SmokeSeer: 3D Gaussian Splatting for Smoke Removal and Scene Reconstruction

Neham Jain, Andrew Jong, Sebastian Scherer, Ioannis Gkioulekas
Carnegie Mellon University
Pittsburgh, PA

nhjain@andrew.cmu.edu, ajong@andrew.cmu.edu, basti@andrew.cmu.edu, igkioule@cs.cmu.edu

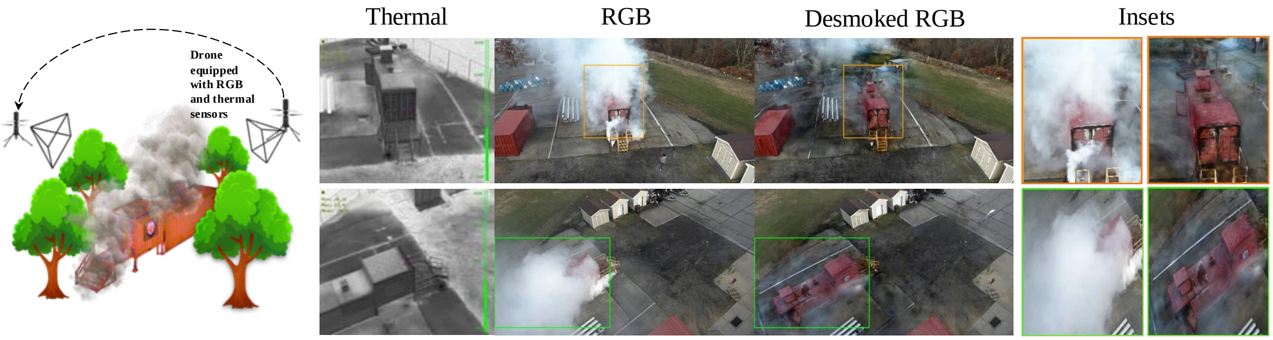

Figure 1. Our method utilizes RGB and thermal images from a drone-mounted sensor to perform simultaneous 3D scene reconstruction and smoke removal using an inverse rendering approach within the 3D Gaussian splatting framework. Insets highlight the effectiveness of our approach in revealing occluded structures.

Abstract

Smoke in real-world scenes can severely degrade image quality and hamper visibility. Recent image restoration methods either rely on data-driven priors that are susceptible to hallucinations, or are limited to static low-density smoke. We introduce SmokeSeer*, a method for simultaneous 3D scene reconstruction and smoke removal from multi-view video sequences. Our method uses thermal and RGB images, leveraging the reduced scattering in thermal images to see through smoke. We build upon 3D Gaussian splatting to fuse information from the two image modalities, and decompose the scene into smoke and non-smoke components. Unlike prior work, SmokeSeer handles a broad range of smoke densities and adapts to temporally varying smoke. We validate our method on synthetic data and a new real-world smoke dataset with RGB and thermal images. We provide an open-source implementation and data on the project website.*[1]

1. Introduction

Reliable visual perception is essential for safety-critical applications such as search and rescue, robot navigation, and industrial inspection. The ability to accurately perceive and reconstruct 3D environments is particularly vital, as it enables precise spatial reasoning and path planning in complex scenarios. For example, firefighters navigating through burning buildings increasingly depend on vision-based systems to maintain situational awareness. However, dense smoke severely compromises these systems, obscuring vital environment details and increasing operational risks. Developing technologies that enable these systems to “see through smoke” is therefore critical for enhancing both safety and operational effectiveness in these hazardous environments.

Though several approaches have targeted the problem of enhancing visibility through scattering media, significant limitations remain. Learning-based approaches that map hazy to clear images require extensive paired datasets and typically process individual frames, thereby ignoring valuable multi-view constraints. Closer to our work, neural rendering approaches such as ScatterNeRF [24] and DehazeNeRF [4] incorporate physical light transport models and operate on multi-view RGB data. However, all these approaches primarily address static haze removal and are

[1] https://imaging.cs.cmu.edu/smokeseer

ill-equipped to handle dense, temporally evolving smoke.

We build an end-to-end system that performs joint 3D scene reconstruction and smoke removal in the presence of dense, temporally evolving smoke. Our method uses images from RGB and thermal cameras, and is effective on real-world smoke data (Figure 1). We build upon 3D Gaussian splatting (3DGS) [14] and decompose a smoke-filled scene into two sets of Gaussians: one representing the smoke part, and another representing the non-smoke part of the scene, which we refer to as the surface Gaussians. This decomposition allows us to render only the surface Gaussians to visualize the scene without smoke.

Performing this decomposition using only RGB images is challenging due to the visual ambiguity between light-reflecting surfaces and light-scattering smoke particles. To address this challenge, we leverage thermal cameras that capture long-wavelength infrared radiation, which is substantially less affected by scattering in smoke than visible light. This property enables thermal sensors to preserve critical spatial information even in dense smoke conditions. However, thermal images are low-resolution, have low contrast, and lack the texture details crucial for object recognition and scene understanding. Our method overcomes this limitation through a joint optimization strategy that fuses the robust spatial cues from thermal data with the rich texture information provided by RGB imagery.

