Spatially-Adaptive Image Restoration using Distortion-Guided Networks

Kuldeep Purohit[1] Maitreya Suin[2] A. N. Rajagopalan[2] Vishnu Boddeti[1]
[1] Michigan State University [2] Indian Institute of Technology Madras

Abstract

We present a general learning-based solution for restoring images suffering from spatially-varying degradations. Prior approaches are typically degradation-specific and employ the same processing across different images and different pixels within. However, we hypothesize that such spatially rigid processing is suboptimal for simultaneously restoring the degraded pixels as well as reconstructing the clean regions of the image. To overcome this limitation, we propose SPAIR, a network design that harnesses distortion-localization information and dynamically adjusts computation to difficult regions in the image. SPAIR comprises of two components, (1) a localization network that identifies degraded pixels, and (2) a restoration network that exploits knowledge from the localization network in filter and feature domain to selectively and adaptively restore degraded pixels. Our key idea is to exploit the non-uniformity of heavy degradations in spatial-domain and suitably embed this knowledge within distortion-guided modules performing sparse normalization, feature extraction and attention. Our architecture is agnostic to physical formation model and generalizes across several types of spatially-varying degradations. We demonstrate the efficacy of SPAIR individually on four restoration tasks- removal of rain-streaks, raindrops, shadows and motion blur. Extensive qualitative and quantitative comparisons with prior art on 11 *benchmark datasets demonstrate that our degradation-agnostic network design offers significant performance gains over state-of-the-art degradation-specific architectures.*

1. Introduction

Images are often degraded during the data acquisition process, especially under non-ideal imaging conditions. Such degradations can be attributed to the medium and dynamics between the camera, scene elements and the illumination. For instance, as shown in Fig. 1 , (1) precipitation leads to snow/rain streaks occupying the volume between the scene and the camera, (2) presence of rain-drops on the camera lens causes significant degradation in scene visibility, (3) relative motion between the camera or scene elements results in motion blur, and (4) harsh illumination conditions can induce harsh shadows. Despite the disparate source of degradations, they share the same underlying motif that affects the image quality, namely, degradation that is spatially-varying in nature. For example, raindrops and shadows degrade monolithic parts of the image depending on their size and location, motion blur varies with scene depth and degree of motion, and rain streaks effect only sparse regions whose orientation depends on the relative rain direction. Fig. 1 shows representative examples of degraded images and respective distortion-maps. It can be seen that a large number of pixels undergo little or no distortion. Another observation is that the amount of distortion and its spatial distribution is different in every image.

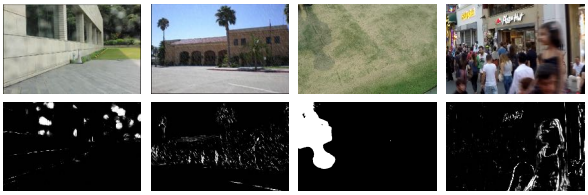


Figure 1. Visualization of degradation masks. The two rows show degraded input images and corresponding predicted masks.

Restoring such images is vital to improve their aesthetic quality as well as the performance of downstream tasks, viz, detection, segmentation, classification, and tracking. Convolutional neural networks (CNNs) currently typify the state-of-the-art for various image restoration tasks. Despite recent progress, existing approaches share several key limitations. Firstly, all layers in their networks are generic CNN layers, which apply the same set of spatially-invariant filters to every degraded image. Such layers are limited in their ability to invert degradations that are highly image-dependent and spatially-varying. Secondly, most network architectures are specifically tailored for individual degradation types as they are based on image formation models. Thirdly, the distortion-localization information embedded in the labeled datasets remains unused or sub-optimally used in all existing solutions.

Static CNN based models trained to directly regress clean intensities from degraded ones, perform poorly when input contains unaffected regions as well as severe inten-

sity distortions in different spatial regions. Conceptually, a stack of fixed learned filters, excelling at restoring pixels degraded with large distortions might not be suitable for reconstructing the texture from unaffected regions. Practically, we observe that such designs often yield poor reconstruction performance (introduce unwanted changes or artifacts on pixels that are not degraded in the input to begin with). The image-dependent nature of spatial distribution and magnitude of distortions only exacerbates the problem faced by static CNNs.

Motivated from the understanding that a restoration network can benefit from adapting to the degradation present in each test image, we propose a distortion-aware model to simultaneously realize the twin goals of restoration and reconstruction. Our spatially-adaptive image restoration architecture (referred to as SPAIR) is suited for any type of degradation which selectively affects parts of the image. It comprises of two components- a distortion-localization network (Net_L) and a spatially-guided restoration network (Net_R). Net_L gathers information from the entire image to estimate a binary mask (localizing high intensity distortions) which steers the processing in Net_R to selectively improve only degraded regions.

The proposed Net_R comprises of 3 distortion-guided blocks- spatial feature modulator (SFM), sparse convolution module (SC) and a custom sparse non-local module (SNL). SFM utilizes the output mask and intermediate features from Net_L to modulate the feature statistics of intermediate features in Net_R. SC and SNL improve features in the spatially-sparse degraded regions in an image-dependent manner, without affecting the features in clean regions. SNL locally restores features in distorted regions by adaptively gathering global context from all clean regions. Our key contributions are:

– A two-stage framework to systematically exploit distortion-localization knowledge for directly addressing the challenges associated with diverse spatially-varying degradations in an interpretable manner. It achieves the twin goals of restoration and reconstruction and works across diverse degradation-types.

– Distortion-guided spatially-varying modulation of features statistics in Net_R with the help of distortion-mask and features from a pretrained Net_L.

– Distortion-guided feature extraction with the help of SC (for local context) and a novel SNL (for global context) modules. These components facilitate spatially-varying restoration while controlling receptive field in an image and location-adaptive manner.

– We demonstrate the versatility of SPAIR by setting new state-of-the-art on **11** synthetic and real-world datasets for various spatially-varying restoration tasks (removing rain-streaks, rain-drops, shadows, and motion blur), outperforming existing approaches designed with task-specific network-engineering. Further, we provide detailed analysis, qualitative results, and generalization tests.

2. Related Works

Adaptive Inference: Adaptive inference techniques [56, 50, 12, 27] have attracted increasing interest since they enable input-dependent alteration of CNN structure. One class of methods dynamically skip subsets of layers in cascaded CNNs during inference [58, 50, 5]. [6] passes sampled pixels (using a random pattern which is fixed during inference) to CNN layers and fills the remaining locations using simple interpolation. Few approaches [13, 12] exploit sparsity in the input image itself using sub-manifold sparse convolutions, but are unsuitable for non-sparse input data.

However, none of these approaches afford the fine-grained spatial-domain control necessary for spatially-varying image restoration at multiple intermediate layers. For instance, the approaches that skip processing of some layers or prune the network still filter the degraded and other image regions with the same parameters. Methods such as [5] are only applicable to cascade of consecutive residual layers, and do not generalize to encoder-decoder designs (typically used for image restoration) where conditionally altering network depth or channel width is non-trivial. The arbitrary rejection of spatial-domain information proposed in [6] is ill-fitted for general restorations tasks.

Raindrop Removal: Solutions for raindrop removal include both classical as well as CNN based approaches. [22] proposed a clustering and median filtering based restoration, while CNN based approaches include, shallow CNNs [4] but with limited performance, a convolutional-LSTM based model for "joint" learning of rain-map and rain-free image [38], and a deeper CNN [31]. [40] instead leveraged physical models of raindrop properties (including closedness and roundness) to estimate drop-probability. In contrast to these methods, SPAIR advocates for a pixel selective and adaptive processing to remove raindrops.

Rain-streak Removal: Conventional deraining methods [3, 72, 29, 33] adopt a model-driven methodology utilizing physical properties of rain and prior knowledge of background scenes into an optimization problem. CNN-based approaches include end-to-end negative residual mapping [8], deeper CNN [62], multi-stage CNNs with recurrent connections [28], CNN for predicting density (heavy, medium, light) during deraining [67], concatenating rain-map for deraining [61]. However, layers in these approaches process all image regions with the same filters (without pixel adaptation). [52] presents model-driven CNN with convolutional dictionary learning.Wang et.al. [54] predicts a rain-map and multiplies it element-wise with feature-maps to enhance them. While SPAIR also utilizes a mask, there are fundamental differences. We estimate a binary mask and utilize it more comprehensively, including

for sparse filtering, attention weight calculation and guiding it to non-degraded image regions. SPAIR significantly differs from rain-guided models of [61, 38] in three aspects. (1) They only concatenate the rain-mask at the input. In contrast, we exploit distortion-mask to only perform convolutions and non-local operations on degradation regions. We also transfer feature statistics from clean to degraded regions at multiple intermediate layers using SFM. (2) They lack global context. SPAIR contains SNL module that adaptively gathers all features values within the clean regions of the image. (3) All pixels are passed through same network with spatially rigid processing, which directly contrasts with our work. Ours is the first approach to exploit explicit degradation-guidance to selectively processes degraded pixels and reduce the effect on unaffected regions, for a variety of spatially-sparse degradations.

Shadow Removal: Early works often erased shadows via user interaction or by transferring illumination from non-shadow regions to shadow regions [14, 23]. More robust results have been achieved using CNN based approaches which include using multiple networks [16], DeshadowNet for illumination estimation in shadow regions [39], stacked conditional GANs [53], ARGAN to detect and remove shadow with multiple steps [2], RIS-GAN [71] to estimate negative residual images and inverse illumination maps for restoration, and finally a cascade of dilated convolutions to jointly estimate shadow-mask and shadow-free image [1]. In contrast to the aforementioned approaches, we propose a two-stage framework wherein the distortion-mask and intermediate learned features of Net_L are employed in a principled manner for region-aware and selective restoration.

Motion Blur Removal: Traditional approaches [26] designed priors on image and motion (eg. locally linear blur kernels were explored in [48, 10]) but with limited success in general 3D and dynamic blurred scenes. Recent CNN-based methods directly estimate the latent sharp image [36], wherein encoder-decoder designs that aggregate features in a coarse-to-fine manner have been proposed [36, 49, 9]. Additionally, [69] explored a design composed of multiple CNNs and RNN and [65] proposed a patch-hierarchical network and stacked its copies along depth to achieve state-of-the-art performance. [37] proposed a recurrent design for efficient deblurring. A limitation shared by all of these methods is the absence of spatially varying adaptive layers. [47] inserts adaptive convolution and attention within the layers of [65] to boost its results. Our distortion-guided sparse architecture performs better than such patch hierachical designs, while generalizing beyond motion-blur and offering consistent gains across other degradations.

Architectures for general Restoration A few solutions have been proposed in literature to address multiple degradation-types. For instance, DuRN [32] make task dependent alterations in their network structure. Similarly, OWAN [46] was proposed to handle multiple degradations present within the same image. However, [46] only addresses simple synthetic degradations that are similar in nature eg. gaussian blur, noise, and jpeg artifacts. SPAIR demonstrates its efficacy on realistic datasets of several physically unrelated degradations which are heavily spatially-varying. In such settings, DuRN and OWAN are quite inferior to our model, as shown in our experiments.

3. Proposed Network Architecture

An image restoration model needs to solve two equally important tasks: (1) locating the areas to restore in an image, and (2) applying the right filtering mechanism to the corresponding regions. While Net_L addresses the former, we realize the latter through a spatially-guided restoration network Net_R. A schematic of SPAIR is shown in Fig. 2. The knowledge from intermediate features of pre-trained Net_L improves Net_R's training, while the mask itself lends adaptiveness to the restoration process. To realize the twin goals of restoration and reconstruction, distortion-guided filtering of the extracted features in Net_R is enabled through SFM (Spatial Feature Modulator), SC (Sparse Convolution), and SNL (Sparse Non Local) modules.

3.1. Distortion Localization Network (Net_L)

To maximize the generalizability of our approach, we adopt the U-Net topology [44] as our CNN backbone (both for localization and restoration networks). Different versions of this are known to be effective for several restoration tasks such as image deblurring [49], denoising [34], and general image-to-image translation [19]. We build a densely connected encoder-decoder structure whose detailed layer-wise description is given in the supplementary. This design delivers competitive performance across all tasks considered and hence acts as a backbone for our Net_R (see Sec. 6). Net_L is a lightweight version of Net_R (with similar structure) since the binary classification (localization) task is simpler than the intensity regression (restoration) task.

Given a degraded image, Net_L produces a single channel mask and is trained using binary cross-entropy loss to match the GT binary mask. For datasets with no ground truth mask, we use the absolute difference between degraded image and clean image, and threshold it to obtain a binary mask, classifying pixels into degraded (value 1) or clean (value 0). Empirically, we observed that Net_R's performance improves when Net_L is trained to predict only the pixels with severe distortions (as opposed to detecting even minute intensity changes). Note that the distortion-map directly correlates with the difficulty of restoration, and it may differ from the physically occuring degradation-distribution. For instance, when physical rain-steaks are equally distributed throughout the image, the distortion-map would contain more non-zero values in the urban tex-

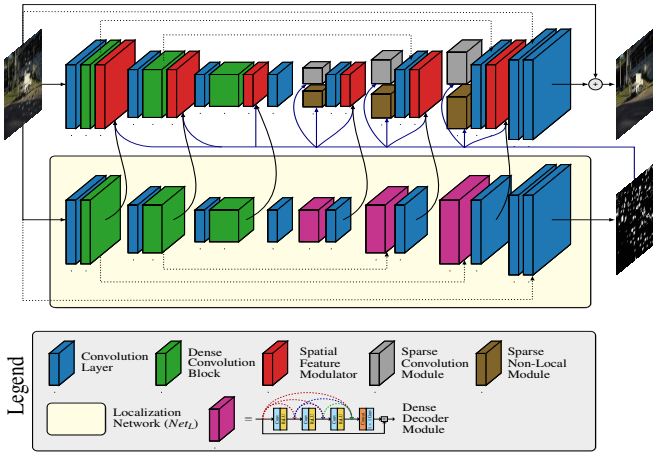


Figure 2. Proposed SPAIR and its components. Net_R is shown at the top and Net_L is shown at the bottom. Connection between the two networks (for SFM) are shown using black arrows.

tured regions than the sky regions (since white rain-streaks do not significantly alter the bright intensities in sky).

