




WaveNet: Wave-Aware Image Enhancement

Jiachen Dang^{†1,2} , Zehao Li^{1,2}, Yong Zhong^{1,2} , Lishun Wang^{1,2} 

¹Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, China

²University of Chinese Academy of Sciences, Beijing, China

Abstract

As a low-level vision task, image enhancement is widely used in various computer vision applications. Recently, multiple methods combined with CNNs, MLP, Transformer, and the Fourier transform have achieved promising results on image enhancement tasks. However, these methods cannot achieve a balance between accuracy and computational cost. In this paper, we formulate the enhancement into a signal modulation problem and propose the WaveNet architecture, which performs well in various parameters and improves the feature expression using wave-like feature representation. Specifically, to better capture wave-like feature representations, we propose to represent a pixel as a sampled value of a signal function with three wave functions (Cosine Wave (CW), Sine Wave (SW), and Gating Wave (GW)) inspired by the Fourier transform. The amplitude and phase are required to generate the wave-like features. The amplitude term includes the original contents of features, and the phase term modulates the relationship between various inputs and fixed weights. To dynamically obtain the phase and the amplitude, we build the Wave Transform Block (WTB) that adaptively generates the waves and modulates the wave superposition mode. Based on the WTB, we establish an effective architecture WaveNet for image enhancement. Extensive experiments on six real-world datasets show that our model achieves better quantitative and qualitative results than state-of-the-art methods. The source code and pretrained model are available at <https://github.com/DeniJsonC/WaveNet>.

CCS Concepts

• Computing methodologies → Image processing;

1. Introduction

Challenges for enhancing degraded images exist not only in the real world but also in computer vision tasks. Poor photographing environments, the improper operation of the photographer, or limitations of camera devices often produce the low-quality photos. These degraded images apply harmful visual effects and bad effects on other computer vision tasks. As a low-level computer vision task, image enhancement provides reliable information for downstream visual decisions. Recently, multiple approaches [ZGM*21, TTZ*22, ZAK*20, CWG*21] based on CNNs, MLP, and Transformer have been designed for image enhancement. These methods perform well on benchmark datasets, but some may cause unacceptable computational costs and weak versatility.

Different from the method that delicately designs the model to achieve high accuracy without thinking of generalization and computational cost, we aim to explore a robust architecture that can perform well with different efficiency for image enhancement. Besides, we desire to improve the representation way of features for dynamically aggregating them according to semantic contents.

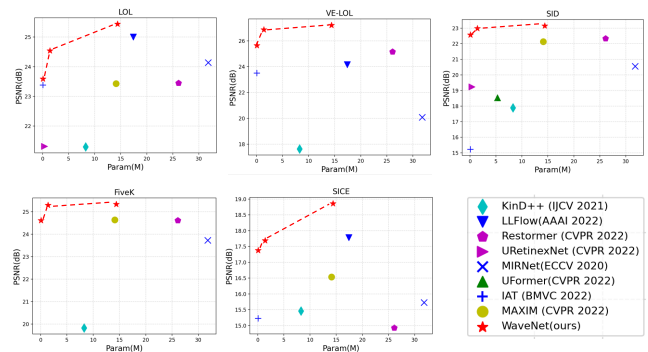


Figure 1: Our method WaveNet achieves the best PSNR-Params trade-off on several image enhancement datasets.

In this paper, our inspiration comes from signal processing. For an image captured from a digital camera, we simply think that every pixel retains the optical information independent of other pixels. *e.g.* Each pixel value has recorded a specific light intensity of the captured optical signal. Thus, a digital image that comes from the real world is a discrete digital matrix generated by sampling, quantizing, and encoding a continuous optical signal set. If we can formulate the optical signal recorded by the pixel, we can enhance

[†] Corresponding author:
Mail: dj.chen112@gmail.com (J. Dang)

the degraded image according to modulating the signals. Thus, we transform the enhancement task into a signal modulation problem.

How to "present" and "modulate" the signal using learning-based approaches is the key to our work. **1)** For "presenting" signals, the Fourier transform theory provides a way of decomposing signals into different frequency sine/cosine waves. Therefore, we describe a pixel as a sample value of a *signal* with three *waves* (Cosine Wave (CW), Sine Wave (SW), and Gating Wave (GW)) to realize the *wave-like* feature representations. For constructing the waves, amplitude and phase are indispensable. The amplitude part measures the maximum intensity of a wave. The phase part represents the initial state of the wave and contains the frequency information. **2)** For "modulating" signals, we have to consider not only a single signal modulation but also aggregating the semantic information of signals in a region. Only when we take into account the semantic information in local and non-local, we better modulate the signal. *e.g.* An individual red pixel provides vague semantic information. We tend to retouch it if the pixel is in an apple, or we tend to remove it as a noise if the pixel is in a banana. Therefore, we need to adaptively modulate these signals with the support of semantic information. Considering above analyses of presenting and modulating waves, we propose the Wave Transform Block (WTB in Figure 3) that allows for efficient and scalable spatial mixing of local and non-local contents and dynamically learns the interaction between waves to enhance the images. Besides, to obtain high-quality results, we use the gating mechanism to control which complementary features should flow forward and allow sub-blocks to focus exclusively on more refinement of image attributes. Furthermore, we build the WaveNet (in Figure 2), an effective architecture using WTB. Section 3 discusses the proposed WaveNet in detail.

Figure 1 shows that the proposed WaveNet achieves the best trade-off between accuracy and complexity on various datasets. For example, WaveNet-T obtains the **23.59dB** on LOL dataset [WWYL18] with only **80k** parameters and outperforms MAXIM [TTZ*22] by **0.16 dB** and IAT [CLG*22] by **0.21dB**. WaveNet-B achieves **25.44dB** compared with the previous state-of-the-art method LLFlow [WY*22] and obtains **0.44dB** gain in PSNR. Besides, WaveNet also achieves strong performance on the VE-LOL [LXY*21], SICE [CGZ18], MIT-Adobe FiveK [BPCD11] and SID [CCXK18] four image enhancement datasets. For high-level vision tasks, we use face detector DSFD [LWW*19] to evaluate dark face detection tasks using images enhanced by various low-light image enhancement methods. Our WaveNet achieves the best results on the DARK FACE dataset [YYR*19].

