Toward Physically Grounded Noise Modeling and Restoration for Uncooled Thermal Imaging

Anonymous authors
Paper under double-blind review

Abstract

Uncooled microbolometer cameras provide compact and low-cost thermal imaging, but their measurements are dominated by complex noise fluctuations. Accurately decomposing these physical noise sources is inherently ill-posed, as many components are strongly correlated and cannot be uniquely identified from limited calibration observations. We propose an identifiability-aware framework for noise modeling and restoration in uncooled thermal imaging. Instead of estimating all physical noise terms, we identify a minimal set of calibration parameters from blackbody flat-field sequences, including per-pixel gain and offset, while temperature/state variation and unresolved correlated components are represented through calibrated perturbation and residual sampling. The remaining correlated components are modeled by residual sampling, enabling physics-consistent noise synthesis. This synthesis process allows restoration networks to be trained using paired data generated from calibrated sensor measurements. Experiments on two VOx microbolometer cameras demonstrate improved restoration quality across distinct camera operating regimes compared with traditional and learning-based restoration baselines. Additional validation using high-frame-rate sequences captured by a cooled LWIR camera and SAM-based segmentation further supports the realism of the synthesized noise and the effectiveness of the proposed pipeline.

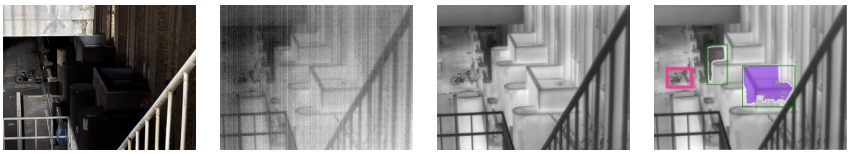
Keywords: thermal imaging, image restoration, noise synthesis, microbolometer sensors, fixed-pattern noise, physics-informed modeling.

1 Introduction

Long-wave infrared (LWIR) imaging in the 8–14 μm band provides passive, temperature-based perception and is widely used in inspection, robotics, environmental monitoring, and other low-light or safety-critical scenarios (Kruse, 2001; Rogalski et al., 2016). Cryogenically cooled detectors provide excellent sensitivity but remain expensive, bulky, and power hungry, limiting their use in embedded and field systems (Rogalski et al., 2016). In contrast, uncooled vanadium-oxide (VOx) or amorphous-silicon (a-Si) microbolometer arrays are compact and low cost, making them the dominant technology for practical thermal cameras (Kruse, 2001; Rogalski et al., 2016). However, this accessibility comes with pronounced sensor artifacts, including fixed-pattern noise (FPN), temporal drift, and scene-dependent bias, making single-frame thermal restoration a persistent challenge.

The degradation of uncooled microbolometers arises from a combination of pixel response nonuniformity, readout offsets, and temporally correlated fluctuations caused by electro-thermal coupling and electronics (Wolf et al., 2016; Oh et al., 2021; Xu et al., 2023), manifesting as spatial banding, blotchy backgrounds, and temperature-dependent artifacts that reduce effective contrast and degrade downstream tasks such as segmentation, tracking, and anomaly detection (Wolf et al., 2016; Oz et al., 2024); robust denoising and nonuniformity correction (NUC) are therefore essential for making uncooled cameras reliable sensing devices rather than rough visualizers. While some works aim at radiometric calibration and precise temperature estimation, many practical applications instead rely on visually consistent thermal images for downstream perception

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b) (c) (d)

Figure 1: Calibration-guided restoration for uncooled thermal imaging. Uncooled microbolometer cameras reveal temperature-dependent scene structure even in complete darkness, but their measurements often suffer from strong nonuniformity and drift artifacts. (a) A standard RGB camera fails to capture scene details under extremely low illumination. (b) The corresponding uncooled LWIR frame exposes thermal structure but exhibits spatial nonuniformity and low-frequency noise. (c) Our method combines calibration-based correction with learning-based restoration to suppress residual structured noise from a single frame. (d) The restored result yields cleaner thermal structure and more reliable inputs for downstream analysis tasks such as segmentation. The proposed pipeline requires only a one-time flat-field calibration and operates on single frames without multi-frame motion or scene-specific optimization.

tasks (Clero et al., 2023; Berg, 2016; Li et al., 2019; Ulmămei et al., 2025); in this work we focus on the latter setting and formulate the problem from the perspective of image restoration rather than absolute temperature recovery.

Existing approaches can be broadly divided into calibration-based and learning-based methods: classical NUC estimates gain and offset using blackbody references and applies linear correction during operation (Zuo et al., 2011; Oz et al., 2024), scene-based extensions update these parameters from image sequences to compensate for drift (Hayat et al., 1999; Ratliff et al., 2002; Orżanowski, 2016), and learning-based methods instead train neural networks to suppress structured noise using supervised or self-supervised data (Saragadam et al., 2021; Liu et al., 2023; Barral et al., 2024). While effective in some scenarios, calibration models often rely on simplified parametric assumptions, and learning approaches typically require large datasets, camera motion, or multiple frames per scene.

In this work we revisit microbolometer noise modeling from the perspective of identifiability. A physically detailed model that decomposes each pixel response into gain, offset, drift, and structured components is attractive in principle, but estimating all latent variables from limited calibration observations is fundamentally ill-posed. Different combinations of gain, offset, and correlated FPN terms can explain the same flat-field statistics, leading to non-unique solutions and unstable parameter estimation.

Instead of fitting an over-parameterized physical model, we adopt a minimal physics-informed formulation that retains only calibration-identifiable parameters: per-pixel gain and offset are estimated from multi-temperature flat-field measurements, while the remaining structured residuals are modeled statistically from temporal sequences collected during calibration. This yields a calibration-and-sampling pipeline—calibrated NUC followed by residual noise synthesized from temperature-dependent statistics—that enables training restoration networks without large real-world datasets.

Because uncooled sensors alone cannot provide high-quality supervision, we additionally collect high-frame-rate thermal sequences using a cooled LWIR camera operating at 1000 fps; the substantially reduced sensor noise of the cooled detector yields reliable references for both training and evaluation on real scenes.

Using this reference data together with the proposed synthesis framework, we train a lightweight U-Net (Ronneberger et al., 2015) restoration model that operates on single frames, keeping the architecture simple so that gains reflect the calibration-and-synthesis pipeline rather than network complexity. Figure 1 illustrates how the pipeline yields cleaner thermal structure and more reliable inputs for downstream vision tasks, and we further evaluate the impact on downstream segmentation using SAM2 (Ravi et al., 2025), showing that improved restoration leads to more stable segmentation results.

To summarize, our contributions are three-fold:

- We revisit microbolometer noise modeling from the perspective of identifiability and show that detailed physical decompositions are fundamentally ill-posed under limited calibration observations.
- We propose a practical calibration-and-sampling framework that estimates calibration-constrained parameters while modeling residual noise statistically to synthesize realistic training data.
- We construct a multi-sensor thermal denoising benchmark and additionally collect high-frame-rate thermal reference sequences using a cooled LWIR camera, enabling comprehensive evaluation on real thermal imagery. Experiments on two VOx sensors demonstrate that the proposed approach consistently outperforms traditional denoising and recent learning-based restoration baselines, while also improving downstream segmentation performance.

2 Related Work

2.1 Noise Physics, Calibration, and Thermal Imaging Data

Uncooled VOx and a-Si microbolometers exhibit spatial nonuniformity and temporal drift from fabrication variability and electro-thermal coupling: pixel response non-uniformity (PRNU) and dark signal non-uniformity (DSNU) arise from membrane geometry and readout-circuitry variations, producing fixed-pattern artifacts (Wolf et al., 2016), while electro-thermal processes further introduce Johnson noise, temperature-fluctuation noise, and strong $1/f$ components whose magnitude depends on integration time and sensor temperature (Oh et al., 2021; Xu et al., 2023; Bianchi et al., 2025), and temperature variation can significantly affect pixel responses beyond static PRNU/DSNU models (Oz et al., 2024). To mitigate these effects, factory pipelines typically perform reference-based nonuniformity correction (NUC) using blackbody sources at multiple temperatures: two-point calibration (Zuo et al., 2011) estimates pixelwise gain and offset, scene-based methods update offsets from video statistics (Hayat et al., 1999; Ratliff et al., 2002; Orżanowski, 2016), and recent work exploits motion-compensated multi-view stacks to suppress fixed-pattern noise (Barral et al., 2024). However, these methods usually assume stable operating conditions or sufficient scene diversity, and often rely on simplified parametric models that generalize poorly across temperature drift and varying environments.

