



OPEN Unsupervised disparity-tolerant algorithm for terahertz image stitching

Xiaojin Wu^{1,2}, Fan Bai³, Lun Li^{1,2}✉, Yuan Gao^{1,2}, Wencheng Wang^{1,2} & Hongfu Cai⁴

Terahertz imaging offers significant potential in areas such as non-destructive testing, security screening, and medical diagnostics. However, due to the immature development of terahertz imaging devices, the field of view remains limited, making it challenging to capture complete target information in a single acquisition. While image stitching techniques can effectively expand the field of view, traditional methods encounter substantial limitations when applied to terahertz images, including low resolution, limited texture features, and inconsistencies arising from parallax. To address these challenges, particularly the parallax inconsistencies in low-resolution terahertz image stitching, we propose an Unsupervised Disparity-Tolerant Terahertz Image Stitching algorithm (UDTATIS). Our approach introduces targeted optimizations for two critical stages: geometric distortion correction and image feature fusion. Specifically, we design a feature extractor and an effective point discrimination mechanism based on the EfficientLOFTR architecture, significantly enhancing feature matching accuracy and robustness. Additionally, we introduce a continuity constraint to ensure the spatial continuity of matched points, thereby mitigating geometric distortions. Furthermore, we develop an improved conditional diffusion model that integrates multi-scale feature fusion with adaptive normalization, refining the transition effects along stitching boundaries. Compared to existing methods, UDTATIS demonstrates superior performance in handling terahertz images characterized by low resolution, limited textures, and parallax, achieving seamless image fusion while maintaining geometric consistency. Extensive quantitative and qualitative evaluations validate that UDTATIS outperforms state-of-the-art stitching algorithms, especially in complex scenes, delivering enhanced visual coherence and structural integrity. Project page: <https://github.com/snow-wind-001/UDTATIS>.

Keywords Terahertz imaging, Image stitching, Unsupervised learning, Diffusion model, Disparity tolerance

Terahertz imaging technology exhibits significant potential in applications such as non-destructive testing, security screening, and medical diagnostics, due to its ability to penetrate various non-conductive materials without generating ionizing radiation¹. However, terahertz imaging devices are typically constrained by a relatively narrow field of view, limiting their effectiveness in large-scale inspection scenarios. Image stitching provides a viable solution by integrating multiple local images into a single panoramic view, thereby extending the operational range of terahertz imaging systems. Despite this, conventional stitching algorithms face considerable challenges when applied to terahertz images, including difficulties in feature extraction, inaccurate image registration, and suboptimal fusion quality^{2–4}.

Challenges in calibration

Due to the unique wavelength characteristics of terahertz imaging, the images are subject to optical and perspective distortions, which can be even more pronounced in terahertz multi-view (compound-eye) imaging. Moreover, the intrinsic and extrinsic parameters of terahertz imaging devices may differ significantly from those of conventional cameras, rendering standard camera calibration techniques ineffective. Therefore, specialized calibration methods tailored to the characteristics of terahertz imaging are required⁵.

¹Institute of Machinery and Automation, Weifang University, Weifang 261061, China. ²Shandong Key Laboratory of Intelligent Manufacturing Technology for Advanced Power Equipment, Weifang, China. ³College of Equipment Engineering, Shenyang Ligong University, Shenyang 110159, China. ⁴Shandong Hongke Hydropower Equipment Co., LTD, Weifang 261021, China. ✉email: ll408907652@163.com

Complexity in registration

Terahertz images exhibit distinctive penetration and reflection properties that differ markedly from those of visible-light images. As a result, feature-based registration methods may struggle to find sufficient stable keypoints for effective alignment, and direct methods may also face difficulties due to the degraded image quality, as shown in Fig. 1. Consequently, more sophisticated algorithms must be developed to address the unique challenges posed by terahertz images⁶.

Difficulties in fusion

In the stitching of terahertz images, the imaging characteristics—such as transparency to certain materials—can cause unusual effects in boundary regions, leading to challenges that traditional image fusion methods, such as alpha blending or Gaussian pyramid blending, cannot easily overcome. Furthermore, although graph-cut-based optimization techniques can identify ideal seams, they may still encounter difficulties when handling terahertz images with complex textures or varying transparency⁷.

Limitations of traditional stitching techniques

For terahertz multi-view images, traditional spatial-domain and frequency-domain stitching techniques may not be applicable due to the significant differences between terahertz and conventional optical images. In particular, region-based methods may struggle with non-uniform illumination and material transparency variations inherent in terahertz images, while feature-based methods may fail to detect sufficient reliable keypoints⁸.

Remaining challenges for THz-oriented deep learning methods

Although CNN- and Transformer-based stitching frameworks have greatly advanced visible-light panoramas, their direct transfer to terahertz data is hindered by three factors. First, most networks presume abundant texture and radiometric consistency, whereas THz images exhibit weak edges, frequency-dependent attenuation, and specular reflections. Second, public training sets for THz imaging are extremely limited; existing models therefore depend on supervised pairs or photometric homogeneity assumptions that do not hold in the 350 GHz band. Third, large-parallax scenes remain problematic because homography-only decoders cannot represent the compound refractive effects of multilayer materials. These gaps motivate us to design an unsupervised, disparity-tolerant pipeline that explicitly filters unreliable matches, enforces spatial continuity, and leverages diffusion-based fusion to restore low-SNR boundaries.

Recent efforts to alleviate these issues include a pyramid self-attention stitcher for hyperspectral microscopy⁹ and a GAN-guided THz mosaic network¹⁰. Nevertheless, both approaches still rely on paired supervision and are evaluated on centimetre-scale samples rather than metre-scale inspection targets. By contrast, our UDTATIS operates without ground-truth mosaics and is validated on large-object THz scans, demonstrating superior robustness under severe texture scarcity and multipath reflections.

In recent years, deep learning-based image stitching methods have achieved remarkable progress; however, most approaches rely on supervised learning, requiring large amounts of paired training data^{11,12}, or struggle to effectively handle scenes with significant parallax¹³. To address the unique characteristics of terahertz images, this paper proposes an Unsupervised Disparity-Tolerant Terahertz Image Stitching algorithm (UDTATIS). UDTATIS integrates the UDIS++¹⁴ framework with the feature extraction and matching capabilities of EfficientLOFTR¹⁵, while introducing several innovations, including effective point discrimination, continuity constraints, and a diffusion model, to further enhance stitching quality.

The UDTATIS algorithm adopts a two-stage design to separately address geometric alignment (Warp) and image fusion (Composition). During the geometric alignment stage, the algorithm employs an improved feature extraction and matching strategy, combined with an effective point discrimination mechanism and continuity constraints, to achieve accurate geometric registration. In the image fusion stage, an enhanced diffusion model, together with multi-scale feature fusion and adaptive normalization techniques, is introduced to optimize the fusion effect, ensuring that the final stitched image maintains geometric consistency while achieving seamless visual transitions.

