GIQE: Generic Image Quality Enhancement via Nth Order Iterative Degradation

Pranjay Shyam[1], Kyung-Soo Kim[1*], and Kuk-Jin Yoon[2*]
[1]Mechatronics, Systems, and Control Laboratory, [2]Visual Intelligence Laboratory
Department of Mechanical Engineering, KAIST, Republic of Korea
{pranjayshyam, kyungsookim, kjyoon}@kaist.ac.kr

Abstract

Visual degradations caused by motion blur, raindrop, rain, snow, illumination, and fog deteriorate image quality and, subsequently, the performance of perception algorithms deployed in outdoor conditions. While degradation-specific image restoration techniques have been extensively studied, such algorithms are domain sensitive and fail in real scenarios where multiple degradations exist simultaneously. This makes a case for blind image restoration and reconstruction algorithms as practically relevant. However, the absence of a dataset diverse enough to encapsulate all variations hinders development for such an algorithm. In this paper, we utilize a synthetic degradation model that recursively applies sets of random degradations to generate naturalistic degradation images of varying complexity, which are used as input. Furthermore, as the degradation intensity can vary across an image, the spatially invariant convolutional filter cannot be applied for all degradations. Hence to enable spatial variance during image restoration and reconstruction, we design a transformer-based architecture to benefit from the long-range dependencies. In addition, to reduce the computational cost of transformers, we propose a multi-branch structure coupled with modifications such as a complimentary feature selection mechanism and the replacement of a feed-forward network with lightweight multiscale convolutions. Finally, to improve restoration and reconstruction, we integrate an auxiliary decoder branch to predict the degradation mask to ensure the underlying network can localize the degradation information. From empirical analysis on 10 datasets covering rain drop removal, deraining, dehazing, image enhancement, and deblurring, we demonstrate the efficacy of the proposed approach while obtaining SoTA performance.

1. Introduction

Image quality plays an important role in the performance of vision-based algorithms designed for tasks such as object detection, semantic segmentation, depth estimation, etc. Hence, an image affected by environmental degradations such as motion blur, illumination variations, rain, fog, snow, and water droplets results in an undesirable performance drop [22, 31]. Despite the nature of degradations, they can be modeled using a common mask-based approach considering that they affect the spatial properties of an image to reduce its quality. However, since the intensity and combinations of degradations co-occurring can be non-uniform, some regions are bound to be affected more than others. Hence, a generic image restoration algorithm should be able to localize and be robust towards spatially varying degradations. While perception algorithms can be made robust to diverse weather conditions either by extending the training dataset [25, 37, 43, 46] or utilizing restoration algorithms as preprocessing step [15, 24, 40] to generate clean images. However, these approaches have their shortcomings as constructing a labeled dataset for high-level perception, diverse enough to account for all variations, is time-consuming and expensive. In contrast, image restoration algorithms are presently degradation-specific (dehazing, deraining, raindrop removal, desnowing, etc.) and do not perform well outside the distribution of training set [11, 16, 42, 44]. Furthermore, as SoTA image restoration algorithms are built upon convolutional neural networks (CNNs), utilization of the same convolution filter for the complete feature would result in weak restoration owing to the co-occurrence of multiple spatially-varying degradations across the image.

(a) (b) (c) (d) (e)

Figure 1. Images generated by proposed Nth order degradation (top) along with corresponding spatial distortion masks (bottom) for a given input having (a) natural rainy droplets and synthetic (b) motion blur, (c) snow, (d) rain, and (e) rain with snow.

*Co-corresponding authors. Listed in alphabetical order.

Yet, from a practical standpoint, having an generic image restoration algorithm would be highly desirable as it would avert extending the dataset for all perception tasks to make them robust to environment variations. Thus in this paper, we focus on blind image restoration as a preprocessing step to ensure the robust performance of perception algorithms in varying environmental conditions.

Due to the inability of CNNs to capture long-range dependencies and their fixed convolutional filters being inappropriate for varying degradations, standard convolutional filters cannot be utilized. Recently, Swin transformers [28] were proposed that can use the advantages of both CNNs and Transformers. Categorically their ability to handle large image resolutions and capture long-range dependencies via a shifted window scheme respectively presents an opportunity to utilize such a mechanism for developing generic image restoration and reconstruction algorithm. However, naively replacing convolutional blocks with transformer modules would result in a substantial increase in redundant computations. Hence in its current form, they cannot be utilized for processing a degraded image. Thus to lower computational cost without reducing performance efficacy, we propose a multiscale architecture that extracts features from different scales representing different feature granularity and subsequently uses transformer modules with various repetitions on each scale. Specifically, we use CSWin [8] wherein the computations of self-attention is decreased by using horizontal and vertical stripes. During experiments, we observe simply concatenating the features extracted by horizontal and vertical filters to be inefficient. Instead, we propose a feature selection module (FSM) that aggregates relevant features and suppresses irrelevant ones. As we deal with image restoration, we observe a self-attention mechanism to result in a high computational cost that isn't observed for high-level perception tasks due to the absence of decoder blocks or utilization of spatially large feature maps. To overcome this, we propose a spatial compression mechanism to replace the multi-head attention.

