FLAME DIFFUSER: GROUNDED WILDFIRE IMAGE SYNTHESIS USING MASK GUIDED DIFFUSION

Hao Wang, Sayed Pedram Haeri Boroujeni
School of Computing
Clemson University
Clemson, SC, USA
hao9, shaerib@g.clemson.edu

Xiwen Chen
School of Computing
Clemson University
Clemson, SC, USA
xiwenc@g.clemson.edu

Ashish Bastola
School of Computing
Clemson University
Clemson, SC, USA
abastol@g.clemson.edu

Huayu Li
Department of Electrical and Computer Engineering
The University of Arizona
Tucson, AZ, USA
hl459@arizona.edu

Abolfazl Razi
School of Computing
Clemson University
Clemson, SC, USA
arazi@clemson.edu

ABSTRACT

The rise of machine learning in recent years has brought benefits to various research fields such as wide fire detection. Nevertheless, small object detection and rare object detection remain a challenge. To address this problem, we present a dataset automata that can generate ground truth paired datasets using diffusion models. Specifically, we introduce a mask-guided diffusion framework that can fusion the wildfire into the existing images while the flame position and size can be precisely controlled. In advance, to fill the gap that the dataset of wildfire images in specific scenarios is missing, we vary the background of synthesized images by controlling both the text prompt and input image. Furthermore, to solve the color tint problem or the well-known domain shift issue, we apply the CLIP model to filter the generated massive dataset to preserve quality. Thus, our proposed framework can generate a massive dataset of that images are high-quality and ground truth-paired, which well addresses the needs of the annotated datasets in specific tasks.

***K*eywords** Generative Models · Image Processing · Prompt Engineering · Diffusion Model · Wildfire Detection · Forestry

1 Introduction

Over the past few years, wildfires have become increasingly devastating, causing widespread destruction to natural environments, human settlements, and infrastructure. Wildfires significantly reduce the quality of air and soil while causing detrimental impacts on the biodiversity of the affected regions [1]. Additionally, they increase the risk of flooding and landslides phenomena and contribute to climate change by releasing greenhouse gases into the atmosphere through the destruction of vegetation. These facts urge the development of more effective fire management strategies. Recent developments in Artificial Intelligence (AI), powered by Deep Learning (DL) methods, have resulted in more accurate data-driven methods to identify and locate fires by processing fire imagery captured by observation towers, satellites, and manned and unmanned aircraft. Meanwhile, object detection plays a significant in various cutting-edge technologies such as autonomous driving, healthcare, military, and forestry. Computer vision tasks rely heavily on object detection algorithms, as they accurately identify and localize instances of a specific class within an image.

In forestry, wildfire detection become a focus in recent decades since precisely locating the heat source in real-time is a huge challenge [2]. Previous works rely on either advanced hardware such as UAV-based dual-channel thermal cameras, or the specialized dataset with human annotation [3]. The former method provided a solution with fast detection in real-time, yet lack of practical implementation since the framework only works on the aerial level and needs time for deployment. The aforementioned method requires hours of human laboring on image annotating and lacks generalizability when applying the trained model to other scenarios due to the well-known domain-shift problem [4].

Classical object detection models such as YOLO, require training to achieve detection on custom datasets. A key challenge is small object detection since the feature information of small objects is different than that of regular objects. Wildfire detection also suffers from this issue, as wildfires are usually small in captured geology images compared to regular camera-captured flames in fire accidents [5]. Furthermore, wildfire images are usually difficult to capture, which makes wildfire datasets precious and lacks annotation.

Recently, language model-based object detection models brought a breakthrough to this area, that pre-trained foundation models can be instantly applied to various tasks without any post-training or fine-tuning [6]. For instance, the detectable classes of zero-shot-based object detection models depend on the text prompt. In other words, a pre-trained zero-shot object detector can respond to any object that is semantically related to the text prompt. This flexibility brings huge advances to various applications, especially practical tasks that utilize large language models [7]. However, for specific objects that are rare and unusual, the foundation model may fail to detect due to the long-tail issue [8].

Diffusion models become a game-changer in the world of generative AI, excelling at creating realistic data. This addresses the widespread issue of data scarcity across diverse fields. Diffusion models offer solutions for data augmentation, thereby enhancing the diversity and volume of datasets, which is essential for training in machine learning tasks. Diffusion models are also key in adapting data for different uses, enabling the safe use of synthetic data that complies with privacy regulations, and even sparking new creative projects [9]. Figure 1 demonstrates that by integrating this method with text-based directions, users can influence both the content and context of the generated images. Nonetheless, this control is very limited, as the textual parameters merely offer guidance for image synthesis at a mathematical level. Achieving precise control through text prompts remains a challenge, as the denoising process typically aims to produce an image that aligns with the textual description, yet it does not always allow for exact manipulation of specific image features (e.g., the number of strips on a flag) [10].

