Rethinking Semantics for Complex-Scene Thermal Image Generation Supplementary Material

Tayeba Qazi[♣*], Ayush Maheshwari[♢], Prerana Mukherjee[♠], Brejesh Lall[♣]
[♣] Indian Institute of Technology Delhi, India
[♢] NVIDIA AI Technology Center, India
[♠] Jawaharlal Nehru University, India

A. Problem Definition

While RGB-to-thermal image synthesis has been extensively studied, a c ritical question remains not only unresolved but largely unexamined, whether incorporating additional contextual information, such as semantic maps, genuinely enhance thermal synthesis quality for complex, multi-object scenes, or is its benefit overstated and confined to simplistic benchmarks? A prevalent hypothesis posits that augmenting RGB input with contextual information, such as semantic segmentation map should provide critical emissivity priors to guide the synthesis of more physically plausible thermal outputs. The intuition is that thermal signatures are highly material-dependent, and explicitly providing this context should simplify the learning problem.

However, this widely-accepted assumption remains largely empirically unverified for complex, multi-object scenarios. Existing studies [2, 5] validating the benefits of semantic context have been confined to overly simplistic, single-object cases. The core objective of this work is to rigorously investigate two fundamental questions:

1. **Context Effectiveness**: Does incorporating additional contextual information (such as semantic segmentation maps) actually improve thermal image synthesis fidelity and quality for complex, multi-object scenes compared to RGB-only approaches?
2. **RGB Channel Analysis**: What is the individual contribution of each RGB channel (Red, Green, Blue) towards thermal generation, and how do these spectral components inform the cross-modal mapping process?

B. Mapping Paradigms

Formally, given a dataset of paired images $\mathcal{D} = (V_i, I_i)_{i=1}^{N}$, where $V_i \in \mathbb{R}^{3\times H\times W}$ is an RGB image and $I_i \in \mathbb{R}^{H\times W}$ is its corresponding thermal counterpart, we investigate two competing mapping paradigms:

RGB-Only Mapping: $\hat{I} = f_\phi(V)$, where $V \in \mathbb{R}^{3\times H\times W}$ is the input RGB image and ϕ denotes the learnable parameters of the mapping function. The use of visible spectrum images as contextual guidance for thermal synthesis is well-established in modern computer vision approaches. Building on pioneering work by ThermalDiff [4], we recognize that RGB imagery provides rich spectral information that correlates strongly with thermal emission properties. Specifically, different surface materials exhibit characteristic reflectance profiles across the red, green, and blue spectral bands that correspond to their thermal emission characteristics. This relationship enables RGB context to provide critical cues related to surface geometry, material composition, scene structure, and environmental conditions.

Context-Augmented Mapping: $\hat{I} = g_\psi(V, S)$, where $S \in \mathbb{R}^{1\times H\times W}$ represents a single-channel semantic map with spatial dimensions $H \times W$ and ψ denotes the learnable parameters of the mapping function g. While RGB images provide implicit material cues through spectral signatures, semantic segmentation maps offer explicit, pixel-wise material identification that directly informs the underlying physics of thermal emission. The fundamental challenge in thermal synthesis lies in accurately predicting material-specific emissivity values ε, which serve as the primary scaling factor for total thermal radiance [1] according to the Stefan-Boltzmann law: $W = \varepsilon\sigma T^4$ where W represents the total power radiated per unit area, T is the absolute temperature (Kelvin) and σ is the Stefan-Boltzmann constant.

Semantic segmentation maps establish a direct mapping from spatial coordinates (u, v) to material class $c \in 1, ..., C$, where each class corresponds to a characteristic emissivity value ε_c: $\varepsilon(u, v) = \mathcal{E}(S(u, v))$ Here, $\mathcal{E} : c \rightarrow \varepsilon_c$ represents a known physical mapping from se-

*Correspondence to: Tayeba Qazi bsz218186@iitd.ac.in

[1] It is important to note that thermal cameras in the LWIR band measure total infrared radiance, which includes both emitted radiation (function of temperature and emissivity) and reflected radiation from the environment. While emissivity is the primary material property affecting thermal emission, reflectivity also plays a role in determining the final thermal signature observed by the camera.