Existing thermal datasets mainly target perception tasks, such as FLIR ADAS (FLIR Systems, 2018), KAIST multispectral (Hwang et al., 2015), and OTCBVS (Hammoud, 2004), but do not provide controlled multi-temperature calibration sequences or paired raw/NUC measurements needed to study sensor-level degradation. Our work addresses this gap by constructing camera-specific calibration datasets with multi-temperature flat fields and paired sequences for evaluating physics-informed thermal restoration.

2.2 Deep Priors, Learning-Based Restoration, and Noise Modeling

Deep learning methods have recently been applied to thermal image restoration. Physics-inspired frameworks such as DeepIR (Saragadam et al., 2021) separate scene radiance from slowly varying nonuniformities using deep priors, TV-DIP (Liu et al., 2023) combines deep image priors with total variation regularization to remove banding artifacts, and other approaches train supervised models on thermal sequences or synthetic noise to suppress fixed-pattern noise and low-light artifacts (Barral et al., 2024). A common limitation of these methods is their reliance on temporal diversity or multi-frame input to disentangle scene structure from sensor nonuniformity, which is often impractical for embedded thermal systems or static scenes; in contrast, we target single-frame restoration and generate supervision directly from calibration data through physics-informed noise synthesis.

Noise modeling has been widely studied in CMOS imaging, where Poisson-Gaussian statistics and calibration-based synthesis are commonly used for realistic RAW denoising (Brooks et al., 2019; Feng et al., 2022; Wei et al., 2020); these works emphasize that accurate calibration can anchor the data distribution, while synthetic noise should reproduce realistic joint statistics rather than simulate every physical mechanism. Similar observations have

been reported for specialized sensors such as EMCCDs (Jiang et al., 2025b) and SWIR cameras (Jiang et al., 2025a), where distribution-oriented noise modeling often proves more practical than full physical simulation (Li et al., 2025). Inspired by these findings, our approach combines calibration-constrained parameters with residual sampling to synthesize realistic thermal noise without requiring a complete electro-thermal sensor simulator.

3 Identifiability-Aware Noise Modeling and Data Synthesis

3.1 Identifiability of Microbolometer Noise Decomposition

Uncooled microbolometer measurements are often modeled using a decomposition that separates pixel responsivity, readout bias, and temporally varying noise components; while such formulations are physically interpretable, estimating every latent component from calibration observations alone is generally ill-posed. Consider a simplified observation model for flat-field measurements under uniform radiance $X(T_t)$:

$$Z_t = GX(T_t) + O + r_t^{\text{row}}\mathbf{1}^\top + \mathbf{1}(c_t^{\text{col}})^\top + d_t\mathbf{1}\mathbf{1}^\top + W_t, \tag{1}$$

where G and O denote pixel-wise gain and offset (PRNU and DSNU), r_t^{row} and c_t^{col} represent row and column readout patterns, d_t models frame-level drift (e.g., $1/f$ components), and W_t denotes stochastic noise.

Vectorizing each frame yields

$$\text{vec}(Z_t) = \mathbf{A}\,\theta_t + \mathbf{w}_t, \tag{2}$$

where θ_t collects the latent parameters (gain, offset, row/column terms, and drift) and $\mathbf{A}$ is the corresponding design matrix determined by the imaging model. Because several components enter the observation through overlapping additive subspaces (e.g., column offsets and global drift), the matrix $\mathbf{A}$ is rank-deficient under flat-field measurements, implying that multiple parameter configurations produce identical observations. Consequently, recovering every physical component independently from flat-field calibration becomes a non-identifiable inverse problem. We demonstrate this ambiguity through controlled toy experiments (see Supplementary Material), where different decompositions achieve comparable calibration errors but generate qualitatively different noise structures. Motivated by this observation, we adopt a minimal formulation that retains only parameters identifiable from calibration measurements, i.e., pixel-wise gain and offset, while modeling the remaining temporal variations statistically; the following sections describe how these calibration parameters are estimated from blackbody measurements and how the remaining noise variability is incorporated into a physics-informed synthesis pipeline for training restoration networks.

3.2 Calibration with a Blackbody Source

Flat-field acquisition. We record flat-field sequences using a temperature-controlled blackbody radiator that provides spatially uniform infrared emission. For each temperature T, the temporal mean frame is denoted $\bar{Z}(T)$; an example of the calibration setup and resulting parameter maps is shown in Figure 2. Per-pixel calibration. Let $\bar{Z}_{ij}(T)$ denote the temporal mean of pixel (i, j) at temperature T. Although thermal radiance follows Planck's law, within the narrow calibration range considered here the radiance-temperature relation can be well approximated as locally linear. We therefore fit a pixel-wise affine model

$$\bar{Z}_{ij}(T_m) = G_{ij}\,T_m + O_{ij}, \quad m = 1, \ldots, M, \tag{3}$$

where G_{ij} and O_{ij} represent the effective gain and offset of pixel (i, j). The parameters are estimated by least-squares regression over the calibration temperatures. Pixels exhibiting saturation or unstable temporal statistics are marked invalid and inpainted from neighboring values. Calibration maps. The resulting gain map G and offset map O characterize spatial responsivity and bias variations across the focal-plane array (FPA), corresponding to PRNU and DSNU components, defining the deterministic part of the sensor response used for subsequent non-uniformity correction.

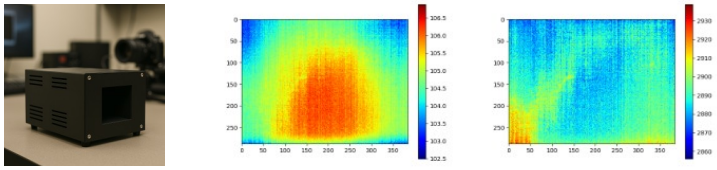


Figure 2: Left: photograph of the blackbody calibration source used in our experiments. Right: example gain (G) and offset (O) maps obtained from multi-temperature blackbody calibration. Spatial patterns reflect responsivity variations and readout-induced bias structures.

3.3 Temporal Noise Characterization (Sampling-Based)

Residual extraction. Given a frame $Z(T,t)$ captured at time t, we remove the calibrated deterministic component

$$N_{\text{temp},ij}(t) = Z_{ij}(T,t) - \big(G_{ij}X(T) + O_{ij}\big). \tag{4}$$

The residual stack collected across temperatures forms a library of real temporal noise samples. These residuals contain low-frequency drift, $1/f$ components, and readout noise typical of uncooled microbolometers (Yadav et al., 2022; Hanson, 2018). Sampling-based modeling. Instead of explicitly decomposing every physical noise component, we model temporal variability by directly sampling from this residual library. This strategy follows recent trends in imaging sensor noise modeling, where combining simple calibration with stochastic residual sampling can reproduce realistic noise statistics without constructing detailed physical simulators (Zhang et al., 2021; Li et al., 2025).

3.4 Extending Calibration to Varying Operating Temperatures

Calibration performed with indoor blackbody sources may not perfectly transfer to real operating conditions, where both ambient temperature and sensor temperature can vary. Controlled chamber experiments in Oz et al. (2024) show that while standard calibration remains approximately valid, residual pixel errors increase as the sensor temperature deviates from the calibration point.

