This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Image restoration refinement with Uformer GAN

Xu Ouyang Ying Chen Kaiyue Zhu Gady Agam
Illinois Institute of Technology
{xouyang3, ychen245, kzhu6}@hawk.iit.edu, agam@iit.edu

Abstract

In this paper, we propose a novel approach for image restoration refinement, aiming to refine the result of restoring a clear original image from a noisy or blurry one. Our proposed method, Uformer GAN, combines the use of Transformer blocks and restoration refinement to achieve superior performance in image restoration tasks. The generator in our Uformer GAN model comprises Transformer blocks followed by a convolution layer. This design allows the model to learn the connections among each pixel of an image and capture local context features. The discriminator, on the other hand, consists of Transformer blocks and convolution blocks to balance the model's capability and efficiency. Additionally, instead of adopting multi-stage networks like other image restoration methods and training them concurrently, we solely focus on training a post-processing network for refined image restoration. This approach reduces the complexity of the overall image restoration process and ensures that our refinement is scalable to various image restoration techniques. We demonstrate the effectiveness of our proposed methods on two datasets: the image deblurring GOPRO dataset and the image denoising SIDD dataset. Our approach shows superior performance compared to other state-of-the-art methods in both datasets.

1. Introduction

Image restoration is a widespread technique used in various computer vision applications, such as image denoising, image deblurring, image deraining and so on. Recently, convolutional neural network-based autoencoders have achieved significant success in image restoration tasks through different architectures and convolution layers (e.g., [4, 18, 21–23, 29]). However, these models have limitations in capturing high-resolution features due to the limited capacity of the convolutional receptive field.

Generative Adversarial Networks (GANs) [11] provide an effective deep neural network framework that can capture data distributions, resulting in sharper and more realistic textures than traditional convolutional models. In GANs model, the generative network maps from a latent space to a data distribution of interest, while the discriminative network distinguishes between generated and real data distributions. GANs have achieved significant success in image generation tasks, such as image deblurring [13], image super-resolution [15], caption generation from images [19], and image generation from captions [3]. In particular, various GANs structures and loss functions have been proposed for image deblurring tasks (e.g., [13], [14], [32]). While GANs improve the upper bound of the capacity of convolutional models, the fundamental problem of the limitation of convolutional receptive fields remains.

The Vision Transformer [9] introduces a self-attention based neural network, which is a novel technique different from convolution-based models in image classification tasks. The transformer builds connections among each pixel of the image, thus overcoming the shortcomings of convolution layers. Inspired by the variant of the Vision Transformer, known as the Swin Transformer [16], [27] designed a Uformer for image restoration tasks. The Uformer is a hierarchical encoder-decoder based on UNet [24] skip connections structure, where convolution layers are replaced with Locally-enhanced Window (Lewin) Transformer blocks. However, it is still an autoencoder model and lacks the capabilities of GANs, which can learn to map from a latent space to a data distribution.

In this paper, we aim to combine the advantages of the Transformer and GANs models, which not only can learn the connections among each pixel of the image but also can learn to map from one data distribution to another. Thus, we propose the Uformer GAN, which has a UNet-shaped structure with Transformer blocks in the generator and a combination of Transformer blocks and convolution layers in the discriminator. Furthermore, to improve the image restoration quality, we use the output of the Uformer as the input of the Uformer GAN directly for training.

In the Uformer GAN model, the Swin Transformer block from [16] is used in both the generator and discriminator. The Swin Transformer block includes non-overlapped windows and shifted window-based self-attention, which

can capture global features and has linear computational complexity, compared to self-attention, which has quadratic computational complexity. To capture local features and reduce the complexity of the neural network, we introduce convolution layers in both the generator and discriminator.

During training, we adopt a distinct approach from other multi-stage training methods [10, 25, 29, 31]. These methods train multiple neural networks in one step and perform back-propagation from the output of the last model to the input of the first model. In contrast, our process involves two steps. First, we train a Uformer for image restoration until it converges. Then, in the second step, we train a Uformer GAN to refine the result from the Uformer. Inspired by semantic segmentation refinement works, CascadePSP [7] and SegFix [28], we are the first to introduce this image restoration refinement method, which offers three advantages: the second step can refine the results from any kind of neural networks for image restoration tasks; the training system is deep enough and can even exceed the capacity of GPU memory since we do not need to do the back-propagation of multiple neural networks simultaneously; this refinement method can prevent the gradient vanishing problem since there are two individual deep neural networks separately trained in two different steps. Additionally, we have skip connections in the first step, so the input of the second step includes all the information of the input of the first step, which means the two-step training can still reach optimal results.