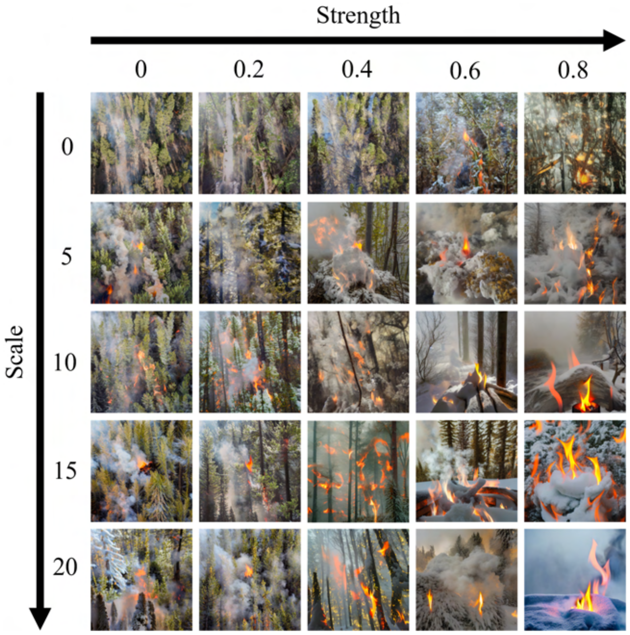

Figure 1: Wildfire image synthesis in terms of text prompt scale and image input strength. Text prompt: 'Photo realistic, wildfires in the snow, flame in the smoke, high-resolution'.

Thus, we propose a framework that can synthesize wildfire images with paired ground truth. The proposed framework utilizes diffusion models with a mask module to guide the synthesis of fire flames, where the mask itself becomes the ground truth of the synthesized images. This eliminates the need for annotation with heavy human laboring on massive datasets. Meanwhile, the diffusion module can generate wildfire images with precisely controlled backgrounds and context based on both the input image and text prompt. This enhances the generalizability of trained models to achieve 'multi-task, one train', and makes it possible to build customized foundation models on other specific areas such as retina vessel segmentation, dendrite pattern detection, and material anomaly detection [11]. Furthermore, we apply the CLIP model to manage the quality of synthesized images [7], to produce a high-quality dataset.

Specifically, the major contributions of this paper are summarized as follows:

- We propose a framework that utilizes an architecture based on the diffusion model to synthesize wildfire images with annotated bounding boxes.
- We implement a mask module that can guide the location of fire flames in image synthesis process.
- We apply the CLIP model to improve the data qualities by filtering unsatisfied images.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work in the field of wildfire datasets, state-of-the-art object detection methods, and text-to-image architecture. In section 3, we present the details of the methodology and implementation of our proposed approach. The experimental results are discussed in Section 4, followed by concluding remarks and future directions in Section 5.

2 Related Work

In this section, we provide background information by reviewing the recent developments in wildfire detection and image analysis.

2.1 Wildfire Datasets

There exist a few publicly available datasets that provide aerial imagery during an active forest fire. We investigated such datasets and found FLAME1[12] and FLAME2 [13] and the D-Fire dataset most appropriate for this study since they provide a comprehensive collection of fire aerial images in both RGB and IR domains.

The FLAME1 dataset provides multiple aerial video recordings collected by drones throughout a prescribed burn operation. Each video has been transformed into individual image frames to enable image-based analysis according to the Frames Per Second (FPS) rate. The dataset includes four different types of visual representations, including normal spectrum, fusion, white-hot, and green-hot schemes. Figure 2 presents some samples of the FLAME1 dataset to provide a better insight into the visual contents of the collected fire images.

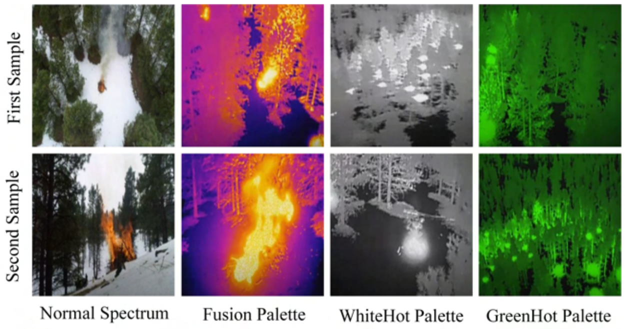

Figure 2: Four examples of different visual representations from the FLAME1 dataset.