Terahertz image data acquisition

This study conducts data acquisition using a self-developed terahertz imaging experimental platform. The constructed experimental system is shown in Fig. 2. The terahertz image acquisition system integrates transmission-mode terahertz imaging technology with a rotational testing approach to perform rotational

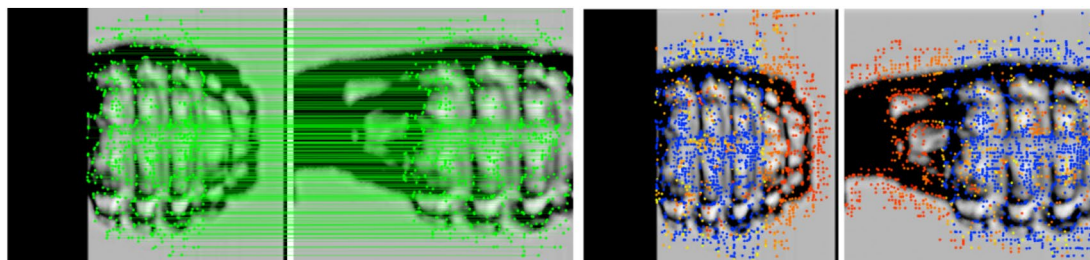


Fig. 1. Complexity of terahertz image registration.

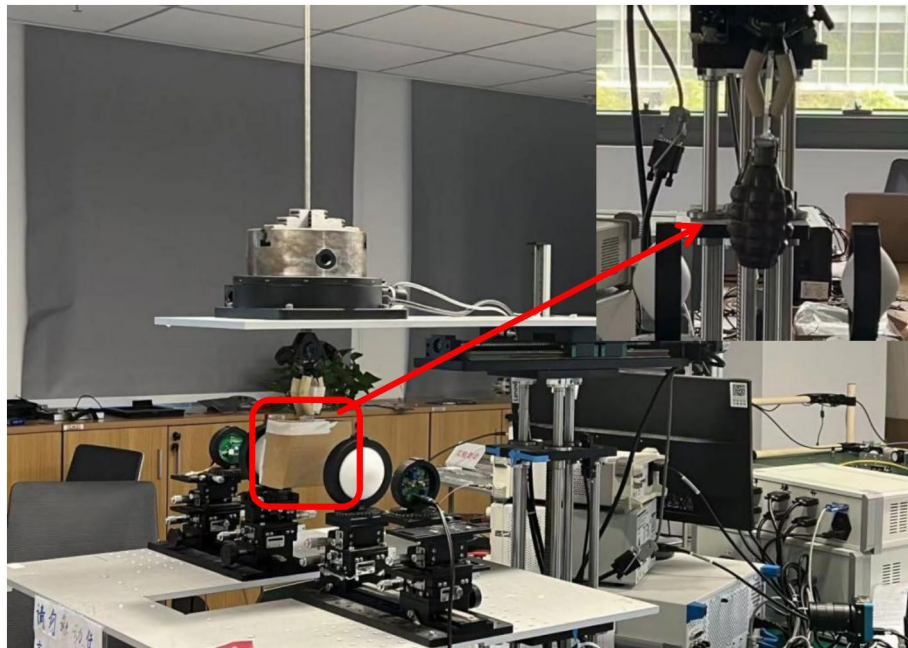


Fig. 2. Terahertz data acquisition experimental system.

imaging analysis of the experimental target (a grenade model), aiming to obtain terahertz images of the target from various angles. Due to the limitations of current terahertz imaging devices, existing focal-plane array (FPA) based terahertz imaging systems are unable to produce high-resolution terahertz images and typically require high-power terahertz sources. Therefore, to achieve higher-quality terahertz images, a point-scanning terahertz imaging system was constructed in this study. However, acquiring a complete image of the target with a point-scanning system often requires a prolonged scanning time and is susceptible to interruptions caused by environmental disturbances. To obtain complete terahertz images under these constraints, terahertz image stitching techniques are employed.

The imaging experimental system employs a scanning lens and a single-pixel detector to perform two-dimensional imaging of the target object (a grenade model). Due to the current immaturity of focal-plane array (FPA) detector technology at 350 GHz, achieving high-resolution and high-dynamic-range imaging remains a significant challenge. Therefore, a single-pixel detector combined with a point-scanning approach is adopted in this study. Although this scanning mechanism has limitations in terms of time efficiency, it offers superior sensitivity and dynamic range compared to existing FPA detectors, making it more suitable for acquiring high-quality terahertz images. In future practical applications, we plan to optimize scanning paths and detection strategies to improve acquisition efficiency.

The scanning lens comprises two controllable mirrors that adjust the direction and position of the terahertz beam. The single-pixel detector integrates a terahertz sensor with a signal processing circuit, enabling reception of signals from different positions.

During the experiment, a CMOS-fabricated chip was used as the terahertz emitter to generate pulses centered at 350 GHz, with an average output power of approximately 0.1 mW. A dedicated lens system, consisting of two vertically rotatable convex lenses with focal lengths of 50 mm (front) and 65 mm (rear), and a diameter of 78.2 mm, was used to focus and expand the terahertz beam. The terahertz receiver was capable of detecting pulses in the 350 GHz frequency band.

Under these experimental conditions, each 256×256 -pixel terahertz image required approximately 6 h to acquire using the constructed single-pixel scanning platform. Image stitching computations were performed on an NVIDIA GeForce RTX 3090 GPU, with a total processing time of approximately 2.5 s per stitching task. An example of the acquired terahertz image of the grenade model is shown in Fig. 3.

System architecture of the unsupervised disparity-tolerant terahertz image stitching algorithm (UDTATIS)

The stitching process of the UDTATIS algorithm is designed in two stages, targeting geometric distortion correction and image feature fusion for low-resolution terahertz images. The overall system architecture is illustrated in Fig. 4. This modular design enables the system to flexibly handle different types of image stitching tasks, and, in particular, allows it to achieve seamless visual effects while maintaining geometric consistency when processing terahertz images with significant disparity.

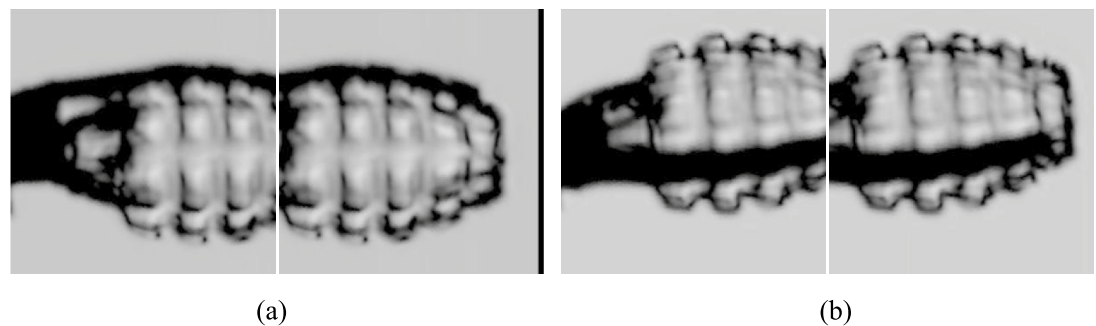


Fig. 3. Partial terahertz images of the acquired target object (grenade model): (a) 45°; (b) 90°.

