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

Neham Jain

CMU-RI-TR-25-34
April 11

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Professor Ioannis Gkioulekas, *chair*
Professor Sebastian Scherer
Bailey Miller

*Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.*

Abstract

The presence of smoke in real-world scenes can severely degrade the quality of images and hamper visibility. Recently introduced methods for image restoration either rely on data-driven priors that are susceptible to hallucination, or are limited to static low-density smoke. We introduce See-Through Smoke, a method to perform simultaneous 3D scene reconstruction and smoke removal from a video capturing multiple views of a scene. To achieve this task, our method uses thermal and RGB images, leveraging the fact that the reduced scattering in thermal images enables us to see through the smoke. We build upon 3D Gaussian splatting to fuse information from the two image modalities, and decompose the scene explicitly into smoke and non-smoke components. Unlike prior approaches, See-Through Smoke handles a broad range of smoke densities and can adapt to temporally varying smoke. We validate our approach on synthetic data and introduce a new real-world multi-view smoke dataset with RGB and thermal images.

Acknowledgments

My journey at CMU over the past two years has been extraordinary, filled with significant professional growth and personal milestones. I am deeply grateful to have crossed paths with so many exceptional individuals who made this experience truly meaningful.

First and foremost, I owe immense gratitude to my advisor, Ioannis Gkioulekas, whose technical proficiency, support, and insightful mentorship have significantly shaped my research and professional trajectory. Ioannis is not only an exceptional researcher but also a kind, hardworking, and inspiring advisor. He is also an amazing instructor, his course on Physics-based Rendering was the most interesting class I took at CMU. Thank you, Ioannis, for your invaluable guidance.

I would like to thank Sebastian Scherer for his guidance on high-level project direction and systems-related challenges. I am also grateful to Bailey Miller for the many insightful discussions that helped refine my research. Special thanks to Andrew Jong, the lead of the wildfire project, for ensuring I had the necessary resources for data collection and for his tremendous support throughout my thesis. I would like to thank Katia, Srujan, Devansh, Yifei, Ian, John, Steve, the firefighting team at the Allegheny County Fire Academy, and other members of the wildfire team for their contributions during the project. I also thank Krishna Mullia for hosting my internship at Adobe Research.

I am thankful to the members of the Imaging Group, especially Arjun, Aswin, Benran, Sreekar, Tanli, Bakari, Bailey, Mani, Dorian, Sagnik, Aswin and others for their engaging discussions and valuable insights. Smith Hall was always buzzing with energy thanks to the people who filled it. From research discussions to casual conversations, Yanbo, Jeff, Alex, Aniket, Anish, Jay, Nikhil, Jialu, Bharath, Grace, Sheng-Yu, Nupur, Qitao, Sally, Himangi and many others provided perspectives and welcome diversions from research.

Adjusting to life in the U.S. was initially challenging, but I was fortunate to find a wonderful circle of friends who quickly became like family. I am especially grateful to my amazing roommate, Srujan, for being my partner in crime through courses, research and life. I would also like to thank Aadesh, Aryan, Dev, Dhruv, Khush, Mehal, Parth, Peya, Rupali,

Sagnik, Sanjali, Sashank, Shashwat, Shivangi, Tirtha, Vihaan, and Yash for making Pittsburgh feel like home and for the unforgettable memories. I would like to thank Lakshya for her unwavering support, this thesis would not have been possible without her.

I am deeply thankful to my cousins, grandparents, and extended family for cheering me on from afar. I owe a special thanks to my brother for his guidance and support throughout my life. I am extremely grateful to my Mom and Dad for their unconditional love, encouragement, and trust in my aspirations. My parents always encouraged me to follow my passion and find my own path, and for that I am truly thankful to them.

Finally, I dedicate this thesis to my late grandmother, Kanta Pokharna, whose memory continues to inspire me every single day.

Funding

This work was supported by National Institute of Food and Agriculture award 2023-67021-39073.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

Reliable visual perception in adverse, smoke-filled environments is a fundamental challenge for robotics and safety-critical applications. In scenarios such as search and rescue, firefighting, and industrial inspection, robots and human operators must navigate and make decisions in conditions where visibility is severely compromised by dense, dynamic smoke. The ability to accurately perceive and reconstruct 3D environments under these conditions is essential for enabling reliable robot localization, robust object detection, and safe path planning, all of which are critical for deploying autonomous systems in real-world, high-risk settings.

For example, firefighters navigating through burning buildings increasingly depend on vision-based systems for maintaining situational awareness. However, dense smoke severely compromises these systems, obscuring vital environmental details and substantially 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.

The core difficulty arises from the physical properties of smoke: unlike thin haze or fog, dense smoke is highly dynamic, spatially heterogeneous, and causes rapid, severe reductions in visibility. Light is both scattered and absorbed and this process is time-varying, making it extremely challenging to recover the true structure and appearance of the underlying scene.

Dense smoke, unlike thin haze or fog, exhibits rapid temporal variations and severe visibility reduction, significantly complicating both visual perception and 3D reconstruction tasks. Traditional visibility enhancement methods often assume static, uniformly distributed haze, conditions which are not met in realistic smoke and fire scenarios. Therefore, an effective system must explicitly model dynamic smoke behavior and account for its spatial variations.

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 [26] and DehazeNeRF [5] incorporate physical light transport models and operate on multi-view RGB data. However, all of the above approaches primarily address static haze removal and are ill-equipped to handle dense, temporally evolving smoke.

We build an end-to-end system that does joint 3D scene reconstruction and smoke removal in the presence of dense, temporally evolving smoke. Our method utilizes images from co-located RGB and thermal cameras and is effective on real-world smoke data. 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 set of surface Gaussians. This decomposition allows us to render only the set of 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 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 recent advances in 3D foundation models 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, our method does not directly rely on any data-driven priors but 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 demonstrate 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.

Solving the problem of robust perception and 3D reconstruction in smoke-filled environments has far-reaching implications. For first responders and firefighters, it can dramatically improve situational awareness, reduce operational risks, and ultimately save lives. For robotics, it enables reliable autonomy in environments previously considered too hazardous for machine perception. The techniques developed here could be extended to other challenging visual conditions, such as fog, dust, or underwater environments, and could be combined with additional sensing modalities like LiDAR or radar for even greater robustness.

Chapter 2

Related Work

2.1 Image-based methods for haze removal

Traditional methods. Koschmieder [16] developed a simplified atmospheric scattering model that describes image formation under haze as a combination of direct attenuation and airlight. This model represents a simplification of the more general radiative transfer equation (RTE) [4], 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 hand-crafted priors to estimate physical parameters in the Koschmieder model. He et al. [11] tries to estimate the attenuation map by leveraging the observation that at least one color channel in all images has very low intensity in haze-free regions. 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 these methods are effective for thin homogeneous haze, they fail in dense smoke scenarios that their priors are not representative of.

Learning-based methods. Some recent methods directly map hazy images to clear images without explicit parameter estimation. Examples include MSRL-

DehazeNet [20] with multi-scale residual learning, collaborative inference frameworks for dense haze in remote sensing [32], and saliency-guided mechanisms for UAV imagery [35].

