FoSp: Focus and Separation Network for Early Smoke Segmentation

Lujian Yao, Haitao Zhao*, Jingchao Peng, Zhongze Wang, Kaijie Zhao

East China University of Science and Technology
{lujianyao, zzwang, kjzhao}@mail.ecust.edu.cn, haitaozhao@ecust.edu.cn, starry-sky@outlook.com

Abstract

Early smoke segmentation (ESS) enables the accurate identification of smoke sources, facilitating the prompt extinguishing of fires and preventing large-scale gas leaks. But ESS poses greater challenges than the conventional object and regular smoke segmentation due to its small scale and transparent appearance, which can result in high miss detection rate and low precision. To address these issues, a *Focus and Separation Network* (FoSp) is proposed. We first introduce a focus module employing bidirectional cascade which guides low-resolution and high-resolution features towards mid-resolution to locate and determine the scope of smoke, reducing the *miss detection rate*. Next, we propose a separation module that separates smoke images into a pure smoke foreground and a smoke-free background, enhancing the contrast between smoke and background fundamentally, and improving segmentation *precision*. Finally, a domain fusion module is developed to integrate the distinctive features of the two modules which can balance recall and precision to achieve high F_β. Furthermore, to promote the development of ESS, we introduce a high-quality real-world dataset called SmokeSeg, which contains more small and transparent smoke images than the existing datasets. Experimental results show that our model achieves the best performance on three available smoke segmentation datasets: SYN70K ($mIoU$: 83.00%), SMOKE5K (F_β: 81.6%) and SmokeSeg (F_β: 72.05%). The code can be found at https://github.com/LujianYao/FoSp.

Introduction

In wildlife, smoke is an important indicator of fire. Early smoke segmentation (ESS) enables rapid identification of the location of the fire (Muhammad, Ahmad, and Baik 2018; Robinson 1979), facilitating the timely extinguishing of the flames by rescue personnel and preventing the occurrence of large fires. In industrial production, ESS can also aid in promptly detecting the location of gas leaks and prevent the spread of toxic and harmful gases (Hsu et al. 2021).

Smoke segmentation is usually defined as a binary segmentation task (*e.g.*, salient object detection (Borji et al. 2015; Qin et al. 2019), small object segmentation (Wang, Zhou, and Wang 2019; Liu et al. 2021)), but the task of ESS is particularly challenging due to three main reasons: ① As a small non-rigid object, early smoke exhibits strong variability, resulting in diverse patterns due to differences in scenes, lighting conditions, and wind intensities. ② As a transparent object, the low contrast will make it difficult to distinguish between the smoke and the background. ③ Unlike other transparent object detection (e.g., glass segmentation (Mei et al. 2022; Yu et al. 2022)), the transparency of early smoke is variable. The conventional smoke image only has a transparent part at the edge, but the early smoke also has transparency in the main body.

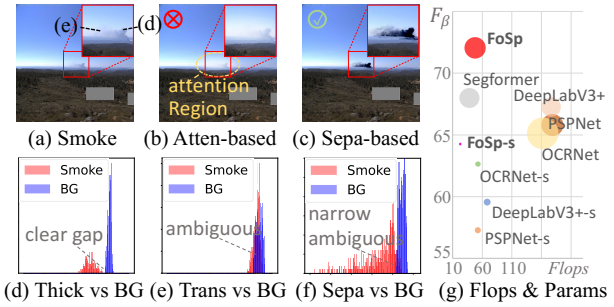

Figure 1: Motivation and Comparison. In this figure, "Sepa" means "Separation" and "BG" means "Background".

The thick part of the smoke behaves like common objects, with distinct edges and noticeable color. However, transparent smoke has low contrast with the background, making it difficult to distinguish. From the perspective of pixel distribution, Fig. 1d illustrates a gap between the thick parts of smoke and the background, making it relatively easy to segment. Conversely, Fig. 1e shows that the distribution of transparent smoke has a large ambiguous overlapping region with the surrounding background. The larger the overlapping region is, the more difficult to distinguish them.

Previous regular smoke segmentation methods (Li et al. 2018; Yuan et al. 2019a,b, 2022) primarily emphasize larger receptive fields to cope with the variability and blurred edges of smoke. However, directly applying binary segmentation can result in high miss detection (incomplete segmentation) and low precision (rough-edge segmentation) as the early smoke is small and transparent.

*Corresponding author

Therefore, we propose a *Focus and Separation Network* (FoSp) to deal with these two problems separately. Firstly, we design a focus module utilizing bidirctional cascade which directs low-resolution and high-resolution features towards mid-resolution to locate and confirm the scope of smoke, effectively reducing the miss detection rate (for ①). Then, a separation module is introduced to enhance the contrast between the smoke foreground and the background, thereby improving precision (for ②). Previous methods (Cao, Tang, and Lu 2022; Jing, Meng, and Hou 2023) have used attention-based local enhancement, but this approach may fail to increase contrast (compare the origin smoke region in Fig. 1a and the attention region in Fig. 1b). Therefore, we refer to the atmospheric scattering model, estimate the original image directly, and use it to calculate a discrepancy to obtain the foreground image. Fig. 1c shows that the separation-based method can precisely enhance the contrast between the smoke and the background. Finally, to prevent overpowering one of the modules (the model tends to over-optimize recall or precision), we further design the fusion module to balance the performance of the two modules to obtain the general F_β (for ③).

Furthermore, we provide a large and real-world dataset called SmokeSeg, which includes 6,144 real smoke images, with a considerable number of them featuring small and transparent smoke. SmokeSeg currently boasts the largest number of real images among publicly available smoke segmentation datasets. Notably, the SmokeSeg contains 4.5 times more real images than SMOKE5K (Yan, Zhang, and Barnes 2022), which only comprises 1,360 real images.