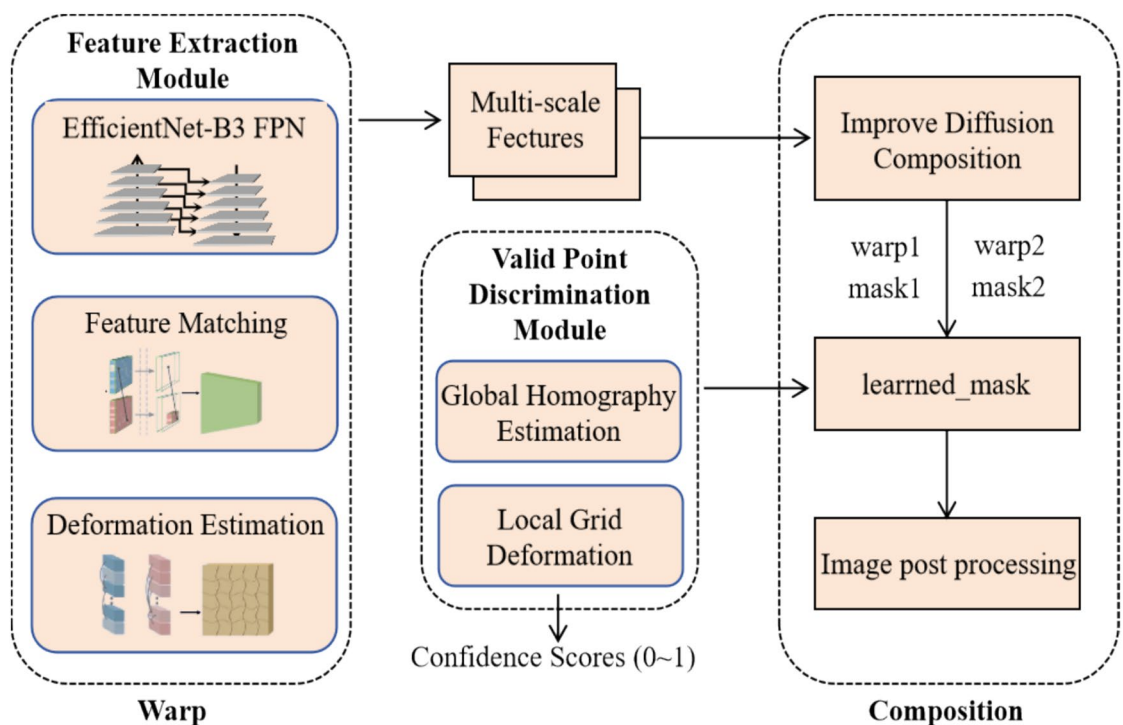


Fig. 4. Overall architecture of the UDTATIS algorithm.

Image stitching process

Geometric distortion correction

To obtain high-quality stitched images, geometric distortion correction must be performed on the input images prior to stitching. Considering the characteristics of terahertz images—such as low resolution, limited texture features, and disparity inconsistencies—we improve the traditional feature extraction and image matching methods used in conventional image stitching algorithms.

The architecture of the improved geometric distortion correction network is illustrated in Fig. 5. This stage consists of four key modules: feature extraction, feature matching, effective point discrimination, and deformation estimation.

Feature extraction

The feature extraction module employs EfficientNet as the backbone network, combined with a Feature Pyramid Network (FPN) to perform multi-scale feature extraction. The mathematical formulation of the feature extraction network is shown in Eq. (1). This design enables the capture of hierarchical features in terahertz images, ranging from low-level texture details to high-level semantic information. Specifically, we adopt EfficientNet-B3 as the feature extractor to generate five feature maps at different scales, corresponding to 1/4, 1/8, 1/16, 1/32, and 1/64 of the original image size, respectively

$$F_i^l = \Phi_l(I_i), i \in \{1, 2\}, l \in \{1, 2, 3, 4, 5\} \quad (1)$$

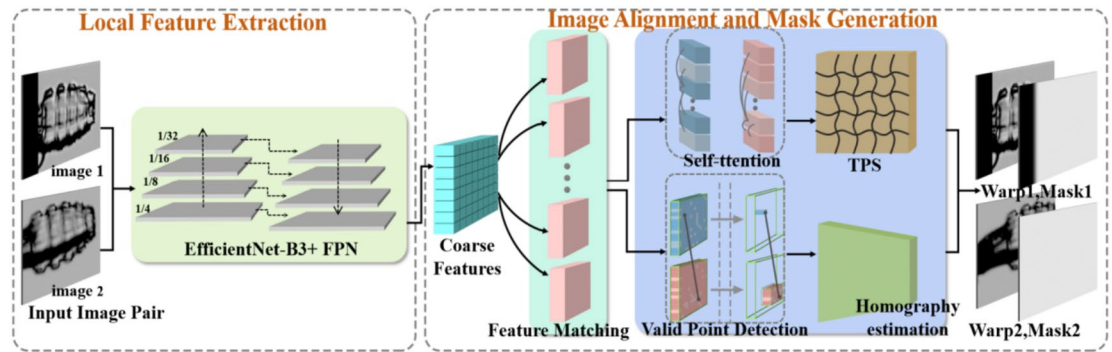


Fig. 5. Network architecture of the warp stage.

here I_i denotes the input image, F_i^l represents the l -th layer feature map extracted from the i -th image, and Φ_l denotes the mapping function of the feature extraction network at the l -th layer.

Feature matching

The feature matching module innovatively adopts an improved Transformer architecture to achieve global context-aware feature matching, overcoming the performance bottlenecks of traditional local descriptor-based methods in low-texture terahertz image scenarios. This module establishes accurate feature correspondences through a three-stage progressive matching strategy: first, in the coarse matching stage, high-dimensional feature maps are projected into a low-dimensional latent space using linear projection. An initial matching hypothesis is then generated by computing a cross-correlation matrix to identify high-response candidate matching pairs, mathematically formulated as follows:

$$C_{ij} = \{\Phi(I_1)_i, \Phi(I_2)_j\} \quad (2)$$

here ϕ denotes the projection function, and i and j represent the indices of feature points.

Subsequently, in the self-attention enhancement stage, a multi-head self-attention mechanism is employed to model long-range dependencies within the feature maps, effectively enhancing the global consistency of feature representations. The computation process can be formulated as follows:

$$\hat{F} = \text{SoftMax}(QK^T/\sqrt{d})V \quad (3)$$

here Q , K , and V represent the query, key, and value matrices of the same image features, respectively.

Finally, in the cross-attention matching stage, a dual-stream cross-attention network is constructed to compute the feature similarity between image pairs. The matching process can be mathematically formulated as follows:

$$M = \text{SoftMax}(Q_1K_2^T/\sqrt{d})V \quad (4)$$

here Q_1 and K_2 represent the query and key matrices of the first and second image features, respectively, d denotes the feature dimension, and M is the matching matrix.

Valid point determination

Due to the uneven distribution of features in terahertz images, some computed matching points may fall in regions with poor texture or be significantly affected by parallax, leading to matching errors. To address this, the warp module is specifically designed with a valid point determination mechanism, enabling the automatic identification and filtering of unreliable matching points.

The valid point determination module consists of a lightweight convolutional neural network, which takes the local feature maps around the matching points as input and outputs a confidence score indicating the reliability of each matching point. For each matching point pair (p_1, p_2) , $k \times k$ feature patches centered at p_1 and p_2 are extracted and concatenated as the input to the determination network. The computation of valid point determination is formulated as follows:

$$s(p_1, p_2) = \Psi(F_1(p_1), F_2(p_2)) \quad (5)$$

here $s(p_1, p_2)$ denotes the confidence score of the matching point pair (p_1, p_2) , $F_i(p_i)$ represents the feature patch around point p_i in the i -th image, and Ψ denotes the determination network.

During training, for each matching point pair in the terahertz images, the RANSAC algorithm is used to estimate the homography matrix H , and the transformation error is then computed (as shown in Eq. 6). The homography transformation error serves as the supervisory signal to generate training labels.

$$e(p_1, p_2) = \|p_1 - Hp_2\|_2 \quad (6)$$

If $s(p_1, p_2) < \tau$, where τ is a predefined threshold, the matching point pair is labeled as a valid point (Label 1); otherwise, it is labeled as an invalid point (Label 0).