Transformer-based architectures have recently shown promising results for dehazing. Li et al. [17] proposed a transformer-based dehazing network that captures long-range dependencies. 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, Fujita et al. [9] developed a GAN-based approach for removing smoke from endoscopic 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) [24] have revolutionized scene representation using continuous volumetric functions. Several works have extended NeRF to handle participating media such as smoke and haze. ScatterNeRF [26] incorporates the Koschmieder model into the NeRF framework, but remains limited to homogeneous haze conditions. DehazeNeRF [5] 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] offers an efficient alternative to NeRF through scene representation using 3D Gaussians, enabling real-time rendering. Dynamic 3DGS [23] extends this framework to handle dynamic scenes, but does not specifically address participating media.

Lastly, recent approaches such as ThermalNeRF [18] and ThermalGaussian splatting [22] 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 [1]. Hwang et al. [12] demonstrated the effectiveness of fusing RGB and thermal information for object detection in adverse weather. Chen et al. [6] proposed a multi-modal framework combining RGB, thermal, and LiDAR data for all-weather autonomous driving.

For 3D reconstruction, Khattak et al. [15] used thermal-inertial odometry for robust localization in smoke-filled environments. Bijelic et al. [3] explored the fusion of multiple sensor modalities for perception through fog and smoke. However, these approaches primarily focus on detection and localization rather than 3D reconstruction and rendering.

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.

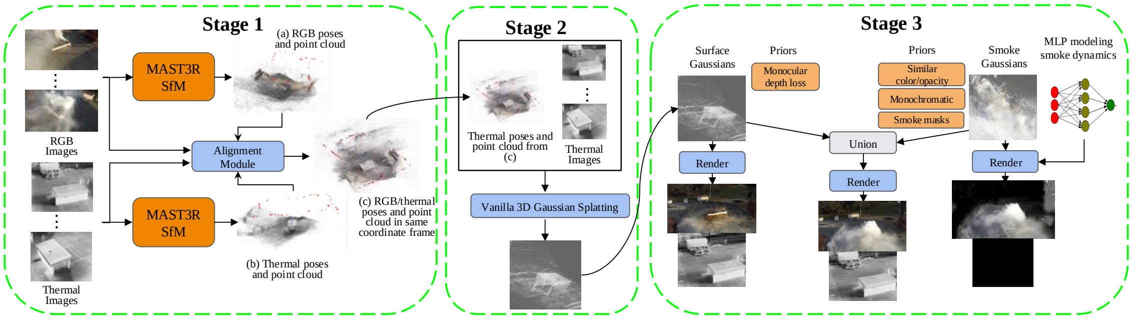


Figure 3.1: An overview of our method.

Chapter 3

Method

Our methodological choices are motivated by the limitations of existing solutions. Using Gaussian splatting for our core representation provides computational efficiency and real-time rendering capability, critical for operational deployments. Also, since 3DGS is a point based representation, it is easier to impose priors and is more interpretable compared to NeRF based representations which rely on neural networks to represent the scene.

The joint use of RGB and thermal imaging leverages complementary sensing capabilities—thermal data robustly capture spatial layout despite smoke, while RGB data provide rich texture necessary for detailed scene understanding. The problem that we are trying to solve is extremely ill-posed. Therefore, incorporating physically motivated priors and depth constraints helps maintain realistic smoke modeling, preventing the model from producing physically implausible results. Our method, depicted in Fig. 3.1, comprises three primary stages:

(1) Camera pose estimation and smoke segmentation.

(2) Initial surface reconstruction from thermal images.

(3) Joint optimization of surface and smoke plume using both RGB and thermal images.

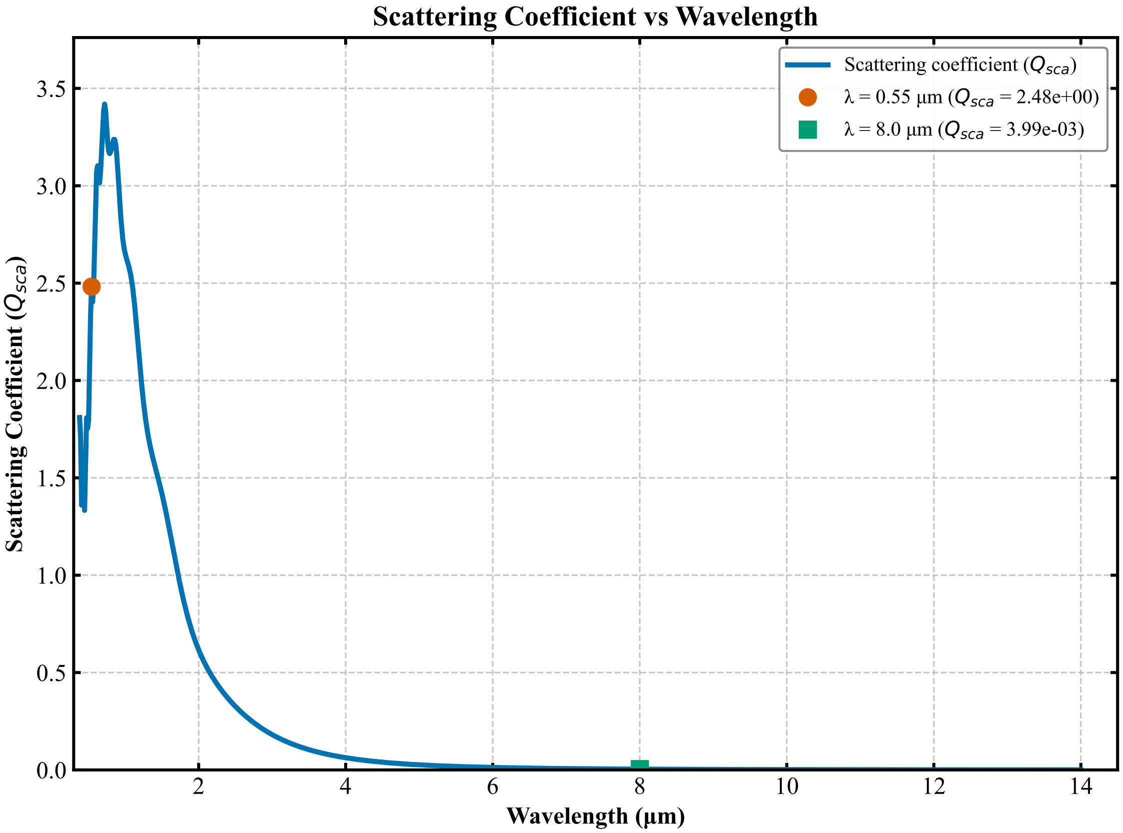


Figure 3.2: Scattering coefficient as a function of wavelength (in μm), illustrating the difference in scattering behavior between visible and infrared spectra.

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 Fig. 3.2. 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 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.