To account for this effect, we incorporate temperature-dependent calibration variability into the synthesis pipeline. Let T_0 denote the nominal calibration temperature. We model the effective offset under temperature variation as

$$\tilde{O}_{ij} = O_{ij} + \epsilon_{ij}(T), \tag{5}$$

where $\epsilon_{ij}(T)$ is a stochastic perturbation whose variance increases with temperature deviation

$$\text{Var}[\epsilon_{ij}(T)] \propto |T - T_0|^{\alpha}. \tag{6}$$

In practice, this effect is simulated by scaling the variance of sampled residual noise according to the simulated temperature deviation. This simple augmentation allows the synthesized data to cover a broader range of operating conditions than those observed during controlled blackbody calibration. Additional implementation details and calibration statistics used for this augmentation are provided in the supplementary material.

3.5 Synthetic Data Generation

Given calibrated gain and offset maps (G, O) together with the residual noise library, we synthesize realistic noisy–clean pairs for network training. Domain alignment. Clean reference images are first normalized to match the statistics of multi-capture references:

$$I_{\text{clean}} \leftarrow \sigma_{\text{ref}} \frac{I_{\text{clean}} - \mu_c}{\sigma_c} + \mu_{\text{ref}}.$$

Inverse–forward NUC pipeline. Because the reference images correspond to NUC-corrected outputs, we first invert the calibration transform to recover the raw sensor domain. Residual

noise is then injected before reapplying the calibration to obtain the final noisy observation. This procedure is conceptually related to the "unprocessing" strategy used for CMOS sensor modeling (Brooks et al., 2019).

$$I_{\text{inv}} = \frac{I_{\text{clean}} - \overline{O}}{\overline{G}} \cdot G + O, \tag{7}$$

$$\tilde{I} = I_{\text{inv}} + N_{\text{sample}}, \tag{8}$$

$$I_{\text{noisy}} = \frac{\tilde{I} - \tilde{O}}{\tilde{G}} \cdot \overline{G} + \overline{O}. \tag{9}$$

We denote by $\bar{G}$ and $\bar{O}$ the spatial mean of the gain and offset maps, respectively. Calibration variability. To emulate sensor drift and temperature-dependent response changes, the calibration maps are randomly perturbed during synthesis:

$$\begin{aligned} \tilde{G} &= G\,(1 + \epsilon_G), \quad \epsilon_G \sim \mathcal{N}(0, \sigma_G^2), \\ \tilde{O} &= O + \epsilon_O, \qquad \epsilon_O \sim \mathcal{N}(0, \sigma_O^2). \end{aligned} \tag{10}$$

These perturbations capture both slow responsivity drift and frame-to-frame offset variations introduced by periodic shutter-based NUC mechanisms in practical cameras. Algorithm summary. The complete calibration and synthesis pipeline is summarized in Algorithm 1 (Appendix A): the procedure first estimates calibration parameters from blackbody flat-field stacks and extracts temporal residual noise, then generates synthetic noisy-clean pairs by sampling residual noise and applying an inverse-forward NUC transformation with perturbed calibration parameters.

4 Experiments and Results

4.1 Dataset Construction

We evaluate the proposed calibration and synthesis framework on two uncooled LWIR microbolometer cameras with distinct noise characteristics (Figure 3). Camera 1 (Vision Sensing VIL384GS, 384×288) provides a raw rev0 mode with no internal correction and a rev4 mode that applies factory non-uniformity correction (NUC), providing a practical camera-internal corrected reference for supervision. Camera 2 (HM-TM06-LF/ANT, 640×512) performs periodic shutter-based NUC: the camera briefly closes an internal shutter to capture a uniform reference frame and update its offset calibration. While this mechanism suppresses long-term drift, it introduces small frame-to-frame bias variations between shutter updates, leaving residual temporal artifacts that still require denoising.

For Camera 1, we collected 15 static indoor and outdoor scenes under stable illumination, each recorded as a 1000-frame sequence in both rev0 (noisy input) and rev4 (clean reference) modes. We additionally captured 52 multi-capture scenes (not necessarily static and therefore not used as paired test data) to construct high-SNR references supporting domain alignment and fine-tuning. Camera 2 contributes 11 training sequences and 5 validation sequences, all captured with its built-in shutter NUC enabled. Before data collection, both cameras underwent the blackbody calibration procedure described in Section 3.2. Although the number of scenes is moderate, temporal drift and spatially correlated noise in each 1000-frame sequence provide diverse noise realizations for learning-based denoising; all data are stored in 16-bit linear RAW format and processed without gamma correction.

In addition to these two uncooled sensors used for training and evaluation, we collected auxiliary high-frame-rate thermal sequences using a cooled scientific LWIR camera (FLIR X6981HS SLS-PH). This system records 1000 fps thermal video with substantially lower readout noise than uncooled sensors, providing high-quality single-frame thermal measurements. We captured 33 scenes containing both static and dynamic content, each with at least 26,000 frames. These sequences are used as high-quality references for the fine-tuning stage and for additional validation experiments on synthesized data.

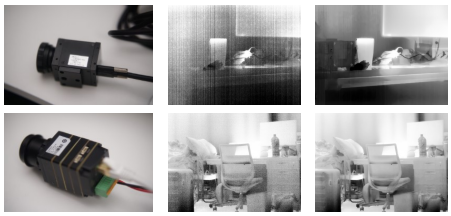


Figure 3: Hardware and dataset overview. Each row shows a camera photograph, an example input frame, and its corresponding corrected or high-SNR reference. Camera 1's rev0 mode exhibits extremely low SNR with strong fixed-pattern and drift noise, while rev4 provides cleaner supervision. Camera 2's shutter-based NUC reduces but does not eliminate low-frequency temporal artifacts.

For training, frames are normalized to $[0, 1]$ and augmented with random horizontal flips and 90° rotations. To further assess generalization, we generate additional synthetic test scenes using our calibration-and-synthesis pipeline; the corresponding quantitative and qualitative results are reported in the supplementary material and exhibit trends consistent with the real-data evaluation.

4.2 Non-Uniformity Correction and GT Curation

NUC for reference frames. To obtain stable supervision, each raw frame I is corrected using the calibrated gain and offset:

$$I_{\text{adj}} = \frac{I - O}{G} \cdot \overline{G} + \overline{O}, \tag{11}$$

which removes pixel-wise responsivity and bias variations while preserving scene radiance and contrast. Perceptual cleaning of static stacks. For static sequences with over 1000 frames, we compute a temporal mean to obtain a preliminary high-SNR reference and refine it with BM3D, with the noise level selected by grid search minimizing BRISQUE. The resulting image serves as a clean ground-truth proxy; any remaining artifacts (e.g., faint horizontal stripes) are excluded via a row-wise confidence mask during training.

4.3 Network Configuration and Training Protocol

We use a lightweight U-Net (Ronneberger et al., 2015) with four encoder–decoder stages and bilinear upsampling, trained with Adam (2×10^{-4}) and a loss combining pixelwise ℓ_1 and VGG-based perceptual terms, keeping the architecture simple so that performance gains reflect the calibration-and-synthesis pipeline rather than model size.

Training follows a two-stage procedure: we first pretrain on grayscale SID images (Chen et al., 2018) with thermal-style degradations injected by our synthesis pipeline, providing diverse low-level priors and robustness against overfitting to the limited calibration data, and then fine-tune using synthetic pairs generated from each camera's own gain, offset, and residual noise library obtained through blackbody calibration. Camera 1 and Camera 2 are trained independently and evaluated only on the test split of the device each was calibrated for, matching typical one-time-calibration deployment. Additional training details are provided in the supplementary material.

4.4 Quantitative Evaluations

We compare our method against both traditional and learning-based baselines: BM3D (Dabov et al., 2007), ELD (Wei et al., 2020), DeepIR (Saragadam et al., 2021), TV-DIP (Liu et al., 2023), and Multiview FPN (Barral et al., 2024). Evaluation employs six complementary metrics covering both distortion and perceptual quality: PSNR and SSIM for pixel fidelity, LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) for deep-feature similarity, and NIQE (Mittal et al., 2013) and MUSIQ (Ke et al., 2021) for no-reference perceptual quality. We report results for both cameras, noting that Camera 1's rev0 mode represents an extremely low-SNR regime that stresses robustness to strong fixed-pattern

and drift noise. Table 1 and Table 2 summarize performance across these six metrics, with best and second-best results highlighted in bold and underlined, respectively.