Continuity constraint and deformation estimation

In terahertz images, due to the presence of parallax, the matching relationships between adjacent regions may become inconsistent, leading to geometric distortions in the stitching results. The continuity constraint mitigates this issue by restricting the variation between neighboring matching points, thereby ensuring the smoothness of the deformation field. The continuity constraint is computed based on the gradient of the feature map, and its mathematical formulation is given as follows:

$$L_{\text{cont}} = \sum_p \in V \|\nabla F_1(p) - \nabla F_2(M(p))\|_1 \quad (7)$$

here V denotes the set of valid points, ∇F_i represents the gradient of the feature map F_i , and $M(p)$ denotes the matching point of p in the second image. By applying this constraint, the gradient variation around the matching points on the feature maps is enforced to remain consistent, thereby promoting spatial continuity.

Due to the immaturity of terahertz imaging device development, the field of view of a single terahertz imaging system is currently limited, making it difficult to capture the entire target object within a single image. Therefore, terahertz imaging experiments are typically conducted by acquiring images from multiple angles around the target. As a result of this multi-view imaging, the captured terahertz images often exhibit deformation.

To address this, a deformation estimation module is incorporated into the design of the image processing network. This module consists of two components: global homography estimation and local grid deformation.

The global homography estimation employs a regression network to directly predict eight homography parameters, providing a foundation for global alignment. The local grid deformation, on the other hand, uses Thin Plate Spline (TPS) interpolation to generate a deformation field based on the matched point pairs, thereby handling local deformations. Specifically, the local grid deformation divides the space into a regular grid and computes the displacement of each grid vertex according to the matching points. The final deformation field is a combination of the global homography transformation and the local grid deformation. The global homography matrix H and the final deformation field function are defined as follows:

$$\begin{cases} H = \Omega(F_1, F_2, M) \\ W(p) = T(H(p) + G(p)) \end{cases} \quad (8)$$

where Ω denotes the homography regression network, which takes the feature maps and the matching matrix as inputs and outputs the homography transformation matrix; $W(p)$ represents the final deformed position of point p ; $H(p)$ denotes the position of p after the homography transformation; $G(p)$ represents the displacement generated by the grid deformation; and T is a constraint function that ensures the deformation results remain within the image boundaries.

Loss function for the warp stage

The training of the Warp stage adopts an unsupervised approach, requiring no paired training data. The loss function consists of multiple components that jointly optimize the network parameters. The overall loss function is defined as follows:

$$L_{W\text{-total}} = \lambda_{\text{homo}} L_{\text{homo}} + \lambda_{\text{mesh}} L_{\text{mesh}} + \lambda_{\text{feat}} L_{\text{feat}} + \lambda_{\text{valid}} L_{\text{valid}} + \lambda_{\text{cont}} L_{\text{cont}} \quad (9)$$

where λ_{homo} , λ_{mesh} , λ_{feat} , λ_{valid} , and λ_{cont} are the weighting coefficients for the homography loss, mesh deformation loss, feature matching loss, valid point discrimination loss, and continuity loss, respectively, used to balance the contributions of each loss component. During training, the default settings adopted in this work are $\lambda_{\text{homo}} = 1.0$, $\lambda_{\text{mesh}} = 1.0$, $\lambda_{\text{feat}} = 0.1$, $\lambda_{\text{valid}} = 0.5$ and $\lambda_{\text{cont}} = 0.2$.

The loss functions for each component are defined as follows:

$$\begin{cases} L_{\text{feat}} = \frac{-\sum_{(p_1, p_2) \in P} \log(e^{(F_1(p_1) \cdot F_2(p_2)) / \tau})}{\sum_{q \in (p_2)} e^{(F_1(p_1) \cdot F_2(p_2)) / \tau}} \\ L_{\text{homo}} = \|F_2 - \text{Warp}(F_1, H)\|_1 \\ L_{\text{mesh}} = \|F_2 - \text{Warp}(F_1, W)\|_1 \\ L_{\text{valid}} = -\sum_{(p_1, p_2) \in P} (y \cdot \log(s(p_1, p_2)) + (1 - y) \cdot \log(1 - s(p_1, p_2))) \\ L_{\text{cont}} = \sum_{p \in V} \|\nabla F_1(p) - \nabla F_2(M(p))\| \end{cases} \quad (10)$$

here P denotes the set of initially matched point pairs, $N(p_2)$ represents the neighborhood of point p_2 , and τ is a temperature parameter that controls the smoothness of the distribution; $\text{Warp}(F_1, H)$ denotes the warping operation of feature map F_1 using the homography matrix H ; $\text{Warp}(F_1, W)$ denotes the warping operation of feature map F_1 using the deformation field W ; y represents the ground-truth label (0 or 1) of the matching point pair, and $s(p_1, p_2)$ represents the confidence score predicted by the discriminative network; V represents the set of valid points, and restricting the computation of continuity constraints to these regions helps avoid enforcing unnecessary continuity in textureless areas.

Through the above process, the feature point alignment problem in terahertz images can be effectively addressed, especially in scenarios with limited texture features and parallax. By comprehensively considering both global and local deformations, and incorporating valid point discrimination and continuity constraints, UDTATIS achieves more accurate geometric alignment, providing a solid foundation for subsequent image fusion.

Image composition and fusion

The image composition and fusion stage (Composition stage) achieves seamless fusion of geometrically aligned images through an improved diffusion model, specifically addressing challenges such as low resolution, low signal-to-noise ratio, and sparse textures in terahertz images. This stage adopts a multi-scale U-Net architecture that integrates self-attention and cross-attention mechanisms to enhance global context modeling capability, and introduces time encoding enhancement and adaptive normalization layers to optimize noise sensitivity and training stability.

The fusion process employs a multi-constraint loss function, including noise prediction, boundary smoothing, mask gradient constraint, perceptual consistency, and multi-scale evaluation, to ensure both detail preservation and natural overall transitions.

Through a mask-guided progressive fusion strategy, the model first generates an initial coarse stitching result and then progressively refines the details, dynamically adjusting loss weights to balance the refinement of global structures and local boundaries, ultimately achieving high-quality stitching results without artifacts.

The network structure of the Composition stage is shown in Fig. 6.

Mask initialization and update strategy

The fusion mask is not randomly initialised. Instead, we compute a data-driven, confidence-aware initial mask M_0 using a lightweight convolutional module (MaskNet) fed with two cues that capture complementary reliability information:

- (i) an absolute intensity-difference map, $D = |\text{warp}_1 - \text{warp}_2|$, which highlights local disagreement and potential seam locations;
- (ii) mid-level semantic features extracted from the penultimate encoder block of the Composition U-Net, $\Phi(\text{warp}_1)$ and $\Phi(\text{warp}_2)$, concatenated channel-wise.

MaskNet outputs a soft confidence field $\tilde{M}_0 \in [0, 1]$. To generate an initial binary paste prior P_0 used in Eq. (16), we apply adaptive thresholding: τ is set to the 85th percentile of \tilde{M}_0 within the overlapping region, and

$$P_0(x, y) = 1 \{ \tilde{M}_0(x, y) \geq \tau \} \quad (11)$$

Morphological opening (3×3) removes isolated noisy pixels. This produces a spatially coherent region from which content in warp_2 may replace that in warp_1 .

