Towards Real-time High-Definition Image Snow Removal: Efficient Pyramid Network with Asymmetrical Encoder-decoder Architecture*

Tian Ye[1†], Sixiang Chen[1†], Yun Liu[2†], Yi Ye[1], JinBin Bai[3], and Erkang Chen[1**]

[1] School of Ocean Information Engineering, Jimei University, Xiamen, China
{201921114031,201921114013,201921114003,ekchen}@jmu.edu.cn
[2] College of Artificial Intelligence, Southwest University, Chonqing, China
yunliu@swu.edu.cn
[3] Department of Computer Science and Technology, Nanjing University, China
jinbin.bai@smail.nju.edu.cn

Abstract. In winter scenes, the degradation of images taken under snow can be pretty complex, where the spatial distribution of snowy degradation varies from image to image. Recent methods adopt deep neural networks to recover clean scenes from snowy images directly. However, due to the paradox caused by the variation of complex snowy degradation, achieving reliable High-Definition image desnowing performance in real time is a considerable challenge. We develop a novel Efficient Pyramid Network with asymmetrical encoder-decoder architecture for real-time HD image desnowing. The general idea of our proposed network is to utilize the multi-scale feature flow fully and implicitly to mine clean cues from features. Compared with previous state-of-the-art desnowing methods, our approach achieves a better complexity-performance trade-off and effectively handles the processing difficulties of HD and Ultra-HD images.

The extensive experiments on three large-scale image desnowing datasets demonstrate that our method surpasses all state-of-the-art approaches by a large margin both quantitatively and qualitatively, boosting the PSNR metric from 31.76 dB to 34.10 dB on the CSD test dataset and from 28.29 dB to 30.87 dB on the SRRS test dataset. The source code is available at https://github.com/Owen718/Towards-Real-time-High-Definition-Image-Snow-Removal-Efficient-Pyramid-Network.

Keywords: Desnowing · Real-time · Asymmetrical Encoder-decoder Architecture.

* This work was supported by Natural Science Foundation of Chongqing, China (Grant No. cstc2020jcyj-msxmX0324), the project of science and technology research program of Chongqing Education Commission of China (Grant No. KJQN202200206), Natural Science Foundation of Fujian Province (Grant No. 2021J01867), the Education Department of Fujian Province (Grant No. JAT190301) and Foundation of Jimei University (Grant No. ZP2020034).
** Corresponding author. †Equal contribution.

1 Introduction

In nasty weather scenes, snow is an essential factor that causes noticeable visual quality degradation. Degraded images captured under snow scenes significantly affect the performance of high-level computer vision tasks [12,14,15,8,9,16].

Snowy images suffer more complex degradation by various factors than common weather degradation,i.e., haze and rain. According to previous works, snow scenes usually contain the snowflake, snow streak, and veiling effect. The formation of snow can be modeled as:

$$\mathbf{I}(x) = \mathbf{K}(x)\mathbf{T}(x) + \mathbf{A}(x)(1 - \mathbf{T}(x)), \tag{1}$$

where $\mathbf{K}(x) = \mathbf{J}(x)(1-\mathbf{Z}(x)\mathbf{R}(x))+\mathbf{C}(x)\mathbf{Z}(x)\mathbf{R}(x)$, $\mathbf{I}(x)$ denotes the snow image, $\mathbf{K}(x)$ denotes the veiling-free snowy image,, $\mathbf{A}(x)$ is the atmospheric light, and $\mathbf{J}(x)$ is the scene radiance. $\mathbf{T}(\mathbf{x})$ is the transmission map. $\mathbf{C}(x)$ and $\mathbf{Z}(x)$ are the chromatic aberration map for snow images and the snow mask, respectively. $\mathbf{R}(x)$ is the binary mask, presenting the snow location information.

As described in Eq.1, the chromatic aberration degradation of snow and the veiling effect of haze are mixed in an entangled way. Existing snow removal methods can be categorized into two classes: model-based methods and model-free methods. For model-based methods [20,11,3], JSTASR [3] tries to recover a clean one from a snow image in an uncoupled way. Utilizing the veiling effect recovery branch to recover the veiling effect-free image, and the snow removal branch to recover the snow-free image. However, the divide and conquer strategy ignores the influence of inner entanglement degradation in snow scenes, and complicated networks by hand-craft design results in unsatisfactory model complexity and inference efficiency. For model-free methods [10,4,17], HDCW-Net [4] proposed the hierarchical decomposition paradigm, which leverages frequency domain operations to extract clean features for a better understanding of the various diversity of snow particles. But the dual-tree complex wavelet limits the inference performance of HDCW-Net, and it still has scope for improvement in its performance.