Our main contributions can be summarized as follows:

- We propose a Focus and Separation Network (FoSp), where the focus module and separation module are introduced to reduce the miss detection rate and improve the precision of early smoke segmentation, respectively.
- We have created a large-scale dataset named *SmokeSeg* for early smoke segmentation, which includes 6144 real images with pixel-wise annotations.
- Our FoSp achieve the best performance on three smoke segmentation datasets: SYN70K ($mIoU$: 83.00%), SMOKE5K (F_β: 81.6%) and SmokeSeg (F_β: 72.05%).

Related Work

Early Smoke Segmentation. Although current semantic segmentation methods (Long, Shelhamer, and Darrell 2015; Lin et al. 2017; Chen et al. 2018; Strudel et al. 2021; Xie et al. 2021; Cheng et al. 2022) are effective at segmenting regular objects (i.e., those with clear outlines and roughly the same shape), these generic algorithms are not suitable for early smoke segmentation due to its varying transparency and small scale. Traditional smoke segmentation methods have mainly focused on extracting high-quality color and texture features (Mahmoud and Ren 2019; Xing et al. 2015; Yuan, Liu, and Zhang 2019), and deep-learning-based methods (Yuan et al. 2022, 2019a,b; Jing, Meng, and Hou 2023) mainly focus on extracting features with larger receptive fields to handle the variability and blurred edges of smoke. However, these methods suffer from high miss detection rate and low precision in segmenting small and transparent early smoke. And many of them (Yuan et al. 2019a,b, 2021) have only been quantitatively evaluated on synthetic datasets. Despite some methods (Jia et al. 2019; Wang et al. 2014) claiming to address early smoke, they still rely on images of regular smoke. Therefore, the existing smoke segmentation methods are not capable of effectively addressing the issue of early smoke.

Prior Attention. Prior attention is a mechanism that adds a branch to the network before the backbone (Uzkent, Yeh, and Ermon 2020; Wang et al. 2020; Xie et al. 2020) or before the final predictions (Najibi, Singh, and Davis 2019; Yang, Huang, and Wang 2022), making the network focus on a specific area in early to obtain finer predictions. However, current methods are mostly used for object detection and are designed to improve *precision*. How to obtain prior attention for segmentation with high *recall* is still to be resolved.

Smoke Segmentation Datasets. Currently, there are two main smoke segmentation datasets: SYN70K (Yuan et al. 2019b) and SMOKE5K (Yan, Zhang, and Barnes 2022). SYN70K is a synthetic dataset comprising 70K images, while SMOKE5K contains 1K real images and 4K synthetic images, with the latter being selected from SYN70K. However, neither of these datasets is specifically designed for early smoke segmentation. The SYN70K comprises entirely of synthetic smoke, which is larger in scale and significantly different from real images. The SMOKE5K has only a small number of real smoke images and a low proportion of early small and transparent smoke. Therefore, there is currently no dataset particularly designed for early smoke segmentation.

Method

Introduction of Focus and Separation (FoSp)

Intuition. We assume that the image pixel $\boldsymbol{i}(x) \in \mathbb{R}^3$ is composed of a smoke-free background component $\boldsymbol{b}(x) \in \mathbb{R}^3$ and a smoke foreground component $\boldsymbol{s}(x) \in \mathbb{R}^3$:

$$\boldsymbol{i}(x) = \boldsymbol{b}(x)\boldsymbol{t}(x) + \boldsymbol{s}(x)(1 - \boldsymbol{t}(x)), \tag{1}$$

where x represents the position of the pixel and $\boldsymbol{t}(x) \in \mathbb{R}^3$ is used to control the density of the smoke foreground in RGB channels. Eq. 1 can be transformed to:

$$\begin{aligned} \boldsymbol{s}(x) &= \frac{\boldsymbol{i}(x) - \boldsymbol{b}(x)\boldsymbol{t}(x)}{1 - \boldsymbol{t}(x)} \\ &= \frac{\boldsymbol{i}(x) - \boldsymbol{b}(x) + \boldsymbol{b}(x) - \boldsymbol{b}(x)\boldsymbol{t}(x)}{1 - \boldsymbol{t}(x)} \\ &= \frac{1}{1 - \boldsymbol{t}(x)}(\boldsymbol{i}(x) - \boldsymbol{b}(x)) + \boldsymbol{b}(x) \\ &= \frac{1}{\boldsymbol{\alpha}(x)}(\boldsymbol{i}(x) - \boldsymbol{b}(x)) + \boldsymbol{b}(x) \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}(x) = 1 - \boldsymbol{t}(x)$. Supposing we can obtain a smoke-free background component $\boldsymbol{b}(x)$, we are capable of adjusting the smoke foreground component $\boldsymbol{s}(x)$ by controlling the value of $\boldsymbol{\alpha}$ to enhance the contrast between foreground and background. Following such intuition, we propose a *Focus and Separation Network* (FoSp) to estimate the background pixel among the smoke region.

