Dual Heterogeneous Complementary Networks for Single Image Deraining

Yuuto Nanba and Hikaru Miyata
Faculty of Science
Yamaguchi University, Japan
b037de@yamaguchi-u.ac.jp

Xian-Hua Han
Graduate School of Science and Technology for Innovation
Yamaguchi University, Japan
hanxhua@yamaguchi-u.ac.jp

Abstract

Single image deraining is an extreme challenge task since it requires to not only recover the spatial detail and high-level contextualized structure of the underlying image but also remove multiple rain layers with various blurring degrees and resolutions. Despite of the great performance advance with the deep learning networks, the dominated researches devote to either constructing deeper and complicated network architecture for recovering reliable detailed texture at the original resolution of the input image or exploiting multi-scale encoder-decode structure for learning semantic context in more larger receptive field while are still far from sufficiency to capture both complementary detailed and semantic contexts. This study proposes a novel dual heterogeneous complementary networks consisting of a main original resolution learning subnet and an auxiliary encoder-decoder subnet for exploring both detailed structure and semantic contexts. Specifically, to capture more plausible intermediate features in dual subnets, we concurrently evaluate the deraining losses of both branches in training phase, and exploit an auxiliary pseudo-label supervised attention module to further guide the feature learning in the main subnet. Moreover, to reconstruct more natural and sharper images, we incorporate multiple losses for network training including an improved MSE, an edge-based loss to recover reliable shape information, and a perceptual loss by evaluating the reconstruction error on the feature map of the learned VGGNet model instead of pixel intensity. Experiments on several benchmark deraining datasets demonstrate great superiority over the state-of-the-arts methods.

1. Introduction

Visibility degradation of images captured in adverse weather such as rain, haze, and fog, causes great loss of the desirable information, which yields harmful effect on the performance of various high-level vision tasks such as image classification, object detection, video surveillance, and aerial robots. To conquer the performance deterioration with the low quality rainy images, removal of the existing rain or raindrops in the rain-polluted observation has drawn considerable research attention in recent years [2, 11, 12, 14, 16, 17, 23, 28], and the existing methods are mainly categorized into traditional optimization-based and deep learning-based paradigms.

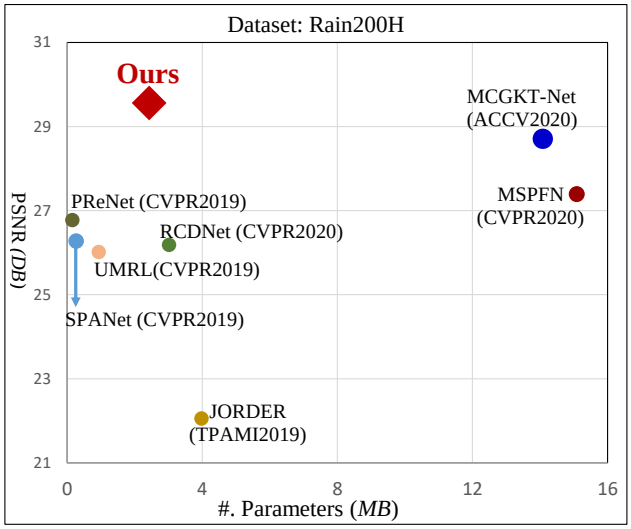


Figure 1. Deraining performance (PSNR) and model size (parameter number) comparisons between various deep models on the Rain200H dataset. Our proposed MCAN can achieve the best performance with the moderate model size.

Traditional methods generally exploit the specific prior knowledge to model the underlying structure of the clear image, and adopt various optimization strategies to iteratively restore the clear image. For example, Chen et al. [2] proposed visual depth guided rain streaks removal method by leveraging the sparsity prior while Luo et al. [16] explored discriminative sparse coding for single image deraining. Further, Kang et al. [12] investigated dictionary learning-based image decomposition for rain streak removal, and Li et al. [14] adopted Gaussian mixture models

to model rain layer priors. Although the deraining performance has been improved with the elaborated prior knowledge, the proper priors for various scenes are changeable and diverse, and the optimization procedure of the prior-regularized deraining model is also complicated.

Recently, driven by the powerful representation capability of deep convolution neural network (CNN), deep learning based methods have became the dominated paradigm for single image deraining, and promising performance has been achieved. Most methods concentrate on building more sophisticated network architectures [4, 5, 13, 19, 20, 22, 26, 34, 37] or designing robust better learning manners [11,17,23,28], and have continuously improved the deraining performance. The dominated pipeline of the current CNN models mainly adopts multiple convolution blocks (CB: Conv layer and activation function pairs) to learn representative feature in the original resolution of the input images, and serially pile up a large number of CBs to increase the network depth for capturing the context dependency in large respective field. However, these serially connected CB models do not explicitly capture the multi-scale representative features and different layers of rain structures, which are the intrinsic attributes of the existed rain in observation. To handle this issue, on one hand, several works [6, 10] investigated to exploit multi-scale information for image deraining, and mainly adopted multiple homogeneous subnets (several main streams) to model different scales of contexts in several synthesized resolution images, respectively. These multiple subnets would lead to more complicated network architectures and high computational cost. On the other hand, motivated by the multi-level learning characteristics of the convolution-based encoder-decoder (E-D) networks, some works [29] leveraged this compact network structure to effectively capture multi-scale contexts in a lightweight way. However, these E-D models utilize the coordinate depth of convolution blocks to learn the representative features on different scales, and may suffer from insufficient modeling capability at the original resolution of the input image since the deraining task aims at recovering all detailed information.