3.2 Background on 3D Gaussian splatting

Recently, 3D Gaussian splatting [14] (3DGS) has emerged as a powerful scene representation that enables faster training and real-time rendering by leveraging rasterization techniques. 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 containing a set of 3D Gaussians $\mathcal{G} = \{g_i\}$. 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 $\boldsymbol{h}_i$ [25]. 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 we provide details about the two 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:

$$T_\sigma(t) = \exp\left(-\int_{t_n}^{t} \sigma(s)\, ds\right), \tag{3.1}$$

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 $\sigma_\mathrm{s}(t)$ for smoke. The combined transmittance is:

$$\begin{aligned} T_{\sigma+\sigma_\mathrm{s}}(t) &= \exp\left(-\int_{t_n}^{t} [\sigma(s) + \sigma_\mathrm{s}(s)], ds\right) && (3.2) \\ &= T_\sigma(t) \cdot T_{\sigma_\mathrm{s}}(t), && (3.3) \end{aligned}$$

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

Chen et al. [5] have shown that the volume rendering equation for a scene with mixed density takes the form:

$$\begin{aligned} C(\mathbf{r}, \mathbf{d}) = &\underbrace{\int_{t_n}^{t_0} c(\mathbf{r}(t), \mathbf{d}) \sigma(t) T_{\sigma+\sigma_\mathrm{s}}(t)\, dt}_{C_\text{surface}} \\ &+ \underbrace{\int_{t_n}^{t_0} c_\mathrm{s}(\mathbf{r}(t)) \sigma_\mathrm{s}(t) T_{\sigma+\sigma_\mathrm{s}}(t)\, dt}_{C_\text{smoke}}. \end{aligned} \tag{3.4}$$

Our dual Gaussian representation directly maps to this equation, where the surface Gaussians $\mathcal{G}$ correspond to the term C_surface with color c and opacity σ, whereas the smoke Gaussians $\mathcal{S}$ correspond to the term C_smoke with color c_s and opacity σ_s. Rendering the union of these two sets $\mathcal{G} \cup \mathcal{S}$ is equivalent to computing the rendering equation equation 3.4.

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

$$C_\text{clear}(\mathbf{r}, \mathbf{d}) = \int_{t_n}^{t_0} c(\mathbf{r}(t), \mathbf{d}) \sigma(t) T_\sigma(t)\, dt. \tag{3.5}$$

Rendering only the surface Gaussians $\mathcal{G}$ is equivalent to computing the rendering equation equation 3.5. By modeling surface and smoke separately, we achieve effective smoke removal through selectively rendering only surface Gaussians.

3.4 Modality-specific representations

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, where $I_k^{\text{RGB}} \in \mathbb{R}^{H \times W \times 3}$ and $I_k^{\text{T}} \in \mathbb{R}^{H' \times W'}$ are captured from the same scene. Each image is associated with a camera pose P_k^{RGB} or P_k^{T} in a common coordinate system.

To account for the different modalities, we define modality-specific appearance attributes for each Gaussian:

- Each surface Gaussians $g_i \in \mathcal{G}$ is characterized by a center position $\boldsymbol{\mu}_i$, a covariance matrix $\boldsymbol{\Omega}_i$, and modality-specific appearance. RGB appearance is encoded using spherical harmonic coefficients $\boldsymbol{h}_i^{\text{RGB}}$, while thermal appearance uses $\boldsymbol{h}_i^{\text{T}}$. We use the same opacity for RGB and thermal modalities. We assume that surfaces are static, thus surface Gaussian parameters are time-invariant.
- Each smoke Gaussian $s_i \in \mathcal{S}$ has position $\boldsymbol{\mu}_i^s$, covariance $\boldsymbol{\Omega}_i^s$, and modality-dependent opacities α_i^{RGB} and α_i^{T}, where $\alpha_i^{\text{T}} \ll \alpha_i^{\text{RGB}}$ due to the properties of smoke described in Section **??**. These properties are time-varying due to the dynamic nature of smoke.

With this notation established, we now detail our three-stage pipeline for reconstructing smoke-obscured scenes.

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 the same coordinate system. 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 impedes reliable feature extraction.

We address these challenges with a three-step approach:

1. *Smoke segmentation:* We use GroundedSAM [28], based on SAMv2 [27], to identify and mask out smoke-affected regions in RGB images. Doing 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] to establish 2D correspondences between RGB-thermal image pairs, then lift these correspondences to 3D using the 2D-3D mappings that we obtain from the per-modality calibration. Doing so enables the estimation of a similarity transform $T \in \mathrm{Sim}(3)$ 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 Gaussian evolve temporally and exhibit modality-dependent opacity, reflecting smoke's varying visibility in RGB versus thermal images (Section **??**). Though in principle we could use Mie theory to model the relationship between opacities, we opt for a more flexible approach with two independent free variables for smoke visibility in each modality.

3.8 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 [34]. 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 multi-layer perceptron (MLP)that takes as input the positions of the 3D Gaussians and a timestep t, and outputs offsets δ_x, δ_y, and δ_z for each Gaussian's position. These offsets transform the canonical 3D Gaussians to the deformed space at each time. We use a bimodal Gaussian distribution to model smoke opacity as a function of time.

3.9 Priors on properties of smoke Gaussians

To facilitate accurate surface-smoke separation and modelling 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:

$$L_{\text{smoke}} = \text{Var}(\{\alpha_i\}_{i \in \mathcal{S}}) + \text{Var}(\{c_i\}_{i \in \mathcal{S}}). \tag{3.6}$$

  This prior is based on 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. Even 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:

  $$L_{\text{mono}} = \sum_{i \in \mathcal{S}} \text{Var}(c_i^{\text{R}}, c_i^{\text{G}}, c_i^{\text{B}}). \tag{3.7}$$

  This prior reflects the physical property that smoke typically appears as a neutral gray color. Unlike colored gases that selectively absorb certain wavelengths, smoke from common sources (e.g., burning organic matter) scatters visible wavelengths roughly equally, resulting in monochromatic appearance. This prior prevents our model from generating implausible colored smoke.

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

  $$L_{\text{depth}} = \|d_i - \hat{d}_i\|_2, \tag{3.8}$$

  where d_i denotes predicted depth on a thermal image using a monocular depth estimation model [33] 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 **??**). 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:

  $$L_{\text{mask}} = \|M_{\text{pred}} - M_{\text{GT}}\|_1, \tag{3.9}$$

  where M_{pred} and M_{GT} denote rendered smoke alpha values 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 only in regions where smoke is present, and prevent them from appearing in smoke-free areas.

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

$$\begin{aligned} L_{\text{total}} = {} & \lambda_{\text{render}} L_{\text{render}} + \lambda_{\text{smoke}} L_{\text{smoke}} + \lambda_{\text{mono}} L_{\text{mono}} \\ & + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{mask}} L_{\text{mask}}. \end{aligned} \tag{3.10}$$

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

Chapter 4

Experimental Evaluation

We conduct extensive quantitative experiments to rigorously evaluate our framework on both synthetic and real-world datasets. Our evaluation protocol includes direct comparison against state-of-the-art methods using standard quantitative metrics, as well as ablation studies to isolate the impact of individual design choices. Specifically, we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) for all methods. PSNR (higher is better) measures pixel-level reconstruction fidelity, SSIM (higher is better) quantifies structural similarity to ground truth, and LPIPS (lower is better) assesses perceptual similarity as judged by deep neural networks. These metrics are computed over all test images in both synthetic and real-world datasets, enabling a comprehensive and objective assessment of image quality and restoration performance.

