GIMO: Generative Image Outpainting for Early Smoke Segmentation

Anonymous CVPR submission

Paper ID 68

Abstract

The critical challenge of early smoke segmentation for bushfire detection is hindered by the inherent transparency, deformable shape, and small size of nascent smoke regions. To overcome this, we propose a novel generative data augmentation framework leveraging image outpainting to simulate fixed-camera perspective transformations, effectively generating smaller smoke instances while preserving existing segmentation labels. We employ a diffusion generative model to outpaint smoke regions, enlarging real-world smoke images with synthetic domain-matched pixels. Experiments conducted on a state-of-the-art baseline demonstrate significant improvements, achieving a ***3%*** *increase in mean Intersection over Union (mIoU) for small smoke and a 0.9% overall mIoU boost. These results highlight the efficacy of our generative data augmentation pipeline in mitigating data scarcity, emphasising its potential for enhancing early wildfire detection and enabling timely deployment of fire services.*

1. Introduction

Bushfires are frequent and devastating natural disasters, resulting in significant loss of human life, wildlife, and environmental damage. Camera-based bushfire detection primarily relies on identifying visible smoke. Early detection of smoke at the initial stages of ignition is crucial, as it facilitates prompt responses by emergency services. Once a fire becomes established, particularly under conditions of high fire danger, extinguishing it becomes extremely challenging or even impossible [17]. Therefore, accurately detecting small, early-stage smoke remains a critical challenge for effective bushfire mitigation.

Early Smoke Segmentation (ESS) has recently become a prominent area of research in wildfire detection [14, 15]. Detecting early-stage smoke is particularly challenging due to its relatively small size in bushfire scenarios, especially when monitoring imagery encompasses expansive landscapes. Consequently, identifying smoke from newly ignited fires under these conditions becomes significantly more difficult.

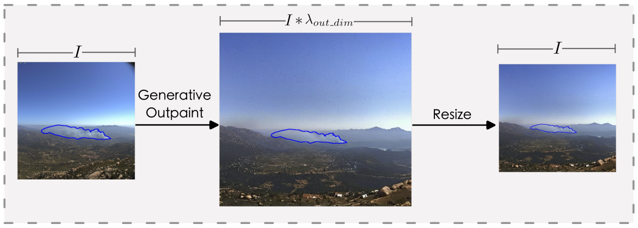

Figure 1. Our GIMO pipeline obtains smaller smoke regions while preserving segmentation masks by using a generative model to outpaint a labelled smoke image.

Although deep learning methods [13, 15, 18, 19] have shown substantial potential to improve detection performance, available data sets often lack sufficient diversity and do not adequately represent the nuances of this specific problem domain. Due to challenges in acquiring large-scale in-the-wild images of bushfire smoke and the labor-intensive nature of manually annotating segmentation masks, existing datasets typically suffer from limited size, insufficient domain-specific scenes, or both [13, 14].

Given the difficulty in obtaining early-stage images with small-size smoke, we investigate whether simulating such conditions by positioning smoke farther away and increasing the proportion of background in images could alleviate the shortage of data for ESS. Outpainting [3, 5, 10] is a task where a source image is extrapolated beyond its original borders using synthetic, semantically-matching pixel content. In this study, we leverage generative frameworks [8, 9] to augment existing smoke datasets by manipulating original images and integrating new backgrounds, while preserving the original segmentation annotations. We conduct experiments using the available labelled smoke segmentation dataset, demonstrating significant improvements in segmentation performance.

We detail our Generative Image Outpainting augmentation framework with our main contributions as follows:

1. We propose a novel generative data augmentation framework designed for simulating fixed-camera perspective transformation on regular smoke images to obtain

smaller smoke regions.

2. We design an augmentation pipeline that leverages an image-conditioned diffusion generative model to outpaint smoke regions while preserving and transforming the old segmentation labels.
3. We test our dataset using a state-of-the-art baseline and show 3% improvement in *mIoU* for small smoke while posting a 0.9% boost in overall *mIoU*. We demonstrate, through these significant boost in performances, that the task of early smoke segmentation is severely impaired by insufficient training data and highlight the efficacy of generative data augmentation in data-scarce scenarios.