To demonstrate the effectiveness of our proposed Uformer GAN for image restoration refinement, we conducted experiments and compared with existing image restoration methods such as the convolution-based HINet [4], transformer-based Uformer [27], and multi-stage based MPRNet [29]. Our experimental results indicate that our proposed method achieved higher PSNR and SSIM scores than the other methods on the image deblurring GOPRO [18] and image denoising SIDD [1] datasets.

In summary, this paper makes the following contributions: 1) We propose a Uformer GAN that combines the benefits of UNet-shaped and transformer-convolution layer based autoencoder in the generator, and a transformer-convolution layer combined model in the discriminator. 2) We introduce image restoration refinement method where we first train a Uformer until convergence, and then fine-tune the results using a Uformer GAN. 3) We conduct experiments on various image restoration tasks and show improvement (0.1 dB on PSNR) over state-of-the-art approaches. Note that the significance of this improvement is consistent with prior improvements in this area.

2. Related Work

MPRNet [29] proposes a three-stage progressive neural network for image restoration tasks. The first two stages apply two autoencoders to learn multi-scale local context, and the third stage employs an original-resolution subnetwork to learn global context. A supervised attention module is added between each stage to refine the results before they are passed to the next stage. Cross-stage feature fusion is also used to propagate contextualized features between stages. However, the system is complex with multiple stages and added modules.

DeblurGAN [13] is a conditional GAN for image restoration tasks. The generator is an autoencoder with residual blocks and a global skip connection, while the discriminator is similar to PatchGAN [12]. Wasserstein GAN with gradient penalty and perceptual loss are used for training. In DeblurGAN v2 [14], the generator uses an Inception-ResNet-v2 with a Feature Pyramid Network, and the discriminator uses a relativistic discriminator with least-square loss. Two branches are integrated into evaluate global and local features, and MSE loss is added to the generator loss to correct both color and texture distortions. However, the problem of the limited convolutional receptive field still exists.

Uformer [27] is the first work to apply vision transformer for image restoration tasks. It uses a Unet-shaped autoencoder with skip-connections. To capture local context, convolutional layers are added to the encoder and transposed convolutional layers are added to the decoder. To capture global context, Locally enhanced Window (LeWin) Transformer blocks are added to each convolutional layer in both the encoder and decoder. These blocks use non-overlapped windows-based self-attention to capture global dependencies of features with linear computational complexity. Additionally, a learnable multi-scale restoration modulator is proposed to adapt to different image degradations in each layer of the decoder. It uses a multi-scale spatial bias to adjust features and restore more details. The use of self-attention mechanism in Uformer leads to state-of-the-art performance in various image restoration tasks. In this paper, we explore the potential of Uformer by building a Uformer GAN, which combines the Uformer model with GANs model.

RFormer [8] introduces a transformer-based Generative Adversarial Network for real fundus image restoration. It constructs both the generator and discriminator using Window-based Self-Attention Blocks (WSABs). However, they utilize multifaceted loss functions, still based on the original GAN loss, and only offer one stage of training. Inspired by this work, we present the Uformer GAN model. Our model employs WGAN loss for both the generator and discriminator. To enhance computational efficiency, we utilize shifted window self-attention, reducing complexity from quadratic to linear. Furthermore, our design incorporates a partially transformer-based discriminator to optimize memory usage.