3.2. Spatially-Guided Restoration Network (Net_R)

As depicted in Fig. 2, Net_R extracts a features pyramid from input degraded image using a cascade of densely connected layers [17]. These features are fed to decoder that generates the restored image. Although correction of small intensities changes can be learnt by simple convolutional layers (basic building block for all prior works), they struggle for spatially-distributed heavy degradations. For such regions, localization based guidance improves restoration quality. We propose 3 modules to employ trained Net_L to convey localization knowledge to Net_R.

Since image generation process requires decoder to learn both reconstruction and restoration, each level of the decoder contains an SC and an SNL module. Note that we refrain from using SC or SNL in the encoder layers of Net_R since that would completely discard the degraded image intensities (which contain partially-useful information). We employ SFM at multiple levels to perform distortion-guided feature normalization to complement SC and SNL modules.

3.2.1 Spatial Feature Modulator (SFM)

SFM fuses the features of Net_R with intermediate features from layers of the pretrained Net_L in an additive manner. We observe that with such feature guidance, early layers of Net_R extract more distortion-aware features that correlate strongly to the degradation-variation within the input image. Since both the networks share a similar encoder-decoder structure, the inputs of all strided convolution layers are fused using SFM, as shown in Fig. 2.

In CNNs, feature normalization is known to be important and complementary to feature extraction. The role of SFM is to perform distortion-guided spatial-varying feature normalization. This complements the distortion-guided feature extraction process using local (SC) and global (SNL) context. SFM module performs adaptive shifting of the feature statistics at degraded locations, which aids the restoration process. Studies [21] show that feature mean relates to global semantic information while variance is correlated to local texture. Inspired from this, our SFM modulates features at degraded locations to match the feature statistics (mean and variance) of clean regions.

Given the fused features F and the predicted mask $\mathcal{M}$, we calculate the modulated features F^S as

$$F^S = \sigma(F, (1-\mathcal{M}))\left(\frac{F \odot \mathcal{M} - \mu(F, \mathcal{M})}{\sigma(F, \mathcal{M})}\right) + \mu(F, (1-\mathcal{M}) \quad (1)$$

The mean operator is $\mu(Q, \mathcal{M}) = \frac{1}{\sum_p \mathcal{M}_p} \sum_p Q_p \odot \mathcal{M}_p$ and the standard deviation is $\sigma(Q, \mathcal{M}) = \sqrt{\frac{1}{\sum_p \mathcal{M}_p} \sum_p (Q_p^2 \odot \mathcal{M}_p - \mu(Q, \mathcal{M})) + \epsilon}$, where subscript p represents 2D pixel location. Since modulation of features is desired only at the degraded locations, the feature output of SFM is $F^S \odot \mathcal{M} + F \odot (1-\mathcal{M})$.

3.2.2 Mask-guided Sparse Convolution (SC)

As discussed earlier, filters of general convolution layers are spatially-invariant and hence are forced to learn the restoration and reconstruction tasks jointly, which impedes the training process and reduces model's performance. SPAIR harnesses the efficacy of mask-guided sparse convolution that facilitates selective restoration of highly degraded regions, and simplifies the learning process. SC (shown in Fig. 2) contains a densely connected set of 6 guided sparse convolution layers followed by a 1×1 convolution to reduce the number of channels. Each unit in SC takes a the input feature map, F and the predicted mask $\mathcal{M}$. Pixels masked as 1 in $\mathcal{M}$ are sampled, and passed through a convolution operation, resulting in a sparse feature map F^S as

$$F_p^S = \begin{cases} 0 & \mathcal{M}_p = 0 \\ \sum_{p' \in R_k} K'_p F_{p+p'} & \mathcal{M}_q, \mathcal{M}_p = 1, \end{cases} \quad (2)$$

where R_k indicates the support region of kernel offsets with kernel size k (e.g., for a 3×3 convolution, $R_k = \{(-1,-1),(-1,0),...,(1,1)\}$ and $k = 3$), and $K \in \mathbb{R}^{C_{in} \times C_{out} \times k \times k}$ denotes convolution weights. Although SC is quite effective for the spatially-varying task at hand, its receptive field is limited to only degraded pixels. We next describe our SNL module which extracts features using global context-aggregation (with distortion-guidance) and complements the role of SC.

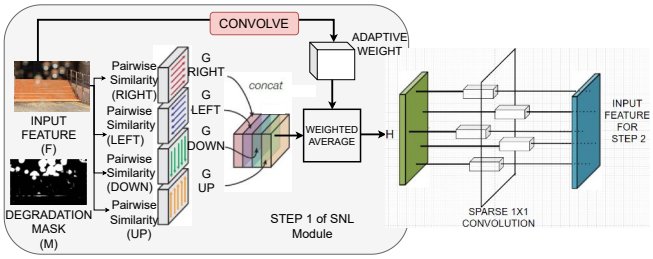


Figure 3. Region-Guided Sparse Non-Local (SNL) Module for degradation-guided context aggregation. Sparse 1x1 convolution connects the two structurally-identical sparse-attention steps.

3.2.3 Region-Guided Sparse Non-Local Module (SNL)

Most computer vision tasks are inherently contextual. Frequently used tools for gathering larger context such as dilated convolution [64], global average or attention pooling [15], or multi-scale methods [49] etc. can enlarge the receptive field beyond simple convolutional layers. Yet, they are not image-adaptive and still cannot utilize the full feature map effectively. In contrast, a single non-local layer [55] is capable of extending the receptive field to the maximum $H \times W$ size adaptively for each image and each pixel in an image. We claim that such a property is well-suited for a spatially-varying restoration model, where the heavily corrupted regions benefit from the ability to gather relevant features from the whole image. Effectiveness of adaptive global context aggregation has also been explored for recognition/segmentation tasks in [55, 41].

We hypothesize that within a restoration model, the non-local context aggregation process can benefit immensely from the knowledge of degraded pixel locations. We intuit that propagating heavily degraded information throughout the spatial domain can be counter-productive. Ideally, an adaptive module should learn to completely ignore irrelevant features, but recent vision models (e.g. image captioning [18]) have shown that this behavior is not practically achieved. They resort to perform additional filtering to remove unnecessary information.

In contrast, the proposed SNL module leverages the distortion mask to control the scope of non-local context aggregation. While restoring degraded pixels, it assigns dynamically estimated non-zero weights to features from only not/less degraded pixel locations, delivering superior performance as it dampens the influence of heavily degraded/corrupted information. Moreover, SNL leaves the features of clean regions unaltered as this operation is performed only on degraded pixel locations.

As illustrated in Fig. 3, SNL comprises of an efficient two-step aggregation approach, with each step comprising of horizontal-vertical scanning of the feature matrix in four fixed directions: left-to-right, top-to-bottom, and vice-versa. Having two steps is important in harvesting full-image contextual information from all pixels. While directional scanning of CNN features has been explored in literature [51, 30], SPAIR introduces a region-guided and sparse non-local module.

We elaborate on the feature aggregation process within first step of the SNL module for horizontal direction (it can be similarly derived for other directions as well). We denote the value at a particular location (i, j) in the feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ as $\mathbf{f}_{i,j} \in \mathbb{R}^{C}$. To model its relationship with all other valid locations (ensuring $\mathcal{M}_{i,j} = 0$) on the right $\mathbf{F}^{right}_{i,j} \in \mathbb{R}^{C \times (W-i)}$, we calculate a pairwise relationship matrix $\mathbf{o}^{right} \in \mathbb{R}^{W-i}$ using softmax as

$$\mathbf{o}^{right} = \text{softmax}(\mathbf{f}_{i,j} \odot \mathbf{F}^{right}_{i,j}) \tag{3}$$

This matrix is then used to weigh the contribution of the features towards the right ($\mathbf{F}^{right}_{i,j}$) as

$$\mathbf{g}^{right}_{i,j} = \mathbf{F}^{right}_{i,j} \odot \mathbf{o} \tag{4}$$

where $\mathbf{g}^{right}_{i,j} \in \mathbb{R}^{C}$. Note that locations with $\mathcal{M}_{i,j} = 0$ are skipped during the above operations and the four directions are parallely executed in the CUDA implementation. Finally, the features from four directions are fused using pixel-wise adaptive weights. These weights $\mathbf{E} \in \mathbb{R}^{4 \times H \times W}$ are generated by feeding the feature $\mathbf{F}$ to another convolution layer. The fused features $\mathbf{h}_{i,j}$ are obtained as

$$\mathbf{h}_{i,j} = \sum_{k \in \Omega}^{4} e^{k}_{i,j} \odot \mathbf{g}^{k}_{i,j} \tag{5}$$

where $e^{k}_{i,j} \in \mathbb{R}^{1}$ is the (k, i, j)-th element of $\mathbf{E}$ and $k \in \{left, right, up, down\}$. The entire process is repeated twice (Fig. 3) to allow each pixel to gather global context.

Sparse 1×1 convolution: To perform feature-refinement between two steps of the SNL module, we introduce a sparse 1×1 convolution. As shown in the subfigure Fig. 3, on the feature locations of interest specified by the binary mask, a point-wise feature representation is extracted. A fully connected layer then accepts and refines the entire stack of these point-wise features. This replaces the 2D convolution on $H \times W$ spatial grid with point-wise 1D convolution on the selected points and facilitates sparse processing.

4. Datasets and Implementation Details

Rain-Streaks: Using the same experimental setups of the recent approaches on image deraining [20], we train our model on 13,712 clean-rain image pairs gathered from multiple datasets [8, 29, 62, 67, 68]. With this single trained model, we perform evaluation on different test sets, including Rain100H [62], Rain100L [62], Test100 [68], Test2800 [8], and Test1200 [67]. We also report the error reduction error for each method relative to the best method

Table 1. Image deraining results. Best and second best scores are **highlighted** and underlined. For each method, relative MSE reduction achieved by SPAIR is reported in parenthesis (see Section 4 for calculation). SPAIR achieves ∼22% improvement over MSPFN [20].

Methods	Test100 [68]		Rain100H [62]		Rain100L [62]		Test2800 [8]		Test1200 [67]		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DerainNet [7]	22.77	0.810	14.92	0.592	27.03	0.884	24.31	0.861	23.38	0.835	22.48 (69.3%)	0.796 (61.3%)
SEMI [57]	22.35	0.788	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.88 (67.8%)	0.744 (69.1%)
DIDMDN [67]	22.56	0.818	17.35	0.524	25.23	0.741	28.13	0.867	29.65	0.901	24.58 (60.9%)	0.770 (65.7%)
UMRL [63]	24.41	0.829	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910	28.02 (41.9%)	0.880 (34.2%)
RESCAN [28]	25.00	0.835	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882	28.59 (37.9%)	0.857 (44.8%)
PreNet [42]	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	29.42 (31.7%)	0.897 (23.3%)
MSPFN [20]	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75 (21.9%)	0.903 (18.6%)
SPAIR (Ours)	**30.35**	**0.909**	**30.95**	**0.892**	**36.93**	**0.969**	**33.34**	**0.936**	**33.04**	**0.922**	**32.91** (0.0%)	**0.926** (0.0%)

Table 2. Quantitative comparisons of models trained and tested on the SPANet [54] and the Rain100H [62] benchmarks.

Dataset	Metric	DSC	GMM	Clear	DDN	RESCAN	PReNet	SPANet	JORDER$_E$	RCDNet$_1$	RCDNet	**SPAIR**
SpaNet [54]	PSNR	34.95	34.30	34.39	36.16	38.11	40.16	40.24	40.78	40.99	41.47	**44.10**
	SSIM	0.9416	0.9428	0.9509	0.9463	0.9707	0.9816	0.9811	**	0.9816	0.9834	**0.9872**
Rain100H [62]	PSNR	13.77	15.23	15.33	22.85	29.62	30.11	25.11	30.50	30.91	31.28	**31.69**
	SSIM	0.3199	0.4498	0.7421	0.7250	0.8720	0.9053	0.8332	0.8967	0.9037	0.9093	**0.9201**

by translating PSNR to RMSE (RMSE $\propto \sqrt{10^{-\text{PSNR}/10}}$) and SSIM to DSSIM (DSSIM $= (1 - \text{SSIM})/2$). We also evaluate SPAIR on SPANet Dataset [54] (real-world rain) containing 2×10^5 training and 1000 testing images.

Raindrop: We use the AGAN dataset [38] with 861 training and 58 test samples. Images were generated by placing a raindrop covered glass between the camera and scene.

Shadow: We evaluate our model using a challenging benchmark ISTD [53] containing 1300 (train) and 540 (test) images (with real shadows and diverse textured scenes).

Motion Blur: We follow the configuration of [47, 66, 25, 49] and use the GoPro [35] dataset containing 2,103 image pairs for training and 1,111 pairs for evaluation. Furthermore, to demonstrate generalizability, we directly evaluate our GoPro trained model on the test set of HIDE [45] and RealBlur [43] datasets. The HIDE dataset is specifically collected for human-aware motion deblurring, containing 2,025 test images. While the GoPro and HIDE datasets are generated by averaging real videos, the blurred images in RealBlur-J dataset are captured in real-world conditions.