Overall, our contribution could be summarized as follows:

- We propose a new way of enhancing feature representation, dubbed *wave-like* feature representation. We aggregate the features with three waves: Cosine Wave (CW), Sine Wave (SW), and Gating Wave (GW).
- We propose the Wave Transform Block (WTB) that is capable of aggregating local and channel information and modeling wave interactions to enhance the original degraded image. We build WaveNet, an effective architecture using WTB for image enhancement.
- Extensive experiments on popular real-world datasets show that our WaveNet achieves SOTA results.

2. Related Work

2.1. Learning-based image enhancement

With the development of deep learning, an amount of research which are learning-based has emerged. Since the ground-breaking methods, LLNet [LAS17] is proposed, learning-based methods have greatly improved. Compared with traditional methods, learning-based methods are more accurate, robust, and faster. These methods used a variety of learning strategies. Most of them used supervised learning, *e.g.* Retinex-Net [WWYL18], DeepUPE [WZF*19a], KinD [ZZG19a], LPNet [LLF*20], DLN [WLSL20], PRIEN [LFH21], and etc. In recent works, MIRNet [ZAK*20] presented parallel multi-resolution convolution streams for extracting multi-scale features to enhance the degraded images. MAXIM [TTZ*22] used two MLP blocks to aggregate local and non-local contents for image restoration. IAT [CLG*22] presented a light-weight network for image enhancement. Nevertheless, these works are unable to achieve promising results on both accuracy and efficiency as the proposed WaveNet.

2.2. Fourier transform

The Fourier transform is proposed to analyze thermal processes and is widely used in various fields due to its excellent performance. In recent years, Fourier transform has been applied in learning-based image processing methods CirCNN [DLW*17] substantially reduces computational complexity and storage complexity owing to FFT-based fast multiplication. FFC [CJM20] is based on Fourier spectral theory and enables models to have non-local receptive fields. Jae-Han Lee *et al.* [LHKK18] uses Fourier frequency domain analysis to estimate single-image depth. Fda [YS20] reduces differences between source and target distributions by exchanging low-frequency spectrum. GFNet [RZZ*21] improves the efficiency by using 2D FFT/IFFT to change the self-attention layer. LaMa [SLM*22] uses fast Fourier convolution to obtain the bigger receptive field. However, traditional Fourier transform in image processing directly employs the algorithm globally to learn a brute-force relationship between pixels, which may ignore the information effectiveness of different regions. The enhanced results rely on the frequency resolution. Therefore, these works apply the 2D-FFT/2D-IFFT operations in their modules, which are limited by strong hypothesis (signal is smooth), artificial priors, high complexities $O(HWC \log(HWC))$, and the ability to dynamically process various inputs. In contrast, the proposed WaveNet utilizes a learning-based method that decomposes pixels into different waves and outperforms these methods in complexities $O(HWC^2)$, accuracy, and efficiency, which are important for model effectiveness.

3. Method

In this section, we first introduce the recent works about periodic function applications in neural networks briefly. Secondly, we take an overview of WaveNet shown in Figure 2. Then, we discuss the details of our model. Finally, two main blocks of WaveNet are illustrated in detail.

3.1. Preliminaries

Over the past decades, there have been some investigations about periodic non-linearities applications. But so far, no research has

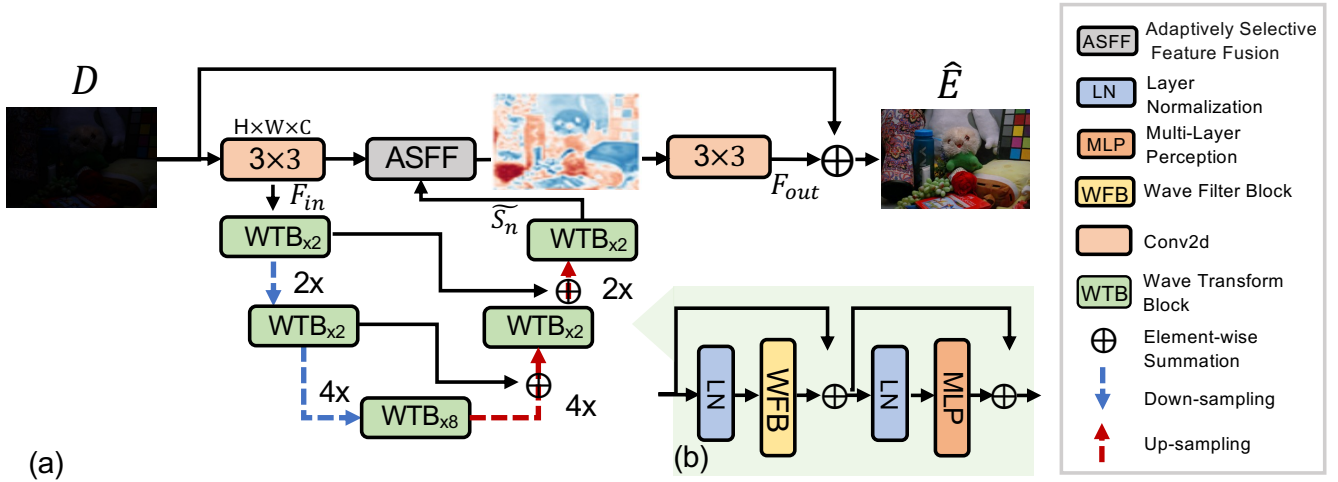


Figure 2: (a) shows the pipeline of the WaveNet. (b) is the Wave Transform Block.

shown that periodic activation functions can robustly outperform other activation functions. Early works used single-layer neural networks to mimic the Fourier transform [Gal88, ZUK*19]. Other works explore neural networks with periodic activations for simple classification tasks [WLC02, PHV16, SRA99] and recurrent neural networks [LZW15, KS97, CMCA97, AM97, SRA99]. Kloczek et al. [KMW*19] used cosine activation functions for image representation. Some works [LJ22, TSM*20] based on implicit neural representation (INR) scheme utilize the Fourier transform to the super-resolution task. However, these works directly introduce periodic functions into their works without convincing insights, nor do they show competitive performance. Unlike these works, we propose the *wave-like* feature representation obtaining satisfactory results on image enhancement.