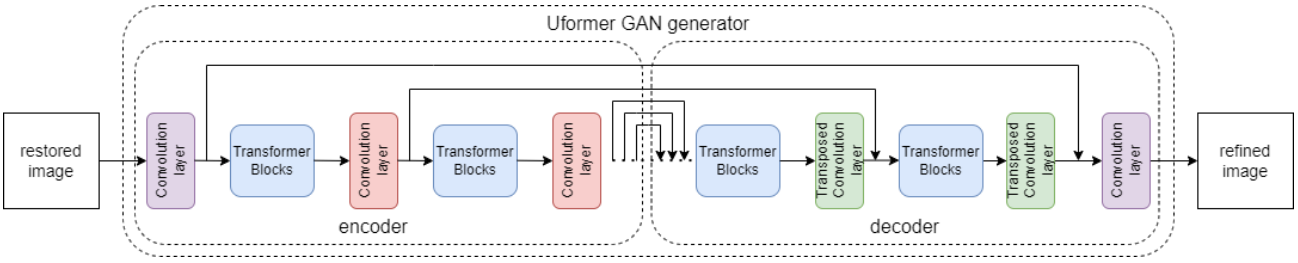

Figure 1. The structure of the generator in the Uformer GAN. The generator has two parts: the encoder which captures multi-scale feature contexts and the decoder which restores features. The feature maps in the decoder are stacked with the corresponding feature maps in the encoder using skip connections. In the encoder, the refined image is processed by a convolution layer and then passed through several Transformer blocks followed by convolution layers. In the decoder, the image is passed through several Transformer blocks followed by transposed convolution layers and finally processed by a convolution layer to obtain the refined image.

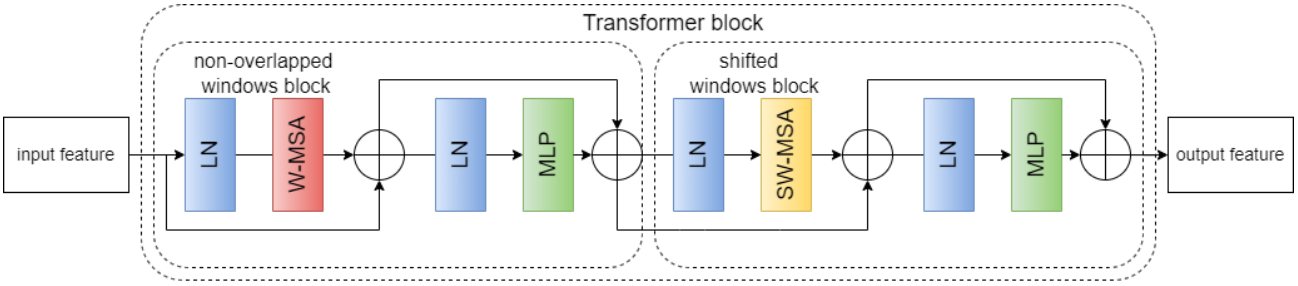

Figure 2. The structure of the Transformer block. There are two window-based blocks: a non-overlapped windows block (NW block) that consists of a standard multi-head self-attention (MSA) module, followed by a 2-layer MLP with GELU nonlinearity in between. A LayerNorm layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. A shifted window block replaces the MSA with a shifted window-based MSA which builds cross-window connections. The remaining design elements are the same as in the NW block.

3. Proposed approach

3.1. Architecture

In this section, we present our Uformer GAN model. The Uformer GAN model consists of two deep neural networks: the generator G, which is used to restore the image, and the discriminator D, which is used to determine whether the generated image from G is real or fake.

The generator, as shown in Figure 1, comprises an encoder that captures multi-scale feature contexts and a decoder that restores features from the encoder. Meanwhile, the generator builds up several skip connections between the encoder and the decoder. In the encoder part of the generator, the image first passes through a convolution layer to obtain non-overlapping patches. These are then reshaped into 2D patch features. Subsequently, the patch feature is passed through several Transformer blocks, in conjunction with Convolution layers.

The Transformer block is used to capture the global feature context. It includes a non-overlapped windows block (NW block) followed by a shifted windows block (SW block) as shown in Figure 2. The NW block consists of a standard multi-head self-attention (MSA) module, followed by a 2-layer MLP with GELU nonlinearity in between. A LayerNorm layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. In the SW block, the MSA is replaced with a shifted window-based MSA which builds cross-window connections and the remaining design elements are the same as in the NW block.

The convolution layer after each Transformer block is used to capture local feature context and downsample the feature for lower-level feature learning. To make the feature suitable for the convolution layer, the 2D patch feature from the Transformer block is reshaped into a 3D feature map. After passing through each convolution layer, the 3D feature map is reshaped back to a 2D patch feature for subsequent Transformer block learning.