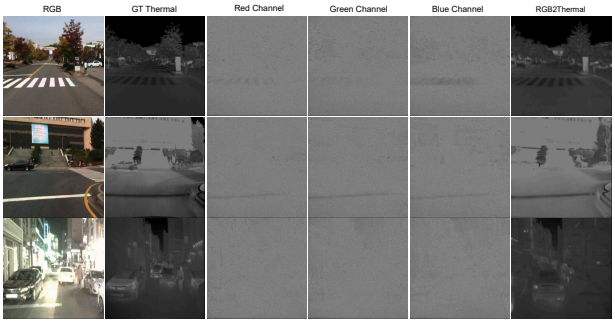

Figure 1. Contribution analysis of Red, Green, and Blue channels for thermal synthesis on the KAIST LWIR dataset [1]. As observed, single channels provide partial cues, red captures coarse reflectance, green encodes sharper textures, and blue enhances edges but only their combination yields synthesized thermal images closely matching ground truth.

mantic categories to material emissivity properties (e.g., $\mathcal{E}$('human skin') ≈ 0.98, $\mathcal{E}$('polished metal') ≈ 0.05).

This explicit material prior addresses fundamental ambiguities that challenge RGB-only approaches. For instance, while dark-colored automotive surfaces (low ε) and dark asphalt paving (high ε) may exhibit nearly identical RGB appearances, they possess radically different thermal emission characteristics due to their material composition. The semantic segmentation map resolves this ambiguity by providing categorical labels ('vehicle' vs. 'road'), thereby enabling physics-informed thermal prediction that incorporates known emissivity priors. This explicit material guidance theoretically enhances synthesis accuracy for complex multi-object scenes where precise material boundaries are critical for accurate temperature field estimation.

C. Investigating the Spectral Contribution of RGB Channels

To quantify the individual contribution of each spectral band to thermal synthesis, we decompose the RGB input into its constituent channels: $V = [V_R, V_G, V_B]$, where $V_c \in \mathbb{R}^{1\times H\times W}$ for $c \in R, G, B$

This decomposition enables targeted analysis of how different wavelengths inform thermal prediction. Each channel captures distinct material properties: red spectra correlate with surface composition and thermal proxies, green bands associate with vegetation and moisture content, and blue channels relate to surface reflectivity and material type. These spectral signatures provide complementary cues for inferring emissivity and thermal behavior.

We analyze the role of individual RGB channels in thermal synthesis using the KAIST Multispectral Dataset [1]. As quantified in Table 4. of the main paper and visualized in Figure 1, each channel provides distinct yet complementary information for thermal prediction.

Red Channel. Operating at the longest visible wavelengths (≈620–750 nm), this channel captures superficial material properties and provides weak thermal priors related to emissivity, but lacks critical structural detail when used in isolation.

Green Channel. Encoding the strongest structural and textural information (≈495–570 nm), this channel offers valuable spatial cues that correlate highly with thermal scene layout, achieving the highest PSNR (23.17 dB) and SSIM (0.63) among individual channels.

Blue Channel. Sensitive to high-frequency edges and contrasts (≈450–495 nm), this channel provides supplementary detail but is highly susceptible to atmospheric scattering, achieving the best LPIPS (0.270) and FID (32.35) scores.

As clearly visualized in Figure 1, single channels provide partial cues, red captures coarse reflectance, green encodes sharper textures, and blue enhances edges but only their combination yields synthesized thermal images closely matching ground truth. This complementary fusion bridges the spectral gap, confirming RGB imagery as a strong contextual prior for thermal synthesis.

D. Statistical Significance Testing

To quantitatively assess the realism of the generated thermal images, we compare the distributions of ground truth (GT) thermal data against those produced by the RGB2Thermal baseline [4] and our RGB+SegMap model (Model 1) using violin plots (Figure 2). The distributions are evaluated using three statistical measures: mean intensity, standard deviation, and kurtosis.

Our analysis reveals that the RGB2Thermal baseline produces a mean intensity slightly higher than the GT, indicating a mild but consistent over-estimation bias. In contrast, the RGB+SegMap model significantly under-estimates the mean, suggesting a substantial loss of thermal fidelity. In terms of variance, the baseline model maintains a standard deviation closely aligned with the GT, demonstrating stable generative diversity. The RGB+SegMap model, however, suffers from a severe collapse in variance, resulting in overly homogeneous and distributionally poor outputs.