During training, the mask is jointly optimised with image synthesis inside the diffusion-denoising loop. Let M_k denote the current soft mask estimate at diffusion timestep k (reverse process). A gradient-based update refines it:

$$M_{k+1} = \text{Clip}_{[0,1]} \left(M_k - \eta_k \frac{\partial L_{\text{total}}}{\partial M_k} \right), \quad k = 0, \dots, T-1 \quad (12)$$

where η_k follows a cosine decay from $\eta_0 = 5 \times 10^{-3}$ to $\eta_T = 5 \times 10^{-4}$ to stabilise late refinements.

Improved diffusion model architecture

The improved diffusion model employs a multi-stage progressive optimization mechanism to effectively address issues such as blurred stitching boundaries, loss of details, and prominent artifacts caused by inaccurate feature

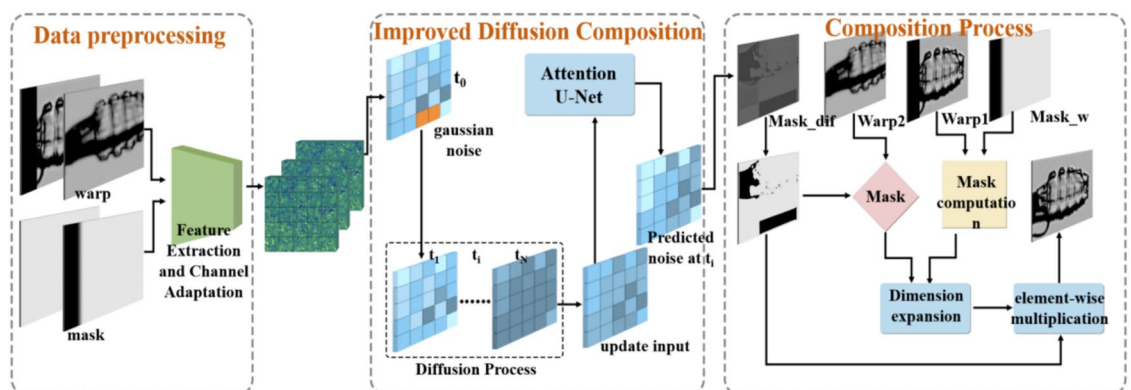


Fig. 6. Network structure of the composition stage.

matching and limited fusion strategies in conventional methods for terahertz images. Its multi-scale U-Net architecture, encompassing both the forward diffusion and reverse denoising processes, can be mathematically described as:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{(1-\beta_t)x_{t-1}}, \beta_t I\right) \quad (13)$$

$$q(x_t|x_0) = N\left(x_t; \sqrt{(\bar{\alpha}_t)x_0}, (1-\bar{\alpha}_t)I\right) \quad (14)$$

where x_0 denotes the original image, x_t denotes the image after adding noise for t steps, β_t represents the noise variance at step t , $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$.

Meanwhile, by incorporating the self-attention mechanism (as shown in Eq. 15), the model can extract global contextual information from low signal-to-noise ratio images, enhancing semantic consistency in texture-sparse regions

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t; t), \sum_\theta(x_t, t)\right) \quad (15)$$

here θ denotes the model parameters, while μ_θ and Σ_θ represent the predicted mean and variance, respectively.

The adaptive normalization layer dynamically adjusts the feature distribution according to the timestep, enhancing the model's robustness to noise and preventing fusion distortions caused by the inherent noise in terahertz images. In addition, the sinusoidal positional encoding strengthens the temporal correlation of timesteps, making the denoising process better aligned with the progressive optimization requirements of image stitching, ultimately producing smooth transitions in boundary regions and reducing noticeable stitching artifacts without manual intervention.

Forward propagation process

In the algorithm proposed in this study, an improved diffusion model processes the image stitching task through a mask-guided progressive blending strategy. Specifically, a pair of warped input images (warp1, warp2) along with their corresponding masks (mask1, mask2) are first provided. Subsequently, a region of size 256×256 is extracted from the left side of the predicted mask generated by the composition network, denoted as mask_out. This extracted region is then binarized to obtain an explicit foreground mask for subsequent processing.

$$P(x, y) = \begin{cases} 1, & (x, y) \in [(0, 256) \times (0, 256)] \wedge C(x, y) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

The binarized mask is then element-wise multiplied with the image warp2 to extract the target content within the designated region. Subsequently, the non-zero regions are pasted onto the left side of the element-wise product of the image warp1 and its corresponding mask mask1, thereby completing the final image composition. The data processing involved in the above procedure can be formally described as follows:

$$I_{out}(x, y) = [1 - P(x, y)](I^{(1)}(x, y)M^{(1)}(x, y)) + P(x, y)I^{(2)}(x, y) \quad (17)$$

In the above formulation, $I^{(1)} \in R^{H \times W \times 3}$ denotes the warp1 image, $I^{(2)} \in R^{H \times W \times 3}$ represents the warp2 image, $M^{(1)} \in [0, 1]^{H \times W}$ corresponds to the binarized and normalized mask1, and $C \in [0, 1]^{H \times W}$ refers to the raw mask output generated by the composition network.

Equation (16) generates a *binary paste prior* PPP from the soft mask output of MaskNet: pixels exceeding the adaptive confidence threshold τ are marked as candidate replacement regions. Equation (17) performs a two-stage compositing: (1) retain high-confidence content from $warp_1$ using its validity mask M_1 ; (2) paste candidate pixels from $warp_2$ wherever $P=1$. The Hadamard products ensure spatial gating, and all masks remain differentiable in training (binary P is relaxed to a smoothed sigmoid form during backpropagation). This operation is conceptually similar to a soft graph-cut seam selection, producing an interpretable initial composite that the diffusion network subsequently refines.

Subsequently, the diffusion model adds controllable noise to the stitched image (as in Eq. 13) and progressively denoises it through the multi-scale U-Net (as in Eq. 14), leveraging the attention mechanism to capture semantic correlations across images and ensure texture continuity at the stitching boundaries.

Loss function of the composition stage

During training, the model optimizes its denoising capability through a noise prediction loss, while simultaneously applying boundary loss and smoothness loss to enforce mask smoothness. Ultimately, the learned dynamic mask enables pixel-level seamless transitions through weighted fusion. This process, combining iterative noise refinement and multi-loss joint optimization, effectively addresses fusion blurring caused by the low signal-to-noise ratio of terahertz images, while preserving structural integrity in weak texture regions, thus significantly enhancing the visual consistency of the stitching results.

The total Composition loss includes mask-specific regularisers that promote interpretability:

$$L_{\text{mask}} = \lambda_{\text{BCE}} \text{BCE}(M_t, \text{blur}(P_0)) + \lambda_{\nabla} \|\nabla M_T\|_1 + \lambda_{\text{TV}} \text{TV}(M_T) + \lambda_{\text{cnt}} \text{H}(M_T) \quad (18)$$

where M_T is the final mask after diffusion, P_0 the initial paste prior, $\text{blur}(\cdot)$ a 5×5 Gaussian to avoid overfitting to hard edges, $\text{TV}(\cdot)$ the total-variation norm, and H the per-pixel Bernoulli entropy (encouraging confident 0/1 decisions). Default weights: $\lambda_{\text{BCE}} = 1.0$, $\lambda_{\text{V}} = 0.5$, $\lambda_{\text{TV}} = 0.25$, $\lambda_{\text{cnt}} = 0.1$. This objective encourages edge-aligned yet spatially smooth masks, reduces speckle noise, and yields interpretable seam selection.