3.2. Pipeline

The mainstream of WaveNet is shown in Figure 2(a). In Figure 2(a), for a degraded image $D \in \mathbb{R}^{H \times W \times 3}$, we first apply a convolutional layer to extract low-level features and expand channels $F_{in} \in \mathbb{R}^{H \times W \times C}$. Then, We use the Wave Transform Blocks (WTB) to transform F_{in} into the *signal-like* features $\tilde{S}_n \in \mathbb{R}^{H \times W \times C}$. n denotes the n -th WTB. We then add an Adaptively Selective Feature Fusion (ASFF) block to connect the shallow feature F_{in} and *signal-like* feature \tilde{S}_n . Using ASFF aims to enhance the feature presentation and weaken the impact of up-down sampling operations on image enhancement. Finally, we apply a convolutional layer to convert the *signal-like* features to a residual image $F_{out} \in \mathbb{R}^{H \times W \times 3}$. The enhanced image is obtained as \hat{E} . The overall process is summarized as:

$$\begin{aligned} F_{out} &= W^{out} H_{ASFF}(F_{in}, \tilde{S}_n), \\ \hat{E} &= D \oplus F_{out}, \end{aligned} \quad (1)$$

where \oplus denotes the element-wise summation, $H_{ASFF}(\ast)$ denotes the ASFF operation. W^{out} denotes the last convolutional layer with filter size 3×3 .

3.3. Wave Transform Block

As shown in Fig. 2(b), The WTB consists of two parts connected through skip-connection. The first part is a Wave Filter Block (WFB), which represents features in the wave form and allows modeling relationships between waves dynamically. The second part is a standard MLP layer to fuse channel information and enhance the transformation ability. The WTB calculation process can be computed as:

$$\begin{aligned} \tilde{S}'_n &= H_{WFB}(\text{LN}(\tilde{S}_{n-1})) \oplus \tilde{S}_{n-1}, \\ \tilde{S}_n &= W^{mlp}(\text{LN}(\tilde{S}'_n)) \oplus \tilde{S}'_n, \end{aligned} \quad (2)$$

where $\tilde{S}'_n \in \mathbb{R}^{H \times W \times C}$ is the output of the first part, $H_{WFB}(\ast)$ represents the WFB operation. $\text{LN}(\ast)$ stands for the Layer Normalization. W^{mlp} denotes the Multilayer Perceptron (MLP) with two FC layers and one PReLU layer.

3.4. Wave Filter Block

In this section, we discuss the key component of our WaveNet. First, we recall the Discrete Fourier Transform (DFT) which proposes non-periodic discrete functions within the specified interval can be split into combinations of periodic functions. Its 1D version can be derived by:

$$X[k] = \sum_{m=0}^{M-1} x[m] \left(\cos\left(\frac{2\pi}{M} km\right) - j \sin\left(\frac{2\pi}{M} km\right) \right), \quad (3)$$

where $x[m]$ is a sequence of M complex numbers, $X[k]$ indicates the spectrum at frequency $\frac{2\pi k}{M}$, and j represents the imaginary unit. It is clear that the spectrum at any frequency has global information in the frequency domain.

Drawing on the idea of 1D-DFT, we adaptively estimate the Fourier coefficients to decompose a pixel into linear combinations of three waves. We take an overview about the formula of *signal-like* maps \tilde{S}_n as follows:

$$\begin{aligned} \tilde{S}_n &= \text{Channel-FC}(\widetilde{CW}_n \oplus \widetilde{SW}_n \oplus \widetilde{GW}_n, W^{fc}) \\ &= W^{fc}(\widetilde{CW}_n \oplus \widetilde{SW}_n \oplus \widetilde{GW}_n), \end{aligned} \quad (4)$$

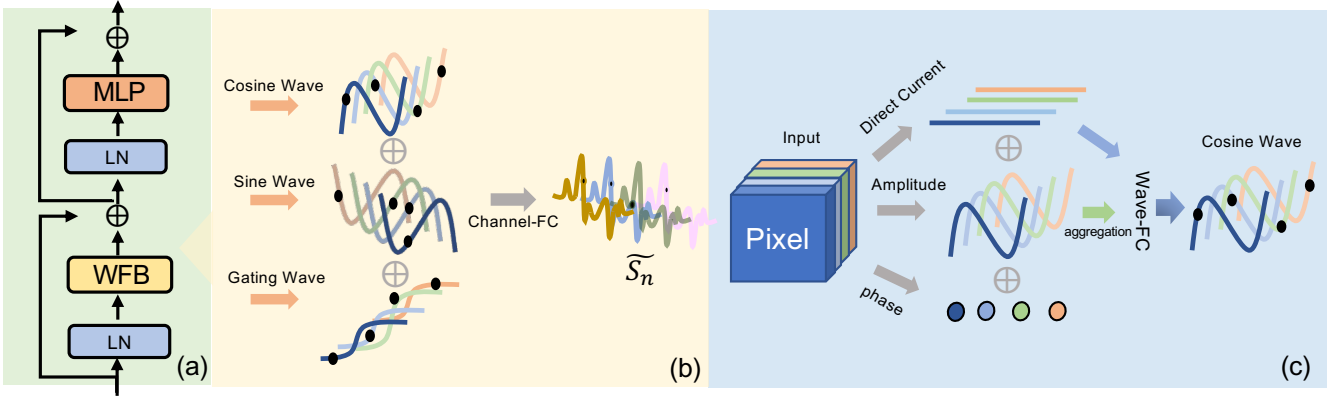


Figure 3: (a) Wave Transform Block. (b) is the schematic of constructing signal-like feature maps with three waves. (c) is the figure of constructing Cosine Wave feature maps.

where W^{fc} indicates the Channel-FC weights. As shown in Figure 3(b), it indicates that *signal-like* maps \tilde{S}_n includes three parts, Cosine Wave maps (\widetilde{CW}_n), Sine Wave maps (\widetilde{SW}_n), and Gating Wave maps (\widetilde{GW}_n). $\widetilde{CW}_n/\widetilde{SW}_n/\widetilde{GW}_n \in \mathbb{R}^{H \times W \times C}$.