Likewise, the FLAME2 dataset includes video recordings and thermal heatmaps captured by infrared cameras. The IR and RGB modes are fully synchronized, so the IR images can be used as a ground truth representation. The captured videos and images are manually annotated by three experts, and labeled frame-wise to provide a comprehensive dataset for various image-based tasks. This dataset includes 53,451 pairs of $254p \times 254p$ RGB/IR images labeled with three classes: (1) Flame with Smoke, (2) Flame with No Smoke, and (3) No Flame with No Smoke. Figure 3 presents some samples of dual RGB/IR images from the FLAME2 dataset to provide a better insight into the dataset's characteristics.

The D-Fire dataset [14] is a comprehensive image dataset designed specifically for machine learning and object detection algorithms in the context of fire and smoke detection. It contains over 21,000 images, categorized into four distinct groups: 1,164 images featuring only fire, 5,867 images with only smoke, 4,658 images showing both fire and smoke, and 9,838 images with neither, serving as a control group. The dataset includes a total of 14,692 bounding boxes for fire and 11,865 for smoke, all annotated in the YOLO format with normalized coordinates. This extensive and well-annotated dataset is particularly valuable for training and evaluating models in fire and smoke detection, making it a significant resource for research and development in this critical area of computer vision and emergency response technology. Figure 4 presents some samples of various images from the D-Fire dataset to provide a better insight into the dataset's characteristics.

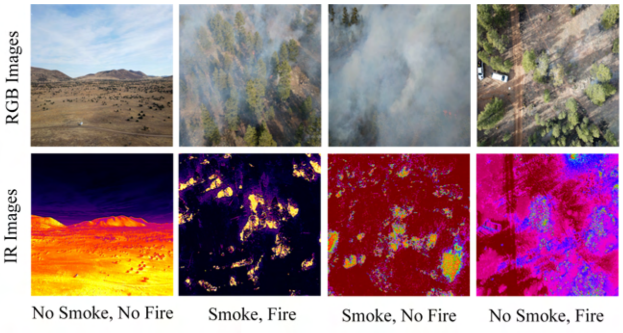

Figure 3: Different examples of dual RGB/IR images as well as their labels from the FLAME2 dataset.

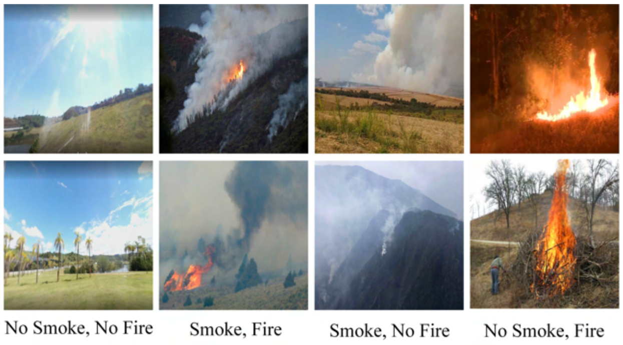

Figure 4: Four examples of different class representations from the D-Fire dataset.

2.2 Object Detection

Over the past few years, object detection has received extensive attention from researchers due to its role in many AI-based monitoring platforms [1]. Many computer vision tasks rely heavily on object detection algorithms, as they play a crucial role in accurately identifying and localizing instances of a specific class within an image. For this purpose, different Deep Learning (DL) is employed various DL networks are employed to enhance the performance of object detection tasks, improving both accuracy and speed while reducing computational costs [3]. Some of the recent and well-known object detection methods include Fast Region-based Convolutional Neural Networks (Fast R-CNN) [15], You Only Look Once (YOLO) [16], Single Shot Multibox Detector (SSD) [17], and Grounding Dino [6]. These methods have demonstrated promising results in achieving the desired performance for object detection tasks in various applications such as fire management [18], traffic analysis [19], and agriculture monitoring [20].

In recent years, several versions of YOLO, from YOLOv1 to YOLOv8, have been successively introduced to enhance the quality of object detection tasks, enabling them to effectively address a wide range of challenges [21]. Generally, YOLO takes an image as input and divides it into the $S \times S$ grid cells. Each grid cell is responsible for predicting an object within the bounding boxes along with the class probability. The YOLO architecture is primarily based on Convolutional Neural Networks (CNNs) to perform feature extraction and prediction tasks. Each version employed different loss functions with distinct characteristics that significantly enhanced the detection performance regarding accuracy, speed, complexity, and computational cost.

Grounding Dino [6] stands as another recent advancement within the field of object detection. The novel idea behind this method is that the Transformer-based detector Dino is concatenated with grounded pre-training, allowing it not only to detect different objects but also to refer to the objects' class names. To achieve this goal, the authors effectively fuse visual and linguistic information to attain the desired results. This technique consists of three main phases, including a feature enhancer, language-guided selection, and cross-modality feature fusion.