Implementation Details [1]**:** The Net_R for each degradation is trained to minimize l_1 reconstruction loss between the output and the GT clean image. Net_L is trained using binary cross entropy loss with respect to the GT binary mask. Each training batch contains randomly cropped RGB patches of size 256×256 from degraded images that are randomly flipped horizontally or vertically. The batch-size was 8 for rain-streak, raindrop, and shadow removal and 16 for deblurring. Both networks use Adam optimizer with initial leaning rate 10^{-4}, halved after every 50 epochs. We use PyTorch library and RTX 2080Ti GPU.

[1]We will publicly release our implementation

Table 3. Raindrop removal results on testset from Qian et al. [38].

Method	Eigen [4]	Pix2pix [19]	AGAN[38]	DuRN[32]	Quan[40]	SPAIR
PSNR	28.59	30.59	31.51	31.24	31.44	**32.73**
SSIM	0.6726	0.8075	0.9213	0.9259	0.9263	**0.9410**

5. Experimental Evaluation

Rain-streak Removal: Following prior art [20], we perform quantitative evaluations (PSNR/SSIM scores) on the Y channel (in YCbCr color space). Table 1 reports the results across all five datasets where SPAIR consistently achieves significant gains over the baselines. Compared to the recent algorithm MSPFN [20], we obtain a performance gain of 2.16 dB (averaged across all datasets). Next, for fair comparison with RCDNet [52], we evaluate SPAIR in their setting (in Table 2) by training and testing on the challenging Rain100H [62] and SPANet [54] (captured in real-world rainy scenes) datasets. While the improvement is modest (0.41 dB) on very heavy rain (Rain100H), it is as large as 3 dB on datasets with low rain density, eg. SPANet and Rain100L (since in this case, we selectively process very few pixels without affecting clean pixels), highlighting the advantage of our distortion-adaptive restoration.

Fig. 4 presents qualitative comparisons on challenging images from Rain100H dataset. Our results exhibit significantly higher visual quality than existing methods which fail to recover background textures (1^{st} row), introduce artifacts (2^{nd} row). SPAIR is robust to changes in scenes and rain densities as it effectively removes rain streaks of different orientations and magnitudes, and generates images that are visually pleasing and faithful to the ground-truth.

Raindrop Removal: Table 3 and Fig. 5 show qualitative and visual comparisons with recent methods [38, 40, 32]. SPAIR outperforms the baselines by a large margin. Our results are visually closer to GT and perceptually better than those of competing methods which often contain artifacts or

Input JORDER [62] Fu et al. [8] RESCAN [28] PReNet [42] SPANet [54] RDCNet [52] SPAIR GT

Figure 4. Qualitative comparison of zoomed-in results on synthetic rainy images from the Rain100H test-set.

Input AGAN [38] DuRN [32] SPAIR GT Input AGAN [38] DuRN [32] SPAIR GT

Figure 5. Qualitative comparisons of results on images from the AGAN testset [38].

Table 4. Shadow removal results on ISTD Dataset [53]. Subscripts S and NS indicate shadow and non-shadow regions, respectively.

Metric	Input	[60]	[14]	[11]	[53]	[16]	[71]	[1]	SPAIR
$RMSE_S$	32.12	19.82	18.95	14.98	10.33	9.48	8.99	8.14	**8.05**
$RMSE_{NS}$	7.19	14.83	7.46	7.29	6.93	6.14	6.33	6.04	**5.47**
RMSE	10.97	15.63	9.30	8.53	7.47	6.67	6.95	6.37	**5.88**

Table 5. Deblurring results. Our method is trained only on the GoPro dataset [35] and directly applied to the test images of HIDE [45] and RealBlur-J [43] datasets. PSNR‡ scores were obtained after training and testing on RealBlur-J dataset.

Method	GoPro [35] PSNR	GoPro [35] SSIM	HIDE [45] PSNR	HIDE [45] SSIM	RealBlur-J [43] PSNR	RealBlur-J [43] SSIM	RealBlur-J [43] PSNR‡
Xu *et al*. [59]	21.00	0.741	-	-	27.14	0.830	
DeblurGAN [24]	28.70	0.858	24.51	0.871	27.97	0.834	
Nah *et al*. [35]	29.08	0.914	25.73	0.874	27.87	0.827	
Zhang *et al*. [69]	29.19	0.931	-	-	27.80	0.847	
DeblurGAN-v2 [25]	29.55	0.934	26.61	0.875	28.70	0.866	29.69
SRN [49]	30.26	0.934	28.36	0.915	28.56	0.867	31.38
Shen *et al*. [45]	-	-	28.89	0.930	-	-	
DBGAN [70]	31.10	0.942	28.94	0.915	-	-	
MT-RNN [37]	31.15	0.945	29.15	0.918	-	-	
DMPHN [66]	31.20	0.940	29.09	0.924	28.42	0.860	
Suin *et al*. [47]	31.85	0.948	29.98	0.930	-	-	
SPAIR	**32.06**	**0.953**	**30.29**	**0.931**	**28.81**	**0.875**	**31.82**

color distortions.

Shadow Removal: We evaluate our shadow-removal model against traditional [14, 11, 60] and learning based methods including ST-CGAN [53], DSC [16], DeShadowNet [39]. Following prior art, results are evaluated in Lab color space using RMSE scores calculated over shadow and non-shadow regions. Fig. 6 and Table 4 show that although CNN-based designs are better than hand-crafted methods, most existing approaches produce shadow boundaries or color inconsistencies. However, SPAIR has minimal artifacts in the shadow boundaries, outperforming the baselines both qualitatively and quantitatively.

Deblurring: We validate our distortion-guided approach for general motion deblurring on 3 benchmarks: GoPro [35], HIDE [45], and the real-world blurred images of a recent RealBlur-J [43]. We report the quantitative comparisons with the existing deblurring approaches in Table 5. Overall, SPAIR performs favorably against other algorithms. Note that inspite of training only on the GoPro, it outperforms all methods including [45] on HIDE, without requiring any human bounding box supervision, thereby demonstrating its strong generalization capability.

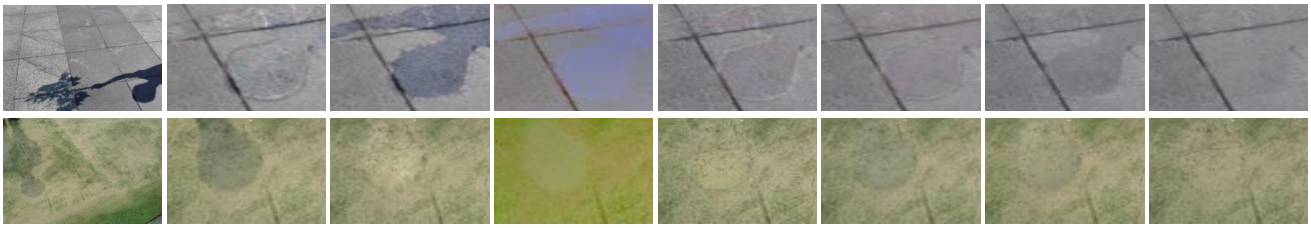
We evaluate models on RealBlur-J [43] testset under two experimental settings: 1) training on GoPro (to test generalization to real images), and 2) training on RealBlur-J. SPAIR obtains performance gain of 0.39 dB over the DMPHN model [66] in setting 1, and 0.44 dB over existing best method for setting 2. Our model's effectiveness is owed to the robustness of the distortion-aware approach.

Visual comparisons on images containing dynamic and 3D scenes are shown in Fig. 7. Often, the results of prior works suffer from incomplete deblurring or artifacts. In contrast, our network demonstrates non-uniform deblurring capability while preserving sharpness. Scene details in the regions containing text, boundaries, and textures are more faithfully restored, making them recognizable.

6. Network Analysis

This work explores the benefits of distortion-localization guided feature modulation and sparse processing for spatially-varying restoration tasks. Table 6 quantifies the effect of individual design choices on performance of SPAIR on the AGAN (raindrop) and GoPro (motion blur) datasets.

To validate our design choices, we implement the following baselines (reported in Table 6). Net1: Dense encoder-decoder network (CNN backbone of our Net_R) with few additional parameters to match Net_L. Net2: Net1 guided by Net_L using SFM. Net3: Net2 with all densely connected



Input [14] [11] [60] [53] [16] [1] SPAIR

Figure 6. Comparison of shadow removal results on ISTD Dataset [53]. Shadow region and boundaries are visible in existing approaches.

Blurred Image Blurred patch SRN DelurGAN-V2 Stack(4)-DMPHN MTRNN Suin et al. SPAIR

Figure 7. Visual comparisons of zoomed-in results of competing deblurring models on images from the GoPro test set [36].

Table 6. Network analysis with PSNR for AGAN and GoPro benchmarks, respectively. DED, SFM, SC, NL, SNL denote dense encoder-decoder, spatial feature modulator, mask-guided sparse convolution, non-local module, and mask-guided sparse non-local module, respectively.

Methods	DED	SFM	SC	NL	SNL	PSNR	
						Raindrop	Motion Blur
Net1	✓					30.72	30.84
Net2	✓	✓				31.48	31.55
Net3	✓	✓	✓			31.83	31.62
Net4	✓	✓	✓	✓		32.19	31.81
SPAIR	✓	✓	✓		✓	**32.73**	**32.06**

convolutional blocks in decoder replaced with SC modules, Net4: Net3 with non-local (NL) layer [55] introduced in the decoder. Net5: Net4 containing the proposed SNL module instead of NL. Good baseline scores of Net1 for both tasks support our backbone design choice.

Effectivenes of SFM: Net2 introduces SFM blocks (Sec. 3.2.1) which guide restoration network using mask and features of Net_L at multiple intermediate levels. The significant improvement in accuracy in comparison to Net1 demonstrates the benefit of degradation guidance.

Effect of SC and SNL modules: Net4 employs the general non-local layer [55] in decoder global context aggregation. Net5 has the same structure as Net4 (sans the NL module), and it feeds the predicted mask as input to the SNL modules. The improvement in behavior and performance is attributed to SNL design which uses explicit distortion-guidance to steer pixel-attention. SNL is more suited than NL for both degraded and clean regions. As explained in Eq. 4, while restoring degraded pixels, SNL assigns dynamically estimated non-zero weights to features originating from only clean pixels in the image. By design, it leaves the features of clean regions unaltered. As reported in Table 6, Net3 vs. Net2 shows the benefits of SC module whereas, Net5 vs. Net3 shows the utility of global context aggregation for restoration. Net5, our final model, shows a significant improvement over CNN baseline (Net1), demonstrating the advantages of our overall solution over static CNNs.

Supplementary Details: We provide additional real results and qualitative comparisons for all four tasks, additional model analysis, detailed operation of modules and layer-wise description of the network in supplementary.

Benefit: Many applications (e.g., autonomous vehicles) involve dealing with rain, shadows, blur etc. at different time instances. Designing architectures that are applicable across multiple tasks, without requiring specialized architecture re-engineering is practically very convenient (potentially facilitating customized hardware design). Our versatile design enables this as only the learned weights vary across degradations while the architecture remains the same.

7. Conclusions

We addressed the single image restoration tasks of removing spatially-varying degradations such as raindrop, rain streak, shadow, and motion blur. We model the restoration task as a combination of degraded-region localization and region guided sparse restoration and propose a guided image restoration framework SPAIR which leverages the features of Net_L for spatial modulation of the features in Net_R using SFM module. We introduce distortion localization awareness in Net_R using sparse convolution module (SC) and sparse non-local attention module (SNL) and show its significant benefits . Extensive evaluation on 11 datasets across four restoration tasks demonstrates that proposed framework outperforms strong degradation-specific baselines. Ablation analysis and visualizations are shown to validate the effectiveness of its key components.