1) wave-like representation: Traditional Fourier transform applied to image processing uses the coordinate as input for 2D-FFT calculation and remaps it into the frequency domain. It strongly relies on artificial priors. *e.g.* The frequency component of basis functions relies on the input resolution and scanning the whole picture. In contrast, we proposed the *wave-like* representation employing a neural network to transform each pixel into different components of waves for simplicity and efficiency. We represent the intermediate *wave-like* representation as follows:

$$\begin{aligned} \widetilde{CW}'_n &= A_n \otimes \cos(\theta_n), & \mathbf{I}_n^{\text{ap}} &= \{A_n, B_n\}, \\ \widetilde{SW}'_n &= B_n \otimes \sin(\alpha_n), & \mathbf{I}_n^{\text{ps}} &= \{\theta_n, \alpha_n\}, \\ \mathbf{I}_n^{\text{dc}} &= \{\bar{A}_n, \bar{B}_n\}, & \mathbf{I}_n^{\text{dc}}/\mathbf{I}_n^{\text{ap}}/\mathbf{I}_n^{\text{ps}} &\in \mathbb{R}^{H \times W \times C}, \end{aligned} \quad (5)$$

where \otimes denotes the element-wise product, \widetilde{CW}'_n and \widetilde{SW}'_n denote the intermediate waves without wave aggregation. $\mathbf{I}_n^{\text{dc}}/\mathbf{I}_n^{\text{ap}}/\mathbf{I}_n^{\text{ps}}$ represents the set including **DC** (\bar{A}_n, \bar{B}_n), **amplitude** (A_n, B_n) and **phase** (θ_n, α_n) for simple expression. The \bar{A}_n and \bar{B}_n denote the DC component which indicates the original information comes from the previous layer. A_n and B_n denote the amplitude term that is a real-value feature representing the content of each wave. θ_n and α_n denote the phase term that includes the current location of a wave and frequency information. We use the *wave-like* feature representation to organize in a structured way to extract deeper abstract regularities in the latent space. Due to the periodicity and differentiable invariance, SIREN [SMB*20] also indicates that using the periodic activation functions can speed up convergence, and periodicity in the latent space enables smooth interpolations and manipulations between different data points. This structured (*wave-like* features) representation encourages the model to learn and generate more coherent and realistic images.

2) Constructing *wave-like* feature maps : To get the *wave-like* maps in Eq. 5, the DC, amplitude and phase are required. Denote S_{n-1} containing j ($j = 1, 2, \dots, H \times W$) signals as $\tilde{S}_{n-1} = [\tilde{s}_{n-1}^1, \tilde{s}_{n-1}^2, \dots, \tilde{s}_{n-1}^j]$, where each signal \tilde{s}_{n-1}^j is a C-dimension

vector. For constructing DC/amplitude/phase components, we get them by a plain channel-FC operation, *i.e.*,

$$\mathbf{I}_n^{\text{dc/ap/ps}^j} = \text{Channel-FC}(\tilde{s}_{n-1}^j, W^{\text{dc/ap/ps}}), \quad (6)$$

where $W^{\text{dc/ap/ps}}$ is the weight with learnable parameters. There are different strategies for DC, amplitude and phase estimation. The most straightforward strategy represents these components with fixed parameters that can be learned during training. However, this way ignores the diversity of different input images. To dynamically capture the particular attributes according to the input feature, we adopt the Channel-FC in Eq. 6 to capture the particular attributes for simplicity and model performance. There are other constructing methods whose impact on the model performance is empirically investigated in the ablation study.

3) Aggregating *wave-like* features: The intermediate waves limit the feature structures in the latent space, which concentrates on modeling more general and regular features. However, as we mentioned in Sec. 1, aggregating waves in local and no-local to provide various semantic information is also essential for *signal-like* feature modulation. The static vector-sum aggregation [FLS11] is popularly employed to calculate wave superposition mode, but it may cause high FLOPs and slow the model speed according. Besides, this static calculation method may not be able to cope with diverse inputs. As the basic operation in CNNs, convolution provides local connectivity and translation equivariance. These properties bring efficiency and generalization to dynamically aggregate waves. To modulate the spatial interactions between different waves, The proposed WaveNet uses different sizes of convolution kernels to gather *wave-like* features in different sizes of regions dynamically, which is also taken as a learnable linear combination of different waves to fit different local features. Thus, the *wave-like* dynamical aggregation can be formulated as follows:

$$\begin{aligned} \widetilde{CW}_n^{\text{agg}} &= W^{\text{cw}}(\widetilde{CW}'_n), \\ \widetilde{SW}_n^{\text{agg}} &= W^{\text{sw}}(\widetilde{SW}'_n), \end{aligned} \quad (7)$$

where W^{cw} and W^{sw} are both learnable convolutional weights. $\widetilde{CW}_n^{\text{agg}}$ and $\widetilde{SW}_n^{\text{agg}}$ are the superposition waves. In addition, we use

