned on the Compton event set. The overall framework of the network is shown in Fig. 2. It consists of three functional modules: an event-set encoder, a conditional neural-field decoder, and an auxiliary coordinate prediction head. The encoder extracts a permutation-invariant global feature representation from a variable number of Compton events. The decoder, conditioned on this feature, predicts the source intensity at an arbitrary query point in continuous three-dimensional space. The auxiliary coordinate prediction head provides a direct spatial localization training signal for the encoder. The design motivation, structural details, and data flow of each module are described below.

1. Input Feature Definition and Grouping

For conventional Compton reconstruction, an event is usually represented as a standard double-layer event vector composed of the three-dimensional position in the scatter layer, the three-dimensional position in the absorber layer, and the energy depositions in the two layers. However, this paper adopts a unified input strategy of "all encodable raw events". Therefore, the network input is no longer limited

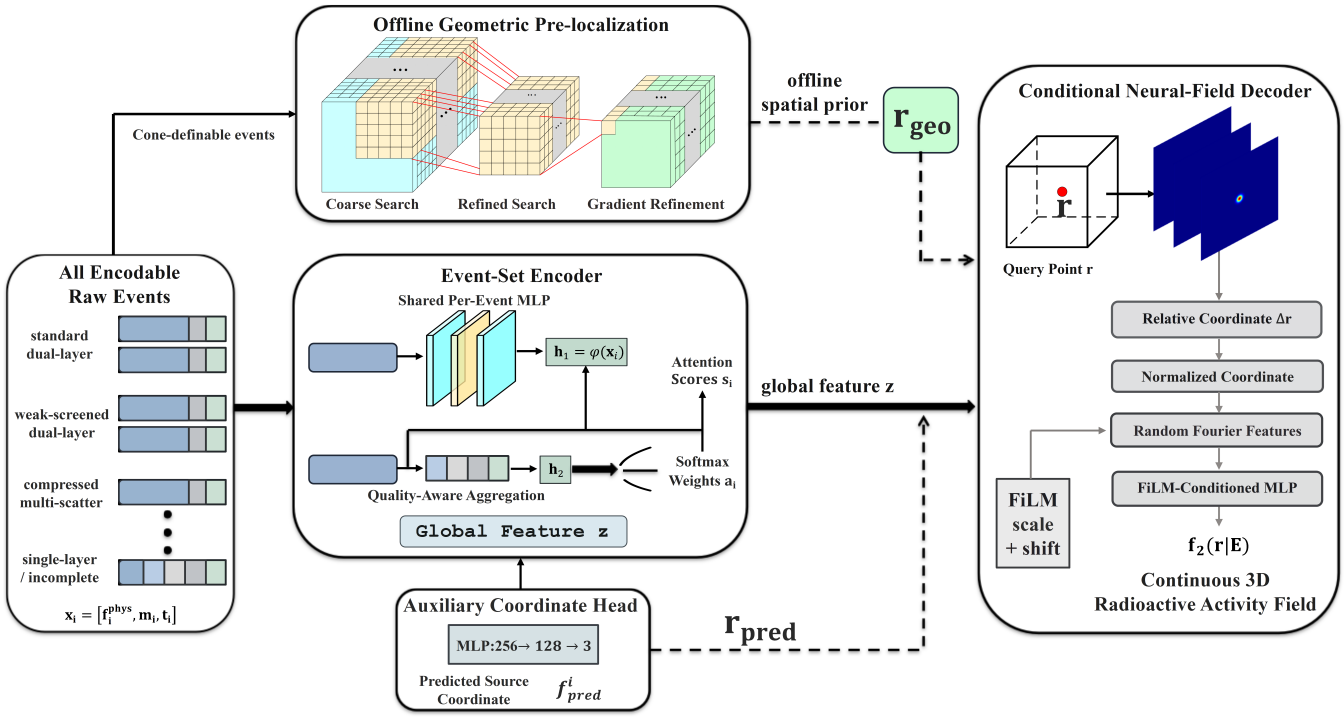

Fig. 2. Overall architecture of the network.

to a fixed-form double-layer event. In particular, for single-layer events, partial-energy-deposition events, or compressed multiple-scattering events, some physical quantities may be missing. If the standard double-layer input format is still forcibly used, it will lead to inconsistent representation or confusion in physical meaning.

For this reason, each raw event is uniformly represented as

$$e_i = (\widetilde{q}_i, m_i, t_i), \tag{16}$$

where $\widetilde{q}_i$ is the basic observation vector after placeholder completion, m_i is the missing-value mask, and t_i is the event-type identifier. The basic physical-feature part retains the directly observable position and energy information, and also introduces, as much as possible, the derived physical quantities related to Compton-scattering geometry. For events that can form a double-layer geometric constraint, this paper further calculates the scattering angle θ_i, its cosine value $\cos\theta_i$, and the unit direction vector $\widehat{\mathbf{u}}_i$ pointing from the scattering point to the absorption point on the basis of the raw observation, so as to explicitly encode the core geometric information of the Compton cone. For events for which the above quantities cannot be completely defined, fixed placeholder values are used for filling, and the missing components are explicitly indicated by the mask vector.

Specifically, the basic physical features are written as

$$f_i^{\mathrm{phys}} = \left[\widetilde{\mathbf{p}}_i^{\,s}, \widetilde{\mathbf{p}}_i^{\,a}, \widetilde{E}_i^{\,s}, \widetilde{E}_i^{\,a}, \widetilde{\cos\theta}_i, \widetilde{\mathbf{u}}_i\right], \tag{17}$$

where $\widetilde{\mathbf{p}}_i^{\,s}$ and $\widetilde{\mathbf{p}}_i^{\,a}$ are the normalized representations of the positions in the scatter layer and the absorber layer, respectively, $\widetilde{E}_i^{\,s}$ and $\widetilde{E}_i^{\,a}$ are the normalized energy depositions, and $\widetilde{\cos\theta}_i$ and $\widetilde{\mathbf{u}}_i$ are the physical quantities derived from the double-layer geometric relationship when they are computable and are completed with placeholder values when they are missing. Correspondingly, the missing-value mask is defined as

$$m_i \in \{0,1\}^{d_m}, \tag{18}$$

where each dimension indicates whether the corresponding physical quantity is truly available; the event-type encoding

$$t_i \in \{0,1\}^{d_t}. \tag{19}$$

is used to indicate whether the event belongs to a standard valid double-layer event, a weakly screened double-layer event, a compressed multiple-scattering event, or a single-layer/incomplete event. The event feature vector finally input to the encoder is written as

$$x_i = [f_i^{\mathrm{phys}}, m_i, t_i]. \tag{20}$$

Compared with using only the standard eight-dimensional raw observables, this design has three advantages. First, for events for which double-layer geometry can be defined, $\cos\theta_i$ and $\mathbf{u}_i$ directly parameterize the key structure of the Compton cone, so that the network does not need to learn the cone geometry completely implicitly from the raw coordinates. Second, through the missing mask, the network can explicitly distinguish the two essentially different situations of "this quantity is zero" and "this quantity is missing and filled by a placeholder", thus avoiding information contamination. Third, through the event-type encoding, the network

can learn the differences in information quality and physical reliability among different event categories in a unified input representation, thus providing a basis for the subsequent attention-weighted aggregation and physical-constraint modeling.

It should be pointed out that the purpose of using a unified event representation in this paper is to preserve the raw detection information under low-count conditions as much as possible within the same tensor-input framework, rather than to assign exactly the same physical-constraint strength to all events. The subsequent network will distinguish the contributions of different events through the mask, type information, and the selective action of the physical loss.