The total loss function for the Composition stage is defined as:

$$L_{\text{c-total}} = \lambda_{\text{boundary}} L_{\text{boundary}} + \lambda_{\text{smooth}} L_{\text{smooth}} + \lambda_{\text{perceptual}} L_{\text{perceptual}} + \lambda_{\text{ms}} L_{\text{ms}} + \lambda_{\text{diffusion}} L_{\text{diffusion}} + L_{\text{mask}} \quad (19)$$

where $\lambda_{\text{boundary}}$, λ_{smooth} , $\lambda_{\text{perceptual}}$, λ_{ms} and $\lambda_{\text{diffusion}}$ are the weighting coefficients for the boundary loss, smoothness loss, perceptual loss, multi-scale loss, and diffusion loss, respectively, balancing the contributions of each loss component. During training, the default settings adopted in this work are $\lambda_{\text{boundary}} = 1.0$, $\lambda_{\text{smooth}} = 1.0$, $\lambda_{\text{perceptual}} = 0.5$, $\lambda_{\text{ms}} = 0.5$, and $\lambda_{\text{diffusion}} = 1.0$.

Meanwhile, to address issues related to training stability and convergence efficiency, an adaptive weight adjustment strategy was incorporated into the network design. By dynamically adjusting the weighting coefficient of the boundary loss-assigning a lower weight during the initial training phase and gradually increasing it as training progresses-the model is guided to first learn global feature distributions before progressively refining local details. Additionally, an upper limit mechanism was imposed on the diffusion loss weight to effectively suppress abnormal gradient fluctuations caused by noise prediction deviations, thereby ensuring training stability in complex scenarios. This combined strategy of progressive optimization and dynamic constraint not only balances the learning priorities between global structure and local precision but also mitigates the risk of training divergence under high-noise conditions.

The loss functions for each component in the Composition stage are as follows:

$$\begin{cases} L_{\text{diffusion}} = \|\varepsilon - \varepsilon_{\theta}(x_t, t)\|_2^2 \\ L_{\text{boundary}} = \sum_{p \in B} \|I_{\text{stitched}}(p) - I_{gt}(p)\|_1 \\ L_{\text{smooth}} = \sum \|\nabla \text{mask}\|_1 \\ L_{\text{perceptual}} = \sum_1 \|\Phi_1(I_{\text{stitched}}) - \Phi_1(I_{gt})\|_2^2 \\ L_{\text{ms}} = \sum_S W_S \|D_S(I_{\text{stitched}}) - D_S(I_{gt})\|_1 \end{cases} \quad (20)$$

The Improved Diffusion Composition (ImprovedDiffusionComposition) achieves high-quality image fusion through multi-level feature interaction and dynamic optimization mechanisms, providing an effective solution to fusion degradation issues caused by resolution limitations and parallax effects in terahertz images.

Experimental results and analysis

Experimental setup

In the experimental setup, the proposed algorithm is validated on the UDIS-D benchmark dataset and a self-collected terahertz-specific dataset. The UDIS-D dataset includes image pairs with complex scenes featuring multiple parallax and diverse texture characteristics, while the self-built dataset focuses on stitching tasks under practical terahertz imaging environments and special scenarios. Note that UDIS-D is a visible-spectrum dataset of natural indoor and outdoor scene pairs; it does not contain terahertz imagery. We therefore use UDIS-D purely as a generic benchmark for cross-method comparison, whereas all THz-specific performance is reported on our self-collected 350 GHz dataset.

The experiments are conducted on an NVIDIA GeForce RTX 3090 GPU platform, with models built using the PyTorch 2.5.1 deep learning framework. The Adam optimizer is employed for training, with an initial learning rate set to $1e-4$. The Warp and Composition stages are optimized separately for 100 epochs each, with batch sizes set to 2 and 4 respectively to balance memory efficiency and gradient stability.

To comprehensively evaluate the algorithm's performance, Peak Signal-to-Noise Ratio (PSNR) is used to quantify pixel-level differences between reconstructed and reference images. The Structural Similarity Index (SSIM) is adopted to assess spatial feature preservation capability. The Boundary Error metric is introduced to precisely measure the geometric consistency of stitching transition areas. Furthermore, subjective visual consistency evaluations are conducted to validate the overall performance of the algorithm in practical application scenarios.

Quantitative evaluation

The quantitative comparison between the proposed UDTATIS method and existing approaches on the UDIS-D benchmark is presented in Table 1. The overall performance of UDTATIS is comparable to that of UDIS++, with only minor differences in PSNR and SSIM (a difference of 0.14 dB and 0.0029, respectively),

Method	PSNR (dB) ↑	SSIM ↑	Boundary error ↓
Stitching ¹⁶	21.68	0.7014	0.1425
UDIS ¹⁷	22.41	0.7136	0.1182
UDIS++ ¹⁴	25.37	0.7358	0.0895
UDTATIS (ours)	25.23	0.7329	0.0931

Table 1. Quantitative comparison of different methods on the UDIS-D test set.

Method	PSNR (dB) ↑	SSIM ↑	Boundary error ↓
Stitching ¹⁷	15.15	0.6342	0.1768
UDIS ⁹	15.38	0.6587	0.1427
UDIS++ ¹⁴	17.45	0.6914	0.1156
UDTATIS (ours)	18.83	0.7218	0.1025

Table 2. Quantitative comparison of different methods on the terahertz image dataset.

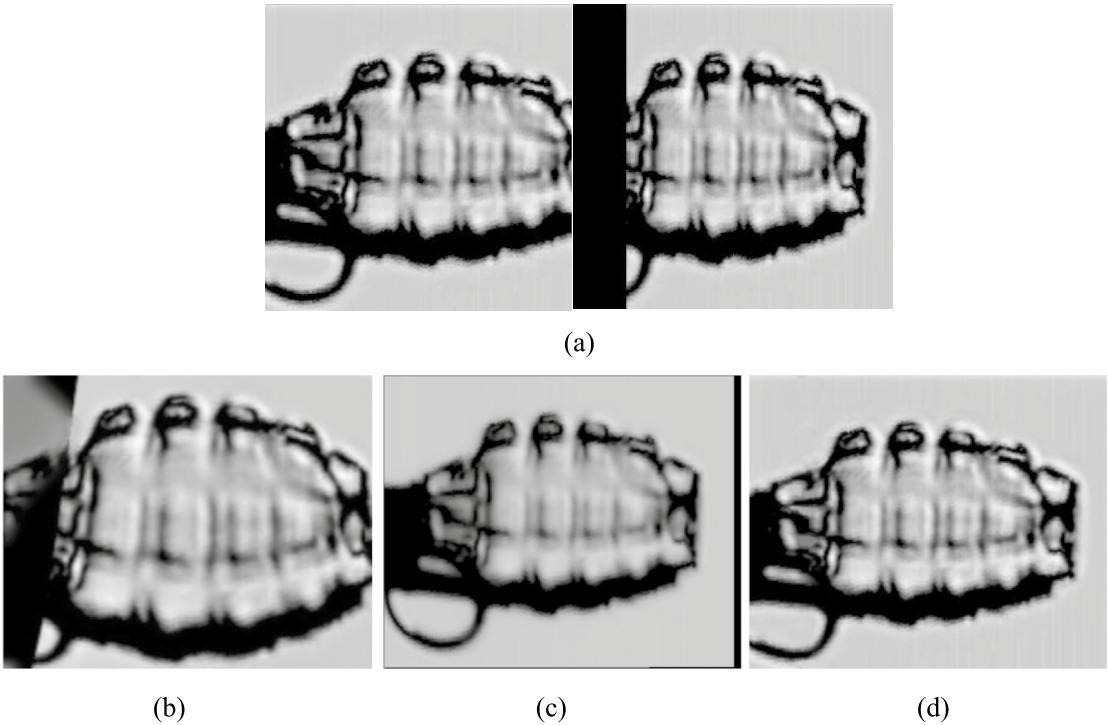


Fig. 7. Visualization comparison of stitching results by different methods. From left to right: input image pair, traditional method, Stitching, UDIS++, and UDTATIS (ours): (a) Input image; (b) Stitching algorithm results; (c) UDIS++ Algorithm results; (d) UDTATIS algorithm results.

