Acta Astronautica 249 (2026) 1369–1384

Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

Research paper

Physics-based diffusion model for thermal image generation in space orbital rendez-vous scenarios

Said Guerazem *, Nabil Aouf

City St George's, University of London, UK

ARTICLE INFO

Keywords:
Artificial Intelligence
Physics-informed diffusion
Image translation
Proximity operations
Spacecraft relative navigation

ABSTRACT

This paper presents an innovative physics-informed diffusion framework for generating thermal infrared imagery from visible-band observations in spacecraft rendezvous for on-orbit servicing (OOS) scenarios. To the best of our knowledge, this is the first learning-based approach to address visible-to-thermal translation tailored to space proximity operations, where strong cross-spectral appearance shifts and space-specific environmental effects limit the applicability of conventional image-to-image methods. The proposed model is trained conditionally using a composite, multi-feature representation designed to bridge the modality gap between visible and thermal bands and to improve robustness across illumination and viewing geometries. Training follows a multistage strategy that couples nominal generation with a refinement phase applied to sampled outputs, yielding improved structural fidelity and reduced artifacts. Finally, we introduce a model-driven physical constraint that embeds the radiometric temperature into the learning objective, guiding the refinement stage toward physically consistent radiance patterns and enhancing realism. Experiments in representative rendezvous configurations demonstrate that the method produces sharper, more plausible thermal renderings while preserving target geometry and its key thermally salient features, supporting downstream perception tasks in rendezvous operations.

1. Introduction

In recent years, space exploration has attracted substantial commercial and institutional interest. As of 2025 Q2, the global space economy has surpassed $600 billion, accompanied by the launch of more than 150 new space-related industries over the same period [1]. This sustained growth underscores the demand for robust and secure technologies across the sector, particularly for *Proximity Operations*, where reliable relative navigation is pivotal for close-range rendezvous (RV) missions.

Relative navigation addresses the estimation of the six degrees-of-freedom (6-DoF) pose — position and orientation — of a target spacecraft with respect to a chaser. Classical approaches have traditionally relied on active sensors such as LiDAR, often at the cost of increased power consumption, mass, and overall system complexity [2]. More recent efforts have shifted toward passive vision systems, especially monocular configurations, due to their favorable payload and energy requirements [3,4]. Existing methods can be broadly categorized as classical or learning-based. Classical techniques typically adopt a model-based strategy, solving the Perspective-*n*-Point (PnP) problem by associating 2D features extracted from images with the known geometry of a 3D target model. Learning-based methods, in contrast, use features learned from data, either via direct end-to-end regression of the 6-DoF pose or via indirect pipelines that predict intermediate features subsequently processed by a PnP solver [5]. Regardless of the chosen approach, the input imagery must exhibit sufficient discriminative content to infer the pose. Intuitively, pose estimation becomes ill-posed when the target is fully occluded or heavily shadowed (e.g., a near-black image in the visible band). This motivates the use of complementary spectral bands in conjunction with the visible domain, most notably long-wave infrared (LWIR) [6,7].

LWIR imagery has been repeatedly investigated for pose estimation in space applications. Thermal features are often more repeatable and robust than their visible-domain counterparts, exhibiting increased resilience under the harsh and highly variable illumination conditions encountered in orbit [7,8].

From a data perspective, the most widely used benchmark for spacecraft pose estimation is the Spacecraft Pose Estimation Dataset (SPEED) [9], released as part of the first European Space Agency Kelvin's Satellite Pose Estimation Challenge (SPEC). While SPEED has been extended to the thermal band through the work of Bechini et al. [10], and has significantly accelerated research by providing high-fidelity imagery of the non-cooperative Tango satellite, its randomized

* Corresponding author.
E-mail address: said.guerazem@citystgeorges.ac.uk (S. Guerazem).

https://doi.org/10.1016/j.actaastro.2026.08.058
Received 13 April 2026; Received in revised form 30 July 2026; Accepted 20 August 2026
Available online 28 August 2026

pose sampling limits its suitability for sequential and time-dependent scenarios such as rendezvous trajectories [6].

Beyond SPEED, existing datasets remain limited. A notable effort is the Astos dataset [6], generated using the ASTOS simulator [11], which includes several RV trajectories of the Envisat satellite across different approaches and operational scenarios. Other commercial solutions include PANGU (Planet and Asteroid Natural Scene Generation Utility) [12], initially developed for synthetic planetary surface imagery and later extended to artificial objects, as well as Airbus SurRender [13], a widely used tool capable of producing multispectral imagery for various space targets (e.g., planets, asteroids, and spacecraft). In contrast, general-purpose simulators such as Blender [14] typically lack the realism required for high-fidelity thermal rendering. Overall, the only publicly available thermal dataset remains the extension built upon SPEED.

To address this gap, we propose a physics-informed Denoising Diffusion Probabilistic Model (DDPM). Our approach builds on a standard denoising U-Net [15], trained to denoise under the conditioning of a multi-feature representation extracted from the corresponding visible image. We further introduce a multi-stage training strategy: after nominal diffusion training, an additional refiner U-Net is employed to post-process sampled outputs, guided by a physics-informed constraint derived from the estimated radiometric temperature. In summary, our contributions are as follows:

- As far as existing research suggests, this work is the first to address visible-to-thermal image translation for space proximity operations using a learning-based approach.
- We introduce a multi-feature conditional diffusion model that, beyond standard edge-based cues, leverages three additional saliency maps extractable directly from the visible image to better bridge the cross-spectral domain gap.
- We propose a multi-stage training strategy that combines standard diffusion denoiser training with a dedicated refinement network. The refinement is driven by a physics-based constraint grounded in target material properties and a gray-body radiometric image-formation model. We further define a *material tracing map* to link material-dependent information to temperature cues in the thermal image, and, to the best of our knowledge, this is the first time such a constraint is introduced in this form.

The remainder of this paper is organized as follows. Section 2 reviews related work on visible-to-thermal translation. Section 3 presents the proposed methodology, including the multi-feature conditional denoising network, the training strategy, and the development of the physics-based constraint. Section 4 reports the experimental evaluation and comparative analysis against existing models. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related work

Research on visible-to-thermal translation has progressed substantially in recent years, driven primarily by terrestrial applications such as autonomous driving, synthetic aperture radar (SAR) imagery, and human surveillance [16–18]. These efforts seek both to augment limited thermal datasets and to enable robust cross-modal matching and detection under challenging illumination conditions.

Generative approaches, and in particular Generative Adversarial Networks (GANs) introduced by Goodfellow et al. [19], have been widely explored for thermal image synthesis [20]. Variants of the original GAN formulation, including conditional GANs [21] and CycleGANs [22], were developed to alleviate well-known issues of the vanilla model, such as training instability and mode collapse [20]. More broadly, GANs as a class of deep generative models (DGMs) have been used to capture complex data distributions and to generate realistic samples for image-to-image translation tasks [16,20,22–24]. The standard GAN framework is cast as a minimax optimization problem [25] between a generator and a discriminator, both implemented as deep neural networks. Achieving a stable solution, often interpreted as a Nash equilibrium, typically requires careful architecture and loss design [20,25].

Several works have proposed GAN-based solutions specifically for visible-to-thermal translation. Li et al. introduced V2IGAN, which employs a modified U-Net backbone [15] augmented with a visual state-space attention module to focus the model on informative regions in the visible image [24,26]. They further incorporated a multi-scale feature contrastive loss to improve robustness, and reported superior performance compared with standard GAN baselines [24]. Ma et al. [16] proposed a GAN formulation that integrates matching losses, including normalized cross-correlation (NCC) and a match-patch (MP) term, to better align structural content between visible and infrared images.

More recently, diffusion models have emerged as a new state-of-the-art family of DGMs [27], outperforming GANs in several high-resolution image synthesis benchmarks [18,28,29], as well as in natural language processing [30,31] and temporal data modeling [32]. Diffusion models operate by progressively corrupting data with noise and then learning the reverse process to recover clean samples. They are characterized by stable training, strong generative performance, and fine control over image content through conditioning, which makes them particularly appealing for image translation problems [27].

Within visible-to-thermal translation, conditional diffusion methods have started to appear. Zhu et al. [18] proposed an edge-guided adversarial conditional diffusion model trained in two complementary phases: the model first learns to denoise conditioned on thermal edge maps and is subsequently adapted to visible-edge conditioning with adversarial training. In parallel, Hu et al. introduced CM-Diff, a conditional DDPM [28,29] designed for bidirectional cross-modality translation between thermal and visible bands. Their framework includes a statistical inference scheme with dual loss functions to enforce consistency in the generated outputs [17].

In the context of physics-based diffusion models, Mao et al. [33] proposed a physics-informed diffusion framework for infrared image generation. Their physical constraint is based on TeV decomposition, which applies the superposition principle to multiple infrared spectral components. This criterion is used to construct a self-supervised network that jointly predicts three maps: a temperature map T, an emissivity map $\mathbf{e}$, and a thermal vector map $\mathbf{V}$, together with a downsampled image. The constraint is enforced both in the **TeV** space and on the reconstructed infrared image. However, their formulation assumes a perfectly transparent atmosphere, and material properties are not explicitly modeled; instead, the physical behavior is learned empirically from the infrared dataset via self-supervision.

Most datasets used to develop and evaluate these methods focus on terrestrial scenarios, such as the LLVIP dataset [34] and the FLIR dataset [35], which have supported progress in cross-modal synthesis and detection. In contrast, thermal datasets tailored to spacecraft relative navigation are extremely limited, which constrains research in this domain.

To the best of our knowledge, the only study addressing synthetic thermal image generation for on-orbit servicing and close-proximity spacecraft operations is the work of Bianchi et al. [36]. Their approach follows the pipeline introduced by Quirino [37] where a detailed 3D model of the target, with explicit material properties, is combined with thermal simulation in OpenFOAM [38] to produce a mesh for radiative modeling, while camera parameters and poses are applied to render paired visible images in Blender. The resulting synthetic multispectral pairs are well suited for the development and validation of multispectral navigation algorithms [36,37].

In summary, AI-based methods dominate visible-to-thermal synthesis in terrestrial settings, whereas for spacecraft applications the prevailing approach remains a classical, domain-specific pipeline that couples thermal simulation with visible rendering to generate spacecraft thermal imagery.

3. Methodology

In this section, we present the formulation of the proposed framework for generating synthetic thermal images from visible-band imagery, with a specific focus on multispectral pose estimation. The objective is to synthesize LWIR images that are spatially aligned with their visible counterparts while preserving the geometric cues required for downstream relative navigation and 6-DoF pose estimation.

3.1. Problem formulation

Let the visible dataset be defined as $\mathcal{D}_{\text{vis}} = \{\mathbf{x}_1^{\text{vis}}, \mathbf{x}_2^{\text{vis}}, \dots, \mathbf{x}_N^{\text{vis}} \mid \mathbf{x}^{\text{vis}} \in \mathbb{R}^{H\times W\times 3}\}$, and the corresponding thermal dataset as $\mathcal{D}_{\text{lwir}} = \{\mathbf{x}_1^{\text{lwir}}, \mathbf{x}_2^{\text{lwir}}, \dots, \mathbf{x}_N^{\text{lwir}} \mid \mathbf{x}^{\text{lwir}} \in \mathbb{R}^{H\times W\times 1}\}$. Each dataset is sampled from the underlying distributions $p_{\text{vis}}(\mathbf{x}^{\text{vis}})$ and $p_{\text{lwir}}(\mathbf{x}^{\text{lwir}})$, respectively. The goal is to train a generative model $\mathcal{Z}$ that learns the conditional distribution $p_{\mathcal{Z}}(\mathbf{x}^{\text{lwir}} \mid \mathbf{x}^{\text{vis}})$, enabling translation from the visible domain to the thermal domain.

For each paired sample $(\mathbf{x}_i^{\text{vis}}, \mathbf{x}_i^{\text{lwir}})$, tight pixel-level alignment between modalities is needed to keep multispectral consistency and to ensure that the generated LWIR images remain usable for pose estimation. To that end, we adopt a Conditional DDPM and condition it on a set of multi-feature inputs extracted from the visible domain. The motivation behind the choice of these conditioning features is detailed in the following subsection.

3.2. Multi-feature conditioning input

Visible and thermal modalities naturally carry different information. The visible spectrum encodes appearance cues such as color and texture, while LWIR captures emitted radiation and typically lacks fine texture and chromatic content. That being said, both modalities share stable structural cues, mainly contours and object boundaries, which are particularly useful for cross-modal translation [17,18]. Building on this observation, we condition the diffusion model using edge-based and saliency-based features extracted from the visible images.

Edges are first extracted from each visible image to preserve geometry during translation. The visible edge map is obtained as:

$$E_{\text{vis}} = \mathcal{E}(\mathbf{x}_{\text{vis}}), \tag{1}$$

where $\mathcal{E}$ is a pretrained multi-scale autoencoder for edge detection [39]. In our setup, $\mathcal{E}$ is pretrained using ground-truth edges generated by a classical edge extraction method applied to our dataset.