To solve the above limitations, this study proposes a novel dual heterogeneous complementary network (DHCN) for single image deraining. Specifically, the proposed network consists of two heterogeneous branches: one main subnet for exploiting the representative feature of the detailed structure at the original resolution of the input image and one auxiliary encoder-decoder subnet for capturing semantic context at multiple scales (resolutions). We establish the main subnet with multiple dual attention guided convolution blocks (DAGCB) to automatically select both effective and important channel of feature and spatial regions for recovering the detailed textures and spatial structures while construct the auxiliary subnet with an encoder-decoder structure to learn the complementary semantic contexts, where the learned multi-scale contexts of the auxiliary subnet have been aggregated with the intermediate features of the main subnet to guide more effective and robust learning in the following blocks. Despite of the employed two branches, our overview DHCN still maintains moderate model size due to the reduced channel number designing in the auxiliary subnet. Fig. 1 shows the compared deraining performances and model sizes between our method and the state-of-the-art (SoTA) deep models. Moreover, we utilize two respective losses to evaluate the learning capability of the dual subnets, and the estimated rain-free image in the auxiliary subnet is adopted to obtain an guided attention map for automatically emphasizing important and effective factors for deraining, dubbed as pseudo-label guided attention module (PLGAM). In addition, to restore more natural and sharper clear image, we incorporate multiple losses for network training including an improved MSE, and an edge-based loss to recover reliable shape information, and a perceptual loss by evaluating the reconstruction error on the feature map of the learned VGGNet model instead of pixel intensity. Experiments on several benchmark deraining datasets demonstrate great superiority over the state-of-the-arts methods.

In summary, our contributions mainly have three-fold:

1) We exploit a novel dual heterogeneous complementary deraining network, where the complementary features for the detailed spatial textures and the semantic contexts can be learned with the main subnet at the original resolution and the auxiliary encoder-decoder subnet at several levels of resolutions, respectively.

2) We leverage the learned semantic contexts and the predicted rain-free image in the auxiliary subnet to guide the feature learning procedure of the main subnet. Specifically, we exploit an aggregation module to transfer the multi-scale semantic contexts of the auxiliary subnet to the main subnet, and a pseudo-label-guided attention module to carry out supervision for the automatic emphasizing of the important and effective features in the main subnet.

3) We incorporate multiple losses for network training including an improved MSE, an edge-based loss and an perceptual loss to recover more nature and sharp rain-free image.

2. Related Work

The goal of the single image deraining is to recover a clear (rain-free) image from its rainy observation, and significant progress has been witnessed benefiting from the evolution of the deep convolution neural networks. This section briefly surveys the deep learning based single image deraining paradigms including the single scale deep models and multi-scale learning networks.

2.1. Single-scale deep models

Removal of rain streaks or drops in a single image is a fundamental but extreme challenging low-level vision task due to its ill-posed nature. Although traditional model-based methods have demonstrated acceptable deraining performance by exploiting various hand-crafted priors based on empirical observation, the employed priors still lack sufficient modeling capability to further boosting performance for diverse deraining scenarios. Recently, deep convolution networks have been popularly applied for single image dearining task, and manifested increasing performance progress via designing complicated and deep architecture. [3–5, 31, 37]. Fu et al. [4, 5] first explored a simple deep CNN to learn mapping relation between the rain-clear images and further extended the simple architecture via incorporating the deep residual-block and a global skip connection for performance boosting. These pioneering works are elaborated to remove the rain structure from the decomposed high-frequency components of the rainy images, and validated significant superiority over the model-based deraining methods. Since the deraining task is to recover all detailed structure and texture in the original scale of the inputted rainy images, the following dominated CNN architectures are evolved with the complex connection and effective module development of high-modeling capability on the single scale. Yang et al. [33] employed a recurrent dilated network for joint rain detection and removal network by recursively leveraging the stage-wise derained results while Li et al. [13] incorporated the squeeze-and-excitation block [9] into the recurrent neural network architecture to automatically learn important feature maps for rain removal. Zhang et al. [37] took the rain density into account to guide the network learning, and Qian et al. [19] automatically learned the attention map to guide the residual map generation. To enhance the visual quality of the derained results, adversarial learning [1] was also exploited by incorporating the discriminated loss with the reconstruction fidelity loss [18, 19] for pursuing more sharp and natural images. Further, Fan et al. [3] integrated a residual-guide network with recursive modules and multi-level supervision to progressively recover derained images whilst Wei et al [28] exploited a semi-supervised deraining paradigm to enhance the generalization capability on unseen rain types. Later, Ren et al. [20] proposed a better and simpler baseline deraining network via incorporating progressive ResNet, recurrent blocks inside and cross stages, and loss functions for boosting performance. Yang et al. [32] investigated a fractal band learning network to capture scale-robust rain features, and Yasarla et al. [35] aggregated the Gaussian process into a semi-supervised learning paradigm for image deraining task. To improve the interpretability, Wang et al. [24] integrated the convolution dictionary learning with the deep networks while Fu et al. [7] exploited a dual graph convolution network with two orthogonal graphs to carry out global relational modeling and reasoning for rain streak removal. In addition, meta-learning-based method [8] has also been investigated to construct a good generalization model for single image deraining. All these existing deep models operating on the original resolution of the input image have to stack many convolution blocks such as decades or hundreds to model long dependency, and thus generally yields large-scale model for performance lifting. This study employs a lightweight single-scale network with several specifically designed modules as the main branch while adopt an auxiliary encoder-decoder branch to capture multi-scale complementary contexts.