References

[1] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, pages 10680–10687, 2020. 3, 7, 8

[2] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10213–10222, 2019. 3

[3] Xinghao Ding, Liqin Chen, Xianhui Zheng, Yue Huang, and Delu Zeng. Single image rain and snow removal via guided l0 smoothing filter. *Multimedia Tools and Applications*, 75(5):2697–2712, 2016. 2

[4] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 633–640, 2013. 2, 6

[5] Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2017. 2

[6] Mikhail Figurnov, Aizhan Ibraimova, Dmitry P Vetrov, and Pushmeet Kohli. Perforatedcnns: Acceleration through elimination of redundant convolutions. In *Advances in Neural Information Processing Systems*, pages 947–955, 2016. 2

[7] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *TIP*, 2017. 6

[8] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 2, 5, 6, 7

[9] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, 2019. 3

[10] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *CVPR*, 2017. 3

[11] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, pages 1–11. Citeseer, 2014. 7, 8

[12] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2

[13] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 2

[14] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012. 3, 7, 8

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[16] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 3, 7, 8

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 4

[18] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 5

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3, 6

[20] Kui Jiang, Zhongyuan Wang, Peng Yi, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020. 5, 6

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[22] M Ramesh Kanthan and S Naganandini Sujatha. Rain drop detection and removal using k-means clustering. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–5. IEEE, 2015. 2

[23] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):431–446, 2015. 3

[24] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. DeblurgGAN: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 7

[25] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 6, 7

[26] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1709, 2016. 3

[27] Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang, and Gao Huang. Improved techniques for training adaptive deep networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1891–1900, 2019. 2

[28] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018. 2, 6, 7

[29] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *CVPR*, 2016. 2, 5

[30] Sifei Liu, Jinshan Pan, and Ming-Hsuan Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *European Conference on Computer Vision*, pages 560–576. Springer, 2016. 5

[31] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 2

[32] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 3, 6, 7

[33] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *ICCV*, 2015. 2

[34] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, pages 2802–2810, 2016. 3

[35] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 6, 7

[36] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, volume 1, page 3, 2017. 3, 8

[37] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, 2020. 3, 7

[38] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018. 2, 3, 6, 7

[39] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4075, 2017. 3, 7

[40] Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2463–2471, 2019. 2, 6

[41] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Standalone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. 5

[42] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019. 6, 7

[43] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 6, 7

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[45] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, 2019. 6, 7

[46] Masanori Suganuma, Xing Liu, and Takayuki Okatani. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2019. 3

[47] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 2020. 3, 6, 7

[48] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015. 3

[49] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 3, 5, 6, 7

[50] Ravi Teja Mullapudi, William R Mark, Noam Shazeer, and Kayvon Fatahalian. Hydranets: Specialized dynamic architectures for efficient inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2018. 2

[51] Francesco Visin, Marco Ciccone, Adriana Romero, Kyle Kastner, Kyunghyun Cho, Yoshua Bengio, Matteo Matteucci, and Aaron Courville. Reseg: A recurrent neural network-based model for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48, 2016. 5

[52] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2020. 2, 6, 7

[53] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 3, 6, 7, 8

[54] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12270–12279, 2019. 2, 6, 7

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 5, 8

[56] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018. 2

[57] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *CVPR*, 2019. 6

[58] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8817–8826, 2018. 2

[59] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, 2013. 7

[60] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012. 7, 8

[61] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1377–1393, 2019. 2, 3

[62] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 2, 5, 6, 7

[63] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *CVPR*, 2019. 6

[64] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 5

[65] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019. 3

[66] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 6, 7

[67] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, 2018. 2, 5, 6

[68] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *TCSVT*, 2019. 5, 6

[69] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018. 3, 7

[70] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, 2020. 7

[71] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12829–12836, 2020. 3, 7

[72] Lei Zhu, Chi-Wing Fu, Dani Lischinski, and Pheng-Ann Heng. Joint bi-layer optimization for single-image rain streak removal. In *Proceedings of the IEEE international conference on computer vision*, pages 2526–2534, 2017. 2

Spatially-Adaptive Image Restoration using Distortion-Guided Networks SUPPLEMENTARY DOCUMENT

Kuldeep Purohit[1] Maitreya Suin[2] A. N. Rajagopalan[2] Vishnu Boddeti[1]
[1] Michigan State University [2] Indian Institute of Technology Madras

Supplementary Details: This document is organized as follows. In Section S1, we provide additional analysis and visualization of various aspects of our model. Experiments on multi-task learning are provided in Section S2, while Section S3 contain additional qualitative results and comparisons on various benchmark datasets for all four tasks. The document concludes with detailed operation of modules in Section S4.

S1. Model Analysis

S1.1. Effect of Spatially-Selective Processing

We perform a few analytical experiments to verify the effectiveness of spatially-selective processing. We vary the threshold (for intensity change to be classified as significant distortion) and calculate the PSNR on thus obtained degraded and non-degraded regions separately. Owing to the adaptiveness of our approach and specialized modules for the current task, we achieve better restoration results for the regions with significant corruption (Fig. S4). This behavior is generally expected from better restoration network design. Interestingly, we observed that, for regions with negligible amount of degradation (non-degraded regions for very low threshold), the input image itself is sufficiently good (Fig. S4). Due to generic processing, most of the previous methods are unable to reconstruct those regions accurately (i.e. they corrupt the background pixels) and result in poor PSNR even in the simplest of regions.

Next, we visualize the error-maps (the difference between degraded and ground-truth image) for the non-degraded regions in Figs. S5 (RainStreak) and Figs. S6 (Raindrop). As we can observe, our method results in the least amount of reconstruction error. The improvement over existing methods in the non-degraded areas shows that SPAIR causes least amount of changes in the non-degraded pixels. Improved restoration in degraded regions is attributed to modules specifically processing the degraded regions. Our method preserves input details which are uncorrupted and is able to improve the quality of the restoration of affected regions.

S1.2. Effect of SFM

We had observed (in Table 6 of main paper) that SFM module is well suited to the spatially-varying task and further improved the performance (Net7). For the rain-streak removal, we have visualized feature statistics of degraded and non-degraded regions in Fig. S7. In the left image, we can see a significant difference in the statistics before SFM. As rainy regions are usually unnaturally brighter, the mean value is much higher than in non-rainy areas. This implies that if we apply a global normalization, the statistics of non-rainy regions will be adversely affected. On the other hand, SFM based normalization only operates on rainy regions and applies an affine transform on it guided by the mean and variance of non-rainy regions. It offers two advantages: a) statistics of degraded regions gets modulated to match the characteristics of non-degraded regions while preserving the content, b) adverse effects of rain on non-rainy regions will be minimized. We can observe in the figure on the right that the statistics of rainy regions have moved closer to the non-rainy regions, resulting in improved performance.

S1.3. Attention Visualization

We have visualized predicted degradation-mask and pixel-level attention maps for raindrop and motion-blur affected images in Fig. S1 and for rain-streak affected images in Fig. S2. As SNL allows a pixel to gather relevant global context adaptively, we can observe that the degraded pixel is mainly focusing on less-corupted regions with similar texture, color structure, etc., which contribute to restoration process. The SNL module brings an improvement in model's behavior and performance by using explicit distortion-guidance to steer pixel-attention. SNL is more suited than general NL for both degraded and clean regions.

S2. Exploring Multi-Task Learning

We also explore an additional benefit of our deisgn. Since our architecture design does not change across restoration tasks, it opens venues for multi-task learning. We perform expriments on jointly learning two restoration

tasks. Among the 4 spatially-varying degradations we consider in this paper, we choose to address Rain-Streaks and RainDrops jointly,s as they are closely related to each other and generally occur in similar environments. We explore the possibility of obtaining a single trained model which can remove raindrops as well as rainstreaks from a given test image. We train SPAIR jointly on the datasets of two tasks: Mixed RainStreak Dataset [9] (used in Table 1 of main paper) and RainDrop Dataset [13] (used in Table 3 of main paper). After training jointly on the two tasks, we evaluate the model on the two benchmarks and compare against existing task specific methods. We also include the recent multi-task restoration model OWAN [17] as a baseline and train it in the same setting as SPAIR.

Specifically, we train our model on 14573 clean-degraded image pairs gathered from rain-streak datasets [4, 11, 22, 24, 25] and the raindrop dataset [13]. With this single trained model (referred to as SPAIR (Joint)), we perform evaluation on different test sets, including Rain100H [22], Rain100L [22], Test100 [25], Test2800 [4], Test1200 [24] and AGAN [13]. The results on the two tasks are reported in Tables S1 and S2. It is evident that SPAIR (Joint) is the first model in literature to achieve state-of-the-art results on Rain-Streak and RainDrop removal tasks, without additional training.

Figure S1. Visualization of pixelwise-relation map for different degraded images. Red box denotes the pixel gathering information. It gathers information from non-degraded pixels in similar textured regions.

S3. Analysis on Distortion Localization Network (Net_L)

We have visualized predicted distortion-mask along with the ground truth distortion-mask in Fig. S3. The close resemblance of the predicted and ground truth degradation maps shows the effectiveness of Net_L.

Supervised vs unsupervised learning: We choose to train Net_L in a supervised fashion, since accurate pixel-level distortion estimation is of key importance in restoration.

Impact of Accuracy of Net_L: Note that, compared to the ground truth mask, there will inevitably be some errors in the predicted one. Although our proposed modules in the decoder leverage the extra guidance from the predicted mask, as there are standard convolution layers too in the decoder and the restoration network is trained with the predicted mask itself, errors in very few pixels of the pre-

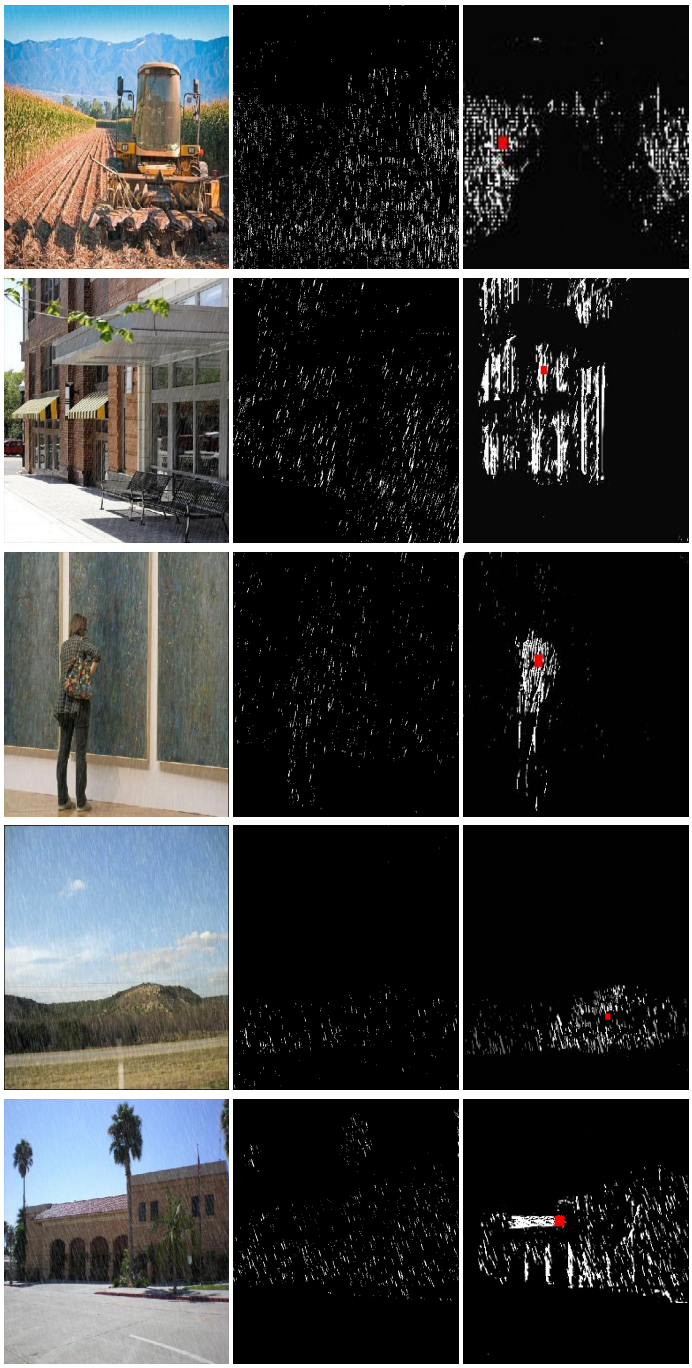


Figure S2. Visualization of attenton on rainstreak-affected images. From left to right: Input image, degradation-mask, and attention-map for a pixel (marked in red) estimated within SNL module.

Table S1. Image deraining results using SPAIR trained jointly for Rain-Streak and Raindrop removal tasks. Best and second best scores are **highlighted** and underlined. SPAIR significantly outperforms baselines methods in both settings: Single task and Joint task learning.

Methods	Test100 [25]		Rain100H [22]		Rain100L [22]		Test2800 [4]		Test1200 [24]		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
MSPFN [9]	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75	0.903
SPAIR (Single)	**30.35**	**0.909**	**30.95**	**0.892**	**36.93**	**0.969**	**33.34**	**0.936**	33.04	0.922	**32.91**	**0.926**
OWAN [17] (Joint)	23.85	0.810	24.46	0.724	28.54	0.878	30.40	0.891	30.09	0.872	27.47	0.835
SPAIR (Joint)	30.33	0.909	30.81	0.892	36.39	0.964	33.34	0.936	**33.10**	**0.925**	32.79	0.925

Table S2. Raindrop removal results (AGAN Dataset [13]) using SPAIR trained jointly for Rain-Streak and Raindrop removal tasks. SPAIR yield better results than existing methods in both settings: Single task and Joint task learning.

Method	Eigen [2]	Pix2pix [8]	AGAN[13]	DuRN[12]	Quan[14]	SPAIR (Single)	OWAN [17] (Joint)	SPAIR (Joint)
PSNR	28.59	30.59	31.51	31.24	31.44	**32.73**	28.44	32.59
SSIM	0.6726	0.8075	0.9213	0.9259	0.9263	**0.9410**	0.841	0.935

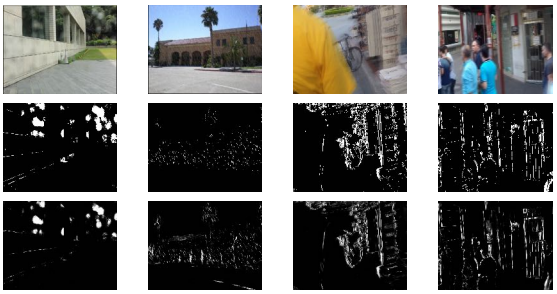


Figure S3. Visualization of degradation mask for different tasks. First, second, third row describes input image, ground-truth mask,predicted mask respectively

dicted mask do not adversely affect the final restoration output. Net_L classifies some regions as non-degraded, while slight intensity distortions may well be present in those areas. Nonetheless, restoration of such distortions is easy and this is achieved through the few non-sparse layers in our network.

We conduct the following experiment to show that at convergence, final performance of Net_R is not very sensitive to small error in the prediction of Net_L. The variation of Net_R's performance with respect to Net_L's cross-entropy loss is shown in Table S3

Table S3. Influence of the acrruacy of Net_L on performance of SPAIR for raindrop removal evaluated using AGAN Dataset [13].