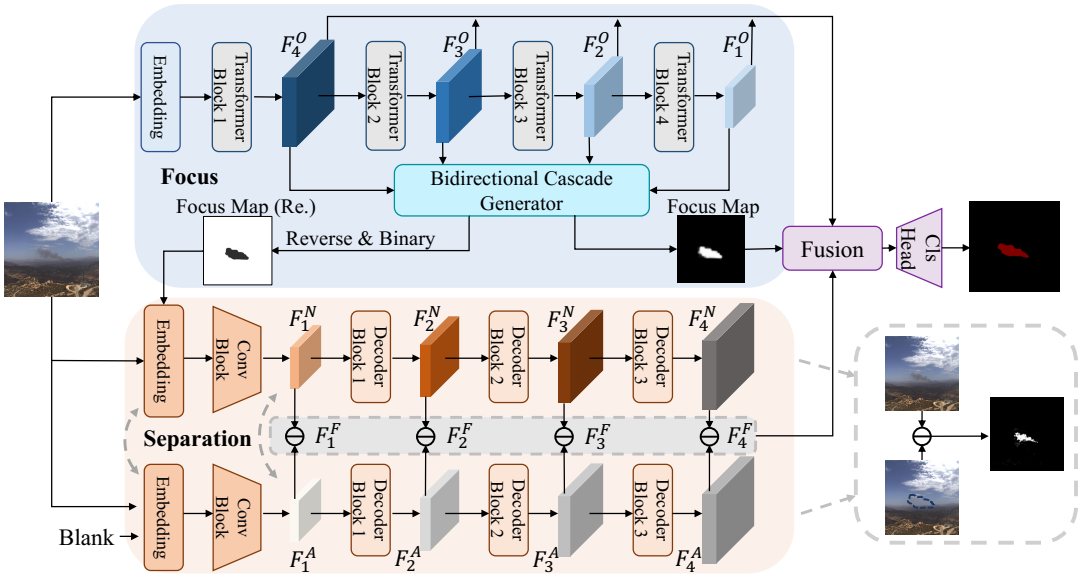

Figure 2: Structure illustration of our Focus and Separation Network (FoSp). The focus module extracts image features and generates a focus map (FM) using a bidirectional cascade generator (BCG). The separation module consists of two inpainters: one to complete the smoke areas with smoke-free backgrounds using the original image and FM, and the other using the original image and a blank image. The smoke foreground features are obtained by subtracting the two features, and then the origin features, FM, and foreground features are fused in the fusion module to obtain the final prediction.

Overview. As shown in Fig. 2 , we first propose a focus module to locate and confirm the scope of smoke. This enables the network to pay closer attention to the smoke and reduce the miss detection rate. Next, a separation module is introduced to split the smoke image into pure smoke foreground and smoke-free background, which enhances the contrast between the foreground and background, thereby improving precision. Finally, a domain fusion module is developed to narrow the domain gap between the features generated by the first two modules, ultimately achieving a balance between recall and precision.

Focus Module

Intuition. Most segmentation methods use pyramid-based *unidirectional* feature integration to obtain a more refined prediction. However, our focus module is designed to achieve a complete smoke area. In this regard, we introduce a novel *bidirectional* feature cascade approach. We consider the smoke as two parts: the low-opacity smoke body and the high-opacity smoke edge. As shown in Fig. 3, low-resolution features can provide an approximate outline of the smoke body, However, the effectiveness of capturing transparent regions of smoke is compromised. Conversely, high-resolution features offer more precise attention and can identify transparent parts of the smoke edge, but cannot fully capture the entire smoke. Therefore, we propose a bidirectional fusion of features that can integrate features from both low and high resolution, resulting in a complete scope of the smoke.

Overview. In the focus module, a bidirectional cascade generator (BCG) is proposed to integrate the original multi-scale image features for locating and confirming the range of smoke, which we refer to as the focus map (FM). As shown in Fig. 3, BCG bidirectionally guides low-resolution features and high-resolution features towards mid-resolution to obtain a complete scope of smoke.

Detail. As shown in Fig 2, in the feature extraction section, MiT-B3 (Xie et al. 2021) is adopted as our backbone, which is a transformer-based backbone with the large receptive field that can handle the variability of smoke. Specifically, we preprocess the image using an embedding layer and extract features at four different scales using four Transformer blocks. These scales range from small to large and are respectively $\mathbf{F}_i \in \mathbb{R}^{\frac{H}{2^{6-i}} \times \frac{W}{2^{6-i}} \times C_i} (i = 1, 2, 3, 4)$. The BCG is illustrated in Fig. 3, which consists of two parts: Low-Mid Cascade (LMC) and High-Mid Cascade (HMC). The two modules are similar, so we will use the LMC module as an example to explain. In LMC, The lowest-resolution feature $\boldsymbol{F}_1 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 1}$ is fed into a layer of Conv to obtain the logits $\boldsymbol{G}_1 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 1}$ and then applied the Sigmoid function to obtain the prediction $\boldsymbol{Q}_1 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 1}$ of the $\boldsymbol{F}_1$. A query-guide approach is adopted to enhance the feature maps by taking the dot product of $\boldsymbol{Q}_1$ and $\boldsymbol{F}_1$, and adding the result to the $\boldsymbol{F}_1$. The cascade process of obtaining the final output logits $\boldsymbol{G}_2 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 1}$ of LMC can be represented using the following equation:

$$\boldsymbol{G}_2 = \text{Conv}(\text{Cat}(\text{U}(\boldsymbol{Q}_1 * \boldsymbol{F}_1 + \boldsymbol{F}_1), \boldsymbol{F}_2), \theta_2), \tag{3}$$

where the θ_2 is the learnable parameters of the Conv and the U refers to the upsampling operation. The HMC follows the same procedure to obtain the final logits $\mathbf{G}_3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 1}$. At last, the two intermediate resolution logits $\mathbf{G}_2$ (LMC) and $\mathbf{G}_3$ (HMC) are fused to obtain the final focus map (FM)

(FM $\in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 1}$):

$$\text{FM} = \text{Sigmoid}(\text{Conv}(\text{Cat}(\boldsymbol{G}_2, \boldsymbol{G}_3), \theta_m)), \tag{4}$$

where the θ_m is the learnable parameters of the Conv.