In the decoder part of the generator, the patch feature is passed through several Transformer blocks together with transposed convolution layers. The Transformer blocks are the same as in the encoder and are used for global feature context learning. The transposed convolution layer after each Transformer block is used to capture local feature

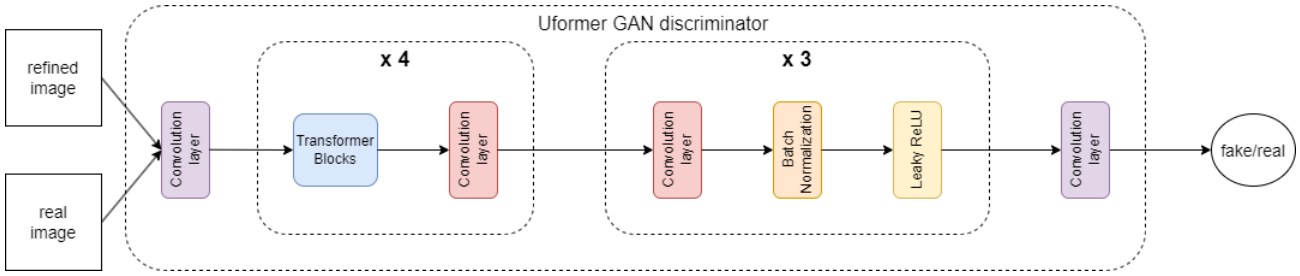

Figure 3. The structure of the discriminator in the Uformer GAN. The discriminator has two stages: four Transformer Blocks followed by a convolution layer which can learn global and local context features; three convolution layers followed by batch normalization layers and Leaky ReLU activation functions to balance model power and efficiency. The feature map is finally passed through a convolution layer to obtain a 1D output for fake or real prediction.

context and upsample the feature for higher-level feature learning. After passing through each transposed convolution layer, the feature map is stacked with the corresponding feature map in the encoder using skip connections. In the end, the feature map is passed through a convolution layer to obtain the restored image.

The discriminator, as shown in Figure 3, is a classifier that is used to distinguish whether the input image is real or fake. The image is first passed through a convolution layer to obtain a feature map, which is then reshaped into a 2D patch feature. The patch feature is then passed through several Transformer blocks together with convolution layers, with the same design as in the generator. The Transformer block is used to capture global feature context, and the convolution layer after the Transformer block is used to capture local feature context and downsample the feature.

To balance memory and computation costs, we propose a solution. The patch feature is reshaped into a 3D feature map. This occurs after passing through four Transformer blocks along with convolution layers. The feature map is then passed through three convolution blocks consisting of a convolution layer, batch normalization (BN) layer, and a Leaky ReLU activation function. The Batch Normalization layer is crucial to ensure that the model converges, as we found that the model will collapse without it. In the end, the feature map is passed through a convolution layer to obtain a 1D output for fake or real prediction.

3.2. Algorithm

In this section, we present our proposed Uformer GAN for image restoration refinement as outlined in Algorithm 1. In the first step, the degraded image x is passed through the Uformer (U) to obtain the restored image y'. We then calculate the Charbonnier loss $L_U(y', y)$ defined as:

$$L_U(y', y) = \sqrt{||y' - y||^2 + \epsilon^2} \quad (1)$$

where $y' = U(x)$, y is the ground-truth image, and ϵ is a constant set to 10^{-3} for all experiments, following the setting in [27]. The parameters of the Uformer, θ_u, are then updated using the Adam optimizer with decoupled weight decay (AdamW) [17] until convergence. The AdamW optimizer decouples weight decay from the gradient-based update, which improves regularization in Adam and has recently been used in Vision Transformer training.

Algorithm 1 Image restoration refinement training

- **parameters**: Uformer (U) parameters (θ_u) in step 1; Uformer GAN generator (G) parameters (θ_g) and discriminator (D) parameters (ψ_d) in step 2.
- **variables**: input image (x) and output image ($y\prime$) in step 1; output image ($y\prime\prime$) in step 2; real image (y) in both steps.

for iteration in step 1 **do**
    compute the loss $L_U(y\prime, y)$
    update the parameters θ_u by AdamW optimizer
end for
for iteration in step 2 **do**
    compute the loss $L_G(y\prime\prime, y)$
    update the parameters θ_g by AdamW optimizer
    compute the loss $L_D(y\prime\prime), y)$
    update the parameters ψ_d by AdamW optimizer
end for