In the first phase, the model extracts multi-scale features from the pair of image/text inputs using image and text backbone architectures. Then, they are fed into the feature enhancer block for cross-modality feature fusion. During the second phase, a language-guided selection module is designed to assist the object detection procedure. It achieves this goal by selecting the most relevant image features that align with the content of the input text. At the last stage, a cross-modality decoder is developed to fuse the image and text features. This involves passing the cross-modality output to the self-attention layers, followed by the image cross-attention layer and the text cross-attention layer to combine image and text features, respectively. Lastly, a Feed-Forward Network (FFN) layer in each cross-modality decoder layer is used to perform further processing and update the cross-modality output.

2.3 Latent Diffusion

Latent diffusion represents an innovative technique in the realm of generative models, which are designed to synthesize novel data samples. These models have been applied in diverse applications, especially in image synthesis [9].

The foundational principle of latent diffusion models is inspired by the diffusion process observed in physics [22]. Different from the conventional diffusion models that directly manipulate the data on the pixel space, latent diffusion models operate within a condensed and abstract representation of the data, known as the latent space. This innovative approach not only enhances the efficiency of the model but also contributes significantly to the generation of diversity outputs [9].

Throughout the training phase, these models demonstrate a capacity to manipulate the data within this latent space. This includes a learning process where a Variational Autoencoder (VAE) is trained to encode the data into the latent space and subsequently decode it, reconstructing it from its latent representation back to its original format in pixel domain [9]. This diffusion and denoise operation in latent space maximizes the efficiency of data synthesis in both computation cost optimization and image quality improvement.

2.4 Image-to-Text Architexture

CLIP stands for Contrastive Language–Image Pre-training, is a cutting-edge model developed by OpenAI that plays an inevitable role in the advancement of text-to-image architectures [7]. This technology essentially bridges the gap between textual descriptions and visual content, enabling machines to understand images in the context of natural language descriptions.

Specifically, CLIP is designed to learn visual concepts from natural language descriptions. It trains on a wide variety of images and their corresponding text captions, allowing it to understand and generate content across different modalities. This multimodal understanding is crucial for creating more sophisticated text-to-image models that can accurately interpret and visualize complex text descriptions. At the heart of CLIP's effectiveness is contrastive learning. This approach involves training the model to recognize which images correspond to which text descriptions among a batch of incorrect pairings. By doing so, CLIP learns a rich representation of images and text that closely aligns with how humans perceive and describe visual content. This learning method significantly improves the model's ability to generate relevant and detailed images based on textual input.

One of the key strengths of CLIP is its ability to generalize across a broad range of tasks without task-specific training data. This versatility means that CLIP can be applied to a wide array of applications, from generating images based on textual descriptions to improving search engines by allowing them to understand the content of images in relation to text queries.

CLIP's capabilities have been instrumental in the development of advanced text-to-image models, such as DALL·E [23], which can generate highly detailed and creative images from textual descriptions. By incorporating CLIP's understanding of the relationship between text and images, these models can produce results that are not only visually impressive but also semantically aligned with the input text.

3 Methodology

In this section, we explain the process of mask generation, image fusion of mask and raw image, and the mask-text-image guided image synthesis.

3.1 Framework

The general architecture of our proposed method is illustrated in Figure 5.

To manage the shift in context effectively, we start by merging the mask layer with the original image early on, before encoding it using a VAE. This ensures that the context and content of the image align closely with the intended outcome. By integrating specific information relevant to the desired context into the generation phase, we aim to accomplish this context shift. We structure this approach around a conditional generative model, allowing for a more targeted and