To effectively leverage the complementary strengths of both modalities, we propose a three-stage approach for smoke removal and 3D scene reconstruction. In the first stage, we leverage advances in 3D foundation models [29] to estimate RGB-thermal poses in the same coordinate frame. In the second stage, we learn the scene's geometry exclusively from thermal images, leveraging their robustness in capturing spatial information even in the presence of dense smoke. In the third stage, we use both RGB and thermal images to optimize two sets of Gaussians, for the smoke and the scene's surfaces. For the surface Gaussians, we rely on initialization from the output of the second stage. For the smoke Gaussians, we use a deformation field to model the temporal variation of smoke, and enforce handcrafted priors based on physical properties of smoke. These choices help ensure that, after optimization, smoke Gaussians exclusively capture the scene smoke, whereas surface Gaussians accurately represent the underlying scene structure.

Unlike prior learning-based dehazing methods, ours does not directly rely on image-to-image learned priors and instead formulates smoke removal as an inverse rendering problem within the 3DGS framework. To the best of our knowledge, this is the first work that jointly uses RGB and thermal images for smoke removal and 3D reconstruction.

Our experiments show state-of-the-art results on both simulated and real-world datasets—collected in partnership with our county's fire department using a field operational drone—for smoke removal and novel view synthesis. Our code and data are publicly available on the project website, to ensure reproducibility and facilitate follow-up research.

2. Related Work

2.1. Image-based methods for haze removal

Traditional methods. Koschmieder [15] developed an atmospheric scattering model that describes image formation under haze as a combination of direct attenuation and airlight. This model is a simplification of the more general radiative transfer equation (RTE) [3], which describes the propagation of light through a medium with scattering and absorption. Though widely used in dehazing methods, the Koschmieder model assumes homogeneous static media, limiting its effectiveness for heterogeneous, dynamic smoke conditions.

Early image restoration approaches relied on handcrafted priors to estimate physical parameters in the Koschmieder model. He et al. [11] tried to estimate the attenuation map by leveraging the observation that in most local patches of haze-free images, at least one color channel has very low intensity. Zhu et al. [36] proposed the color attenuation prior, modeling the depth of the scene through the difference between brightness and saturation. Berman et al. [2] developed a non-local method based on the observation that colors in haze-free images form tight clusters in RGB space. Though effective for thin homogeneous haze, these methods fail in dense smoke scenarios, for which their priors are ill-suited.

Learning-based methods. Some recent methods map hazy to clear images without explicit parameter estimation. Examples include MSRL-DehazeNet [19], collaborative inference frameworks for dense haze in remote sensing [30], and saliency-guided mechanisms for UAV imagery [35].

Transformer-based architectures have recently shown promising results for dehazing. Zamir et al. [33] proposed Restormer, an efficient transformer for high-resolution image restoration including dehazing. Guo et al. [10] introduced a hybrid CNN-transformer architecture that combines local and global feature extraction. Despite these advances, most learning-based methods process individual frames independently, ignoring valuable temporal and multi-view information that could enhance smoke removal performance.

Specific to smoke removal, Salazar-Colores et al. [27] developed an image-to-image translation approach guided by an embedded dark channel for desmoking laparoscopy surgery images. However, this and other similar methods typically require paired training data (smoke versus smoke-free), which is challenging to obtain in real-world scenarios, especially for temporally varying smoke.

2.2. Neural representations for participating media

Neural radiance Fields (NeRF) [22] have revolutionized scene representation using continuous volumetric functions. Several works have extended NeRF to handle participating media such as smoke and haze. ScatterNeRF [24] incorporates the Koschmieder model into the NeRF framework, but remains limited to homogeneous haze conditions. DehazeNeRF [4] can handle heterogeneous media but not dynamic smoke. These methods have primarily focused on static haze removal and do not address the more challenging problem of temporally varying smoke—our focus.

3D Gaussian splatting (3DGS) [14] is an efficient alternative to NeRF through scene representation using 3D Gaussians, enabling real-time rendering. Dynamic 3DGS [21] extends this framework to dynamic scenes, but does not specifically address participating media.

Lastly, recent approaches such as ThermalNeRF [18] and ThermalGaussian splatting [20] incorporate thermal imaging into neural rendering frameworks but do not tackle the problem of imaging through smoke.

2.3. Multi-modal sensing

Multi-modal sensing has emerged as a promising direction for robust perception in challenging environments. Thermal imaging, which captures long-wavelength infrared radiation, is less affected by smoke and haze compared to RGB cameras [9]. Hwang et al. [12] demonstrated the effectiveness of fusing RGB and thermal information for object detection in adverse weather. Li et al. [17] proposed an RGB-thermal object tracking benchmark demonstrating the value of thermal information for robust perception.

Our work bridges these research areas by explicitly modeling temporally varying smoke separately from scene geometry within the 3DGS framework. Unlike previous approaches, ours leverages the complementary strengths of RGB and thermal imaging to achieve 3D reconstruction and smoke removal without requiring paired training data.

3. Method

We introduce *SmokeSeer*, a framework for simultaneous 3D scene reconstruction and smoke removal using RGB-thermal image pairs. Our approach leverages the complementary strengths of RGB (texture-rich) and thermal (smoke-penetrating) modalities to address the challenges of dense, dynamic smoke in safety-critical applications. Our method comprises three stages (Figure 3): (1) camera pose estimation and smoke segmentation, (2) initial surface reconstruction from thermal images, and (3) joint optimization of surface and smoke using both RGB and thermal images.