In the second step, the Uformer GAN attempts to capture the data distribution using the restored image y'. GANs are modeled as a min-max two-player game between a discriminator network $D_\psi(x)$ and a generator network $G_\theta(z)$. The optimization problem in GANs is defined as:

$$\min_G \max_D V(G, D) = \mathbb{E}x \sim pdata[f(D(x))] + \mathbb{E}z \sim platent[f(-D(G(z)))] \quad (2)$$

where $G : Z \to X$ maps from the latent space Z to the input space X; $D : X \to \mathbb{R}$ maps from the input space to a classification of the example as fake or real; and $f : \mathbb{R} \to \mathbb{R}$ is a concave function. The GANs model will reach the

global optimal when $p_{gen} = p_{data}$, where p_{gen} is the generative data distribution and p_{data} is the real data distribution. However, in practice, GANs have the well-known problem of training instability. To alleviate this problem, we use the Wasserstein GAN loss [2] in our Uformer GAN model obtained when using $f(x) = x$.

Following the above GANs training processing, the restored image y' is passed through the generator G of the Uformer GAN to obtain the refined image y''. Both the refined image y'' and the ground-truth image y are then passed through the discriminator D of the Uformer GAN. However, since the restored image y' is already of high quality, the discriminator struggles to distinguish it from the ground-truth image. As a result, we train the generator and discriminator separately. During the initial stages of training, the generator loss does not include an adversarial loss. The early-stage generator loss $L_{G_i}(y'', y)$ is defined as:

$$L_{G_i}(y'', y) = L_{cha}(y'', y) + 100.0 * L_{mse}(y'', y) \quad (3)$$

Where L_{cha} is the Charbonnier loss, which is effective at handling outliers and improving performance, and L_{mse} is the mean square error, which helps correct color and texture distortions. The discriminator loss $L_D(y'', y)$ is also computed as:

$$L_D(y'', y) = L_{adv}(D(y'')) - L_{adv}(D(y)) \quad (4)$$

Where L_{adv} is the WGAN loss for the discriminator. The generator parameters θ_G and the discriminator parameters ψ_D are updated using the AdamW optimizer. After several training epochs, we introduce a new generator loss $L_G(y'', y)$ which is defined by:

$$L_G(y'', y) = L_{adv}(D(y'')) + L_{cha}(y'', y) + 100.0 * L_{mse}(y'', y) \quad (5)$$

Where L_{adv} is the WGAN loss for the generator, which aims to learn the distribution of real images Y from the distribution of restored images Y'. The generator parameters θ_G and the discriminator parameters ψ_D are updated using the adaptive update strategy proposed in [20]. This strategy balances the number of updates for the generator and discriminator based on the change ratios of $L_D(y'', y)$ and $L_G(y'', y)$. Unlike traditional fixed-number updating strategies, this strategy updates the generator or discriminator by comparing the weighted loss change ratios, which are the differences between the current and previous losses. This can accelerate the training convergence and reach the optimal solution. Our two-step training strategy also allows reaching the optimal solution by adding the input image to the end of the network in the first step, which means we still keep the data flow in the second step.

4. Experimental results

4.1. Dataset

We evaluate the image deblurring task using the GOPRO dataset [18]. This dataset is designed for dynamic scene deblurring and was created by taking 240 fps videos with GOPRO4 Hero Black camera, and then averaging a number of successive latent frames to produce blurry images. The dataset includes 3,214 blur/sharp image pairs, each with a resolution of 720x1080. We split these image pairs into a training dataset of 2,103 pairs and a testing dataset of 1,111 pairs, following the method outlined in [27]. During training, we randomly crop the images to a size of 256x256 as the input for the neural network.

We evaluate the image denoising task using the SIDD dataset [1]. This dataset contains 30,000 noisy images from 10 scenes under different lighting conditions, taken with several different smartphone cameras. For our image denoising experiments, we use the SIDD Medium dataset which consists of 320 noisy/clear standard RGB (sRGB) image pairs, with two image pairs from each scene. We split each image into several 256x256 size patches, generating a training dataset of 96,000 pairs and a testing dataset of 1,280 pairs, following the method outlined in [27].

4.2. Implementation details

We employ various data augmentation techniques when loading images into the neural network. We use the AdamW optimizer for training both the Uformer and the Uformer GAN. The initial learning rate is set to $2e-4$, and a cosine learning rate decay is employed to gradually decrease it. In the first step of training, a warm-up period of 10 epochs is utilized for the Uformer. We follow the original Uformer-B architecture with depths $1, 2, 8, 8$ in the encoder and $8, 8, 2, 1$ in the decoder, and set the window size to 8x8 in all Transformer blocks. In the second step of training, we use a warm-up period of 50 epochs for the Uformer GAN, and separately update the generator and discriminator during this period, since the discriminator is difficult to train with already high-performing input images y'. The generator has the same depths and window size as in Uformer. The discriminator has depths $1, 2, 8, 8$ Transformer blocks with a window size of 8x8.