Separation Module

Intuition. The region of FM shares similarities with the adjacent pixels, while the region of smoke can be perceived as an external element that is not compatible with the background. We leverage the contextual features surrounding FM to complement the features within the FM area, thereby aligning them with the neighboring features (i.e., smoke-free background). Consequently, by subtracting the original image, the foreground smoke can be extracted.

Overview. In the separation module, we utilize the FM to obtain the smoke region and apply the inpainting technique (a.k.a Image completion) to fill the smoke area with a smoke-free background using the surrounding scene (particularly the area around the smoke) as a reference. We then subtract the smoke-free background from the original smoke region to derive the pure smoke foreground. Furthermore, to preserve information and maintain consistency with high-level segmentation tasks, we separate the foreground and background at the feature level.

Detail. As shown in Fig. 2, In order to separate the foreground and background of smoke images at the feature level, the entire separation module is composed of two weight-shared inpainters.

1) In the top branch, the original image $\boldsymbol{I}$ and focus map **FM** is first input into the embedding layer EB$(\cdot)$ for feature integration and then encoded into a latent feature $\boldsymbol{F}_1^N \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_1}$ through the Conv block.

$$\boldsymbol{F}_1^N = \text{Conv}(\text{EB}(\boldsymbol{I}, \text{FM}), \theta_I), \tag{5}$$

where θ_I represents the pre-trained Conv parameters of the inpainting network.

2) Different from the top branch, in the bottom branch, the input is the original image $\boldsymbol{I}$ and a $\boldsymbol{Blank}$ image, without any mask.

$$\boldsymbol{F}_1^A = \text{Conv}(\text{EB}(\boldsymbol{I}, \boldsymbol{Blank}), \theta_I), \tag{6}$$

where $\boldsymbol{F}_1^A \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_1}$.

3) Then the latent features of the two branches are sent to three decoder blocks and three different scale inpainting-network-domain feature maps are obtained.

$$\boldsymbol{F}_i^{A/N} = \text{Dec}_{i-1}(\boldsymbol{F}_{i-1}^{A/N}), i = 2, 3, 4, \tag{7}$$

where $\boldsymbol{F}_i^{A/N} \in \mathbb{R}^{\frac{H}{2^{6-i}} \times \frac{W}{2^{6-i}} \times C_i}$.

4) Finally, we calculate the L^1-norm of the four scale feature maps in both the top branch and the bottom branch to get the foreground features $\boldsymbol{F}_i^F$, forming a feature list and feeding it into the domain fusion module.

$$\boldsymbol{F}_i^F = \beta * \|\boldsymbol{F}_i^N - \boldsymbol{F}_i^A\|_1, i = 1, 2, 3, 4 \tag{8}$$

where β is the gain factor and we set $\beta = 10$.

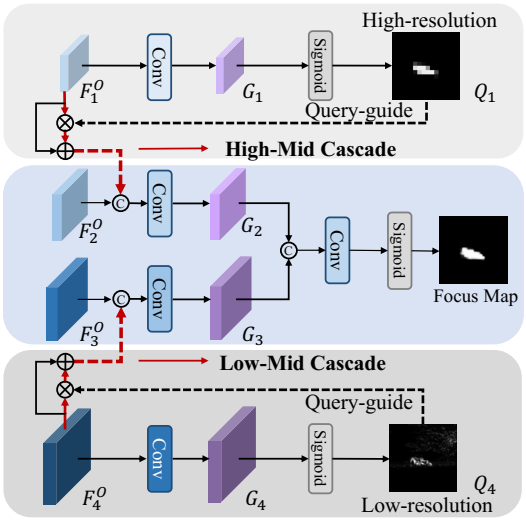

Figure 3: Structure of bidirectional cascade generator.

Domain Fusion

Intuition. Due to the foreground features generated by the inpainters, the origin features and foreground features actually belong to different feature domains. Simply adding features does not effectively merge the high-quality features of both, and this operation may even damage the original feature domain due to different information represented by different channels. Therefore, we design a domain fusion module to achieve the fusion of both domain features.

Detail. As shown in Fig. 4, the domain fusion module is divided into three domains: origin domain, foreground domain, and fusion domain. The origin domain is composed of features at different scales generated by the focus module, and we utilize the focus map to enhance the features of the smoke region. The foreground domain is composed of four distinct scale foreground features generated by the separation module. In each stage, the features of the two domains and the feature of *previous fusion stage* are concatenated and then fused with an MLP. The features of the two domains are hierarchically merged in this way.

Loss Function

The loss function consists of two parts: basic loss and *Focus* loss. In conjunction with our proposed focus module, we introduce a *Focus* loss, which provides supervisory information to the prediction logits $\boldsymbol{G}_i (i = 1, 2, 3, 4)$ of multiple resolution feature maps.

$$\mathcal{L}_{focus} = \lambda_f \mathcal{L}_{BCE}(\text{FM}, \mathcal{Y}) + \sum_{i=1}^{4} \lambda_i (\mathcal{L}_{BCE}(\boldsymbol{G}_i, \mathcal{Y})), \tag{9}$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy, $\mathcal{Y}$ is the ground truth, λ_f and λ_i are the weighting factors and we set $\lambda_f = \lambda_i = 0.1$.

For the basic segmentation loss, we use the standard binary cross-entropy. The final objective function is the sum of the two losses:

$$\mathcal{L} = \mathcal{L}_{focus} + \lambda_b \mathcal{L}_{BCE}(\boldsymbol{P}, \mathcal{Y}), \tag{10}$$

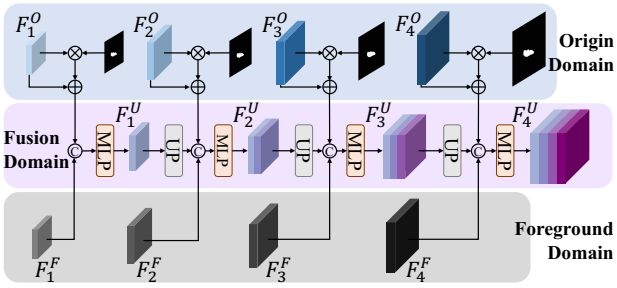

Figure 4: Structure of domain fusion module.