While image restoration is an extensively studied topic, prior works under-utilize the paired samples by proposing an end-to-end architecture. Concretely, a spatial distortion mask that represents the location of affected pixels isn't utilized. We highlight that designing the restoration algorithm that also predicts the spatial distortion mask as auxiliary output during training would aid the network in identifying locations that are affected. One of the challenges faced in training and evaluating practical image restoration algorithms is the absence of paired datasets having multiple co-occurring degradations involving motion blur, illumination variations, fog, rain, and water droplets. Thus, we utilize the Cityscapes [7] and its synthetic variants containing fog [39], and rain [17] degradation for training and evaluating restoration quality as well as its effect on downstream tasks such as semantic segmentation. We utilize Pix2PixHD [48] to consider distortions caused by water droplets. We summarize our contributions as follows,

- We propose an image restoration and reconstruction architecture that is able to recover images affected by blind distortion combination.
- To ensure degradation is accurately localized, we integrate an auxiliary degradation prediction branch to enhance the restoration performance.
- To enable realistic distortions we propose a N^{th} order degradation model that recursively applies a set of degradations.
- We propose a Feature Selection Module and Spatial Compression Mechanism to reduce the computations of the CSWin Transformer module.
- We examine the effect of image restoration vis-a-vis extended training on downstream tasks towards achieving robust performance.

2. Related Works

2.1. Image Restoration and Enhancement

Image restoration and enhancement are extremely researched areas with different methods being developed to independently recover images affected by degradations such as varying illumination conditions [19, 47], motion blur [23, 26, 54, 55], rain [14, 36], fog [9, 20, 35, 42, 51] and raindrop [34]. Current SoTA assumes prior information about the degradation and thus restores a degraded image by following either a model-based multi-stage approach [2, 57] or an end-to-end approach [1, 4, 13] to directly generate a restored image. However, in real scenarios, an unknown combination of such degradations might affect an image, adversely affecting the performance of the restoration algorithm. Furthermore, as vision-based algorithms are being increasingly used for high-level perception tasks, a degraded image results in a significant performance drop of such algorithms and is highly undesirable.

The current approach of designing a unique algorithm for each algorithm is inefficient and thus encourages researchers to examine the possibility of having a common restoration architecture. Towards this goal, [53] proposed multiscale cascaded CNNs and demonstrated the performance of the same architecture for multiple restoration and reconstruction tasks such as image enhancement, super-resolution, and image denoising. However, it is still trained for limited scenarios and thus does not perform well when the distribution of the test set is different from the training set or when test degradation is not covered in the training set. Following similar motivation [11] proposed an unsupervised approach for restoring images by disentangling the image into clean and noisy latent spaces to generate noisy images and subsequently restore them.

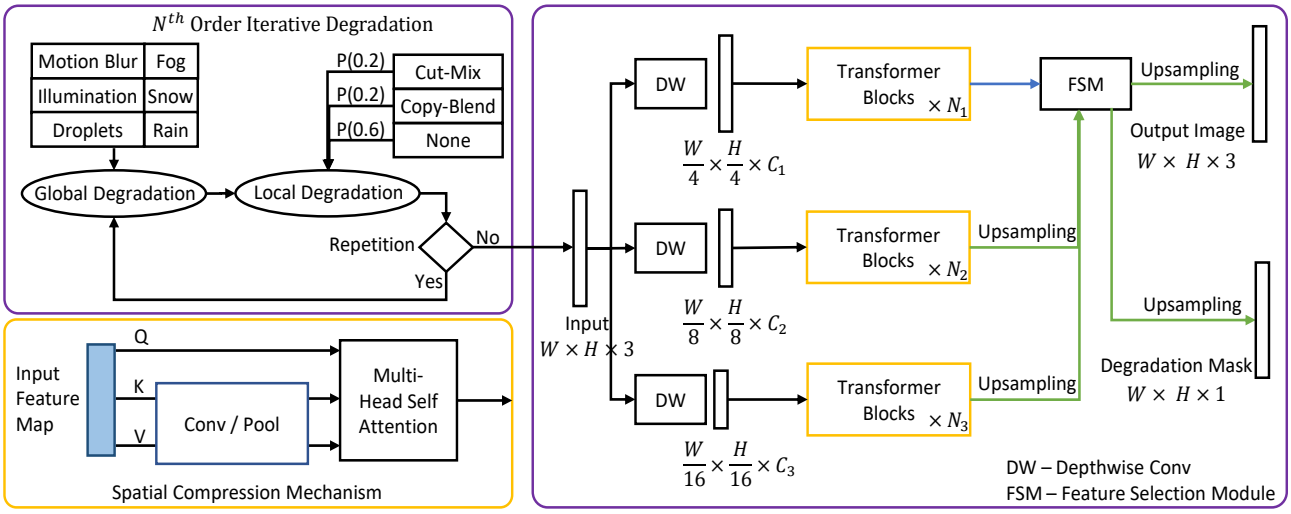


Figure 2. Overview of the proposed restoration and reconstruction architecture with Nth order iterative degradation mechanism and Spatial Compression Mechanism for Multi-Head Self-Attention within the Transformer.

Recently different transformer-based algorithms were proposed for performing high-level vision tasks such as object detection [3], image classification [10, 28] and semantic segmentation [50]. These works achieved SoTA performance, benefiting from the ability of the transformer mechanism to model long-range dependencies. Aiming to leverage these characteristics, different works [5, 49] were proposed. IPT [5] proposed a multi-head and multi-tail approach wherein each head and tail performs a specific task, however, such an assumption doesn't hold for real images. Uformer [49] proposed replacing the convolutional blocks within the UNet [38] architecture with transformer modules and achieved SoTA performance in rain removal and image denoising tasks.