and a small discrepancy in boundary error (0.0036). The UDTATIS method is specifically designed to address the characteristics of terahertz images, such as low signal-to-noise ratios and unique frequency-band signals. In contrast, the UDIS-D benchmark may be more representative of general-purpose scenarios. As a general algorithm, UDIS++ may benefit from a model architecture or training strategy that offers stronger generalization across diverse datasets. On the other hand, UDTATIS prioritizes robustness in terahertz image processing, which may result in slightly compromised performance on non-specialized datasets.

On the terahertz image dataset, the advantages of UDTATIS are even more pronounced. The testing results on the terahertz image dataset are shown in Table 2. A comparison of the data in Table 2 reveals that the UDTATIS method is better suited for handling terahertz image stitching tasks characterized by limited texture features and significant parallax.

Qualitative evaluation

Due to the underdeveloped nature of terahertz imaging devices, the terahertz images currently obtainable through experiments are of relatively low resolution. As a result, traditional stitching algorithms (such as DeepIS and UDIS) fail to achieve satisfactory stitching performance. However, the Stitching method and UDIS++ have been verified to work with the terahertz dataset collected in this study. A comparison of the image stitching results produced by the proposed UDTATIS algorithm, the Stitching method, and UDIS++ on the terahertz dataset is shown in Fig. 7.

As shown in the comparison results of Fig. 7, UDTATIS demonstrates the best performance in handling stitching boundaries, with virtually no visible artifacts. Traditional stitching algorithms fail to achieve satisfactory results on the terahertz dataset, exhibiting noticeable discontinuities and distortions at the boundaries. While UDIS++ produces relatively smooth stitching results, there is still room for improvement in preserving image details and ensuring seamless boundary transitions. In contrast, UDTATIS achieves the most natural and

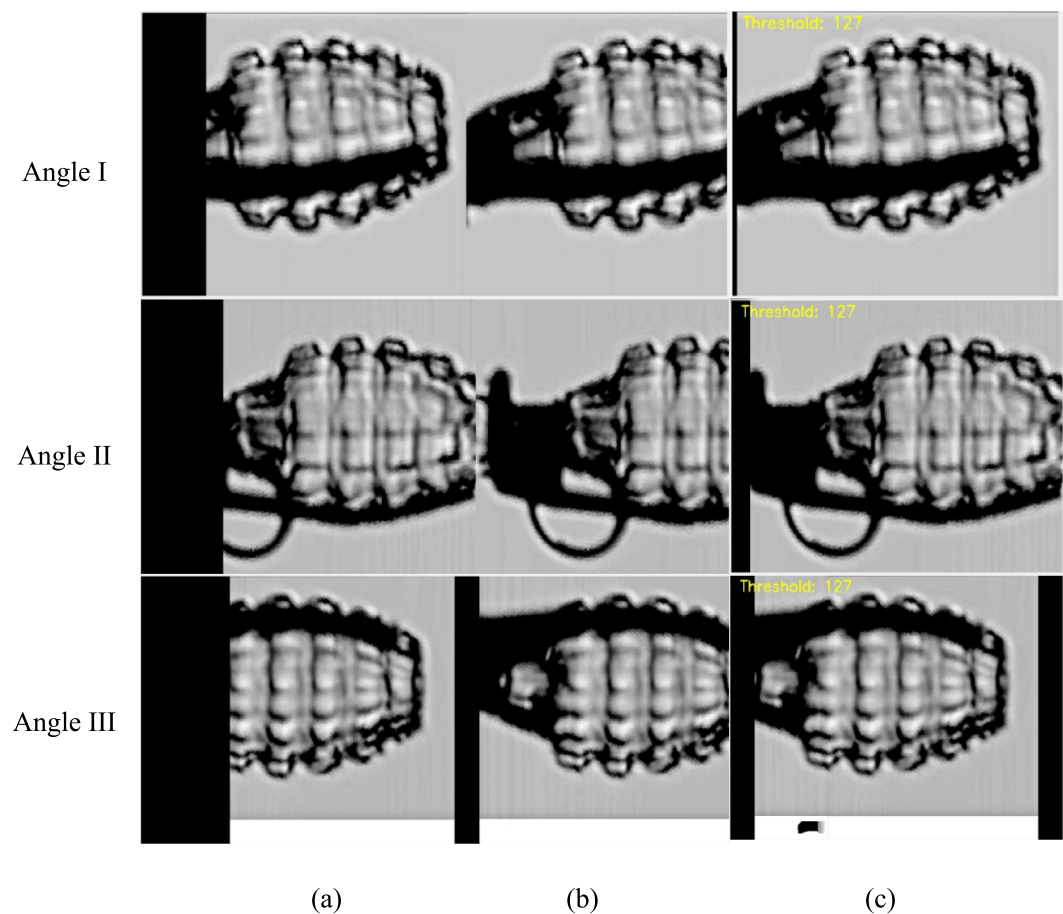


Fig. 8. Stitching results of terahertz images collected from different angles: (a) Input image 1; (b) Input image 2; (c) Splicing result.

Method variants	PSNR (dB) ↑	SSIM ↑	Boundary error ↓
Baseline model	17.85	0.6912	0.1292
Effective point discrimination	17.96	0.6998	0.1164
Continuity constraint	18.04	0.6948	0.1113
Diffusion model	18.52	0.7136	0.1075
Adaptive normalization	18.32	0.7082	0.1145
UDTATIS	18.83	0.7218	0.1025

Table 3. Ablation study results of UDTATIS components.

seamless fusion, while maintaining the original image’s structural integrity and fine details. The stitching effects of the collected terahertz images from different angles are shown in Fig. 8.

Ablation study results and analysis

To further evaluate the contribution of each module to the THz image stitching performance, ablation experiments were conducted, and the results are shown in Table 3. The baseline model achieved 17.85 dB in PSNR, 0.6912 in SSIM, and 0.1292 in Boundary Error.

After introducing the effective point discrimination mechanism, PSNR increased to 17.96 dB and Boundary Error decreased by 14.2%, highlighting its critical role in filtering out mismatched points. The addition of the diffusion model further boosted the PSNR to 18.52 dB and reduced Boundary Error by 16.8%, demonstrating the significant effect of multi-scale fusion on boundary optimization.

The complete UDTATIS model ultimately achieved optimal performance of 18.83 dB, 0.7218, and 0.1025, with a 20.7% reduction in boundary error compared to the baseline, systematically proving the effectiveness of the collaborative module design in improving stitching quality.

To explicitly clarify the interaction between the diffusion model and adaptive normalization (AN), it is important to highlight that the synergistic effect primarily arises from time-step conditioned feature modulation.