2.2. Multi-scale deep models

In order to model multiple rain streak layers and wide varieties of image contents, multi-stage methods, which are usually implemented in a progressive manner. [6, 36, 38], have been exploited for single image deraining. For example, Zheng et al. [38] employed three separate subnets to remove the heavy rain in a coarse-resolution level of the pyramid first and then reduce the light rain in the high-resolution level, where the encoder-decoder architectures were adopted in the two coarse levels while a plain residual CNN network was used in the high-resolution level. Jiang et al. [10] presented a multi-scale progressive fusion network (MSPFN) via collaboratively modeling the rain streaks from multiple scales with the pyramid representation. Most multi-stage paradigms generally adopt several independent subnets (several main-streams) to capture high-representative contexts on multi-scale input images, and unavoidably yield complicated and high-cost CNN models. Motivated by great success of the encoder-decoder architectures such as U-Net [21] and FCN [15] in semantic image segmentation, Yamamichi [29] introduced a simply-implemented and naturally multi-scale deraining framework with a single encoder-decoder subnet, and proposed a multi-level context gating knowledge transfer network (MCGKT-Net). To lift the deraining performance, this method leveraged the interactive learning between the learned features of encoder and decoder paths for internal knowledge transfer, and further incorporated the external knowledge in other task domains. Despite of the simple architecture, MCGKT-Net demonstrated promising deraining performance. However, MCGKT-Net employed the baseline encoder-decoder architecture by doubling the feature channel number along with the increased scale while the interactive learning adopted convLSTM unit for capturing the relation between features, where both would yield large number of parameters. Moreover, the identical convolution block are usually used in different scales of the encoder-decoder network and may cause insufficient representation capability in the original input resolution, which is crucial

to recover the detailed texture of the clear image. This study employs a lightweight encoder-decoder subnet with the reduced channel number in the coarse levels as an auxiliary branch for capturing multi-scale of contexts in larger receptive field while leverage a serially connected convolution blocks as the main branch for learning the fine and detailed texture context at the original resolution. With the dual heterogeneous branches, our proposed framework has the advantage of high modeling capability in complementary detail structure and high-representative contexts, and thus boosts the deraining performance.

3. Proposed dual heterogeneous complementary network

In this section, we present in detail the proposed dual heterogeneous complementary deraining network (DHCN). DHCN mainly consists of a shared shallow module, a main subnet for learning the representative features at the original resolution, an auxiliary encoder-decoder subnet to learn multi-scale semantic contexts for enhancing the complementary modeling capability of the main subnet, a simple aggregation module to transfer the multi-scale context of the auxiliary subnet to the main subnet, and a pseudo-label guided attention module, which leverages the predicted rain-free image in the auxiliary subnet to automatically learn a supervised attention map. The conceptual framework of our proposed DHCN is illustrated in Fig. 2.

On the whole, an input rainy image firstly passes through the shared shallow module with two vanilla convolution layers to transform the input RGB channel to features $\mathbf{X}_s$, and then the transformed features are imported into both main and auxiliary subnets, respectively, to further extract more representative contexts at various resolutions. Since the auxiliary subnet aims to assist the feature learning of the main subnet, the multi-scale contexts extracted from the encoder and decoder paths are aggregated with intermediate features of the main subnet to enhance the modeling capability. Moreover, the predicted rain-free image by the auxiliary subnet is leveraged to our proposed pseudo-label guided attention module to learn a supervision attention map for further selecting and emphasizing the important features while the output of the main subnet is as the final estimation of the rain-free image. What is more, in the dual main and auxiliary subnets, we incorporate multiple losses for our network training. Next, we substantiate the network architectures of the dual heterogeneous networks: the main and auxiliary subnets, the proposed pseudo-label guided attention module and the adopted multiple losses.

3.1. The dual heterogeneous networks

The existing deraining CNN models mainly follows two architecture designs: a single-scale feature modeling network and a multi-scale context exploitation pipeline with an encoder-decoder structure. The single-scale CNN model piles up a serial of convolution layers at the original resolution of the input, and aims to generate the reliable clear image with spatial details [11,17,23,28], which usually suffers from semantic robustness in the predicted output due to the limited receptive field. In contrast, the multi-scale exploitation models [6, 10] gradually decrease the resolution of the input to construct multi-scale representations, and employ similar convolution blocks on all scales for feature learning at the encoder path. Then, the decoder path progressively applies deconvolution/up-sampling operations to recover the original resolution, and exploits the semantic contexts using the convolution blocks on large receptive field. Although these encoder-decoder models have powerful capability of multi-scale information modeling, they are insufficient to capture spatial details due to the repeated use of resolution decreasing operation. To handle the inherent limitations of the aforementioned CNN models, this study proposes a novel dual heterogeneous network by leveraging the complementary feature modeling capability of two existing pipelines to generate both contextually reliable and spatially accurate rain-free image. Our dual heterogeneous networks are composed of the main and auxiliary subnets with different network architectures to exploit complementary features and contexts, where the main subnet aims to learn the representative features of the spatial details at the original resolution while the auxiliary subnet is used to learn multi-scale contexts at several levels of resolutions. Next, we will present the detailed architectures of the main and auxiliary subnets.

The main subnet: for simplicity, we pile up a serial of same blocks, dubbed as dual attention guided convolution blocks (DAGCB), to configure our main subnet, which can preserve all fine details from the input image to the predicted output. Since the main subnet aims to capture the spatial details at the same resolution as the input, no down-sampling or resolution-reduction operation has been used. Moreover, in order to automatically emphasize important and effective feature and regions, we incorporate both channel and spatial attentions with the vanilla convolution layer, and propose a dual attention guided convolution blocks (DAGCB) as the basic component of the main subnet. The schematic of DAGCB is illustrated in Fig. 3(a). Given a feature map $\mathbf{X}$, the DAGCB first transforms it to $\hat{\mathbf{X}}$ using two convolution layers, and then employs channel and spatial attention modules, which mainly consist of global average pooling (GAP), convolution layer and sigmoid activations, to learn both attention maps: $\mathbf{A}_C$ and $\mathbf{A}_S$. The channel and spatial attention guided feature maps: $\mathbf{X}_C$ and $\mathbf{X}_S$ can be expressed as:

$$\mathbf{X}_C = \hat{\mathbf{X}} \odot \mathbf{A}_C, \mathbf{X}_S = \hat{\mathbf{X}} \odot \mathbf{A}_S, \tag{1}$$

where $\odot$ denotes the element-wise multiplication. Finally,

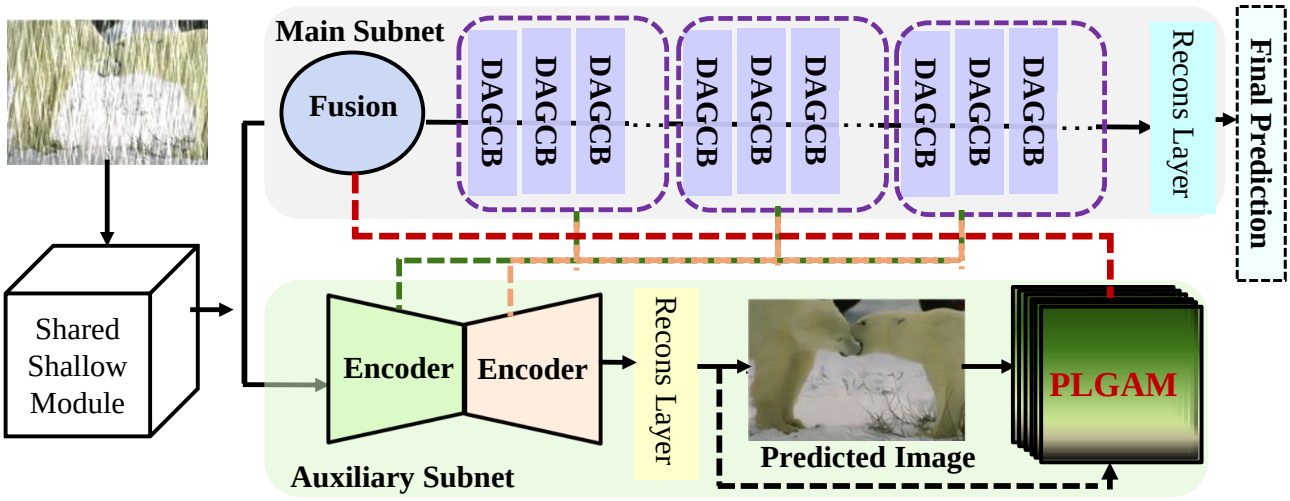
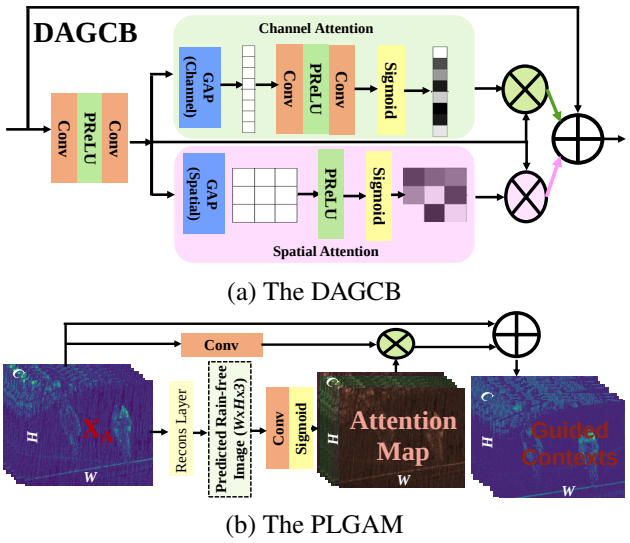


Figure 2. The conceptual architecture of our proposed DHCN.



(a) The DAGCB

(b) The PLGAM

Figure 3. The architecture of the dual attention guided convolution block (DAGCB) and the pseudo-label guide attention module (PLGAM).

the output of the DAGCB is the aggregated results of the input and the attention guided features:

$$\bar{\mathbf{X}} = \mathbf{X} + \mathbf{X}_C + \mathbf{X}_S \tag{2}$$

After multiple stacked DAGCBs, we adopt a reconstruction layer to estimate the final rain-free image.

The auxiliary subnet: We construct the auxiliary subnet using a Unet-like architecture, which mainly consists of symmetric encoder and decoder paths. In overview, both encoder and decoder paths in the auxiliary are composed of S blocks, and each block contains 2 convolution layers with 3*3 kernels following the PReLU activation function after each layer. Since the encoder-decoder subnet serves as an auxiliary role to exploiting multi-scale contexts instead of preserving all potential information in each scale, we retain similar channel number of the learned feature maps in all scales instead of channel doubling with scale increasing, which can greatly reduce the parameters in the auxiliary subnet. Moreover, we also employ a point-wise convolution layer instead of the simple skip connection to transform the feature map of the encoder to the corresponding decoder path.

In detail, given the output $\mathbf{X}_s$ of the shared shallow module, the auxiliary subnet firstly extracts the representative features $\mathbf{Y}_0^E$ at the $0-th$ order scale using two vanilla convolution layers, and then employs a down-sampling operation using a vanilla convolution with stride parameter 2 to obtain the first-order scale of input feature $\mathbf{X}_1^E = f_{Conv3s2}(\mathbf{Y}_0^E)$. Let's denote the output features of the encoder and decoder paths as $[\mathbf{Y}_0^E, \mathbf{Y}_1^E, \ldots, \mathbf{Y}_S^E]$ and $[\mathbf{Y}_0^D, \mathbf{Y}_1^D, \ldots, \mathbf{Y}_S^D]$, the input: $\mathbf{X}_s^D$ of $s-th$ scale block in the decoder path can be formulated as:

$$\mathbf{X}_s^D = f_{Cat}(f_{Pw}(\mathbf{Y}_{s+1}^E), f_{Up}(\mathbf{Y}_{s+1}^D)) \tag{3}$$

where f_{Up} represents the bilinear up-sampling operation, f_{Pw} is the point-wise convolution to transform the learned feature map to the corresponding decoder while f_{Cat} denotes the concatenation layer. Finally the output of $0-th$ order scale in the decoder path is imported to a reconstruction layer to provide an auxiliary estimation of the rain-free image. What is more, we also unify the size of the encoder and decoder's outputs: $[\mathbf{Y}_0^E, \mathbf{Y}_0^E, \ldots, \mathbf{Y}_S^E]$ and $[\mathbf{Y}_0^D, \mathbf{Y}_1^D, \ldots, \mathbf{Y}_S^D]$, and aggregate them together to be

feed-backed to three intermediate DAGCBs for providing the complementary contexts of the main subnet.