In addition, we incorporate a *multi-scale saliency representation* inspired by the visual space-domain decomposition proposed by Liu et al. [40]. Although originally developed for multispectral image fusion, the fine-to-coarse decomposition provides perceptual cues that are complementary to edges and beneficial for cross-modal synthesis. In practice, this gives the model more guidance than relying on edges alone, and avoids having to start from thermal edges as an initial condition as in prior work [18].

Given a visible image $\mathbf{I}_{\text{vis}}$ represented in grayscale, we compute saliency maps across multiple scales. The coarse-scale representation is generated through Gaussian smoothing, where the scale space is obtained by convolving the image with a Gaussian kernel of standard deviation σ. The first scale is computed as:

$$F_{\text{vis}}^1 = \text{Gaussian}(\sigma) * \mathbf{I}_{\text{vis}}, \tag{2}$$

where Gaussian$(\cdot)$ denotes Gaussian filtering and $\sigma = 5$, following [40].

To obtain coarser representations, the smoothed image is progressively downsampled and filtered:

$$F_{\text{vis}}^2 = \text{Gaussian}(\sigma) * \text{coarse}(F_{\text{vis}}^1), \tag{3}$$

$$F_{\text{vis}}^3 = \text{Gaussian}(\sigma) * \text{coarse}(F_{\text{vis}}^2), \tag{4}$$

where coarse$(\cdot)$ denotes a 4× downsampling operation that reduces resolution and enhances global context.

For numerical consistency, each scale map F_{vis}^i is normalized by measuring the deviation from its average gray level:

$$f_{\text{vis}}^i = \frac{1}{M_i N_i} \sum_{x=1}^{M_i} \sum_{y=1}^{N_i} F_{\text{vis}}^i(x,y), \tag{5}$$

$$\mathbf{S}_{\text{vis}}^i(x,y) = \left| F_{\text{vis}}^i(x,y) - f_{\text{vis}}^i \right|, \tag{6}$$

where f_{vis}^i denotes the mean intensity of the ith scale and (M_i, N_i) are its spatial dimensions. The resulting saliency maps $\mathbf{S}_{\text{vis}}^i$ highlight visually distinctive regions at multiple scales, which improves the model's perceptual awareness during denoising.

The final conditional input is formed by combining a single-channel grayscale visible image $\mathbf{I}_{\text{vis}}$, its edge map E_{vis}, and the set of saliency maps $\mathbf{S}^{\text{vis}} = \{\mathbf{S}_{\text{vis}}^1, \mathbf{S}_{\text{vis}}^2, \mathbf{S}_{\text{vis}}^3\}$. These complementary features provide both structural and perceptual context to the diffusion model. Unlike standard diffusion frameworks [29,41,42], where conditioning often comes from class labels, text prompts, or paired natural images, our conditioning is explicitly feature-driven and tailored to the cross-spectral relationships required by the translation problem in this context. Before concatenation, all saliency maps are resized back to the original image resolution so that the final conditioning tensor has size $H \times W \times 5$.

3.3. Diffusion model architecture and training

In the following, we present the diffusion formulation we rely on, the corresponding neural parameterization, and the training procedure used to learn the reverse dynamics and generate new samples. The probabilistic setup makes the optimization objective explicit, and clarifies what the network is actually minimizing.

3.3.1. Diffusion formulation

We adopt a conditional variance-preserving diffusion model to map visible-band observations to thermal imagery. Let $\mathbf{x}_0^{\text{lwir}} \in \mathbb{R}^{H\times W\times 1}$ denote a clean thermal image and $\mathbf{x}_c$ the associated conditioning tensor derived from the visible modality. In our case:

$$\mathbf{x}_c = \mathbf{x}^{\text{vis}} \oplus \mathcal{E}(\mathbf{x}^{\text{vis}}) \oplus \mathbf{S}^{\text{vis}} \in \mathbb{R}^{H\times W\times 5}. \tag{7}$$

Following the discrete DDPM formulation [29,41], the forward noising process defines a Markov chain $\{\mathbf{x}_t^{\text{lwir}}\}_{t=0}^T$ with Gaussian transitions:

$$q(\mathbf{x}_t^{\text{lwir}} \mid \mathbf{x}_{t-1}^{\text{lwir}}) = \mathcal{N}\Big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1}^{\text{lwir}},\, \beta_t \mathbf{I}\Big), \quad t = 1, \dots, T, \tag{8}$$

where $\{\beta_t\}_{t=1}^T$ is a noise schedule and T is the length of the diffusion process. Marginalizing over the intermediate states yields the closed-form expression:

$$q(\mathbf{x}_t^{\text{lwir}} \mid \mathbf{x}_0^{\text{lwir}}) = \mathcal{N}\Big(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0^{\text{lwir}},\, (1-\bar{\alpha}_t)\mathbf{I}\Big), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s), \tag{9}$$

so that a noisy sample at time t can be obtained as:

$$\mathbf{x}_t^{\text{lwir}} = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0^{\text{lwir}} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{10}$$

The denoising network is trained to predict the injected noise from $(\mathbf{x}_t^{\text{lwir}}, t, \mathbf{x}_c)$. We parameterize this predicted noise as $\hat{\epsilon}_\theta(\mathbf{x}_t^{\text{lwir}}, t, \mathbf{x}_c)$ and optimize the standard DDPM objective:

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}}\left[\left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta\left(\mathbf{x}_t^{\text{lwir}}, t, \mathbf{x}_c\right)\right\|_2^2\right], \tag{11}$$

with t sampled uniformly from $\{1, \dots, T\}$ unless otherwise specified. This discrete-time formulation coincides with conventional training in image diffusion models and is directly applied here to our visible-to-thermal setting [29,41].

Algorithm 1 Conditional DDPM training (single iteration)

Require: Thermal image: $\mathbf{x}_0^{\text{lwir}}$, Conditioning tensor: $\mathbf{x}_c = \mathbf{x}^{\text{vis}} \oplus \mathcal{E}(\mathbf{x}^{\text{vis}}) \oplus \mathbf{S}^{\text{vis}}$, Noise schedule: $\{\beta_t\}$, Denoising network: $\hat{\epsilon}_\theta$

Sample $t \sim \text{Uniform}\{1, \dots, T\}$

Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

Compute $\bar{\alpha}_t \leftarrow \prod_{s=1}^{t}(1-\beta_s)$

Form noisy input $\mathbf{x}_t^{\text{lwir}} \leftarrow \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0^{\text{lwir}} + \sqrt{1-\bar{\alpha}_t}\,\epsilon$

Predict noise $\hat{\epsilon}_\theta \leftarrow \hat{\epsilon}_\theta(\mathbf{x}_t^{\text{lwir}}, t, \mathbf{x}_c)$

Compute loss $\mathcal{L} \leftarrow \|\epsilon - \hat{\epsilon}_\theta\|_2^2$

Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$

Since this discrete formulation admits a continuous-time representation as a variance-preserving stochastic differential equation (SDE) [43], we can also sample by numerically integrating the associated probability-flow ordinary differential equation (ODE). This makes the network $\hat{\epsilon}_\theta$, trained with the discrete DDPM loss, compatible with DPM-Solver [44]. The latter treats the probability flow as an initial-value problem on a continuous time horizon and integrates from a terminal noise level back to $t = 0$ using a small number of adaptive steps. In practice, this allows us to sample high-fidelity thermal images significantly more efficiently than standard DDIM sampling [41].

Training algorithm

The training procedure for the first-stage diffusion model follows standard DDPM training, augmented with conditioning on visible-derived features. A single training step is summarized in Algorithm 1.

At inference time, we freeze the trained diffusion model and use DPM-Solver to integrate the probability-flow ODE, producing a coarse thermal prediction conditioned on the visible input. This coarse sample is then refined by a second-stage U-Net, trained separately as described later, to obtain the final high-fidelity thermal output. The Stage 01 conditional diffusion training scheme is summarized in Fig. 3(b).

Diffusion backbone ϵ_θ

The first stage of our system is a conditional diffusion model built on a U-Net backbone [15,29]. The input to the network is the concatenation of a noisy thermal image $\mathbf{x}_t^{\text{lwir}} \in \mathbb{R}^{B\times 1\times H\times W}$ and the conditioning tensor $\mathbf{x}_c \in \mathbb{R}^{B\times C_{\text{cond}}\times H\times W}$ defined earlier. These features are processed jointly through the encoder–decoder hierarchy, while a time embedding is injected to modulate all residual blocks.

Overall, the design follows common diffusion backbones, with a base width of 64 feature maps and channel multipliers reaching 512 at the bottleneck. Self-attention layers [45] are inserted at coarse spatial resolutions after pre-activation residual blocks (GroupNorm, SiLU, and convolution) as this configuration is known to stabilize training under the heavy-tailed noise distribution induced by the diffusion process [46,47]. Attention is used to capture long-range structure at low resolutions [28] without incurring the quadratic cost of full-resolution attention (see Table 1).

Table 1
Architecture of the conditional diffusion U-Net ϵ_θ. The network operates on a noisy thermal input concatenated with a conditioning tensor of $C_{\text{cond}} = 5$ channels (visible, edges, and saliency maps). All encoder/decoder blocks consist of stacked time-conditioned residual blocks with GroupNorm and SiLU nonlinearities; attention is applied at the lowest resolutions.

Block	Resolution	Channels
Input/conditioning	$H \times W$	$1 + C_{\text{cond}}$
Enc-0	$H \times W$	64
Enc-1	$H/2 \times W/2$	128
Enc-2 (attn)	$H/4 \times W/4$	256
Enc-3 (attn)	$H/8 \times W/8$	256
Bottleneck (attn)	$H/16 \times W/16$	512
Dec-3 (attn)	$H/8 \times W/8$	256
Dec-2 (attn)	$H/4 \times W/4$	256
Dec-1	$H/2 \times W/2$	128
Dec-0	$H \times W$	64
Output	$H \times W$	1

3.4. Physics-based constraint

We now develop the theoretical framework used to impose the physics-based constraint on the generated thermal image. The core idea is to estimate a radiometric temperature map of the target from a model-based quantity, and then use that estimate to guide refinement.

It is important to clarify that the objective of this constraint is not to perform an absolute thermometric inversion of the scene. Instead, the aim is to construct a radiometric consistency signal that preserves the material and geometric structure of the thermal image. In this work, the thermal images are normalized images generated under a consistent sensor model, and therefore the derived temperature quantity is used as a temperature proxy for loss construction rather than as an independent Kelvin-scale measurement.

We consider a microbolometer thermal camera operating in the LWIR spectrum from 8 μm to 14 μm, a sensor type with established space heritage and sensitivity to incoming heat flux at the pixel level [7, 37]. The outgoing heat flux emitted in this band by a surface element at temperature T is:

$$q_{\text{out}}(T) = \pi \int_0^\infty \varepsilon(\lambda)\mathbf{B}(\lambda, T)\mathbf{R}(\lambda)\, d\lambda, \tag{12}$$

$$\mathbf{B}(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}, \tag{13}$$

where $\varepsilon(\lambda)$ is emissivity, $\mathbf{B}(\lambda, T)$ is Planck's law, h is Planck's constant, c is the speed of light, and $\mathbf{R}(\lambda)$ accounts for transmissivity (atmospheric attenuation). For space applications, transmissivity can be taken as $\mathbf{R}(\lambda) = 1$, hence:

$$q_{\text{out}}(T) = \pi \int_0^\infty \varepsilon(\lambda)\mathbf{B}(\lambda, T)\, d\lambda. \tag{14}$$

This assumption is important. In prior physics-based approaches, including Mao et al. [33], it is already adopted for terrestrial cases, and it holds even more naturally in space where atmospheric absorption is not present.

The integral form above is used here as a gray-body radiometric surrogate for the LWIR image-formation process. This keeps the derivation connected to thermal radiation physics, while avoiding the need to explicitly model a full wavelength-dependent detector response for each rendered image.

For an infinitesimal mesh area on the target dA_t, the emitted flux is distributed over a hemisphere with total amount given by (14). An infinitesimal camera pixel area dA_p intercepts only a portion of that flux, governed by the solid angle between the two surfaces. Let $\mathbf{r}$ denote the vector between the two surface elements, and $\mathbf{n}_p$ and $\mathbf{n}_t$ the normals of the pixel and target surface, respectively. As shown in Fig. 1, the projected areas introduce cosine terms through the angles α_t and α_p between each normal and the line-of-sight direction.