2. *Event-Set Encoder*

The event-set encoder is responsible for extracting a fixed-dimensional global scene representation $\mathbf{z}$ from the variable-length raw event set

$$\mathcal{E}_{\mathrm{raw}} = \{x_i\}_{i=1}^{M}. \tag{21}$$

Since Compton events themselves do not have sequential-order meaning, the encoder must satisfy two basic requirements: first, the output should be permutation-invariant to the ordering of the input events; second, it should be able to handle event sets with significantly varying numbers in different scenes, so as to adapt to the input differences from extremely sparse to relatively high statistics. Different from the conventional setting that only processes standard double-layer events, the input in this paper contains both events that can form strong geometric constraints and events that provide only weak auxiliary information. Therefore, the encoder should also have the ability to adaptively model event quality and information completeness.

a. Per-event feature mapping. For each uniformly represented event vector $x_i = [f_i^{\mathrm{phys}}, m_i, t_i]$, it is first mapped to a high-dimensional latent representation by a multilayer perceptron with shared parameters:

$$h_i = \phi(x_i), \qquad h_i \in \mathbb{R}^{d_h}, \tag{22}$$

where $\phi(\bullet)$ is an event-level feature extractor composed of multilayer fully connected layers, and the SiLU activation function is used between layers. Parameter sharing ensures that the same mapping is applied to all events, thus satisfying the permutation equivariance required for set modeling. Since the input already explicitly contains the missing-value mask m_i and the event-type encoding t_i, this per-event mapping process extracts not only geometric and energy features, but also simultaneously learns event completeness, event quality, and differences among event categories.

b. Quality-aware attention aggregation. After the per-event latent representation is obtained, this paper uses attention-weighted aggregation rather than simple mean pooling. The reason is that different events contribute differently to reconstruction: standard valid double-layer events usually provide strong Compton geometric constraints, whereas single-layer and incomplete events mostly carry only auxiliary structural information. If all events are assigned the same weight, the effective contribution of high-quality events is easily weakened. To this end, a scalar attention score is calculated for each event:

$$s_i = g(h_i), \tag{23}$$

where $g(\bullet)$ is a two-layer perceptron. Since h_i itself already contains physical features, the missing mask, and event-type information, s_i actually reflects the comprehensive contribution of this event to the current scene representation. Then, the aggregation weight is obtained by Softmax normalization:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{M} \exp(s_j)}. \tag{24}$$

Finally, the global feature representation is obtained by weighted summation:

$$\mathbf{z} = \sum_{i=1}^{M} \alpha_i h_i. \tag{25}$$

The above aggregation method has three roles. First, it naturally satisfies permutation invariance, because no matter how the order of the input events changes, the result of Softmax weight normalization and weighted summation remains unchanged. Second, it can handle variable-length inputs, because the output dimension is determined only by the latent representation dimension d_h and is independent of the total number of events M. Third, it equips the model with an explicit "quality-aware" capability: the network can adaptively increase the contribution of informative events and suppress the interference of low-reliability events to the global representation according to factors such as whether the event can form a stable cone constraint, whether there is severe missing information, and whether it belongs to a weak-quality event.

It should be emphasized that the attention aggregation here is not equivalent to artificially "performing hard screening again". Different from traditional event screening, this paper does not directly discard weak-quality events before encoding. Instead, they are incorporated into a unified set representation, and the network adaptively learns their relative contributions during training. This design allows the model to preserve the information gain brought by all-event utilization while avoiding feature contamination caused by low-quality events entering the subsequent reconstruction without discrimination.

3. *Auxiliary Coordinate Prediction Head*

In the conditional encoder–decoder framework, the training signal of the encoder comes entirely from the back-propagation of the physical loss function at the decoder side. However, this indirect gradient path faces a difficulty in the early stage of training: the decoder has not yet learned how to use the global feature $\mathbf{z}$, which causes the gradient signal

returned to the encoder to be weak and noisy, and the encoder may degenerate into an approximately identity mapping that outputs almost indistinguishable feature representations for different scenes.

To alleviate the problem of weak physical-loss gradients in the early stage of training, this paper attaches an auxiliary coordinate prediction head on top of the encoder, which directly regresses a coarse source-position estimate from the global feature $\mathbf{z}$:

$$\widehat{\mathbf{r}}_{\mathrm{pred}} = \mathrm{MLP}_{\psi}(\mathbf{z}) \in \mathbb{R}^3. \tag{26}$$

This prediction head is implemented by a two-layer fully connected network (256→128→3), with the SiLU activation function used in the middle. This coordinate prediction is not used as the final output. Instead, it is used to provide directional guidance in the early stage of training and to construct relative coordinates during decoding. It should be pointed out that the weight of the auxiliary loss is set to $\lambda_{\mathrm{coord}} = 0.001$, which is far smaller than the scale of the main physical loss. Its role is to provide directional guidance for the encoder rather than to impose a strong constraint, so as to ensure that the feature representation of the encoder during training is dominated by the physical loss rather than locked by the coarse-localization target.

4. *Conditional Neural-Field Decoder*

The decoder is the core component of the proposed method. It is responsible for predicting the source intensity at an arbitrary query point $\mathbf{R} \in \mathbb{R}^3$ in continuous three-dimensional space, conditioned on the global feature $\mathbf{z}$ and the predicted coordinate $\widehat{\mathbf{r}}_{\mathrm{pred}}$ output by the encoder. The decoder design integrates three key techniques: relative-coordinate representation, random Fourier feature encoding, and the FiLM conditioning mechanism.

a. Relative-Coordinate Representation Traditional neural implicit-field methods directly use absolute spatial coordinates as input. This requires the network to implicitly learn the positional information of the source in the whole reconstruction space. This representation has a fundamental difficulty in cross-scene generalization: the source positions in different scenes are different, and the network needs to relearn the mapping from absolute coordinates to intensity for each new source position.

This paper proposes to transform the input of the decoder from absolute coordinates to relative coordinates that take the predicted source position as the origin:

$$\Delta\mathbf{r} = \mathbf{r} - \widehat{\mathbf{r}}_{\mathrm{pred}}. \tag{27}$$

This transformation changes the object learned by the decoder from "the shape of the source distribution in absolute space" to "the point spread function (PSF) relative to the source position". Since the PSFs in different scenes are highly similar in shape (all are concentrated distributions centered on the source), the relative-coordinate representation significantly reduces the difficulty of cross-scene generalization. To maintain numerical stability, the relative coordinates are normalized by the reconstruction-space scale factor:

$$\widetilde{\Delta\mathbf{r}} = \frac{\Delta\mathbf{r}}{\mathbf{s}_{\mathrm{coord}}}, \tag{28}$$

where $\mathbf{s}_{\mathrm{coord}} = [L_x, L_y, L_z]$ is the reconstruction-space size along each axis.

In addition, in order to make the decoder aware of the absolute position of the source in the reconstruction space (because a source located at the detector edge and one located at the center may have different PSF shapes), the normalized absolute-position information $\widetilde{\mathbf{r}}_{\mathrm{abs}}$ of the predicted coordinate is also provided to the decoder as an auxiliary input:

$$\widetilde{\mathbf{r}}_{\mathrm{abs}} = 2 \cdot \frac{\widehat{\mathbf{r}}_{\mathrm{pred}} - \mathbf{r}_{\min}}{\mathbf{r}_{\max} - \mathbf{r}_{\min}} - 1 \in [-1, 1]^3, \tag{29}$$

where $\mathbf{r}_{\min}$ and $\mathbf{r}_{\max}$ are the lower and upper bounds of the reconstruction space, respectively.

b. Random Fourier Feature Encoding Recent studies have shown that standard multilayer perceptrons have a preference for learning low-frequency signals first, which makes it difficult for the network to represent high-frequency spatial details, such as the sharp peak structure of a point source. To overcome this limitation, this paper introduces random Fourier feature encoding to map the normalized relative coordinates to a high-dimensional frequency space.

Let $B \in \mathbb{R}^{3\times d_B}$ be a frequency matrix randomly sampled from the Gaussian distribution $\mathcal{N}(0, \sigma_B^2)$ and fixed during training, where $d_B = 128$ and $\sigma_B = 2.0$. The Fourier feature encoding is defined as

$$\gamma(\widetilde{\Delta\mathbf{r}}) = \Big[\sin(2\pi\widetilde{\Delta\mathbf{r}}B), \cos(2\pi\widetilde{\Delta\mathbf{r}}B)\Big] \in \mathbb{R}^{2d_B}, \tag{30}$$

where the sin and cos functions act element-wise. The frequency scale σ_B controls the range of spatial frequencies covered by the encoding: $\sigma_B = 2.0$ achieved the best balance between sharp peak representation and training stability in the experiments. Although a larger σ_B helps capture higher-frequency spatial details, it tends to introduce overfitting noise under sparse-data conditions.

c. FiLM Conditioning Mechanism The decoder needs to dynamically adjust its spatial prediction behavior according to the event features of different scenes, which are encoded in the global feature $\mathbf{z}$. This paper uses the Feature-wise Linear Modulation (FiLM) mechanism to realize this conditioning process. The design motivation is as follows: compared with simple feature concatenation (directly concatenating $\mathbf{z}$ with spatial features), FiLM realizes conditioning by applying an affine transformation to hidden-layer features, enabling the global feature to independently scale and shift each channel of the spatial features, thus providing stronger modulation ability and more flexible feature interaction.

The decoder backbone consists of a four-layer fully connected network with hidden dimension $d_h = 512$. The input is the concatenated vector of the Fourier encoding, the normalized relative coordinates, and the normalized absolute

position:

$$\mathbf{h}_0 = [\gamma(\widetilde{\Delta\mathbf{r}}), \widetilde{\Delta\mathbf{r}}, \widetilde{\mathbf{r}}_{\text{abs}}] \in \mathbb{R}^{2d_B+3+3}. \tag{31}$$

For the lth layer ($l = 1, 2, 3, 4$), the forward computation is executed according to the following steps:

$$\mathbf{h}'_l = \text{SiLU}(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l)\,, \tag{32}$$

$$[\gamma_l, \beta_l] = \mathbf{V}_l \mathbf{z} + \mathbf{c}_l, \tag{33}$$

$$\mathbf{h}_l = \gamma_l \odot \mathbf{h}'_l + \beta_l, \tag{34}$$

where $\mathbf{W}_l \in \mathbb{R}^{d_h \times d_{\text{in},l}}$ and $\mathbf{b}_l$ are the weight and bias of the lth layer, $\mathbf{V}_l \in \mathbb{R}^{2d_h \times d_f}$ and $\mathbf{c}_l$ are the weight and bias of the FiLM-parameter generation layer, $\gamma_l, \beta_l \in \mathbb{R}^{d_h}$ are the modulation scale factor and shift factor, respectively, and $\odot$ denotes element-wise multiplication.