While our primary focus is on these quantitative metrics, we note that further quantitative evaluation could be performed on downstream tasks relevant to robotics and scene understanding. For example, desmoking and 3D reconstruction can be quantitatively assessed by measuring improvements in:

- **Object Detection Accuracy:** Evaluated by measuring improvements in detecting people or objects in desmoked images.
- **Segmentation Quality:** Assessed by the quality of semantic and instance segmentation results on desmoked scenes.
- **SLAM Performance:** Quantified by the accuracy of pose estimation pipelines

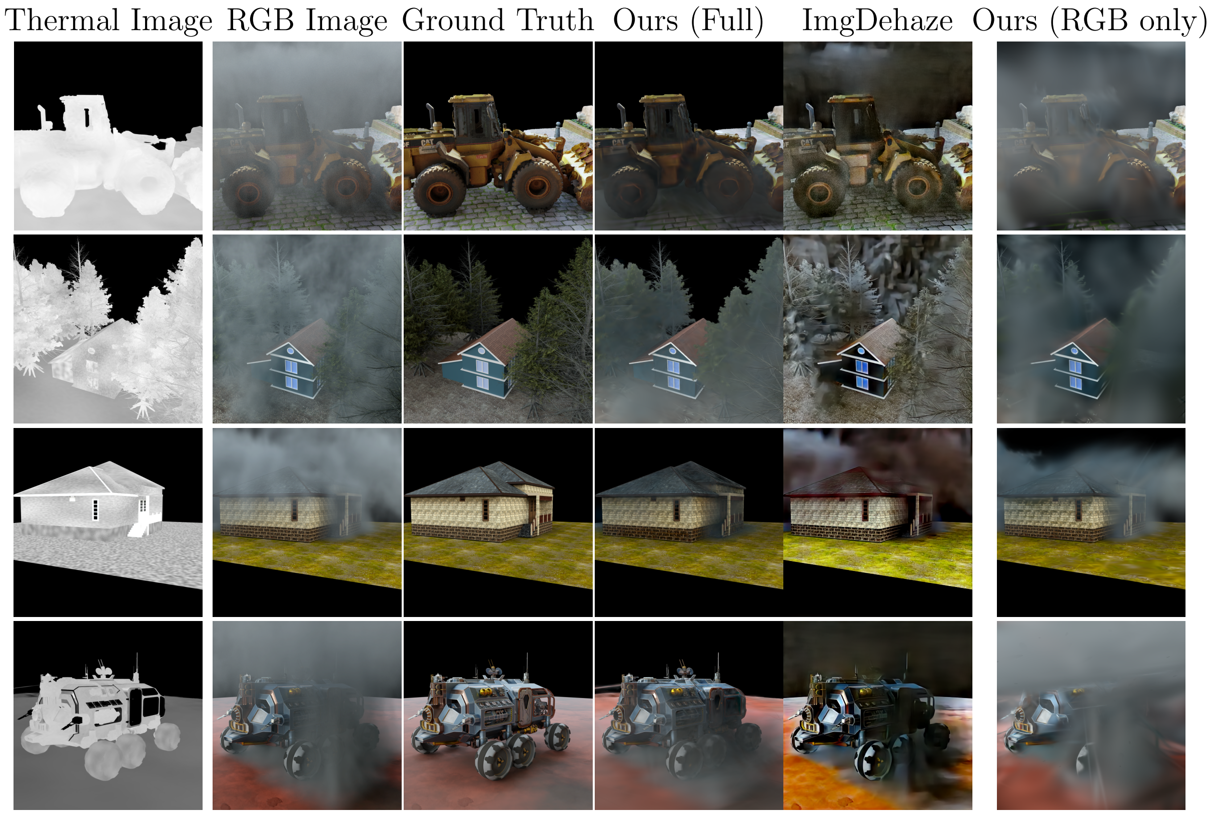


Figure 4.1: Qualitative results on the synthetic dataset. From left to right: Thermal Image, RGB Image, Ground Truth, Ours (Full), ImgDehaze, and Ours (RGB only). Our method effectively removes smoke while maintaining the structural integrity and texture details of the scene.

when using desmoked images.

- **3D Reconstruction Fidelity:** Determined by how closely the reconstructed 3D models match the ground-truth in terms of completeness and geometric accuracy. We could use metrics such as Chamfer Distance to evaluate the quality of the reconstructed 3D models.

Incorporating these quantitative downstream metrics in future work would provide a more application-driven evaluation of our method's effectiveness in real-world scenarios.

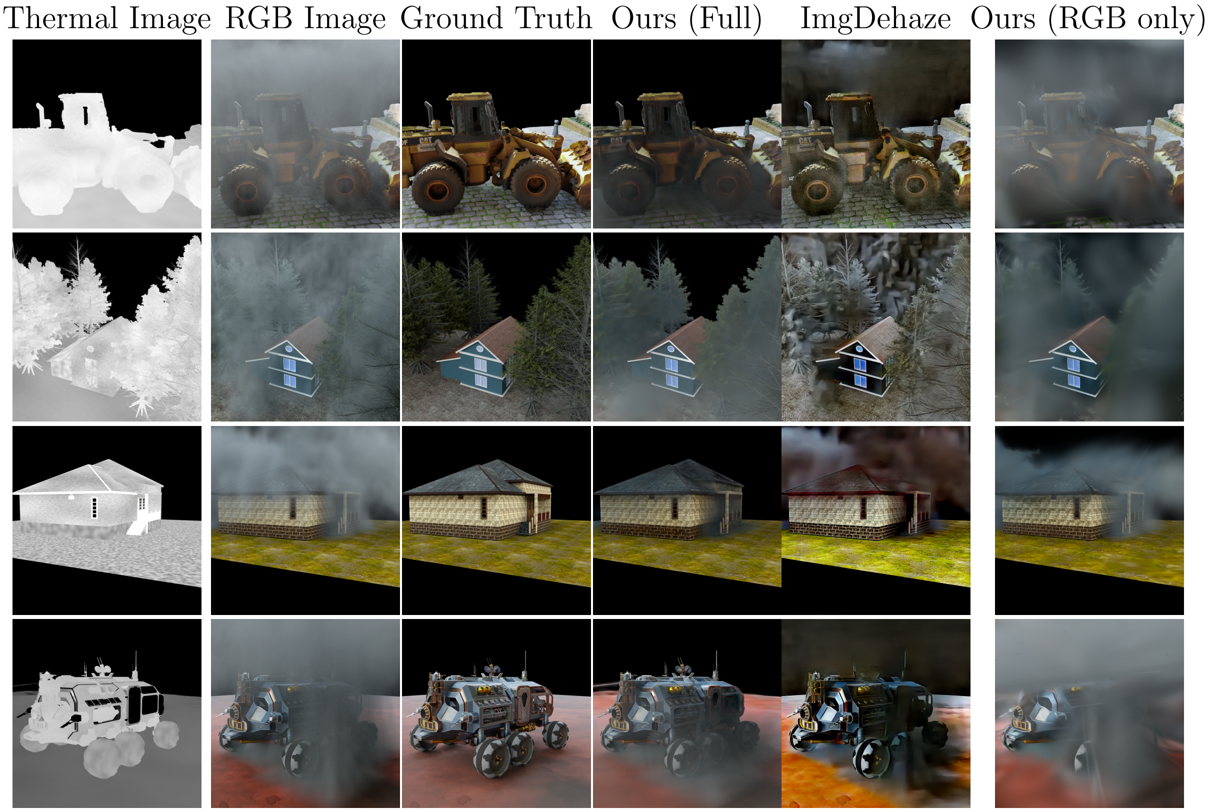


Figure 4.2: Qualitative results on the real dataset. From left to right: Thermal Image, RGB Image, Ground Truth, Ours (Full), ImgDehaze, and Ours (RGB only). Our method effectively removes smoke while maintaining the structural integrity and texture details of the scene.