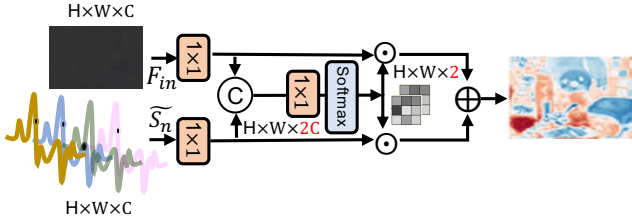


Figure 4: Adaptively Selective Feature Fusion

Wave-FC to adaptively linearly combine the basis waves in Eq. 7 to adaptively assign weights to waves and enhance the representation capacity. Like the coefficients in Eq. 3. *i.e.*,

$$\begin{aligned} \widetilde{CW}_n / \widetilde{SW}_n &= \text{Wave-FC}(\mathbf{I}_n^{\text{agg}}, \mathbf{I}_n^{\text{dc}}, \Theta^{\text{agg}}, \Theta^{\text{dc}}) \\ &= \Theta^{\text{dc}} \mathbf{I}_n^{\text{dc}} \oplus \Theta^{\text{agg}} \mathbf{I}_n^{\text{agg}}, \quad \mathbf{I}_n^{\text{agg}} = \{\widetilde{CW}_n^{\text{agg}}, \widetilde{SW}_n^{\text{agg}}\}, \end{aligned} \quad (8)$$

where $\Theta^{\text{dc}}/\Theta^{\text{agg}}$ is the learnable coefficient map. $\mathbf{I}_n^{\text{agg}} \in \mathbb{R}^{H \times W \times C}$ is the wave aggregation features. As shown in Figure 3(c), we obtain the final *wave-like* features after Wave-FC operation.

4) Gating Wave: Since the information of CW and SW is periodic, we design the Gating Wave (GW) for handling some non-periodic features. We use the gating mechanism which is activated with Tanh non-linearity to enhance the feature richness of the network. The \widetilde{GW}_n is formulated as:

$$\begin{aligned} \text{Gating}(\widetilde{S}_{n-1}) &= \diamond(\widetilde{S}_{n-1}) \otimes W^{\text{gw}} \tanh(\diamond(\widetilde{S}_{n-1})), \\ \widetilde{GW}_n &= \text{Wave-FC}(\diamond(\text{Gating}(\widetilde{S}_{n-1})), \widetilde{S}_{n-1}, \Theta^{\text{gw}}, \Theta^{\text{dc}}), \end{aligned} \quad (9)$$

where \diamond denotes the Channel-FC in Eq. 4 for simple expression and Θ^{gw} is the learnable weight to aggregate local information. The \widetilde{GW}_n retains valid information from the previous layer and provides complementary features to the next layer. The final \widetilde{S}_n is obtained by Eq. 4. The whole WTB operation indicate that the output \widetilde{S}_{n-1} from the previous WTB will be modulated by the next WTB.

3.5. Adaptively Selective Feature Fusion

The U-shape methods are proposed to balance the accuracy and computational cost. (which is common in low-level vision tasks because of the high-resolution inputs) However, the U-shape models may lose detail information after many up-down sampling operations. Thus, we design the ASFF to adaptively fuse shallow features and *signal-like* features. As shown in Figure 4, We use Channel-FC to compress the *signal-like* maps contents and shallow features, and capture the intra-channel interactions. Furthermore, we employ a Softmax to generate the intra-channel attention maps. Finally, we blend the two branch outputs with attention maps and use the element-wise product to adaptively build the relationship between shallow two kinds of features. In Figure 2 we can see that the output from ASFF goes to a 3×3 convolutional layer to transform the *signal-like* linear-projection hybrid feature maps into the final residual image F_{out} . Overall, benefitting from the proposed adaptive *wave-like* representation, our WaveNet achieves high performance with acceptable computation-consuming.

3.6. Loss Function

Given an output image \hat{E} and a ground truth image E , we use a signal-based loss function to guide our *signal-like* feature repre-

sentations. Specifically, the loss function L for WaveNet consists of PSNR loss L_{psnr} , MS-SSIM loss L_{ssim} and Edge loss L_{edge} . *i.e.*

$$\begin{aligned} L &= \lambda_1 \cdot L_{\text{psnr}}(\hat{E}, E) + \lambda_2 \cdot L_{\text{ssim}}(\hat{E}, E) \\ &\quad + \lambda_3 \cdot L_{\text{edge}}(\hat{E}, E), \end{aligned} \quad (10)$$

where λ_1 , λ_2 and λ_3 are the weight coefficients used to make trade-off for the loss function L from our experiments.

4. Experiment

In this section, we test our WaveNet on the popular real-world benchmark datasets for image enhancement and dark face detection. Various ablation studies provided in **supplemental material** demonstrate performance and effectiveness of *wave-like* modeling.

4.1. Dataset and Experimental Setup

The proposed WaveNet is tested on six datasets including five low/normal real-captured datasets and one high-level application dataset. The datasets for image enhancement include LOL [WWYL18], VE-LOL [LXY*21], SICE [CGZ18], SID [CCXK18] and MIT-Adobe FiveK [BPCD11]. DARK FACE [YYR*19] is composed of various scenes with faces taken in the dark. *Note that we use 'expert C' as FiveK ground truth and the Sony subset following the script provided by SID [CCXK18]* to transfer the low-light images from RAW to RGB for our training and testing.

Implementation Details: WaveNet is end-to-end trainable and requires no pretraining on large datasets. For data augmentation, we trained the network using random horizontal-vertical flips, rotation, and MixUp [ZCDLP17]. We set the Adam optimizer [KB14] with an initial learning rate of 1×10^{-4} , which is steadily decreased to 10^{-6} with the cosine annealing decay [LH16]. The values of $\lambda_1/\lambda_2/\lambda_3$ in Eq. 10 are 0.33/0.33/0.1. The proposed WaveNet is trained on a single NVIDIA RTX 3090 using the PyTorch.

By varying the width and depth of the model, we build 3 models with different parameters and computational costs, denoted as WaveNet-T, WaveNet-S, and WaveNet-B sequentially.

Metrics: To evaluate the performance of WaveNet, we adopt Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the main evaluation metrics, which are classic to measure the difference between output and ground truth.