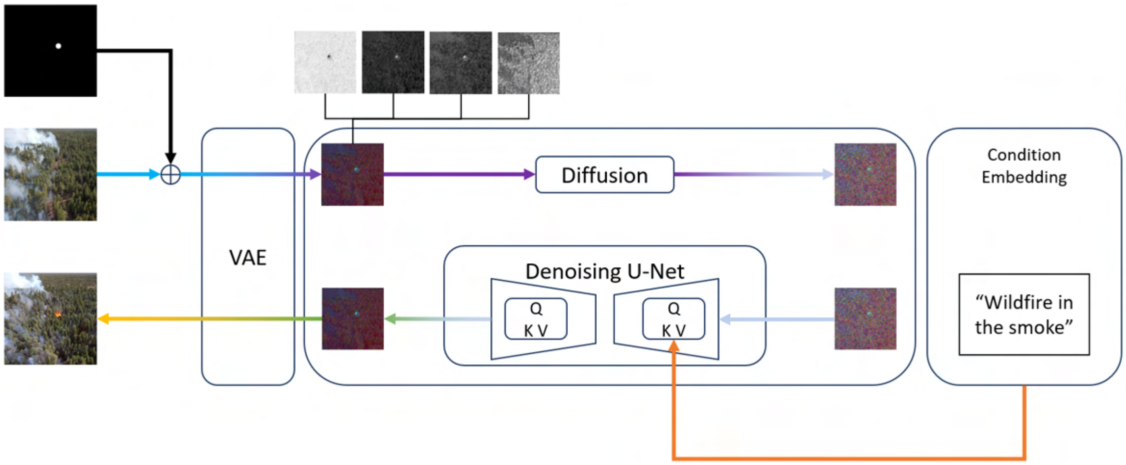

Figure 5: Architecture of proposed method.

controlled generation of data. This can be formalized as:

$$\hat{z} = \epsilon(x \odot m)$$
$$\tilde{x} = \mathcal{D}(\hat{z}; c)$$

Where: m is the control matrix containing target context-specific information, x denotes the raw image, ϵ represents the VAE encoding process, $\mathcal{D}$ represents the VAE decoding process, c is the conditional variable (e.g., text prompt), $\hat{z}$ represents the latent variables, and $\tilde{x}$ is the synthesized data sample.

By manipulating the data generation process, it becomes possible to intentionally induce a context shift of the synthesized data, thereby aligning it with the expectation of the target data.

3.2 Mask Generation

The first step in our approach involves generating masks that define specific areas of interest in the synthesized data. We use fundamental mathematical algorithms such as Gaussian functions, shapes, and filters to create these masks. These masks guide the generation of data with desired attributes, allowing us to control the context during the image synthesis. Figure 6 shows the process of wildfire image synthesis with randomly generated masks.

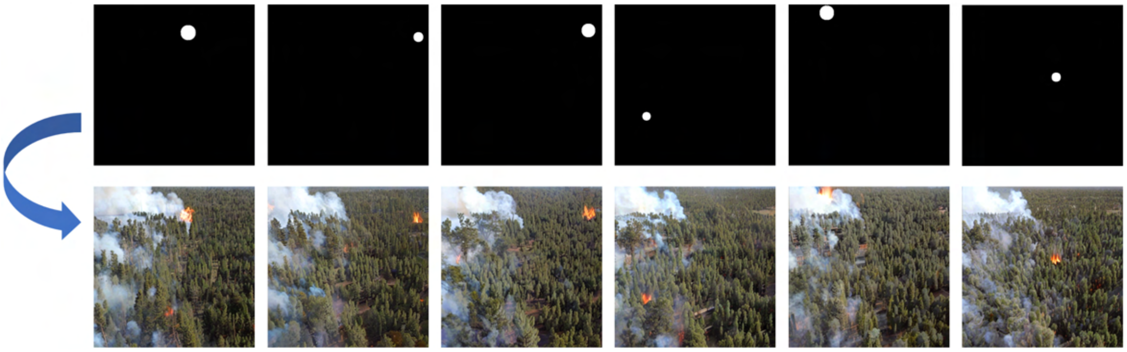

Figure 6: Wildfire image synthesis with random generated masks.

With the generated masks, we proceed to synthesize data using the diffusion model. This model takes both latent variables and masks as inputs to produce realistic synthesized data. The masks act as guides, shaping the features of the generated objects.

To achieve the context shift, we apply the control matrix c to influence the generated samples before the encoding. After the fusion of both control matrix c and the input image, we apply the original encoder to generate the latent variable $\hat{z}$. By manipulating c, we can guide the position prior to the context (e.g., flame element) in the generated image to better match the expectation.

3.3 Data Filtering using CLIP

We leverage the text-to-image mapping capabilities of CLIP to assess the content within the bounding boxes of synthesized images. The classification results from CLIP determine the class of the image patches contained within these bounding boxes. Specifically, we implemented custom class labels for CLIP to use in predicting the categories of these bounding boxes. We exclude images if the identified class for a bounding box is not related to wildfires, or if the confidence level of the classification is too low. This ensures that the constructed dataset remains focused and relevant to our specific research needs. As shown in Figure 7, images with non-wildfire classes will be excluded after the CLIP prediction.

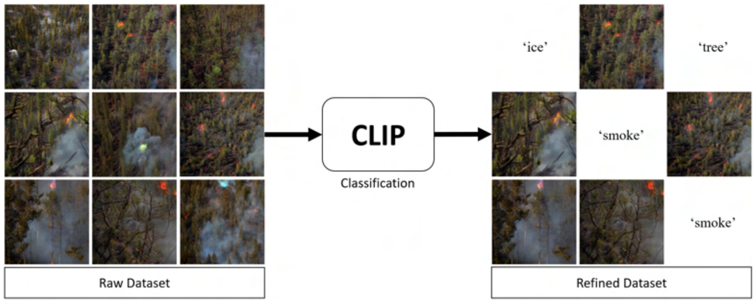

Figure 7: Using CLIP classification as a data filter.