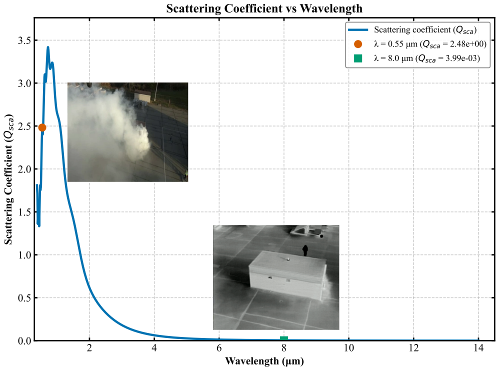

Figure 2. Scattering coefficient as a function of wavelength (in μm). We calculate the scattering coefficient using size, refractive index using standard values for organic matter found in smoke particles [1]. The scattering coefficient is significantly higher in the visible spectrum (0.38–0.7 μm) compared to the long-wave infrared spectrum (8–14 μm). Inset images which are taken from a drone at roughly the same time illustrate this effect: in visible light (left), smoke strongly obscures the scene, while in thermal infrared (right), the underlying structure is clearly visible.

3.1. Use of thermal images

Mie theory [8] provides a framework for understanding how different types of particles—such as smoke particles—interact with electromagnetic radiation at different wavelengths. For smoke particles of a given size and refractive index, we can use the Mie theory equations to characterize their wavelength-dependent scattering behavior, as illustrated in Figure 2. This analysis reveals a crucial insight: smoke particles predominantly scatter wavelengths in the visible spectrum (0.38–0.7 μm), where RGB cameras operate. However, in the long-wave infrared (LWIR) spectrum (8–14 μm) utilized by thermal cameras, scattering effects from smoke particles are negligible. This property allows thermal imaging to penetrate smoke and reveal underlying surface geometry otherwise obscured in RGB imagery.

In practice, smoke exhibits two key thermal behaviors. First, smoke is largely transparent in the long-wave infrared (LWIR) spectrum because heat dissipates rapidly as smoke moves away from the fire source. This transparency enables thermal cameras to capture clear views of scene geometry even when RGB cameras are completely occluded by dense smoke. However, in regions extremely close to the fire source, hot smoke can become emissive and appear as a thermal source rather than remaining transparent, which our method accounts for in the joint optimization stage.

3.2. Background on 3D Gaussian splatting

Given a collection of posed images $\{I_k\}_{k=1}^K$, $I_k \in \mathbb{R}^{H \times W}$ captured from a scene, 3DGS aims to reconstruct a representation $\mathcal{G}$ of the scene as a set of 3D Gaussians $\mathcal{G} = \{g_i\}$.

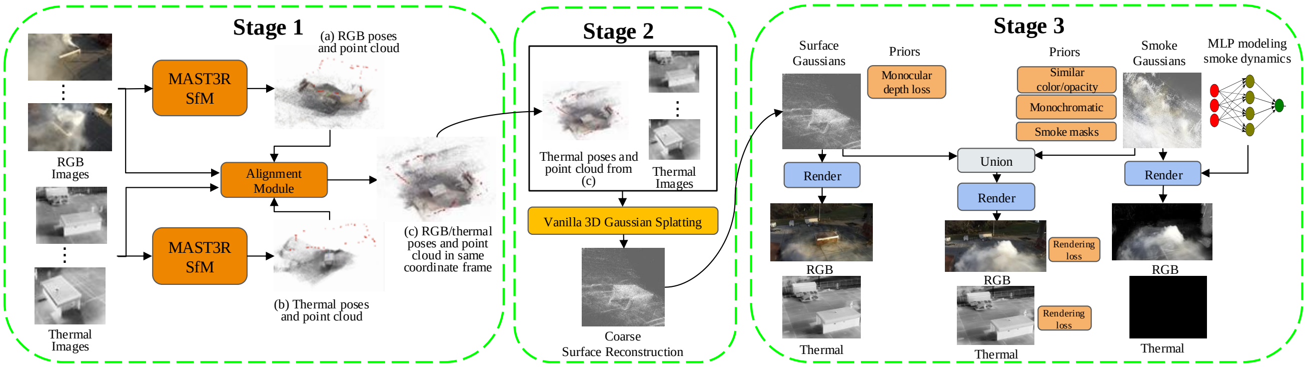

Figure 3. An overview of our method, SmokeSeer. The framework consists of three primary stages: (1) Camera pose estimation and smoke segmentation, (2) Initial surface reconstruction from thermal images, and (3) Joint optimization of surface and smoke plume using both RGB and thermal images.

Each Gaussian primitive g_i is characterized by a center position $\boldsymbol{\mu}_i$, a symmetric positive-definite covariance matrix $\boldsymbol{\Omega}_i$, an alpha value α_i, and appearance attributes encoded using spherical harmonic coefficients $\boldsymbol{h}_i$ [23]. Unlike approaches requiring different representations for surfaces (e.g., meshes or implicits) and volumes (e.g., voxel grids), Gaussian primitives can represent both surfaces and smoke, simplifying optimization and rendering.

3.3. Modeling scattering media using Gaussians

We decompose the smoke-filled scene into two sets of Gaussians: surface Gaussians $\mathcal{G}$ representing surfaces in the scene, and smoke Gaussians $\mathcal{S}$ capturing the dynamic smoke plume. Before detailing these sets, we explain how to render images using Gaussian primitives.