Single image snow removal methods have made remarkable progress recently. Yet, there are few studies on the efficient single image snow removal network, which attract us to explore the following exciting topic:

How to design an ***efficient*** *network to* ***effectively*** *perform single image desnowing?*

Most previous learning-based methods [4,3,17] hardly achieve real-time inference efficiency with HD resolution, even running on the expensive advanced graphics processing unit. We present a detailed run-time comparison in the experiment section to verify this point. Most methods can not perform real-time processing ability to handle High-Definition degraded images. In this manuscript, we propose an efficient manner to perform **HD(1280×720)** resolution image snow removal in real time, which is faster and more deployment-friendly than previous methods. Moreover, our method is the first desnowing network that can handle $UHD(4096 \times 2160)$ image processing problem.

Previous desnowing networks [4,3] usually only have 2 to 3 scale-level, which limits the desnowing performance and inferencing efficiency. Different from the

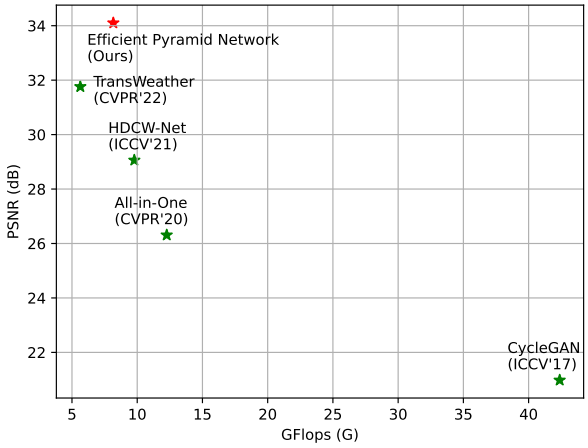

Fig. 1: Trade-off between performance vs: number of operations on CSD (2000) [4] testing dataset. Multi-adds are calculated on 256×256 image. The results show the superiority of our model among existing methods

previous mainstream designs of desnowing CNN architecture, the proposed network owns five scale levels, which means the smallest feature resolution of the feature flow in our pyramid network is only $\frac{1}{32}$ of input snow image. Sufficient multi-scale information brings impressive representation ability. Furthermore, the efficient and excellent basic block is also a significant factor for network performance and actual efficiency. Thus we propose a Channel Information Mining Block (**CIMB**) as our basic block to mine clean cues from channel-expanded features, which is inspired by NAFNet [2]. Besides that, we propose a novel External Attention Module (EAM) to introduce reliable information from original degraded images to optimize feature flows in the pyramid architecture. The proposed EAM can adaptively learn more useful sample-wise features and emphasize the most informative region on the feature map for image desnowing.

Motivated by the proposed efficient and effective components, our method achieves excellent desnowing performance. Compared with the previous best method transweather [17], the proposed method has better quantitative results (**31.76dB/0.93 vs. 34.10dB/0.95**) on CSD [4] dataset. And as shown in Figure. 1, compared with previous state-of-the-art desnowing methods, our method achieves a better complexity-performance trade-off.

The main contributions of this paper are summarized as follows:

- a) We propose a Channel Information Mining Block (CIMB) to explore clean cues efficiently. Furthermore, the External Attention Module (EAM) is proposed to introduce external instructive information to optimize features. Our ablation study demonstrates the effectiveness of CIMB and EAM.

- b) We propose the Efficient Pyramid Network with asymmetrical encoder-decoder architecture, which achieves real-time High-Definition image desnowing.
- c) Our method achieves the best quantitative results compared with the state-of-the single image desnowing approaches.

2 Related Works

Traditional methods usually make assumptions and use typical priors to handle the ill-posed nature of the desnowing problem. One of the limitations of these prior-based methods is that these methods can not hold well for natural scenes containing various snowy degradation and hazy effect. Recently, due to the impressive success of deep learning in computer vision tasks, many learning-based approaches have also been proposed for image desnowing [11,3,4,10]. In these methods, the key idea is to directly learn an effective mapping between the snow image and the corresponding clean image using a solid CNN architecture. However, these methods usually involve complex architecture and large-kernel size of essential convolution components and consume long inferencing times, which cannot cover the deployment demand for real-time snow image processing, especially for High-Definition (HD) and Ultra-High-Definition (UHD) images.

The first desnowing network is named DesnowNet [11], which focuses on removing translucent and snow particles in sequence based on a multi-stage CNN architecture. Li *et al.* propose an all-in-one framework to handle multiple degradations, including snow, rain, and haze. JSTASR [3] propose a novel snowy scene imaging model and develop a size-aware snow removal network that can handle the veiling effect and various snow particles. HDCW-Net [4] performs single image snow removal by utilizing the dual-tree wavelet transform and designing a multi-scale architecture to handle the various degradation of snow particles. Most previous methods are model-based, which limits the representation ability of CNNs. Moreover, the wavelet-based method is not deployment-friendly for applications.