The initialization strategy of the FiLM-parameter generation layer is crucial to training stability. In this paper, the weight of $\mathbf{V}_l$ is initialized as a zero matrix, and in the bias $\mathbf{c}_l$, the scaling part is initialized as an all-ones vector and the shifting part is initialized as an all-zeros vector, namely,

$$\gamma_l|_{t=0} = \mathbf{1}, \qquad \beta_l|_{t=0} = \mathbf{0}. \tag{35}$$

This "identity initialization" strategy ensures that the FiLM layer is an identity transform at the beginning of training ($\mathbf{h}_l = \mathbf{h}'_l$), and the behavior of the decoder is equivalent to that of a standard MLP without conditioning. As training proceeds, the network gradually learns meaningful conditional modulation, thereby avoiding the early-stage training instability that may be caused by random initialization.

d. Output Layer and Non-negativity Constraint After the final hidden-layer feature is linearly mapped to a scalar output, a non-negativity constraint is imposed through the Softplus activation function:

$$f_\theta(\mathbf{r}) = \text{Softplus}\big(\mathbf{w}_{\text{out}}^T \mathbf{h}_4 + b_{\text{out}}\big) + \epsilon, \tag{36}$$

where $\text{Softplus}(x) = \ln(1 + e^x)$, and $\epsilon = 10^{-6}$ is a small constant used to prevent numerical underflow. The output bias b_{out} is initialized to 1.0, so that the source-intensity prediction is a positive-valued distribution in the initial stage of training, thereby avoiding gradient vanishing caused by the saturation region of Softplus near zero. The reason for choosing Softplus rather than ReLU as the output activation is its global differentiability. The gradient of ReLU is discontinuous at zero. When the predicted intensity of a large number of query points is zero, which is very common in sparse-source scenarios, large gradient dead zones will appear and hinder the optimization of the loss function in these regions.

5. Loss Function and Training Strategy

This section derives the physics-constrained loss function required for network training on the basis of Compton-scattering dynamics and maximum-likelihood-estimation theory. Different from standard emission tomography that relies on line-integral projection, Compton imaging uses the complete interaction information of single photons to infer the source origin, which requires establishing a specific likelihood model based on geometric constraints.

6. Theoretical Basis of the Physics-Constrained Loss Function

The image reconstruction of a Compton camera is essentially an inverse problem of reconstructing the radioactive source distribution under incomplete and noisy event observations. Different from conventional supervised learning, which relies on a large number of labeled images, the training objective established in this paper does not require the network output to match a reference image pixel by pixel. Instead, it requires the predicted source distribution to explain, as much as possible, the physically consistent information in the observed event set.

Let the continuous three-dimensional source distribution output by the neural network be

$$\rho_\theta(\mathbf{r} \mid \mathcal{E}_{\text{raw}}), \qquad \mathbf{r} \in \Omega, \tag{37}$$

where θ is the network parameter, $\mathcal{E}_{\text{raw}}$ is the input raw event set, and Ω is the reconstruction region. Since this paper is more concerned with the relative spatial shape of the source distribution rather than the absolute total intensity, it is normalized into the form of a probability density:

$$\bar{\rho}_\theta(\mathbf{r}) = \frac{\rho_\theta(\mathbf{r})}{\int_\Omega \rho_\theta(\mathbf{r})\, d\mathbf{r} + \varepsilon}, \tag{38}$$

where ε is a numerical-stability term. The normalized $\bar{\rho}_\theta(\mathbf{r})$ can be interpreted as the relative probability distribution of the photon emission position in the reconstruction space.

It should be emphasized that the input in this paper is "all encodable raw events", but not all raw events can be fully characterized by a unified Compton-cone geometric model. Therefore, at the physical-modeling level, the raw event set is further divided into two parts:

$$\mathcal{E}_{\text{raw}} = \mathcal{E}_{\text{cone}} \cup \mathcal{E}_{\text{aux}}, \qquad \mathcal{E}_{\text{cone}} \cap \mathcal{E}_{\text{aux}} = \varnothing, \tag{39}$$

where $\mathcal{E}_{\text{cone}}$ denotes the subset of events for which the Compton-scattering geometric constraint can be stably defined according to the position and energy information, and $\mathcal{E}_{\text{aux}}$ denotes the subset of events such as single-layer events, partially incomplete events, or other events that cannot stably form cone constraints but can still provide auxiliary structural information. The former is used to construct the explicit physical likelihood. The latter participates in the representation learning of the event set through the unified input representation, but does not directly enter the cone forward-likelihood term.

For any event $e_i \in \mathcal{E}_{\text{cone}}$, its scattering point, absorption point, and energy deposition can be used to calculate the scattering angle θ_i. According to the Compton-scattering relation, this event corresponds to a cone geometric constraint with the scattering point as the cone vertex, the scattering direction as the cone axis, and θ_i as the half-opening angle.

In other words, if a candidate source distribution can assign a higher probability mass near the cones corresponding to these events, it is more likely to explain the current observation. Based on this, the reconstruction problem is formulated as a maximum-likelihood-estimation problem based on event geometric consistency.

The statistical assumption of the physical constraint in this paper is built on the approximate independence of events under sparse detection conditions. Since the incident count rate is low, the explicit temporal correlations introduced by detector dead time and pulse pile-up can be approximately ignored. Therefore, for events in $\mathcal{E}_{\mathrm{cone}}$, the joint likelihood can be approximately factorized into the product of the marginal likelihoods of single events. It should be noted that this independence assumption applies only to the geometrically interpretable subset of events that enter the physical likelihood, and it does not mean that all raw events in $\mathcal{E}_{\mathrm{raw}}$ are regarded as following the same cone-observation model. Through the strategy of "unified input of raw events + explicit physical constraints on geometrically interpretable events", this paper combines all-event utilization with interpretable physical modeling.

7. *Differentiable Cone Forward Operator*

For an event $e_i \in \mathcal{E}_{\mathrm{cone}}$ that can define a Compton geometric constraint, its observation probability is closely related to the probability mass of the source distribution near the corresponding cone surface, which is the basic starting point of the Compton-imaging forward problem. Let the geometrically interpretable observable of the ith event be

$$q_i = (\mathbf{p}_i^s, \mathbf{p}_i^a, E_i^s, E_i^a), \tag{40}$$

where $\mathbf{p}_i^s$ is the scattering point and $\mathbf{p}_i^a$ is the absorption point. According to the Compton relation, the scattering angle θ_i can be obtained, and the cone axis is determined by the direction vector from the scattering point to the absorption point. Therefore, for any spatial position $\mathbf{r} \in \Omega$, its unit direction vector with respect to the scattering point is defined as

$$\mathbf{v}_i(\mathbf{r}) = \frac{\mathbf{r} - \mathbf{p}_i^s}{\|\mathbf{r} - \mathbf{p}_i^s\| + \varepsilon_r}. \tag{41}$$

where ε_r is a numerical-stability term used to prevent the denominator from being zero. If $\mathbf{r}$ lies on the ideal cone surface, it should satisfy $\angle(\mathbf{v}_i(\mathbf{r}), \mathbf{u}_i) = \theta_i$, where $\mathbf{u}_i$ is the unit vector of the cone axis.

In the actual detection process, the scattering angle is not strictly deterministic, but is affected by factors such as energy resolution and Doppler broadening. Let the total angular uncertainty of the ith event be $\sigma_{\theta,i}$, then it can be written as

$$\sigma_{\theta,i}^2 = \sigma_{\theta,i,\mathrm{ER}}^2 + \sigma_{\theta,i,\mathrm{DB}}^2, \tag{42}$$

where $\sigma_{\theta,i,\mathrm{ER}}$ denotes the scattering-angle uncertainty caused by the finite energy resolution of the detector, and $\sigma_{\theta,i,\mathrm{DB}}$ denotes the Doppler-broadening uncertainty caused by the momentum distribution of bound electrons in the target material. This paper follows the detector-parameter settings described above and approximately models these two terms, thus obtaining the effective angular-resolution width corresponding to each event.