3.2. The pseudo-label guided attention module

As introduced above, the auxiliary subnet can estimate an intermediate rain-free image, denoting it as $\mathbf{G}_A \in \mathbf{R}^{W\times H\times 3}$, which can be potentially leveraged for further enforcing the modeling capability of the main subnet. This study exploits a pseudo-label guided attention module (PLGAM) for making full use of the auxiliary estimation and boosting the deraining performance of the main subnet. The schematic diagram of PLGAM is illustrated in Fig. 3(b). Since the output image $\mathbf{G}_A$ of the auxiliary subnet is estimated from the decoder's feature $\mathbf{Y}_0^D$ using the reconstruction layer via minimizing the reconstruction error of the ground truth image, and thus we dub it as pseudo label. The PLGAM aims at leveraging the pseudo label to automatically learn a supervision attention map for refining the auxiliary subnet's contextual feature $\mathbf{Y}_0^D$, and then the resulted attention guided context is aggregated with the shallow feature $\mathbf{X}_s$ as the input of the main subnet. Specifically, with the last stage output $\mathbf{Y}_0^D \in \mathbf{R}^{W\times H\times C}$ of the auxiliary subnet's decode, a reconstruction layer is adopted for providing an estimation $\mathbf{G}_A$ of the rain-free image. Next, per-pixel attention map based on the pseudo label $\mathbf{G}_A$ are generated using a point-wise convolution followed by the sigmoid activation, and the generated attention mask is then utilized to re-calibrate the transformed features from $\mathbf{Y}_0^D$ using a point-wise convolution, which results in an attention-guided feature map. With a residual connection structure, the attention-guided feature is added with $\mathbf{Y}_0^D$ to produce the output of the PLGAM, which is passed to the main subnet for further processing.

3.3. Multiple loss functions

Given the predicted rain-free images $\mathbf{G}_M$ and $\mathbf{G}_A$ of the main and auxiliary subnets and the corresponding ground-truth $\mathbf{G}$, we construct the loss using both predictions: $\mathbf{G}_M$ and $\mathbf{G}_A$ to optimize our DHCDN in an end-to-end learning manner as the followings:

$$\mathcal{L}_{Total} = \mathcal{L}_M(\mathbf{G}_M, \mathbf{G}) + \mathcal{L}_A(\mathbf{G}_A, \mathbf{G}) \tag{4}$$

In most works, mean squared error (MSE) is commonly used for formulate the losses: $\mathcal{L}_M$ and $\mathcal{L}_A$ to conduct network training. However, the simple MSE loss usually produces blurry and over-smoothed visual effect due to the loss of high-frequency textures with the squared penalty. In this study, we adopt an improved MSE loss, which is more tolerant to small errors and has better convergence property during training [20]. Further, to preserver more reliable edge information and natural image recovering, we also incorporate an additional edge loss to constrain the high-frequency component and a perceptual loss via feature comparison of pre-trained VGGNet instead of per-pixel comparison. The integrated loss is expressed as the following:

$$\mathcal{L}_* = \mathcal{L}_*^{MSE} + \mathcal{L}_*^{Edge} + \mathcal{L}_*^{Per} \tag{5}$$

where $*$ denotes M or A. Concretely, we formulate the multiple losses

$$\begin{aligned} \mathcal{L}_*^{MSE} &= \sqrt{(\mathbf{G}_* - \mathbf{G})^2 + \epsilon^2} \\ \mathcal{L}_*^{Edge} &= \sqrt{(Lap(\mathbf{G}_*) - Lap(\mathbf{G}))^2 + \epsilon^2} \\ \mathcal{L}_*^{Per} &= \sum_i \|\phi_i(\mathbf{G}_*) - \phi_i(\mathbf{G})\| \end{aligned} \tag{6}$$

where ϵ is a penalty coefficient, and is empirically set to 10^{-3} in our experiment. $Lap(\cdot)$ represents the edge maps using via the Laplacian operator while $\phi_i(\cdot)$ means the feature extraction of $i-th$ block from the pre-trained VGGNet using ImageNet dataset. In our experiment, we extract the feature map from 3, 4 and 5 convolution blocks of the VGGNet for computing the perceptual loss.

4. Experimental Results

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed dual heterogeneous complementary deraining network. We first present the experimental setting, and then provide the comparisons with the state-of-the-art deraining methods and ablation study.