Table 1: Quantitative comparison on Camera 1 dataset (15 scenes × 100 frames).

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑
BM3D (Dabov et al., 2007)	20.64	0.745	0.532	0.422	11.99	27.32
ELD (Wei et al., 2020)	20.45	0.870	0.325	0.282	6.85	20.41
DeepIR (Saragadam et al., 2021)	24.55	<u>0.932</u>	<u>0.165</u>	<u>0.184</u>	7.67	25.25
TV-DIP (Liu et al., 2023)	21.48	0.800	0.434	0.359	8.99	27.96
Multiview FPN (Barral et al., 2024)	20.79	0.772	0.463	0.388	10.44	<u>28.25</u>
Ours	<u>25.13</u>	0.918	0.204	0.236	5.91	27.71
Ours (finetuned)	27.78	0.951	0.129	0.174	<u>6.42</u>	29.38

Table 2: Quantitative comparison on Camera 2 dataset (11 scenes × 100 frames).

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑
BM3D (Dabov et al., 2007)	30.39	0.742	0.386	0.216	8.10	32.61
ELD (Wei et al., 2020)	30.73	0.783	0.347	0.196	7.57	31.94
DeepIR (Saragadam et al., 2021)	24.99	<u>0.944</u>	0.218	0.200	8.31	29.24
TV-DIP (Liu et al., 2023)	21.42	0.757	0.402	0.251	8.09	<u>33.96</u>
Multiview FPN (Barral et al., 2024)	30.50	0.748	0.381	0.220	8.27	32.43
Ours	<u>30.98</u>	0.896	<u>0.186</u>	<u>0.142</u>	4.79	33.76
Ours (finetuned)	34.43	0.959	0.119	0.125	<u>6.95</u>	38.77

Our physics-informed synthesis consistently outperforms existing baselines across both sensors. On Camera 1, the fine-tuned model improves PSNR by 2.65 dB over the non-fine-tuned variant, together with improved perceptual scores such as MUSIQ. Compared with Multiview FPN—which relies on multi-frame alignment and view redundancy—our single-frame approach attains comparable or superior results without geometric constraints, yielding sharper textures and more accurate thermal gradients. These consistent improvements highlight that accurate physics-guided synthesis, even when paired with a lightweight U-Net backbone, can rival or surpass more complex architectures as long as the data generation process faithfully reflects real sensor statistics. Additional quantitative and qualitative comparisons are provided in the supplementary material and exhibit the same performance trends.

4.5 Qualitative Comparisons

Figure 4 shows representative scenes of varying temperature contrast and background complexity. BM3D suppresses white noise but smooths mid-frequency texture; ELD introduces mild checkerboard artifacts from its CMOS-specific prior; DeepIR and TV-DIP recover local contrast but leave residual column structure; Multiview FPN mitigates large-scale banding yet over-smooths high-gradient regions. Our method produces spatially uniform and temporally stable reconstructions that preserve material boundaries and fine-scale temperature variations, yielding perceptually more coherent and physically plausible outputs.

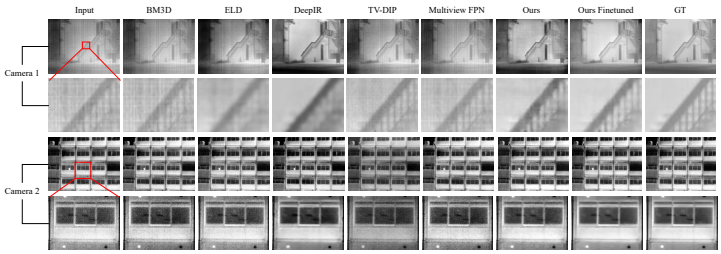


Figure 4: Qualitative comparison on representative thermal scenes. Columns correspond to methods (BM3D, ELD, DeepIR, TV-DIP, Multiview FPN, Ours, Ours finetuned, Ground Truth). Our single-frame model removes both fixed-pattern and flicker noise while retaining microstructure.

4.6 Ablation Study

We conduct ablations on Camera 1 to quantify each component's contribution: w/o Calibration—replace measured (G, O) with uniform values; w/o Residual Sampling—use Gaussian

white noise; w/o Perturbation—omit Gaussian perturbations on (G, O); w/o Fine-tuning—train only on synthetic data.

Table 3: Ablation results on Camera 1 (15 scenes × 100 frames).

Configuration	PSNR↑	SSIM↑
Full model (finetuned)	27.78	0.951
w/o Fine-tuning	25.13	0.918
w/o Perturbation	25.01	0.912
w/o Residual Sampling	24.37	0.903
w/o Calibration	22.59	0.871

Removing calibration or realistic residual sampling causes a clear PSNR/SSIM drop, confirming that sensor-aware modeling is crucial. Eliminating perturbation slightly weakens robustness, while omitting fine-tuning limits adaptation to real thermal distributions; each proposed component thus contributes meaningfully to the final performance.

4.7 Downstream Evaluation: Segmentation Robustness

To verify downstream utility, we apply the promptable SAM2 (Ravi et al., 2025) model to noisy, denoised, and clean inputs over 20 scenes from both cameras. As shown in Figure 5, segmentation on noisy frames yields fragmented masks and uncertain boundaries, whereas denoised images lead to coherent regions and sharper contours. Although ground-truth masks are unavailable, we observed clear qualitative improvement in boundary continuity and region stability after denoising, suggesting that physically grounded noise suppression provides more reliable structural cues for high-level inference.

5 Discussion

Combining calibration-derived sensor parameters with data-driven learning provides a practical route for thermal image restoration: estimating identifiable gain and offset through blackbody calibration and modeling the remaining variability via residual sampling avoids ill-posed decomposition of correlated noise sources while preserving real sensor statistics. Exposing the network to statistically faithful degradations, rather than reproducing every physical noise mechanism, proves sufficient for robust learning: high-frame-rate cooled LWIR data provide a distinct thermal-content domain for validating restoration, while downstream segmentation tests suggest improved restoration benefits higher-level vision tasks.

Future work will explore multi-view thermal capture combined with NeRF-like reconstruction for higher-fidelity references in dynamic scenes, and explicit modeling of the thermal time constant of uncooled microbolometers to address temporal lag and ghosting artifacts.

6 Conclusion

We presented an identifiability-aware framework integrating blackbody calibration with residual noise sampling, isolating identifiable parameters and modeling remaining variability through data-driven synthesis to produce realistic noisy–clean training pairs. Experiments on two uncooled LWIR cameras show consistent gains over classical and learning-based baselines, with cooled-data validation and downstream segmentation supporting the effectiveness and generalization of the calibrated synthesis pipeline as an alternative to physical simulation.

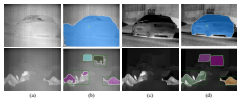


Figure 5: Impact of denoising on downstream segmentation using SAM2 (Ravi et al., 2025). From left to right: (a) noisy input, (b) segmentation on noisy, (c) our denoised image, (d) segmentation on denoised. Denoising improves boundary continuity and mask coherence, turning fragmented regions in (b) into more coherent object areas in (d).

References

Arnaud Barral, Pablo Arias, and Axel Davy. Fixed pattern noise removal for multi-view single-sensor infrared camera. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1669–1678, 2024. doi: 10.1109/WACV56127.2024.00110.

Amanda Berg. Detection and tracking in thermal infrared imagery., 2016.

Lucia Bianchi, Michele Bechini, Matteo Quirino, and Michèle Lavagna. Synthetic thermal image generation and processing for close proximity operations. Acta Astronautica, 226: 611–625, 2025. ISSN 0094-5765. doi: https://doi.org/10.1016/j.actaastro.2024.10.061. URL https://www.sciencedirect.com/science/article/pii/S0094576524006362.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In CVPR, pp. 11036–11045, 2019.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In CVPR, pp. 3291–3300, 2018.