4.1 Datasets

4.2 Synthetic Dataset

To enable quantitative evaluation with ground truth, we create a synthetic dataset using Blender's Mantaflow [30] based smoke simulator. This dataset consists of 10 scenes: 5 object-level scenes (chair, ficus, hotdog, lego, and racecar from the NeRF synthetic dataset [24]) and 5 large-scale indoor scenes (kitchen, living room, office, bedroom, and bathroom).

For each scene, we generate 150 frames of both RGB and thermal images with dynamic smoke evolution, capturing the temporal characteristics of real smoke.

4.3 Real-World Dataset

In collaboration with our county's fire department, we collect a real-world dataset using a Spirit drone equipped with co-located RGB and thermal cameras. The payload on the drone does not provide synchronized timestamps between the RGB and thermal cameras which makes pose estimation challenging. The RGB camera captures 4K resolution images, while the thermal camera (FLIR Boson) captures lower-resolution (640×512) thermal imagery. We collect data in 2 different environments. For each environment, we use controlled smoke bombs to generate dense smoke which is representative of a fire situation faced by firefighters. The drone follows pre-planned trajectories to capture multi-view data of each scene. The real-world dataset presents several challenges not present in the synthetic data due to the imperfect alignment between RGB and thermal cameras, the random motion of smoke due to wind and motion blur due to drone movement. These challenges make our real-world dataset a rigorous test for smoke removal and 3D reconstruction algorithms.

Chapter 5

Implementation Details

For training, we use the Adam optimizer with a learning rate of 1×10^{-3} for position and opacity parameters, and 1×10^{-4} for rotation and scaling parameters. The loss weights are set to $\lambda_{\text{smoke}} = 0.1$, $\lambda_{\text{mono}} = 0.05$, $\lambda_{\text{depth}} = 0.5$, and $\lambda_{\text{mask}} = 0.2$. We train our models on an NVIDIA RTX 4090 GPU.

Our method takes around 10 minutes for Stage 2 and 30 minutes for Stage 3 for a typical scene. For a typical scene, our model uses approximately 500K surface Gaussians and 300K smoke Gaussians. The final trained model enables real-time rendering at over 60 FPS on an RTX 4090 GPU.

5.1 Baseline Methods

We compare our approach against the following baselines:

- **ImgDehaze + 3DGS**: A two-stage approach where we first apply a state-of-the-art single-image dehazing method (ConvIR []) to each RGB frame, then train a standard 3DGS model 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.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Vanilla 3DGS	23.6	0.85	0.15
ImgDehaze + 3DGS	24.2	0.86	0.14
VidDehaze + 3DGS	24.5	0.87	0.13
ScatterNeRF	24.1	0.86	0.14
DehazeNeRF	24.8	0.88	0.12
Ours (RGB only)	25.4	0.89	0.12
Ours (Full)	**25.8**	**0.90**	**0.11**

Table 5.1: Quantitative results averaged over the synthetic dataset for novel view synthesis. Our full method consistently outperforms all baselines across all metrics.

5.2 Results on Synthetic Data

Table 5.1 presents the quantitative results for novel view synthesis on our synthetic dataset. Our full method consistently outperforms all baselines across all metrics. The improvement is particularly significant in scenes with heavy smoke, where our method achieves a PSNR gain of up to 3.2dB over Vanilla 3DGS.

Figure 4.1 shows qualitative results on the synthetic dataset. Our method successfully removes smoke while preserving fine details in the scene. In contrast, baseline methods either fail to completely remove smoke or introduce artifacts in the reconstructed scene.

5.3 Results on Real Data

5.3.1 Ablation Studies

We conduct several ablation studies to validate our design choices and analyze the contribution of each component in our framework.

5.3.2 Contribution of Individual Components

Table 5.2 presents an ablation study on the contribution of individual components in our framework. Each component provides a measurable improvement in performance,

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
Base (RGB only, no priors)	24.1	0.86	0.14
+ Smoke Consistency Prior	24.6	0.87	0.13
+ Monochromaticity Prior	24.9	0.88	0.13
+ Mask Alignment Prior	25.4	0.89	0.12
+ Thermal (no Depth Consistency)	25.5	0.89	0.12
+ Depth Consistency Prior (Full)	**25.8**	**0.90**	**0.11**

Table 5.2: Ablation study showing the impact of each component in our framework. Starting from a baseline using only RGB images without any priors, we progressively add each component. Each addition improves performance, with the full model achieving the best results. The smoke consistency and monochromaticity priors help in better smoke separation, while thermal information and depth consistency enable more accurate geometry reconstruction.

with the combination of all components yielding the best results. The smoke consistency and monochromaticity priors contribute significantly to the quality of smoke removal, while the thermal information and depth consistency prior are crucial for accurate geometry reconstruction.

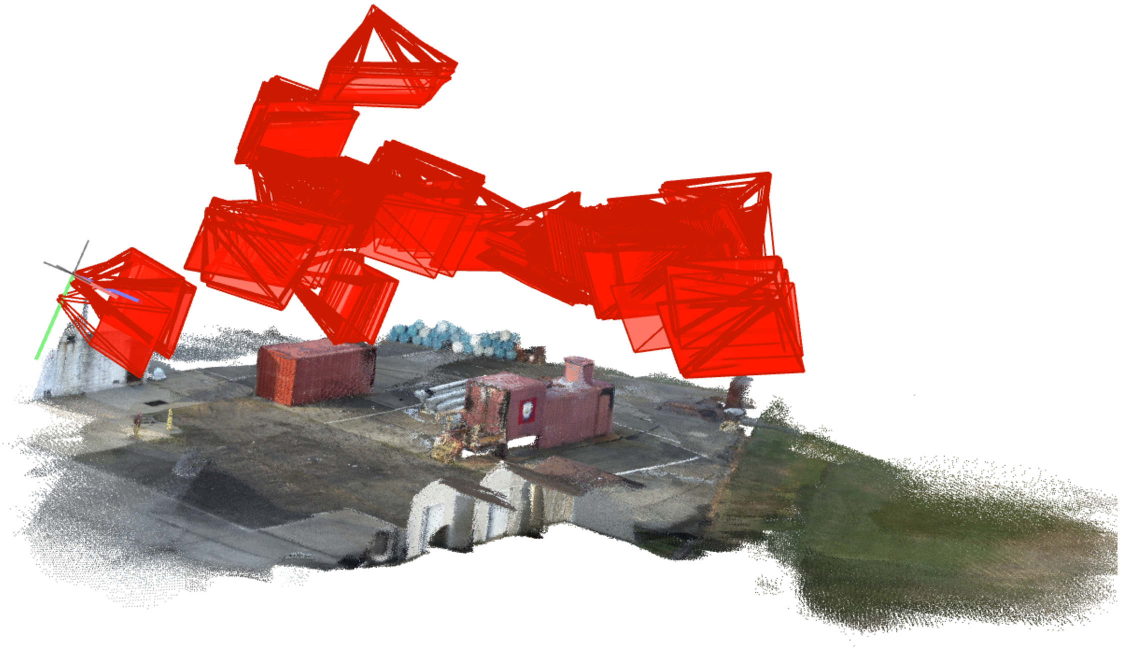


Figure 5.1: Qualitative results of VGGT [31] on the red container scene from our real-world dataset. VGGT identifies smoke as low-confidence points and enables effective removal of smoke through confidence-based filtering, resulting in a clean reconstruction of the underlying scene geometry.