We first define the transmittance function, which is central to volumetric rendering. For a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ starting at position $\mathbf{o}$ in direction $\mathbf{d}$, the transmittance $T_\sigma(t)$ represents the probability that the ray travels from its origin to point $\mathbf{r}(t)$ without obstruction. It is defined as:

$$T_\sigma(t) \coloneqq \exp\left(-\int_{t_n}^{t} \sigma(s)ds\right), \tag{1}$$

where $\sigma(s)$ is the density function along the ray, and t_n is the near-plane distance. In a scene with both surfaces and smoke, we have two density functions: $\sigma(t)$ for surfaces and $\sigma_\text{s}(t)$ for smoke. The combined transmittance is:

$$\begin{aligned} T_{\sigma+\sigma_\text{s}}(t) &= \exp\left(-\int_{t_n}^{t} [\sigma(s) + \sigma_\text{s}(s)]\, ds\right) & (2)\\ &= T_\sigma(t) \cdot T_{\sigma_\text{s}}(t), & (3)\end{aligned}$$

which represents the probability of the ray reaching point $\mathbf{r}(t)$ without hitting either a surface or smoke particles.

Chen et al. [4] have shown that the volume rendering equation for a scene with mixed density takes the form:

$$\begin{aligned} C(\mathbf{r}, \mathbf{d}) = &\underbrace{\int_{t_n}^{t_0} c(\mathbf{r}(t), \mathbf{d})\sigma(t)T_{\sigma+\sigma_\text{s}}(t)\, dt}_{C_\text{surface}} \\ &+ \underbrace{\int_{t_n}^{t_0} c_\text{s}(\mathbf{r}(t))\sigma_\text{s}(t)T_{\sigma+\sigma_\text{s}}(t)\, dt,}_{C_\text{smoke}} \end{aligned} \tag{4}$$

where t_0 is the far-plane distance. Our dual Gaussian representation directly maps to this equation, where the surface Gaussians $\mathcal{G}$ correspond to C_surface with color c and opacity σ, whereas the smoke Gaussians $\mathcal{S}$ correspond to C_smoke with color c_s and opacity σ_s. Rendering the union $\mathcal{G} \cup \mathcal{S}$ is equivalent to computing the rendering equation (4).

The rendering equation for the clear-view surfaces without smoke interference is given by:

$$C_\text{clear}(\mathbf{r}, \mathbf{d}) = \int_{t_n}^{t_0} c(\mathbf{r}(t), \mathbf{d})\sigma(t)T_\sigma(t)\, dt. \tag{5}$$

Rendering only the surface Gaussians $\mathcal{G}$ is equivalent to computing the rendering equation (5). By modeling surface and smoke separately, we achieve effective smoke removal through selectively rendering only surface Gaussians.

3.4. Modality-specific representations

Building on the Gaussian representation described in Section 3.2, we extend the model to handle both RGB and thermal modalities. We use $\{I_k^\text{RGB}\}_{k=1}^{K_\text{RGB}}$ and $\{I_k^\text{T}\}_{k=1}^{K_\text{T}}$ to denote our RGB and thermal image collections respectively, with associated camera poses P_k^RGB and P_k^T.

For our dual Gaussian representation:

- Surface Gaussians $\mathcal{G}$ maintain the parameters from Section 3.2 but with modality-specific spherical harmonic coefficients ($\boldsymbol{h}_i^\text{RGB}$, $\boldsymbol{h}_i^\text{T}$), and modality-shared opacity α.

- Smoke Gaussians $\mathcal{S}$ have modality-specific spherical harmonic coefficients and opacities ($\alpha_i^{\text{RGB}} \gg \alpha_i^{\text{T}}$), reflecting the physical properties described in Section 3.1. In addition, they are time-varying to capture smoke dynamics.

3.5. Stage 1: Generating segmentation masks and obtaining poses

In this stage, our objective is to estimate camera poses for RGB and thermal images in a common coordinate system. Accurate cross-modal poses are a prerequisite for any multi-view fusion; without them, correspondence and consistency losses are ill-defined. This task is challenging due to the different sensor responses between these modalities, which complicates cross-modal feature matching. Additionally, the featureless appearance and dynamic smoke in RGB images impede reliable feature extraction.

We address these challenges with a three-step approach:

1. *Smoke segmentation:* We use GroundedSAM [26], based on SAMv2 [25], to identify and mask out smoke-affected regions in RGB images. Doing so ensures we match only features from reliable, smoke-free areas.
2. *Independent 3D reconstructions:* We run MAST3R-SfM [7] independently on RGB and thermal images. Using the masks from the previous step, we discard matches in the smoke regions of RGB images. Though MAST3R-SfM handles RGB-RGB and thermal-thermal matching well, it struggles with RGB-thermal matching.
3. *Cross-modal registration:* We use MINIMA [13], which is specialized for cross-modality matching, to establish 2D correspondences between RGB-thermal image pairs. We then lift these correspondences to 3D using the 2D-3D mappings from the per-modality calibration. Doing so enables the estimation of a similarity transform $T \in \text{Sim}(3)$ that aligns the RGB and thermal coordinate systems.

3.6. Stage 2: Reconstructing the scene using thermal images

In this stage, we obtain a first reconstruction of the scene geometry using only thermal images, which are minimally affected by smoke. We run vanilla 3D Gaussian splatting [14] on the thermal sequence, which outputs a smoke-free representation of the scene geometry. The surface reconstruction is coarse due to the low resolution of thermal images, but serves as a reliable initialization for our surface Gaussians.

3.7. Stage 3: Fusing RGB-thermal information and refining geometry

In the final stage, we jointly optimize surface and smoke Gaussian sets using both RGB and thermal images:

- *Surface Gaussians:* Initialized from Stage 2, these Gaussians remain static and maintain identical opacity across modalities. We augment them with spherical harmonic coefficients to capture RGB appearance.
- *Smoke Gaussians:* Randomly initialized within the scene bounds, these Gaussians evolve temporally and exhibit modality-dependent opacity, to model smoke's varying visibility in RGB versus thermal images (Section 3.1). Though in principle we could use Mie theory to model opacities, we opt for a more flexible approach with two independent variables for smoke visibility in each modality.