Kaoutar Clero, Mohamed Nadour, Said Ed-Diny, Achalhi Mohammed, Mouhamed Cherkaoui, Hamd Abdelali, Sanaa El Fkihi, Intissar Benzakour, Rziki Said, Hicham Tagemouati, and François Bourzeix. A review of the use of thermal imaging and computer vision for pattern recognition. In Computer Science & Information Technology (CS & IT), pp. 121–136, 11 2023. doi: 10.5121/csit.2023.132108.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing, 16(8):2080–2095, 2007. doi: 10.1109/TIP.2007.901238.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Hansen Feng, Lizhi Wang, Yuzhi Wang, and Hua Huang. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In ACM Multimedia, pp. 1436–1444, 2022. doi: 10.1145/3503161.3548186.

FLIR Systems. Free FLIR thermal dataset for algorithm training. https://oem.flir.com/solutions/automotive/adas-dataset-form/, 2018. Accessed: 2024-11-01.

Riad I. Hammoud. OTCBVS benchmark dataset collection. http://vcipl-okstate.org/pbvs/bench/, 2004. Accessed: 2025-09-26.

Charles M. Hanson. Implications of 1/f noise in uncooled thermal imaging. In Infrared Technology and Applications XLIV, volume 10624 of Proc. SPIE, pp. 106241C, 2018. doi: 10.1117/12.2306334.

Majeed M. Hayat, Sergio N. Torres, Ernest E. Armstrong, Stephen C. Cain, and Brian Yasuda. Statistical algorithm for nonuniformity correction in focal-plane arrays. Applied Optics, 38(5):772–780, 1999. doi: 10.1364/AO.38.000772.

Soonmin Hwang, Jaesik Park, Namil Kim, Youngchan Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In CVPR, pp. 1037–1045, 2015.

Haiyang Jiang, Hongjun Wang, and Yinqiang Zheng. See the invisible with swir: Learning to enhance via synthetic noise modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 7449–7458, October 2025a.

Haiyang Jiang, Tetsuichi Wazawa, Imari Sato, Takeharu Nagai, and Yinqiang Zheng. Revolutionizing EMCCD denoising through a novel physics-based learning framework for noise modeling. In The Thirteenth International Conference on Learning Representations, 2025b. URL https://openreview.net/forum?id=vmulbBDCan.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5128–5137, 2021. doi: 10.1109/ICCV48922.2021.00510.

Paul W. Kruse. Uncooled Thermal Imaging: Arrays, Systems, and Applications. SPIE Press, Bellingham, WA, 2001. ISBN 9780819441296.

Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night:edge-conditioned cnn for thermal image semantic segmentation, 2019. URL https://arxiv.org/abs/1907.10303.

Feiran Li, Haiyang Jiang, and Daisuke Iso. Noise modeling in one hour: Minimizing preparation efforts for self-supervised low-light raw image denoising. In CVPR, pp. 5699–5708, 2025.

Kang Liu, Honglei Chen, Wenzhong Bao, and Jianlu Wang. Thermal imaging spatial noise removal via deep image prior and step-variable total variation regularization. Infrared Physics & Technology, 134:104888, 2023. doi: 10.1016/j.infrared.2023.104888.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. IEEE Signal Processing Letters, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Jaesub Oh, Hyeong sub Song, Jongcheol Park, and Jong-Kwon Lee. Noise improvement of a-si microbolometers by the post-metal annealing process. Sensors, 21(20):6722, 2021. doi: 10.3390/s21206722.

Tomasz Orżanowski. Nonuniformity correction algorithm with efficient pixel offset estimation for infrared focal plane arrays. SpringerPlus, 5(1831):1–13, 2016. doi: 10.1186/s40064-016-3534-1.

Navot Oz, Nir Sochen, David Mendlovic, and Iftach Klapp. Estimating temperatures with low-cost infrared cameras using physically-constrained deep neural networks. Opt. Express, 32(17):30565–30582, Aug 2024. doi: 10.1364/OE.531349. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-32-17-30565.

Bradley M. Ratliff, Majeed M. Hayat, and J. Scott Tyo. An algebraic algorithm for nonuniformity correction in focal-plane arrays. Journal of the Optical Society of America A, 19 (9):1737–1747, 2002. doi: 10.1364/JOSAA.19.001737.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Antoni Rogalski, Piotr Martyniuk, and Mariusz Kopytko. Challenges of small pixel infrared detectors: A review. Reports on Progress in Physics, 79(4):046501, 2016. doi: 10.1088/0034-4885/79/4/046501.

O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351 of LNCS, pp. 234–241. Springer, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a. (available on arXiv:1505.04597 [cs.CV]).

Vishwanath Saragadam, Akshat Dave, Ashok Veeraraghavan, and Richard G. Baraniuk. Thermal image processing via physics-inspired deep networks. In IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 4057–4065, 2021. arXiv:2108.07973.

Andrei-Alexandru Ulmămei, Taddeo D'Adamo, Costin-Emanuel Vasile, and Radu Hobincu. Human detection in uav thermal imagery: Dataset extension and comparative evaluation on embedded platforms. Journal of Imaging, 11(12), 2025. ISSN 2313-433X. doi: 10.3390/jimaging11120436. URL https://www.mdpi.com/2313-433X/11/12/436.

Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In CVPR, pp. 2758–2767, 2020.

Alejandro Wolf, Jorge E. Pezoa, and Miguel Figueroa. Modeling and compensating temperature-dependent non-uniformity noise in IR microbolometer cameras. Sensors, 16(7):1121, 2016. doi: 10.3390/s16071121.

Yanan Xu, Mukti Rana, and Kevin Díaz-Aponte. Electro-thermal numerical analysis of microbolometer over various kinds of design structure under adjustable thermal conductance in the microelectromechanical system. Heliyon, 9(4):e14886, 2023. doi: 10.1016/j.heliyon.2023.e14886.

P. V. Karthik Yadav, Isha Yadav, B. Ajitha, Abraham Rajasekar, Sudha Gupta, and Y. Ashok Kumar Reddy. Advancements of uncooled infrared microbolometer materials: A review. Sensors and Actuators A: Physical, 342:113611, 2022. doi: 10.1016/j.sna.2022.113611.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pp. 586–595, 2018.

Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4593–4601, October 2021.

Chun Zuo, Qian Chen, Tianyang Qian, and Shaohua Feng. Two-point non-uniformity correction algorithm for infrared focal plane arrays based on gray compensation. In Infrared Technology and Applications XXXVII, volume 8012 of Proc. SPIE, pp. 80123U, 2011. doi: 10.1117/12.882718.

AI Use Statement

The authors used ChatGPT, Claude, and Codex for research discussion, literature verification, coding assistance, and manuscript editing. These tools did not independently validate the scientific claims. The authors verified the experimental results, citations, code changes, and final manuscript, and take full responsibility for the submission.

This supplemental document provides additional analyses, implementation details, and experimental results that complement the main paper. Section B presents a toy experiment illustrating the identifiability limitations of decomposing structured thermal noise from flat-field statistics. Section C describes implementation details of the calibration procedure, residual noise library construction, and temperature-aware synthesis pipeline. Section D provides training details of the proposed model. Section E reports additional quantitative and qualitative evaluations on synthetic test scenes. Section A gives the full pseudocode of the calibration-guided noise synthesis pipeline summarized in the main paper.

A Calibration-guided noise synthesis: full algorithm

Algorithm 1 gives the complete pseudocode of the calibration-and-synthesis pipeline summarized in Section 3.5 of the main paper: it first estimates per-pixel gain and offset from blackbody flat-field stacks and extracts the temporal residual noise library, and then generates synthetic noisy–clean pairs by sampling residual noise and applying the inverse–forward NUC transformation with perturbed calibration parameters.