2.2. Reducing Computational Complexity of a Transformer

While transformers have demonstrated superior results due to their capability to model long-range dependencies, their high computational requirement due to the self-attention mechanism is undesired. As computational complexity of the self-attention mechanism is quadratic to the size of the input feature map. Hence for images, naively applying self-attention would result in excessive redundant computations. To overcome such computational bottlenecks, Swin Transformer [28] proposed a shifted window approach to compute self-attention in a localized region and increase the receptive field via shifted windows. CSWin [8] devised a more efficient approach by using cross-shaped windows that divides the feature map into horizontal and vertical stripes and subsequently performed self-attention in parallel. This approach outperformed prior transformer architectures in terms of both computations and performance.

2.3. Generic Image Restoration and Reconstruction

Owing to its desired characteristics, different works have proposed methodologies that could be used for generic image restoration and reconstruction. Notably, Dual Residual Networks [27] proposed different residual architectures that could be tweaked to perform different restoration tasks. Building upon it [45] examined a strategy to perform multiple operations in parallel to restore various degradations. However, since these approaches assume a known degradation model, they cannot be directly applied to images with multiple degradations with unknown mix ratios found in natural conditions.

3. Methodology

3.1. Architecture Overview

We summarize the proposed framework in Fig. 2 and refer to it as GIQE. Importantly we emphasize two mechanisms to achieve a generic image restoration and reconstruction network *i.e.* (1) Nth Order Degradation to generate synthetic training samples mimicking natural conditions and (2) multi-scale transformer-based image restoration and reconstruction pipeline.

3.2. Optimizing Transformer Mechanism

As transformer architectures involve huge computational complexity on account of self-attention mechanisms, we first propose two *tricks* to improve the performance, while reducing the computational complexity. First, we reduce the spatial resolution of features used for the self-attention mechanism. Second, replacing the MLP layers after the self-attention mechanism with a multi-scale feature extraction module to improve local information content.

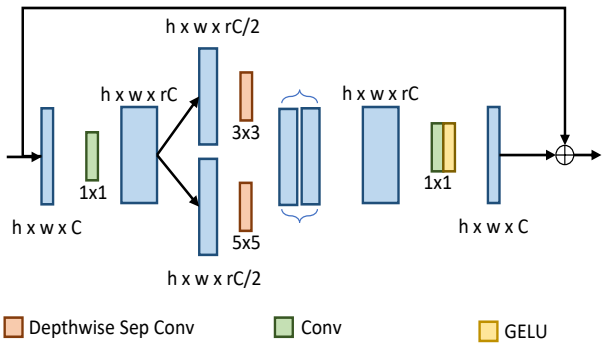


Figure 3. Schematic of Multi-Scale Feature Extraction Mechanism

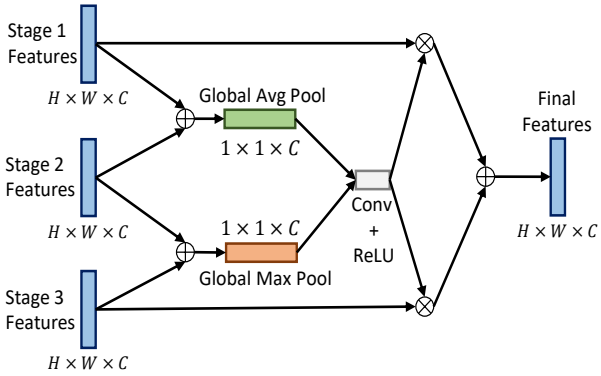


Figure 4. Schematic of Feature Selection Module

Spatial Compression Mechanism : Multi-Head Self-Attention within the transformer encoder works on an input feature $(X \in R^{C\times h\times w})$ with height (h), width (w) and channels (C) respectively. Hence we first summarize the computational complexity for global multi-head self-attention (GMSA) along with recent window-based self-attention mechanisms that ensure reduced computations *i.e.* Swin Transformer (wherein each window has $M \times M$ patches) and CSWin Transformer (where the local self-attention is based on horizontal and vertical stripes with a width of sw),

$$\Omega(GMSA) = whC(4C + 2wh) \quad (1)$$

$$\Omega(Swin) = whC(4C + 2M^2) \quad (2)$$

$$\Omega(CSWin) = whC(4C + sw*h + sw*w) \quad (3)$$

As the spatial resolution of input feature space directly affects the computational complexity, reducing it before the projection layer to obtain Key (K), Query (Q) and Value (V) vectors would result in reduced computations. Hence we propose two approaches *i.e.* either using average pooling (with size s) or simply reshaping the feature map (by a factor s) to increase the number of channels that are subsequently reduced using the projection layer in MSA. Hence the complexity reduction achieved after downsampling the feature map (R_s) can be calculated as,

$$\Omega(\beta GMSA) = whC\left(4CR_s^2 + \frac{2wh}{R_s^2}\right) \quad (4)$$

$$\Omega(\beta Swin) = whC(4CR_s^2 + 2M^2) \quad (5)$$

$$\Omega(\beta CSWin) = whC\left(4CR_s^2 + \frac{sw*h + sw*w}{R_s}\right) \quad (6)$$

It should be highlighted that using pooling operation to reduce the feature maps results in computations becoming linear, instead of quadratic. Furthermore, since we downsample the input feature map only for keys and values, both Swin and CSWin transformer architectures can be subsequently used to ensure greater computational efficacy.