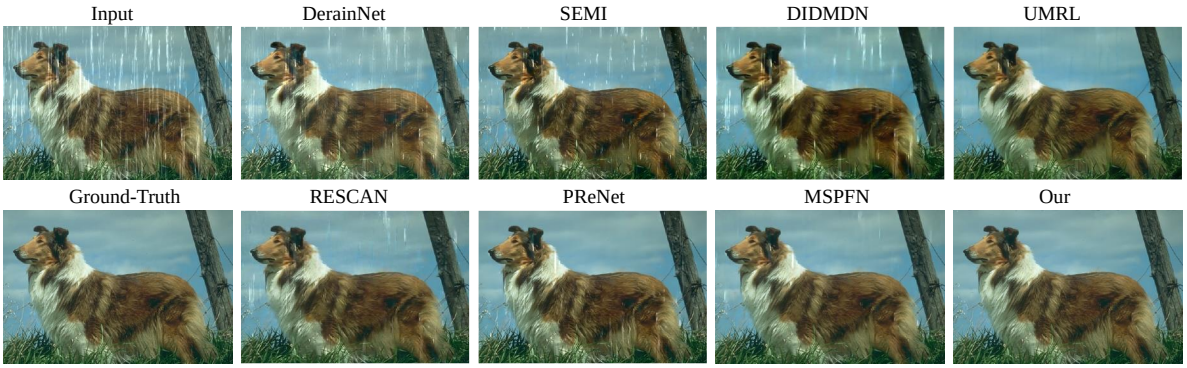
4.1. Experimental setting

Following the work [20], we carry out network training using the collected clean/rain images about 13700 pairs, and obtain a CNN model for evaluating the deraining performance on five rainy datasets: Rain 200L [31], Rain 200H [31], and Rain800 [20], Rain1200 [37] and Rain2800 [5]. The images in Rain200L has light rain and is relatively easy dataset while the images in Rain 200H are contaminated by more heavy rain with different shapes, directions, and sizes, and thus is the most challenging dataset in deraining community. The rainy images in Rain800 are synthesized by adding fine rain streak to the clean images, and have the fine-grained streaks with noise-like structures. The rainy images in Rain1200 dataset are generated with different levels of rainy density under light, medium and heavy rain conditions while the rainy images in the Rain2800 dataset are synthesized with 14 types of rain patterns for one clean image.

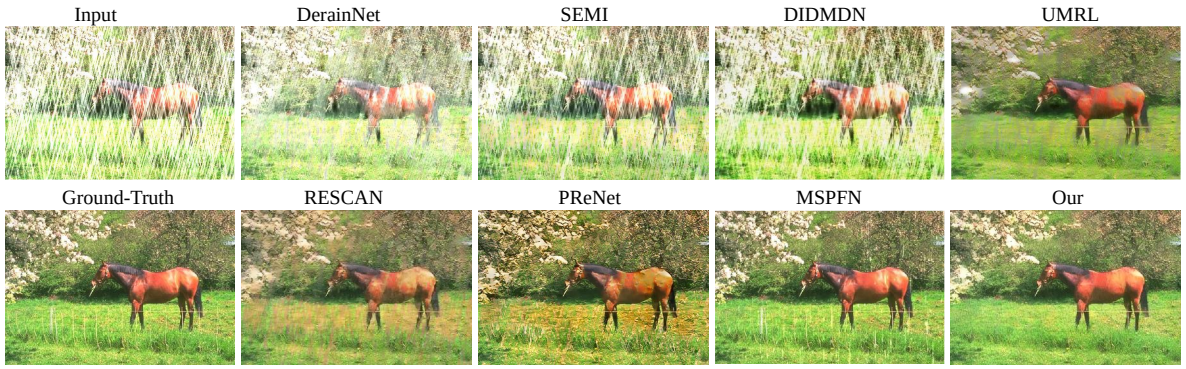
Two evaluation metrics, i.e. PSNR and SSIM [27], are adopted to assess the performance of our deraining method quantitatively. SSIM evaluates the image structure difference and is more consistent with human perceptual measure. We use pytorch to train and test our proposed method. In the training process, we crop 256×256 patches from the

Table 1. Average PSNR/SSIM comparison on five deraining datasets. Red and blue colors are used to indicate top 1^{st}, 2^{nd} performance.

Methods	Rain1200	Rain200L	Rain200H	Rain800	Rain2800	#.Parameters	Times (s)
Rainy	23.64/0.727	26.71/ 0.834	13.79/0.367	22.18/0.663	25.16/0.782	-	-
DerainNet [5]	23.39/0.832	34.46/0.957	26.11/0.792	22.78/0.803	24.31/0.861	58,175	0.61
JORDER [25]	24.32/0.862	34.95/0.959	22.05/0.727	22.24/0.776	29.03/0.888	4,169,024	0.43
SEMI [28]	26.06/ 0.822	25.03/0.842	16.56/0.486	22.35/0.788	24.43/0.782	-	-
DIDMDN [37]	29.66/0.899	35.40/0.962	26.61/0.824	22.53/0.812	28.13/0.867	135,800	0.53
RCDNet [30]	29.81/0.859	35.28/0.971	26.18/0.836	24.59/0.821	33.04/0.946	3,166,355	-
RESCAN [13]	30.54/0.879	29.80/0.881	26.75/0.835	24.99/0.830	31.29/0.904	499,668	0.61
UMRL [34]	30.55/0.910	29.18/0.923	26.01/0.832	24.41/0.829	29.97/0.905	984,356	2.02
PReNet [20]	31.49/0.910	32.44/0.950	26.77/0.858	24.79/0.849	31.75/0.916	168,963	0.156
SPANet [26]	31.84/0.900	35.79/0.965	26.27/0.865	22.41/0.838	28.46/0.880	283,716	1.72
MSPFN [10]	32.06/0.913	31.64/0.925	27.39/0.843	27.01/0.851	32.85/0.930	15,823,424	-
Ours	31.85/0.900	36.37/0.970	29.56/0.883	29.46/0.896	33.04/0.930	2,553,105	0.27



(a) one example image from the Rain200L dataset



(b) one example image from the Rain200H dataset

Figure 4. Compared visual results with the state-of-the-art models.

training samples, and adopt Adam to optimize our network. The networks are trained with 4×10^5 iterations, a batch size of 16, and the learning rate is set as 2×10^{-4}.