4 Experimental Results

In this section, we show the qualitative results by illustrating the influence of different parameters on the context of synthesized images. Furthermore, we discuss the impact and interpretation of implemented modules on the synthesized data.

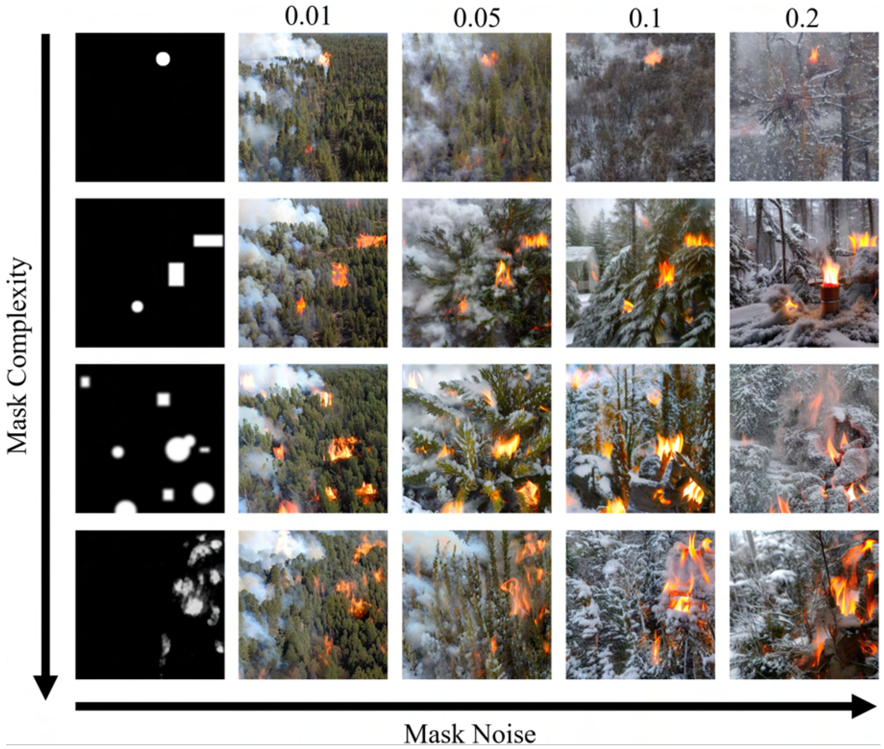

Figure 8: Wildfire image synthesis in terms of mask noise level and mask complexity.

4.1 Context Control with Diffusion Settings

As shown in Figure 8, the change of control parameters instantly impacts the context of synthesized images. Specifically, the mask shape and position determine the flame position on the synthesized image. By controlling the mask shape, the number of shapes, and the position of shapes, the flame on the synthesized image can be manipulated precisely.

Meanwhile, the variation in mask noise level impacts the content consistency of the synthesized image which is guided by the text prompt. Different than the text prompt scale, which could interrupt the text guide strength, that does not consider the physical logic of data synthesis and may harm the realism of the generated image. On the other hand, the masking noise conducts the overall style of images by reducing the amount of information that is propagated from the raw image, thereby the realism of the image remains.

As shown in Figure 9, we apply the thermal mask of the FLAME 2 dataset as the input mask, and the output images synthesize wildfires whose position is close to the mask.

