When the fog is dense,IRUNET is all you need

Zhijian Zhou[1]

Abstract

This paper proposes a computer vision model called IRUNET, which combines infrared images with RGB images and uses prior knowledge based on the UNet architecture and self-attention mechanism to remove haze from images.

Firstly, we introduce the IRUNET model in detail, including its structure and operating mechanism. Then, we discuss the principle of fusing infrared images and RGB images, and analyze the advantages and disadvantages of this method.

Subsequently, we elaborate on how the IRUNET model utilizes prior knowledge to more accurately remove haze. We design a new self-attention mechanism that can better preserve the details and texture features of images during the dehazing process while reducing potential noise introduced during the process.

Finally, we demonstrate the superiority of the IRUNET model through extensive experiments. Compared with traditional dehazing methods, our model can better remove haze while preserving image details, and fusing infrared images can improve image clarity and contrast.

In summary, the IRUNET model is an efficient, accurate, and practical computer vision model that can play an important role in the field of dehazing

Keywords

IRUNET, Infrared images, Self-attention mechanism, Dehazing

1. Introduction

When taking photos or videos, we often encounter interference from fog, which leads to a decrease in image or video quality, making them unclear. This kind of interference not only affects our daily life, but also poses great challenges to many application fields such as robot navigation, autonomous driving, and video monitoring. Therefore, eliminating fog in images and videos has become a hot research area.

Currently, many dehazing algorithms have been proposed, such as physics-based methods, traditional image processing methods, and deep learning-based methods. However, these methods still have some limitations, such as difficulty in handling complex scenes, being easily affected by lighting changes and noise, and poor dehazing effect.

To address these issues, this paper proposes a deep learning model named IRUNET. IRUNET combines infrared and RGB images and uses a self-attention mechanism to remove fog from the image to enhance its clarity and contrast. This paper provides a detailed introduction of IRUNET's network architecture and implementation, and conducts extensive experimental

20307140081@fudan.edu.cn (Z. Zhou)
https://github.com/ZhijianZhou/IRUNET (Z. Zhou)
18001699516 (Z. Zhou)

CEUR Workshop Proceedings (CEUR-WS.org)

evaluations on IRUNET to demonstrate its superiority in dehazing effectiveness compared to several state-of-the-art dehazing methods. The experimental results show that IRUNET outperforms other dehazing methods in terms of image quality metrics such as PSNR, SSIM, and visual quality, indicating its effectiveness in removing fog from images.

2. Related work

2.1. Dark channel prior

Kaiming He proposes a simple yet effective image prior called the "dark channel prior" for single image haze removal. The dark channel prior is based on the statistics of outdoor images and by applying it to the haze imaging model, single image haze removal becomes simpler and more effective. However, the method may fail to work for certain specific images, such as those where the scene objects are inherently similar to the atmospheric light and no shadow is cast on them. Additionally, like other haze removal methods, it is limited by the applicable scope of the haze imaging model.

2.2. Dehazeformer

Yuda Song proposes DehazeFormer, which is a Transformer-based method for image dehazing that includes improvements such as modified normalization layer, activation function, and spatial information aggregation scheme. The authors trained multiple variants of DehazeFormer on various datasets and demonstrated its effectiveness. However, one limitation of using Transformers is the high hardware requirements needed to train and run these models.

2.3. Gunet

Yuda Song also proposed gUNet, a compact dehazing network, by making minimal modifications to the popular U-Net architecture. They replaced U-Net's convolutional blocks with residual blocks featuring the gating mechanism and fused the main path and skip connection feature maps using the selective kernel. The authors demonstrated that gUNet outperforms state-of-the-art methods on multiple image dehazing datasets with significantly reduced overhead. Additionally, they conducted extensive ablation studies to verify key designs that contributed to the improved performance of image dehazing networks. However, gUNet did not consider incorporating prior knowledge.

3. IRUNET

Figure 1 shows the overall architecture of IRUNET. Our model can be viewed as a 7-stage U-Net variant, with each stage consisting of a stack of Conv blocks proposed by us. One difference from U-Net is that we added a three-layer convolutional block during the second downsampling layer, to fuse the information from the infrared (IR) image, which also undergoes a parallel structure with the RGB layer.Additionally, we have incorporated the Channel Squeeze-and-Excitation