To embed the above physical constraint into a differentiable training objective, this paper first introduces the Klein–Nishina (KN) differential scattering cross section as an angular prior probability weight. Let the incident photon energy be $E_i^s + E_i^a$ and the scattered photon energy be E_i^a. Then the KN cross-section weight factor is defined as

$$W_{\mathrm{KN}}(\theta_i) = \left(\frac{E_i^a}{E_i^s + E_i^a}\right)^2 \cdot \left(\frac{E_i^s + E_i^a}{E_i^a} + \frac{E_i^a}{E_i^s + E_i^a} - \sin^2\theta_i\right). \tag{43}$$

Combined with this physical cross section, a soft cone kernel with KN weighting is constructed to measure the geometric consistency between any position $\mathbf{r}$ and event e_i:

$$K_i(\mathbf{r}) = W_{\mathrm{KN}}(\theta_i) \cdot \frac{1}{\|\mathbf{r} - \mathbf{p}_i^s\|^2 + \varepsilon_r} \cdot \exp\left(-\frac{\left[\arccos\big(\mathbf{v}_i(\mathbf{r}) \cdot \mathbf{u}_i\big) - \theta_i\right]^2}{2\sigma_{\theta,i}^2 + \varepsilon_\theta}\right). \tag{44}$$

This modified kernel not only ensures spatial geometric consistency, but also strictly follows the quantum-mechanical scattering-probability distribution. Here, ε_θ is a numerical-stability term. This kernel consists of two parts: the exponential term is used to describe the consistency between position $\mathbf{r}$ and the cone geometry, and its Gaussian form reflects the soft-constraint broadening caused by the scattering-angle measurement error; the distance term $(\|\mathbf{r} - \mathbf{p}_i^s\|^2 + \varepsilon_r)^{-1}$ is used to approximately characterize the geometric-sensitivity correction in radiation transport, so as to suppress pseudo-high responses caused only by the increase in cone-surface area at large distances.

It should be noted that this soft cone kernel is defined only for events in $\mathcal{E}_{\mathrm{cone}}$. For events in $\mathcal{E}_{\mathrm{aux}}$ for which the scattering geometry cannot be stably determined, this paper does not forcibly construct a unified cone kernel and does not directly incorporate them into the forward likelihood. Instead, they play an auxiliary role at the representation level through the event-set encoder described above. This avoids incorrectly interpreting events without clear cone-physics meaning as standard Compton geometric constraints.

8. *Event Matching Degree and Physical Loss*

Given the predicted source distribution $\bar{\rho}_\theta(\mathbf{r})$, the matching degree of the ith geometrically interpretable event is defined as the weighted integral of the source distribution over its soft cone kernel:

$$s_i = \int_\Omega \bar{\rho}_\theta(\mathbf{r}) K_i(\mathbf{r})\, d\mathbf{r}, \qquad e_i \in \mathcal{E}_{\mathrm{cone}}. \tag{45}$$

This quantity can be understood as follows: if the predicted source distribution assigns a high probability mass near the positions consistent with the event geometry, then s_i is large, indicating that the current reconstruction result is more capable of explaining this event; otherwise, if the source distribution is inconsistent with the cone constraint corresponding to this event, then s_i is small.

Considering that differences may still exist in the physical reliability and information completeness of different events, this paper further introduces an event weight w_i into the physical loss. A higher weight is assigned to standard valid double-layer events; for events that can define a cone but have weaker quality, a relatively lower weight can be assigned. Therefore, the joint likelihood based on $\mathcal{E}_{\mathrm{cone}}$ can be written as

$$p(\mathcal{E}_{\mathrm{cone}} \mid \bar{\rho}_\theta) \approx \prod_{e_i \in \mathcal{E}_{\mathrm{cone}}} (s_i + \varepsilon_s)^{w_i}, \tag{46}$$

where ε_s is a stability term used to prevent numerical underflow. The corresponding negative log-likelihood form is

$$\mathcal{L}_{\mathrm{phys}} = -\sum_{e_i \in \mathcal{E}_{\mathrm{cone}}} w_i \log(s_i + \varepsilon_s). \tag{47}$$

It can thus be seen that the physical loss in this paper does not impose a unified cone-consistency constraint on all raw events. Instead, it explicitly optimizes only the geometrically interpretable event subset, while other incomplete events influence the global conditional feature $\mathbf{z}$ through the unified input representation and indirectly participate in the prediction of the source field. This design of "all-event utilization at the representation level + selective physical constraints at the loss level" is an important feature that distinguishes this paper from traditional strictly screened reconstruction and purely supervised image reconstruction.

Compared with conventional image-supervision loss, the above physical loss has three advantages. First, its optimization objective depends only on the observed events and the differentiable forward operator, and does not require image-level labels. Therefore, it belongs to physics-constrained training without image labels. Second, the network-parameter update is directly driven to improve the ability of the predicted source distribution to explain the real observed events, thereby explicitly satisfying the consistency requirement of Compton-scattering geometry. Third, since the constraint comes from the event-level physical relationship rather than a fixed image-label distribution, this method has a more reasonable basis for generalization under low-count and unseen-energy conditions.

It should be further stated that the current version of this paper does not explicitly construct an independent background-distribution term for random coincidences, non-Compton background, or severely misordered events. Instead, their influence is jointly weakened through event preprocessing, unified encoding, event-quality weighting, and attention aggregation. Therefore, the strict applicable object of the current physical loss in this paper is still mainly events that can form geometrically interpretable constraints. For more complex raw-background modeling, it can still be extended on this basis into an explicit mixed-likelihood model in future work.

9. *Regularization Term*

Under extremely sparse-data conditions ($N < 10$), relying only on the physical constraint may lead to pathological solutions, such as isolated noise points or overly sharpened spikes. To improve the stability of the solution, total variation (TV) regularization is introduced:

$$\mathrm{TV}(f_\theta) = \sum_{\mathbf{r}} \Big(|\nabla_x f_\theta(\mathbf{r})|^2 + |\nabla_y f_\theta(\mathbf{r})|^2 + |\nabla_z f_\theta(\mathbf{r})|^2 + \delta^2 \Big)^{1/2}. \tag{48}$$

where ∇_x, ∇_y, and ∇_z denote the discrete gradient operators along the three coordinate directions in three-dimensional space, respectively, and $\delta = 10^{-6}$ is a smoothing parameter to ensure differentiability. TV regularization promotes piecewise smooth solutions, effectively suppresses checkerboard artifacts and isolated noise points, and at the same time preserves the sharp features of source boundaries.

The final optimization objective is a weighted combination of the physical-constraint term and the TV regularization term:

$$\mathcal{L}_{\mathrm{total}} = \alpha \mathcal{L}_{\mathrm{phys}} + \tau\, \mathrm{TV}(f_\theta). \tag{49}$$

The hyperparameters α and τ are determined by grid search on the validation set. In the search space $\alpha \in \{0.5, 1.0, 2.0\}$ and $\tau \in \{0.001, 0.01, 0.1\}$, $\alpha = 1.0$ and $\tau = 0.01$ achieved the best validation performance.

10. *Auxiliary Geometric Guidance Loss*

As described in Section 2.4.3, the auxiliary coordinate prediction head of the encoder outputs $\widehat{\mathbf{r}}_{\mathrm{pred}}$, which requires a direct training signal. This paper takes the geometric prelocalization result $\mathbf{r}_{\mathrm{geo}}$ (see Section 2.3 for details) as the target and defines a mean-squared-error loss:

$$\mathcal{L}_{\mathrm{coord}} = \|\widehat{\mathbf{r}}_{\mathrm{pred}} - \mathbf{r}_{\mathrm{geo}}\|^2. \tag{50}$$

The weight of this loss is set to $\lambda_{\mathrm{coord}} = 0.001$, which is far smaller than the scale of the physical NLL loss. This design ensures that the auxiliary loss provides directional guidance for the encoder in the early stage of training, while the feature learning is dominated by the physical loss in the later stage of training, thus avoiding the representation capability of the encoder being locked by the coarse-localization target.

11. *Optimizer and Learning-Rate Strategy*

The Adam optimizer is used, with the initial learning rate $\eta_0 = 10^{-4}$ and the momentum parameters $\beta_1 = 0.9$ and

$\beta_2 = 0.999$. The weight-decay coefficient is set to $\lambda = 10^{-5}$ to provide mild L_2 regularization and prevent overfitting caused by excessively large weights. To provide more refined parameter adjustment in the later stage of training, a cosine-annealing learning-rate schedule is adopted:

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\left(1 + \cos\left(\frac{t}{T}\pi\right)\right), \quad (51)$$

where t is the current training step, T is the total number of training steps, $\eta_{\max} = 10^{-4}$, and $\eta_{\min} = 10^{-6}$. Compared with stepwise decay, this strategy is smoother. The learning rate starts from $\eta_{\max}$ and decays from fast to slow, which is conducive to helping the network escape from local optima and converge to a better solution in the later stage of training.