3.7.1 Modeling the dynamic smoke

Our approach explicitly accounts for the temporal evolution of smoke, which is critical for applications such as firefighting where smoke behavior is dynamic and unpredictable. Accounting for smoke motion enables more accurate surface reconstruction in areas temporarily occluded by passing smoke, and improves separation of surface and smoke. We model the dynamics of smoke following the deformable 3D Gaussians framework [32]. This framework uses 3D Gaussians in a canonical space, along with a deformation field to model motion over time. To model this field, we use a multilayer perceptron (MLP) that takes as input the positions of the 3D Gaussians and a timestep t, and outputs offsets in position, scale, and rotation. These offsets transform the canonical 3D Gaussians to the deformed space at each time. We use a bimodal Gaussian distribution following [16] to model smoke opacity as a function of time.

3.7.2 Priors on properties of smoke Gaussians

To facilitate accurate surface-smoke separation and modeling of realistic smoke behavior, during optimization we use priors motivated by physical properties of smoke:

- *Smoke consistency:* We minimize variance in opacity and color across smoke Gaussians:

$$L_{\text{smoke\_alpha}} = \text{Var}(\{\alpha_i\}_{i\in\mathcal{S}}) \tag{6}$$

$$L_{\text{smoke\_color}} = \text{Var}(\{c_i\}_{i\in\mathcal{S}}) \tag{7}$$

  This prior reflects the physical observation that smoke particles in a local region typically have similar optical properties. In real smoke, particles of similar size and composition would have nearly identical opacity and scattering properties. By enforcing consistency across smoke Gaussians, we prevent unrealistic variations from arising during optimization. Though the loss should ideally apply to Gaussians in local neighborhoods, we found that applying it across all Gaussians works well in practice.
- *Monochromaticity:* We enforce consistent color channels across smoke Gaussians:

$$L_{\text{mono}} = \sum_{i\in\mathcal{S}} \text{Var}(c_i^{\text{R}}, c_i^{\text{G}}, c_i^{\text{B}}). \tag{8}$$

  This prior reflects the physical property that smoke typically appears as a neutral gray color. It prevents our model from generating implausible colored smoke.

- *Depth consistency:* We align the surface Gaussians with monocular depth cues:

$$L_{\text{depth}} = \|d_i - \hat{d}_i\|, \tag{9}$$

where d_i denotes predicted depth on a thermal image using a monocular depth estimation model [31] and $\hat{d}_i$ is the rendered depth from the surface Gaussians using thermal camera parameters. This prior leverages the smoke-penetrating property of thermal imaging (Section 3.1). Since thermal images are minimally affected by smoke, they provide reliable depth cues for the underlying surface geometry, helping to prevent surface Gaussians from being incorrectly positioned in smoke-occluded regions.
- *Mask alignment:* The alpha values of smoke Gaussians should be consistent with the masks from Stage 1:

$$L_{\text{mask}} = \|M_{\text{pred}} - M_{\text{GT}}\|_1, \tag{10}$$

where M_{pred} and M_{GT} are the pixel-wise accumulated alpha values of the rendered smoke Gaussians and segmentation masks, respectively. This prior ensures spatial consistency between our reconstructed smoke volume and the observed smoke regions in input images. It helps constrain the optimization to place smoke Gaussians in only regions with smoke present, and prevent them from appearing in smoke-free ones.

The total optimization loss is a weighted sum of these physically-motivated priors and a standard rendering loss for RGB and thermal images:

$$\begin{aligned} L_{\text{total}} = {} & \lambda_{\text{render}} L_{\text{render}} + \lambda_{\text{smoke\_alpha}} L_{\text{smoke\_alpha}} \\ & + \lambda_{\text{smoke\_color}} L_{\text{smoke\_color}} + \lambda_{\text{mono}} L_{\text{mono}} \\ & + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{mask}} L_{\text{mask}}. \end{aligned} \tag{11}$$

This formulation enables separation of scene geometry from smoke, while maintaining physical consistency across the RGB and thermal modalities.

4. Experimental evaluation

We evaluate our method on synthetic and real-world datasets to demonstrate its effectiveness for smoke removal and 3D scene reconstruction. We compare against state-of-the-art methods, and validate our design choices through ablation studies. We provide implementation details in the supplement, and video results on the project website.

4.1. Datasets

Synthetic dataset. For quantitative evaluation with ground truth, we create a synthetic dataset using Blender's Mantaflow [28] smoke simulator. The dataset comprises 10 scenes: 5 object-level scenes from the NeRF synthetic dataset [22], and 5 large-scale scenes. For each scene, we generate 150 RGB and thermal frames with dynamic smoke.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
ImgDehaze + 3DGS	14.42	0.37	0.318
Ours (RGB only)	15.08	0.38	0.326
Ours (Full)	**19.92**	**0.76**	**0.247**

Table 1. Quantitative results on the synthetic dataset for novel view synthesis. Our full method outperforms all baselines.