Using this convention, the heat flux emitted by dA_t and intercepted by dA_p can be written as:

$$dQ_{tp} = \frac{1}{\pi} q_{\text{t}}(T) \frac{\cos(\alpha_t)\, dA_t\ \cos(\alpha_p)\, dA_p}{\|\mathbf{r}\|^2}, \tag{15}$$

or equivalently:

$$dQ_{tp} = \frac{1}{\pi} q_{\text{t}}(T) \frac{(\mathbf{n}_t \cdot \mathbf{r}_{tp})(\mathbf{n}_p \cdot \mathbf{r}_{pt})}{\|\mathbf{r}\|^4}\, dA_t\, dA_p. \tag{16}$$

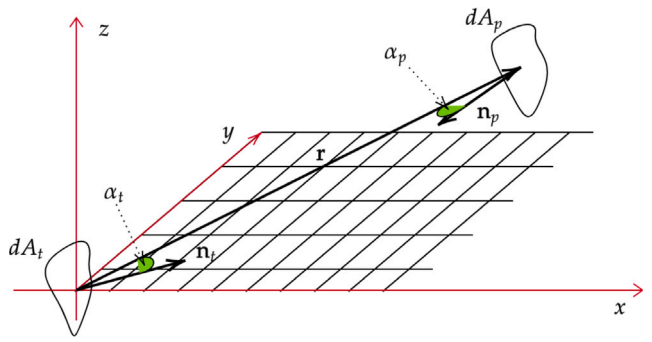

Fig. 1. Mathematical setup for the physics-based constraint. dA_p is an infinitesimal pixel area and dA_t is an infinitesimal target mesh area. $\mathbf{n}_p$ and $\mathbf{n}_t$ denote the corresponding surface normals.

Following the same formulation as in [37], and given the small size of pixels and their close spacing, we generalize this to the camera surface for each target element dA_t:

$$dQ_{tc} = \frac{1}{\pi} q_t(T) \frac{(\mathbf{n}_t \cdot \mathbf{r}_{tc})(\mathbf{n}_c \cdot \mathbf{r}_{ct})}{\|\mathbf{r}\|^4} \, dA_t \, dA_c, \tag{17}$$

which discretizes to:

$$\Delta Q_{tc} = \frac{1}{\pi} q_t(T) \frac{(\mathbf{n}_t \cdot \mathbf{r}_{tc})(\mathbf{n}_c \cdot \mathbf{r}_{ct})}{\|\mathbf{r}\|^4} \, \Delta A_t \Delta A_c. \tag{18}$$

Adding the view-factor approximation and considering a single emissivity value per mesh face [37], the heat flux per unit target area emitted toward the camera becomes:

$$q_{tc} = \frac{\Delta Q_{tc}}{\Delta A_t} = \varepsilon \left(\int_0^\infty \mathbf{B}(\lambda, T) \, d\lambda \right) \frac{(\mathbf{n}_t \cdot \mathbf{r}_{tc})(\mathbf{n}_p \cdot \mathbf{r}_{ct})}{\|\mathbf{r}\|^4} \, \Delta A_c, \tag{19}$$

and the integral is the Stefan–Boltzmann law:

$$\int_0^\infty \mathbf{B}(\lambda, T) \, d\lambda = \frac{2\pi^5 k^4}{15 h^3 c^2} T^4 = \sigma \cdot T^4. \tag{20}$$

Replacing this in (19) gives:

$$q_{tc} = \varepsilon \sigma T^4 \frac{(\mathbf{n}_t \cdot \mathbf{r}_{tc})(\mathbf{n}_p \cdot \mathbf{r}_{ct})}{\|\mathbf{r}\|^4} \, \Delta A_c. \tag{21}$$

Using a simple thermal camera response model, the relation between the digital number $I(x, y)$ in the thermal image and the heat flow q_{tc} is:

$$I(x, y) = a \cdot q_{tc} + b + \eta(x, y), \tag{22}$$

where a is the fixed gain of the sensor response, b is the fixed offset, and $\eta(x, y)$ denotes the sensor noise. We keep these terms in the derivation first, and then return to their practical role after the physics loss is defined.

Substituting the previous expression for q_{tc} yields:

$$\begin{aligned} I(x, y) = a \Bigg[& \varepsilon(x, y) \, \sigma \, T^4 \, \frac{\Delta A_c}{\|\mathbf{r}(x, y)\|^4} \\ & \times \big(\mathbf{n}_t(x, y) \cdot \mathbf{r}_{tc}(x, y)\big) \, \big(\mathbf{n}_c(x, y) \cdot \mathbf{r}_{ct}(x, y)\big) \Bigg] \\ & + b + \eta(x, y). \end{aligned} \tag{23}$$

Here (x, y) denotes the pixel location in the image. Each pixel corresponds to a surface material with emissivity $\varepsilon(x, y)$, observed under a given pose, which is why the geometric terms also inherit the (x, y) dependence.

If the sensor response is calibrated, the gain- and bias-corrected radiometric signal can be written as:

$$\bar{I}(x, y) = \frac{I(x, y) - b - \eta(x, y)}{a}. \tag{24}$$

Using this corrected signal, the temperature relation becomes:

$$\begin{aligned} T^4(x, y) = & \frac{\|\mathbf{r}(x, y)\|^4}{\sigma \, \Delta A_c \, \varepsilon(x, y)} \\ & \times \frac{\bar{I}(x, y)}{\big(\mathbf{n}_t(x, y) \cdot \mathbf{r}_{tc}(x, y)\big) \, \big(\mathbf{n}_c(x, y) \cdot \mathbf{r}_{ct}(x, y)\big)}. \end{aligned} \tag{25}$$

This remains valid as long as the surface is not a perfect non-emitter ($\varepsilon \neq 0$) and the relevant mesh element is visible (angles approaching $\pm\frac{\pi}{2}$ would mean the surface does not appear in the image). We can therefore write:

$$T^4(x, y) = \mathcal{M}(x, y) \cdot \bar{I}(x, y), \tag{26}$$

with:

$$\begin{aligned} \mathcal{M}(x, y) = & \frac{\|\mathbf{r}(x, y)\|^4}{\sigma \, \Delta A_c \, \varepsilon(x, y)} \\ & \times \frac{1}{\big(\mathbf{n}_t(x, y) \cdot \mathbf{r}_{tc}(x, y)\big) \, \big(\mathbf{n}_c(x, y) \cdot \mathbf{r}_{ct}(x, y)\big)}. \end{aligned} \tag{27}$$

(26) means that we can estimate a radiometric temperature proxy from geometry and material properties: the local pose, the relative surface normals, and the emissivity. Since this holds for every pixel coordinate (x, y), we can write the matrix form:

$$\mathbf{T}^4 = \mathbf{M} \odot \bar{\mathbf{I}}^{\text{lwir}}, \tag{28}$$

where $\mathbf{M}$ is what we call the *material tracing map*. This name reflects the fact that it gathers material properties and viewing geometry into a single representation.

The term "tracing" also refers to how the supervisory signal for this map is constructed in Blender. Depth, material labels, and surface normals are rendered from known viewpoints and combined through (27). However, the material tracing map is not assumed to be known during inference. At inference time, the map is estimated from the visible image by a learned network, and it is then used as a physics-guidance signal rather than as an externally provided measurement of target emissivity, distance, or surface normals. Additionally, an important point should be clarified. This model-based formulation does not imply a cooperative target. While prior geometric or material information can be used offline to train the estimator, no active target signal or ground-truth tracing map is assumed during inference [5].

We start by constructing a dedicated dataset. The process begins with a 3D model of the target; in our case, an Envisat model. We place viewpoints around the target covering azimuth and elevation in 10° increments, with radii in $\{50, 55, \ldots, 100\}$ meters. From each viewpoint, we render a set of `.EXR` outputs, notably a depth map, a material index map, and surface normals. These renders are then combined through the model in (27) using a Hadamard product $\odot$ (element-wise multiplication). The resulting dataset contains rendered visible images paired with their material tracing maps as targets.

This dataset is used only to train the tracing-map estimator. It is not used to provide ground-truth tracing maps during the final thermal-image generation experiments. This is important because the rendezvous thermal dataset and the Blender tracing-map dataset are not pixel-aligned in a way that would justify applying Blender ground-truth tracing maps directly to the thermal labels. Therefore, in the proposed pipeline, the tracing map used in the physics constraint is always the estimated map produced from the visible image.

We train a U-Net-based architecture (similar in spirit to the refinement model) that takes the rendered visible image and predicts the corresponding material tracing map. To reduce dynamic range effects and improve numerical stability, we predict the fourth-root of the map, and then use it to estimate $\mathbf{T}$ rather than $\mathbf{T}^4$. We also apply output normalization, alongside online data augmentation on the input, to enforce better generalization. The loss function is a Charbonnier loss [48]:

$$\mathcal{L}_{\text{charb}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\big(\hat{\mathbf{M}}_i - \mathbf{M}_i\big)^2 + \epsilon^2}, \tag{29}$$

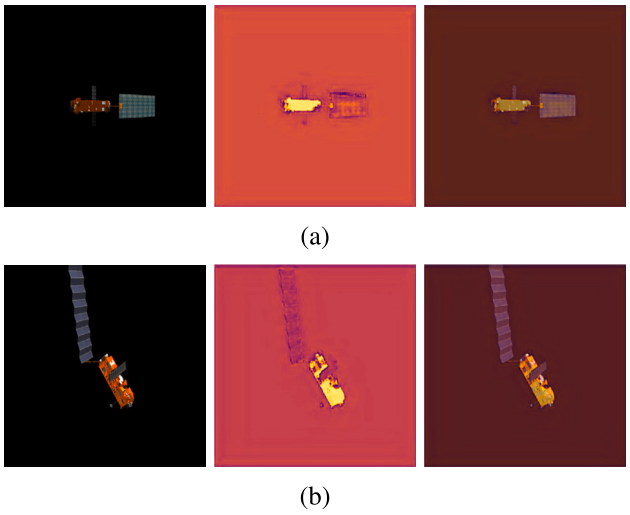

Fig. 2. Qualitative examples of the tracing map estimator applied to Astos samples. In each row, the left image is the visible RGB input, the middle image is the predicted trace map, and the right image is the trace map overlaid on the RGB image. Since Astos is not pixel-aligned with the Blender tracing-map dataset, these examples show predicted maps only.

Table 2
Representative emissivity values used in the Blender model for constructing the material tracing map dataset. Values are indicative and serve to enable the physics-based optimization of the refinement network.

Blender material	ϵ
Gold Foil	0.03
Orange Foil	0.03
Teflon	0.88
Satellite Base	0.15
Satellite Dish	0.05
Reflective Solar Panel	0.05
Solar Panel base	0.10

where $\hat{\mathbf{M}}$ is the estimated tracing map, $\mathbf{M}$ is the ground truth, and $\epsilon > 0$ is a small constant.

Edges are also enforced through the Image Gradient Difference Loss (GDL) [49]:

$$\mathcal{L}_{\text{edge}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left|\partial_x\hat{\mathbf{M}}_i - \partial_x\mathbf{M}_i\right| + \left|\partial_y\hat{\mathbf{M}}_i - \partial_y\mathbf{M}_i\right|\right). \tag{30}$$

The combined loss is then:

$$\mathcal{L} = \mathcal{L}_{\text{charb}} + \lambda_{\text{edge}}\,\mathcal{L}_{\text{edge}}, \tag{31}$$

with $\lambda_{\text{edge}} = 0.05$. Training is performed until convergence. Table 2 reports a subset of the emissivity values used in the Blender model, and Fig. 2 shows qualitative results of the estimator on samples from the Astos dataset [6]. Finally, the overall training setup of the material tracing-map estimator is summarized in Fig. 3(a).

To further validate the material tracing-map estimator, we evaluate it on the held-out Blender validation split used during its training. This validation set is generated with the same rendering procedure as the training data, but with viewpoints chosen to test interpolation rather than exact repetition of the training grid. In particular, azimuth and elevation are organized in 10° bins, and the validation set includes samples from every angular bin. The camera distance is sampled between the radii used for training, for example between 50 m and 55 m, so that the estimator is evaluated on intermediate ranges while preserving pixel alignment between the rendered visible image and the tracing-map target (see Table 3).

Table 3
Training and validation configuration of the material tracing-map estimator.

Item	Configuration
Input/target	RGB render → material tracing map
Resolution	512 × 512
Total paired renders	7128
Training/validation split	6415/713
Split ratio/seed	0.9/42
Validation sampling	All 10° azimuth/elevation bins
Validation distance	Random intermediate radii between training distances
Target representation	Normalized $\mathbf{M}^{1/4}$
Loss	Charbonnier + 0.05 GDL
Backbone	U-Net, 64 base channels, 2 residual blocks
Attention resolutions	16 and 8
Parameters	22.87M
Checkpoint used	EMA weights of the best validation model

Table 4
Quantitative evaluation of the material tracing-map estimator on the aligned Blender validation split. Metrics are reported in the normalized fourth-root tracing-map space used during training.