3 Proposed Method

This section will first introduce the Channel Information Mining Block (CIMB) and External Attention Module. Then, we present the proposed pyramid architecture. Worth noting that profit from the asymmetrical encoder-decoder design and efficient encoder blocks, the proposed method is faster and more effective than previous CNN methods.

3.1 Channel Information Mining Block

Previous classical basic block in desnowing networks [11,3,4] usually utilize large kernel-size convolution and frequency-domain operations. In contrast with the

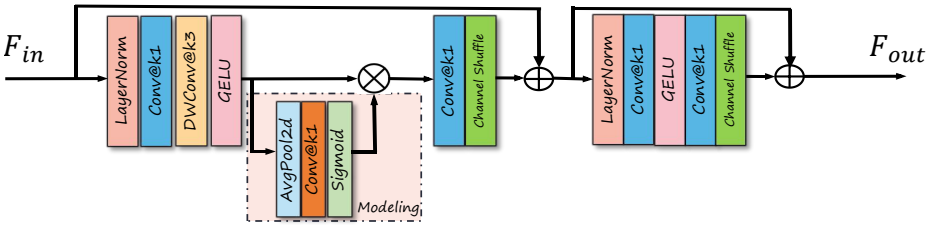

Fig. 2: The proposed Channel Information Mining Block (**CIMB**). The *DWConv* denotes the depth-wise convolution operation, and the kernel size of the convolution is denoted by $@k*$. The design of our CIMB is motivated by NAFNet [2].

previous design, the proposed Channel Information Mining Block (CIMB) focus on how to effectively mine clean cues from incoming features with minimum computational cost.

As shown in Fig. 2, let's denote the input feature as F_{in}^{c} and output feature as F_{out}^{c}. The computational process of CIMB can be presented by:

$$F_{out} = CIMB(F_{in}). \tag{2}$$

Our design is simple and easy to implement in a widely-used deep learning framework. We utilize Layer Normalization to stabilize the training of the network and use the convolution with the kernel size of 1×1 to expand the channel of feature maps from c to αc, where the α is the channel-expand factor, is set as 2 in our all experiments. The channel-expanding way is crucial to motivate state-of-the-art performance because we found that high-dimension information mining is significantly suitable for single image desnowing. We further utilize channel information modeling to model the distribution of snow degradation and explore clean cues in the channel dimension. The channel-expanding design is motivated by NAFNet [2], which is an impressive image restoration work. Nevertheless, it ignores achieving efficient information interaction across different channels. We deeply realize the lack of effective channel interaction and introduce channel shuffle operation to achieve efficient and effective channel information interaction. Our ablation study demonstrates that the clever combination of channel-expanding and channel-shuffle achieves better performance with little computational cost by channel-shuffle.

3.2 External Attention Module

In the past few years, more and more attention mechanisms [7,18,18] have been proposed to improve the representation ability of the convolution neural network. However, most attention module [7,18,6,13] only focus on exploring useful information from incoming features, which ignores capturing and utilizing implicit instruction information from original input images. Moreover, information loss in multi-scale architecture cannot be avoided; thus, we utilize down-sampled snow

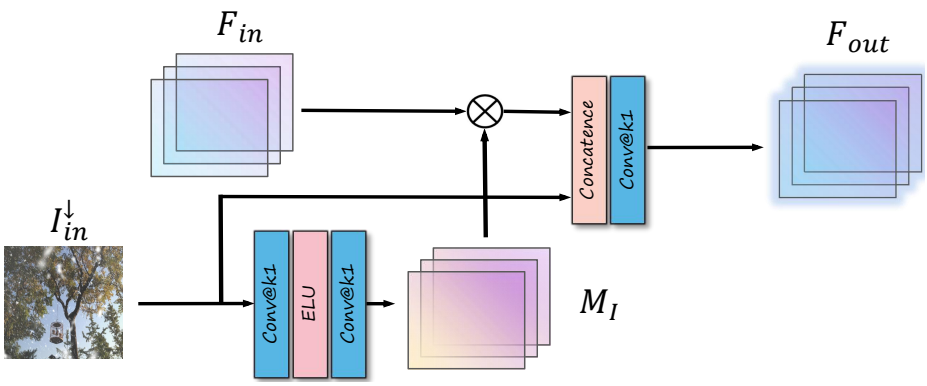

Fig. 3: The proposed External Attention Module (**EAM**). The kernel size of the convolution is denoted by @$k*$.The tensor size of the external attention map M_I is the same as the incoming feature F_{in} and generated feature F_{out}.

degraded images to introduce original information, relieving the information loss by multi-scale feature flow design.