2. Method

We detail the process of our proposed **G**enerative **IM**age **O**utpainting (GIMO) data augmentation framework in Figure 2. GIMO leverages an image-guided generative process to produce in-domain smoke images targeting small smoke. Let $X \subseteq \mathbb{R}^{H \times W \times C}$ be the input image space and $Y \subseteq \mathbb{R}^{H \times W \times 1}$ be the segmentation label space. Let the population of image-label pairs be denoted by $Z = X \times Y$. Given a training set $S \in Z$ with n_S independent and identically distributed (i.i.d) samples from Z, we use a trained diffusion generative model to augment image-label pairs from S and obtain an augmented set $\widetilde{S}$. GIMO transforms a given input pair $(x \in X, y \in Y)$ through an out-painting process that preserves image dimensions $GIMO : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W \times C}$. We use a conditional diffusion model $G(.)$ to transform images from S and extract n_G i.i.d samples to form a new augmented dataset $\widetilde{S}$ where n_G is a hyper-parameter that represents the number of new image-label pairs we wish to generate. To train smoke segmentation models, we combine the original along with our augmented dataset to obtain $S_{combined} = S + \widetilde{S}$. More specifically, given an input smoke image x and label y, we obtain an outpainted image $\widetilde{x}$ and $\widetilde{y}$ using the following steps:

Staged Canvas. We define a hyper-parameter λ_{out_dim} that controls the dimensions of outpainted image with respect to the original. For simplicity, we apply λ_{out_dim} equally to H and W. We then pad our source image x and its corresponding label y to the outpainted image size:

$$\begin{aligned} H_{out} &= H * \lambda_{out\_dim} \\ W_{out} &= W * \lambda_{out\_dim} \end{aligned} \tag{1}$$

We create a white RGB image canvas $\widetilde{x} \in \mathbb{R}^{H_{out} \times W_{out} \times C}$ where we randomly position the source image x. Similarly, we create a label image canvas $\widetilde{y} \in \mathbb{R}^{H_{out} \times W_{out} \times 1}$ where we match the positioning to ensure that the smoke segmentation label in $\widetilde{y}$ aligns with the smoke image region in $\widetilde{x}$. With this step, our final segmentation label image is prepared and we have staged our RGB image canvas ready for generative outpainting.

Obtain Scene Prompt. Along with our staged canvas image, we acquire a text prompt to help guide the generator to "fill in" missing pixels in our staged canvas $\widetilde{x}$. Since text prompts for diffusion models strongly affect the resulting generation and are often highly engineered [11], in order to obtain specific prompts suited for text-to-image diffusion models, we use InternLM2 [1] which is a pre-trained open-source Vision Large Language Model that outputs high-quality prompts that describe the input image scene.

Synthetic Outpaint. Armed with our staged canvas image and a text prompt that describes our desired image scene, we pass these to ControlNet [22] to obtain an outpainted image $\widetilde{x}$. We find that the initial output $\widetilde{x}$ contains some artefacts particularly around the border of the source image x. To refine the threshold between the original and synthetic pixels as well as preserve the details of the source image, we pass $\widetilde{x}$ through the diffusion model again but with an Image Prompt adapter [16], along with a mask overlaying the position of the original image pixels, to provide more context regarding the image scene. This helps focus the model's attention on the fine-grained features and textures present in the source input image x. Finally, we resize the outpainted image to the source image size $H \times W$.

Generate Augmented Dataset. The process above is repeated until we generate the desired n_G number of outpainted images for a new dataset $\widetilde{S}$.

3. Experiments

3.1. GIMO implementation

We employ a pre-trained Stable Diffusion XL [7] with a ControlNet arm [22] that is conditioned for outpainting. All images in SmokeSeg are of dimensions 512×512. We obtain the desired out-painted image size using equation 1. We experiment and show results for $\lambda_{out\_dim} = 1.5, 2.0, 2.5$. For ControlNet, there are two main hyper-parameters that we optimised empirically: conditioning scale and number of inference steps. The conditioning scale dictates the strength of the model guidance conditioned on the input image where a lower value enables greater freedom in reworking the scene. As we want the diffusion model to be strongly conditioned on the source input image, we set it to its maximum value: 1.0. We show effects of various conditional scale values in Figure 3. We found that 15 denoising inference steps were ideal to achieve generated pixels that matched the texture and semantics of the original while keeping the inference time manageable. An NVIDIA RTX 3090 takes 9.96 seconds to outpaint one image.