12. *Progressive Training Strategy*

Considering that the reconstruction difficulty of extremely sparse data ($N < 10$) is significantly higher than that of moderately sparse data ($N \in [10, 100]$), we adopt a progressive training strategy:

a. Stage 1 (Epoch 0–50). Samples with a relatively large number of events ($N \in [100, 1000]$) are used for training. In this stage, the network learns the basic patterns of Compton imaging and the cone geometric constraints, and establishes a stable feature representation.

b. Stage 2 (Epoch 50–100). Moderately sparse samples ($N \in [10, 100]$) are gradually introduced. The lower bound of the number of events decays linearly from 50 to 10, so that the network gradually adapts to sparser input conditions.

c. Stage 3 (Epoch 100–150). Extremely sparse samples ($N \in [2, 10]$) are added for joint training. In this stage, samples of each sparsity level are randomly sampled according to a uniform distribution, so as to ensure the balanced performance of the network over the whole sparsity range.

This progressive strategy avoids the network falling into poor local optima in the early training stage due to the high noise and strong uncertainty of extremely sparse data, and guides the network to establish robust feature representations from easy to difficult by gradually increasing the task difficulty.

III. EXPERIMENTS AND RESULTS

This section provides a unified description of the experimental configuration, including the hardware and software environment, the sparse-data experimental design, the evaluation metric system, and the implementation details of the comparison methods. To ensure that the experimental results accurately reflect the true performance of the proposed method under low-count conditions with raw unscreened-event input, all experiments in Chapter 3 were conducted under the same detector model, simulation parameters, and spatial sampling rules as those described in Chapter 2.

A. Experimental setup

1. *Hardware and software environment*

All experiments were carried out on a workstation equipped with three NVIDIA GeForce RTX 4090 GPUs. The central processing unit was an Intel Core i9-14900KS, and the system memory was 64 GB. The deep learning models were implemented in PyTorch 1.12.0, with CUDA 12.1 and cuDNN 8.9.2. The Python version was 3.8.10. The conventional reconstruction algorithms (SBP and MLEM) were implemented based on NumPy 1.24.3 and SciPy 1.10.1, and were accelerated by just-in-time compilation using Numba. The Monte Carlo simulation was built on Geant4 11.1.1 with multithreaded parallel acceleration. The complete training of Sparse-Aware ComptonNet required approximately 25 h for 150 epochs.

2. *Sparse-data experimental design*

To systematically evaluate the reconstruction performance of the proposed method under low-count conditions, we designed experiments covering multiple sparsity levels. The sparsity of a scenario was characterized by the number of valid double-layer events N retained after the standard screening criteria (see Section 2.2.3). It should be noted that, in the dual-source and multi-energy three-source experiments, N denotes the number of valid events corresponding to each independent point source. Here, N is used only as a unified sparsity scale to describe the reconstruction difficulty of different physical scenarios, and is not equal to the actual total number of input events for all methods. For the conventional methods SBP and MLEM, the input consisted of N valid double-layer events satisfying the standard screening conditions. For the proposed SACN and the 3D-UNet baseline, the input consisted of the full set of encodable raw events generated in the same physical simulation that produced these N valid events.

Therefore, the main focus of the present experiments is not the fairness of different algorithm bodies under exactly matched inputs, but the performance difference between two data-usage strategies under the same underlying physical scenario, namely, *strict-screening reconstruction* and *all-event-utilization reconstruction*. This experimental design aims to answer the following question: under extremely sparse conditions, can retaining as many encodable raw events as possible and exploiting them through unified representation and selective physical constraints support source reconstruction more effectively than the conventional strict-screening strategy?

Training was performed using only single-energy Cs-137 point-source data, in order to examine the transfer ability of the model from a single training isotope to unseen energies and more complex source scenarios. Co-60, 511 keV annihilation radiation, and multi-source or multi-energy combined scenarios were not used in training and were employed only for performance evaluation during testing. Under each sparsity level, three kinds of experiments were conducted:

1. single-source experiments (Cs-137, 662 keV), used to evaluate basic localization accuracy, spatial resolution, and background suppression ability;

2. dual-source experiments (511 keV annihilation radiation + Co-60), used to verify multi-source separation ability and source-interference suppression ability;

3. multi-energy three-source experiments (662 keV Cs-137, 511 keV annihilation radiation, and 1173 keV Co-60), used to evaluate the overall reconstruction performance under complex radiation-field conditions.

These settings enabled us to systematically analyze the performance differences among methods or data-usage strategies, from single-source to multi-source and from single-energy to multi-energy scenarios, under a unified sparsity scale.

3. *Evaluation metrics*

To comprehensively quantify the reconstruction performance of different methods, we adopted an evaluation system covering four aspects: spatial resolution, localization accuracy, image quality, and computational efficiency. It should be noted that the core output of SACN is a continuous three-dimensional source field, whereas the outputs of SBP, MLEM, and U-Net are discrete reconstruction images. To ensure consistency in the evaluation representation, the reconstruction results of all methods were mapped to the same evaluation form. Specifically, for SACN, the predicted continuous three-dimensional source field was sampled on a fixed spatial grid and then projected by maximum response along the detector normal (z direction), generating a two-dimensional evaluation map consistent with those of the comparison methods. For the dual-source and three-source scenarios, local peak detection was first performed on the two-dimensional evaluation map, and the reconstructed peaks were then matched to the true source positions by the nearest-neighbor rule, from which the average PA was calculated. For the conventional methods and U-Net, their reconstruction maps were used directly. The subsequent FWHM, PA, and CNR were all computed on this unified evaluation representation.

The spatial-resolution metric, full width at half maximum (FWHM), was used to evaluate the spatial resolution of the reconstructed point source. It was obtained by fitting a two-dimensional Gaussian function to the source region in the reconstruction image:

$$G(x,y) = A\exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right] + B, \quad (52)$$

where (x_0, y_0) denotes the fitted peak position, σ_x and σ_y denote the standard deviations in the x and y directions, respectively, A is the peak amplitude, and B is the background baseline. The FWHM is then calculated as

$$\mathrm{FWHM}_x = 2.355\,\sigma_x, \qquad \mathrm{FWHM}_y = 2.355\,\sigma_y. \quad (53)$$

The final FWHM is defined as the geometric mean in the x and y directions,

$$\mathrm{FWHM} = \sqrt{\mathrm{FWHM}_x\,\mathrm{FWHM}_y}, \quad (54)$$

and a smaller FWHM indicates better spatial resolution. For an ideal point source, the FWHM reflects the intrinsic resolution limit of the imaging system.

The reconstruction-accuracy metric, position accuracy (PA), was used to evaluate the deviation between the reconstructed source position and the true position:

$$\mathrm{PA} = \sqrt{(x_{\mathrm{true}} - x_{\mathrm{recon}})^2 + (y_{\mathrm{true}} - y_{\mathrm{recon}})^2}, \quad (55)$$

where $(x_{\mathrm{true}}, y_{\mathrm{true}})$ is the true source position and $(x_{\mathrm{recon}}, y_{\mathrm{recon}})$ is the reconstructed position determined by Gaussian fitting. A smaller PA indicates more accurate source localization. Under sparse-data conditions, PA is a key metric for evaluating the practical utility of a method.

The image-quality metric, contrast-to-noise ratio (CNR), was used to evaluate the signal-to-noise ratio of the reconstruction image:

$$\mathrm{CNR} = \frac{I_{\mathrm{signal}} - I_{\mathrm{background}}}{\sigma_{\mathrm{background}}}, \quad (56)$$

where I_{signal} is the mean intensity within a circular source region centered at the reconstruction peak and having a radius of $1.5 \times$ FWHM, and $I_{\mathrm{background}}$ and $\sigma_{\mathrm{background}}$ are the mean intensity and standard deviation, respectively, of the background region located more than $3 \times$ FWHM away from the source center. A higher CNR indicates clearer source–background separation and is beneficial to subsequent source detection and quantitative analysis.

The computational-efficiency metric was the reconstruction time, defined as the total elapsed time from the input of Compton-event data to the output of the reconstruction image, including data preprocessing, network inference, and post-processing. For the deep learning methods, the reconstruction time includes only the inference stage and does not include the training time.

4. *Comparison methods*

Three representative methods were selected for performance comparison, covering three categories of reconstruction methods: analytical reconstruction, statistical iterative reconstruction, and deep learning reconstruction.