Multi-Scale Feature Extraction (MSFE) : Standard transformer architecture contains a feed-forward network (FFN) that is not able to fully extract the local information content which is necessary for image restoration tasks. This affects the performance since information of neighboring pixels can guide the restoration and reconstruction of affected pixels. Hence we replace the FFN block with a Multi-Scale Feature Extraction mechanism that comprises Depth-wise Separable Convolutions [6] with filter size of 3 and 5, following [49]. When replacing the FFN with the proposed MSFE, we first increase the feature dimension of the input, by a factor r, using a projection layer which is subsequently split into 2 parts, corresponding to filters of sizes 3 and 5. GELU activation function is subsequently used after each convolutional layer. The complete mechanism is summarized in Fig. 3.

3.3. Multi-Scale Image Restoration

The proposed image reconstruction and restoration network is designed following a multi-scale approach to enable high-quality reconstructed images. Unlike UFormer that down-samples the encoded transformer features at stage l to be used as inputs for transformer at stage $l-1$, we design the same common backbone and modify the hyperparameters of the architecture such as the number of transformer blocks, channels, heads, and window size. Features from high-scaled images would be rich in global image semantics, wherein features from low-scaled images would have rich local image semantics. Fusion of these features would ensure restored images are both globally and locally coherent. To merge these features effectively, we propose the feature selection module.

Feature Selection Module : Only the pixels requiring restoration should be considered for joint restoration and reconstruction, while irrelevant ones should be reconstructed. Thus, the multi-scale architecture should be able to aggregate features following this notion. However, element-wise summation or concatenation doesn't ensure such efficient feature merging. Hence an adaptive mechanism is required that can perform these tasks without the significant computational overhead. We propose a feature selection module that first matches the feature dimensions using a 1x1 con-

volution and transposed convolution of size 2 with stride 2. Subsequently, features are aggregated via element-wise summation, followed by global max and average pooling operations, enhanced using 1x1 convolutions. A sigmoid operation is subsequently performed to generate a channel-wise attention map to identify channels containing important features. The overall architecture of the feature selection module is summarized in Fig. 4. With this mechanism, we focus on capturing the relevant features for image restoration, whereas regions that don't require any restoration can be passed using skip connections.

Auxiliary Decoder Branch : As SOTA restoration algorithms aim to generate a restored and reconstructed image directly, they have to localize, identify the quantum of degradation jointly, and predict the approximation of restored pixels. Hence during optimization, the network is tasked to perform these tasks simultaneously, resulting in sub-optimal optimization. As we use a paired dataset to restore and reconstruct a degraded image, we can integrate a secondary decoder that estimates the degraded regions in binary classification to aid the training process. Since the spatial location of the degraded pixels would be the same for both decoders, the auxiliary decoder branch can follow the same architecture as the feature selection module and would thus complement the main restoration and reconstruction branch. We obtain such a mask (I_{Mask}) by simply subtracting the input degraded image ($I_{Noisy} \in R^{W \times H \times 3}$) with clean output image ($I_{GT} \in R^{W \times H \times 3}$) having a width ($W$) and ($H$). We then perform max pooling operation along the channel dimension to result in a map of spatial resolution $W \times H$. The resultant difference map is subsequently thresholded based on pixel intensity at location (x, y) following,

$$I_{mask} = I_{GT} - I_{Noisy} \quad where \begin{cases} 0 & if\ I(x,y) = 0 \\ 1 & if\ otherwise \end{cases} \tag{7}$$

3.4. Nth Order Degradation

Since a natural image can contain a variety of degradations with varying intensity, *e.g.*, a rainy driving scene contains both dynamic rain and motion blur. To ensure the consistent performance of a generic restoration and reconstruction algorithm, we require a paired training dataset that contains a large degradation space covering a range of degradation combinations. However, capturing such a dataset can be excruciating and impossible, as commonly used paired datasets focus on a specific degradation. Thus we present a synthetic degradation model that could be coupled with a real degraded image to generate a variety of non-linear combinations of degradations. Hence, we propose an iterative degradation mechanism that generates synthetic non-linear degraded image (I_{Noisy}) by introducing deformations ($D(x)$; $x \in$ *motion blur, noise, fog, snow, rain, illumination variations or None*) recursively (r *times*) onto a clean image (I_{GT}). While these degradations deform the complete training sample, we additionally introduce localized degradation ($LD(.)$) using randomly selected Cut-Mix [52] and Copy-Blend [41] or *None* operations with probability 0.2, 0.2, and 0.6. The pipeline can be mathematically represented as,

$$I_{Noisy} = [\ LD(\ D(x,\ I_{GT}))]^r \tag{8}$$

We summarize the pipeline in Fig. 2 with qualitative samples included in Fig. 1 with additional details included in supplementary. As an additional enhancement mechanism, we observe that sharpening the ground truth images following a random gaussian blur filter to improve edge information. Subsequently, when this information is used as ground truth, restored images have higher edge information. The mechanism to generate sharper ground truth images can be summarized as,

$$I_{Sharp-GT} = I_{GT} + \alpha * (I_{GT} - I_{Blurred}) \tag{9}$$

where α represents the weighted addition and blurred images are generated using gaussian blur with filter size randomly chosen between $\in [3, 13]$.

3.5. Loss Formulation

Following prior works [27, 49], we use a combination of pixel-wise (L1) and structural similarity loss (SSIM) for the image reconstruction and restoration branch, whereas, for the auxiliary decoder branch, we use binary cross-entropy loss. To ensure stability during training we use an additive term ε within L1 loss and set it as 10^{-6}.

$$\begin{aligned} L = \lambda_1 * \sqrt{||I_{GT} - I_{Restored}||^2 + \varepsilon} + SSIM(I_{GT} - I_{Restored}) \\ + \lambda_2 * BCE(I_{Mask-GT}, I_{Mask-Restored}) \end{aligned} \tag{10}$$

During our experiments, we fix λ_1 and λ_2 to 1.