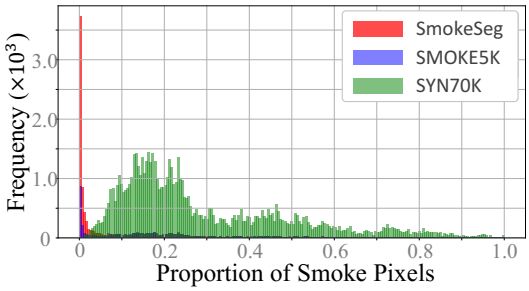

Figure 5: Distribution of smoke pixel ratio in an image of three datasets. This indicates that SmokeSeg contains a much higher proportion of small smoke images than the other two datasets.

where $\boldsymbol{P}$ is the final prediction of model, λ_b is the weighting factor and we set $\lambda_b = 0.5$.

Experiments

SmokeSeg Dataset

There are two main smoke segmentation datasets available: SYN70K (Yuan et al. 2019b) and SMOKE5K (Yan, Zhang, and Barnes 2022). SYN70K comprises 70k synthetic images, while SMOKE5K is a dataset containing 1,360 real images and 4k synthetic images, with the latter being selected from SYN70K. However, neither of these datasets is tailored specifically for *early smoke segmentation*. SYN70K is a synthetic dataset with lots of images of smoke, but the smokes in it are large and prominent, which is quite different from real smoke. SMOKE5K has a limited number of real smoke images, and a low proportion of these are early smoke images, making it difficult to conduct experiments specifically targeting early smoke segmentation.

To address these issues, we contribute *SmokeSeg* dataset. SmokeSeg consists of 6,144 real images (the raw smoke images are sourced from FigLib (Dewangan et al. 2022).), which has the largest number of real images with pixel-wise annotations in any publicly available smoke segmentation dataset. The number of real images in SmokeSeg is 4.5 times larger than that of SMOKE5K, which only comprises 1,360 real images. The majority of these images in SmokeSeg are early smoke, characterized by small scale and transparent appearance. Fig. 5 shows the distribution of smoke pixel ratios in three datasets, with smaller ratios indicating smaller smoke. It can be seen that the SYN70K dataset hardly contains small smoke images, while SMOKE5K only has a small fraction of them. In contrast, the majority of our SmokeSeg dataset is composed of small smoke images.

Implementation Details

Datasets. We conduct experiments on three large smoke datasets: SYN70K (Yuan et al. 2019b), SMOKE5K (Yan, Zhang, and Barnes 2022) and our SmokeSeg. To make a fair comparison, we train our models on SYN70K and SMOKE5K and test them on their respective test sets. Then we provide a new benchmark on our SmokeSeg dataset.

Evaluation Metrics. Following previous smoke segmentation literature (Yan, Zhang, and Barnes 2022; Yuan et al. 2019b), we evaluate our model utilizing the mMse ($\mathcal{M}$) and $mIoU$ metric on SYN70K and adopt mMse ($\mathcal{M}$) and F-measure (F_β) on SMOKE5K. For our *SmokeSeg*, we employ the *three metrics* mentioned above (F_β, $mIoU$, $\mathcal{M}$) to comprehensively evaluate the performance of the model.

Training Details. We implement our FoSp on MMSegmentation with a single NVIDIA RTX 3090Ti GPU. Each image is resized to 512×512. Random crop and random flip are adopted during the training. We use the AdamW (Loshchilov and Hutter 2017) optimizer and set the learning rate to 6e-5 with 0.01 weight decay. We train 40k iterations on SMOKE5K and SmokeSeg, and 80k iterations on SYN70K, with all batch sizes set to 6.

Comparison with State-of-the-Art Methods

SmokeSeg. As there is no complete open-source smoke segmentation code, we have chosen several strong baselines (Chen et al. 2018; Zhao et al. 2017; Yuan, Chen, and Wang 2020; Zhang et al. 2018; Xie et al. 2021) that perform well in semantic segmentation for comparison. In order to explore the abilities of various methods for the smoke of different scales, especially for early smoke, we divide the test set into three parts based on the proportion of smoke pixels in an image, namely small, medium, and large, and test them separately. We use the evaluation metric of "small" to measure the early smoke segmentation capability of our model.

$$\begin{cases} \text{Small} : \delta < 0.5\% \\ \text{Medium} : 0.5\% < \delta < 2.5\% \\ \text{Large} : \delta > 2.5\% \end{cases}, \tag{11}$$

where δ is the smoke pixel ratio in an image.