As shown in Fig. 3, the External Attention Module is a plug-and-play architectural unit that can be directly used in CNNs for image restoration tasks. Specifically, we perform a series of operations on the down-sampled $I_{in}^{\downarrow}$ to generate the external attention map $\mathcal{M}_I^{H\times W\times C}$:

$$\text{Conv} \circ \text{ELU} \circ \text{ Conv with } 1\times 1 : I_{in}^{\downarrow} \rightarrow \mathcal{M}_I^{H\times W\times C}, \tag{3}$$

where the $I_{in}^{\downarrow}$ is the down-sampled original image, whose spatial size is the same as the incoming feature map F_{in}. Then we multiply the $\mathcal{M}_I^{H\times W\times C}$ with the F_{in}:

$$F_{\text{att}}^{H\times W\times C} = \mathcal{M}_I^{H\times W\times C} \cdot F_{\text{in}}^{H\times W\times C}, \tag{4}$$

where the $F_{\text{att}}^{H\times W\times C}$ is the scaled feature map. And we further utilize $I_{in}^{\downarrow}$ and channel-wise compression to introduce the original information:

$$\text{Concatence: } I_{in}^{\downarrow} + F_{att}^{H\times W\times C} \rightarrow F_{fusion}^{H\times W\times (C+3)}. \tag{5}$$

And a convolution with a kernel size of 3×3 is used to compress the dimension of F_{fusion} and get the final output feature F_{out}:

$$\text{Conv:} F_{fusion}^{H\times W\times (C+3)} \rightarrow F_{out}^{H\times W\times C}. \tag{6}$$

The benefits of our EAM are twofold: i) Fully utilize original degradation information to instruct the feature rebuilding explicitly. ii) Relief the information loss by repeat down-up sampling in multi-scale architecture. Please refer to our experiments section for the ablation study about the External Attention Module, which demonstrates the effectiveness of our proposed EAM.

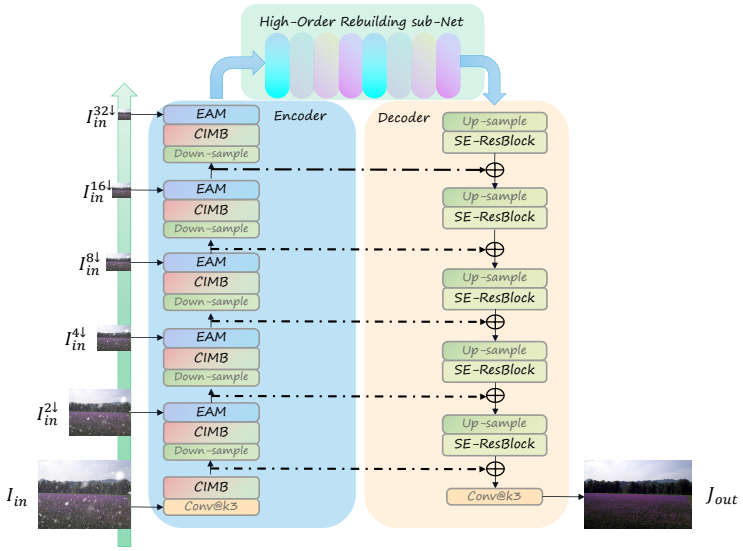

Fig. 4: Overview of the proposed Efficient Pyramid Network.

3.3 Efficient Pyramid Network with Asymmetrical Encoder-decoder Architecture

Efficient Pyramid Network. U-Net style architectures can bring sufficient multi-scale information compared with single-scale-level architectures. However, the fully symmetrical architecture of U-Net style architectures results in redundant computational costs. Unlike previous mainstream designs [2,19] in the image restoration area, we develop an efficient pyramid architecture to explore clean cues from features of 5 scale levels, which is a more efficient way to handle the complex degradation, aka, snowy particles, and uneven hazy effect.

As shown in Fig. 4, our efficient pyramid network consists of three parts: encoder, decoder, and High-Order Rebuilding Sub-Net. Every scale-level encoder block only has a CIMB and an EAM in the encoder stage. For the decoder, we utilize ResBlock with SE attention [7] as our basic decoder block. Our decoder is fast and light. In the following sections, we further present the straight idea about Asymmetrical E-D design and High-Order Rebuilding Sub-Net.

Asymmetrical Encoder-decoder Architecture. The symmetrical encoder-decoder architecture has been proven work well in many CNNs [19,2,4]. Most methods tend to add more conv-based blocks in every scale level with a symmetrical design to improve the model performance further. However, directly utilizing symmetrical architecture is not the best choice for our aims for the following reasons. First, we aim to process ***HD*** *(*1280×720*)* snowy images in real time, so we have to make trade-offs between performance and model complexity. Second, we found that a heavy encoder with a light decoder has a better representation

ability to explore clean cues than a symmetrical structure. Thanks to the asymmetrical ED architecture, our Efficient Pyramid Network performs well on HD snow images in real time.