Region	MAE ↓	RMSE ↓	PSNR ↑	Charbonnier ↓
Full image	0.000266	0.004294	47.3424 dB	0.001243
Foreground	0.010871	0.027447	31.2302 dB	0.011126
Background	0.0000099	0.000836	61.5535 dB	0.001004
Boundary region	0.016767	0.042598	27.4122 dB	0.017167

The results in Table 4 show that the estimator accurately reconstructs the normalized fourth-root tracing map on unseen aligned renders, reaching a full-image MAE of 2.66×10^{-4} and a PSNR of 47.34 dB. Since most of the image corresponds to background pixels where the tracing-map target is close to zero, we also report foreground and boundary-region errors. These regions are more demanding because they contain the spacecraft structure, material transitions, and projected geometric discontinuities that are later used by the physics-based refinement loss.

As expected, the boundary region gives the highest error, with an MAE of 1.68×10^{-2} and a PSNR of 27.41 dB, while the foreground region reaches an MAE of 1.09×10^{-2}. The mean GDL over the validation images is 4.59×10^{-4}, showing that the estimator preserves the main spatial gradients of the tracing map. This is important because the physics constraint does not only depend on the global magnitude of $\hat{\mathbf{M}}$, but also on its ability to localize target shape changes and material-dependent transitions.

We further report material-wise errors in Table 5. Material ID 0 corresponds to the background and dominates the pixel count, which explains its very low error. The remaining IDs correspond to target materials and are therefore more informative for the tracing-map estimator. Across the foreground material IDs, the MAE remains in the range of 5.18×10^{-3} to 6.14×10^{-2}, with the lowest errors obtained on the most represented target materials. Higher errors are observed on rare material IDs, which is expected because these regions occupy fewer pixels and often correspond to small components or sharp material transitions. This material-wise behavior is consistent with the foreground and boundary-region errors reported above, and confirms that the estimator remains accurate on the spacecraft regions that are later used by the physics-based refinement loss.

These validation results are reported only on the aligned Blender data, where ground-truth tracing maps are physically available. They are not used as ground-truth tracing-map scores on the Astos thermal-generation test set, since the Astos thermal images and the Blender tracing-map renders are not pixel-aligned. In the final visible-to-thermal generation pipeline, the tracing map is therefore always predicted from the visible image and used as a learned physics-guidance signal during refinement.

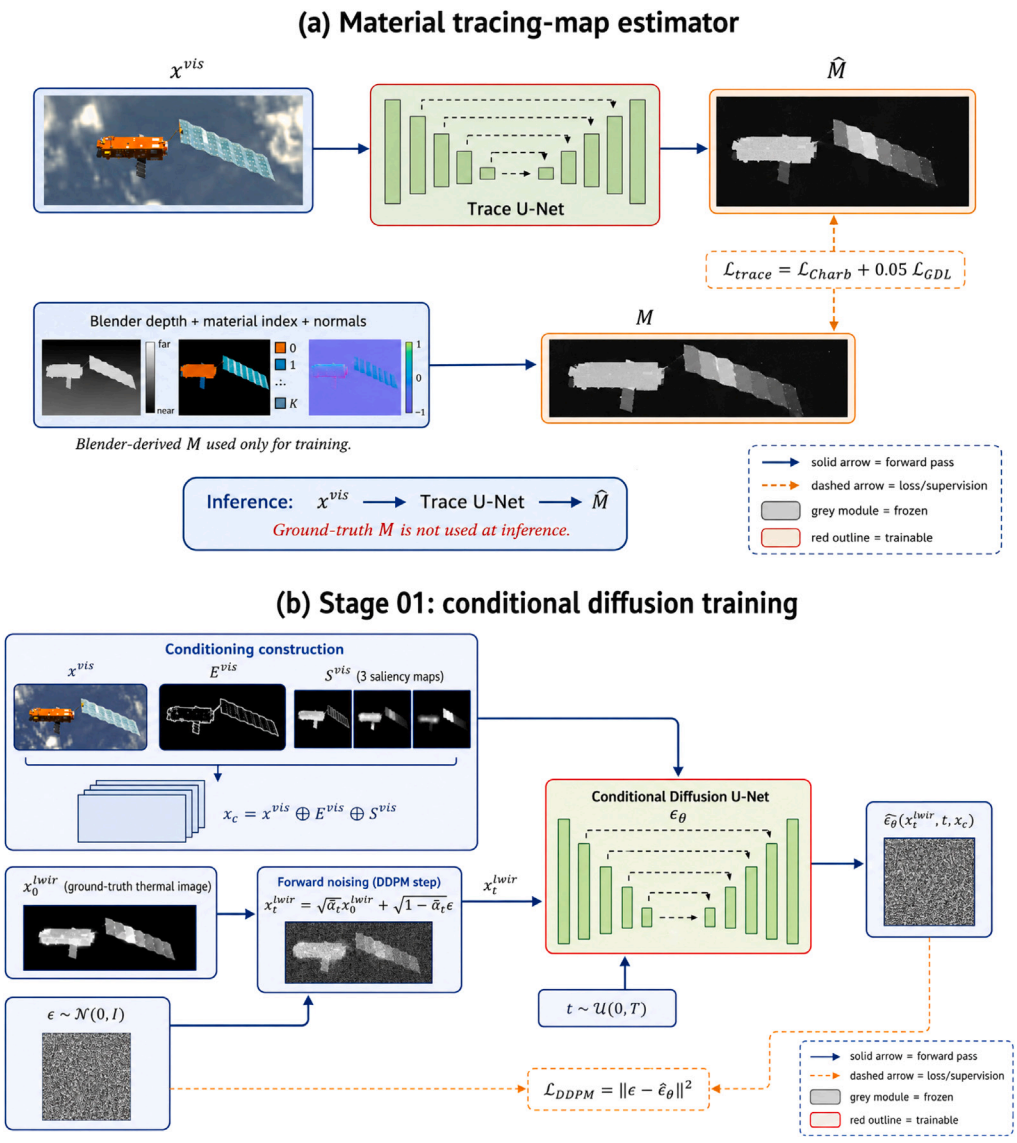

Fig. 3. Overview of the auxiliary tracing-map estimator and the Stage 01 conditional diffusion training. (a) The material tracing-map estimator is trained using Blender-derived supervision from depth, material-index, and surface-normal renders. (b) The Stage 01 conditional diffusion model is trained with the standard DDPM noise-prediction objective using the five-channel visible-derived conditioning tensor. No physics loss is applied during this stage.

Table 5
Material-wise evaluation of the material tracing-map estimator on the aligned Blender validation split. Metrics are reported in the normalized fourth-root tracing-map space used during training. ID 0 corresponds to the background.

ID	Pixels	MAE ↓	RMSE ↓	PSNR ↑	ID	Pixels	MAE ↓	RMSE ↓	PSNR ↑
0	182 309.22	0.000008	0.000590	64.5861	11	1873	0.042287	0.071715	22.8878
1	548	0.032578	0.056871	24.9022	12	15 288	0.018675	0.041197	27.7027
2	9870	0.033372	0.052945	25.5236	13	4054	0.040102	0.073297	22.6982
3	41 790	0.018546	0.034119	29.3402	14	2426	0.042078	0.077180	22.2499
4	86 519	0.022545	0.047848	26.4028	15	16 036	0.024853	0.051770	25.7184
5	1587	0.038728	0.064418	23.8199	16	1 338 160	0.007832	0.017134	35.3228
6	170	0.061418	0.077764	22.1844	17	1 269 563	0.005183	0.015105	36.4179
7	1 063 907	0.017278	0.039155	28.1443	18	8155	0.028804	0.061216	24.2627
8	324 028	0.008811	0.025913	31.7295	19	221 230	0.009415	0.024607	32.1788
9	3337	0.048975	0.082181	21.7045	20	119 247	0.012619	0.028012	31.0532
10	38 579	0.025666	0.050352	25.9596	21	33 183	0.025257	0.047553	26.4564

Once the estimated material tracing map is available, the same map is applied to both the generated thermal image and the reference thermal image. This gives two radiometric temperature proxies:

$$\hat{\mathbf{T}}_{\text{gen}} = \left(\hat{\mathbf{M}} \odot \hat{\mathbf{I}}_{01}^{\text{lwir}}\right)^{1/4}, \qquad \hat{\mathbf{T}}_{\text{gt}} = \left(\hat{\mathbf{M}} \odot \mathbf{I}_{01}^{\text{lwir}}\right)^{1/4}, \tag{32}$$

where $\hat{\mathbf{I}}_{01}^{\text{lwir}}$ and $\mathbf{I}_{01}^{\text{lwir}}$ denote the generated and reference thermal images mapped to $[0, 1]$, and $\hat{\mathbf{M}}$ is the estimated material tracing map. The physics-based refinement loss is then applied between these two proxies:

$$\mathcal{L}_{\text{physics}} = \left\| \hat{\mathbf{T}}_{\text{gen}} - \hat{\mathbf{T}}_{\text{gt}} \right\|_1 . \tag{33}$$

We can now explain the practical choice of the camera-response terms in (22). In principle, a, b, and the statistics of $\eta(x, y)$ can be obtained through sensor calibration. In the present work, however, the physics constraint is applied to normalized thermal images generated under the same camera model. Therefore, the same response model affects both branches of (33): the generated image and the reference image.

The gain a does not introduce an additional spatial constraint. After bias correction, it appears as a global scale factor in the calibrated signal. In the fourth-root formulation used in (32), this becomes a constant multiplicative factor $a^{-1/4}$ applied to both temperature proxies. Such a constant factor only rescales $\mathcal{L}_{\text{physics}}$ and can be absorbed into the physics-loss weight. Therefore, explicitly changing a would be equivalent to reweighting the same loss term rather than introducing a new physical effect.

The offset b is also not a learnable or image-dependent quantity in this setting. It defines the origin of the camera digital response and is removed when forming the calibrated intensity in (24). Since the generated and reference thermal images are compared in the same normalized image domain, using $b = 0$ corresponds to working directly with the bias-corrected normalized signal. Introducing an arbitrary non-zero b without an independent calibration would shift both branches of the loss by the same fixed sensor offset and would not provide additional information to guide the refinement network.

The noise term $\eta(x, y)$ is different from a and b because it is not a deterministic calibration constant. We assume it is zero-mean, i.e., $\mathbb{E}[\eta(x, y)] = 0$, and define the default physics constraint using the mean sensor response. This is why the main formulation uses $\eta(x, y) = 0$. Adding a random realization of the noise directly inside (33) would make the supervision target stochastic rather than correcting a systematic physical factor. For this reason, sensor noise is not included in the default loss, but its effect can be studied separately through controlled robustness ablations using representative microbolometer noise levels.

With this formulation, setting $a = 1$, $b = 0$, and $\eta(x, y) = 0$ should be understood as working in the normalized, mean-response radiometric domain, not as claiming that a real detector has unit gain, zero electronic offset, and no noise. The quantities ϵ, $\mathbf{r}$, $\mathbf{n}_t$, and $\mathbf{n}_c$ are used to generate supervised tracing-map targets in the rendered auxiliary dataset, while the operational pipeline only receives the visible image and predicts $\hat{\mathbf{M}}$ from it. The resulting physics constraint therefore remains compatible with the image-to-image generation setting while preserving the material and geometric structure of the thermal image-formation model.

3.5. Refinement network architecture and training

This subsection completes the proposed architecture by introducing the refinement network and its training setup. The optimization framework is then defined by combining classical reconstruction terms with the physics-based constraint introduced above, in order to refine the coarse samples produced by the diffusion model.

Table 6
Architecture of the refinement U-Net $\mathcal{R}_\psi$. The network follows the same multi-scale hierarchy as the diffusion backbone but does not use time conditioning, as it operates on a deterministic coarse sample concatenated with the conditioning tensor. Attention is applied at the lowest resolutions.

Block	Resolution	Channels
Input/conditioning	$H \times W$	$1 + C_{\text{cond}}$
Enc-0	$H \times W$	64
Enc-1	$H/2 \times W/2$	128
Enc-2 (attn)	$H/4 \times W/4$	256
Enc-3 (attn)	$H/8 \times W/8$	256
Bottleneck (attn)	$H/16 \times W/16$	512
Dec-3 (attn)	$H/8 \times W/8$	256
Dec-2	$H/4 \times W/4$	256
Dec-1	$H/2 \times W/2$	128
Dec-0	$H \times W$	64
Output	$H \times W$	1

Refinement U-net $\mathcal{R}_\psi$

The second stage of our system is a time-independent refinement network $\mathcal{R}_\psi$ that operates on the coarse thermal sample produced by the diffusion model. Given a coarse prediction $\tilde{\mathbf{x}}_0^{\text{lwir}}$ from the DPM-Solver sampler and the same conditioning tensor $\mathbf{x}_c$, the refiner predicts a residual correction:

$$\hat{\mathbf{x}}_0^{\text{lwir}} = \tilde{\mathbf{x}}_0^{\text{lwir}} + \mathcal{R}_\psi\left(\tilde{\mathbf{x}}_0^{\text{lwir}}, \mathbf{x}_c\right). \tag{34}$$

Equivalently, the model is trained to correct and refine the first thermal sample produced by DPM-Solver.