3.6. Training Methodology

For training the proposed framework on different datasets having either single or multiple degradations, we follow a common training pipeline. Specifically, we use AdamW [29] with momentum coefficients as 0.9 and 0.999 and weight decay as 0.02. The training image resolution is fixed to 128 $\times$ 128 with a batch size of 4 using 2 Nvidia 3090 GPUs with an initial learning rate of 2e-4. The learning rate is adjusted following the cosine annealing with a minimum learning rate set to 2e-6. The complete network is trained for 400 epochs. Apart from the proposed Nth order degradation, we randomly rotate the image by 90°, 180° or 270°. Furthermore, to ensure generating degradations does not lead to computational bottlenecks, based on empirical evaluation, we limit the value of N to 5.

Table 1. Ablation studies of different mechanisms on the GOPRO-I, GOPRO-II (r), and GOPRO-III datasets. Higher values of PSNR and SSIM denote better performance

Algorithm	GOPRO-I	GOPRO-II (r)	GOPRO-III	# Params
	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	(x10^6)
Input	25.64 / 0.79	21.24 / 0.61	12.68 / 0.42	-
Baseline	30.78 / 0.89	22.19 / 0.64	13.49 / 0.40	14.6
Varying Order Degradation				
+ Aug (N=1)	30.77 / 0.89	23.66 / 0.65	15.19 / 0.51	14.6
+ Aug (N=3)	30.71 / 0.89	25.37 / 0.69	19.23 / 0.55	14.6
+ Aug (N=5)	30.69 / 0.88	26.49 / 0.71	24.14 / 0.57	14.6
Spatial Compression				
+ SC (Pool)	30.87 / 0.88	26.43 / 0.71	24.09 / 0.57	12.4
+ SC (Conv)	30.91 / 0.89	26.71 / 0.72	24.23 / 0.58	13.6
Feature Enhancement				
+ MSFE (3)	31.16 / 0.89	26.05 / 0.73	24.59 / 0.61	12.7
+ MSFE (5)	31.24 / 0.89	26.08 / 0.73	24.62 / 0.62	12.8
+ MSFE (3, 5)	31.46 / 0.89	26.24 / 0.75	24.69 / 0.63	13.1
+ SDB	32.03 / 0.90	26.39 / 0.76	24.76 / 0.65	13.1
+ Scale (=2)	32.37 / 0.92	26.77 / 0.79	24.93 / 0.66	19.8
+ Scale (=3)	32.79 / 0.93	26.91 / 0.80	24.99 / 0.67	25.1
+ GT Sharpen	33.05 / 0.93	27.05 / 0.80	25.07 / 0.68	25.1
CSWin based	33.21 / 0.94	26.42 / 0.81	25.19 / 0.68	24.6
Longer Training x100 Epochs				
Swin based	33.37 / 0.94	27.11 / 0.82	25.42 / 0.70	25.1
CSWin based	33.48 / 0.94	27.64 / 0.82	25.98 / 0.72	24.6

4. Experimental Analysis

4.1. Datasets and Evaluation Metrics

We choose both natural and synthetic single degradation datasets for Motion Blur, Rain, Snow, Fog, Illumination Variations, water droplets for our experiments. As there lacks a dataset that captures multiple degradations and its corresponding paired clean image, we extend the cityscapes dataset to have varying combinations and intensities of the degradations mentioned above. To generate fog we use the framework proposed in [58] while for illumination variation we use model proposed in [30], whereas for water droplets we a pix2pixHD model [48] trained using [33] dataset. For motion blur, rain, and snow augmentations, we use imgaug library [21]. We elaborate on the data generation process in the supplementary. As it is difficult to ascertain the image quality of restored images, we use high-level perception tasks such as semantic segmentation to determine the impact of restoration and reconstruction. To evaluate the performance of restored images, we use PSNR and SSIM as evaluation metrics where the PSNR is calculated on Y channel of the YCbCr image.

4.2. Ablation Studies

Training transformer models is a time-consuming process, thus, we first evaluate the contribution of different mechanisms proposed to determine an effective and efficient baseline algorithm. For this task, we utilize the GOPRO [32] dataset to represent natural motion blur and add synthetic augmentations such as rain, snow, illumination change, and None with a probability of 0.25. Furthermore, we randomly switch the input degraded image with a clean ground truth image to ensure the network can perform image reconstruction. We chose GOPRO dataset due to a large amount of natural paired training and test images, in which synthetic degradations can be included to generate image pairs having multiple degradations. To evaluate the performance in single and multiple degradations, we use 3 versions of the GOPRO test set. This allows us to determine the peak performance in single vs. multiple degradation conditions. GOPRO-I contains the standard motion blur images as input, whereas GOPRO-II(r/s/i/f/n) contains synthetic rain or snow or illumination variation or fog or noise, alongside motion blur. Finally, GOPRO-III contains all the variations (rain, snow, noise, illumination change, fog) with varying intensities. We summarize the quantitative results in Tab. 1 and qualitative results for GOPRO-II(r) image in Fig. 5 with remaining results included in supplementary.