High-Order Rebuilding Sub-Net. It is critical for an image restoration network to exploit clean cues from the latent features effectively. Here, we design an effective High-Order Rebuilding Sub-Net to further rebuild clean features in high-dimension space. Our High-Order Rebuilding Sub-Net comprises 20 Channel Information Mining Block of 512 dimensions. Due to latent features only having $\frac{1}{32}\times$ resolution size than original input images, our High-Order Rebuilding Sub-Net quickly performs clean cues mining. And channel information exploration ability in the latent layer provides a significant gain from Table.7.

4 Loss Function

We only utilize the Charbonnier loss [1] as our basic reconstruction loss:

$$\mathcal{L}_{\text{char}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left\|J_{out}^{i} - J_{gt}^{i}\right\|^{2} + \epsilon^{2}}, \tag{7}$$

with constant ϵ empirically set to $1e^{-3}$ for all experiments. And J_{gt}^{i} denotes the ground-truth of J_{out}^{i} correspondingly.

5 Experiments

5.1 Datasets and Evaluation Criteria

We choose the widely used PSNR and SSIM as experimental metrics to measure the performance of our network. We train and test the proposed network on three large datasets: CSD [4], SRRS [3] and Snow100k [3], following the benchmark-setting of latest desnowing methods [4] for authoritative evaluation. Moreover, we re-train the latest bad weather removal method TransWeather (CVPR'22) [17] to make a better comparison and analysis. Worth noting that we reproduce the TransWeather-based for a fair comparison, provided by the official repository of TransWeather [17].

5.2 Implementation Details

We augment the training dataset by randomly rotating by 90,180,270 degrees and horizontal flip. The training patches with the size 256×256 are extracted as input paired data of our network. We utilize the AdamW optimizer with an initial learning rate of 4×10^{-4} and adopt the CyclicLR to adjust the learning rate progressively, where on the mode of triangular, the value of gamma is 1.0, base momentum is 0.9, the max learning rate is 6×10^{-4} and base learning

Table 1: Quantitative comparisons of our method with the state-of-the-art desnowing methods on CSD,SRRS and Snow 100K desnowing datasets (PSNR(dB)/SSIM). The best results are shown in **bold**, and second best results are underlined.

Method	CSD(2000)		SRRS (2000)		Snow 100K (2000)		#Param	#GMacs
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
(TIP'18)Desnow-Net [11]	20.13	0.81	20.38	0.84	30.50	0.94	26.15 M	1717.04G
(ICCV'17)CycleGAN [5]	20.98	0.80	20.21	0.74	26.81	0.89	7.84M	42.38G
(CVPR'20)All-in-One [10]	26.31	0.87	24.98	0.88	26.07	0.88	44 M	12.26G
(ECCV'20)JSTASR [3]	27.96	0.88	25.82	0.89	23.12	0.86	65M	-
(ICCV'21)HDCW-Net [4]	29.06	0.91	27.78	0.92	31.54	0.95	6.99M	9.78G
(CVPR'22)TransWeather [17]	31.76	0.93	28.29	0.92	31.82	0.93	21.9M	5.64G
Ours	**34.10**	**0.95**	**30.87**	**0.94**	**33.62**	**0.96**	66.54M	8.17G

rate is the same as the initial learning rate. We utilize the Pytorch framework to implement our network with 4 RTX 3080 GPU with a total batch size of 60. For channel settings, we set the channel as $[16, 32, 64, 128, 256, 512]$ in each scale-level stage respectively.

5.3 Performance Comparison

In this section, we compare our Efficient Pyramid Network method with the state-of-the-art image desnowing methods of [11,3,4], classical image translation method of [5] and the bad weather removal methods of [10,17].

Visual Comparison with SOTA methods We compare our method with the state-of-the-art image desnowing methods on the quality of restored images, presented in Fig. 5 and 6. Our approach generates the most natural desnowing images compared to other methods. The proposed method effectively restores the degraded area of the snow streaks or snow particles in both synthetic and authentic snow images.

Quantitative Results Comparison In Table.1, we summarize the performance of our Efficient Pyramid Network and SOTA methods on CSD [4], SRRS [3] and Snow 100K [11]. Our method achieves the best performance with 34.10dB PSNR and 0.95 SSIM on the test dataset of CSD. Moreover, it achieves the best performance with 30.87dB PSNR, 33.62 dB PSNR, and 0.94 SSIM, 0.96 SSIM on SRRS and Snow100k test datasets.

Run-time Discussion In Table.2, 3, 4, we present detailed run-time and model-complexity comparison with different processing resolution settings. Worth noting that TransWeather [17] can not handle **UHD**(4096×2160) degraded images, as shown in Table.3, although it is minorly faster than ours when the input image is HD or smaller size. As shown in Table.2, our Efficient Pyramid Network achieves real-time performance when processing HD resolution images.