4.2. Quantitative Evaluation

Note that we obtain these results in Table 1 either from the respective papers or by running the respective public code and full-resolution testing. **Low-light Image Enhancement:** LOL [WWYL18] and SID [CCXK18] include extremely dark images with lots of noise. It is challenging to complete the low-light image enhancement task on these datasets. We tested our methods on this dataset, and the results are shown in Table 1. From Table 1, the proposed WaveNet-B achieves **0.44/0.88** dB gain in PSNR over the previous best model LFFlow [WYY*22] and Restormer [ZAK*22] on LOL/SID. WaveNet-T is a lightweight model. It shows promising performance of quality and efficiency. Our WaveNet-T gains better (**23.59/22.57**dB) result with only **80k** parameters compared to the currently popular lightweight method IAT [CLG*22] and image restoration model MAXIM [TTZ*22]. WaveNet-S provides a balance of accuracy and efficiency, which

Table 1: Quantitative comparison on the LOL, MIT-Adobe FiveK, SID, VE-LOL and SICE in terms of Params, FLOPs, PSNR and SSIM. The best and second-best results of the evaluated methods are **highlighted** and underlined.

Method	Efficiency		LOL		FiveK		SID(Real)		VE-LOL(Real)		SICE	
	Params (M) ↓	FLOPs (G) ↓	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
URetinexNet [WWZ*22]	0.34	57	21.32	0.834	23.51	0.826	19.24	0.588	21.09	0.858	13.57	0.621
LLFlow [WWY*22]	17.42	287	<u>24.99</u>	0.868	-	-	-	-	24.15	<u>0.895</u>	<u>17.78</u>	0.742
Uformer [WCB*22]	5.29	12	<u>16.36</u>	0.507	23.89	0.906	18.54	0.577	18.82	0.771	15.12	0.577
SID [CCXK18]	7.76	13.73	14.35	0.436	16.77	0.589	16.97	0.591	13.24	0.442	11.23	0.387
RetinexNet [WWYL18]	0.84	587.4	16.77	0.562	12.31	0.671	16.48	0.578	15.47	0.567	15.9	0.705
Restormer [ZAK*22]	26.11	87.7	23.45	0.83	24.52	0.926	22.34	0.638	25.16	0.882	14.93	0.749
MIRNet [ZAK*20]	31.79	785	24.14	0.83	23.73	0.925	20.56	0.611	20.08	0.82	16.23	0.763
IAT [CLG*22]	<u>0.09</u>	1.5	23.38	0.809	-	-	16.32	0.565	23.50	0.824	15.23	0.525
MBLLEN [LLWL18]	20.47	20	17.9	0.702	19.78	0.825	20.56	0.567	18.01	0.715	15.69	0.658
MAXIM [TTZ*22]	14.14	216	23.43	0.863	24.64	0.913	22.13	0.596	22.86	0.818	16.54	0.782
KinD++ [ZGM*21]	8.28	692	21.3	0.822	19.83	0.784	17.89	0.572	15.63	0.699	15.46	0.678
KinD [ZZG19b]	8.02	35	20.87	0.79	14.54	0.741	18.02	0.583	14.74	0.641	14.52	0.534
IPT [CWG*21]	115.63	1,188	16.27	0.504	-	-	20.68	0.566	19.80	0.813	14.53	0.561
DRBN [YWF*20]	0.58	38	19.55	0.746	13.35	0.378	19.02	0.577	20.13	0.82	13.43	0.536
DeepUPE [WZF*19b]	1.02	21.1	14.38	0.446	23.04	0.893	17.01	0.604	13.27	0.452	10.97	0.355
RF [KY20]	21.54	46.23	15.23	0.452	19.21	0.652	16.44	0.596	14.05	0.458	11.42	0.391
DeepLPF [MMM*20]	1.77	5.86	15.28	0.473	20.01	0.661	16.02	0.587	14.10	0.48	11.68	0.378
Sparse [YWH*21]	2.33	53.26	17.2	0.64	21.2	0.756	18.68	0.606	20.06	0.816	15.21	0.617
FIDE [XYL20]	8.62	28.51	18.27	0.665	22.57	0.831	18.34	0.578	16.85	0.678	13.79	0.541
3D-LUT [ZCL*20]	0.6	7.7	16.35	0.585	25.21	0.922	20.11	0.592	17.59	0.721	15.69	0.733
WaveNet-T(Ours)	0.08	<u>4.49</u>	23.59	0.839	24.62	0.933	22.57	0.673	25.64	0.876	17.38	0.784
WaveNet-S(Ours)	1.4	82.5	24.54	0.856	<u>25.30</u>	<u>0.927</u>	<u>22.98</u>	<u>0.702</u>	<u>26.86</u>	<u>0.872</u>	17.69	<u>0.785</u>
WaveNet-B(Ours)	14.4	162	25.44	<u>0.864</u>	25.34	0.924	23.15	0.726	27.21	0.899	18.86	0.795

Table 2: Quantitative comparison of mAP of face detection in the dark.

Method	DSFD [LWW*19]	+MAXIM [TTZ*22]	+LLFlow [WWY*22]	+Restormer [ZAK*22]	+IAT [CLG*22]	+WaveNet(Ours)
mAP	0.163	<u>0.228</u>	0.143	0.212	0.222	0.238

achieves competitive results on all datasets.

Image Enhancement: MIT-Adobe FiveK [BPCD11] is a real-world image enhancement dataset containing a wide variety of scenes for training and testing. As shown in Table 1, our WaveNet achieves the best and second-best scores and surpasses all the baselines. Our WaveNet-B achieves **0.70dB** gain in PSNR at most compared to the MAXIM [TTZ*22].

Cross-dataset Evaluation: To evaluate our method’s generality and effectiveness in real no-reference degraded images, we take the cross-dataset validation method. *e.g.* Our method is trained on the LOL dataset, and the model is directly tested on the testing set of VE-LOL (Real Part) and SICE. From the results in Table 1 we can see that our method significantly outperforms other trained methods in all metrics and provides a **2.05/1.08dB** promotion on VE-LOL/SICE compared with previous SOTA methods. From another perspective, our WaveNet can provide a more stable and general result without retraining.