We first examine the performance contribution of different mechanisms by fixing the transformer hyperparameters of GIQE *i.e.*, window size to 8, channel size i.e. C_1, C_2, C_3 to 120, 96, 48 respectively and number of self-attention heads to 6. Furthermore we fix the number of transformer blocks (N_1, N_2, N_3) for each scale as 36, 24 and 12 respectively. We provide the results for different model architectures in supplementary. Following this, we use the Swin transformer and determine the optimal framework hyperparameters. We first train the baseline for deblurring using GOPRO dataset without using the proposed degradation pipeline and evaluate performance on GOPRO-I, GOPRO-II(r), and GOPRO-III datasets. As the degradations within GOPRO-II(r) and GOPRO-III aren't included in the training dataset, the restoration quality is poor. However, when degradation space is increased by using the proposed N^{th} order degradation model, we observe improved restoration quality for other degradations while slightly decreasing for when the image has only motion-blur. Specifically for 1^{st} order degradation (Aug (N=1)), we observe a performance boost of +1.47db for GOPRO-II(r) and +1.7db for GOPRO-III, however -0.01db performance drop on GOPRO-I is also observed. We observe that increasing the degradation order improves performance on multiple degradations, with reduced performance on a single degradation dataset. We attribute this drop in performance to arise from the reduced amount of images within the dataset. While we observed improved performance when increasing the order of degradation, *i.e.*, +1.71db for GOPRO-II(r), the increase stagnates when increasing the degradation space to 5^{th} order *i.e.*, 1.12db. Furthermore, we observe the data loader becoming the bottleneck by consuming more time to process the inputs vis-a-vis the GIQE. Hence we limit the degradation order to 5.

After obtaining the degradation order and baseline, we

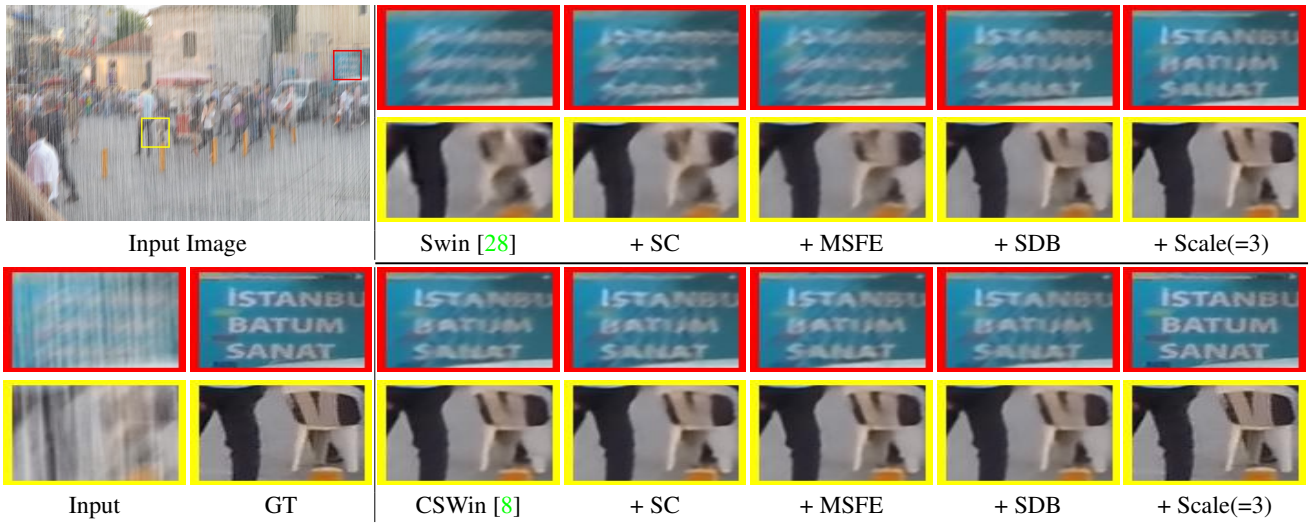


Figure 5. Qualitative evaluation of combining different mechanisms on restoration and reconstruction quality using Swin [28] and CSWin [8] transformer based models on GOPRO-II(r) subset

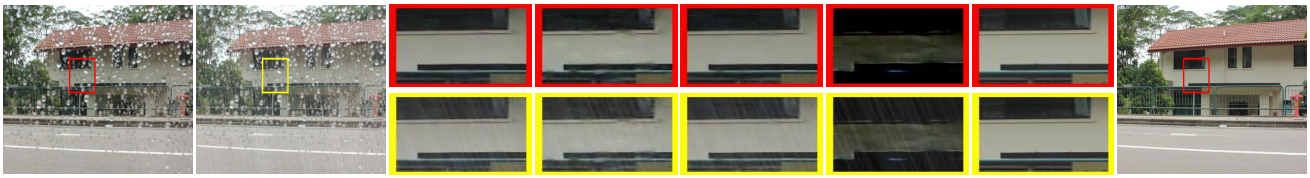
examine the impact of the spatial compression (SC) module via a 2D convolution filter or simply using an average pooling operation. While the pooling operation results in a linear computational complexity and a reduced number of parameters ($\sim$2.2M), it also results in a slight performance drop (-0.28db) compared to a convolutional filter-based spatial reduction. In addition, we observe SC mechanism to improve restoration performance under all degradation conditions surpassing the baseline (+0.18db, +0.22db for pooling and convolution-based SC, respectively). However, emphasizing computational complexity reduction, we choose the pooling operation for a spatial reduction since increasing the number of transformer mechanisms, or self-attention heads wouldn't increase the computations substantially. Following this, we modify the FFN within the transformer mechanism with the MSFE block and observe using separable convolutions with filter sizes 3 and 5 to result in higher performance (+0.25db, +0.33db, +0.55db for size 3, 5, and both respectively for GOPRO-I dataset) while increasing the number of parameters (0.3, 0.4, 0.7 M).