Epoch	20	50	70	90
BCE Loss ($\times 10^{-2}$)	8.4	6.06	5.95	5.92
PSNR	32.08	32.67	32.73	**32.73**

S4. Additional qualitative comparisons

Rain-Streaks: Figs. S8, S9 show additional qualitative results and comparisons state-of-the-art methods on Rain100H Dataset (Table 2 of main paper). Existing methods suffer from visible rain streaks or texture-smearing along rain direction. In comparison, our results are visually more pleasing, while being faithful to the ground-truth image. Fig. S10 contains comparisons of all methods trained on the combined RainStreak Dataset (Table 1 of main paper). Further, we evaluate SPAIR trained on the combined RainStreak Dataset (Table 1 of main paper) on real-world rainy images (taken from internet) in Fig. S11. It is evident that few rain streaks remain visible and background remains unclear in the results of all existing methods while our approach generates satisfactory deraining results.

Raindrop: We show additional results on the test-set of AGAN dataset in Figs. S12,S13,S14. We also include comparisons on a real-world image in Fig. S15 . Visually, we can observe significant improvement over prior works.

Shadow Removal: Fig. S16 provides additional qualitative comparisons on shadow removal showing that most existing approaches produce shadow boundaries or color inconsistencies. In contrast, SPAIR has minimal artifacts in the shadow boundaries, outperforming the baselines both qualitatively and quantitatively.

Motion Blur: While the GoPro and HIDE datasets are generated by averaging consecutive frames of real high frame-rate videos, the blurred images in RealBlur-J dataset are captured in real-world conditions. In Fig. S17 we provide compariosns of our results with the best results from [16]. In Figs. S18-S22, we provide additional results and visual comparisons of our architectures with competing methods on the GoPro deblurring benchmark. The visual results show that our results closely mimic the ground-truth sharp images, while producing artifact-free results in regions containing challenging blur. Improvements over prior methods become more pronounced on images affected with large blur.

S5. Network Details

Network layer details are given in Fig. S23.

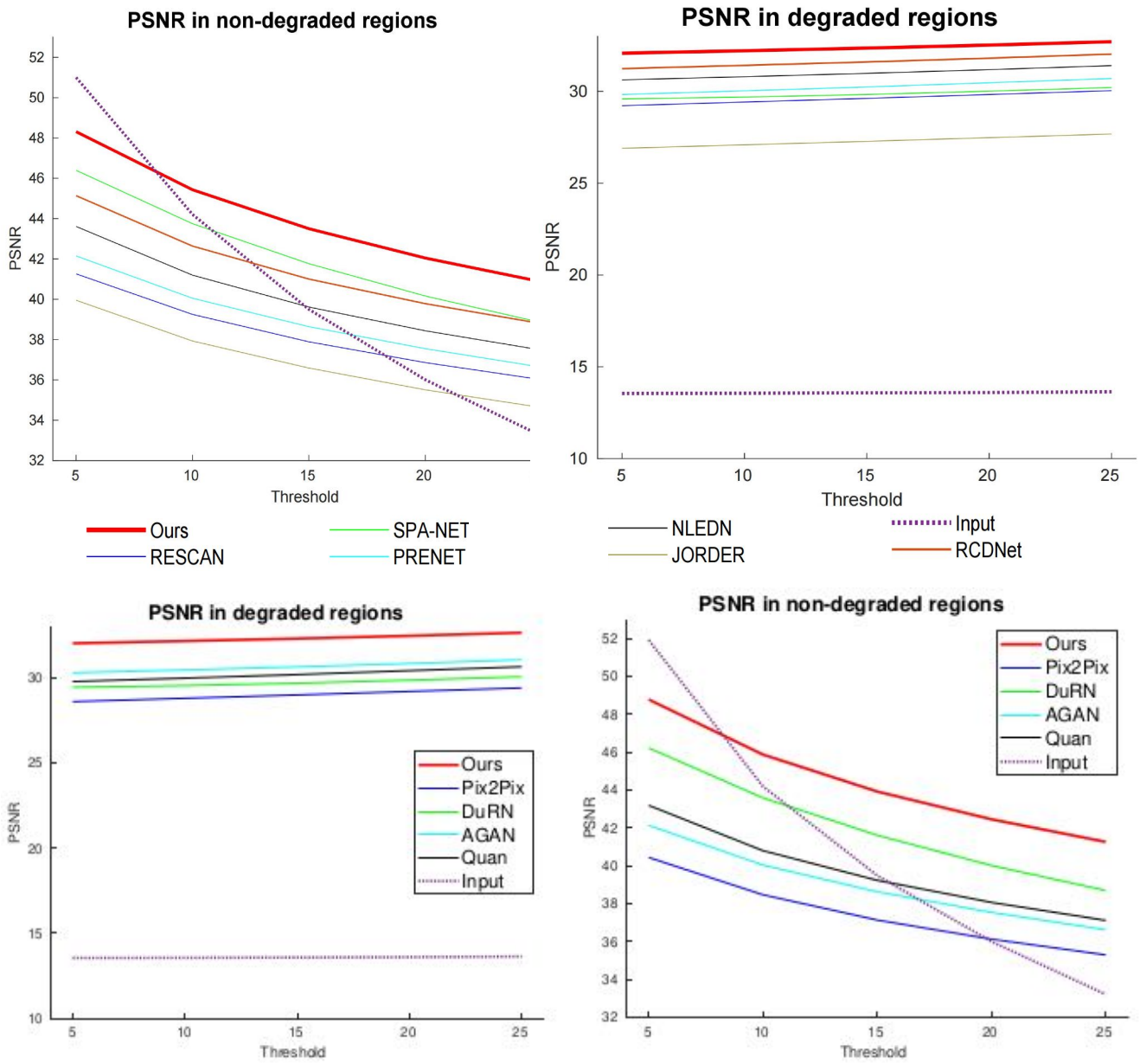


Figure S4. Comparison with baseline methods using PSNR scores in degraded and non-degraded regions for two tasks: rainstreak removal and raindrop removal. (Best viewed in color).

References

[1] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, pages 10680–10687, 2020. 13

[2] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 633–640, 2013. 3, 10, 13

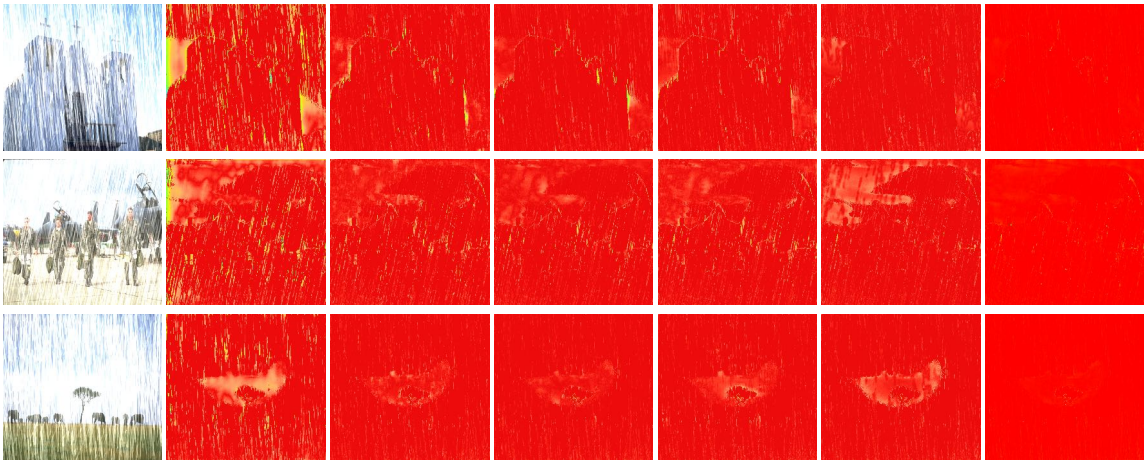
[3] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *TIP*, 2017. 9

[4] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, 2017. 2, 3, 9

[5] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, pages 1–11. Citeseer, 2014. 13

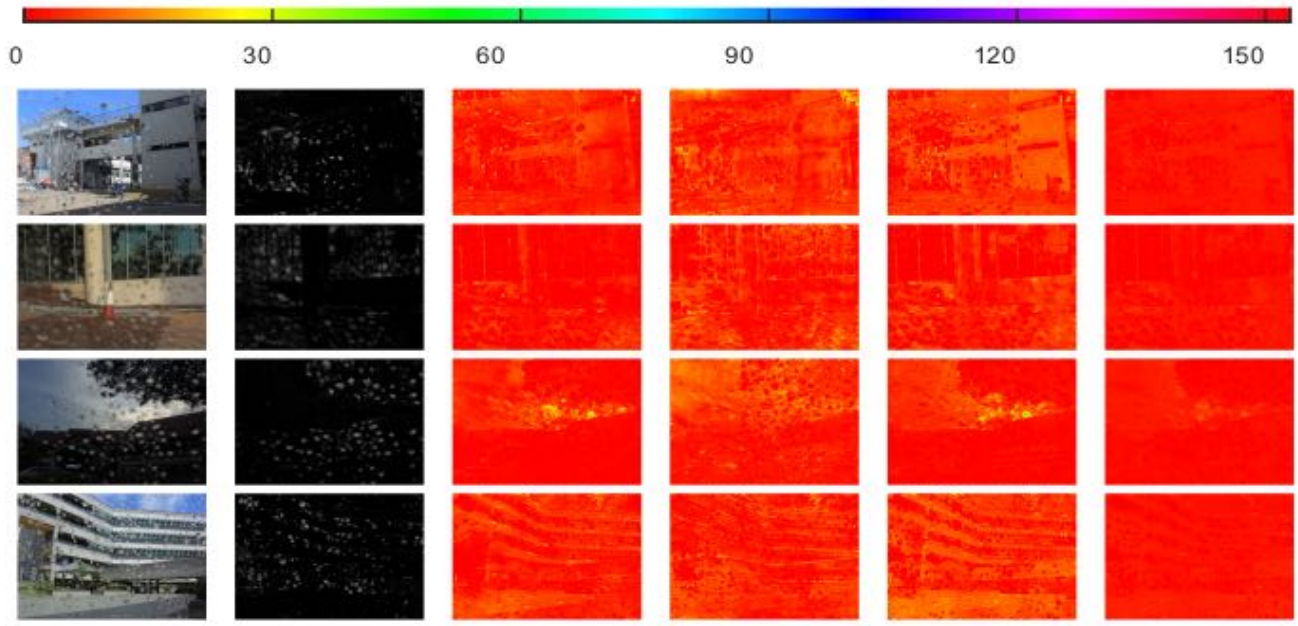
[6] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012. 13

[7] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-



Input DDN JORDER PreNet RESCAN SPANet Ours

Figure S5. Visualization of errors only on the pixels belonging to non-degraded regions (on images affected with Rain-Streaks from Rain100H dataset). From left to right: Input Image, DDN, JORDER, PRENET, RESCAN, SPANET, OURS. (Best viewed in color and zoomed-in).



Input Mask DuRN AGAN Quan Ours

Figure S6. Comparison with baseline methods using error maps calculated only in non-degraded regions within images affected with Raindrops. (Best viewed in color).

Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 13

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3, 10, 13

[9] Kui Jiang, Zhongyuan Wang, Peng Yi, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020. 2, 3, 9

[10] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, 2018. 9

[11] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *CVPR*, 2016. 2

[12] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki

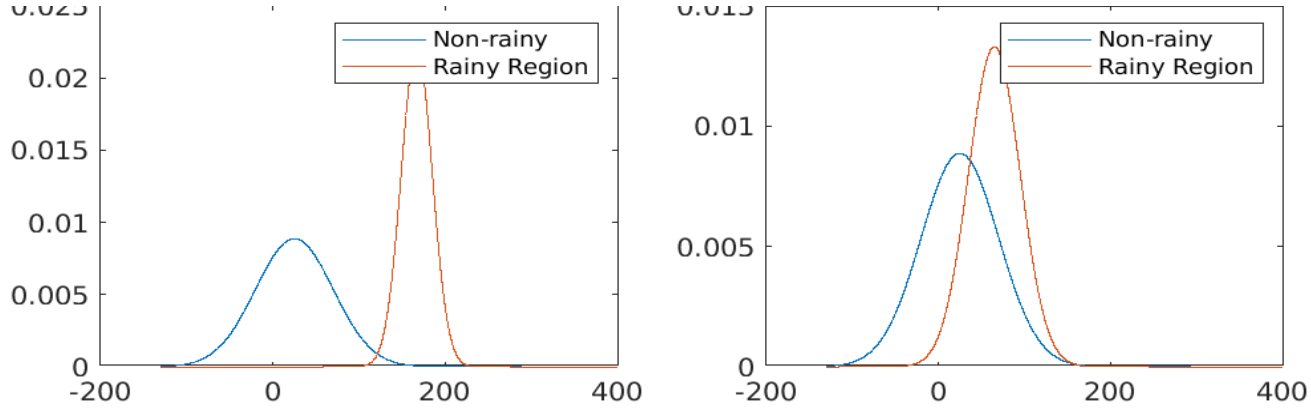


Figure S7. To demonstrate the effectiveness of Spatial Feature Modulator (SFM), we calculate feature statistics (mean and variance) for pixels belonging to rainy and non-rainy regions, separately. (a) Shows the distribution on features before SFM. The significant difference between the statistics is evident. (b) Shows the statistics obtained after SFM, wherein the rainy and non-rainy distributions have come significantly closer.

Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 3, 11, 12

[13] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018. 2, 3, 10, 11, 12, 13

[14] Yuhui Quan, Shijie Deng, Yixin Chen, and Hui Ji. Deep learning for seeing through window with raindrops. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2463–2471, 2019. 3, 10, 13

[15] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 2019. 9

[16] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 3, 14

[17] Masanori Suganuma, Xing Liu, and Takayuki Okatani. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2019. 2, 3

[18] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2020. 8

[19] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 13

[20] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12270–12279, 2019. 8, 9

[21] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012. 13

[22] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 2, 3

[23] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *CVPR*, 2019. 9

[24] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, 2018. 2, 3, 9

[25] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *TCSVT*, 2019. 2, 3

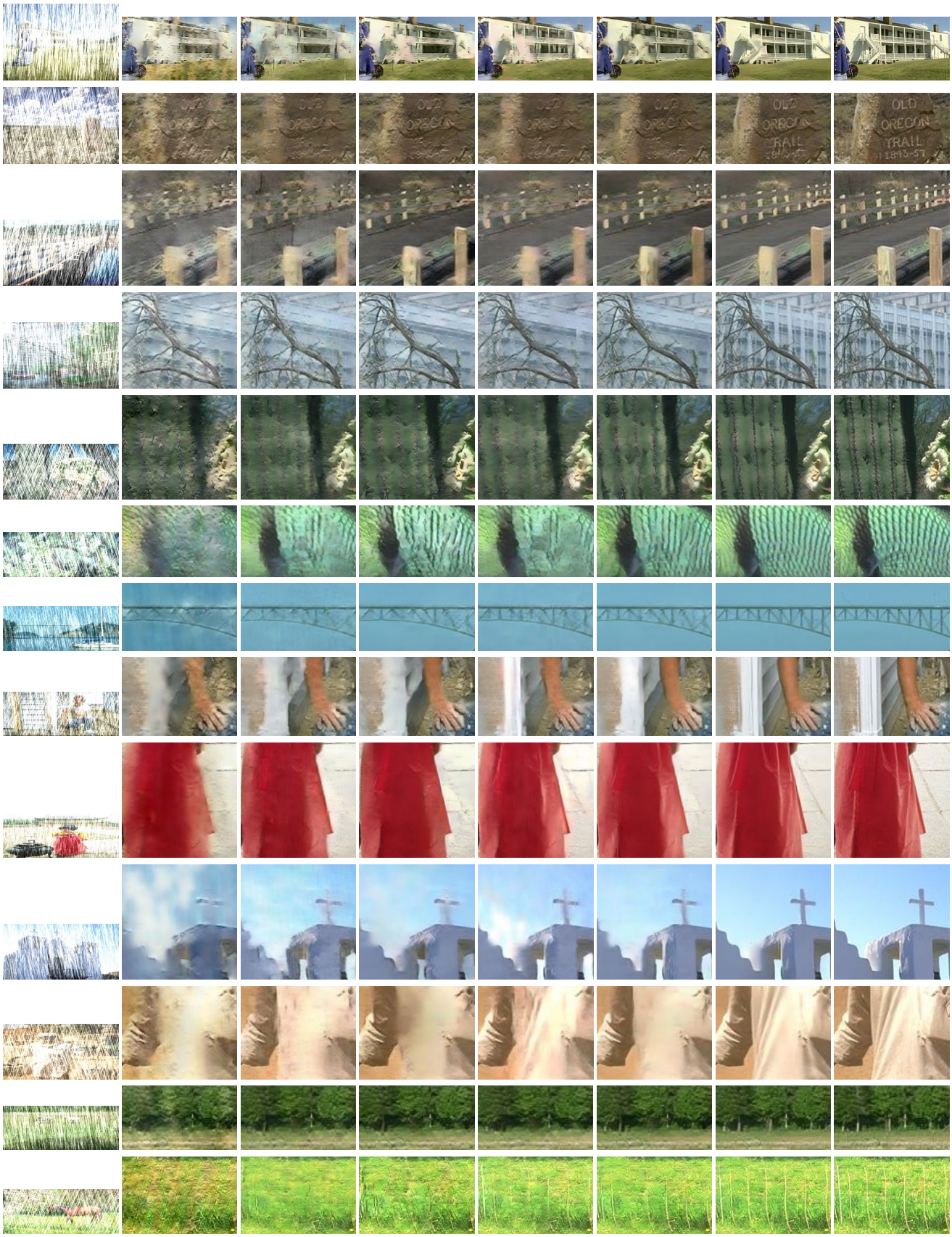


Figure S8. Qualitative comparison of results on test images from the Rain100H test-set (corresponding to Table 2 of main paper).

(a) Input (b) SPA-Net [20] (b) RCDNet [18] (c) Ours (d) GT

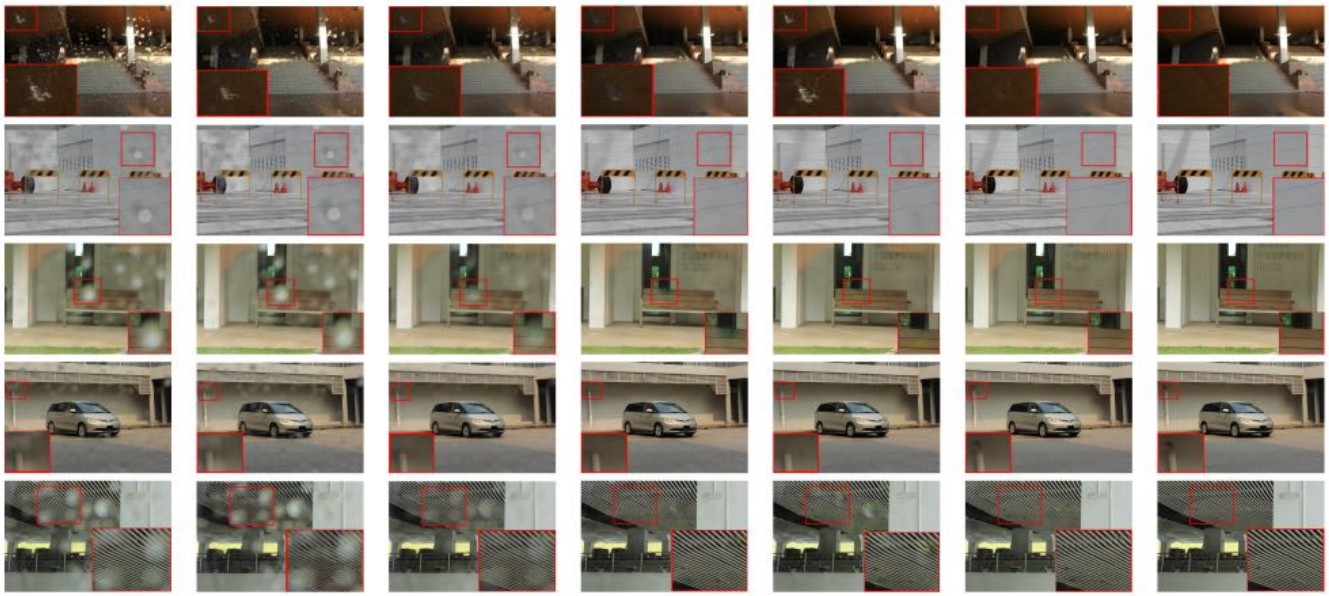
Figure S9. Visual comparisons on real rain-affected images from the SPANet dataset [20] (corresponding to Table 2 of main paper).

Input DerainNet [3] UMRL [23] RESCAN [10] PreNet [15] MSPFN [9] SPAIR GT

Figure S10. Qualitative comparisons on test images from various benchmarks considered in Table 1 of main paper.

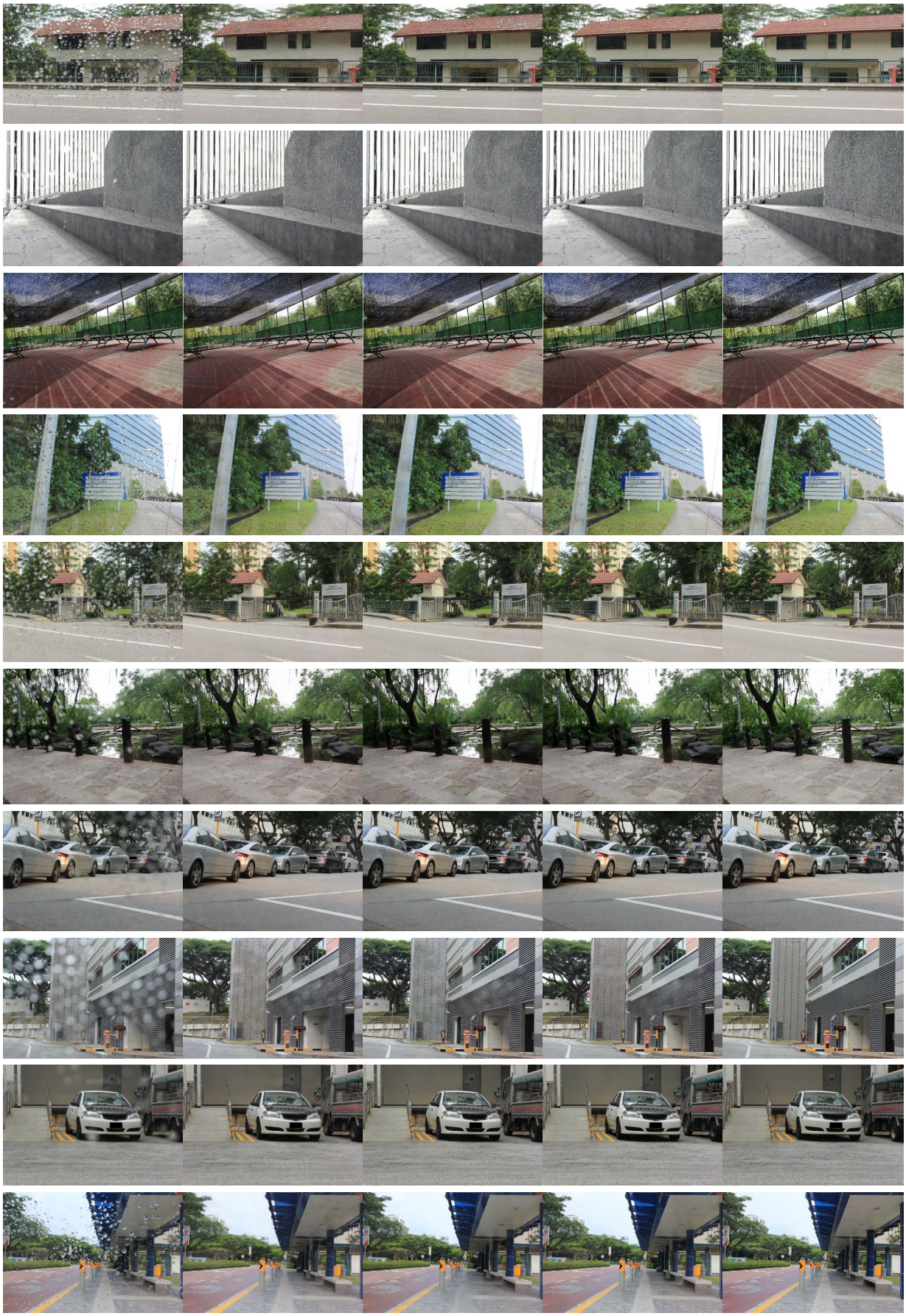
(a) Input (b) DDN [4] (c) DID [24] (d) RESCAN [10] (e) SPA-Net[20] (f) Ours

Figure S11. Qualitative comparisons on real-world rainy images from internet.



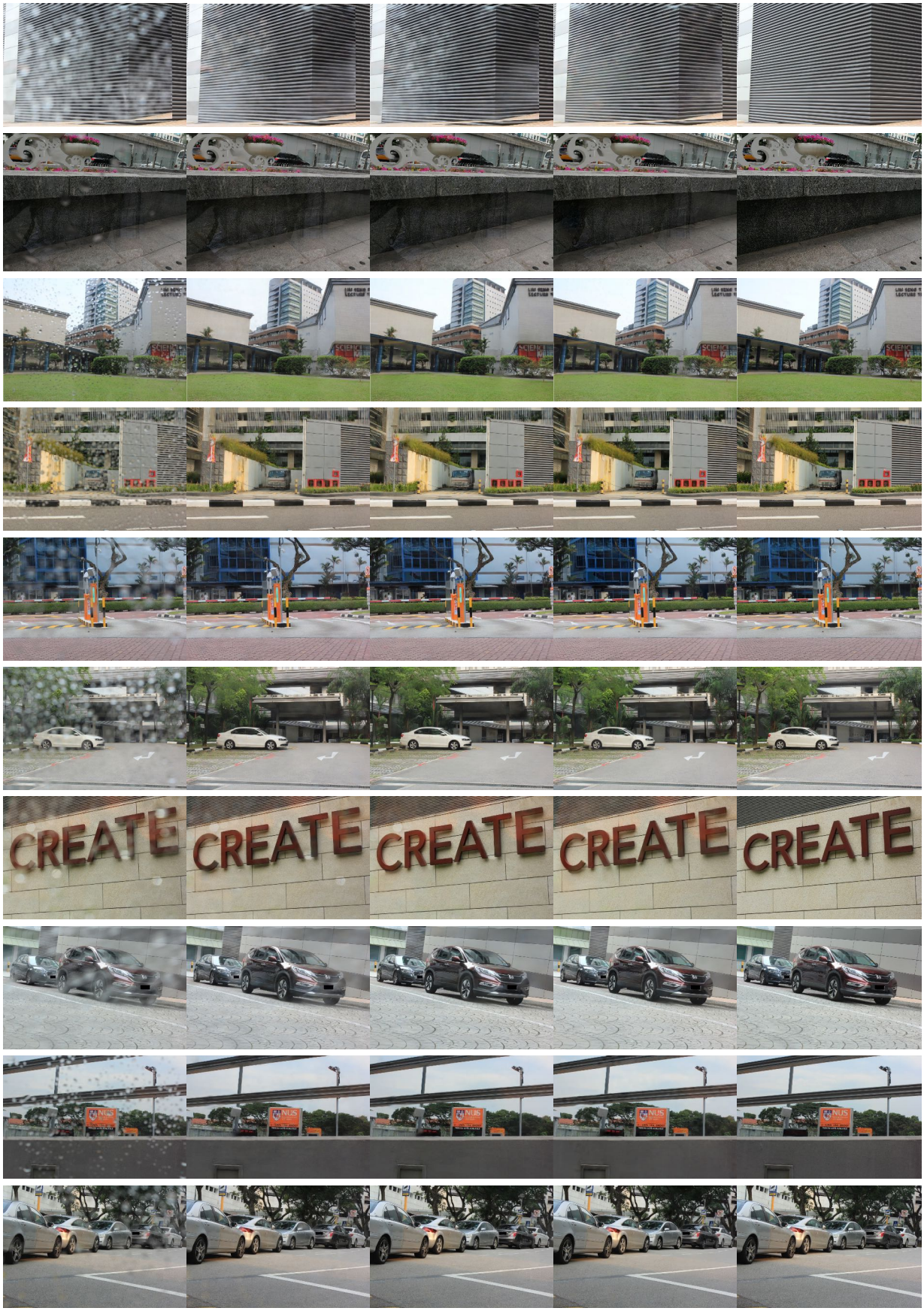
(a) Input (b) Eigen [2] (c) Pix2pix [8] (d) A-GAN [13] (e) Quan et al. [14] (f) Ours (g) GT

Figure S12. Qualitative comparisons of results on test images from the AGAN testset [13].