Regarding kurtosis, the baseline yields moderately heavy-tailed distributions, retaining a degree of fine-scale variability. The RGB+SegMap model also produces heavy-tailed distributions but exhibits less stable peakedness, implying that semantic conditioning introduces training instabilities and leads to unreliable distributional alignment.

The training progression of the RGB+SegMap model further highlights its instability: it begins with extreme over-estimation and high dispersion, briefly approaches GT alignment mid-training, but ultimately diverges toward under-estimation and variance collapse. This erratic behavior indicates that semantic conditioning, while potentially

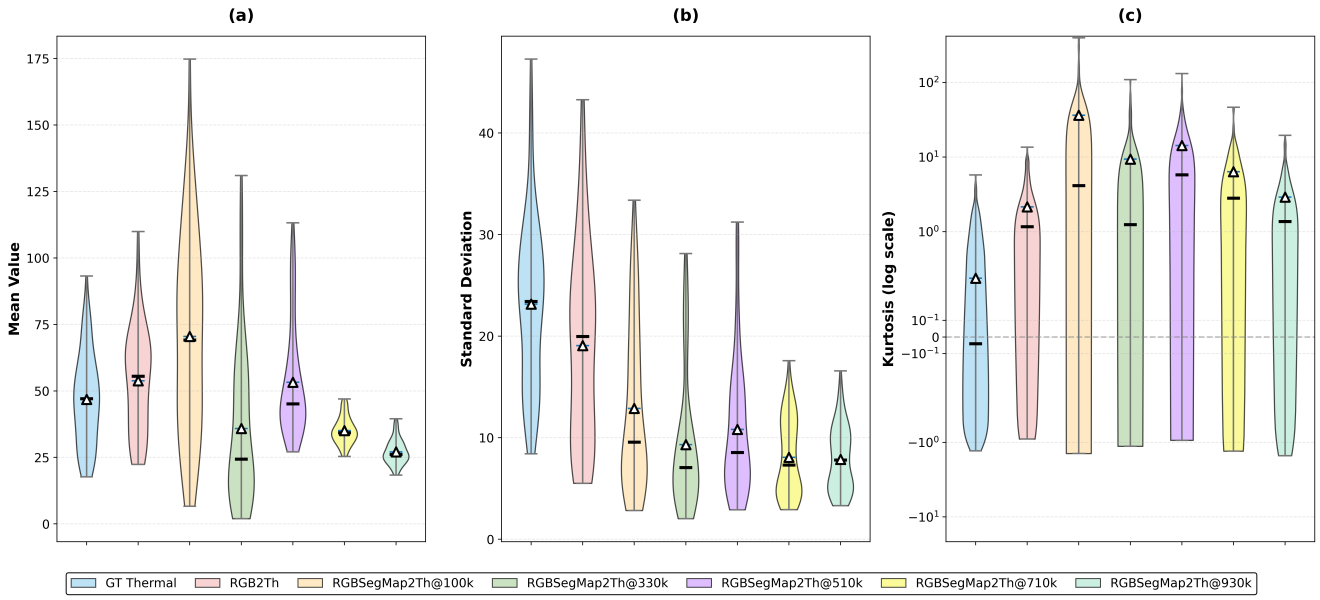

Figure 2. Statistical comparison of thermal image distributions: ground truth (GT), RGB2Th baseline [4], and RGB+SegMap2Th (Model 1) at various training stages. Violin plots show distributions of mean intensity, standard deviation, and kurtosis. RGB2Th maintains consistent but biased statistics, while RGB+SegMap2Th shows optimal alignment at intermediate training (510k) but deteriorates with extended training (930k).

informative, adversely affects convergence stability compared to the RGB-only baseline.

E. Understanding thermal image generation via visual analysis

In Figure 3, Figure 4 and Figure 5, we present visual comparison of the denoising trajectories, layer-wise DF-CAM visualizations and effect of timestep sampling strategy, respectively, for the RGB2Th baseline [4] and the RGB+SegMap2Th model (Model 1).