We then examine the effect of introducing the auxiliary decoder branch (SDB) within the training process and observe improved restoration performance for all degradation types *i.e.* +0.57db, +0.15db, and +0.07db. As the SDB mechanism is only utilized during training, it does not have an effect on the overall number of parameters of GIQE. We subsequently examine the effect of including multiscale branches whose features are merged using the feature selection module. We observe the performance to improve by +0.34db and +0.76db as we increase the number of scaled versions to 3 *i.e.* 1/8 and 1/16 scales. Finally, we examine an additional *trick* of sharpening the ground truth by using a randomized gaussian filter. This approach seems to improve the performance by +0.26db, +0.14db, and +0.08db for different datasets. As our evaluation was based on the Swin transformer, we now replace it with the CSWin transformer and observer improved performance with a reduced number of parameters for the same setting. Additionally, we observe that the presence of multiple degradations reduces the overall performance of the restoration algorithm, due to a large degradation space from which the restoration mapping needs to be learned. Hence we examine the role of increased training by $\times$100 epoch, we observe longer training to significantly improve the restoration performance when multiple degradations are affecting the image. We refer to the CSWin based multiscale GIQE version for remaining evaluations.

4.3. Image Restoration and Reconstruction

4.3.1 Single Degradation Restoration

We first compare the performance of the proposed mechanism with SoTA algorithms for restoring images affected by single degradation and summarize the results for Deblurring on GOPRO [32] in Tab. 2. We also examine the domain invariance characteristics ensured by the proposed training mechanism and transformer network. Additionally, we compare the performance with SoTA using RealBlur [18] dataset by inferencing models using GOPRO-pretrained weights. Based on these results, we observe the proposed GIQE to surpass previous SoTA for deblurring on GOPRO dataset while undergoing lower performance drop compared to SoTA when evaluated on images outside training distribution, *i.e.*, Real Blur dataset. We believe this characteristic to arise from the dual contribution of the degradation model and Transformer architecture. As highlighted by prior works [41,56] the ability of a restoration algorithm depends majorly on the degradation diversity within the training dataset. It should be noted the proposed GIQE outper-



Raindrop | Raindrop + Rain | A.GAN [34] | MPRNet [54] | DuRN [26] | Uformer [49] | GIQE | GT

Figure 6. Qualitative comparison of SoTA algorithms for raindrop removal and raindrop and rain removal on Raindrop [34] and Rain1400 [12] dataset.

Table 2. Performance comparison with SoTA image deblurring algorithms.

Algorithm	GoPRO [32]	RealBlur [18]	GMACs	# Params
	PSNR / SSIM	PSNR / SSIM	($x10^9$)	($x10^6$)
Input	25.64 / 0.79	26.55 / 0.80	-	-
DMPHN [55]	27.98 / 0.84	26.04 / 0.79	825.1	5.4
DeblurGANv2 [23]	28.92 / 0.89	26.68 / 0.81	411.55	5.0
MPRNet [54]	31.84 / 0.92	26.68 / 0.82	11169.5	20.1
DuRN [26]	28.00 / 0.85	26.37 / 0.81	3416.9	**3.7**
Uformer [49]	32.27 / 0.90	**30.74 / 0.88**	1235.2	27.3
GIQE	**33.48** / 0.94	30.18 / 0.84	**814.8**	24.6

Table 3. Ablation Performance comparison with SoTA raindrop removal algorithms.

Algorithm	Raindrop	Raindrop + Rain1400	GMACs	# Params
	PSNR / SSIM	PSNR / SSIM	($x10^9$)	($x10^6$)
Input	21.41 / 0.75	11.69 / 0.57	-	-
A.GAN [34]	23.68 / 0.75	19.79 / 0.59	531.9	**6.2**
Pix2PixHD [48]	24.05 / 0.69	17.43 / 0.58	412.9	182.4
DuRN [26]	23.91 / 0.75	17.88 / 0.62	332.9	10.1
EfficientDerain [14]	23.72 / 0.75	20.12 / 0.70	**296.2**	27.3
MPRNet [54]	24.19 / 0.79	20.47 / 0.66	4643.85	20.1
Uformer [49]	23.51 / 0.64	17.09 / 0.60	926.4	27.3
GIQE	**25.18 / 0.82**	**23.41 / 0.75**	465.6	24.6

forms transformer-based UFormer by a margin of 1.21db on GOPRO dataset while demonstrating higher robustness towards domain shifts. We observe similar performance characteristics on other single degradation restoration tasks such as deraining, desnowing, dehazing, and image enhancement We present necessary quantitative and qualitative results in supplementary.

4.3.2 Multiple Degradation Restoration

We subsequently evaluate the performance of SoTA and GIQE towards a more realistic scenario wherein multiple degradations co-exist. For this, we use the Raindrop dataset [34] and introduce synthetic rain following Rain1400 [12] design. We summarize the quantitative performance in Tab. 3 and qualitative performance in Fig. 6. To ensure SoTA can handle multiple datasets, we retrain them on the Raindrop dataset augmented by Rain1400 rain streak pattern following the training methodology proposed for each SoTA. We observe AGAN and DuRN perform unsatisfactorily when evaluating them on both images from the quantitative and qualitative results. We concur this to arise from their degradation specific architecture, whereas other SoTA were able to enhance the images considerably, however they weren't able to restore them to the quality when only raindrop was present. We believe this arises from the drawback of using convolutional filters, which makes them inflexible towards spatially varying degradations. We include comprehensive evaluation in supplementary wherein we observe similar patterns when evaluating on different degradation combinations.