SmokeSeg [14] is real dataset geared for early smoke segmentation. It has 3,355 *small*, 1442 *medium* and 547 *large* smoke images for training. We experiment with different

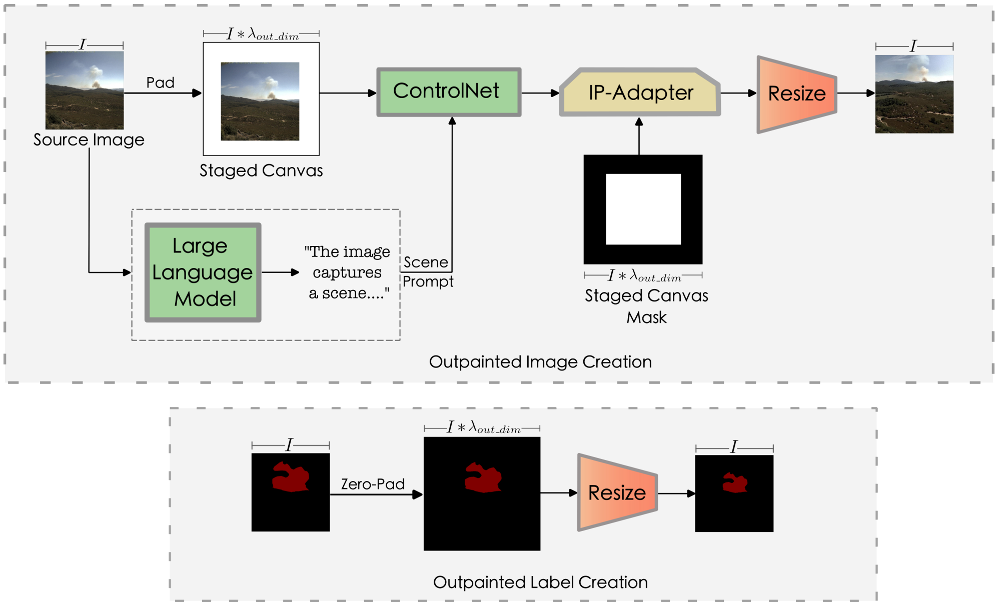

Figure 2. Illustration of our proposed GIMO pipeline. We obtain outpainted image after passing our staged canvas through a ControlNet Diffusion model and then refining it with an IPAdapter. The output is then resized to input image size. The corresponding segmentation label is obtained by zero-padding the original mask to outpainted size then resized to input image size.

	Small			Medium			Large			Total		
Method	$F_1 \uparrow$	$mIoU \uparrow$	$mMse \downarrow$	$F_1 \uparrow$	$mIoU \uparrow$	$mMse \downarrow$	$F_1 \uparrow$	$mIoU \uparrow$	$mMse \downarrow$	$F_1 \uparrow$	$mIoU \uparrow$	$mMse \downarrow$
FCN [6]	55.30	41.58	0.0013	70.72	57.31	0.0043	71.64	58.22	0.0233	65.41	51.89	0.0089
PSPNet [23]	55.27	42.01	0.0012	71.44	58.13	0.0042	72.15	58.54	0.0224	65.80	52.42	0.0086
EncNet [21]	58.09	44.86	0.0012	71.16	57.74	0.0042	71.54	57.96	0.0232	66.54	53.15	0.0088
DeepLabv3+ [2]	57.00	43.70	0.0013	73.34	59.80	0.0039	72.79	59.39	0.0219	67.26	53.85	0.0084
CCNet [4]	53.94	40.59	0.0015	69.79	56.26	0.0046	72.31	59.01	0.0231	64.45	51.42	0.0090
OCRNet [20]	53.60	41.04	0.0012	72.24	59.34	0.0039	72.25	59.16	0.0224	65.13	52.66	0.0085
SegFormer [12]	58.32	45.72	0.0017	74.34	61.37	0.0038	72.50	58.95	0.0235	67.99	54.98	0.0088
Trans-BVM [13]	60.64	47.01	0.0013	72.97	59.51	0.0040	71.77	58.79	0.0223	69.12	55.57	0.0085
FoSP [14]	72.74	59.46	0.0014	78.57	66.76	0.0043	82.29	71.26	0.0201	77.70	65.58	0.0073
FoSP + GIMO (Ours)	**74.06**	**61.27**	**0.0010**	**79.25**	**67.46**	**0.0042**	**82.36**	**71.48**	**0.0195**	**78.17**	**66.16**	**0.0075**

Table 1. Results showing the gain in performance using GIMO on our baseline FoSp compared against other methods. We show improvements across the board for all smoke sizes, with the performance for *small* smoke showing substantial increase.

outpainting schedules to achieve our aim of improving performance for small smoke segmentation. When our staged canvas undergoes the out-painting and then the refinement network, there are minute changes to the scene structure and/or texture.