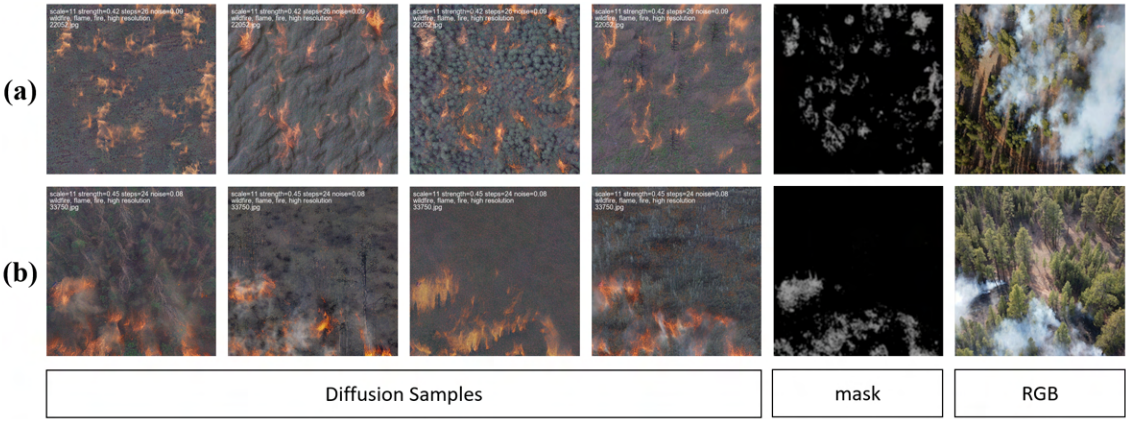
(a)
scale=11 strength=0.42 steps=26 noise=0.09
wildfire, flame, fire, high resolution
22052.jpg
scale=11 strength=0.42 steps=26 noise=0.09
wildfire, flame, fire, high resolution
22052.jpg
scale=11 strength=0.42 steps=26 noise=0.09
wildfire, flame, fire, high resolution
22052.jpg
scale=11 strength=0.42 steps=26 noise=0.09
wildfire, flame, fire, high resolution
22052.jpg
(b)
scale=11 strength=0.45 steps=24 noise=0.08
wildfire, flame, fire, high resolution
33750.jpg
scale=11 strength=0.45 steps=24 noise=0.08
wildfire, flame, fire, high resolution
33750.jpg
scale=11 strength=0.45 steps=24 noise=0.08
wildfire, flame, fire, high resolution
33750.jpg
scale=11 strength=0.45 steps=24 noise=0.08
wildfire, flame, fire, high resolution
33750.jpg
Diffusion Samples
mask
RGB

Figure 9: Wildfire image synthesis with thermal masks from FLAME2.

4.2 Dataset Construction

Figure 10 shows the selected images to represent the quality of our proposed dataset.

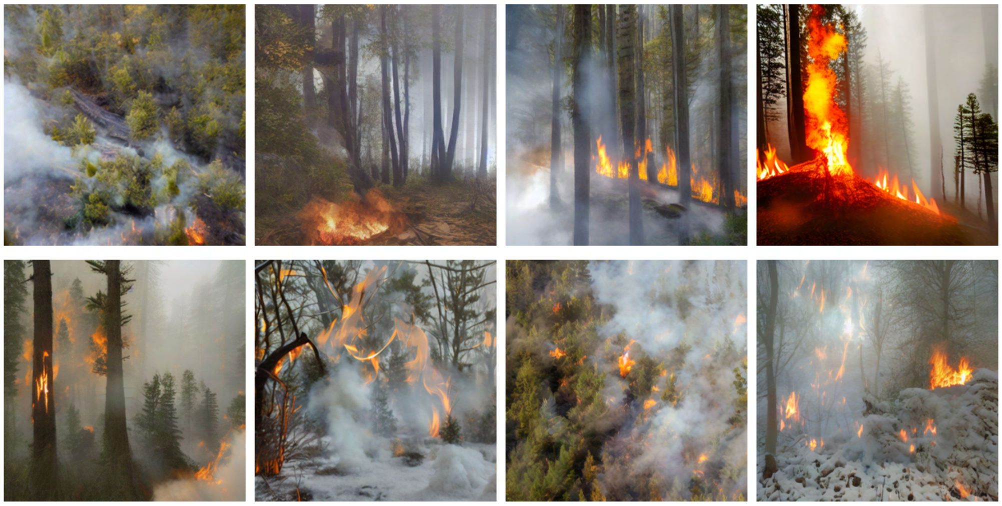

Figure 10: Selected images from the constructed dataset.

As illustrated in Figure 11, the synthesized data now features bounding boxes that correspond accurately with its masks. This combined annotation approach not only points out the precise location of wildfires within the images but also labels corresponding objects with a measure of ground confidence. By manipulating image styles through text prompts, we have been able to create a substantial annotated dataset.

The resources for accessing the code and dataset are provided at the link: `https://arazi2.github.io/aisends.github.io/project/flame`

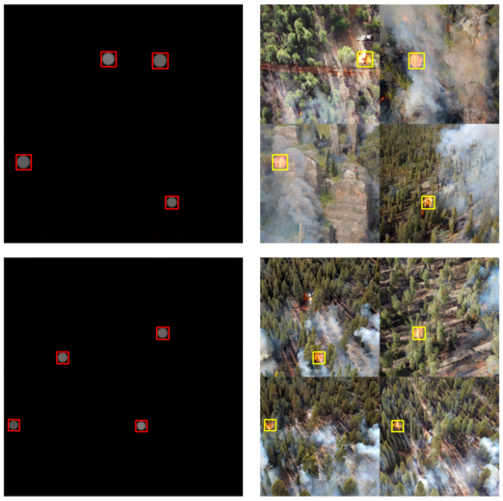

Figure 11: Joint annotation between masks and synthesized images.