As shown in Table 1, the comparison is divided into two parts, with the top part of the table showing the comparison of methods with lightweight backbones, and the bottom part showing the comparison of methods with medium-sized backbones. From the comparison of the lightweight methods, our FoSp using the least number of parameters surpasses other methods with higher parameter counts in the vast majority of cases. In the comparison of the medium-sized methods, our FoSp significantly outperforms all other methods, especially in the comparison of *small smoke*, where our method is 7.71% higher than the second-best

	Total			Small			Medium			Large			
Method	$F_\beta \uparrow$	$mIoU \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$mIoU \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$mIoU \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$mIoU \uparrow$	$\mathcal{M} \downarrow$	Params
FCN-s (2015)	54.12	42.06	0.0105	33.93	25.00	0.0017	64.08	50.65	0.0051	67.24	52.96	0.0271	9.71
DeepLabv3+-s (2018)	59.56	46.71	0.0095	44.06	32.77	0.0014	68.35	54.76	0.0046	68.37	54.46	0.0246	15.23
PSPNet-s (2017)	57.28	44.76	0.0096	40.86	30.70	0.0015	63.77	50.09	0.0051	69.72	55.68	0.0245	13.61
OCRNet-s (2020)	62.64	50.11	**0.0087**	50.99	39.47	0.0013	71.47	55.09	0.0043	70.45	**57.33**	**0.0225**	12.07
SegFormer-s (2021)	60.73	48.19	0.0097	48.21	35.18	0.0015	70.22	56.81	0.0042	67.52	54.18	0.0256	6.38
FoSp-s (ours)	**64.26**	**51.44**	0.0089	**51.35**	**39.57**	**0.0012**	**72.47**	**58.95**	**0.0041**	**70.60**	57.31	0.0234	5.79
FCN (2015)	65.41	51.89	0.0089	55.30	41.58	0.0013	70.72	57.31	0.0043	71.64	58.22	0.0233	49.48
PSPNet (2017)	65.80	52.42	0.0086	55.27	42.01	0.0012	71.44	58.13	0.0042	72.15	58.54	0.0224	48.96
EncNet (2018)	66.54	53.15	0.0088	58.09	44.86	0.0012	71.16	57.74	0.0042	71.54	57.96	0.0232	35.87
DeepLabv3+ (2018)	67.26	53.85	0.0084	57.00	43.70	0.0013	73.34	59.80	0.0039	72.79	59.39	0.0219	43.58
CCNet (2019)	64.45	51.42	0.0090	52.94	40.59	0.0015	69.79	56.26	0.0046	72.31	59.01	0.0231	49.81
OCRNet (2020)	65.13	52.66	0.0085	52.60	41.04	0.0012	72.24	59.34	0.0039	72.25	59.16	0.0224	70.37
SegFormer (2021)	67.99	54.98	0.0088	58.32	45.72	0.0017	74.34	61.37	0.0038	72.50	58.95	0.0235	44.60
FoSp (ours)	**72.05**	**59.03**	**0.0079**	**66.03**	**52.68**	**0.0010**	**75.90**	**62.99**	**0.0037**	**74.98**	**62.24**	**0.0208**	47.52

Table 1: Performance on the test set of SmokeSeg. The best results: bold. The second-best results: underline.

	DS01		DS02		DS03	
Method	$\mathcal{M}$	$mIoU$	$\mathcal{M}$	$mIoU$	$\mathcal{M}$	$mIoU$
SMD	0.32	62.88	0.34	61.50	0.33	62.09
TBFCN	0.30	66.67	0.32	65.85	0.31	66.20
LRN	0.31	66.43	0.31	67.71	0.30	67.46
ESPNet	-	61.85	-	61.90	-	62.77
LKM	0.27	75.82	0.28	74.93	0.27	75.39
RefineNet	0.25	77.16	0.26	76.75	0.25	77.52
PSPNet	0.24	78.71	0.25	78.01	0.24	78.39
CCL	0.23	78.87	0.25	77.95	0.24	78.55
DFN	0.23	80.87	0.24	79.90	0.23	80.60
DSS	0.27	71.04	0.29	70.01	0.29	69.81
W-Net	0.27	73.06	0.25	73.97	0.26	73.36
CCENet	-	74.67	-	75.24	-	76.01
Trans-BVM	0.12	-	0.14	-	0.13	-
FoSp (ours)	**0.10**	**83.00**	**0.11**	**81.81**	**0.10**	**82.80**

Table 2: Performance on three test sets of SYN70K.

SegFormer (Xie et al. 2021) on F_β, demonstrating that our method can better recognize early smoke. Fig. 1g shows the parameters (the size of the bubble) and computational complexity of different models.

SYN70K. We evaluate our model on their three synthetic test sets (DS01, DS02, and DS03). As shown in Table 2, our model outperforms the previous state-of-the-art method Trans-BVM on all three test sets.

SMOKE5K. We evaluate our model on their test set of 400 real smoke images. As demonstrated in Table 3, our approach surpasses previous state-of-the-art Trans-BVM and achieves 2.5% improvement on F_β.

Ablation Study

To further investigate the role of each module in our proposed FoSp, we conduct ablation experiments on the SmokeSeg dataset. As shown in Table 4, we utilize SegFormer (Xie et al. 2021) as the baseline for our study. In order to investigate the influence of individual modules on the miss detection rate and precision of early smoke detection, we augment the evaluation metrics by including recall (for miss detection rate) and precision. F_β is adopted as the comprehensive evaluation metric.

Method	F_β	$\mathcal{M}$
F3Net (Wei, Wang, and Huang 2020)	67.0	0.004
BASNet (Qin et al. 2019)	73.3	0.005
SCRN (Wu, Su, and Huang 2019)	76.9	0.003
ITSD (Zhou et al. 2020)	77.4	0.003
UCNet (Zhang et al. 2020)	78.7	0.003
Trans-BVM (Yan, Zhang, and Barnes 2022)	79.1	0.002
FoSp (ours)	**81.6**	**0.002**

Table 3: Performance on the test set of SMOKE5K.

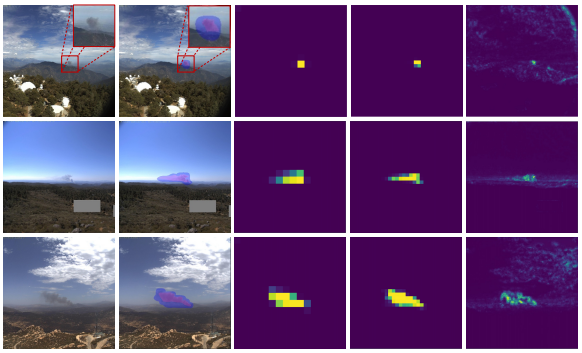

Figure 6: Qualitative results of focus module. The blue region represents the focus map (FM), while the red denotes the ground truth (GT).