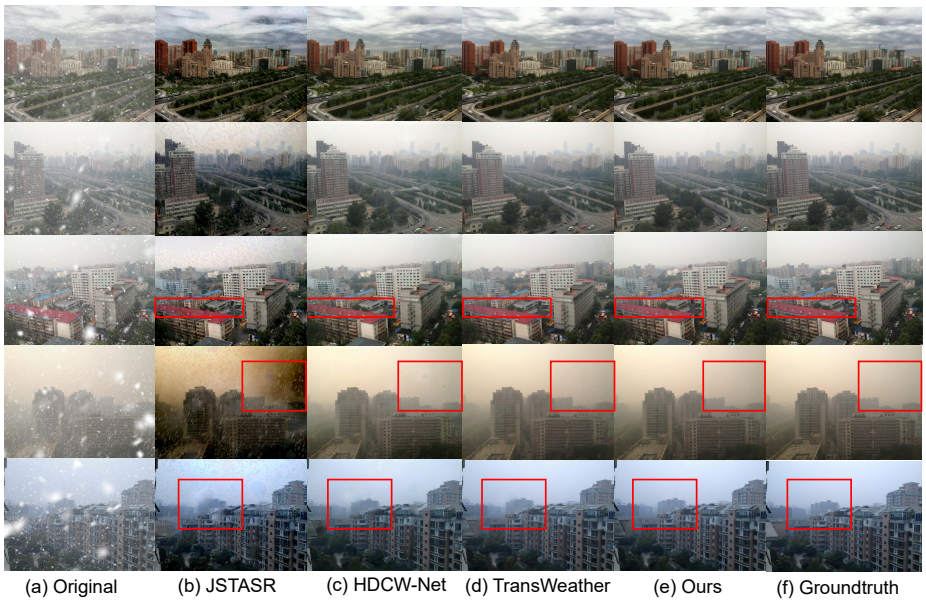

Fig. 5: Visual comparisons on results of various methods (b-e) and our proposed network(e) on synthetic winter photos. Please zoom in for a better illustration.

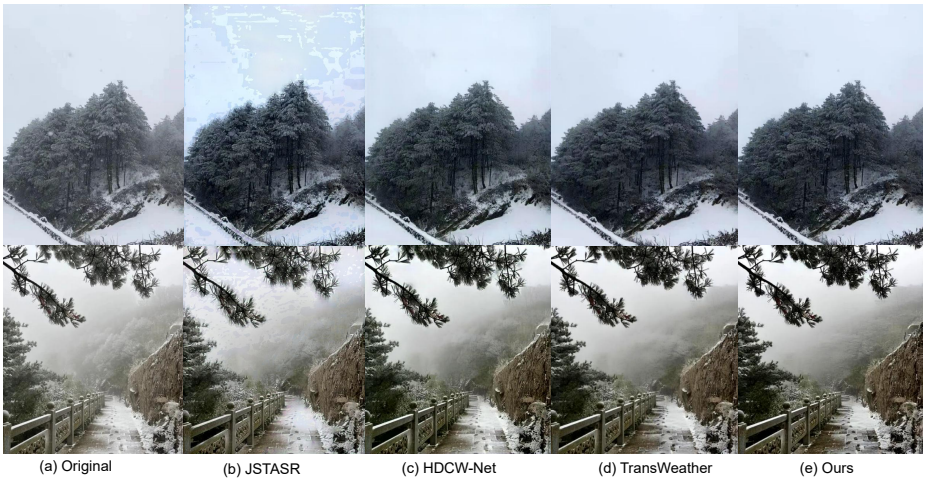

Fig. 6: Visual comparisons on results of various methods (b-d) and our method(e) on real winter photos. Please zoom in for a better illustration.

Table 2: **Comparison of Inference time, GMACs (fixed-point multiply accumulate operations performed persecond) and Parameters when input HD (1280 × 720) images**. Our method achieves the best runtime-performance trade-off compared to the state-of-the-art approaches. The time reported in the table corresponds to the time taken by each model to feed-forward an image of dimension 1280 × 720 during the inference stage. We perform all inference testing on an A100 GPU for a fair comparison. Notably, we utilize the *torch.cuda.synchronize()* API function to get accurate feed forward run-time.

Method	Inf. Time(in s)	GMACs(G)	Params(M)
TransWeather [17]	0.0300	79.66	21.9
Ours	0.0384	113.72	65.56

Table 3: **Comparison of Inference time, GMACs (fixed-point multiply accumulate operations performed persecond) and Parameters when input UHD (4096 × 2160) images**. We perform all inference testing on an A100 GPU for a fair comparison.

Method	Inf. Time(in s)	GMACs(G)	Params(M)
TransWeather [17]	Out of Memory	-	21.9
Ours	0.23	1097.03	65.56

Table 4: **Comparison of Inference time, FPS and Parameters when input small (512×672) images**. Following the testing platform of HDCW-Net [4], we perform all inference testing on a RTX 1080ti GPU for a fair comparison.