(a) Input (b) A-GAN [13] (c) DuRN [12] (d) Ours (e) GT

Figure S13. Qualitative comparisons of results on images from the AGAN testset [13].

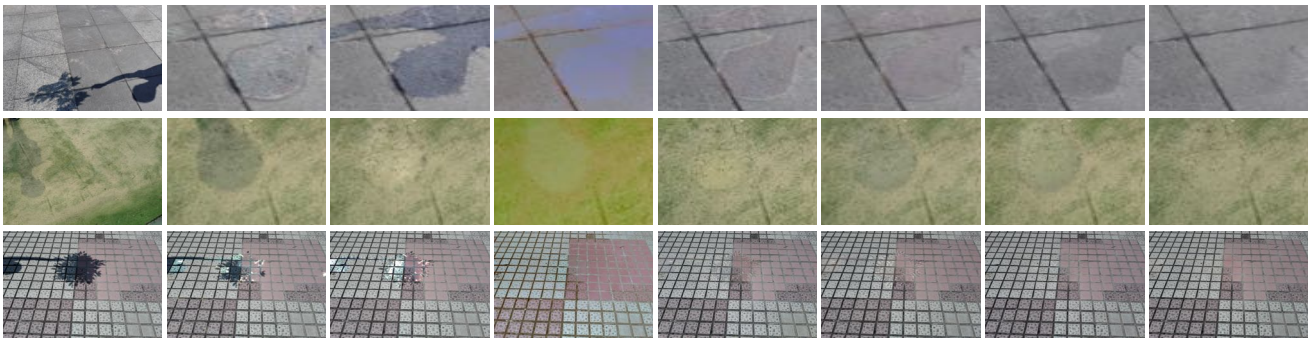


(a) Input (b) A-GAN [13] (c) DuRN [12] (d) Ours (e) GT

Figure S14. Qualitative comparisons of results on images from the AGAN testset [13].

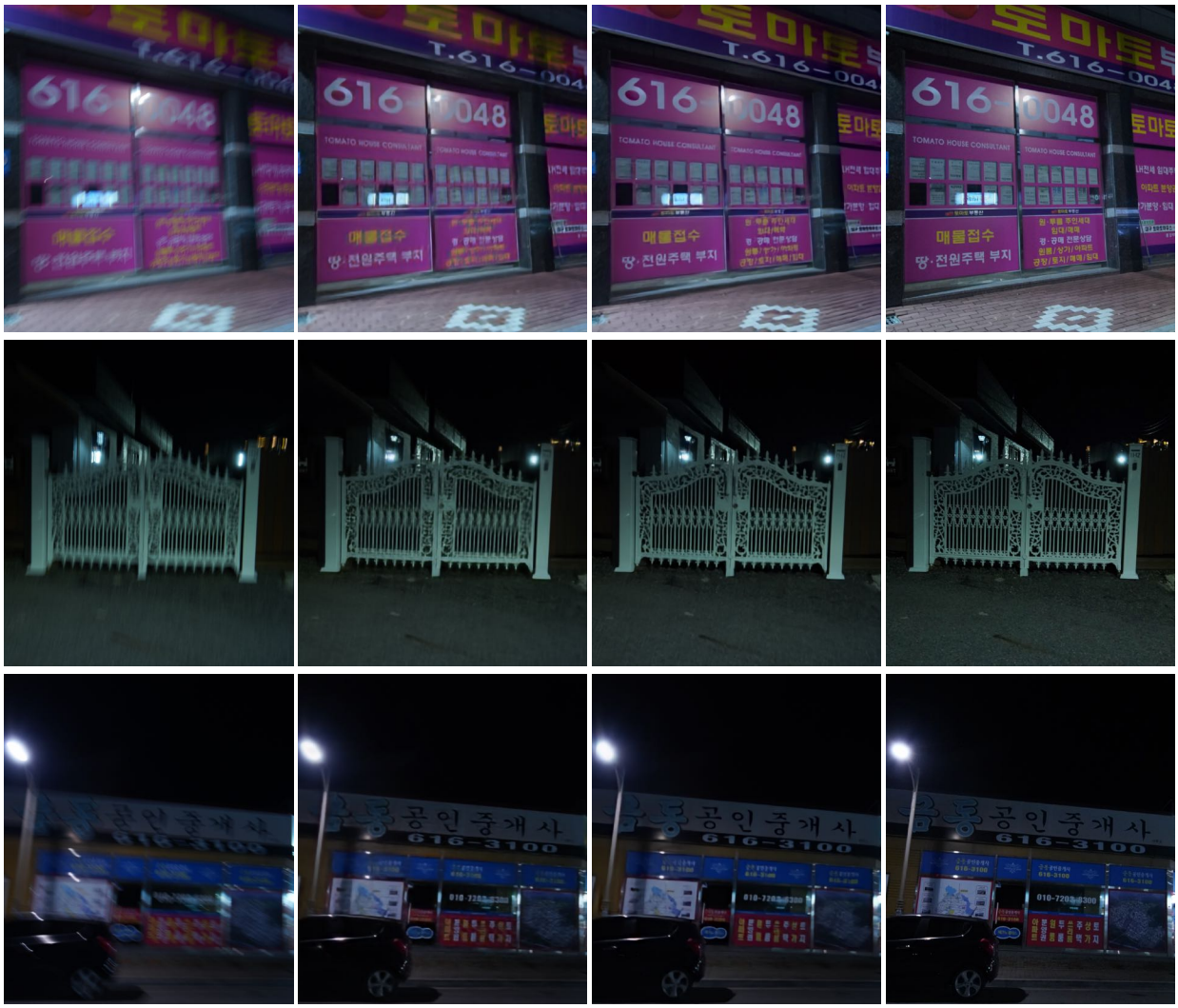
(a) Input (b) Eigen [2] (c) Pix2pix [8] (d) A-GAN [13] (e) Quan et al. [14] (f) Ours

Figure S15. Qualitative comparisons of results on a real-world image from the AGAN testset [13].



Input [6] [5] [21] [19] [7] [1] SPAIR

Figure S16. Comparison of shadow removal results on ISTD Dataset [19]. Shadow region and boundaries are visible in existing approaches.



(a) Input (b) [16] (c) SPAIR (d) GT

Figure S17. Visual comparisons on real-world blurred images from the RealBlurJ dataset [16] (corresponding to Table 2 of main paper).

Figure S18. Visual comparison for deblurring on images from GoPro test-set. The figure shows the full sized images along with zoomed-in patches corresponding to the Blurred image, results of DeblurGANv2, MS-CNN, DeblurGAN, SRN, Stack(4)-DMPHN, Our Result, and Ground-truth, respectively.

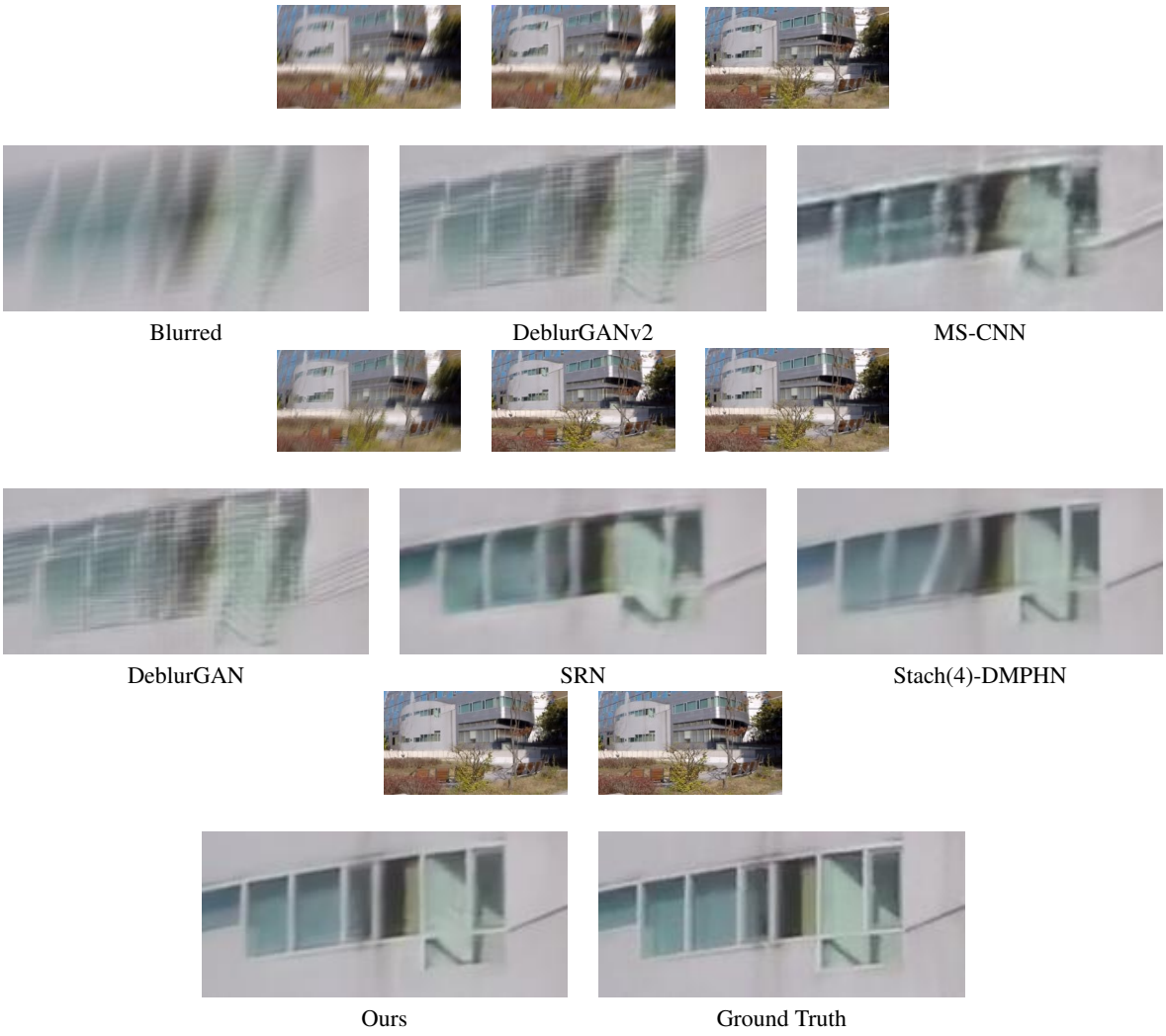


Figure S19. Visual comparison for deblurring on images from GoPro test-set. The figure shows the full sized images along with zoomed-in patches corresponding to the Blurred image, results of DeblurGANv2, MS-CNN, DeblurGAN, SRN, Stack(4)-DMPHN, Our Result, and Ground-truth, respectively.