Effect of Focus Module. Methods (b) and (c) in Table 4 demonstrate the role of the focus loss and focus module. It can be observed that after incorporating the focus loss and focus module, a substantial enhancement on recall is observed, particularly in the segmentation of small smoke particles. As shown in Fig. 6, high-resolution features can capture fine details of the smoke edges, whereas low-resolution features can capture the overall characteristics of the smoke body. By establishing bidirectional cascade between these two types of features, a focus map can be generated to fully

Module						Metrics			Metrics (small)		
Method	Baseline	$\mathcal{L}_{focus}$	Focus	Separation	Fusion	Recall	Precision	F_β	Recall	Precision	F_β
(a)	✓					71.77	71.34	67.99	60.82	63.29	58.32
(b)	✓	✓				73.53	71.55	68.90	65.06	63.06	60.99
(c)	✓	✓	✓			**75.60**	71.50	70.00	**68.52**	62.86	62.23
(d)	✓	✓	✓	✓		73.82	**75.50**	71.50	66.24	**71.16**	65.77
(e)	✓	✓	✓	✓	✓	74.60	75.20	**72.05**	68.24	68.56	**66.03**

Table 4: Ablation study of our FoSp.

Method	F_β	$mIoU$	$\mathcal{M}$
SegFormer	67.99	54.98	0.0088
Foreground only	68.22	54.94	0.0087
FoSp (image-level)	71.04 ↑ 3.05	57.81 ↑ 2.83	0.0082
FoSp (feature-level)	**72.05** ↑ 4.06	**59.03** ↑ 4.05	**0.0079**

Table 5: Experiments on different separation levels.

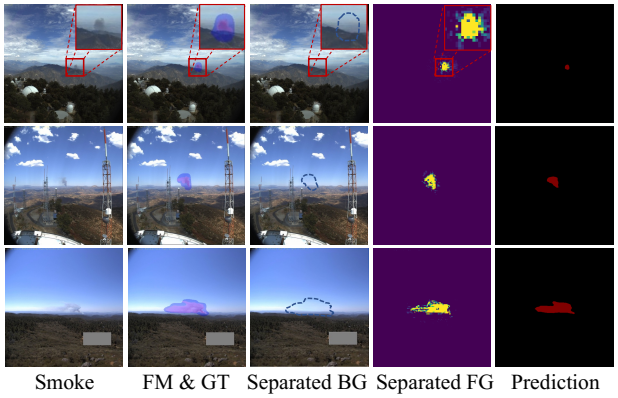

Figure 7: Qualitative results of separation module.

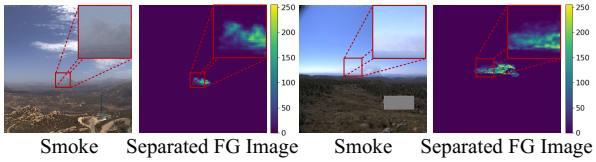

Figure 8: Qualitative results of image-level separation.

cover the smoke area and suppress background interference.

Effect of Separation Module. Method (d) in Table 4 demonstrates the role of the separation module. This method of separating the *smoke foreground* from a smoke-free background results in a notable increase in precision, particularly in the segmentation of small smoke particles, where precision improved by 7.87% compared to the baseline and by 8.30% compared to the focus module only. The foreground feature in Fig. 7 demonstrates that the extracted smoke foreground feature closely resembles the shape of the smoke itself, particularly the fine edges of the smoke. To further explore the role of the foreground features, we conduct an experiment in which only the foreground features are used for prediction in the decoder head. As shown in Table 5, the "Foreground only" method even outperforms the SegFormer (Xie et al. 2021) which uses the full original features on F_β metric. This demonstrates the immense potential of the foreground feature in our FoSp.

Effect of Domain Fusion Module. Method (e) in Table 4 has demonstrated the effectiveness of domain fusion. After incorporating the focus module, the model is inclined towards improving recall, while adding the separation module tends to improve precision. Furthermore, after including the domain fusion module, a better balance between these two metrics is achieved, resulting in the best performance on F_β metric. Compared to the baseline, the model shows a significant improvement of 4.09% (F_β) on the entire test set, and an even more pronounced improvement of 7.71% on the test set of small smoke particles.

Feature-level Separation *vs* Image-level Separation. Fig. 8 has shown the performance of the image-level separation. Although image-level separation methods can split the foreground of smoke more precisely, it is computationally expensive to restore the complete image, and downsampling is required to match the feature dimension of the original image. We have conducted experiments on both feature-level and image-level separation methods, as shown in Table 5. It can be observed that the FoSp using image-level separation outperformed SegFormer, which first proves the effectiveness of the FoSp concept. However, compared to the feature-level separation FoSp, the image-level FoSp results in a lower F_β and $mIoU$. Therefore, we think feature-level separation is the optimal choice for our FoSp.

Conclusion

In this paper, we propose a FoSp for early smoke segmentation (ESS). Concentrating on the formulation of smoke, we first determine the scope of the smoke, and then separate the smoke foreground from the smoke-free background, increasing the contrast between the background and foreground to obtain the complete and refined segmentation map. Furthermore, we provide a SmokeSeg dataset for ESS to promote the development of this field. Surprisingly, we find that sometimes using image-level separation can provide a certain degree of transparency information at the edges of smoke, as shown in Fig. 8. This has surpassed the scope of binary segmentation and we believe this method can be applied to more sophisticated segmentation tasks such as image matting.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62173143 and Grant 61973122.