Architecturally, $\mathcal{R}_\psi$ reuses the diffusion U-Net hierarchy but omits the time-embedding pathway. As such, there is no dependence on the diffusion time index: the network only sees a deterministic coarse input and the conditioning tensor. The refinement U-Net is instantiated with the same base width, channel multipliers, and attention placement as the diffusion model.

Table 6 summarizes the architecture. The input now consists of the coarse thermal prediction concatenated with the conditioning tensor. As in the diffusion backbone, residual blocks are built from GroupNorm, SiLU, and convolutional layers, while attention is used at the coarsest resolutions.

Using a second U-Net to refine diffusion outputs serves two practical purposes. First, it decouples global generative modeling from local restoration: the diffusion backbone ϵ_θ handles sampling a plausible global thermal configuration under the forward–reverse diffusion dynamics, while $\mathcal{R}_\psi$ focuses on sharpening edges, correcting local artifacts, and enforcing the physics and feature-based constraints that are not explicitly present in standard diffusion training. This separation has proven effective in high-resolution synthesis, super-resolution, and 3D completion [50–52]. Second, by removing time conditioning, the refinement model is easier to optimize while remaining expressive, since it effectively operates at a fixed "noise level" given by the coarse sample distribution.

Conditioning the refiner on the same visible-derived tensor $\mathbf{x}_c$ ensures that corrections remain consistent with the external modality. In practice, this improves alignment between visible edges and thermal boundaries, enhances reconstruction of fine structures, and strengthens adherence to the physics-based constraint during training. The lightweight transition from ϵ_θ to $\mathcal{R}_\psi$ also keeps the two-stage design easy to reproduce and analyze in ablations.

In all experiments, both U-Nets operate on 512×512 inputs with base width 64 and two residual blocks per resolution, which provided a good balance between fidelity and computational cost for thermal image generation. The Stage 02 physics-guided refinement scheme is shown in Fig. 4.

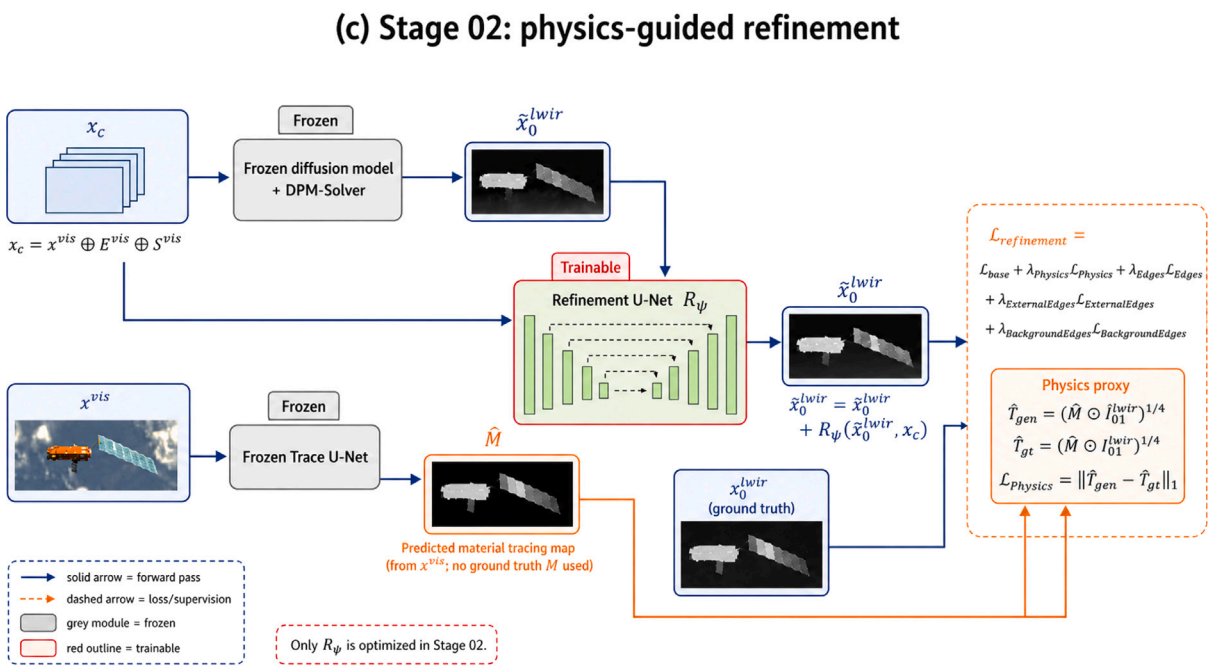

Fig. 4. Overview of the Stage 02 physics-guided refinement scheme. The trained diffusion model and the tracing-map estimator are frozen, while only the refinement network R_ψ is optimized. The frozen diffusion sampler produces a coarse LWIR estimate $\tilde{x}_0^{\text{lwir}}$ from the visible-derived conditioning tensor, and the frozen Trace U-Net predicts the material tracing map $\hat{M}$ from the visible image. The refinement network then produces the final thermal image $\hat{x}_0^{\text{lwir}}$, which is supervised using reconstruction, edge, background, and physics-based losses.

Optimization

For the second-stage optimization, we apply different constraints to the refinement network, separating losses on the foreground (the target spacecraft) from losses on the background (empty space, partial Earth, or full Earth). This choice follows directly from the modeling in the physics-based constraint: the tracing-map estimator is designed to capture target-specific geometry and material properties, and the assumptions made above are valid for the target in space, not for the background. Also, for space relative navigation, the background mainly acts as clutter; the relevant algorithm still needs to isolate the target and recover its pose from the frame.

We start with a standard $\mathcal{L}_1$ reconstruction term between the refined prediction and the ground truth:

$$\mathcal{L}_{\text{base}} = \left\| \hat{\mathbf{x}}_0^{\text{lwir}} - \mathbf{x}_0^{\text{lwir}} \right\|_1 . \tag{35}$$

Using the material tracing map, we can estimate temperature on the target region of the generated thermal image. The corresponding physics-based loss, applied to the foreground only, is:

$$\mathcal{L}_{\text{Physics}} = \left\| \hat{\mathbf{T}}_{\text{Target}} - \mathbf{T}_{\text{Target}} \right\|_1 . \tag{36}$$

Crisp edges are also needed on the target. We therefore build a model similar to [39] to extract spacecraft edges only. The additional foreground edge loss is:

$$\mathcal{L}_{\text{Edges}} = \left\| \mathcal{E}(\hat{\mathbf{x}}_0^{\text{lwir}}) - \mathcal{E}(\mathbf{x}_0^{\text{lwir}}) \right\|_1 . \tag{37}$$

External edges, i.e., the outer contour capturing the overall target silhouette, are further emphasized through:

$$\mathcal{L}_{\text{External Edges}} = \left\| \mathcal{E}_{\text{ext}}(\hat{\mathbf{x}}_0^{\text{lwir}}) - \mathcal{E}_{\text{ext}}(\mathbf{x}_0^{\text{lwir}}) \right\|_1 , \tag{38}$$

as they are particularly important for pose estimation.

Finally, we add an Image GDL term, but only on the background:

$$\mathcal{L}_{\text{Background Edges}} = \frac{1}{N} \sum_{i=1}^{N} \left(\left| \partial_x \hat{\mathbf{B}}_i - \partial_x \mathbf{B}_i \right| + \left| \partial_y \hat{\mathbf{B}}_i - \partial_y \mathbf{B}_i \right| \right), \tag{39}$$

where $\hat{\mathbf{B}}$ and $\mathbf{B}$ denote the background regions of $\hat{\mathbf{x}}_0^{\text{lwir}}$ and $\mathbf{x}_0^{\text{lwir}}$, respectively. The background is therefore not ignored. It is not assigned the foreground material-tracing physics model, but it remains constrained as an appearance- and pairing-consistency component through the full-frame reconstruction term and the background GDL. This is important because the generated images are intended to remain useful as paired visible/LWIR samples for downstream relative navigation algorithms.

The final combined objective for refinement training is:

$$\begin{aligned} \mathcal{L}_{\text{Refinement}} = \mathcal{L}_{\text{base}} &+ \lambda_{\text{Physics}} \cdot \mathcal{L}_{\text{Physics}} + \lambda_{\text{Edges}} \cdot \mathcal{L}_{\text{Edges}} \\ &+ \lambda_{\text{External Edges}} \cdot \mathcal{L}_{\text{External Edges}} \\ &+ \lambda_{\text{Background Edges}} \cdot \mathcal{L}_{\text{Background Edges}} . \end{aligned} \tag{40}$$

The loss weights are set to $\lambda_{\text{Physics}} = 100$, $\lambda_{\text{Edges}} = 20$, $\lambda_{\text{External Edges}} = 25$, and $\lambda_{\text{Background Edges}} = 0.05$. The complete training and inference configuration is summarized in Table 7. Stage 01 trains the conditional diffusion model with the standard noise-prediction objective. Stage 02 then freezes the diffusion model and optimizes only the refinement network using the objective in (40). This separation ensures that the diffusion model learns a conditional thermal prior, while the refinement network is responsible for correcting sampled outputs and enforcing the additional reconstruction, edge, background, and physics-based constraints.

All training experiments were performed on an NVIDIA V100 Tensor Core GPU. Runtime and computational-cost measurements are reported separately in Section 4 using the inference benchmark setting described there.

Regarding the Astos dataset, the data is composed of 13 rendezvous trajectories with different tumbling and approach profiles [6]. To reduce the correlation between adjacent frames, the data are not split using consecutive individual frames. Instead, each trajectory is divided into intervals of 100 frames, and the train, validation, and test sets are constructed from these intervals. This interval-based split increases the visual and pose difference between samples assigned to different subsets compared with a purely frame-level split, although we note that

Table 7
Training and inference configuration of the proposed pipeline.

Configuration	Stage 01 diffusion model	Stage 02 refinement model
Input resolution	512 × 512	512 × 512
Input/condition	Noisy LWIR image + 5-channel visible tensor	Coarse LWIR image + 5-channel visible tensor
Output	Predicted Gaussian noise	Refined LWIR image
Training dataset	Astos paired visible/LWIR images	Astos paired visible/LWIR images
Data split	13 trajectories split into 100-frame intervals, 80/10/10	Same split as Stage 01
Training epochs	300	100
Diffusion timesteps	$T = 1000$	Frozen Stage 01 model
Sampling method	–	DPM-Solver (3rd order)
Sampling steps	–	10
Noise schedule	Cosine schedule, $m = 0.008$	Frozen Stage 01 schedule
Optimizer	AdamW	AdamW
Peak learning rate	2×10^{-4}	1×10^{-4}
Warmup steps	1500	500
Weight decay	10^{-4}	10^{-4}
Batch size	64	64
EMA decay	0.999	0.999
Main objective	DDPM noise-prediction MSE	(40)
Training hardware	NVIDIA V100 Tensor Core GPU	NVIDIA V100 Tensor Core GPU

Table 8
Quantitative comparison against state-of-the-art visible-to-thermal translation methods on Astos. ↑ indicates higher is better; ↓ indicates lower is better. Best results are shown in bold.

Model	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Pix2Pix	37.6673	0.9548	0.0232	34.1598 dB
CycleGAN	78.9041	0.9528	0.0464	26.9977 dB
VSAIT	73.1865	0.9482	0.0636	26.9728 dB
ECDM	95.0828	0.3577	0.6307	19.2171 dB
LPTN	200.1153	0.9309	0.1522	25.3317 dB
Ours	**25.5570**	**0.9838**	**0.0161**	**40.3766** dB

Table 9
Quantitative comparison between the coarse diffusion output and the final refined output of our pipeline. All metrics improve significantly after refinement, both in terms of structure and perceptual similarity, as well as distributional closeness to the ground truth.

Model	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Diffusion coarse output	69.5192	0.4347	0.4577	14.5102 dB
Refined output	25.5570	0.9838	0.0161	40.3766 dB

the end of one interval may still resemble the beginning of the following one due to the continuous nature of rendezvous trajectories. The same split is used for the diffusion model, the refinement model, the direct U-Net ablations, and the evaluated baselines.

4. Results

In the following section, we present the results obtained with the proposed physics-based diffusion model. We first describe the evaluation metrics and report the main comparison against state-of-the-art visible-to-thermal baselines. We then add controlled experiments to isolate the effect of the five-channel conditioning tensor, direct regression, the refinement network, and the physics-based constraint. Finally, we report computational cost and discuss the practical trade-off between image fidelity and inference time.