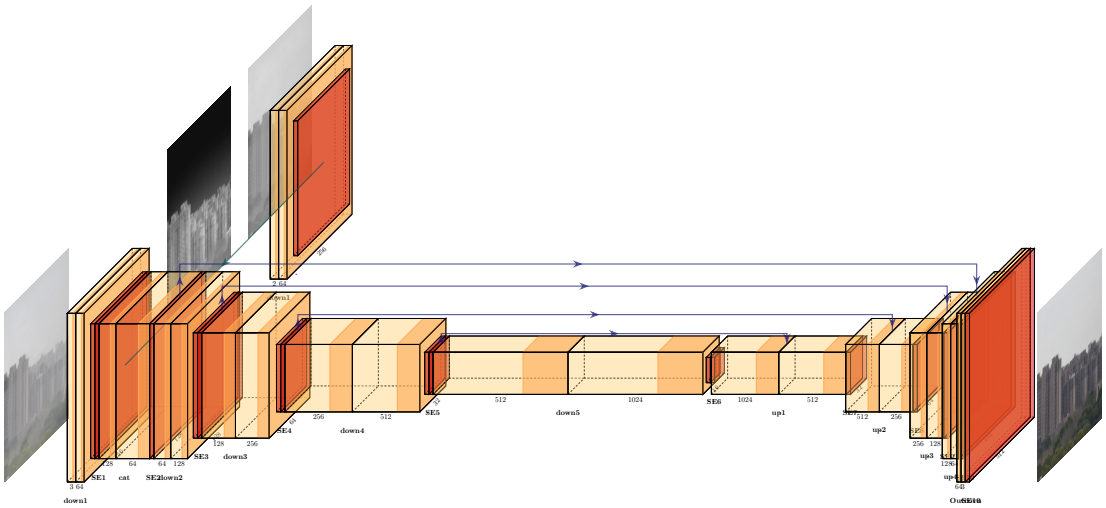

Figure 1: IRUNET: A state-of-the-art image dehazing network based on the U-Net architecture, with spatial attention and infrared image fusion techniques

Block (SE) after each convolutional layer to explicitly model the interdependence between channels and adaptively recalibrate the feature responses along the channel direction.

3.1. Infrared fusion

From the Figure 2 images, it can be observed that after applying Fourier transform and color inversion, the infrared image is closer to the original image in the high-frequency components and better preserves the texture and edge information. This is because the principle of infrared imaging is based on receiving the infrared radiation energy emitted by the detected object, which is less affected by weather and lighting conditions and is not affected by PM2.5 pollution, and can produce clear images. Therefore, using the fusion of infrared and RGB images can better restore the texture and detail features of the original image in the task of image dehazing. subfigure

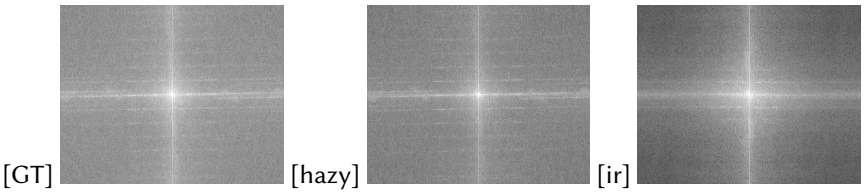

Figure 2: The image above shows the frequency domain images obtained by applying the Discrete Fourier Transform (DFT) to the original image (leftmost), the hazy image (middle), and the flipped infrared image (rightmost), after being centered.

We can clearly compare the differences between the two images from Figure 3. Compared to the hazy image, the infrared image has a higher sky discrimination because the infrared camera receives less thermal radiation from the sky. From another perspective, we hope that by using

the image of the infrared channel, the network can learn how to distinguish between sky and foreground independently, which can be considered as a priori knowledge of thermal radiation channel. This has an advantage over the dark channel prior, which divides the dark channel based on the brightness of the image and is greatly affected by other light-colored pixel blocks in the image that are not sky. Therefore, from the comparison in Figure 4, we can see that the visually our image after dehazing is superior to that obtained using the dark channel prior

[a]  [b] 

Figure 3: On the left are grayscale images of the infrared image, and on the right are grayscale images of the hazy image

[a]  [b]  [c] 

Figure 4: Image a is the original image, image b is the result obtained using the dark channel prior method, and image c is the result generated by our IRUNET.

The reason for choosing to fuse infrared and RGB image information at the second convolutional layer is that in the Unet structure, the higher the layer, the more high-frequency information is preserved, making it more suitable for fusing the better-preserved high-frequency information from the infrared image into the RGB image. If fusion is performed at deeper layers, the low-frequency signals in the infrared image will interfere with the dehazing process of the image, making it difficult for the network to converge and the dehazed image generated by the network at deeper fusion layers may not conform to our normal perception, as shown in Figure 5.