We find that when outpainting then resizing images that already contain small smoke, results can sometimes distort the smoke region or remove parts of it (see Figure 4). However, this change in smoke shape is not reflected in its corresponding segmentation map. Hence, we only select *medium* and *large* images from the SmokeSeg for outpainting to ensure faithful scaling of smoke regions while preserving the boundary such that the corresponding segmentation still holds a valid annotation.

By introducing randomness regarding where the original image is placed on the staged canvas, and by virtue of the randomness in scene generation of diffusion models, we have the flexibility of reusing the same source image and obtaining a unique outpainted image every time. This allows us to scale our GIMO pipeline to obtain a dataset of desired size. We generate 3,555 additional small smoke images and combine that with the original SmokeSeg training set of 5,344 images to create our GIMO dataset.

3.2. Baselines and Evaluation Metrics.

We evaluate our GIMO on the test set of SmokeSeg [14] by training a recent smoke segmentation model, FoSp [14]. We download a local copy of its source code and train on GIMO while matching its recorded training settings. We

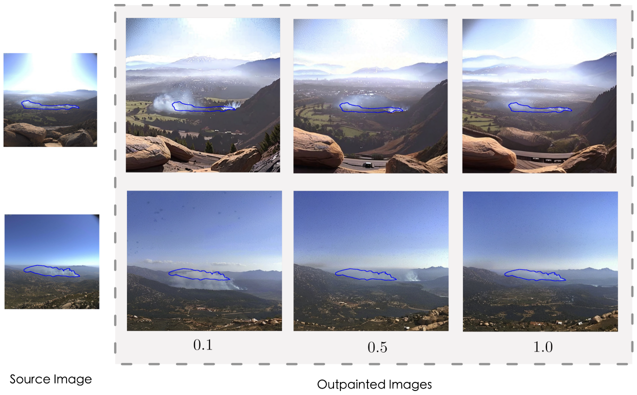

Figure 3. Figure showing effect of different conditioning scale for ControlNet generation guidance. A higher value produces more faithful out-painted image with the original smoke region intact.

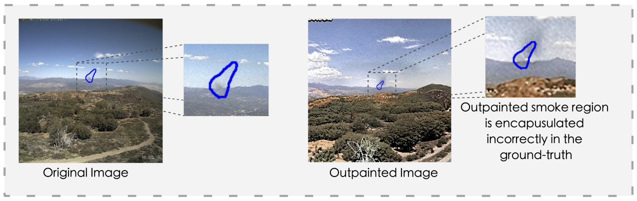
Original Image
Outpainted Image
Outpainted smoke region is encapusulated incorrectly in the ground-truth

Figure 4. Figure showing outpainting result on original *small* smoke image. The outpainted smoke ground-truth (right image) fails to properly preserve the original *small* smoke regions (left image).

examine the efficacy of our dataset on FoSP using their published results on SmokeSeg [14]. We adhere to the evaluation metrics used in the original papers of our baseline. We report results using mean mean squared error ($mMse$), mean Intersection over Union ($mIoU$), F-measure (F_1) for FoSp.

3.3. Main Results

In table 1, GIMO exhibits consistently significant improvements on our baseline FoSp. Training on GIMO allows FoSp to acheive a strong 3% improvement on $mIoU$ for small smoke images. Moreover, there are notable improvements for medium (+ 1.05%) and large (+0.3%) smoke categories as well despite SmokeSeg containing mostly small smoke images and GIMO focusing solely on small smoke. Overall, GIMO enables an $mIoU$ improvement of 0.9% averaged over the entire SmokeSeg dataset.

These results demonstrate the effectiveness of GIMO in improving performance for small smoke images. Furthermore, GIMO as a data augmentation pipeline has the flexibility that allows it to specifically target smoke images of a specific size in a truly scalable manner. The learned distribution of our out-painting and refining generative models that are used to fill in pixels within our canvas means that each out-painted image contains unique scene content.