4.2. Quantitative Evaluation

Comparison with the SoTA models: We compare our proposed DHCN with the state-of-the-art methods, including deep detail network (DerainNet) [5], JORDER [25], semi-supervised transfer learning (SEMI) [28], density-aware deraining (DIDMDN) [37], the rain convolutional

Table 2. Ablation results.

Models	Rain200L		Rain200H		Rain800		Rain1200		Rain2800		Average		
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	#.Params
w/o SA	32.22	0.926	28.25	0.852	26.53	0.871	32.37	0.913	32.42	0.926	30.36	0.897	1,530,033
SinNet	33.90	0.944	28.68	0.857	28.27	0.876	31.56	0.910	32.74	0.929	31.03	0.903	1,530,058
+ $\mathcal{L}_*^{Per}$	35.33	0.965	28.92	0.876	27.56	0.880	31.70	0.912	32.59	0.927	31.22	0.912	1,530,058
+ PLGAM	33.24	0.937	28.69	0.856	28.14	0.873	31.20	0.910	32.84	0.930	30.97	0.901	1,552,058
+ PLGAM+ $\mathcal{L}_*^{Per}$	34.32	0.957	28.49	0.866	26.08	0.877	32.28	0.915	32.54	0.926	30.74	0.908	1,553,058
DualNet	34.29	0.953	29.06	0.865	28.27	0.880	32.53	0.914	32.90	0.931	31.41	0.909	2,531,105
+ $\mathcal{L}_*^{Per}$	35.44	0.960	29.45	0.882	27.94	0.880	32.18	0.908	32.70	0.929	31.54	0.912	2,531,105
+ PLGAM	34.66	0.953	29.45	0.869	28.93	0.883	32.07	0.909	32.99	0.932	31.61	0.909	2,553,105
+ PLGAM+ $\mathcal{L}_*^{Per}$	36.37	0.969	29.56	0.883	29.46	0.896	31.85	0.900	33.04	0.930	32.06	0.916	2,553,105

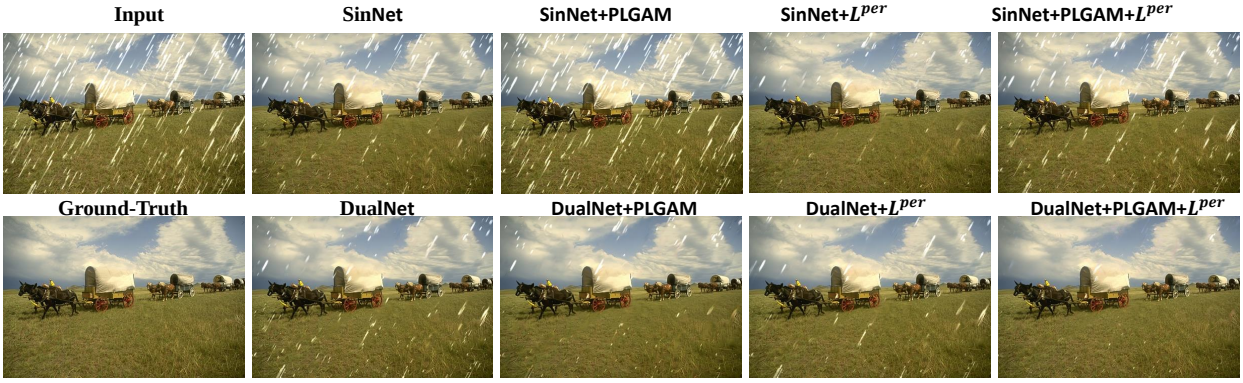


Figure 5. Compared visual results w/ or w/o the proposed modules in our DHCN.

dictionary network (RCDNet) [24], recurrent squeeze-and-excitation context aggregation net (RESCAN) [13], uncertainty guided multi-scale residual learning (UMRL) [34], progressive deraining network (PreNet) [20], spatial attentive network (SPANet) [26], and multi-scale progressive fusion network (MSPFN) [10]. The quantitative metrics, model sizes and running times for a 512×512 image using our proposed model and the compared deraining methods are manifested in Table 1. From Table 1, we can observe that our proposed DHCN has illustrated better or comparable SSIM and PSNR in all datasets. Moreover, the visual results with our network and different SoTA methods on two example images have been shown in Fig. 4. From Fig. 4, we can see that the proposed model can restore clearer results.

Ablation Study: Next, we evaluate the effectiveness of different proposed modules. Specifically, we denote the single main subnet as SinNet, and assess the deraining performance by incorporating different modules such as removal of the spatial attention (SA) in the DAGCB, the perceptual loss, the PLGAM via directly predicting the pseudo label from the shared shallow feature using a simple reconstruction block without the auxiliary subnet. It should noted that the simple reconstruction block for the pseudo label would cause unreliable estimation, and thus the incorporation with the SinNet only may degrade the deraining performance. Moreover, we dub the incorporated main and auxiliary subnets as DualNet, and then evaluate the performance w/ or w/o additional modules: PLGAM and the perceptual loss. The ablation results are illustrated in Table 2, which manifests our proposed DHCN achieves best performance. The visual results of one example image are given in Fig. 5.

5. Conclusions

This study proposed a novel dual heterogeneous complementary deraining network, which is composed of two subnets: a main original resolution feature learning subnet and an auxiliary subnet with encoder-decoder structure. To make full use of the intermediate prediction in the auxiliary subnet, we exploited a pseudo-label guided attention module for supervising the important information exploitation. Moreover, we also incorporated multiple losses for our end-to-end network learning. Extensive experiments demonstrated that our proposed method achieved superiority performance over the SoTA methods.