In both image deblurring and image denoising tasks, we use two popular evaluation metrics: the structural similarity index (SSIM) [26] and the peak signal to noise ratio (PSNR) to measure the similarity between the restored images and the target images. SSIM is based on computing the mean, variance and co-variance of a variety of windows of two different images, while PSNR is based on the inverse of the mean squared error (MSE). Generally, higher values for both metrics indicate that the restored images are more similar to the real images.

Figure 4. Comparison of our Uformer GAN with the Uformer for deblurring task using three different blur/sharp image pairs from the GOPRO dataset. The first column shows the blur image, the second column shows the corresponding sharp image, the third column shows the deblurred image produced by the Uformer, and the fourth column shows the refined deblurred image produced by our Uformer GAN.

4.3. Qualitative evaluation

Figure 4 shows four different visualized results on the image deblurring GoPro dataset. We compared our Uformer GAN with the baseline method Uformer model. As we can see, in the first row, Uformer GAN gets more texture details on the bottom of bag; in the second row, Uformer GAN gets more texture details on the top right of background; in the third row, Uformer GAN synthesizes darker wall on the left region; in the fourth row, Uformer GAN generates brighter lines between two vehicles. Figure 5 shows four different visualized results on the image denoising SIDD dataset. We also compared our Uformer GAN with the Uformer model.