Chapter 6

Conclusions

6.1 Limitations and Future Work

While our method achieves state-of-the-art results in smoke removal and 3D reconstruction, several limitations remain that present opportunities for future work. One promising direction is the adoption of recent feed-forward transformer-based architectures, such as VGGT [31], for both smoke removal and 3D reconstruction. These models can infer multiple 3D scene attributes in a single forward pass by leveraging alternating frame-wise and global attention mechanisms, thus eliminating the need for iterative optimization and enabling scene processing in under 10 seconds. In our preliminary experiments, we applied VGGT to smoke-filled RGB images from our real-world dataset which we had collected in collaboration with the fire department. VGGT was able to identify smoke as low-confidence points due to its dynamic nature, and subsequent filtering of these points by confidence effectively removed the smoke. Qualitative results on the red container scene are shown in Figure 5.1. Unlike our optimization-based approach, a VGGT-style model could jointly process RGB and thermal data to directly reconstruct smoke-free scenes, potentially improving computational efficiency and temporal consistency in dynamic smoke scenarios. This approach aligns with the broader trend of unifying multiple 3D tasks within a single model architecture trained on diverse datasets.

Currently, we model the temporal evolution of smoke using a deformation field, but do not explicitly incorporate physics-based priors such as the laws of fluid dynamics.

Integrating priors based on the Navier-Stokes equations could enable more accurate modeling of smoke's physical behavior in future work. Another limitation is the need to carefully balance multiple loss terms during optimization. Incorporating diffusion model priors along with feed-forward models could provide stronger guidance for reconstructing regions that are heavily occluded by smoke, which would be especially valuable in extreme cases where both RGB and thermal information are unreliable.

6.2 Summary

We introduced See-Through Smoke, a novel framework for simultaneous 3D reconstruction and smoke removal in dynamic, smoke-filled environments using both RGB and thermal imagery. Our method integrates physically motivated priors, efficient 3D Gaussian splatting, and multimodal data fusion to address the severe visibility challenges posed by dense, evolving smoke.

The main technical contributions of this work are as follows:

- A 3D Gaussian Splatting (GS) framework capable of decomposing a scene into two sets of Gaussians: one modeling the underlying surfaces and the other representing the smoke. Leveraging 3D-GS enables efficient 3D scene reconstruction and real-time rendering.
- A joint optimization strategy that incorporates thermal images, significantly enhancing smoke removal performance. Since thermal wavelengths are largely unaffected by smoke particles, they preserve surface clarity and provide reliable spatial cues for reconstruction.
- Robustness to varying smoke densities and the ability to handle temporally evolving smoke, making the method applicable to a wide range of real-world scenarios.

Our experiments on real-world firefighting datasets validate the practical effectiveness of our approach for emergency response applications. By releasing our code and dataset, we aim to support further research in vision through scattering media and multimodal scene understanding.

Future directions include exploring alternative sensor modalities, such as LiDAR or radar, to further improve robustness in extremely dense smoke. Additionally, evaluating the system in realistic firefighting or search-and-rescue simulations, with input from emergency personnel, could yield valuable insights and help tailor the system to operational needs.

In summary, See-Through Smoke offers a promising step toward enhanced perception in smoke-obscured environments, with the potential to improve both robotic autonomy and human safety. We anticipate that this work will contribute to the development of comprehensive multimodal perception systems capable of reliable operation under severe visibility constraints.

Appendix A

SmokeSeer Supplementary

This document provides additional visualizations and implementation details that complement the main paper. We provide qualitative results of our method as videos on our synthetic and real-world datasets inside an html file in the supplementary material.

A.1 Smoke Segmentation Results

Grounded-SAM [28] integrates Grounding DINO [19], an open-set object detector, with the Segment Anything Model (SAMv2) [27], to facilitate text-driven object detection and segmentation. We utilized this framework to automatically generate segmentation masks for smoke by inputting the prompt ”smoke.” These masks are crucial for reliable feature matching in Stage 1 of our pipeline and also serve as supervision for the mask alignment loss in Stage 3. Figure A.1 shows examples of the smoke segmentation masks generated in Stage 1 of our pipeline.

A.2 Feature Matching Comparison

Figure A.2 compares traditional SIFT matching [21] with MAST3R-SfM [7] for feature matching in smoke-affected scenes. The comparison highlights how SIFT matching fails for thermal images which have low texture, while MAST3R-SfM, combined with

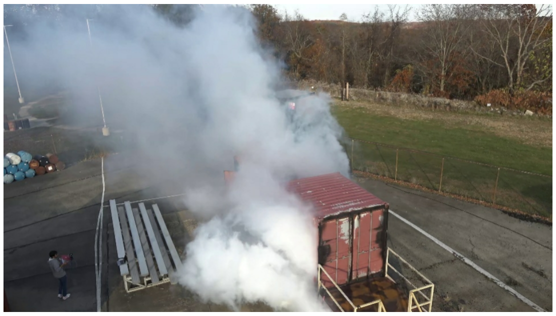
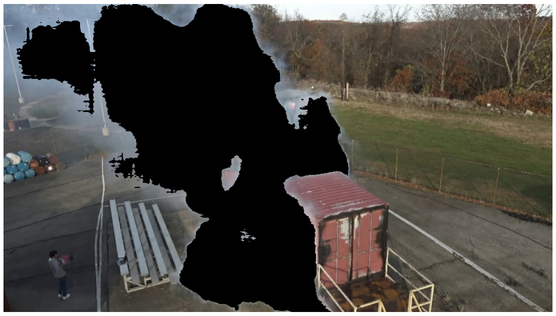
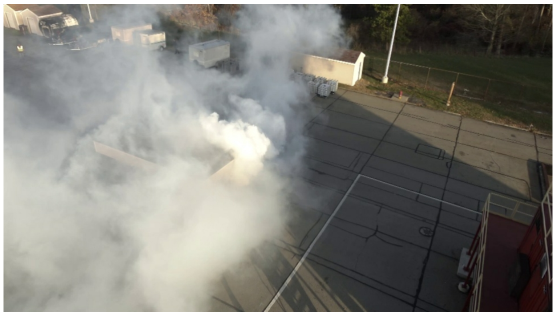
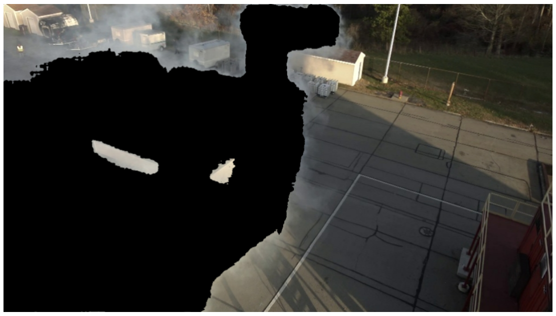
Input RGB	RGB with Smoke Mask
	
	

Table A.1: Output of smoke segmentation pipeline. From left to right: Input RGB images with smoke, generated smoke masks. Note how the segmentation model accurately identifies smoke regions even with varying density and illumination.

smoke masking, provides more reliable correspondences. This robust matching is essential for the accurate camera pose estimation required in Stage 1 of our pipeline.