Algorithm 1: Calibration-Guided Noise Synthesis

Input: flat-field stacks $Z(T,t)$, clean image I_{clean}

// Calibration
for each temperature T do
compute $\bar{Z}(T) = \text{mean}_t\, Z(T,t)$
estimate (G, O) by pixel-wise linear regression
for each frame (T,t) do
compute $N_{\text{temp}}(t) = Z(T,t) - (G\,X(T) + O)$

// Synthesis
normalize I_{clean}
compute $I_{\text{inv}} = \dfrac{I_{\text{clean}} - \bar{O}}{\bar{G}} \cdot G + O$
sample $N_{\text{sample}} \sim N_{\text{temp}}$
perturb $\tilde{G} = G(1+\epsilon_G), \quad \tilde{O} = O + \epsilon_O$
compute $I_{\text{noisy}} = \dfrac{I_{\text{inv}} + N_{\text{sample}} - \tilde{O}}{\tilde{G}} \cdot \bar{G} + \bar{O}$

Output: $(I_{\text{noisy}}, I_{\text{clean}})$

B Toy experiment: Limitations of decomposing structured thermal noise using flat-field statistics

The main paper adopts a deliberately minimal calibration strategy: we estimate only the well-constrained gain and offset maps from flat-field means, and treat the remaining spatio-temporal variation as an empirically sampled residual field. This choice avoids relying on strong assumptions about how structured thermal noise should be factorized (e.g., into row-, column-, or global components). To further motivate this design, we include here a controlled toy experiment demonstrating that even with abundant flat-field data, attempting to infer a full variance decomposition from pixel-level statistics can become highly unstable once the underlying noise sources exhibit correlated structure.

B.1 Motivation

Classical treatments of fixed-pattern and readout noise often begin with covariance-based reasoning: if certain components are independent (e.g., row-only or column-only processes), then specific covariance patterns can separate them. Such independence-based estimators act as a natural baseline for assessing identifiability from flat-field data. However, real uncooled microbolometer sensors do not guarantee such independence, i.e., row- and column-wise variations can be jointly influenced by shared latent factors (e.g., thermal drift, mechanical coupling, per-frame bias variation). This raises the question: if one tries to decompose thermal noise under the independence assumptions often used in classic statistical derivations, how sensitive is the result to correlated structure?

The following toy experiment is designed to answer this question in isolation, in a setting where ground truth is known. It is not intended to mimic the full complexity of real sensors, but rather to illustrate why recovering a physically meaningful decomposition from flat-field statistics alone is inherently ill-posed.

B.2 Toy model with tunable correlation

We simulate a simplified thermal-style noise signal

$$x_{tij} = c_{t,j} + r_{t,i} + u_t + f_{ij} + j_{tij} + T, \tag{12}$$

on a $1000 \times 48 \times 96$ grid, with five components mirroring those found in real sensors: column-wise noise $c_{t,j}$, row-wise noise $r_{t,i}$, a global temporal drift u_t, a static FPN map f_{ij}, Johnson noise j_{tij}, and a constant bias. The key parameter is a correlation coefficient $\rho \in [0,1]$ that determines how strongly the row and column processes are tied together through a shared per-frame latent variable z_t. When $\rho = 0$, the model satisfies the independence assumptions commonly invoked in variance-based derivations; when $\rho > 0$, these assumptions are systematically violated.

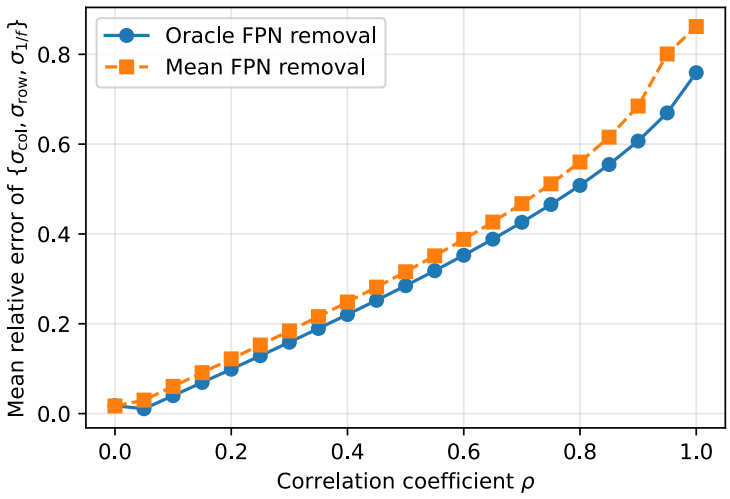


Figure 6: Mean relative error vs. row/column correlation ρ in the toy model. At $\rho = 0$, the independence assumption holds and the estimator recovers $(\sigma_{\text{col}}, \sigma_{\text{row}}, \sigma_{1/f})$ with small error ($\approx$1–2%). As ρ increases, the inferred decomposition becomes progressively distorted: correlated row/column variability is absorbed into the global term, causing $\widehat{\sigma}_{1/f}$ to inflate and $\widehat{\sigma}_{\text{col}}, \widehat{\sigma}_{\text{row}}$ to shrink. Even with oracle FPN removal, the error exceeds 75% at $\rho = 1$; the mean-based version grows even faster due to mild aliasing.

All components are generated with known standard deviations

$$(\sigma_{\text{col}}, \sigma_{\text{row}}, \sigma_{1/f}, \sigma_{\text{fpn}}, \sigma_{\text{john}}) = (7, 5, 10, 8, 3),$$

so ground-truth values are available for quantitative comparison.

B.3 Independence-based covariance estimator

To probe identifiability, we apply a simple estimator based on classical covariance relations: if the structured components in (12) were independent, one could estimate

$$\sigma^2_{\text{col}},\ \sigma^2_{\text{row}},\ \sigma^2_{1/f}$$

from the variance, same-row covariance, same-column covariance, and cross-row/column covariance of flat-field frames. This estimator is not meant to be a full physical reconstruction, nor is it used in the main paper; rather, it serves as a diagnostic tool for studying how the inferred "decomposition" reacts to increasing correlation ρ.

We evaluate two scenarios:

- Oracle FPN removal: subtracting the true static FPN map.
- Mean-based FPN removal: subtracting the temporal mean, as commonly done in flat-field correction.

For each ρ, we measure the mean relative error across the three structured components

$$\mathcal{E} = \tfrac{1}{3} \sum_{k \in \{\text{col,row,1/f}\}} |\widehat{\sigma}_k - \sigma_k| \,/ \sigma_k.$$

B.4 Results

Figure 6 shows how $\mathcal{E}$ varies as ρ increases from 0 to 1.

Key observations:

- When $\rho = 0$, identifiability is good. The independence relations hold, and the estimator accurately recovers the three structured sigmas.

- For moderate ρ, the decomposition drifts. Correlation between row and column terms makes them statistically indistinguishable from the global temporal component; the estimator reallocates variance incorrectly.
- For large ρ, the decomposition collapses. At $\rho \to 1$, almost all structure is absorbed by the global term regardless of FPN removal, leading to substantial wrong estimation.

Notably, the temporal-mean FPN estimate itself remains accurate throughout (the RMSE decreases from ≈ 0.30 at $\rho = 0$ to ≈ 0.10 at $\rho = 1$), indicating that the failure is not due to deficient FPN estimation but the intrinsic ambiguity in decomposing correlated structured noise using second-order statistics alone.

B.5 Interpretation

This toy experiment supports the main paper's modeling philosophy. Flat-field stacks strongly constrain the gain and offset maps, but the remaining spatio-temporal variability cannot, in general, be assigned uniquely to row-, column-, and global components when correlated structure is present. Thus it is more stable to calibrate the parts that are identifiable (gain/offset) and treat the rest as an empirically sampled residual rather than forcing a potentially misleading physical decomposition.

C Implementation details of calibration and temperature-aware synthesis

This section provides additional implementation details for the blackbody calibration procedure, residual noise library construction, and temperature-aware augmentation used in the synthesis pipeline.

C.1 Blackbody calibration protocol

Flat-field calibration data are collected using a temperature-controlled blackbody source under approximately uniform-emission conditions. For each calibration temperature T_m, the camera records a long temporal sequence after thermal stabilization, and the temporal mean frame $\bar{Z}(T_m)$ is used for subsequent pixel-wise regression. The calibration temperatures range from 20°C to 40°C with a step size of 5°C, resulting in five temperature levels in total. For each temperature setting, Camera 1 records a sequence of 10000 frames, Camera 2 a sequence of 1000 frames, after the blackbody emission becomes thermally stable.