4.4. Supplementary Materials

We examine the effect of restoration and reconstruction algorithms on high-level perception tasks *i.e.* semantic segmentation in natural conditions wherein multiple degradations co-exist, while also discussing the limitations. In summary, we demonstrate that it is more efficient to integrate an image restoration and reconstruction algorithm vis-a-vis retraining a model with deeper backbone.

5. Conclusion

In this paper, we argue that the fixed convolutional filters restrict the restoration and reconstruction quality owing to their inability to cope with spatially varying degradations. Furthermore, this restricts their ability to process images that are affected simultaneously by multiple degradations that are extensively encountered in natural conditions. To solve this issue, we propose a transformer-based image restoration algorithm provided with synthetic degraded images as input. For ensuring the complexity of degradation and its dynamic nature is accurately captured in training samples, we propose an N^{th} order, iterative degradation model. Additionally, we suggest two tricks to reduce the computational cost of the transformer model while increasing its ability to capture local features. We highlight that when training a restoration algorithm, the underlying network is expected to localize, quantify the magnitude and type of degradation and restore the affected regions. Thus, increasing the complexity of the training cycle. We integrate an auxiliary decoder branch performing binary classification to aid in training to identify degraded regions within an image. We conduct extensive experiments on both single and multiple degradation datasets to demonstrate the efficacy of the proposed approach.

6. Acknowledgement

This research was supported in part by KAIST-KU Joint Research Center, KAIST, Korea (N11200035) and Hankook Pixel (G06210065).

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint arXiv:1909.06581*, 2019. 2

[2] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[4] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 2

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021. 2, 3, 7

[9] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10729–10736, 2020. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Wenchao Du, Hu Chen, and Hongyu Yang. Learning invariant representation for unsupervised image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14483–14492, 2020. 1, 2

[12] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3855–3863, 2017. 8

[13] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 2

[14] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, and Yang Liu. Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. In *AAAI*, 2021. 2, 8

[15] Mazin Hnewa and Hayder Radha. Object detection under rainy conditions for autonomous vehicles: a review of state-of-the-art and emerging techniques. *IEEE Signal Processing Magazine*, 38(1):53–67, 2020. 1

[16] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013. 1

[17] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[18] Jucheol Won Sunghyun Cho Jaesung Rim, Haeyun Lee. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7, 8

[19] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 2

[20] Eunsung Jo and Jae-Young Sim. Multi-scale selective residual learning for non-homogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 507–515, June 2021. 2

[21] Alexander B. Jung. imgaug. https://github.com/aleju/imgaug, 2018. [Online; accessed 30-Oct-2018]. 6

[22] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8828–8838, 2020. 1

[23] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 2, 8

[24] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 1

[25] Tong Liu, Zhaowei Chen, Yi Yang, Zehao Wu, and Haowei Li. Lane detection in low-light conditions using an efficient data enhancement : Light conditions style transfer. In *2020 IEEE intelligent vehicles symposium (IV)*, 2020. 1

[26] 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. 2, 8

[27] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 3, 5

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 3, 7

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *arXiv preprint arXiv:1908.00682*, 2019. 6

[31] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 1

[32] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6, 7, 8

[33] Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7087–7093. IEEE, 2019. 6

[34] Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 8

[35] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11908–11915, 2020. 2

[36] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[37] Eduardo Romera, Luis M. Bergasa, Kailun Yang, Jose M. Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1312–1318, 2019. 1

[38] 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

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018. 2

[40] Pranjay Shyam, Antyanta Bangunharcana, and Kyung-Soo Kim. Retaining image feature matching performance under low light conditions. In *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, pages 1079–1085. IEEE, 2020. 1

[41] Pranjay Shyam, Sandeep Singh Sengar, Kuk-Jin Yoon, and Kyung-Soo Kim. Evaluating copy-blend augmentation for low level vision tasks. *arXiv preprint arXiv:2103.05889*, 2021. 5, 7

[42] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Towards domain invariant single image dehazing. *arXiv preprint arXiv:2101.10449*, 2021. 1, 2

[43] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Weakly supervised approach for joint object and lane marking detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2885–2895, 2021. 1

[44] Pranjay Shyam, Kuk-Jin Yoon, Sandeep Singh Sengar, and Kyung-Soo Kim. Lightweight hdr camera isp for robust perception in dynamic illumination conditions via fourier adversarial networks. In *The 32nd British Machine Vision Conference, BMVC 2021*. British Machine Vision Association (BMVA), 2021. 1

[45] 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

[46] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690A. International Society for Optics and Photonics, 2019. 1

[47] Li-Wen Wang, Zhi-Song Liu, Wan-Chi Siu, and Daniel P.K. Lun. Lightening network for low-light image enhancement. *IEEE Transactions on Image Processing*, 2020. 2

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6, 8

[49] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 3, 4, 5, 8

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 3

[51] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 193–202, June 2021. 2

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 5

[53] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. *arXiv preprint arXiv:2003.06792*, 2020. 2

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 2, 8

[55] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 8

[56] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *arxiv*, 2021. 7

[57] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[58] Yanfu Zhang, Li Ding, and Gaurav Sharma. Hazerd: an outdoor scene dataset and benchmark for single image dehazing. In *2017 IEEE international conference on image processing (ICIP)*, pages 3205–3209. IEEE, 2017. 6