Efficiency Analysis: It is worth noting that efficiency term are tested on a 256×256 image. Compared with IAT [CLG*22], which is also a lightweight model, WaveNet-T with 12% fewer parameters than IAT [CLG*22] achieves better PSNR and SSIM scores on all datasets. Compared with the previous SOTA LLFlow [WWY*22], WaveNet-T has **220**× fewer parameters and **64**× fewer FLOPs but achieves its **94%** accuracy on LOL [WWYL18] dataset. Compared to Restormer [ZAK*22], whose FLOPs is close to our WaveNet-S, our WaveNet-S has **18.6**× fewer parameters and faster than

it. Besides, Our WaveNet-S provides better generality in cross-dataset validation. With acceptable computational costs, WaveNet-B achieves the best results on all datasets in Table 1. Overall, our WaveNet achieves the best trade-off on accuracy and efficiency. The superiority of WaveNet implies that the proposed module WTb has a large potential and modulating the feature representation.

4.3. Qualitative Evaluation

The qualitative comparisons of images are given in Figure 5. The higher PSNR and SSIM values in Table 1 indicate that our methods can better restore the color and better reserve details. From the Figure 5 we can see that our WaveNet keeps more fine details, more natural color consistency, higher contrast, and less noise compared with other methods. Specifically, it can be observed in cross-dataset validation visual results (4th and 5th row) in Figure 5 that Our WaveNet not only suppresses overexposure in regions of high brightness but also enhances natural colors and details in regions of shadows. Hence, compared with other methods, the proposed WaveNet use waveform feature representations achieving excellent performance in generality and qualitative evaluation.

4.4. High-level Vision Evaluation:

To validate the performance of our WaveNet in high-level vision tasks, we use the DARK FACE dataset [YYR*19] composed of images with faces taken in the dark. The Dual Shot Face Detector (DSFD) [LWW*19] trained on the WIDER FACE dataset

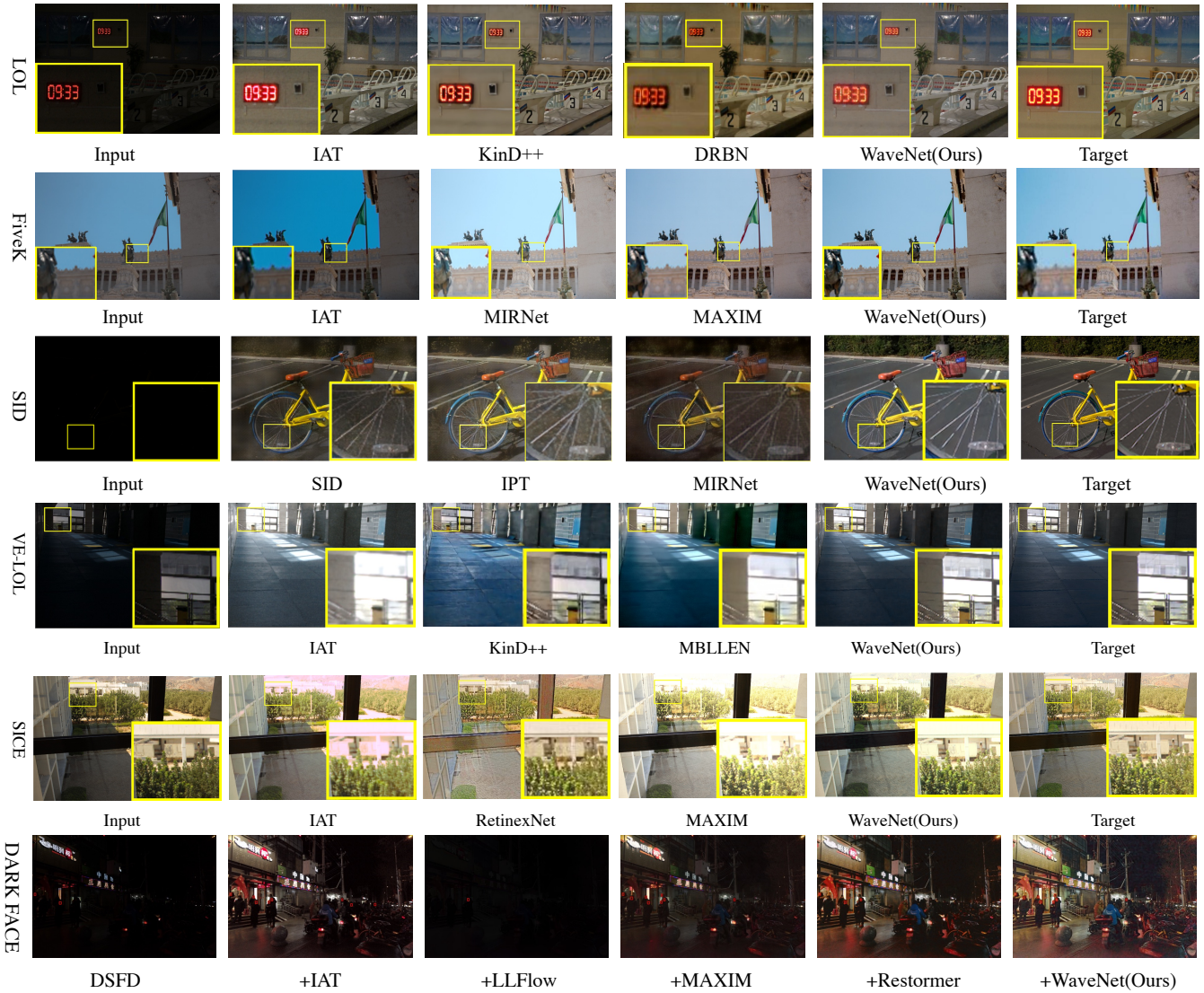


Figure 5: Visual comparison with the state-of-the-art methods on six datasets. (Zoom in for better view)

[YLLT16] is used as the face detector. We take the last 500 images of DARK FACE [YYR*19] training set enhanced by different methods and feed them to the DSFD [LWW*19]. The results shown in Table 2 validate that our WaveNet gains the best score. The bottom row in Figure 5 shows that DSFD detected the most faces using the WaveNet-enhanced images.