As a representative conventional analytical reconstruction method, the SBP algorithm is simple to compute but suffers from severe star-shaped artifacts in the reconstructed image. In the implementation, the cone surface corresponding to each Compton event was projected onto the reconstruction plane, and the reconstruction image was generated by superposing events. As a representative statistical iterative reconstruction method, MLEM was optimized based on a Poisson statistical model. The number of iterations was set to 80, and the convergence threshold was set to 10^{-6}. As a supervised

deep learning baseline, a U-Net architecture was adopted for image reconstruction. The network contained four encoder layers and four decoder layers. Each layer consisted of convolution, normalization, and nonlinear activation. Max pooling was used for downsampling, bilinear interpolation was used for upsampling, and skip connections were used to fuse multiscale features between the encoder and decoder. Unlike SBP and MLEM, the input of U-Net also followed the all-event-utilization strategy. In the implementation, each event was first encoded as a feature vector according to the unified event-encoding scheme proposed in this work, namely,

$$x_i = [f_i^{\mathrm{phys}},\, m_i,\, t_i], \tag{57}$$

and was then transformed into a regular feature-map representation processable by the network through a projection layer and tensorization operation. U-Net output a single-channel reconstruction image of size 256×256. U-Net was trained in an image-supervised manner, using the true source-distribution images as labels. Unlike SBP and MLEM, the input of U-Net was not the N valid events after strict screening, but the same set of all encodable raw events as that used by SACN, so that the effect of the all-event-utilization strategy could be compared within the deep learning paradigm.

B. Single-source reconstruction experiment

Single-source reconstruction is the most basic application scenario of the Compton camera and is also the basic experiment for evaluating localization accuracy, spatial resolution, and background-suppression ability under low-count conditions. In this section, a Cs-137 point source was used for single-source reconstruction, with the source position set to $(-80\ \mathrm{mm},\ 80\ \mathrm{mm},\ -100\ \mathrm{mm})$. Under different sparsity levels, SBP, MLEM, 3D-UNet, and the proposed SACN were used for reconstruction, and the evaluation metrics were computed. The results are listed in Table 1.

As shown in Table 1, under the current simulation setting, SACN achieved better overall performance at all sparsity levels, and its advantage became more obvious under extremely low-count conditions. When the number of valid events was only $N = 2$, SBP mainly showed the superposition of discrete cone trajectories and could hardly form a stable concentrated peak for localization. Although MLEM could focus the response through iterative updating, it was still prone to noise amplification and pseudo-peaks under severely underdetermined conditions. Under the same condition, U-Net was more likely to produce over-smoothed responses or response drift. In contrast, SACN was still able to recover a relatively concentrated single-peak response and showed more stable results in PA and CNR. As the number of valid events increased from $N = 2$ to $N = 10$, $N = 50$, and $N = 100$, the performance of all methods improved overall. However, SACN maintained a more balanced performance in FWHM, PA, and CNR, indicating that the proposed method has good adaptability to statistical fluctuations.

Figure 3 shows typical reconstruction results of different methods under different sparsity conditions. It can be seen that, under extremely low-count conditions, SBP mainly preserves the cone-superposition structure and cannot form a clear dominant peak; MLEM can produce a local high response in some samples, but is often accompanied by pseudo-peaks and background fluctuations; and U-Net is more likely to output blurred or shifted response distributions on sparse samples. In contrast, SACN can reconstruct a single-source response with more accurate position, more concentrated peak, and relatively cleaner background over the statistical range from $N = 2$ to $N = 100$. This result indicates that, under the current setting, the event-level physics-consistency training objective can provide stable constraints for sparse-data reconstruction.

This stability is related to the training mechanism of the proposed method. The optimization target of SACN is not to simply fit a fixed label image, but to require the predicted source distribution to better explain the geometric-consistency information in the observed events from the perspective of physics. Therefore, compared with U-Net, which mainly relies on image-supervised mapping, SACN shows better adaptability when the statistics vary substantially. Compared with MLEM, which relies on iterative updating, SACN is less susceptible to interference from local pseudo-peak structures under extremely low-count conditions.

In terms of computational efficiency, the inference-stage reconstruction time of SACN is on the same order as that of U-Net and is significantly shorter than that of MLEM, which requires multiple iterations. Overall, in the current single-source experiment, SACN achieves a good balance among reconstruction accuracy, image quality, and computational efficiency, which provides a basis for the subsequent analysis of

TABLE 1. Quantitative performance comparison for single-source reconstruction.

N	Method	FWHM	PA (mm)	CNR	RT (s)
2	SBP	1.91	19.52	1.06	2
2	MLEM	19.89	20.09	0.04	1096
2	U-Net	205.15	467.53	1.52	3
2	**SACN**	**13.65**	**4.89**	**53.27**	**2**
10	SBP	28.83	21.90	0.78	2
10	MLEM	18.26	19.48	7.91	2826
10	U-Net	57.19	10.99	1.59	3
10	**SACN**	**9.33**	**3.27**	**72.01**	**2**
50	SBP	93.53	32.53	0.74	3
50	MLEM	21.28	7.95	15.28	6861
50	U-Net	47.97	12.17	2.65	4
50	**SACN**	**7.57**	**2.82**	**67.21**	**3**
100	SBP	55.86	24.30	0.85	3
100	MLEM	22.53	8.45	15.93	8305
100	U-Net	42.13	9.15	2.41	5
100	**SACN**	**7.30**	**3.31**	**76.07**	**3**

more complex multi-source scenarios.

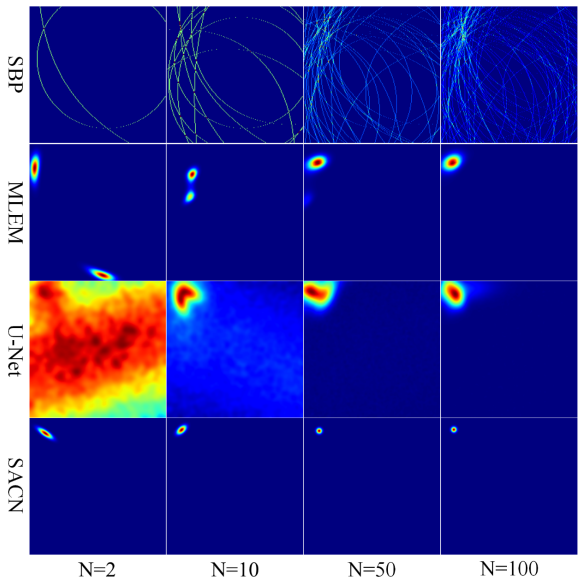

Fig. 3. Typical reconstruction results of different methods in the single-source experiment under different sparsity levels.

TABLE 2. Quantitative performance comparison for dual-source reconstruction.

N	Method	FWHM (mm)	PA (mm)	CNR
2	SBP	43.78	108.65	0.06
2	MLEM	16.32	12.85	37.19
2	U-Net	200.0	116.45	1.36
2	**SACN**	**15.62**	**4.60**	**45.89**
10	SBP	147.78	101.62	0.19
10	MLEM	16.17	8.46	42.22
10	U-Net	65.46	15.75	4.99
10	**SACN**	**10.06**	**3.61**	**52.91**
50	SBP	69.92	40.13	2.14
50	MLEM	15.93	17.32	4.70
50	U-Net	39.92	12.10	11.05
50	**SACN**	**7.32**	**3.23**	**56.78**
100	SBP	34.33	3.99	9.80
100	MLEM	10.00	3.88	4.67
100	U-Net	35.29	24.67	10.82
100	**SACN**	**6.93**	**3.13**	**61.52**

C. Dual-source reconstruction experiment

Multi-source separation ability is an important metric for evaluating the spatial resolution and practical value of a Compton camera. In the dual-source scenario, a limited number of Compton events must simultaneously constrain two spatially adjacent radiation sources, making the reconstruction problem more underdetermined than that in the single-source case. This is especially likely to cause peak merging, pseudo-peak proliferation, and localization deviation under low-count conditions. For this reason, this section constructs a dual-source reconstruction experiment consisting of a 511 keV annihilation radiation source at $(30\ \mathrm{mm}, 10\ \mathrm{mm}, -80\ \mathrm{mm})$ and a Co-60 source at $(40\ \mathrm{mm}, -10\ \mathrm{mm}, -80\ \mathrm{mm})$. The performance of each method in the multi-source separation scenario is evaluated, and the results are shown in Table 2 and Fig. 4.

From the quantitative results in Table 2, it can be seen that, under the current test setting, SACN shows good dual-source separation ability at different statistical levels. Under higher-statistics conditions, SACN can distinguish the two adjacent source peaks relatively accurately, and its PA is better than that of MLEM and U-Net. When the statistics further decrease to $N = 50$, $N = 10$, or even lower, MLEM and U-Net are more likely to exhibit source merging, dominant-peak deviation, or pseudo-peak interference. It should be noted that, under the extremely sparse condition ($N = 2$, that is, a total of 4 valid events in the dual-source scenario), SACN can still achieve a small localization error and preserve structural separation between the two sources. Although the cone surfaces corresponding to 4 events can, in ideal analytical geometry, barely intersect to form a discrete point solution, conventional methods still diverge easily and produce severe cross artifacts under real detection conditions with physical errors such as energy-resolution broadening. By integrating all-event information, including multiple-scattering and incomplete events, the proposed method provides additional implicit geometric constraints, thereby significantly improving spatial resolution and localization stability under underdetermined conditions. Recent related studies, such as ComptonNet, have also confirmed that all-event deep learning models that break the conventional screening criteria can achieve more accurate source-distribution estimation than conventional methods when dealing with extremely low-count statistics, such as single-digit events.