Following Eq. (3) in the main paper, for each pixel (i, j) we fit an affine response model

$$\bar{Z}_{ij}(T_m) = G_{ij} T_m + O_{ij},$$

where G_{ij} and O_{ij} denote the effective gain and offset parameters, respectively. The parameters are estimated by least-squares fitting across all calibration temperatures.

Before regression, a small median filter is applied to each frame to suppress persistent hot or dead pixels that could otherwise introduce unstable values during the pixel-wise fitting process. This preprocessing step improves numerical robustness of the least-squares estimation while preserving the overall flat-field structure used for calibration.

In practice, this procedure produces camera-specific gain and offset maps that capture the dominant deterministic response nonuniformity of each sensor. As discussed in the main paper, these maps represent the identifiable part of the sensor model and serve as the basis for both non-uniformity correction and subsequent synthetic data generation.

C.2 Residual noise library construction

After gain-offset calibration, each observed frame is decomposed into a deterministic calibrated component and a residual term, following Eq. (4) in the main paper. Concretely, after subtracting the fitted gain-offset response from each flat-field frame, the remaining residual stack is stored as a library of real sensor noise samples.

These residual frames contain the structured fluctuations not explicitly parameterized in the minimal calibration model, including low-frequency drift, row/column-correlated readout artifacts, and stochastic sensor noise. Rather than imposing an analytic factorization of these effects, we directly reuse the measured residuals during synthesis.

Residual sampling is performed at the full-frame level. During synthesis, one residual frame is randomly drawn from the residual stack and added to the simulated observation so that the synthesized noisy images preserve the spatial statistics and temporal variability observed in real sensor measurements.

Residual frames collected at different calibration temperatures are pooled together in a single library without explicit temperature conditioning. Temperature-dependent effects are instead handled through the stochastic parameter perturbation described in the following subsection, which acts as an additional augmentation mechanism.

In total, the calibration dataset contains residual stacks from five temperature levels (20°C–40°C). For Camera 1, each temperature level provides 10 000 frames, resulting in about 50 000 residual samples in total. For Camera 2, each temperature level provides 1 000–1 500 frames, resulting in 5 000–7 500 residual samples.

In the current implementation, residual sampling is performed without fitting an additional analytic distribution to the residual stack. This keeps the synthesis process closely tied to the measured calibration statistics and avoids introducing unnecessary modeling assumptions for the non-identifiable part of the sensor behavior.

C.3 Temperature-dependent augmentation

To emulate operating conditions beyond the nominal calibration point, we introduce temperature-dependent perturbations to the calibration parameters during synthesis, following the stochastic formulation described in Eqs. (5)–(6) and Eq. (10) of the main paper.

In practice, offset variation is the dominant source of temperature-induced mismatch in uncooled thermal cameras. Empirically, residual drift and shutter-update inconsistency often appear as slowly varying bias fluctuations rather than large multiplicative gain changes. Therefore, we primarily model temperature-dependent effects through stochastic perturbations of the offset map, while gain perturbations are kept comparatively small.

To guide the design of the augmentation model, we analyze the empirical relationship between pixel bias and the temperature difference between the camera and the observed scene using the calibration dataset reported in (Oz et al., 2024). For each frame, we approximate the pixel bias by the difference between the measured frame mean and the blackbody reference signal. The temperature difference is approximated by the gap between the camera FPA temperature and the blackbody temperature.

Figure 7 visualizes the normalized relationship between these two quantities in log–log scale. After excluding degenerate cases where either quantity approaches zero, the majority of samples follow an approximate power-law trend. In particular, the observed scaling roughly spans slopes between 1 and 3 in the log–log domain.

Motivated by this empirical trend, we model the variance of the offset perturbation as a power-law function of the temperature difference:

$$\mathrm{Var}(\epsilon_O) \propto |T - T_0|^{\alpha},$$

where T denotes the temperature associated with the sampled calibration frame and T_0 represents a reference operating temperature. The exponent α is randomly sampled from the range $[1, 3]$ during synthesis.

To simulate a variety of operating conditions, the reference temperature T_0 is randomly sampled from a broad range (10°C–50°C). This allows the synthetic data generation process to cover a wide range of possible temperature mismatches between the camera and the observed scene.

The proportionality constant in the variance scaling is chosen through empirical normalization so that the resulting perturbation magnitude roughly matches the range of temperature-

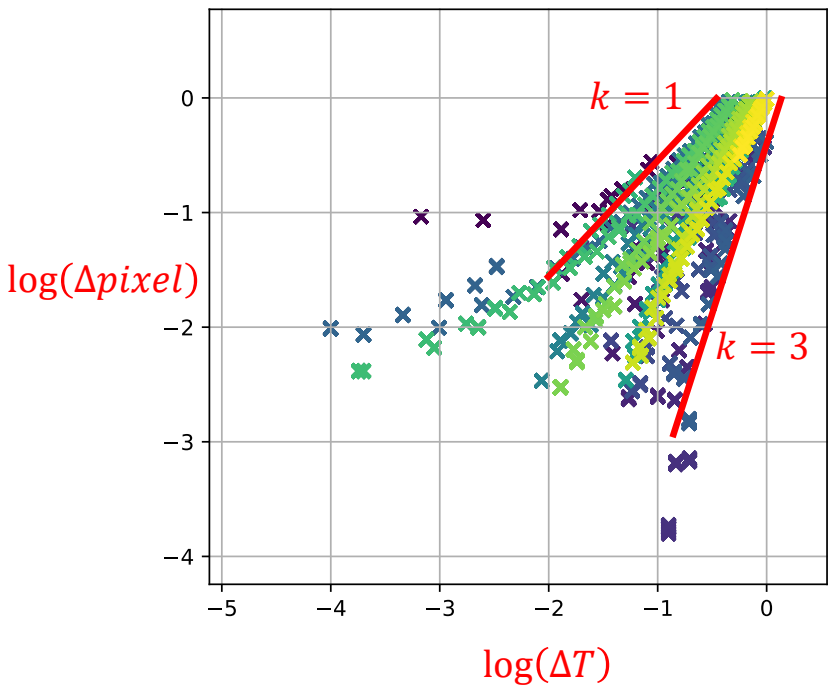


Figure 7: Empirical scaling between temperature difference and pixel bias. Using the calibration dataset from (Oz et al., 2024), we compute the frame-wise pixel bias as the difference between the measured frame mean and the blackbody reference signal. The temperature difference is approximated by the gap between the camera FPA temperature and the blackbody temperature. After normalization and log transformation (with a small offset added to avoid numerical instability), the resulting scatter distribution follows an approximate power-law trend. Most samples fall within slopes roughly between 1 and 3 in the log–log domain.

induced bias variations observed in the calibration data. This ensures that the synthetic perturbations remain within realistic operating regimes while still providing sufficient variability for training.

These perturbations are applied only during synthetic data generation and do not modify the calibrated gain or offset maps themselves. Their role is to broaden the support of the synthetic training distribution while keeping the synthesized noise statistically consistent with the empirical behavior observed in real sensor measurements.

D Training details

We provide additional training details for reproducibility. Unless otherwise stated, all models use the lightweight U-Net backbone described in the main paper and are optimized using Adam.