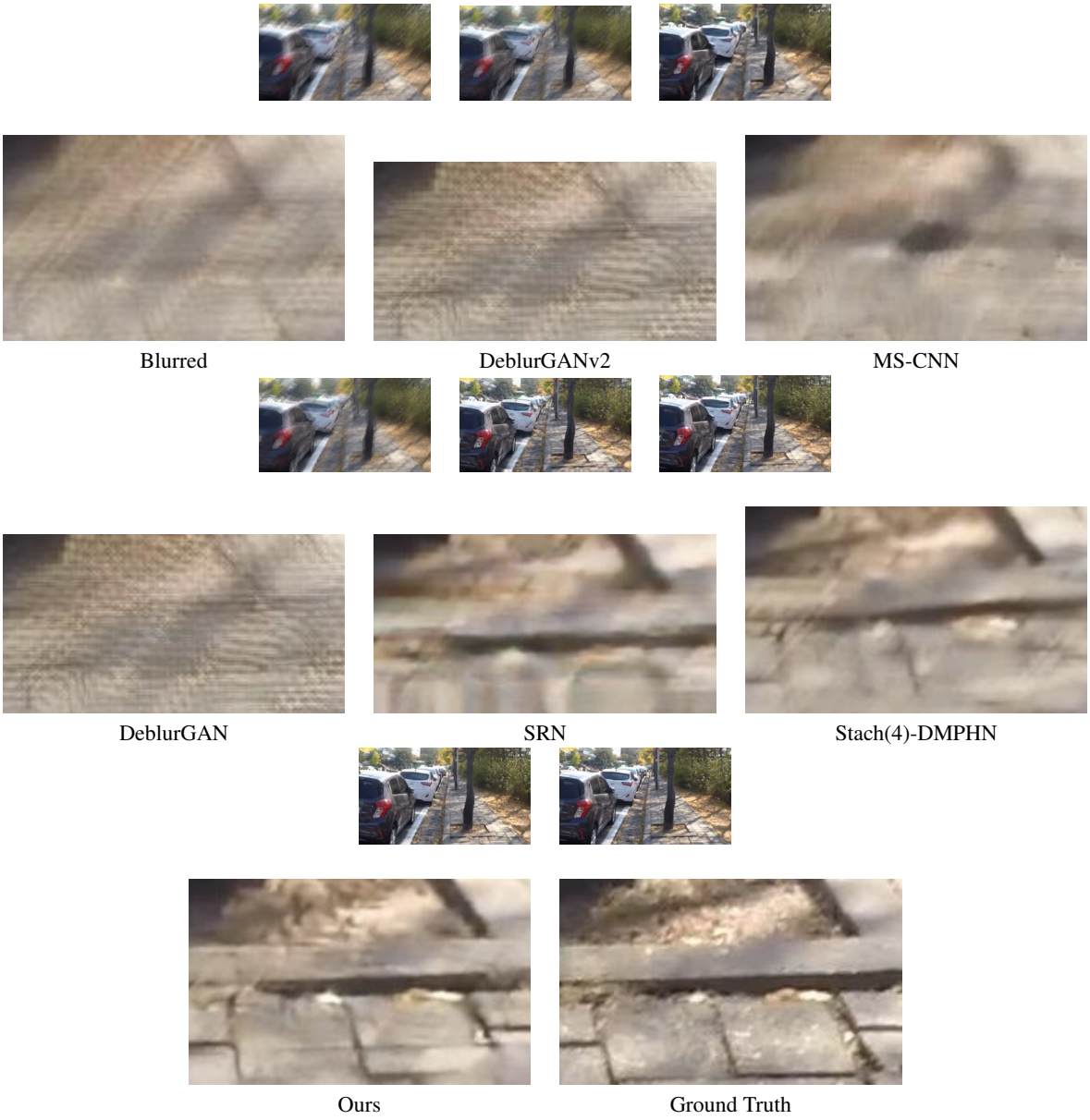


Figure S20. Visual comparison for deblurring on images from GoPro test-set. The figure shows the full sized images along with zoomed-in patches corresponding to the Blurred image, results of DeblurGANv2, MS-CNN, DeblurGAN, SRN, Stack(4)-DMPHN, Our Result, and Ground-truth, respectively.

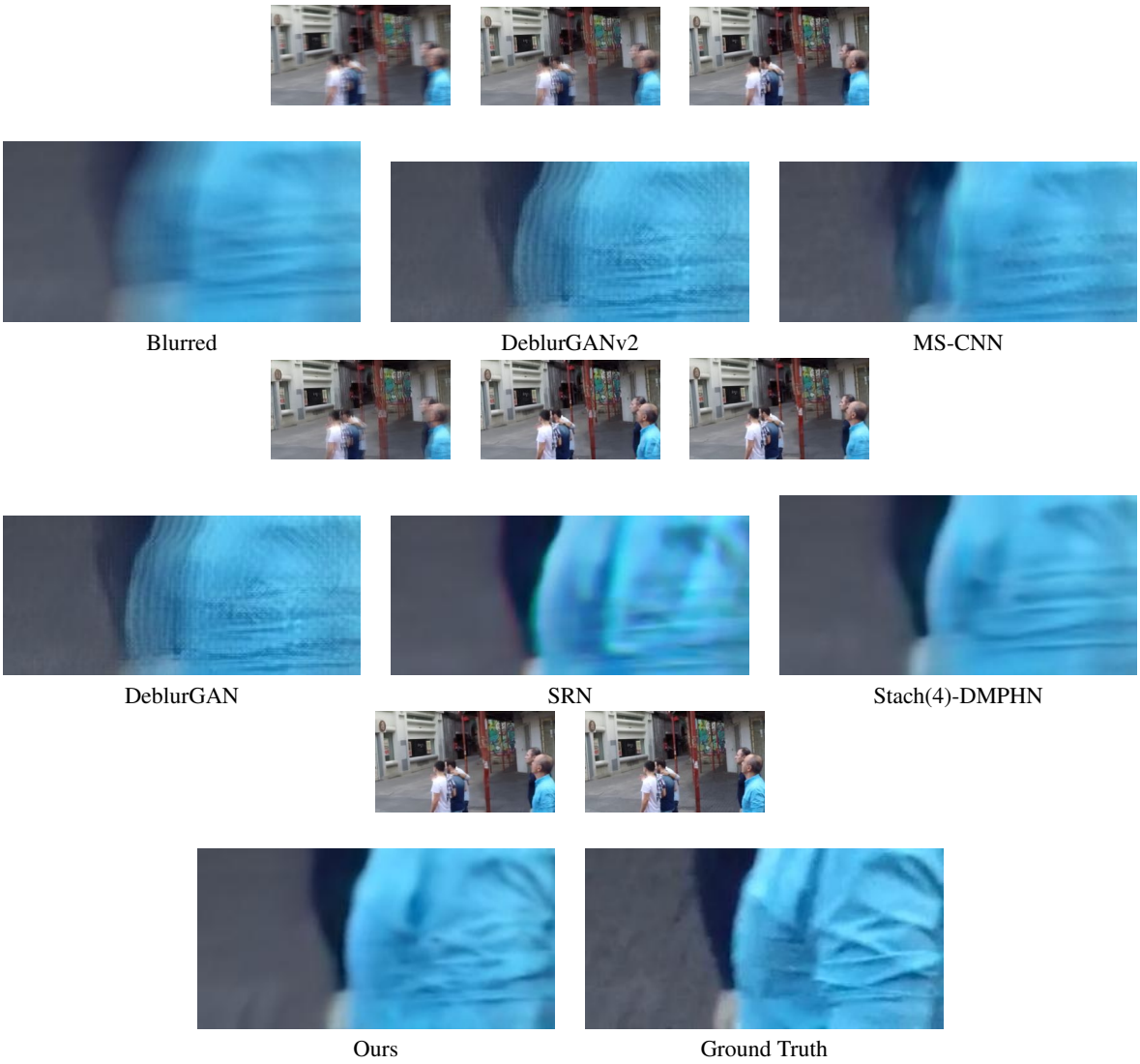


Figure S21. Visual comparison for deblurring on images from GoPro test-set. The figure shows the full sized images along with zoomed-in patches corresponding to the Blurred image, results of DeblurGANv2, MS-CNN, DeblurGAN, SRN, Stack(4)-DMPHN, Our Result, and Ground-truth, respectively.

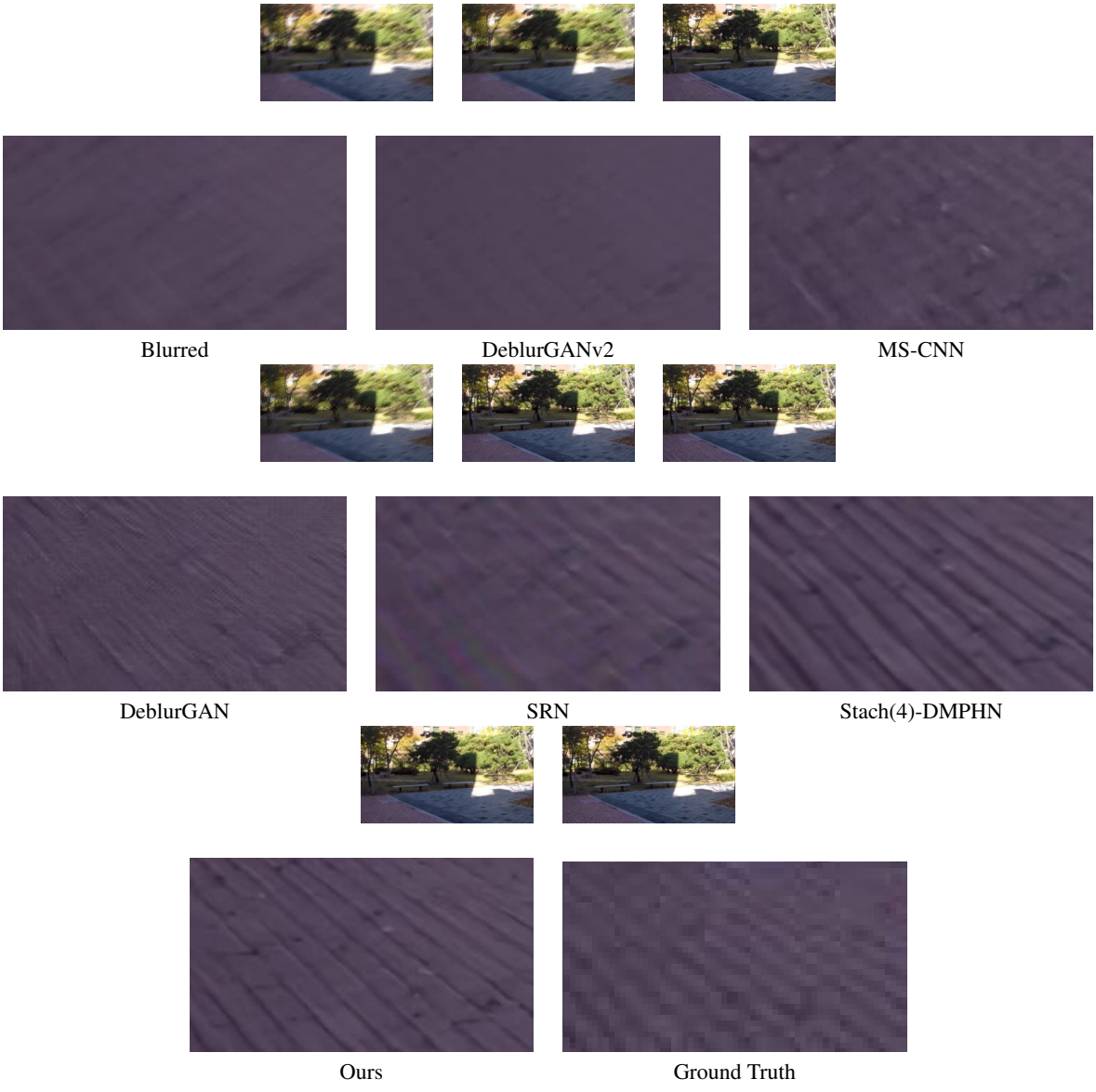


Figure S22. Visual comparison for deblurring on images from GoPro test-set. The figure shows the full sized images along with zoomed-in patches corresponding to the Blurred image, results of DeblurGANv2, MS-CNN, DeblurGAN, SRN, Stack(4)-DMPHN, Our Result, and Ground-truth, respectively.

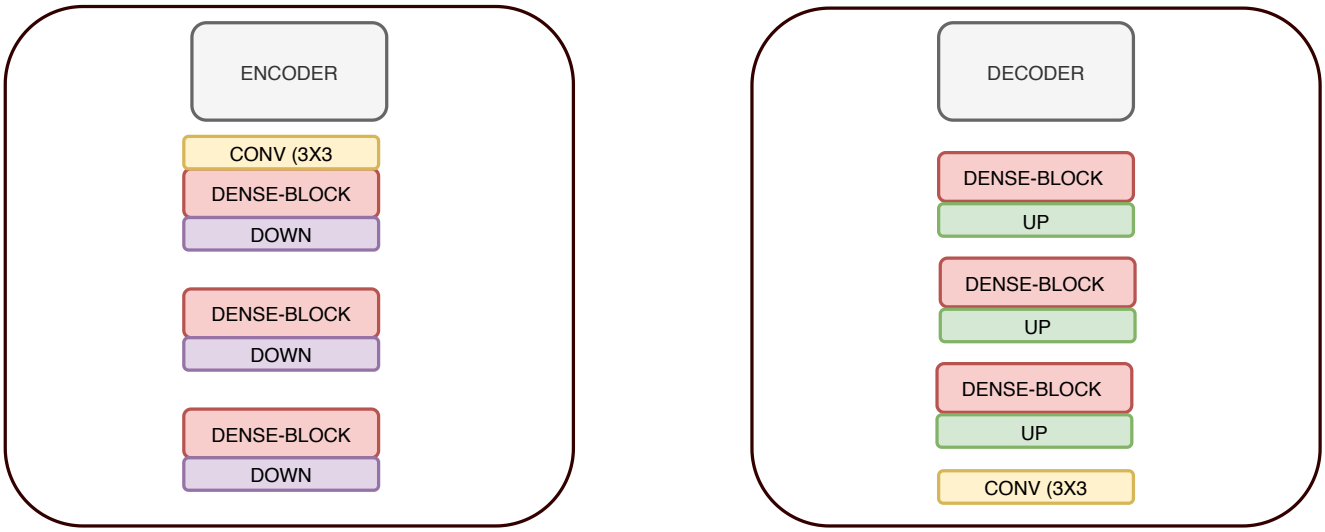


Figure S23. Layerwise details of encoder decoder within Net_L.