5 Conclusion

The current solution for specific object detection reveals the need for data expansion in the existing world. However, the existing datasets in a few specific research areas are precious and lack annotation. Conventional labeling and data analysis can be labor-intensive and time-consuming, which may lead to unexpected consequences in such tasks.

Flame Diffuser as demonstrated, effectively addresses the current issue. By integrating diverse backgrounds and contexts into the image synthesis process, ensures that the models are robust and adaptable to various real-world scenarios. This approach not only enhances the quality and utility of the dataset but also significantly reduces the time and effort required for manual annotation. Consequently, this leads to more efficient and effective development of fire detection models, which is crucial for early detection and prevention of wildfires, ultimately contributing to the preservation of ecosystems and the protection of human and wildlife habitats. The proposed framework indicates the potential of diffusion models in solving complex challenges in data generation and paves the way for similar advancements in other areas of object detection and machine learning research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number CNS-2204445. Special thanks to USDA Forest Service and Kaibab National Forest administration for providing access to the prescribed fire.

References

[1] S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. ONeill, P. Z. Fule, A. Watts, N.-M. T. Kokolakis, and K. G. Vamvoudakis, "A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," *arXiv preprint arXiv:2401.02456* , 2024.

[2] X. Chen, B. Hopkins, H. Wang, L. O'Neill, F. Afghah, A. Razi, P. Fulé, J. Coen, E. Rowell, and A. Watts, "Wildland fire detection and monitoring using a drone-collected rgb/ir image dataset," *IEEE Access* **10**, pp. 121301–121317, 2022.

[3] S. P. H. Boroujeni and A. Razi, “Ic-gan: An improved conditional generative adversarial network for rgb-to-ir image translation with applications to forest fire monitoring,” *Expert Systems with Applications* **238**, p. 121962, 2024.

[4] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, “Adaptive risk minimization: Learning to adapt to domain shift,” *Advances in Neural Information Processing Systems* **34**, pp. 23664–23678, 2021.

[5] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, “Uav-yolo: Small object detection on unmanned aerial vehicle perspective,” *Sensors* **20**(8), p. 2238, 2020.

[6] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv preprint arXiv:2303.05499* , 2023.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[8] W. Yu, T. Yang, and C. Chen, “Towards resolving the challenge of long-tail distribution in uav images for object detection,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3258–3267, 2021.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[10] M. N. Everaert, M. Bocchio, S. Arpa, S. Süsstrunk, and R. Achanta, “Diffusion in style,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2251–2261, 2023.

[11] X. Chen, H. Wang, A. Razi, M. Kozicki, and C. Mann, “Dh-gan: a physics-driven untrained generative adversarial network for holographic imaging,” *Optics Express* **31**(6), pp. 10114–10135, 2023.

[12] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, “Aerial imagery pile burn detection using deep learning: The flame dataset,” *Computer Networks* **193**, p. 108001, 2021.

[13] B. Hopkins, L. O’Neill, F. Afghah, A. Razi, E. Rowell, A. Watts, P. Fule, and J. Coen, “Flame 2: Fire detection and modeling: Aerial multi-spectral image dataset,” 2022.

[14] P. V. A. de Venancio, A. C. Lisboa, and A. V. Barbosa, “An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices,” *Neural Computing and Applications* **34**(18), pp. 15349–15368, 2022.

[15] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[16] T. Diwan, G. Anirudh, and J. V. Tembhurne, “Object detection using yolo: Challenges, architectural successors, datasets and applications,” *multimedia Tools and Applications* **82**(6), pp. 9243–9275, 2023.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37, Springer, 2016.

[18] Y. Wang, C. Hua, W. Ding, and R. Wu, “Real-time detection of flame and smoke using an improved yolov4 network,” *Signal, Image and Video Processing* **16**(4), pp. 1109–1116, 2022.

[19] M. Sha and A. Boukerche, “Performance evaluation of cnn-based pedestrian detectors for autonomous vehicles,” *Ad Hoc Networks* **128**, p. 102784, 2022.

[20] S. Khan and L. AlSuwaidan, “Agricultural monitoring system in video surveillance object detection using feature extraction and classification by deep learning techniques,” *Computers and Electrical Engineering* **102**, p. 108201, 2022.

[21] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, “A review of yolo algorithm developments,” *Procedia Computer Science* **199**, pp. 1066–1073, 2022.

[22] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems* **33**, pp. 6840–6851, 2020.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, 2021.