The batch size is set to 8. Training is performed on full-resolution images rather than on randomly cropped patches. The training pipeline consists of two stages: (1) pretraining on synthetic degradations generated from external grayscale content using the proposed noise synthesis pipeline, and (2) camera-specific fine-tuning using the calibrated gain, offset, and residual noise statistics of each thermal sensor. During pretraining, clean source images are resized or cropped to match the target camera resolution before noise synthesis. Specifically, images are prepared at 384×288 for Camera 1 and 640×512 for Camera 2. Synthetic noisy-clean pairs are then generated at the corresponding full-image resolution. During

Table 4: Synthetic evaluation on Camera 1. Quantitative comparison on synthetic noisy–clean pairs generated using Camera 1 calibration and residual statistics. The best score in each column is bolded and the second best is underlined.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑
BM3D	20.77	0.721	0.484	0.387	12.72	30.93
ELD	18.83	0.841	0.378	0.287	6.21	22.38
DeepIR	<u>27.51</u>	<u>0.944</u>	<u>0.172</u>	<u>0.189</u>	7.57	26.15
TV-DIP	20.28	0.718	0.489	0.383	11.86	29.97
Multiview FPN	20.67	0.718	0.494	0.394	12.66	<u>30.55</u>
Ours	24.48	0.897	0.201	0.198	5.05	28.07
Ours (finetuned)	33.71	0.968	0.079	0.112	<u>5.99</u>	30.41

fine-tuning, images captured by the target thermal camera are used directly, so the spatial resolution remains unchanged. When the 1000 fps thermal reference data are used for fine-tuning, they are cropped to the same target resolution as the corresponding camera before training.

The initial learning rate is 2×10^{-4}. Models are trained for 5000 epochs. Training is performed on two NVIDIA RTX 3090 GPUs and takes approximately 16 hours to complete.

E Additional results on synthetic test scenes

In the main paper, synthetic noisy–clean pairs generated by our calibration-and-synthesis pipeline serve two purposes: (1) establishing the statistical fidelity of the synthetic degradations, and (2) providing a pretraining signal before domain adaptation on real thermal data. Here we present comprehensive quantitative evaluations on two held-out synthetic test sets (one per camera), covering six complementary quality metrics.

E.1 Experimental setup

For each camera, we construct a synthetic test set by combining clean grayscale images (distinct from those used in training or validation) with the calibrated gain/offset maps and the residual noise library. The synthesis process follows exactly the same inverse–noise–forward pipeline described in the main paper, ensuring that the synthetic test inputs are produced under identical statistical assumptions.

We evaluate the same baselines as in the real-data experiments: BM3D (Dabov et al., 2007), ELD (Wei et al., 2020), DeepIR (Saragadam et al., 2021), TV-DIP (Liu et al., 2023), and Multiview FPN (Barral et al., 2024), as well as our proposed physics-informed model and its fine-tuned variant.

We report six metrics that capture distortion, perceptual distance, and no-reference image quality:

- distortion metrics: PSNR, SSIM;
- perceptual metrics (lower is better): LPIPS and DISTS;
- no-reference metrics (higher is better except NIQE): NIQE, MUSIQ.

All results are computed on synthetic noisy-clean pairs with the clean image serving as ground truth.

E.2 Quantitative results

Tables 4 and 5 summarize the quantitative results for Camera 1 and Camera 2, respectively. For clarity, the best score in each column is bolded, and <u>the second best score is underlined</u>.

On Camera 1, DeepIR achieves the strongest performance among existing methods, with high distortion scores (PSNR $\approx$ 27.5 dB, SSIM $\approx$ 0.94) and competitive perceptual quality. Our physics-informed model, trained purely on synthetic pairs, delivers comparable

Table 5: Synthetic evaluation on Camera 2. Quantitative comparison on synthetic noisy–clean pairs generated using Camera 2 calibration and residual statistics. The best score in each column is bolded and the second best is underlined.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑
BM3D	38.85	0.887	0.188	0.140	6.91	37.92
ELD	36.83	0.945	0.111	0.093	5.13	38.70
DeepIR	25.61	0.946	0.208	0.193	8.17	30.79
TV-DIP	22.37	0.868	0.216	0.156	6.68	38.74
Multiview FPN	41.70	0.890	0.173	0.115	6.64	38.07
Ours	37.70	0.962	0.088	0.088	5.41	39.41
Ours (finetuned)	40.49	0.981	0.071	0.092	6.69	41.85

perceptual performance while using a lightweight architecture. After fine-tuning using synthetic data aligned with the real-domain validation distribution, the model achieves the best PSNR, SSIM, LPIPS, and DISTS among the compared methods, improving PSNR to 33.7 dB and SSIM to 0.97 and attaining the lowest perceptual distances (LPIPS, DISTS), while remaining competitive on the no-reference metrics (NIQE, MUSIQ).

On Camera 2, classical methods such as BM3D and ELD obtain high PSNR values due to the milder noise distribution. Nevertheless, our physics-informed model substantially improves perceptual metrics and SSIM while maintaining competitive PSNR. After fine-tuning, it further improves PSNR to 40.5 dB and SSIM to 0.98, and achieves the highest MUSIQ score, indicating that the synthetic degradations produced by our pipeline not only resemble real sensor noise but also effectively support model optimization.

E.3 Qualitative results on synthetic test scenes

Beyond quantitative metrics, we also visualize representative qualitative results on synthetic test scenes for both cameras. Figure 8 shows example crops from the synthetic test sets, including the clean reference, the synthetically degraded input, several baselines (BM3D (Dabov et al., 2007), ELD (Wei et al., 2020), DeepIR (Saragadam et al., 2021), TV-DIP (Liu et al., 2023), and Multiview FPN (Barral et al., 2024)), and our method before and after fine-tuning.

On Camera 1, the synthetic noise is strong enough to obscure fine structural details and low-contrast textures. Classical denoisers reduce variance but tend to over-smooth edges or leave residual streak-like patterns, while DeepIR and Multiview FPN better recover structure at the cost of occasional artifacts. Our physics-informed model produces sharper edges and more coherent texture, and the fine-tuned version further suppresses residual noise without introducing hallucinated details, closely matching the clean reference.

On Camera 2, where the synthetic noise is comparatively milder, differences are more subtle but still visible. BM3D and ELD retain global contrast yet slightly blur thin structures and small high-frequency patterns. In contrast, our model preserves both contrast and texture, and the fine-tuned variant yields the cleanest backgrounds and most stable object boundaries, consistent with the improvements observed in PSNR/SSIM and perceptual metrics.

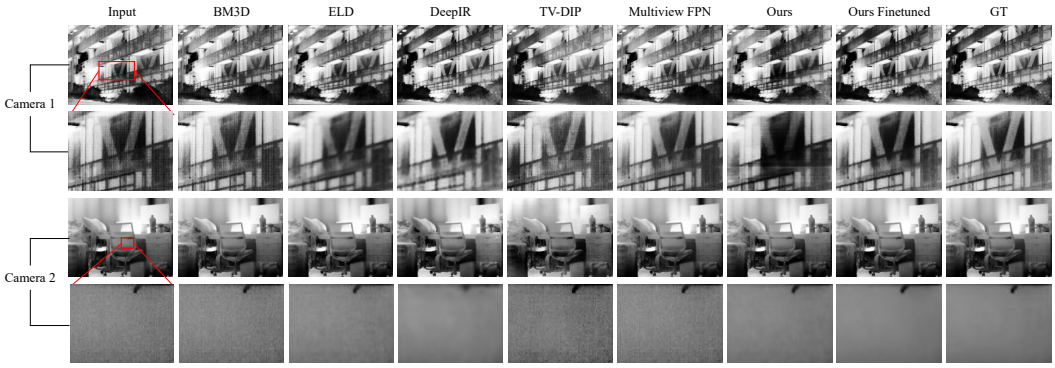


Figure 8: Qualitative results on synthetic test scenes for Camera 1 and Camera 2. For each camera, we show (from left to right): the synthetic noisy Input, denoised outputs from BM3D (Dabov et al., 2007), ELD (Wei et al., 2020), DeepIR (Saragadam et al., 2021), TV-DIP (Liu et al., 2023), Multiview FPN (Barral et al., 2024), our physics-informed model, and its fine-tuned variant, followed by the clean GT. Each top row shows the full-frame result, and the corresponding bottom row shows a magnified crop (highlighted in red in the first column). Across both cameras, classical baselines reduce noise but either blur high-frequency structure or leave residual row/column patterns, while our model better preserves edges and textures; the fine-tuned version produces the closest match to GT with minimal residual noise.