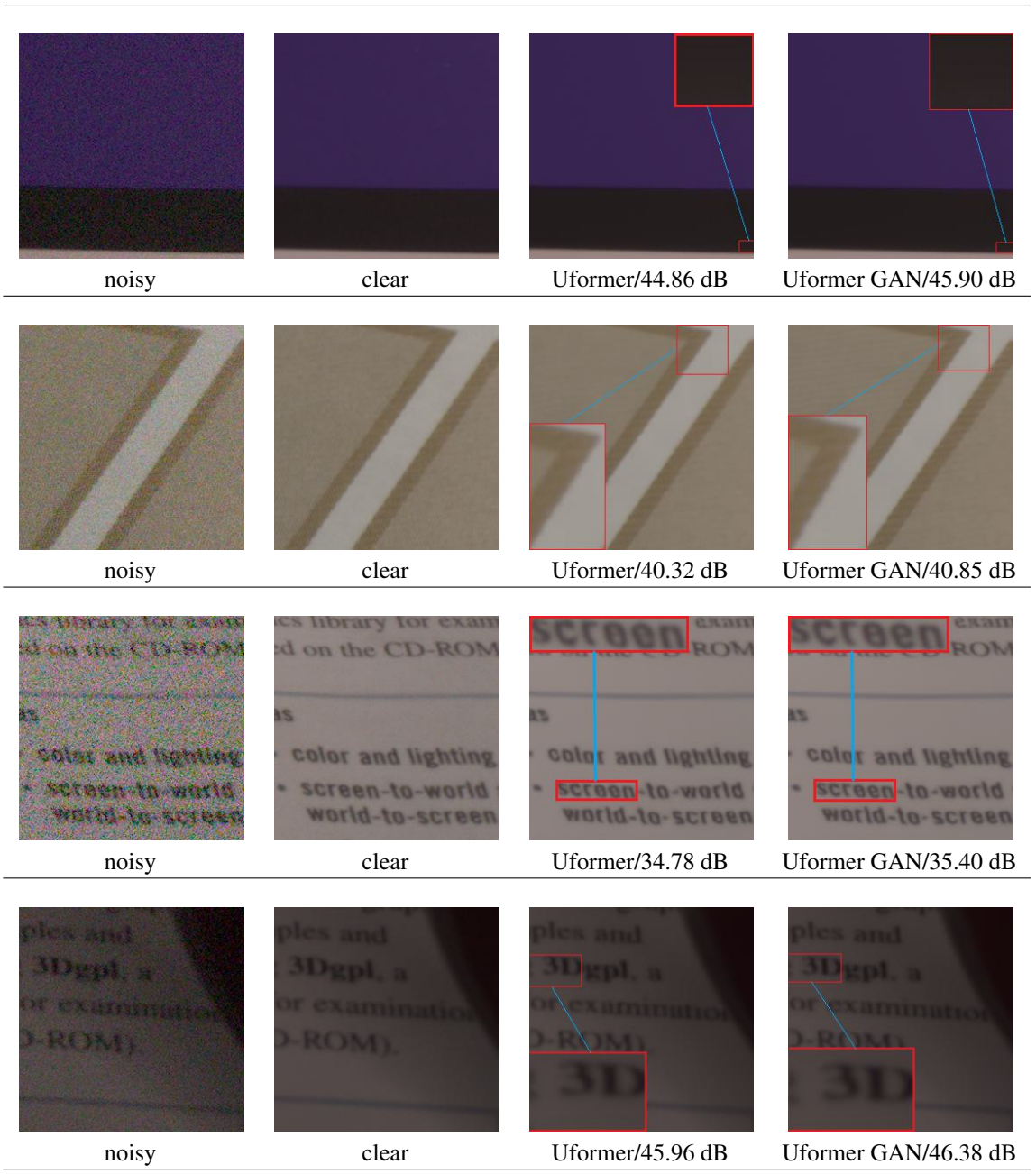

Figure 5. Comparison of the results of our Uformer GAN and the Uformer for image denoising using three different noisy/clear image pairs from the SIDD dataset. The first column shows the noisy image, the second column shows the clear image, the third column shows the denoised image generated by the Uformer, and the fourth column shows the refined denoised image generated by our Uformer GAN.

As we can see, in the first row, Uformer GAN synthesizes clear grids near the edge; in the second row, Uformer GAN generates stripes which color is more similar than Uformer; in the third row, Uformer GAN synthesizes a clear lowercase 'e' without connection between the tail and the middle line; in the fourth row, Uformer GAN generates more contents near the edge. In conclusion, our approach results in clearer and more detailed textures in image deblurring task and in less noise and fewer texture distortions in image denoising task.

Metrics	Dataset	DBGAN	MPRNet	HINet	Uformer	Uformer GAN (refine)
PSNR↑	GOPRO	31.10	32.66	32.71	33.06	33.17
SSIM↑	GOPRO	0.942	0.959	0.962	0.967	0.970
PSNR↑	SIDD	35.78	39.72	39.99	39.89	39.97
SSIM↑	SIDD	0.919	0.959	0.958	0.960	0.962

Table 1. Comparison of our novel two-step Uformer GAN and four previous state-of-the-art methods on the GOPRO dataset and the SIDD dataset. Each column represents a different image restoration model. The first and third row are the PSNR scores of each method, and the second and fourth row are the SSIM scores of each method. Our Uformer GAN for refinement achieves the best performance compared to all other methods on the GOPRO dataset, and the best SSIM and approximate best PSNR on the SIDD dataset.

4.4. Quantitative evaluation

We compare our Uformer GAN with GAN-based models DeblurGAN-v2 [14] and DBGAN [32], a convolution-based model HINet [4], a multi-stage model MPRNet [29], and a transformer-based model Uformer [27]. We evaluate the image deblurring task on the GOPRO dataset, and the image denoising task on the SIDD dataset. We use PSNR and SSIM scores, which are well-known image restoration performance evaluation metrics. We show the evaluation results for both image debluring and image denoising task in Table 1. As observed, our novel Uformer GAN achieves the highest PSNR and SSIM scores compared to other methods on both datasets. As can be observed, our Uformer GAN achieves the highest PSNR and SSIM scores compared to other methods on the GOPRO dataset. And our Uformer GAN reaches the highest SSIM scores and approximate highest PSNR scores. In summary, our approach results in better performance than other image restoration methods.

4.5. Ablation Study

We investigate the effects of GAN model and image restoration refinement on the image deblurring GOPRO dataset and the image denoising SIDD dataset. We train in four different ways: a one-step training that trains a Uformer to obtain the restored image; a one-step training that trains a Uformer GAN model to obtain the restored image; a two-step training that trains a Uformer until convergence in step 1, then trains another Uformer with the restored image from step 1; and a two-step training that trains a Uformer until convergence in step 1, then trains a Uformer GAN with the restored image from step 1. We show the ablation study results in Table 2. Notably, the one-step Uformer GAN approach outperformed the singular Uformer method. The two-step Uformer model also surpassed the one-step Uformer, while the combination of two-step Uformer and Uformer GAN excelled over the standalone two-step Uformer. In essence, the ablation study highlights the efficacy of our Uformer GAN and the advantages of adopting a two-step training paradigm for superior image restoration.