Real-world dataset. In collaboration with our county's fire department, we collected a real-world dataset using a Spirit drone equipped with roughly co-located RGB and thermal cameras. There is no time synchronization between the frames captured by the RGB and thermal cameras on the drone, which makes relative pose estimation challenging. We do not report quantitative metrics for the real-world dataset as obtaining true ground truth is challenging in such environments. Instead, we provide an approximation which we refer to as "Reference" in the figures. We provide more details in the supplement. This dataset presents several challenges not found in synthetic data, including: imperfect alignment between RGB and thermal cameras, unpredictable smoke motion due to wind, and motion blur from drone movement. These factors make our real-world dataset a rigorous benchmark for evaluating the practical utility of smoke removal algorithms in safety-critical applications.

4.2. Baseline methods

We compare three methods:

- *ImgDehaze + 3DGS*: A two-stage approach that first applies a state-of-the-art single-image dehazing method (ConvIR [6]) to each RGB frame, then uses deformable 3DGS [32] on the dehazed images.
- *Ours (RGB only)*: Our approach using only RGB images (Stage 3 without thermal input).
- *Ours (Full)*: Our complete approach using both RGB and thermal images.

We could not compare with DehazeNeRF [4], the prior work closest to ours, due to lack of open-source code.

4.3. Results

Synthetic data results. Table 1 presents quantitative results for novel view synthesis on our synthetic dataset, using the PSNR, SSIM, and LPIPS [34] metrics. Our full method outperforms all baselines across all metrics, especially in scenes with heavy smoke where we achieve a PSNR gain of up to 4.8 dB over the RGB-only approaches.

Figure 4 shows qualitative results on the synthetic dataset. Our method removes smoke while preserving fine details in the scene. In contrast, baseline methods either fail to completely remove smoke or introduce artifacts.

Real-world data results. Figure 5 demonstrates our method's effectiveness on real-world data. The improve-

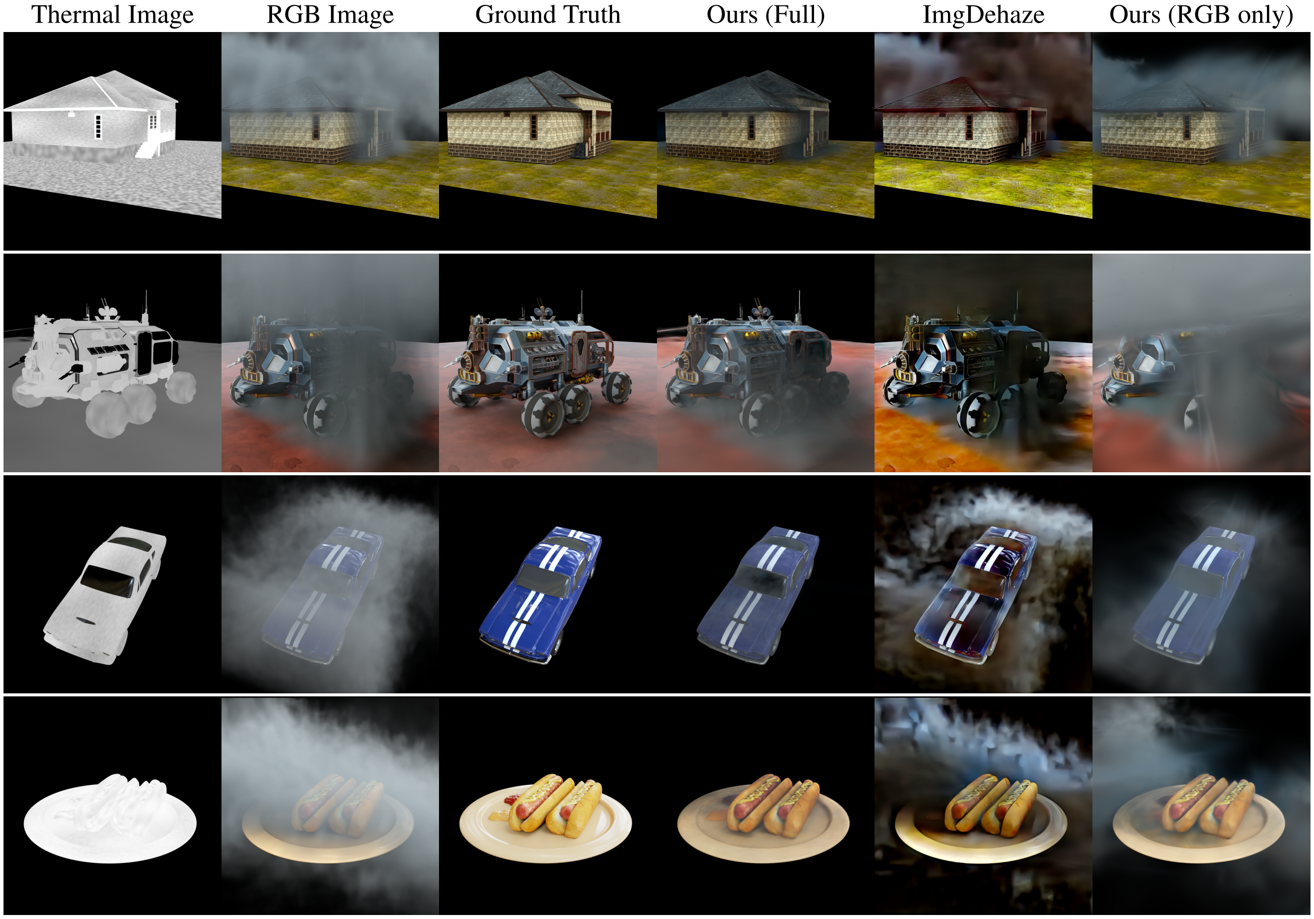

Figure 4. Qualitative results on the synthetic dataset. Our full method effectively removes smoke while preserving structural and texture details, outperforming RGB-only approaches.