References

Borji, A.; Cheng, M.-M.; Jiang, H.; and Li, J. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12): 5706–5722.

Cao, Y.; Tang, Q.; and Lu, X. 2022. STCNet: spatiotemporal cross network for industrial smoke detection. *Multimedia Tools and Applications*, 81(7): 10261–10277.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Dewangan, A.; Pande, Y.; Braun, H.-W.; Vernon, F.; Perez, I.; Altintas, I.; Cottrell, G. W.; and Nguyen, M. H. 2022. FIgLib & SmokeyNet: Dataset and Deep Learning Model for Real-Time Wildland Fire Smoke Detection. *Remote Sensing*, 14(4): 1007.

Hsu, Y.-C.; Huang, T.-H. K.; Hu, T.-Y.; Dille, P.; Prendi, S.; Hoffman, R.; Tsuhlares, A.; Pachuta, J.; Sargent, R.; and Nourbakhsh, I. 2021. Project RISE: recognizing industrial smoke emissions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14813–14821.

Jia, Y.; Du, H.; Wang, H.; Yu, R.; Fan, L.; Xu, G.; and Zhang, Q. 2019. Automatic early smoke segmentation based on conditional generative adversarial networks. *Optik*, 193: 162879.

Jing, T.; Meng, Q.-H.; and Hou, H.-R. 2023. SmokeSeger: A Transformer-CNN coupled model for urban scene smoke segmentation. *IEEE Transactions on Industrial Informatics*, 1–12.

Li, X.; Chen, Z.; Wu, Q. J.; and Liu, C. 2018. 3D parallel fully convolutional networks for real-time video wildfire smoke detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 89–103.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

Liu, Y.; Sun, P.; Wergeles, N.; and Shang, Y. 2021. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172: 114602.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mahmoud, M. A. I.; and Ren, H. 2019. Forest fire detection and identification using image processing and SVM. *Journal of Information Processing Systems*, 15(1): 159–168.

Mei, H.; Dong, B.; Dong, W.; Yang, J.; Baek, S.-H.; Heide, F.; Peers, P.; Wei, X.; and Yang, X. 2022. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12622–12631.

Muhammad, K.; Ahmad, J.; and Baik, S. W. 2018. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing*, 288: 30–42.

Najibi, M.; Singh, B.; and Davis, L. S. 2019. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9745–9755.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7479–7489.

Robinson, D. A. 1979. Smoke Detection: Critical Element of a University Residential Fire Safety Program. *Journal of the American College Health Association*, 27(5): 265–66.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7262–7272.

Uzkent, B.; Yeh, C.; and Ermon, S. 2020. Efficient object detection in large images using deep reinforcement learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1824–1833.

Wang, H.; Zhou, L.; and Wang, L. 2019. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8509–8518.

Wang, S.; He, Y.; Zou, J. J.; Zhou, D.; and Wang, J. 2014. Early smoke detection in video using swaying and diffusion feature. *Journal of Intelligent & Fuzzy Systems*, 26(1): 267–275.

Wang, Y.; Lv, K.; Huang, R.; Song, S.; Yang, L.; and Huang, G. 2020. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. *Advances in Neural Information Processing Systems*, 33: 2432–2444.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12321–12328.

Wu, Z.; Su, L.; and Huang, Q. 2019. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7264–7273.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, Z.; Zhang, Z.; Zhu, X.; Huang, G.; and Lin, S. 2020. Spatially adaptive inference with stochastic feature sampling and interpolation. In *European conference on computer vision*, 531–548. Springer.

Xing, D.; Zhongming, Y.; Lin, W.; and Jinlan, L. 2015. Smoke image segmentation based on color model. *Journal on Innovation and Sustainability RISUS*, 6(2): 130–138.

Yan, S.; Zhang, J.; and Barnes, N. 2022. Transmission-Guided Bayesian Generative Model for Smoke Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3009–3017.

Yang, C.; Huang, Z.; and Wang, N. 2022. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13668–13677.

Yu, L.; Mei, H.; Dong, W.; Wei, Z.; Zhu, L.; Wang, Y.; and Yang, X. 2022. Progressive glass segmentation. *IEEE Transactions on Image Processing*, 31: 2920–2933.

Yuan, C.; Liu, Z.; and Zhang, Y. 2019. Learning-based smoke detection for unmanned aerial vehicles applied to forest fire surveillance. *Journal of Intelligent & Robotic Systems*, 93(1): 337–349.

Yuan, F.; Dong, Z.; Zhang, L.; Xia, X.; and Shi, J. 2022. Cubic-cross convolutional attention and count prior embedding for smoke segmentation. *Pattern Recognition*, 131: 108902.

Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; and Li, X. 2019a. A wave-shaped deep neural network for smoke density estimation. *IEEE transactions on image processing*, 29: 2301–2313.

Yuan, F.; Zhang, L.; Xia, X.; Huang, Q.; and Li, X. 2021. A gated recurrent network with dual classification assistance for smoke semantic segmentation. *IEEE Transactions on Image Processing*, 30: 4409–4422.

Yuan, F.; Zhang, L.; Xia, X.; Wan, B.; Huang, Q.; and Li, X. 2019b. Deep smoke segmentation. *Neurocomputing*, 357: 248–260.

Yuan, Y.; Chen, X.; and Wang, J. 2020. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 173–190. Springer.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7151–7160.

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; and Barnes, N. 2020. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8582–8591.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, H.; Xie, X.; Lai, J.-H.; Chen, Z.; and Yang, L. 2020. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9141–9150.