Method	Inf. Time(in s)	FPS	Params(M)
JSTASR [3]	0.87	1.14	65
HDCW-Net [4]	0.14	7.14	6.99
Ours	0.0584	17.11	65.56

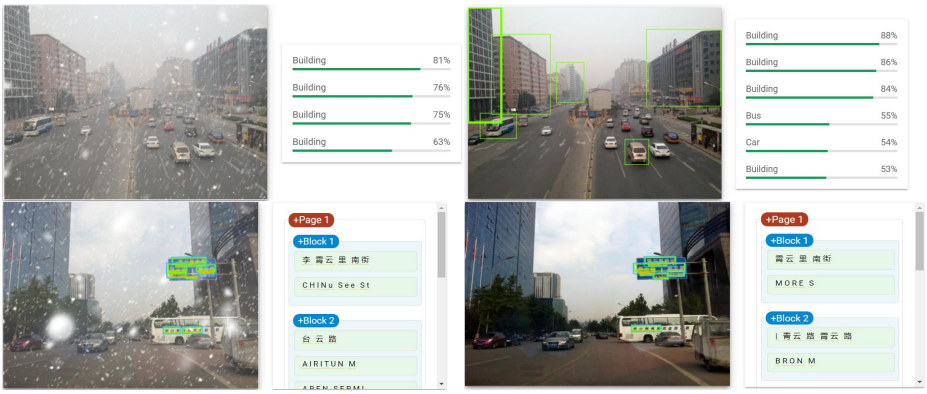

Fig. 7: Synthetic snow images (left) and corresponding desnowing results (right) by the proposed method with the high-level vision task results and confidences supported by Google Vision API.

Model Complexity and Parameter Discussion For real-time deployment of CNNs, computational complexity is an important aspect to consider. To check the computational complexity, we tabulate the GMAC in Table.1 for previous

state-of-the-art methods and our proposed method when the input image is 256×256. We note that our network has less computational complexity when compared to previous methods. Notably, Desnow-Net is highly computationally complex than ours, even though it has less number of parameters.

5.4 Quantifiable Performance for High-level Vision Tasks

As shown in Fig. 7, we offer subjective but quantifiable results for a concrete demonstration, in which the vision task results and corresponding confidences are both supported by Google Vision API. Our comparison illustrates that these snowy degradations could impede the actual performance of high-level vision tasks. And our method could significantly boost the performance of high-level vision tasks.

5.5 Ablation Study

For a reliable ablation study, we utilize the latest desnowing dataset,*i.e.*, CSD [4] dataset as the benchmark for training and testing in all ablation study experiments.

Table 5: Configurations of the proposed Channel Information Mining Block.

Metric	wo LN	GELU→ReLU	wo Channel Shuffle	Ours
CSD(2000) PSNR/SSIM	33.79/0.93	33.76/0.94	32.61/0.93	34.10 /0.95

Configurations of the Channel Information Mining Block In Table.5, we present the quantitative results of different configuration settings for CIMB. Specifically, we remove the Layer Normalization (**wo LN**), replace the GELU with ReLU (**GELU → ReLU**) and remove the channel shuffle operation(**wo Channel Shuffle**). From the results of Table.5, We found that LayerNorm is essential to stabilizing the training process, GELU and LN can provide a certain improvement on PSNR and SSIM, which is in line with NAFNet [2]. For the channel shuffle, we observe that it attracts an obvious gain compared with LN or GELU. Therefore, we believe that the channel interaction by channel shuffle operation benefits information mining in high-dimension space.

Table 6: Verification for the proposed External Attention Module.

Metric	wo Concat $I_{in}^{\downarrow}$	$I_{in}^{\downarrow} \rightarrow F_{in}$	ELU → ReLU	Ours
CSD(2000) PSNR/SSIM	32.16 /0.93	32.89 /0.94	33.96 /0.94	34.10 /0.95

Verification of key designs for the proposed External Attention Module. In Table.6, we present our verification about designs of the External Attention Module. (a)**wo Concat** $I_{in}^{\downarrow}$. To verify the effectiveness of introducing down-sampled original images, we remove the setting of image-feature fusion by concatence. (b)$I_{in}^{\downarrow} \rightarrow F_{in}$. To verify the key idea that generates an external attention map from the original degraded image to optimize current scale features, we replace the I_{in} with incoming F_{in}. (c) **ELU** $\rightarrow$ **ReLU**. We replace the ELU with $ReLU$ to explore the influence of the non-linear function on network performance. we remove the image-feature fusion by concatence. We believe that the clean external cues from the degraded image are key to improving performance instead of a simple attention mechanism. In addition, we demonstrate the effectiveness of external information from the results. Compared with the feature, the original degradation information can instruct the feature rebuilding and alleviate the information loss. We also notice ELU and ReLU non-linear activation almost have no impact on performance.