4.1. Performance metrics

To validate our model's performance for visible-to-thermal generation under the chosen set of conditioning features, we rely on the most common evaluation metrics used in the literature: FID [53], PSNR [54], LPIPS [55], and SSIM [56]. For the implementation, FID is computed using an Inception-v3 network [57], while LPIPS is computed using a VGG backbone [58]. Since both networks expect three-channel inputs, each thermal image is duplicated into three channels for feature extraction. Unless stated otherwise, all scores are reported over a batch of 1000 test images.

Quantitative results for the full pipeline and for the coarse diffusion output are reported in Table 9, with representative qualitative examples shown in Fig. 6.

We also evaluate the proposed approach against representative state-of-the-art baselines: Pix2Pix [23], CycleGAN [22], VSAIT [59], ECDM [18], and LPTN [60]. For all methods, we use the grayscale visible image as the conditioning input. Each baseline is trained following the procedure described in its original work, and for the same number of iterations as our model to keep the comparison fair. Qualitative comparisons are shown in Fig. 5, while quantitative results are summarized in Table 8. As shown in the table, our method achieves the best performance across all four metrics on the Astos dataset, reflecting improvements in both structural fidelity and perceptual similarity.

4.2. Effect of multi-channel conditioning on baseline models

In the initial comparison against state-of-the-art visible-to-thermal translation methods, the baseline models were evaluated using only the grayscale visible image as input. However, the proposed method uses a richer conditioning tensor that includes the grayscale visible image, an edge map, and three saliency maps. To make the comparison more balanced and to verify whether the performance gain is only caused by the additional input channels, we retrain the competing methods with the same five-channel conditioning tensor. This additional experiment isolates the effect of the input representation on the baseline models, while keeping the proposed architecture unchanged.

The results are reported in Table 10. All methods are evaluated on the same 1000 test images as the earlier test, and the input tensor is composed of the visible grayscale image, the edge map, and the three saliency maps. As expected, some methods benefit from the additional conditioning information. Pix2Pix improves compared with its grayscale-only version, reaching 35.80 dB PSNR and 0.9884 SSIM. LPTN also improves in terms of FID compared with the grayscale-input setting, although its PSNR and LPIPS remain lower than those of Pix2Pix.

The five-channel experiment shows that simply increasing the amount of visible-domain conditioning is not sufficient to obtain the same behavior across all architectures. Pix2Pix benefits the most from the additional information and becomes the strongest feed-forward baseline in this setting. However, CycleGAN, VSAIT, and ECDM do not show the same improvement, which suggests that the ability to exploit edge and saliency information depends strongly on the architecture and training

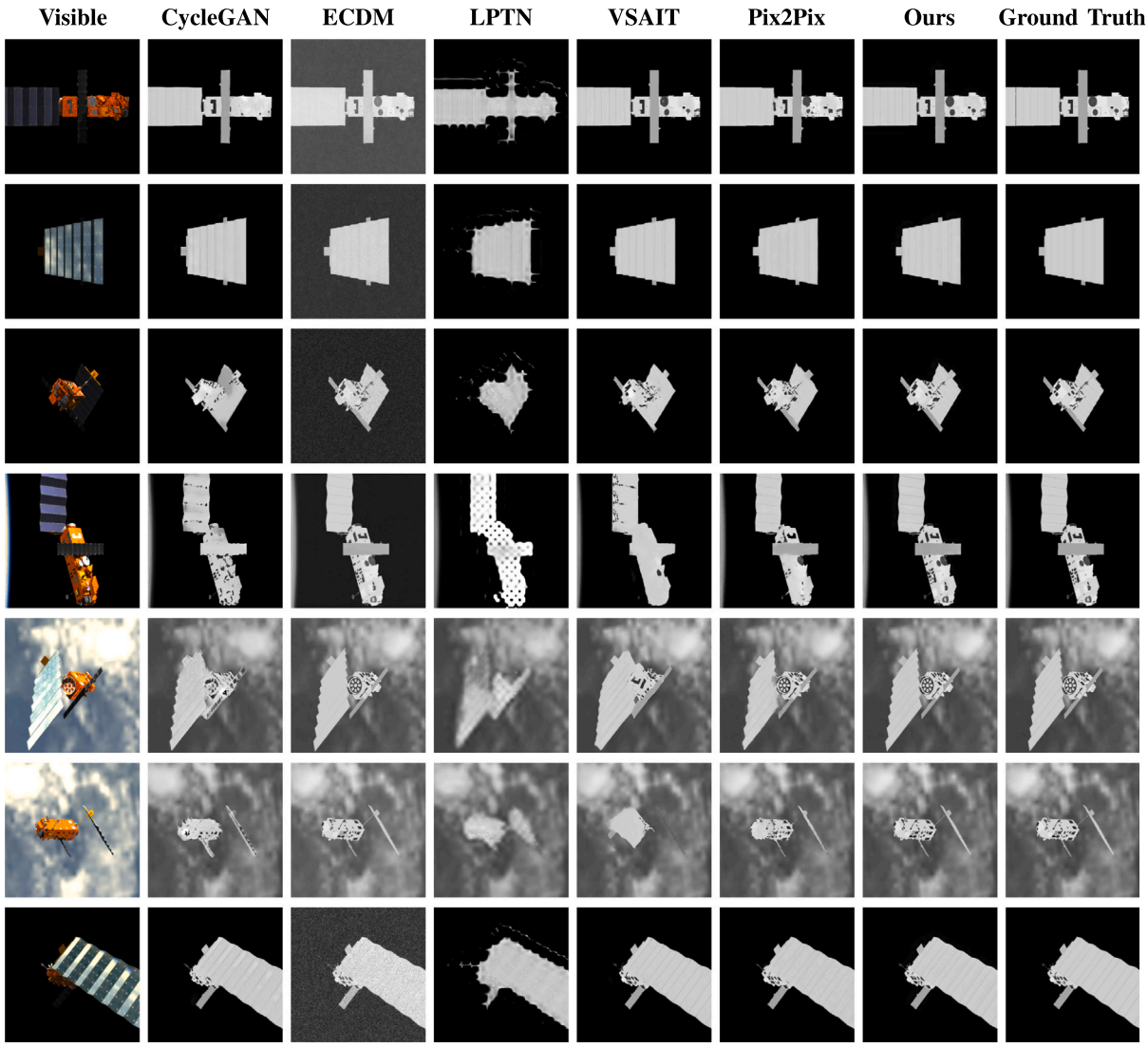

Fig. 5. Qualitative comparison on Astos test samples. Our method produces thermal images that remain tightly aligned with the visible structure and closely match the ground truth appearance. The improvement introduced by the refinement stage over the coarse diffusion output is particularly visible on thin structures and target boundaries. All the images are zoomed-in for better display of the details present on the images.

Table 10
Quantitative comparison of state-of-the-art baselines trained with the same five-channel conditioning tensor used by the proposed diffusion stage. The input is composed of the grayscale visible image, one edge map, and three saliency maps. All results are reported over the 1000 test images.

Model	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Pix2Pix	30.9008	0.9884	0.0210	35.7997 dB
CycleGAN	141.6925	0.9193	0.1486	23.0328 dB
LPTN	130.5298	0.9466	0.1124	26.2335 dB
VSAIT	308.7804	0.2522	0.6798	11.8974 dB
ECDM	125.7830	0.3129	0.7104	14.3955 dB

objective. In particular, the degradation observed for VSAIT and ECDM indicates that adding channels without an explicit mechanism to use them consistently can make the optimization more difficult.

This experiment therefore confirms that the improvement of the proposed method cannot be attributed only to the presence of additional input channels. The proposed pipeline combines multi-channel conditioning with diffusion-based generation, refinement, and the physics-based constraint. The five-channel baseline comparison shows that the conditioning tensor alone improves some baselines, especially Pix2Pix, but does not close the gap with the full physics-based diffusion pipeline reported in the main comparison.

4.3. Ablation study: Direct regression, refinement, and physics

After evaluating the effect of the five-channel conditioning tensor on the baseline models, we further isolate the contribution of the main components of the proposed pipeline. The goal of this ablation is to distinguish between three effects: direct supervised visible-to-thermal regression, diffusion-based generation followed by refinement, and the material-tracing physics constraint. To this end, we compare a direct U-Net trained with the same five-channel input, its physics-constrained version, the coarse diffusion output, the full cascade without the physics loss, and the complete proposed model. All variants are evaluated under the same test protocol.

The results in Table 11 first show that direct regression is already a strong baseline when the same visible-derived conditioning tensor is used. The direct U-Net reaches 36.95 dB PSNR and 0.9430 SSIM, confirming that the visible image, edge map, and saliency maps contain enough structural information to learn a significant part of the synthetic visible-to-thermal mapping. Adding the material-tracing physics

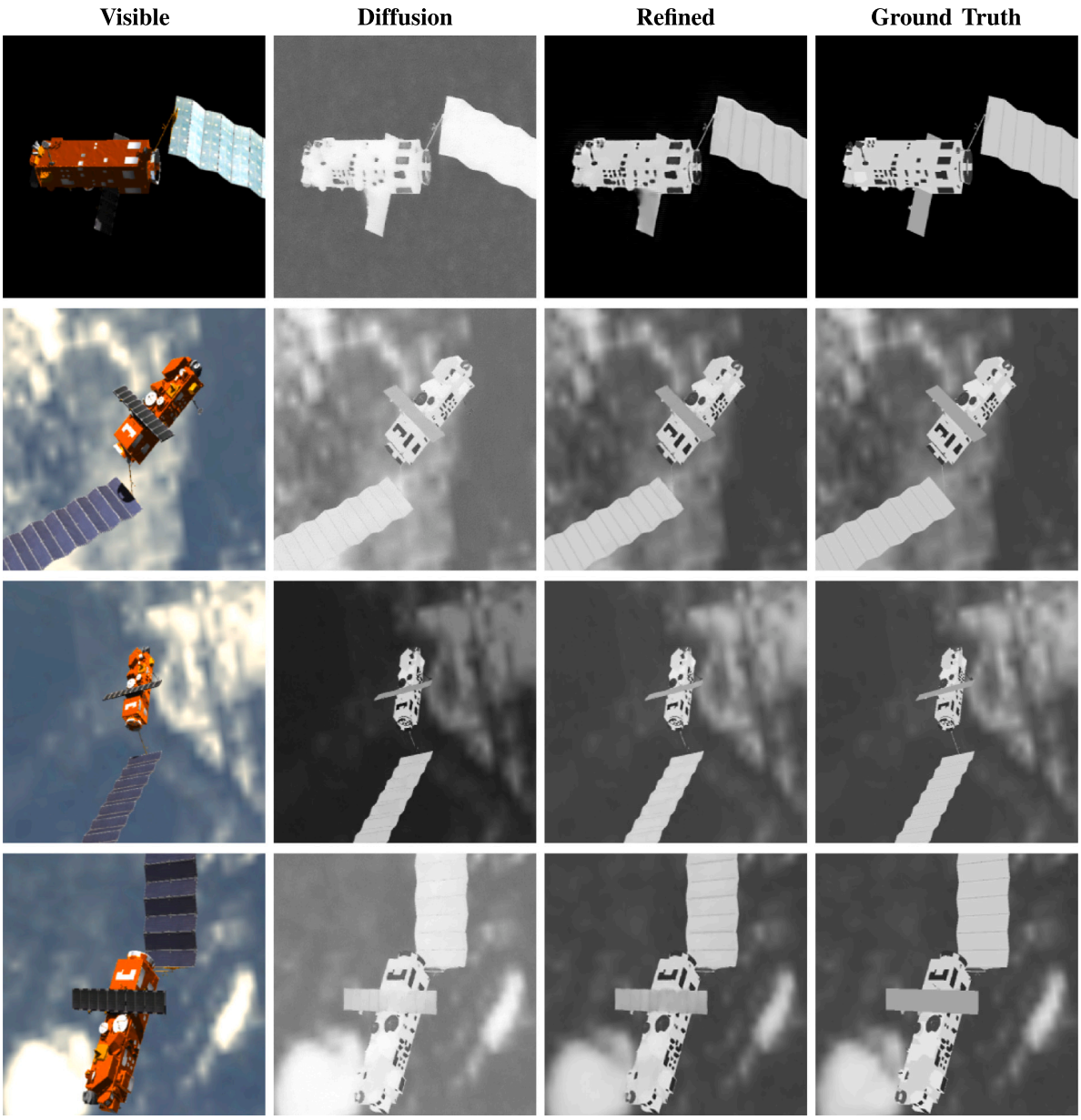

Fig. 6. Zoomed-in qualitative results on Astos test samples. The refinement stage improves the coarse diffusion output by sharpening target boundaries, reducing background artifacts, and better matching the intensity distribution of the ground truth thermal images.