3.2. SEblock

The core idea of the SE (Squeeze-and-Excitation) block is to use a Squeeze operation to compress the feature map into a global feature vector, and then use an Excitation operation to weight each channel, enhancing the expression of useful information and reducing the influence of irrelevant information. Specifically, the Squeeze operation usually employs global average pooling to compress the feature map into a 1D vector, while the Excitation operation consists of two fully connected layers, where the first layer is used for dimensionality reduction, and the second layer is used for dimensionality restoration and generating channel-wise attention

Figure 5: Bad dehazed images generated by the network used for deep fusion of infrared and RGB images during training.

weights. Finally, the channel-wise attention weights are applied to the original feature map, and each channel is weighted by the corresponding attention weight using element-wise multiplication, resulting in an enhanced feature map with important features emphasized.

In the original structure of Unet, the number of channels can reach up to 1024 after multiple rounds of convolution. However, some channels may contain less important information. Therefore, as shown in Figure 1, we introduced SE blocks after the convolutional layers in IRUNET, with the aim of allowing the network to learn the importance of different channels, enhancing the expressiveness and performance of the model, reducing the impact of useless information, and thus improving the robustness and generalization ability of the model. Additionally, we also added SE blocks after the fusion of infrared and RGB image information, hoping that the model can autonomously select important channels from both sources and integrate their information to restore the original image to the greatest extent possible.

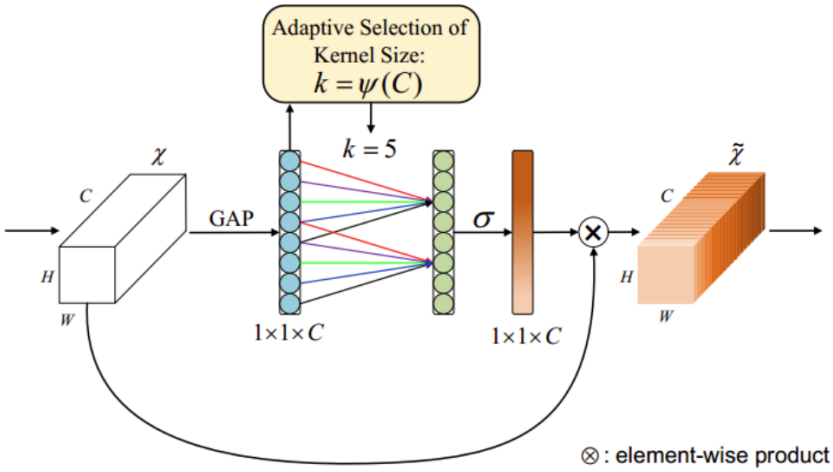

Figure 6: The above figure illustrates the specific operations of SE block.

4. Experience

4.1. Data set

The dataset used in this study was obtained from the Image Processing and Computer Vision course at Fudan University. In order to facilitate the reproducibility of the experiments, we fixed the random seed to 3407 and split the dataset into training, validation, and testing sets with an 8:1:1 ratio. We also generated the haze-free images with the minimum impact on saturation and hue in the HSV domain for the hazy images, which were later input to the infrared image network together with the infrared images.

4.2. Data augmentation

In terms of data augmentation, we divide the input into image blocks with a patch size of $256 * 256$. The purpose of doing this is to reduce GPU memory pressure. Directly inputting large images into the network will occupy too much GPU memory. Dividing the images into small patches can reduce GPU memory pressure and accelerate model training through batch processing, which is beneficial for training IRUNET on personal GPUs. It can also provide more training samples, increase data diversity, and improve the generalization ability of the model. In addition, it can reduce the risk of model overfitting, because the model needs to learn more local features rather than global features of the entire image. By dividing the image into small patches, data can be further augmented by random translation to improve the generalization ability of the model.

4.3. Optimization strategy

For the loss function, we used the L1 loss function. Compared with the L2 loss function, L2 tends to penalize larger errors while having a higher tolerance for smaller errors. This ignores