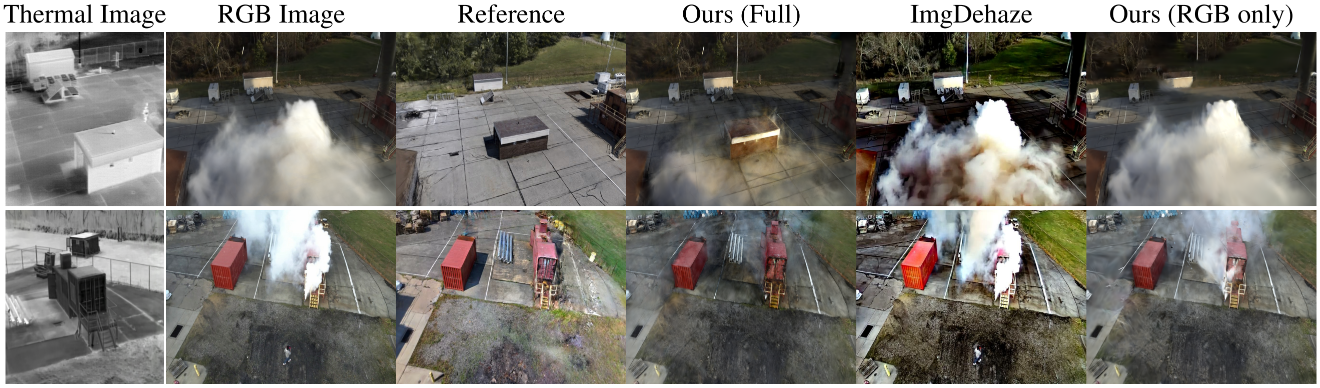

Figure 5. Qualitative results on the real-world dataset. Our method successfully removes smoke in challenging real-world conditions while preserving scene details.

ment is particularly noticeable in regions with dense smoke, where baseline methods struggle to reconstruct the scene geometry. Our method fuses complementary information in RGB and thermal modalities, to improve smoke removal while preserving texture details critical for scene understanding.

Though real-world reconstructions exhibit some artifacts, e.g., residual wisps of smoke or reduced color saturation un-

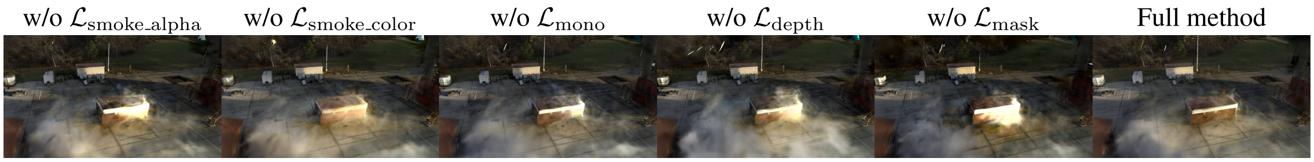
w/o $\mathcal{L}_{\text{smoke\_alpha}}$
w/o $\mathcal{L}_{\text{smoke\_color}}$
w/o $\mathcal{L}_{\text{mono}}$
w/o $\mathcal{L}_{\text{depth}}$
w/o $\mathcal{L}_{\text{mask}}$
Full method

Figure 6. Visual comparison of ablation configurations. Each image shows the result of removing a specific prior from our full method. Our method is able to recover the bricks in the wall of the house better than other configurations.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
w/o $\mathcal{L}_{\text{smoke\_alpha}}$	18.96	0.68	0.312
w/o $\mathcal{L}_{\text{smoke\_color}}$	19.02	0.71	0.289
w/o $\mathcal{L}_{\text{mono}}$	19.78	0.76	0.248
w/o $\mathcal{L}_{\text{depth}}$	19.88	0.75	0.252
w/o $\mathcal{L}_{\text{mask}}$	19.82	0.74	0.251
Ours (Full)	**19.92**	**0.76**	**0.247**

Table 2. Ablation study showing the impact of each component in our framework. Each prior contributes to the overall performance, with the depth consistency prior having the most significant impact.

der extreme occlusion (Figure 5), they represent a substantial improvement in situational awareness. For first responders, the ability to discern room layout, locate doorways, and identify obstacles even at reduced fidelity turns an unusable, smoke-obscured video stream into an actionable 3D map.

4.4. Ablation study

Table 2 shows an ablation study on individual components in our framework. Each component provides a measurable performance improvement, and the combination of all components yields the best results. The depth consistency prior ($\mathcal{L}_{\text{depth}}$) significantly improves performance, highlighting the importance of leveraging thermal information for accurate geometry reconstruction in smoke-filled environments.

Figure 6 provides a visual comparison of different ablation configurations. Without the smoke consistency priors ($\mathcal{L}_{\text{smoke\_alpha}}$ and $\mathcal{L}_{\text{smoke\_color}}$), the model struggles to separate smoke from surfaces. Without the monochromaticity prior ($\mathcal{L}_{\text{mono}}$), the model generates unrealistic colored smoke. The depth consistency prior ($\mathcal{L}_{\text{depth}}$) is important for real-world scenes where the camera poses might be noisy. The mask alignment prior ($\mathcal{L}_{\text{mask}}$) helps localize smoke and place smoke Gaussians in the correct location.