A.3 Cross-Modal Registration Details

Figure A.3 visualizes the cross-modal registration process using MINIMA [13] for aligning RGB and thermal coordinate systems. We find that running COLMAP [29] on our data fails to register the images and gives a degenerate result. Running MAST3R-SfM [7] does give better results than COLMAP and is able to register all images. However, it is still not perfect and there is some misalignment. The visualization shows the 2D correspondences established between RGB and thermal image pairs and the resulting aligned point clouds. This cross-modal registration is crucial for our approach, as it allows us to get the RGB and thermal images in the same coordinate system which is necessary for running the subsequent 3D

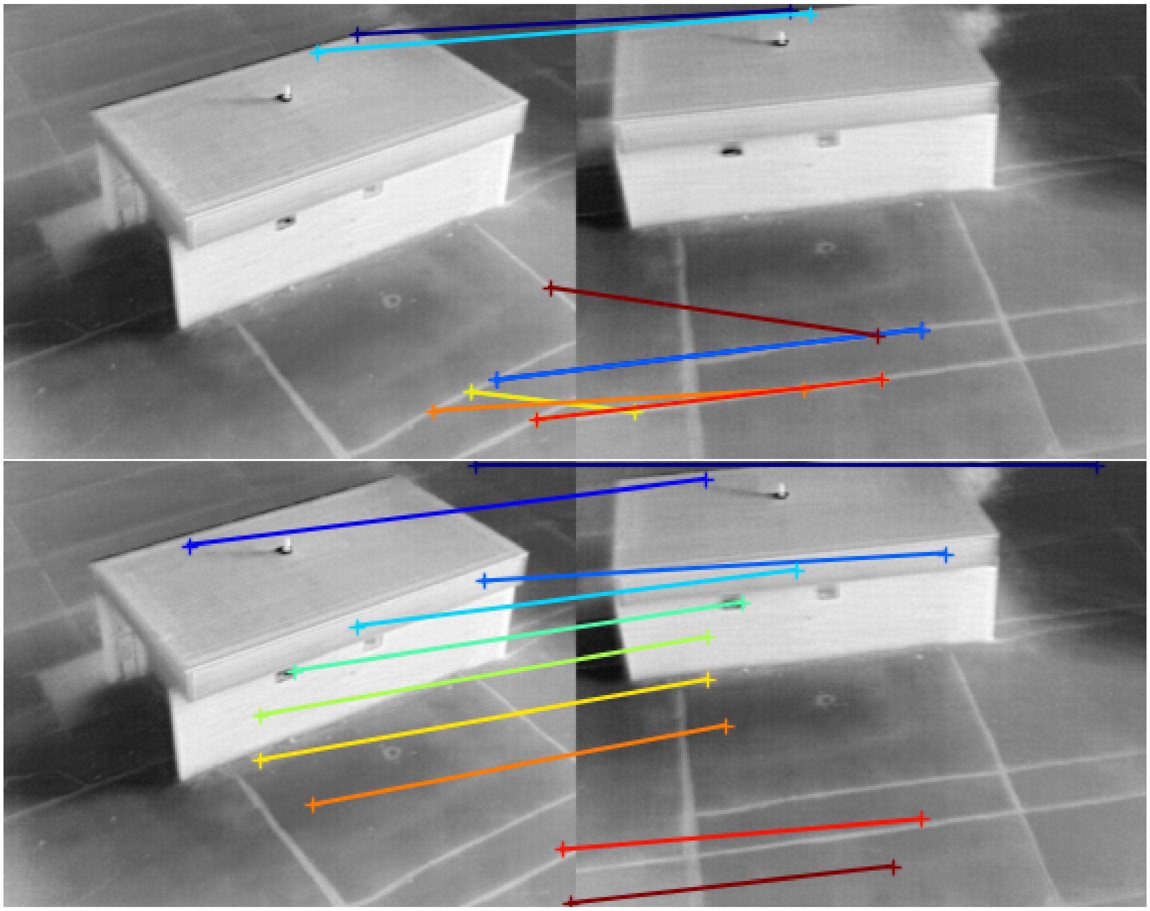


Table A.2: Feature matching comparison between images of the same modality (thermal). Top: Traditional SIFT feature matching fails in thermal images due to low texture and contrast in thermal images. Bottom: MAST3R-SfM provides more reliable correspondences by leveraging learning-based features that are more robust to the challenges of thermal imagery.

reconstruction pipeline.

A.4 Implementation Details

This section provides additional technical details about our implementation to facilitate reproducibility. All experiments were conducted on a workstation with an NVIDIA RTX 4090 GPU, 128GB RAM, and an Intel i9-13900K CPU. We implemented our framework in PyTorch, building upon the official 3D Gaussian Splatting repository[1]. For deformable Gaussian splatting, we adapted the implementation from Yang et al. [34]. We use the default parameters and configurations for 3DGS and Deformable 3DGS.

A.4.1 Joint RGB-Thermal Optimization

For Stage 3, we use the following hyperparameters:

- $\lambda_{\text{render}} = 1.0$
- $\lambda_{\text{smoke\_alpha}} = 0.1$
- $\lambda_{\text{smoke\_color}} = 0.05$
- $\lambda_{\text{mono}} = 0.1$
- $\lambda_{\text{depth}} = 2.0$
- $\lambda_{\text{mask}} = 0.5$

A.5 Creating of reference image for real world experiments

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. To create this reference, we collected additional smoke-free RGB images of the same scenes in a separate drone flight. We then performed the following steps: (1) reconstructed the smoke-free scene using 3DGS, (2) obtained the camera poses of smoke-filled and smoke-free image sets in the same coordinate frame, and (3) rendered novel views of the smoke-free reconstruction using the camera poses from the smoke-filled sequence. These reference images serve as an approximate benchmark, though they are not perfect ground truth since the poses are noisy and environmental conditions may have changed between captures.