The visualization results in Fig. 4 further reflect the difference among methods in the dual-source scenario. SBP is strongly affected by star-shaped artifacts and cannot clearly resolve the dual-source structure. MLEM can form two high-response regions in some cases, but is often accompanied by additional pseudo-peaks and background fluctuations. U-Net can roughly indicate the regions of the two sources, but bridging responses or peak shifts are more likely to appear between the two sources. In contrast, the reconstruction results of SACN usually present two more independent peak structures whose positions are closer to the ground truth, indicating that it has a stronger advantage in suppressing source-to-source interference.

This phenomenon is related to the modeling mechanism of the proposed method. In the dual-source scenario, events generated by different sources are naturally mixed in the observation domain, while a single event usually provides only one

limited geometric constraint. Conventional methods are more likely to produce local pseudo-solutions under such mixed constraints. Pure image-supervised methods are more likely to learn an overly smooth average mapping when the training distribution is limited. SACN, in contrast, uses event-level physics-consistency constraints so that the predicted source field must remain consistent with multiple event geometries simultaneously. In this way, without presetting the number of sources, dual-source separation is achieved through the structural resolvability of the reconstructed field itself.

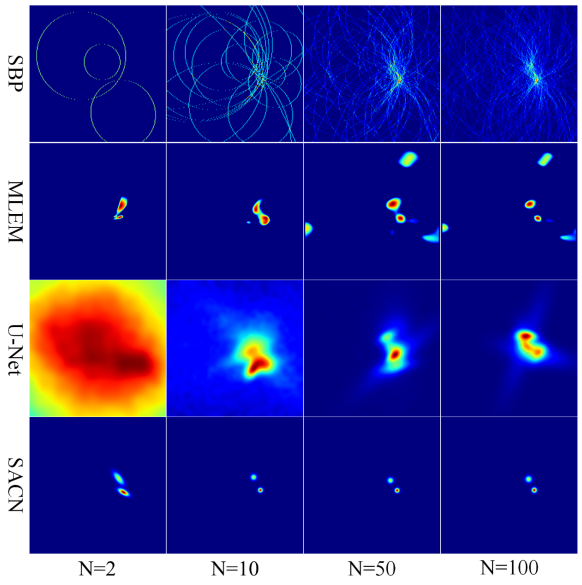

Fig. 4. Typical reconstruction results of different methods in the dual-source experiment under different sparsity levels.

D. Multi-energy three-source reconstruction experiment

Multi-energy multi-source reconstruction is one of the most challenging scenarios for practical Compton-camera applications, because the positions and structural responses of multiple sources with different energies must be recovered simultaneously under sparse-data conditions. In this section, a reconstruction experiment containing three sources with different energies is designed to simulate complex radiation fields in nuclear-medical imaging and radionuclide monitoring.

The experimental configuration includes three point sources: a 662 keV Cs-137 source located at $(10\ \mathrm{mm}, 10\ \mathrm{mm}, 30\ \mathrm{mm})$; a 511 keV positron-annihilation source (used to simulate a PET tracer) located at $(-20\ \mathrm{mm}, 30\ \mathrm{mm}, 30\ \mathrm{mm})$; and a Co-60 source located at $(20\ \mathrm{mm}, -10\ \mathrm{mm}, 30\ \mathrm{mm})$. For the Co-60 source, the G4RadioactiveDecay module was specifically enabled in Geant4 to faithfully simulate the physical process in which 1.173 MeV and 1.332 MeV cascade gamma photons are emitted simultaneously within an extremely short coincidence window. Faced with this type of complex coincidence event, which readily produces multiple hits (pile-up) and overlapping topological structures, the proposed SACN model robustly handles these multiple-scattering events through unified input-feature encoding, especially the event-type identifier and missing mask, thereby avoiding the information loss caused by conventional strict energy-window screening. This scenario covers the main energy ranges in typical Compton-camera applications and sets a geometrical layout that is challenging but theoretically separable.

TABLE 3. Quantitative performance comparison for multi-energy three-source reconstruction.

N	Method	FWHM (mm)	PA (mm)	CNR
2	SBP	67.01	76.52	0.38
2	MLEM	8.72	21.29	22.79
2	U-Net	200.0	72.76	1.08
2	**SACN**	**11.78**	**5.52**	**58.17**
10	SBP	165.83	86.86	0.41
10	MLEM	13.66	56.37	16.35
10	U-Net	66.61	17.46	5.41
10	**SACN**	**10.35**	**2.61**	**66.33**
50	SBP	69.64	27.45	2.33
50	MLEM	16.49	20.39	13.67
50	U-Net	36.58	19.91	20.46
50	**SACN**	**7.29**	**3.31**	**68.01**
100	SBP	63.86	4.34	5.16
100	MLEM	31.87	48.39	23.23
100	U-Net	45.72	18.25	12.98
100	**SACN**	**7.03**	**2.61**	**77.74**

Table 3 gives the quantitative results of different methods in the three-source scenario, and Fig. 5 shows the corresponding typical reconstruction visualizations. The results show that, under the multi-energy three-source condition, the performance of both SBP and MLEM further decreases compared with those in the single-source and dual-source scenarios. Because SBP directly superposes the cone trajectories corresponding to different events, it is more likely to produce complex background mixing and star-shaped artifacts. Although MLEM can form responses in some regions, it is more likely to show peak diffusion, additional pseudo-peaks, and mutual interference between adjacent true peaks under multi-source coupling conditions. In this scenario, U-Net can recover several high-response regions, but its description of the relative positional relationship among the three sources is not sufficiently accurate, and obvious structural fusion occurs between some sources.

In contrast, under the current setting, SACN can reconstruct three relatively independent response peaks whose positions are closer to the true values, and the overall background artifacts are also relatively weaker. The results in

Fig. 5 indicate that SACN is better than the other comparison methods in preserving peak structures in the multi-source scenario, especially in terms of inter-source separation and background suppression. It should be noted that the current evaluation still focuses on spatial-position recovery and structural separation ability. The quantitative recovery of source activity at different energies has not yet been studied separately. Therefore, the term *multi-energy* here mainly reflects the complexity of reconstruction under mixed event statistics rather than strict spectroscopic energy quantification.

It should be emphasized that the model in this work was trained only with single-energy Cs-137 point-source data, whereas Co-60, 511 keV annihilation radiation, and the three-source mixed scenario were not used in training. Therefore, the experiments in this section not only verify the separation ability of the method under multi-source conditions, but also further test its transfer performance under unseen energies and higher scene complexity. From the current results, SACN does not show obvious performance collapse when unseen energies and more sources are introduced into the test scenario. This indicates that the training strategy based on event-level physics consistency improves the adaptability of the model to complex scenarios to a certain extent.

Taken together, the single-source, dual-source, and multi-energy three-source experiments show that, under the current simulator, detector model, and test scenarios, SACN exhibits good reconstruction performance and structural stability. Compared with the conventional reconstruction methods based on strict screening, the all-event-utilization strategy combined with physical constraints shows clear potential. Compared with the purely image-supervised U-Net baseline, SACN shows better robustness and physical consistency under unseen energies and multi-source conditions.

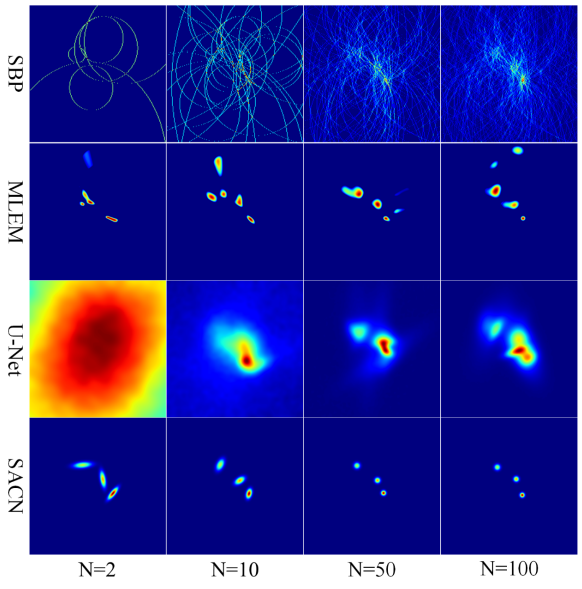

Fig. 5. Typical reconstruction results of different methods in the multi-energy three-source experiment under different sparsity levels.