3.4. Ablation Study

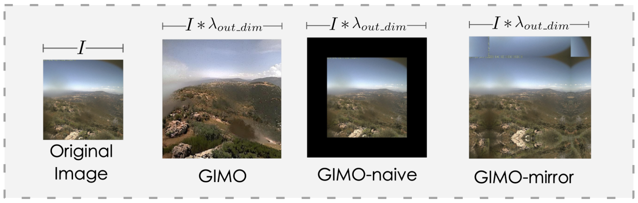

Figure 5. Figure showing our generative GIMO image compared with our two naive baselines: GIMO-naive (which simply pads the borders) and GIMO-mirror (which mirrors the pixels across the borders).

Naive Baselines. To demonstrate the importance of incorporating generated synthetic pixels onto our staged canvas, we prepare naive versions of GIMO pipeline. As the simplest case, we use the staged canvas, as is, with zero pixel values across the padded pixel region. We refer to this version as GIMO-naive. Additionally, we prepare a version of GIMO where the source image pixels are mirrored across the borders and refer to this as GIMO-mirror (See figure 5 for illustration). We maintain dataset size across all versions. We train FoSp on all versions of our baseline datasets and show results in Table 2. We expect the best performance when the input image data is consistent with real world scenes. Hence our naive baselines perform worse than our generative out-painted dataset as zero-padding or pixel-mirroring represent real world smoke images.

Dataset	F_1	$mIoU$	$mMse$
GIMO-naive	70.01	57.67	0.0011
GIMO-mirror	70.03	58.67	0.0011
GIMO	**74.06**	**61.27**	**0.0010**

Table 2. Results comparing naive baselines with our generative GIMO dataset on small test set of Smoke-Seg. All models are trained on FoSp for equal number of epochs.

4. Conclusion

We present a scalable generative data augmentation pipeline that strongly improves performance by over 3% for segmentation of small smoke images corroborating our hypothesis that there is a tangible lack of sufficient training data in this domain. By leveraging pre-trained generative diffusion models to enlarge existing image scenes, we are able to efficiently generate unique smoke images while preserving the original segmentation labels. Our outpainting approach, while effective in enlarging scenes and targeting various smoke region sizes, inherently preserves the original smoke shape and consistency thereby limiting the diversity of generated smoke instances. This limitation motivates the exploration of a more powerful approach in the future: the creation of entirely synthetic smoke images using generative diffusion models.

References

[1] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. 2

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[3] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 1

[4] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 3

[5] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. *arXiv preprint arXiv:2104.03963*, 2021. 1

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[9] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[10] Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, and Yi Yang. Sketch-guided scenery image outpainting. *IEEE Transactions on Image Processing*, 30:2643–2655, 2021. 1

[11] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022. 2

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 3

[13] Siyuan Yan, Jing Zhang, and Nick Barnes. Transmission-guided bayesian generative model for smoke segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3009–3017, 2022. 1, 3

[14] Lujian Yao, Haitao Zhao, Jingchao Peng, Zhongze Wang, and Kaijie Zhao. Fosp: Focus and separation network for early smoke segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6621–6629, 2024. 1, 2, 3, 4

[15] Lujian Yao, Haitao Zhao, Jingchao Peng, Zhongze Wang, and Kaijie Zhao. Dsa: Discriminative scatter analysis for early smoke segmentation. In *European Conference on Computer Vision*, pages 467–484. Springer, 2025. 1

[16] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[17] Marta Yebra, Nick Barnes, Colleen Bryant, Geoffrey J Cary, Salman Durrani, Jia-Urnn Lee, David Lindenmayer, Robert Mahony, Roslyn Prinsley, Philippa Ryan, et al. An integrated system to protect australia from catastrophic bushfires. *TheAustralian Journal of Emergency Management*, 36 (4):20–22, 2021. 1

[18] Feiniu Yuan, Lin Zhang, Xue Xia, Boyang Wan, Qinghua Huang, and Xuelong Li. Deep smoke segmentation. *Neurocomputing*, 357:248–260, 2019. 1

[19] Feiniu Yuan, Kang Li, Chunmei Wang, and Zhijun Fang. A lightweight network for smoke semantic segmentation. *Pattern Recognition*, 137:109289, 2023. 1

[20] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 3

[21] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 3

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 2

[23] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3