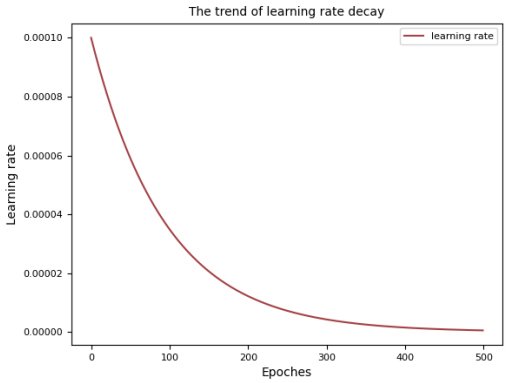

Figure 7: The trend of learning rate decay

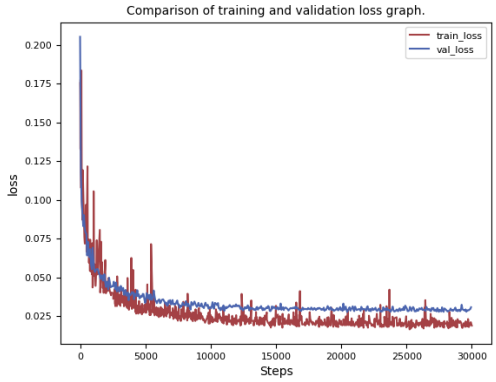

Figure 8: Comparison of training and validation loss graph

the essential structure of the image and can result in artifacts and larger blocks.
We used the Adam optimizer and conducted multiple experiments to set the learning rate. Ultimately, we adopted a strategy of exponential decay for the learning rate,as shown in the Figure 7. This strategy is intended to achieve fast convergence in the early stages of training and then gradually decrease the step size to reduce oscillations and make the training process more stable.

$$lr = init\_lr * decacy\_rate^{\frac{global\_steps}{decay\_steps}}$$

As shown in the Figure 8,we monitored the validation loss during training, and if it stopped decreasing or started to increase, we stopped training. This prevented the model from overfitting to the training set and allowed it to generalize better to new data. We ultimately selected the model trained around 400 epochs as our final model.

5. result

5.1. Evaluate

PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are two commonly used metrics for evaluating the quality of images. PSNR measures the difference between the original image and the reconstructed image by calculating the ratio of the maximum possible power of the original image to the power of the difference between the two images. A higher PSNR value indicates a higher similarity between the two images.

SSIM, on the other hand, evaluates the structural similarity between two images. It compares the luminance, contrast, and structure of the original and the reconstructed images, and calculates the index based on the differences between them. A higher SSIM value indicates a higher similarity between the two images.

Both PSNR and SSIM are widely used in the field of image processing and computer vision to evaluate the performance of image restoration, denoising, and enhancement algorithms.

As shown in Table 1, we applied the dark channel prior (DCP), the original Unet model trained on our training set, and two pre-trained models provided by GUnet (rehaze and haze4K), as well as the IRUNET model, to perform dehazing on our test images and calculated their PSNR and SSIM scores.

Table 1
The following table shows the scores obtained by different models on our test set.

Model	PSNR	SSIM
DCP	17.06	0.856
Unet	24.93	0.857
rshaze	31.79	0.853
haze4k	34.73	0.936
IRUNET	34.95	0.977

The performance of IRUNET is significantly superior to the dark channel prior method and the Unet model trained with the same parameters, and slightly better than the two pre-trained models of Gunet, as shown in Table 1. This overall confirms the advantage of IRUNET in fusing infrared images, utilizing prior knowledge, and employing channel-wise self-attention mechanisms over other models.

Figure 9: A

Figure 10: B

Figure 11: C

Figure 12: D

5.2. Image show

We selected A-G images from the test set to showcase the dehazing performance of IRUNET, which consists of a total of seven image categories.

Figure 13: E

Figure 14: F

Figure 15: G

6. Discussion and Limitation

Our model still has the following issues

1. As Figure 16 shows,IRUNET cannot handle the halo effect very well. Halos are relatively rare in the dataset, partly because the shorter wavelength of light emitted by halos makes them penetrate through fog more easily and the infrared images cannot capture this information. As a result, the performance of our model on images with halos in the test set is poor.
2. Due to the limitations of infrared images, our model has not been tested on a public benchmark dataset, and we have not compared its performance with other models on a level playing field. We have only tested it on our own dataset, which has certain limitations.
3. Due to the limitation of experimental equipment, we only had one portable laptop with a 3060 graphics card to train the model. We did not have enough time and hardware

Figure 16: Images with halo effect

resources to conduct comprehensive comparative experiments. If we had done enough experiments, we believe that IRUNET could have shown its full potential.

References

[1] Kaiming He, Jian Sun, and Xiaoou Tang. "Single Image Haze Removal Using Dark Channel Prior". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.12 (2011), pp. 2341–2353. DOI: 10.1109/TPAMI.2010.168.

[2] David Picard. "Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision". In: *arXiv preprint arXiv:2109.08203* (2021).

[3] Yuda Song et al. "Rethinking Performance Gains in Image Dehazing Networks". In: *arXiv preprint arXiv:2209.11448* (2022).

[4] Yuda Song et al. "Vision transformers for single image dehazing". In: *IEEE Transactions on Image Processing* 32 (2023), pp. 1927–1941.

[5] Hang Zhao et al. "Loss functions for image restoration with neural networks". In: *IEEE Transactions on computational imaging* 3.1 (2016), pp. 47–57.

[3] [4] [2] [5] [1]