Table 11
Ablation study of direct regression, refinement, and the physics-based constraint. The direct U-Net variants use the same five-channel conditioning tensor as the proposed diffusion stage but do not use diffusion sampling or refinement. The full model corresponds to the proposed diffusion-refinement cascade.

Model	Diffusion	Refinement	Physics	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Direct U-Net	✗	✗	✗	40.3526	0.9430	0.0202	36.9473 dB
Direct U-Net + Physics	✗	✗	✓	31.5306	0.9642	0.0195	38.8887 dB
Diffusion coarse output	✓	✗	✗	69.5192	0.4347	0.4577	14.5102 dB
Full model without Physics	✓	✓	✗	45.4410	0.7066	0.0316	34.0978 dB
Full model + Physics	✓	✓	✓	**25.5570**	**0.9838**	**0.0161**	**40.3766** dB

constraint to this direct model improves all metrics, increasing PSNR to 38.89 dB and reducing FID from 40.35 to 31.53. This confirms that the proposed physics term is not only useful inside the final cascade, but also acts as a general thermal-structure regularizer for visible-to-thermal generation.

The same trend appears more clearly in the proposed cascade. The coarse diffusion output alone captures a first thermal estimate, but it remains insufficient in terms of pixel-level accuracy and structural fidelity. Once the refinement network is added, the output becomes significantly closer to the reference thermal image. Without the physics term, the full cascade reaches 34.10 dB PSNR and 0.0316 LPIPS, already improving over the coarse diffusion sample. However, when the physics-based constraint is included, the final model reaches the best performance across all metrics, with 25.56 FID, 0.9838 SSIM, 0.0161 LPIPS, and 40.38 dB PSNR.

This ablation therefore supports two conclusions. First, the improvement cannot be attributed only to the additional five-channel input, since a direct U-Net using the same conditioning remains below the

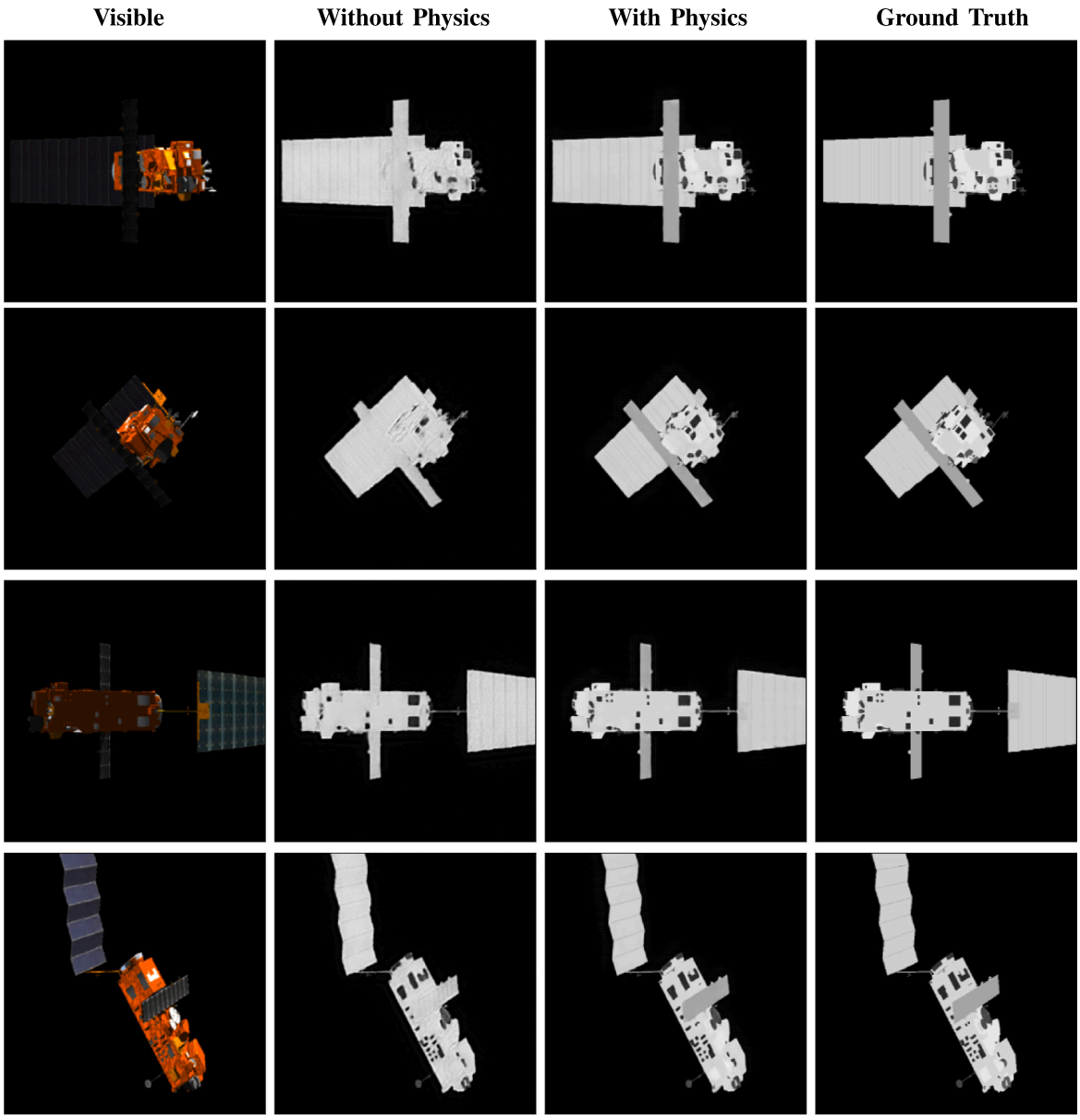

Fig. 7. Zoomed-in qualitative comparison of refinement trained with and without the physics-based constraint. The centered crops focus on the spacecraft target to better highlight local material-dependent thermal patterns and small structural details. Style-based optimization preserves the global structure, but it can miss temperature-related cues on small components and material transitions. The physics-based term improves material-consistent intensity patterns while keeping geometric alignment.

complete model. Second, the material tracing map consistently improves both the direct regressor and the diffusion-refinement cascade. The best results are obtained only when the generative diffusion prior, the refinement network, and the physics-based constraint are combined, which confirms the role of the proposed multi-stage formulation in producing high-fidelity and physically consistent synthetic thermal images (see Fig. 7).

4.4. Robustness to microbolometer noise in the physics constraint

The default physics-based constraint is defined using the mean camera response, where the stochastic sensor noise term is set to zero. To evaluate the sensitivity of the refinement stage to this assumption, we add controlled zero-mean microbolometer noise inside the physics-guidance term and keep the rest of the pipeline unchanged. This experiment tests whether the material-tracing constraint remains useful when the radiometric consistency signal is perturbed by representative detector noise levels.

Table 12
Robustness of the proposed full pipeline to controlled microbolometer noise injected in the physics constraint. The no-noise model corresponds to the default mean-response formulation used in the main experiments.

Model	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Full model, no added noise	**25.5570**	**0.9838**	**0.0161**	**40.3766** dB
Full model + 40 mK noise	44.7666	0.8964	0.0380	36.1834 dB
Full model + 60 mK noise	58.5462	0.6887	0.0682	33.4083 dB
Full model + 80 mK noise	65.1152	0.5984	0.0840	32.0983 dB

The results in Table 12 show a clear monotonic degradation as the injected noise level increases. With 40 mK noise, the model still

Table 13
Computational comparison of the proposed method and state-of-the-art visible-to-thermal translation baselines. All methods are evaluated at 512 × 512 resolution with batch size equal to one. The proposed method is also reported stage-by-stage to show the contribution of the diffusion sampler and the refinement network.

Model	Cond. ch.	Steps	Params (M)	FLOPs (G)	Latency (ms) ↓	FPS ↑
Pix2Pix	5	1	54.40	143.29	25.78	38.79
CycleGAN	5	1	11.37	454.91	150.39	6.65
LPTN	5	1	0.31	3.66	7.97	125.42
VSAIT	5	1	23.16	534.70	156.79	6.38
ECDM	5	10	56.58	51 559.09	15 817.62	0.063
Proposed diffusion sampler	5	10	56.51	13 974.45	14 533.76	0.069
Proposed refinement network	5	1	22.87	1073.88	411.39	2.43
Proposed full pipeline	5 + 5	10	79.44	15 048.32	14 620.29	0.068

preserves a relatively high PSNR of 36.18 dB, but SSIM decreases to 0.8964 and FID increases to 44.77. At 60 mK and 80 mK, the degradation becomes stronger, indicating that the physics-guided refinement is sensitive to perturbations in the radiometric consistency signal.

This behavior supports the use of the mean-response formulation as the default training objective. The gain and bias terms define deterministic calibration factors, whereas microbolometer noise directly perturbs the physics target used by the refinement loss. Injecting stronger noise therefore weakens the material-tracing guidance and progressively reduces structural and perceptual fidelity. The ablation also provides a practical robustness estimate for the proposed constraint. Therefore, the model remains functional under moderate noise, but accurate radiometric conditioning is also important for preserving the full benefit of the physics-based refinement.

4.5. Computational cost

To complement the reconstruction and perceptual metrics reported above, we also evaluate the computational cost of the proposed method and the competing baselines. This is important because, in a spacecraft rendezvous setting, the generated thermal images are not only expected to be visually accurate, but should also remain computationally tractable for integration in simulation, testing, and navigation pipelines.

All models are evaluated at a resolution of 512 × 512 with a batch size of one on an NVIDIA Quadro P3200 GPU. For the feed-forward baselines, a single forward pass is used. For diffusion-based methods, the reported cost includes the complete sampling procedure with 10 solver steps. In the proposed method, we report three entries: the first-stage diffusion sampler, the second-stage refinement network, and the complete cascade. This separation makes it possible to identify the cost of the generative stage and the additional overhead introduced by the refinement step.

The number of parameters, floating-point operations, inference time, and throughput are reported in Table 13. The FLOPs are computed for one 512 × 512 input image. For the proposed diffusion sampler, the denoising-network cost is multiplied by the number of solver steps. For the complete pipeline, the reported cost is obtained by summing the cost of the diffusion sampler and the refinement network. Since FLOP counting can vary slightly depending on the profiler and the supported operators, inference time is also reported as the main practical indicator of runtime cost.

From Table 13, LPTN is the lightest model in terms of both parameter count and latency, requiring only 0.31M parameters and 7.97 ms per image. Pix2Pix also remains relatively efficient, with 54.40M parameters but only 25.78 ms per image. In contrast, CycleGAN and VSAIT have higher latency despite having fewer parameters than Pix2Pix in the case of CycleGAN, which shows that parameter count alone does not fully characterize runtime behavior.

The diffusion-based models are naturally more expensive because inference requires multiple denoising evaluations rather than a single feed-forward pass. ECDM requires 15.82 s per image with 10 sampling steps. The proposed first-stage sampler requires 14.53 s under the same number of steps, while the refinement stage adds 411.39 ms. Therefore, the full proposed cascade reaches 14.62 s per image. This confirms that most of the inference time is concentrated in the diffusion sampling stage, while the refinement network contributes a relatively small fraction of the total runtime.

Although the complete proposed method is slower than single-pass GAN and CNN-based baselines, this cost must be interpreted together with the image-quality results reported in Tables 8–11. The cascade formulation provides the best reconstruction and perceptual performance, while the refinement network is responsible for most of the improvement from the coarse diffusion output to the final thermal image. Therefore, the proposed method favors generation accuracy and physics consistency over real-time throughput. This trade-off is acceptable for the targeted use case, where the main objective is to generate high-fidelity synthetic thermal imagery for spacecraft rendezvous analysis, dataset generation, and validation of downstream perception algorithms.

4.6. Applicability and synthetic-to-real gap

The experiments reported in this work are performed on synthetic Astos data. The resulting PSNR, SSIM, LPIPS, and FID values should therefore be interpreted as in-domain synthetic results, rather than as a direct guarantee of the same performance on real in-orbit thermal imagery. Real LWIR images can be affected by effects that are only partially represented in the present dataset, including sensor noise, optical blur, distortion, platform vibration, stray light, calibration mismatch, and imperfect background radiance. These factors may reduce the quantitative performance when transferring directly from synthetic data to real observations.

This limitation is also linked to the current state of available data. To the best of our knowledge, there is no publicly available real in-orbit dataset providing pixel-aligned visible/LWIR image pairs for supervised visible-to-thermal generation in spacecraft rendezvous scenarios. The same difficulty also exists in the synthetic domain, since high-fidelity spacecraft thermal rendering tools are generally expensive, difficult to access, or tied to proprietary simulation environments. The objective of the proposed method is therefore to provide a practical synthetic thermal image generation pipeline starting from rendered visible images, with the aim of supporting downstream tasks such as relative navigation and multispectral perception.