Table 7: Ablation study of the High-Order Rebuilding Sub-Net.

Block Num.	**0**	**6**	**12**	**Ours**
CSD(2000) PSNR/SSIM	**29.76 /0.93**	**32.41/0.93**	**33.69/0.94**	**34.10/0.95**
Params.(M)	**12.81**	**28.64**	**44.46**	**65.56**

Ablation study of the High-Order Rebuilding Sub-Net. In Table.7, we explore the influence of the depth of High-Oder Rebuilding Sub-Net. We found that deeper architecture has better performance on image desnowing. Due to 5 level down-sampling design, simply stacking more blocks cannot result in an unacceptable computational burden.

Table 8: Ablation study of the Efficient Pyramid Network.

Metric	**CIMB $\rightarrow$ SE-ResBlock**	**wo EAM**	**wo HOR Sub-Net**	**Ours**
CSD(2000) PSNR/SSIM	**32.94/0.94**	**31.79/0.93**	**29.76/0.93**	**34.10/0.95**

Ablation study of the Efficient Pyramid Network. To verify the effectiveness of each proposed component, we present comparison results in Table.8. (a) **CIMB** $\rightarrow$ **SE-ResBlock**. We replace the proposed CIMB with SE-ResBlock [7]. (b) **wo EAM**. We remove the proposed External Attention Module from our complete neural network. (d) **wo HOR Sub-set.** We remove the HOR Sub-net, and only reserve our encoder and decoder. Each proposed component is necessary for our Efficient Pyramid Network. We observe that the proposed basic

block CIMB has superiority compared to SE-ResBlock [7], which is based on the spatial and channel modeling, and the interaction between channels in high-dimension space. Besides, we also demonstrate the necessity of EAM and HOR Sub-Net. Our framework achieves SOTA performance in real time.

Effectiveness of Scale levels of Efficient Pyramid Network . We also carry out the discussion on the number of the semantic level and presented the results in Table.9. We explore different levels in our framework and demonstrate that using five semantic levels is best for our design. Specifically, too many scale levels can cause the feature size in the latent layer to be too small, so it will lose much information. On the other hand, too few scale levels will limit the speed of model inference due to resolution.

Table 9: Ablation study of Scale levels of Efficient Pyramid Network.

Metric	$\frac{1}{16}$x	$\frac{1}{64}$x	Ours
PSNR/SSIM	**33.11/0.94**	**32.08/0.93**	**34.10/0.95**

6 Limitations

Compared with previous lightweight desnowing methods, for instance, HDCW-Net (only 6.99M params.). The proposed Efficient Pyramid Network has better desnowing performance, but the much bigger parameters of our network will result in difficulty when deployed in edge devices.

7 Conclusion

In this work, we propose an Efficient Pyramid Network to handle High-Definition snow images in real time. Our extensive experiment and ablation study demonstrate the effectiveness of our proposed method and proposed blocks.

Although our method is simple, it is superior to all the previous state-of-art desnowing methods with a considerable margin on three widely-used large-scale snow datasets. We hope to further promote our method to other low-level vision tasks such as deraining and dehazing.

References

1. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing. vol. 2, pp. 168–172. IEEE (1994) 8

2. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022) 3, 5, 7, 12
3. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: European Conference on Computer Vision. pp. 754–770. Springer (2020) 2, 4, 8, 9, 11
4. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4196–4205 (2021) 2, 3, 4, 7, 8, 9, 11, 12
5. Engin, D., Genç, A., Kemal Ekenel, H.: Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 825–833 (2018) 9
6. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722 (2021) 5
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018) 5, 7, 13, 14
8. Huang, X., Ge, Z., Jie, Z., Yoshie, O.: Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10750–10759 (2020) 2
9. Lan, M., Zhang, Y., Zhang, L., Du, B.: Global context based automatic road segmentation via dilated convolutional neural network. Information Sciences **535**, 156–171 (2020) 2
10. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3175–3185 (2020) 2, 4, 9
11. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018) 2, 4, 9
12. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2056–2063 (2013) 2
13. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11908–11915 (2020) 5
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 2
15. Shafiee, M.J., Chywl, B., Li, F., Wong, A.: Fast yolo: A fast you only look once system for real-time embedded object detection in video. arXiv preprint arXiv:1709.05943 (2017) 2
16. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. Advances in neural information processing systems **26** (2013) 2
17. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. arXiv preprint arXiv:2111.14813 (2021) 2, 3, 8, 9, 11

18. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 5
19. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10551–10560 (2021) 7
20. Zheng, X., Liao, Y., Guo, W., Fu, X., Ding, X.: Single-image-based rain and snow removal using multi-guided filter. In: International Conference on Neural Information Processing. pp. 258–265. Springer (2013) 2