E.1. Regions recovered by the diffusion model in terms of semantic and detail level

We analyze the denoising trajectories of our baseline (RGB2Th) and segmentation-conditioned model (RGB+SegMap2Th). The baseline exhibits coherent, progressive refinement: coarse scene structures (e.g., vehicle contours, road layout) emerge by mid-denoising steps ($t \approx 1000$), with finer details consolidating in the final steps ($t < 200$). In contrast, the segmentation-conditioned model delays semantic recovery, suppressing object boundaries until later stages. This results in intermediate instability as the model struggles to reconcile the noisy segmentation prior with latent scene structure. While both models converge to plausible outputs, the baseline achieves smoother semantic alignment Figure 3.

E.2. Visual concepts prioritized by a given context

We employ Diffusion Gradient-weighted Class Activation Mapping (DF-CAM) [3] to visualize internal feature representations. The RGB2Thermal model [4] demonstrates a hierarchical focus: down-sampling layers emphasize global structural cues like road boundaries and vehicle silhouettes; mid-layers build semantic and spatial coherence; and up-sampling layers refine fine-grained details.

Conversely, the RGB+SegMap2Th model shows distorted prioritization. Its down- and mid-layer activations overfit to segmentation boundaries, focusing on noisy semantic regions at the expense of holistic contextual understanding. This impairs detail refinement in up-sampling layers, resulting in fragmented activations and diminished reconstruction quality. These findings confirm that while the baseline learns a balanced, hierarchical representation, the segmentation-conditioned model's over-reliance on its semantic prior disrupts this process, leading to reduced consistency and inferior output Figure 4.

E.3. Visual concept implied at timestep t

Analysis using non-uniform sampling [3] reveals fundamental differences in timestep usage. The RGB2Th baseline [4] exhibits a structured process: early timesteps ($t > 1500$) capture global semantics and thermal layout, while late timesteps ($t < 400$) refine local textures and contrast.

In contrast, the RGB+SegMap2Th model shows no coherent learning phase specialization. Early sampling fails to establish a robust semantic layout due to the noisy segmen-

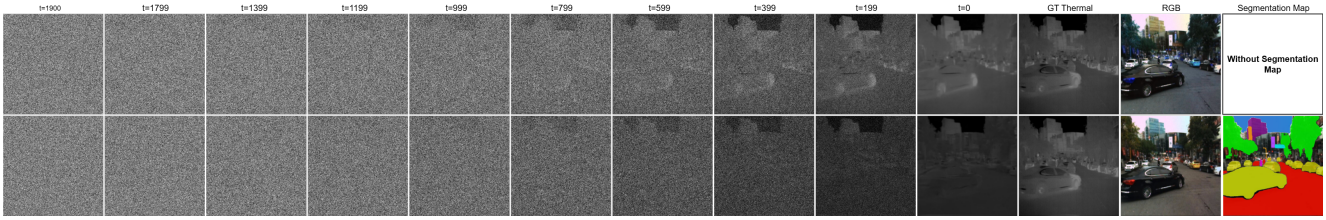

Figure 3. Visual comparison of the denoising trajectories for the RGB2Th baseline [4] (top) and the RGB+SegMap2Th model, Model 1 (bottom). The process evolves from noise ($t = 2000$) to the final generated image ($t = 0$). The rightmost column shows the conditioning input for each row: RGB image for the baseline, and both the RGB and segmentation map for Model 1. The baseline model shows more coherent and progressive semantic structuring compared to Model 1, which exhibits delayed and less stable recovery due to its noisy segmentation prior. (Note: Brightness and contrast of RGB+SegMap2Th samples are enhanced for visualization).