IV. CONCLUSION

To address the problems of unstable reconstruction under low-count conditions, information loss caused by conventional strict event screening, and the dependence of image-supervised methods on labels in Compton cameras, this paper proposes SACN, a reconstruction framework based on event-level physical consistency. Rather than simply treating Compton reconstruction as an empirical mapping from measurement data to images, this paper formulates it as an inverse problem driven by a physical forward model: the source distribution output by the network should explain, as much as possible, the geometric consistency information contained in the observed events. Based on this idea, this paper embeds a differentiable Compton forward model into the training objective and achieves physically constrained training for sparse data without requiring image labels.

In terms of method implementation, this paper uses a conditional neural implicit field to represent the continuous three-dimensional source distribution, and uses an event-set encoder to uniformly represent all encodable raw events. At the same time, a geometric pre-localization mechanism is introduced to provide coarse guidance for continuous-space optimization, so as to alleviate the search difficulty under extremely sparse conditions. Furthermore, this paper imposes physical likelihood constraints, at the loss level, on events for which geometric constraints can be stably defined, while single-layer or incomplete events are mainly used for auxiliary representation learning. In this way, a unified modeling framework is formed, namely, full-event utilization at the representation level and selective physical constraints at the loss level.

The experimental results show that, under the current simulator, detector model, and test scenarios, the proposed method exhibits good overall performance in single-source, double-source, and multi-energy three-source tasks. Especially under low-count conditions, SACN is overall superior to the comparison methods selected in this paper in terms of localization accuracy, background suppression, and preservation of multi-source structure, indicating that event-level physical consistency constraints play a positive role in sparse Compton reconstruction. In addition, the model is trained only with single-energy point-source data of Cs-137, yet still maintains relatively stable performance in unseen isotopes and more complex multi-source scenarios, indicating that the method has a certain degree of cross-energy transfer capability under the current setting.

It should be pointed out that the conclusions of this paper are currently mainly based on simulation validation, and the main experiments focus on comparing two data-usage strategies, namely, “strict-screening reconstruction” and “full-event-utilization reconstruction”, rather than a completely fair comparison of algorithm bodies under exactly the same input conditions. Meanwhile, this paper has not yet explicitly established a unified mixed-observation model for random coincidence, complex background, and stronger non-ideal detector effects, and has not carried out systematic research on spectroscopic quantitative recovery in multi-energy scenar-

ios.

Overall, this work shows that, in the problem of sparse Compton imaging, shifting reconstruction from a supervised learning paradigm that depends on image labels to a label-free optimization paradigm driven by a physical forward model is a promising research path. Future work will further focus on real experimental data validation, matched-input controlled comparison, explicit modeling of complex background events, reconstruction of extended-source scenarios, and joint correction of non-ideal detector response effects, so as to promote the practical application of this method in nuclear medicine imaging and radiation source detection.

V. REFERENCES

[1] Kim SM, Lee JS. A comprehensive review on Compton camera image reconstruction: from principles to AI innovations. Biomedical Engineering Letters. 2024;14(6):1175-1193. doi:10.1007/s13534-024-00418-8.

[2] Parajuli RK, Sakai M, Parajuli R, Tashiro M. Development and Applications of Compton Camera—A Review. Sensors (Basel). 2022;22(19):7374. doi:10.3390/s22197374.

[3] Llosá G, Rafecas M. Hybrid PET/Compton-camera imaging: an imager for the next generation. European Physical Journal Plus. 2023;138(3):214. doi:10.1140/epjp/s13360-023-03805-9.

[4] Yamaya T, Tashima H, Takyu S, Takahashi M. Whole Gamma Imaging: Challenges and Opportunities. PET Clinics. 2024;19(1):83-93. doi:10.1016/j.cpet.2023.08.003.

[5] Sakai M, et al. Experimental study on Compton camera for boron neutron capture therapy applications. Scientific Reports. 2023;13(1):22883. doi:10.1038/s41598-023-49955-9.

[6] Torres-Sánchez P, et al. The potential of the i-TED Compton camera array for real-time boron imaging and determination during treatments in Boron Neutron Capture Therapy. Applied Radiation and Isotopes. 2025;217:111649. doi:10.1016/j.apradiso.2024.111649.

[7] Tsukamoto H, et al. Development of an omnidirectional rotating Compton camera for imaging 177Lu radioactive contamination. PLOS ONE. 2025;20(6):e0325586. doi:10.1371/journal.pone.0325586.

[8] Kim D, Yan L, Shimazoe K, Takahashi H, Ogane K, Yoshino M, Kamada K, Uenomachi M. Demonstration of in-vivo simultaneous 3D imaging with 18F-FDG and Na131I using Compton–PET system. Scientific Reports. 2024;14(1):20946. doi:10.1038/s41598-024-71750-3.

[9] Wu C, Zhang S, Li L. First Demonstration of Compton Camera Used for X-Ray Fluorescence Imaging. IEEE Transactions on Medical Imaging. 2023;42(5):1314-1324. doi:10.1109/TMI.2022.3226329.

[10] Wu C, Zhang S, Li L. An accurate probabilistic model with detector resolution and Doppler broadening correction in list-mode MLEM reconstruction for Compton camera. Physics in Medicine & Biology. 2022;67(12):125017. doi:10.1088/1361-6560/ac73d2.

[11] Barrientos L, et al. Gamma-ray sources imaging and test-beam results with MACACO III Compton camera. Physica Medica. 2024;117:103199. doi:10.1016/j.ejmp.2023.103199

[12] Daniel G, Gutierrez Y, Limousin O. Application of a deep learning algorithm to Compton imaging of radioactive point sources with a single planar CdTe pixelated detector. Nuclear Engineering and Technology. 2022;54(5):1747-1753. doi:10.1016/j.net.2021.10.031.

[13] Yao Z, Shi C, Tian F, Xiao Y, Geng C, Tang X. Technical note: Rapid and high-resolution deep learning-based radiopharmaceutical imaging with 3D-CZT Compton camera and sparse projection data. Medical Physics. 2022;49(11):7336-7346. doi:10.1002/mp.15898.

[14] Kazemi Kozani M, Magiera A. Machine learning-based event recognition in SiFi Compton camera imaging for proton therapy monitoring. Physics in Medicine & Biology. 2022;67(15):155012. doi:10.1088/1361-6560/ac71f2.

[15] Muñoz E, Ros A, Borja-Lloret M, Barrio J, Dendooven P, Oliver JF, Ozoemelam I, Roser J, Llosá G. Proton range verification with MACACO II Compton camera enhanced by a neural network for event selection. Scientific Reports. 2021;11(1):9325. doi:10.1038/s41598-021-88812-5.

[16] Pérez-Curbelo J, Roser J, Muñoz E, Barrientos L, Sanz V, Llosá G. Improving Compton camera imaging of multi-energy radioactive sources by using machine learning algorithms for event selection. Radiation Physics and Chemistry. 2025;226:112166. doi:10.1016/j.radphyschem.2024.112166.

[17] Sato S, Tanaka KS, Kataoka J. ComptonNet: A direct reconstruction model for Compton camera. Applied Physics Letters. 2024;124(25):253702. doi:10.1063/5.0213950.

[18] Long Z, Jiang X. A Physics-Constrained Deep Learning Method for Compton Cameras 3-D Imaging. IEEE Transactions on Nuclear Science. 2026. doi:10.1109/TNS.2025.3624815.

[19] Ikeda T, Takada A, Abe M, Yoshikawa K, Tsuda M, Ogio S, Sonoda S, Mizumura Y, Yoshida Y, Tanimori T. Development of convolutional neural networks for an electron-tracking Compton camera. Progress of Theoretical and Experimental Physics. 2021;2021(8):083F01. doi:10.1093/ptep/ptab091

[20] Zhu Y, Liu Y, Zhang Y, Liang D. Implicit neural representation for medical image reconstruction. Physics in Medicine & Biology. 2025;70(12):12TR01. doi:10.1088/1361-6560/addfa5.

[21] Shen L, Pauly J, Xing L. NeRP: Implicit Neural Representation Learning With Prior Embedding for Sparsely Sampled Image Reconstruction. IEEE Transactions on Neural Networks and Learning Systems. 2024;35(1):770-782. doi:10.1109/TNNLS.2022.3177134.

[22] Lee J, Baek J. Iterative reconstruction for limited-angle CT using implicit neural representation. Physics in Medicine & Biology. 2024;69(10):105008. doi:10.1088/1361-6560/ad3c8e.

[23] He Y, Ruan D. An implicit neural deformable ray model for limited and sparse view-based spatiotemporal reconstruction. Medical Physics. 2025;52(6):3959-3969. doi:10.1002/mp.17714.

[24] Wang Y, Liang N, Wang S, Guo J, Zhang X, Zheng Z, Cai A, Li L, Yan B. Implicit neural prior-guided diffusion for spectral CT reconstruction. Medical Physics. 2025;52(7):e17946. doi:10.1002/mp.17946.

[25] Yu M, Ahn J, Baek J. Continuous representation-based reconstruction for computed tomography. Medical Physics.

2025;52(7):e17849. doi:10.1002/mp.17849.