The synthetic-to-real gap is not limited to the absence of real paired data. Real LWIR images may also include sensor-specific non-uniformity, calibration drift, optical blur, distortion, motion smear, stray light, background-radiance mismatch, and differences between the assumed and actual thermal state of spacecraft materials [6,7,36]. These effects can alter both local target contrast and the global intensity distribution, and may reduce synthetic-domain metrics such as PSNR, SSIM, LPIPS, and FID when the model is applied directly to real

observations. A practical transfer path would therefore combine sensor-aware degradation, domain randomization, calibration-informed pre-processing, and eventual fine-tuning or validation on laboratory or in-orbit paired visible/LWIR data.

Future work should address this gap by extending the training data with sensor-aware degradations and, more importantly, by acquiring real paired visible/LWIR data under controlled laboratory or in-orbit conditions. Such data would make it possible to evaluate the proposed framework under real radiometric and optical effects, and to adapt the model from synthetic thermal generation toward real-domain thermal image synthesis.

5. Conclusion

In this paper, we introduced a physics-informed diffusion model tailored for thermal image generation in spacecraft rendezvous scenarios. We first highlighted the growing importance of advancing space technologies, and the role that thermal imagery can play in space perception, while pointing out the persistent lack of publicly available multispectral datasets suitable for relative navigation and pose estimation.

The proposed framework relies on conditioning a diffusion model using a multi-feature representation extracted from the visible band. Training follows the standard DDPM objective during the first stage, and is then extended with a second refinement stage where coarse samples produced by DPM-Solver are further improved under multiple constraints. These constraints include style-based terms as well as a model-based physics constraint linked to the target spacecraft. In particular, we introduce a material tracing map that combines material emissivity and geometric viewing configuration to construct a radiometric temperature proxy from the thermal image, and we use it as a supervisory signal during refinement. Experimental results show that this learning objective leads to a significant improvement in the quality of generated samples, both visually and in terms of distributional similarity to the ground truth, with additional benefits for downstream perception tasks. We further isolated the effects of the refinement stage and the physics-based constraint, and we provided a comparative study against existing methods, which demonstrates the superiority of the proposed approach within the considered setup.

Future work should investigate extending the dataset generation pipeline beyond Envisat to additional spacecraft models, in order to build broader multispectral benchmarks for space pose estimation. In particular, integrating targets such as the Tango satellite would provide a complementary reference scenario. The concept of material tracing maps could also be explored further, including using such maps directly as conditioning inputs to the generative model. Extending the same idea to background modeling may also benefit robustness in more complex scenes. Finally, it would be important to study the behavior of the proposed framework across different image resolutions, since Astos provides high-resolution rendering, and evaluating lower-resolution regimes would help quantify applicability to other sensing configurations and operational constraints.

CRediT authorship contribution statement

Said Guerazem: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Conceptualization. **Nabil Aouf:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Space Foundation Editorial Team, The Space Report 2025 Q2 Highlights Record $613 Billion Global Space Economy for 2024, Driven by Strong Commercial Sector Growth, Space Foundation, 2025-07-22, URL https://www.spacefoundation.org/2025/07/22/the-space-report-2025-q2/. (Accessed 8 December 2025).

[2] W. Fehse, Automated Rendezvous and Docking of Spacecraft, in: Cambridge Aerospace Series, Cambridge University Press, 2003.

[3] R. Szeliski, Computer Vision: Algorithms and Applications, first ed., Springer-Verlag, London, U.K., 2011.

[4] M.A. Fischler, R.C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.

[5] J. Song, D. Rondao, N. Aouf, Deep learning-based spacecraft relative navigation methods: A survey, Acta Astronaut. 191 (2022) 22–40.

[6] D. Rondao, N. Aouf, M.A. Richardson, ChiNet: Deep recurrent convolutional learning for multimodal spacecraft pose estimation, IEEE Trans. Aerosp. Electron. Syst. 59 (2) (2023) 937–949, http://dx.doi.org/10.1109/TAES.2022.3193085.

[7] Ö. Yılmaz, N. Aouf, E. Checa, L. Majewski, M. Sanchez-Gestido, Thermal analysis of space debris for infrared-based active debris removal, Proc. Inst. Mech. Eng. Part G: J. Aerosp. Eng. 233 (2017) 095441001774091, http://dx.doi.org/10.1177/0954410017740917.

[8] D. Rondao, N. Aouf, M.A. Richardson, O. Dubois-Matra, Benchmarking of local feature detectors and descriptors for multispectral relative navigation in space, Acta Astronaut. 172 (2020) 100–122.

[9] M. Kisantal, S. Sharma, T.H. Park, D. Izzo, M. Märtens, S. D'Amico, Satellite pose estimation challenge: Dataset, competition design, and results, IEEE Trans. Aerosp. Electron. Syst. 56 (5) (2020) 4083–4098.

[10] M. Bechini, M. Lavagna, P. Lunghi, Dataset generation and validation for spacecraft pose estimation via monocular images processing, Acta Astronaut. (ISSN: 0094-5765) 204 (2023) 358–369, http://dx.doi.org/10.1016/j.actaastro.2023.01.012, URL https://www.sciencedirect.com/science/article/pii/S0094576523000127.

[11] F. Cremaschi, S. Winter, V. Rossi, A. Wiegand, Launch vehicle design and GNC sizing with ASTOS, in: International Conference on Astrodynamics Tools and Techniques, ICATT, 2016, Conference paper (ESA Indico PDF). URL https://indico.esa.int/event/111/contributions/293/attachments/440/485/Cremaschi_ICATT_2016.pdf.

[12] S.M. Parkes, I. Martin, M. Dunstan, D. Matthews, Planet surface simulation with PANGU, in: SpaceOps 2004 (Eighth International Conference on Space Operations), 2004, http://dx.doi.org/10.2514/6.2004-592-389, URL https://arc.aiaa.org/doi/10.2514/6.2004-592-389.

[13] J. Lebreton, R. Brochard, M. Baudry, G. Jonniaux, A. Hadj Salah, K. Kanani, M. Le Goff, A. Masson, N. Ollagnier, P. Panicucci, A. Proag, C. Robin, Image simulation for space applications with the SurRender software, in: 11th International ESA Conference on Guidance, Navigation & Control Systems, 2021, Conference paper (Airbus PDF). URL https://www.airbus.com/sites/g/files/jlcbta136/files/2021-11/SurRender_Lebreton_ESA-GNC_2021.pdf.

[14] Blender Online Community, Blender - a 3D modelling and rendering package, Blender Foundation, Amsterdam, The Netherlands, 2024, URL https://www.blender.org.

[15] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[16] D. Ma, S. Li, J. Su, Y. Xian, T. Zhang, Visible-to-infrared image translation for matching tasks, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17 (2024) 18199–18213, http://dx.doi.org/10.1109/JSTARS.2024.3468456.

[17] B. Hu, C. Gao, S. Liu, J. Guo, F. Chen, F. Liu, CM-diff: A single generative network for bidirectional cross-modality translation diffusion model between infrared and visible images, 2025, arXiv preprint arXiv:2503.09514.

[18] G. Zhu, H. Pan, Q. Wang, C. Tian, C. Yang, Z. He, Data generation scheme for thermal modality with edge-guided adversarial conditional diffusion model, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 10544–10553.

[19] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[20] B. Ghojogh, A. Ghodsi, F. Karray, M. Crowley, Generative adversarial networks and adversarial autoencoders: Tutorial and survey, 2021, arXiv preprint arXiv:2111.13282.

[21] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[22] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[23] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[24] B. Li, D. Ma, F. He, Z. Zhang, D. Zhang, S. Li, Infrared image generation based on visual state space and contrastive learning, Remote. Sens. (ISSN: 2072-4292) 16 (20) (2024) http://dx.doi.org/10.3390/rs16203817, URL https://www.mdpi.com/2072-4292/16/20/3817.
[25] D.-Z. Du, Minimax and its applications, in: R. Horst, P.M. Pardalos (Eds.), Handbook of Global Optimization, Springer US, Boston, MA, ISBN: 978-1-4615-2025-2, 1995, pp. 339–367, http://dx.doi.org/10.1007/978-1-4615-2025-2_7, URL https://doi.org/10.1007/978-1-4615-2025-2_7.
[26] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, Y. Liu, Vmamba: Visual state space model, Adv. Neural Inf. Process. Syst. 37 (2024) 103031–103063.
[27] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. 56 (4) (2023) 1–39.
[28] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[29] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[30] J. Austin, D.D. Johnson, J. Ho, D. Tarlow, R. Van Den Berg, Structured denoising diffusion models in discrete state-spaces, Adv. Neural Inf. Process. Syst. 34 (2021) 17981–17993.
[31] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, M. Welling, Argmax flows and multinomial diffusion: Learning categorical distributions, Adv. Neural Inf. Process. Syst. 34 (2021) 12454–12465.
[32] K. Rasul, A.-S. Sheikh, I. Schuster, U. Bergmann, R. Vollgraf, Multivariate probabilistic time series forecasting via conditioned normalizing flows, 2020, arXiv preprint arXiv:2002.06103.
[33] F. Mao, J. Mei, S. Lu, F. Liu, L. Chen, F. Zhao, Y. Hu, PID: Physics-informed diffusion model for infrared image generation, Pattern Recognit. (ISSN: 0031-3203) 169 (2026) 111816, http://dx.doi.org/10.1016/j.patcog.2025.111816, URL https://www.sciencedirect.com/science/article/pii/S0031320325004765.
[34] X. Jia, C. Zhu, M. Li, W. Tang, W. Zhou, LLVIP: A visible-infrared paired dataset for low-light vision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3496–3504.
[35] T. FLIR, Free Teledyne FLIR thermal dataset for algorithm training, 2019, https://www.flir.com/oem/adas/adas-dataset-form. (Accessed 18 September 2025).
[36] L. Bianchi, M. Bechini, M. Quirino, M. Lavagna, Synthetic thermal image generation and processing for close proximity operations, Acta Astronaut. (ISSN: 0094-5765) 226 (2025) 611–625, http://dx.doi.org/10.1016/j.actaastro.2024.10.061, URL https://www.sciencedirect.com/science/article/pii/S0094576524006362.
[37] M. Quirino, Novel Thermal Images Generator for Autonomous Space Proximity Operations (Ph.D. thesis), Politecnico di Milano, 2023, URL https://www.politesi.polimi.it/handle/10589/213812.
[38] OpenFOAM Foundation / OpenCFD Ltd., OpenFOAM: The open source CFD toolbox, 2025, https://openfoam.org/. (Accessed 20 September 2025).
[39] C. Shi, J. Lu, Q. Sun, J. Zhou, W. Huang, R. Xia, Multi-scale auto-encoder for edge detection, IEEE Access 10 (2022) 116253–116260, http://dx.doi.org/10.1109/ACCESS.2022.3218149.
[40] Y. Liu, L. Dong, W. Ren, W. Xu, Multi-scale saliency measure and orthogonal space for visible and infrared image fusion, Infrared Phys. Technol. (ISSN: 1350-4495) 118 (2021) 103916, http://dx.doi.org/10.1016/j.infrared.2021.103916, URL https://www.sciencedirect.com/science/article/pii/S1350449521002887.
[41] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2021.
[42] Z. Wang, H. Zheng, P. He, W. Chen, M. Zhou, Diffusion-gan: Training gans with diffusion, 2022, arXiv preprint arXiv:2206.02262.
[43] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations, 2021.
[44] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps, in: Advances in Neural Information Processing Systems, 2022.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[46] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, 2016.
[47] D.P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1×1 convolutions, in: Advances in Neural Information Processing Systems, 2018.
[48] J.T. Barron, A general and adaptive robust loss function, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4331–4339.
[49] M. Mathieu, C. Couprie, Y. LeCun, Deep multi-scale video prediction beyond mean square error, 2015, arXiv preprint arXiv:1511.05440.
[50] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[51] C. Saharia, W. Chan, J. Saxena, L. Li, J. Whang, E. Denton, S.K.S. Ghasemipour, M. Norouzi, D. Fleet, T. Salimans, et al., Image super-resolution via iterative refinement, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[52] Z. Lyu, Z. Kong, X. Xu, L. Pan, D. Lin, A conditional point diffusion-refinement paradigm for 3d point cloud completion, 2021, arXiv preprint arXiv:2112.03530.
[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[54] F.A. Fardo, V.H. Conforto, F.C. De Oliveira, P.S. Rodrigues, A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms, 2016, arXiv preprint arXiv:1605.07116.
[55] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[56] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.
[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[58] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[59] J. Theiss, J. Leverett, D. Kim, A. Prakash, Unpaired image translation via vector symbolic architectures, in: European Conference on Computer Vision, Springer, 2022, pp. 17–32.
[60] J. Liang, H. Zeng, L. Zhang, High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9392–9400.