5. Conclusion

In this paper, we propose a Uformer GAN model and image restoration refinement method. The Uformer GAN is a combination of Transformer and Convolution based neural networks which can capture both global and local context features. Additionally, the image restoration refinement method trains a Uformer in the first step, and then trains a Uformer GAN in the second step using the output from the first step as input, which can help refine the restored image. We demonstrate that our method outperforms the Uformer in both image deblurring and image denoising tasks. In future work, we plan to apply our GAN-based refinement method to different pretrained models for various image generation tasks, such as Restormer [30], NAFNet [5], and Diffusion models [6] for image restoration refinement.

Metrics	Dataset	Uformer	Uformer GAN	Uformer (**refine**)	Uformer GAN (**refine**)
PSNR↑	GOPRO	33.06	33.09	33.13	33.17
SSIM↑	GOPRO	0.967	0.967	0.969	0.970
PSNR↑	SIDD	39.89	39.92	39.95	39.97
SSIM↑	SIDD	0.960	0.959	0.962	0.962

Table 2. Ablation study measuring the effect of the Uformer GAN model and the two-step training strategy on image deblurring (GOPRO dataset) and the image denoising (SIDD dataset). Each column represents different Uformer models and training strategies. The first and third row are the PSNR scores of each method, and the second and fourth row are the SSIM scores of each method. Our Uformer GAN for refinement achieves the best performance compared to all other methods.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 2, 5

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017. 5

[3] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. Improving image captioning with conditional generative adversarial nets. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019. 1

[4] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 182–192, 2021. 1, 2, 8

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision – ECCV 2022*, pages 17–33, Cham, 2022. Springer Nature Switzerland. 8

[6] Zheng Chen, Yulun Zhang, Ding Liu, bin xia, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *Advances in Neural Information Processing Systems*, pages 29114–29125. Curran Associates, Inc., 2023. 8

[7] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8887–8896, 2020. 2

[8] Zhuo Deng, Yuanhao Cai, Lu Chen, Zheng Gong, Qiqi Bao, Xue Yao, Dong Fang, Wenming Yang, Shaochong Zhang, and Lan Ma. Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4645–4655, 2022. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[10] Longtao Feng, Xinfeng Zhang, Xiang Zhang, Shanshe Wang, Ronggang Wang, and Siwei Ma. A dual-network based super-resolution for compressed high definition video. In *Advances in Multimedia Information Processing – PCM 2018*, pages 600–610, 2018. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 1

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 2

[13] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018. 1, 2

[14] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8877–8886, 2019. 1, 2, 8

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 1

[16] 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. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 4

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 257–265, 2017. 1, 2, 5

[19] Xu Ouyang, Xi Zhang, Di Ma, and Gady Agam. Generating image sequence from description with lstm conditional gan. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2456–2461, 2018. 1

[20] Xu Ouyang, Ying Chen, and Gady Agam. Accelerated wgan update strategy with loss change rate balancing. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2545–2554, 2021. 5

[21] Kuldeep Purohit and A. N. Rajagopalan. Region-adaptive dense network for efficient motion deblurring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 11882–11889, 2020. 1

[22] Wenqi Ren, Jiawei Zhang, Jinshan Pan, Sifei Liu, Jimmy S. Ren, Junping Du, Xiaochun Cao, and Ming-Hsuan Yang. Deblurring dynamic scenes via spatially varying recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3974–3987, 2022.

[23] Wenqi Ren, Senyou Deng, Kaihao Zhang, Fenglong Song, Xiaochun Cao, and Ming-Hsuan Yang. Fast ultra high-definition video deblurring via multi-scale separable network. *International Journal of Computer Vision*, 2023. 1

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation.

In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 1

[25] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3603–3612, 2020. 2

[26] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[27] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17662–17672, 2022. 1, 2, 4, 5, 8

[28] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506, 2020. 2

[29] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14816–14826, 2021. 1, 2, 8

[30] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming–Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 8

[31] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5971–5979, 2019. 2

[32] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Björn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743, 2020. 1, 8