5. Conclusion

We presented SmokeSeer, a framework for joint 3D scene reconstruction and smoke removal in dynamic smoke-filled environments. Our key insight is to leverage the complementary strengths of RGB and thermal imaging to decompose the scene into its surface and smoke components. We achieved this using a 3DGS-based inverse rendering pipeline to optimize separate smoke and surface Gaussians, appropriately regularized to account for their different physical properties. We demonstrated our method on synthetic datasets and real-world environments representative of firefighting settings. Our experiments in real-world firefighting scenarios demonstrate practical viability for emergency response applications. By publicly releasing our code and dataset, we aim to establish a foundation for future research in vision through smoke and multimodal scene understanding.

Limitations and future work. Though our method achieves state-of-the-art performance in smoke removal and 3D reconstruction, several limitations remain. First, we model the temporal evolution of smoke using a deformation field, without explicitly incorporating physics-based priors such as fluid dynamics. Future work could integrate priors based on the Navier-Stokes equations to better capture smoke's physical behavior [5]. Second, our method requires careful balancing of multiple loss terms during optimization (Equation 11), and cannot handle very dense smoke that might occlude the scene completely. Future work could incorporate generative priors to provide stronger guidance at regions heavily occluded by smoke.

Acknowledgments. We thank Ian Higgins and John Keller for help with flying the drone during data acquisition; Sreekar Ranganathan for help with the camera hardware and data collection setup; and Jeff Tan and Nikhil Keetha for helpful discussions. This work was supported by National Institute of Food and Agriculture award 2023-67021-39073; Defense Science and Technology Agency contract #DST000EC124000205; and Alfred P. Sloan Research Fellowship FG202013153 for Ioannis Gkioulekas. This work used Bridges-2 at the Pittsburgh Supercomputing Center, through allocation cis220039p from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation awards 2138259, 2138286, 2138307, 2137603, and 2138296; and National Artificial Intelligence Research Resource Pilot-2211.

References

[1] E. Alonso-Blanco, A. I. Calvo, R. Fraile, and A. Castro. The influence of wildfires on aerosol size distributions in rural areas. *The Scientific World Journal*, 2012:735697, 2012. 3

[2] Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1674–1682, 2016. 2

[3] Subrahmanyan Chandrasekhar. *Radiative Transfer*. Dover Publications, New York, 1960. Unabridged and slightly revised version of the work first published in 1950. 2

[4] W. Chen, W. Yifan, S. Kuo, and G. Wetzstein. Dehazenerf: Multiple image haze removal and 3d shape reconstruction using neural radiance fields. In *3DV*, 2024. 1, 3, 4, 6

[5] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. Physics informed neural fields for smoke reconstruction with sparse data. *ACM Trans. Graph.*, 41(4), 2022. 8

[6] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Revitalizing convolutional network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[7] Bardienus Pieter Duisterhof, Lojze Žust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *ArXiv*, abs/2409.19152, 2024. 5

[8] Jeppe Revall Frisvad, Niels Jørgen Christensen, and Henrik Wann Jensen. Computing the scattering properties of participating media using lorenz-mie theory. In *ACM SIGGRAPH 2007 Papers*, page 60, New York, NY, USA, 2007. Association for Computing Machinery. 3

[9] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision and Applications*, 25 (1):245–262, 2014. 3

[10] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Image dehazing transformer with transmission-aware 3d position embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):25–39, 2022. 2

[11] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2341–2353. IEEE, 2011. 2

[12] 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. 3

[13] Xingyu Jiang, Jiangwei Ren, Zizhuo Li, Xin Zhou, Dingkang Liang, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 5

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3, 5

[15] H. Koschmieder. *Theorie der horizontalen Sichtweite*. Keim & Nemnich, 1924. 2

[16] Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. In *Proceedings of the Neural Information Processing Systems*, 2024. 5

[17] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 3

[18] Yvette Y Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. ThermalNeRF: Thermal radiance fields. In *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2024. 3

[19] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Multiscale residual learning for single image dehazing. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1366–1371. IEEE, 2019. 2

[20] Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermalgaussian: Thermal 3d gaussian splatting, 2024. 3

[21] Jonathon Luiten, Vincent Leroy, Julian Ost, Fabian Manhardt, Francis Engelmann, Deva Ramanan, and Federico Tombari. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22398–22408, 2023. 3

[22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3, 6

[23] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. 4

[24] Andrea Ramazzina, Mario Bijelic, Stefanie Walz, Alessandro Sanvito, Dominik Scheuble, and Felix Heide. Scatternerf: Seeing through fog with physically-based inverse neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17957–17968, 2023. 1, 3

[25] 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, 2024. 5

[26] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5

[27] Sebastián Salazar-Colores, Hugo M. Jiménez, César J. Ortiz-Echeverri, and Gerardo Flores. Desmoking laparoscopy surgery images using an image-to-image translation guided by an embedded dark channel. *IEEE Access*, 8:208898–208909, 2020. 2

[28] Nils Thuerey and Tobias Pfaff. MantaFlow, 2018. *http://mantaflow.com*. 6

[29] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2

[30] Wenjing Wang, Yuan Yuan, Qi Wu, Xiangyu Li, and Yanyun Zhang. Dynamic collaborative inference for dense haze removal in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 2

[31] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 6

[32] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 5, 6

[33] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 2

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[35] Xiaolong Zhao, Yingjie Jiang, Weilong Ding, Feng Huang, and Wenbing Tao. Saliency-guided image dehazing for uav imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 2

[36] Qingsong Zhu, Jiaming Mai, and Ling Shao. Fast single image haze removal algorithm using color attenuation prior. *IEEE Transactions on Image Processing*, 24(11):3522–3533, 2015. 2