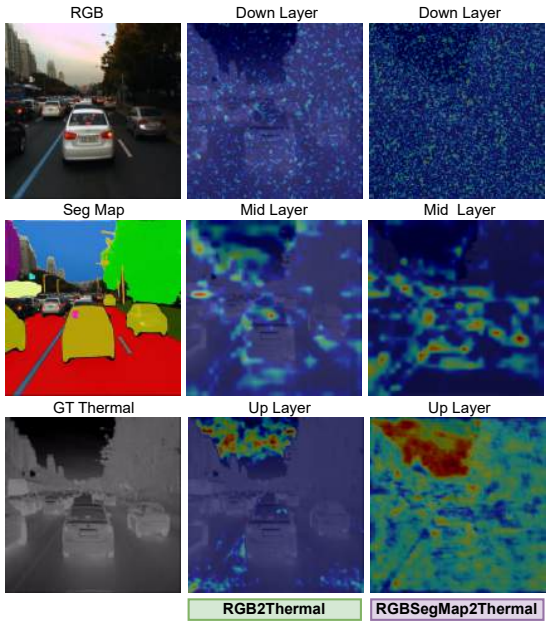

Figure 4. Layer-wise DF-CAM (Diffusion Gradient-weighted Class Activation Mapping) visualizations of U-Net feature activations. The visualizations show contextual inputs (RGB or RGB+SegMap) with corresponding ground truth thermal imagery. Activations from the down, mid, and up layers of the RGB2Th baseline [4] model emphasize global structural cues in early layers and progressively refine semantic and thermal details in later stages. In contrast, activations from the RGB+SegMap2Th model display over-sensitivity to segmentation boundaries and fragmented refinement, highlighting the disruptive influence of noisy semantic priors on thermal image generation.

tation prior, while late sampling generates discordant local details decoupled from global structure. This absence of hierarchical learning demonstrates that the semantic prior provides no useful information, instead acting as a disruptive source of noise Figure 5.

References

[1] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 2

[2] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[3] Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248:123231, 2024. 3

[4] Tayeba Qazi, Brejesh Lall, and Prerana Mukherjee. Thermaldiff: A diffusion architecture for thermal image synthesis. *Journal of Visual Communication and Image Representation*, 111: 104524, 2025. 1, 2, 3, 4, 5

[5] Doan Thinh Vo, Phan Anh c, Nguyen Nhu Thao, and Huong Ninh. An approach to synthesize thermal infrared ship images. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 1

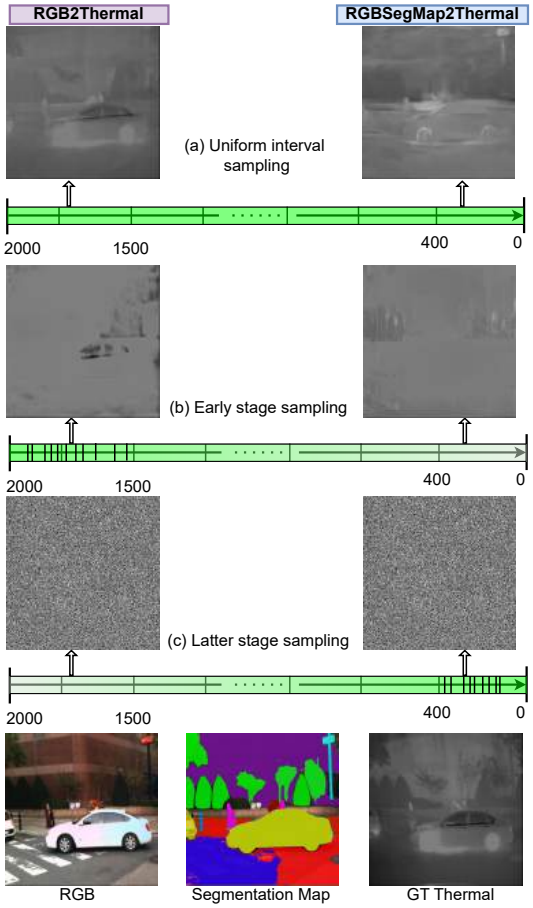

Figure 5. Effect of timestep sampling strategy on generation quality. The RGB2Th baseline [4] (left) maintains a coherent structure across all strategies. The RGB+SegMap2Th model (right) is highly unstable: early sampling ($t > 1500$) fractures semantics by overfitting to the segmentation map, while late sampling ($t < 400$) produces incoherent local details. The results confirm that the additional segmentation context does not enhance generated image fidelity and introduces a critical dependency on balanced (uniform) training for stability. (Note: Brightness and contrast of RGB+SegMap2Th samples are enhanced for visualization.)