References

[1] S. Bengio, I. J. Goodfellow, and A. Kurakin. Adversarial machine learning at scale. *arXiv:1611.01236*, 2017. 3

[2] D. Chen, C. Chen, and L. Kang. Visual depth guided color image rain streaks removal using sparse coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 24:1430–1455, 2014. 1

[3] Z.W. Fan, H.F. Wu, X.Y. Fu, and Y. Huang. Residual-guide feature fusion network for single image deraining. *ACMMM*, 2018. 3

[4] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26:2944–2956, 2017. 2, 3

[5] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley. Removing rain from single images via a deep detail network. *CVPR*, 2017. 2, 3, 6, 7

[6] X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley. Lightweight pyramid networks for image deraining. *TNNLS*, 31:1794 – 1807, 2019. 2, 3, 4

[7] X.Y. Fu, Q. Qi, Z.J. Zha, Y.R. Zhu, and X.B. Ding. Rain streak removal via dual graph convolutional network. *AAAI*, 2021. 3

[8] X. Gao, Y. Wang, J. Cheng, M. Xu, and M. Wang. Meta-learning based relation and representation learning networks for single-image deraining. *Pattern Recognition*, 120:108124, 2021. 3

[9] J Hu., S. Li, and G. Sun. Squeeze-and-excitation networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[10] K. Jiang, Z.y. Wang, P. Yi, C. Chen, B.j. Huang, Y.m. Luo, J.y. Ma, and J.j. Jiang. Multi-scale progressive fusion network for single image deraining. *CVPR*, page 8346–8355, 2020. 2, 3, 4, 7, 8

[11] X. Jin, Z.b. Chen, J.x. Lin, Z.k. Chen, and W. Zhou. Unsupervised single image deraining with self-supervised constraints. *ICIP*, page 2761–2765, 2019. 1, 2, 4

[12] L. Kang, C. Lin, and Y. Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE Transactions on Image Processing*, 21:1742–1755, 2012. 1

[13] X. Li, J.l. Wu, Z.c. Lin, H. Liu, and H.b. Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. *ECCV*, pages 254–269, 2018. 2, 3, 7, 8

[14] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown. Rain streak removal using layer priors. *CVPR*, page 2736–2744, 2016. 1

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[16] Y. Luo, Y. Xu, and H. Ji. Removing rain from a single image via discriminative sparse coding. *ICCV*, page 3397–3405, 2015. 1

[17] P. Mu, J. Chen, R.s. Liu, X. Fan, and Z.x. Luo. Learning bilevel layer priors for single image rain streaks removal. *IEEE Signal Processing Letters*, 16:307–311, 2019. 1, 2, 4

[18] J.C. Pu, X.S. Chen, L. Zhang, Q.H. Zhou, and Y. Zhao. Removing rain based on a cycle generative adversarial network. *ICIEA*, 2018. 3

[19] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu. Attentive generative adversarial network for raindrop removal from a single image. *CVPR*, page 2482–2491, 2018. 2, 3

[20] D.w. Ren, W.m. Zuo, Q.h. Hu, P.f. Zhu, and D.y. Meng. Progressive image deraining networks: a better and simpler baseline. *CVPR*, page 3937–394, 2019. 2, 3, 6, 7, 8

[21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, page 234–241, 2015. 3

[22] G.q. Wang, C.m. Sun, and A. Sowmya. Erlnet: Entangled representation learning for single image deraining. *ICCV*, page 5644–5652, 2019. 2

[23] H. Wang, Y.c. Wu, Q. Xie, Q. Zhao, Y. Liang, S.j. Zhang, and D.y. Meng. Structural residual learning for single image rain removal. *Knowledge-Based Systems*, page 106595, 2020. 1, 2, 4

[24] H. Wang, Q. Xie, Q. Zhao, Y. Liang, and D.Y. Meng. Rcdnet: An interpretable rain convolutional dictionary network for single image deraining. *arXiv:2107.06808*, 2021. 3, 8

[25] H. Wang, Q. Xie, Q. Zhao, and D.Y. Meng. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:1377–1393, 2020. 7

[26] T.y Wang, X. Yang, K. Xu, S.z. Chen, Q. Zhang, and R. WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. *CVPR*, page 12270–12279, 2019. 2, 7, 8

[27] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612. 6

[28] W. Wei, D.y. Meng, Q Zhao, Z.b. Xu, and Y. Wu. Semi-supervised transfer learning for image rain removal. *CVPR*, page 3877–3886, 2019. 1, 2, 3, 4, 7

[29] K. Yamamichi and X.-H. Han. Mcgkt-net: Multi-level context gating knowledge transfer network for single image deraining. *ACCV*, page 68–83, 2020. 2, 3

[30] W.H. Yang, R. T. Tan, J.A. Feng, Z.M. Guo, S.C. Yan, and J.Y. Liu. Rcdnet: a model-driven deep neural network for single image rain removal. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[31] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan. Deep joint rain detection and removal from a single image. *CVPR*, 2017. 3, 6

[32] W.H. Yang, S.Q. Wang, and J.Y. Liu. Removing arbitrary-scale rain streaks via fractal band learning with self-supervision. *IEEE Transactions on Image Processing*, 42:6759–6772, 2020. 3

[33] Y.Z. Yang, W. Ran, and H. Lu. Rddan: A residual dense dilated aggregated network for single image deraining. *ICME*, 2020. 3

[34] R. Yasarla and V. M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. *CVPR*, page 8405–8414, 2019. 2, 7, 8

[35] R. Yasarla, V. A. Sindagi, and V. M Patel. Syn2real transfer learning for image deraining using gaussian processes. *CVPR*, 2020. 3

[36] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Multi-stage progressive image restoration. *ACM Multimedia Conference*, page 14821–14831, 2018. 3

[37] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. *CVPR*, page 695–704, 2018. 2, 3, 6, 7

[38] Y. P. Zheng, X. Yu, M. M. Liu, and S. L. Zhang. Residual multiscale based single image deraining. *BMVC*, 2019. 3