AN dynamically adjusts feature statistics—mean and variance—based on the diffusion timestep, thus precisely controlling the denoising process at different noise levels. Specifically, in early, high-noise diffusion steps, AN enlarges the dynamic range of features to improve denoising effectiveness; in later, low-noise steps, AN stabilizes feature representations, preserving critical fine details and minimizing distortion. Moreover, this time-adaptive feature adjustment provides significant benefits in terms of improved gradient flow and numerical stability, preventing gradient explosion and numerical overflow during training, especially relevant for sensitive image stitching tasks. Quantitatively, the incorporation of AN demonstrates an overall improvement of model convergence quality, enriched feature representation, and optimized stitching boundaries, which collectively enhance the visual smoothness and continuity of THz image mosaics. On our dataset, the diffusion model alone improves PSNR by 0.48 dB (from 18.04 to 18.52 dB). With AN, the PSNR slightly adjusts to 18.32 dB, a minor reduction reflecting a beneficial regularization effect rather than diminished capability, as evidenced by improved boundary consistency and perceptual quality. Thus, the combined deployment of diffusion fusion and adaptive normalization is particularly advantageous for THz image stitching due to their complementary strengths in noise handling, detail retention, and boundary refinement.

Conclusion and future work

This study proposes an unsupervised disparity-tolerant terahertz image stitching algorithm (UDTATIS), which achieves joint optimization of geometric alignment and image fusion through a two-stage architecture. In the Warp stage, by integrating the EfficientLOFTR feature extractor, valid point discrimination mechanism, and continuity constraint, the method significantly improves matching accuracy and deformation field smoothness in low-resolution and texture-sparse scenarios. In the Composition stage, an improved diffusion model is introduced, employing a multi-scale U-Net architecture and adaptive normalization techniques to effectively address issues of hard boundary transitions and noise sensitivity during fusion. Experimental results demonstrate that the proposed algorithm achieves superior performance on both the UDIS-D benchmark and the custom terahertz dataset, particularly excelling in boundary error suppression and weak texture structure preservation compared to existing methods.

Future work will focus on several key directions. For model lightweighting and computational efficiency, we plan to apply knowledge distillation and network pruning strategies to reduce model complexity while maintaining performance, and explore transformer-based sparse attention mechanisms to improve inference speed. For multimodal fusion, we aim to develop a shared latent space learning framework that jointly aligns visible and terahertz features using contrastive learning, enabling cross-domain enhancement in weakly textured areas. In addition, we will investigate online stitching pipelines with progressive refinement modules to support real-time applications and extend the algorithm's robustness to complex scenes involving large disparities and dynamic occlusions. These efforts are expected to promote the practical deployment of terahertz imaging in industrial inspection and medical diagnostics.

Data availability

The algorithm and related data supporting the findings of this study are available on GitHub at: <https://github.com/snow-wind-001/UDTATIS>.

Received: 19 May 2025; Accepted: 18 August 2025

Published online: 25 August 2025

References

- Bai, F., Li, L., Wang, W. & Wu, X. DETransMVSNet: Research on terahertz 3D reconstruction of multi-view stereo network with deep equilibrium transformers. *IEEE Access* **11**, 146042–146053. <https://doi.org/10.1109/ACCESS.2023.3342847> (2023).
- Wu, X. et al. A 3D reconstruction of terahertz images based on the FCTMVSNet algorithm. *IEEE Access* **12**, 108975–108985. <https://doi.org/10.1109/ACCESS.2024.3439358> (2024).
- Wang, W., Yin, B., Li, L., Li, L. & Liu, H. A low light image enhancement method based on dehazing physical model. *CMES-Comput. Model. Eng. Sci.* <https://doi.org/10.32604/cmescs.2025.063595> (2025).
- Wang, W., Yuan, X., Wu, X. & Liu, Y. Fast image dehazing method based on linear transformation. *IEEE Trans. Multimed.* **19**(6), 1142–1155. <https://doi.org/10.1109/TMM.2017.2652069> (2017).
- Zarrinkhat, F. et al. Calibration alignment sensitivity in corneal terahertz imaging. *Sensors* **22**(9), 3237. <https://doi.org/10.3390/s22093237> (2022).
- Wang, H. J. et al. Image registration method using representative feature detection and iterative coherent spatial mapping for infrared medical images with flat regions. *Sci. Rep.* **12**, 7932. <https://doi.org/10.1038/s41598-022-11379-2> (2022).
- Oluwasanmi, A. et al. Attention autoencoder for generative latent representational learning in anomaly detection. *Sensors* **22**(1), 123. <https://doi.org/10.3390/s22010123> (2022).
- Li, X. et al. High-throughput terahertz imaging: progress and challenges. *Light Sci. Appl.* **12**, 233. <https://doi.org/10.1038/s41377-023-01278-0> (2023).
- Jiang, Y. et al. Classification of unsound wheat grains in terahertz images based on broad learning system. *IEEE Trans. Plasma Sci.* **52**(10-Part2), 4973–4982. <https://doi.org/10.1109/TPS.2024.3390777> (2024).
- Jiang, Y. et al. Adaptive compressed sensing algorithm for terahertz spectral image reconstruction based on residual learning. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **281**, 121586. <https://doi.org/10.1016/j.saa.2022.121586> (2022).
- Gao, G., Zhao, S., Wang, X. & Sun, X. An image retrieval method based on SIFT feature extraction. In *Proceedings of SPIE 13288, Fourth International Conference on Computer Graphics, Image, and Virtualization (ICCGIV 2024)*, 1328812. <https://doi.org/10.1117/12.3045302> (2024).
- Megha, V. & Rajkumar, K. K. Seamless panoramic image stitching based on invariant feature detector and image blending. *Int. J. Image Graph. Signal Process. (IJIGSP)* **16**(4), 30–41. <https://doi.org/10.5815/ijigsp.2024.04.03> (2024).
- Nie, L. et al. A view-free image stitching network based on global homography. *J. Vis. Commun. Image Represent.* **73**, 102950 (2020).
- Nie, L., Lin, C., Liao, K. et al. Parallax-tolerant unsupervised deep image stitching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7399–7408 (2023).

15. Wang, Y., He, X., Peng, S. et al. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 21666–21675 (2024).
16. Weber, L. A. & Schenk, D. Automatische Zusammenführung zertrennter Konstruktionspläne von Wasserbauwerken. *Bautechnik* **99**, 330–340. <https://doi.org/10.1002/bate.202200010> (2022).
17. Nie, L., Lin, C., Liao, K., Liu, S. & Zhao, Y. Unsupervised deep image stitching: Reconstructing stitched features to images. *IEEE Trans. Image Process.* **30**, 6184–6197. <https://doi.org/10.1109/TIP.2021.3092828> (2021).

Acknowledgements

During the preparation of this manuscript, the authors used ChatGPT for the purpose of manuscript translation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Author contributions

X.J.W. (Xiaojin Wu), F.B. (Fan Bai) and L.L. (Lun Li), responsible for the paper writing, data statistics and analysis; W.C.W. (Wencheng Wang) is responsible for the structure design of the paper; Y.G. (Yuan Gao) is responsible for data collation and statistics; H.F.C. (Hongfu Cai) was responsible for the overall paper review.

Funding

This work was supported by the Natural Science Foundation of Shandong Province. Author L.L. (Lun Li) received funding under Grant Number ZR2024MA055, and author W.C.W. (Wencheng Wang) received funding under Grant Number ZR2023MF047. The full name of the funder is Natural Science Foundation of Shandong Province. The funder's website is: <http://kjt.shandong.gov.cn/>

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025