[1] https://github.com/graphdeco-inria/gaussian-splatting

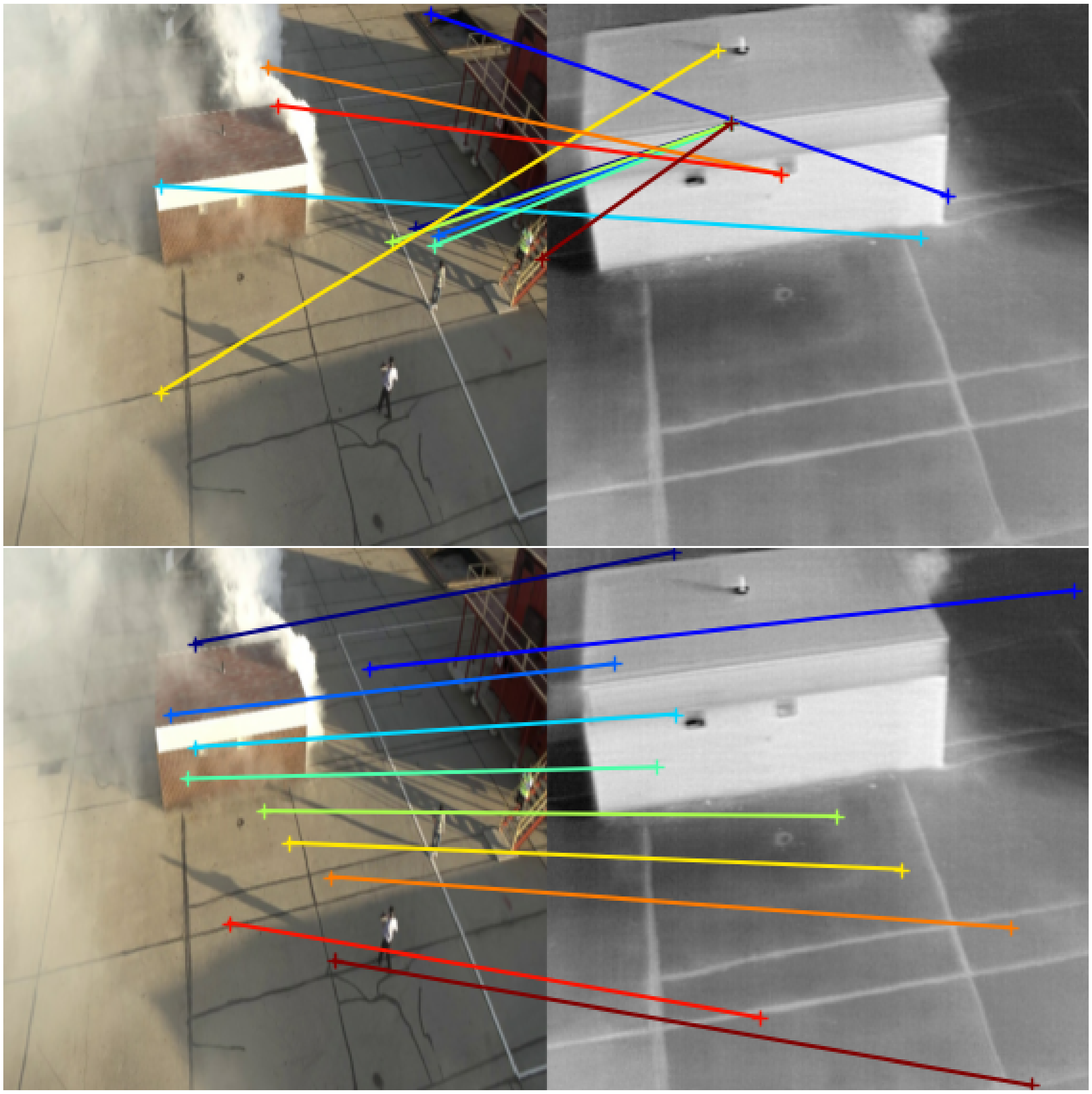


Table A.3: Cross-modal feature matching comparison. Top: Traditional SIFT feature matching fails between RGB and thermal images due to fundamental differences in appearance across modalities. Bottom: MINIMA [13] provides more reliable correspondences by explicitly addressing cross-modal challenges, enabling accurate registration of the two sensor types.

Bibliography

[1] Junaid Amin, Muhammad Sharif, Nargis Gul, Mudassar Raza, Muhammad Almas Anjum, Muhammad Waseem Nisar, and Syed Ahmad Chan Bukhari. Thermal imaging systems for real-time applications in smart cities: A review. *Journal of Real-Time Image Processing*, 19:283–302, 2022. 2.3

[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. 2.1

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 2.3

[4] Subrahmanyan Chandrasekhar. *Radiative Transfer.* Dover Publications, New York, 1960. Unabridged and slightly revised version of the work first published in 1950. 2.1

[5] 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, 2.2, 3.3

[6] Zhixiang Chen, Jiabo Li, Yinhao Zhang, Yifan Peng, Guanyu Jiang, Xiangyang Yu, and Huanqiang Zeng. All-weather autonomous driving: Can we learn from thermal imaging? *IEEE Transactions on Intelligent Transportation Systems*, 23 (8):10965–10976, 2021. 2.3

[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. URL `https://api.semanticscholar.org/CorpusID:272988049`. 2, A.2, A.3

[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*, SIGGRAPH '07, page 60, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781450378369. doi:

10.1145/1275808.1276452. URL https://doi.org/10.1145/1275808.1276452. 3.1

[9] Satoshi Fujita, Masahiro Oda, Noriaki Shimizu, Holger R Roth, Yuichiro Hayashi, Masahiro Ito, Takayuki Kitasaka, Kazunari Misawa, and Kensaku Mori. Smoke removal in endoscopic images using physics-based color enhancement and gan-based restoration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, pages 3–13. Springer, 2022. 2.1

[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. 2.1

[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*, volume 33, pages 2341–2353. IEEE, 2011. 2.1

[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. 2.3

[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. (document), 3, A.3, A.3

[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), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/. 1, 2.2, 3.2, 3.6

[15] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Thermal-inertial slam for the environments with challenging illumination. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5123–5129. IEEE, 2023. 2.3

[16] H. Koschmieder. *Theorie der horizontalen Sichtweite*. Beiträge zur Physik der freien Atmosphäre. Keim & Nemnich, 1924. URL https://books.google.com/books?id=FAOgHAAACAAJ. 2.1

[17] Zheyan Li, Yunxuan Guo, Yiqi Yan, Di Huang, and Xiaoguang Wang. You only need attention: Transformer-based dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18573–18582, 2022. 2.1

[18] Yvette Y Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. Ther-

malNeRF: Thermal radiance fields. In *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2024. 2.2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. A.1

[20] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Multi-scale residual learning for single image dehazing. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1366–1371. IEEE, 2019. 2.1

[21] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. A.2

[22] Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermalgaussian: Thermal 3d gaussian splatting. *arXiv preprint arXiv:2409.07200*, 2024. 2.2

[23] 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. 2.2

[24] 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. 2.2, 4.2

[25] 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*, SIGGRAPH '01, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 158113374X. doi: 10.1145/383259.383317. URL https://doi.org/10.1145/383259.383317. 3.2

[26] 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, October 2023. 1, 2.2

[27] 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. URL https://arxiv.org/abs/2408.00714. 1, A.1

[28] 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. 1, A.1

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. A.3

[30] Nils Thuerey and Tobias Pfaff. MantaFlow, 2018. *http://mantaflow.com*. 4.2

[31] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. (document), 5.1, 6.1

[32] 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. 2.1

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3.9

[34] 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. 3.8, A.4

[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. 2.1

[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. 2.1