5. Conclusion

This paper proposes an architecture for image enhancement tasks, dubbed WaveNet, which aims to extract more spatial details, better color accuracy, and a higher contrast from the original degraded images. Inspired by signal processing, we formulate the enhancement task as a signal modulation problem. We use a learning-based method to adaptively construct and modulate the signals come from digital images. Specifically, we utilize the Fourier transform theory to decompose a pixel into three wave functions (Cosine Wave (CW), Sine Wave (SW), Gating Wave (GW)) for feature charac-

terization. The amplitude and phase are essential for constructing *wave-like* features. Amplitude is the original real-valued feature, and phase modulates the relationship between the varying inputs and fixed weights in WaveNet. Specifically, we propose the Wave Transform Block (WTB) to dynamically construct the *wave-like* feature maps and modulate the wave interactions (wave superposition) locally and non-locally. Furthermore, WaveNet enhances various inputs by adaptively adjusting the amplitude and phase of each wave in the constructed *signal-like* feature maps. The extensive experiments on the benchmark datasets validate the effectiveness of our WaveNet for image enhancement.

Acknowledgement. This work was supported by Science and Technology Service Network Initiative (KFJ-ST-S-QYZD-2021-21-001), Sichuan Science and Technology Program (2019ZDZX0006, 2020YFQ0056), and the Talents by Sichuan provincial Party Committee Organization Department.

References

- [AM97] ALQUÉZAR MANCHO R.: Symbolic and connectionist learning techniques for grammatical inference, 1997. 3
- [BPCD11] BYCHKOVSKY V., PARIS S., CHAN E., DURAND F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011* (2011), IEEE, pp. 97–104. 2, 5, 6
- [CCXK18] CHEN C., CHEN Q., XU J., KOLTUN V.: Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018). 2, 5, 6
- [CGZ18] CAI J., GU S., ZHANG L.: Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing* 27, 4 (2018), 2049–2062. 2, 5
- [CJM20] CHI L., JIANG B., MU Y.: Fast fourier convolution. *Advances in Neural Information Processing Systems* 33 (2020), 4479–4488. 2
- [CLG*22] CUI Z., LI K., GU L., SU S., GAO P., JIANG Z., QIAO Y., HARADA T.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *BMVC* (2022), p. 238. 2, 5, 6
- [CMCA97] CHOUËKI M. H., MOUNT-CAMPBELL C. A., AHALT S. C.: Implementing a weighted least squares procedure in training a neural network to solve the short-term load forecasting problem. *IEEE Transactions on Power systems* 12, 4 (1997), 1689–1694. 3
- [CWG*21] CHEN H., WANG Y., GUO T., XU C., DENG Y., LIU Z., MA S., XU C., XU C., GAO W.: Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 12299–12310. 1, 6
- [DLW*17] DING C., LIAO S., WANG Y., LI Z., LIU N., ZHUO Y., WANG C., QIAN X., BAI Y., YUAN G., ET AL.: Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices. 395–408. 2
- [FLS11] FEYNMAN R. P., LEIGHTON R. B., SANDS M.: *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, vol. 1. Basic books, 2011. 4
- [Gal88] GALLANT: There exists a neural network that does not make avoidable mistakes. 657–664. 3
- [KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv.org* (dec 2014). [arXiv:1412.6980v9](https://arxiv.org/abs/1412.6980v9). 5
- [KMW*19] KLOCEK S., MAZIARKA Ł., WOŁCZYK M., TABOR J., NOWAK J., ŚMIEJA M.: Hypernetwork functional image representation. 496–510. 3
- [KS97] KOPLON R., SONTAG E. D.: Using fourier-neural recurrent networks to fit sequential input/output data. *Neurocomputing* 15, 3-4 (1997), 225–248. 3
- [KY20] KOSUGI S., YAMASAKI T.: Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In *AAAI* (2020). 6
- [LAS17] LORE K. G., AKINTAYO A., SARKAR S.: Lnet: A deep auto-encoder approach to natural low-light image enhancement. *Pattern Recognition* 61 (2017), 650–662. 2
- [LFH21] LI J., FENG X., HUA Z.: Low-light image enhancement via progressive-recursive network. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 11 (2021), 4227–4240. 2
- [LH16] LOSHCHEV I., HUTTER F.: Sgdr: Stochastic gradient descent with warm restarts. *arXiv.org* (aug 2016). [arXiv:1608.03983v5](https://arxiv.org/abs/1608.03983v5). 5
- [LHKK18] LEE J.-H., HEO M., KIM K.-R., KIM C.-S.: Single-image depth estimation based on fourier domain analysis. 330–339. 2
- [LJ22] LEE J., JIN K. H.: Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 1929–1938. 3
- [LLF*20] LI J., LI J., FANG F., LI F., ZHANG G.: Luminance-aware pyramid network for low-light image enhancement. *IEEE Transactions on Multimedia* 23 (2020), 3153–3165. 2
- [LLWL18] LV F., LU F., WU J., LIM C.: Mblen: Low-light image/video enhancement using cnns. In *BMVC* (2018), vol. 220, p. 4. 6
- [LWW*19] LI J., WANG Y., WANG C., TAI Y., QIAN J., YANG J., WANG C., LI J., HUANG F.: Dsfed: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5060–5069. 2, 6, 7
- [LXY*21] LIU J., XU D., YANG W., FAN M., HUANG H.: Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision* 129 (2021), 1153–1184. 2, 5
- [LZW15] LIU P., ZENG Z., WANG J.: Multistability of recurrent neural networks with nonmonotonic activation functions and mixed time delays. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46, 4 (2015), 512–523. 3
- [MMM*20] MORAN S., MARZA P., MCDONAGH S., PARISOT S., SLABAUGH G.: Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 6
- [PHV16] PARASCANDOLO G., HUTTUNEN H., VIRTANEN T.: Taming the waves: sine as activation function in deep neural networks, 2016. 3
- [RZZ*21] RAO Y., ZHAO W., ZHU Z., LU J., ZHOU J.: Global filter networks for image classification. *Advances in neural information processing systems* 34 (2021), 980–993. 2
- [SLM*22] SUVOROV R., LOGACHEVA E., MASHIKHIN A., REMIZOVA A., ASHUKHA A., SILVESTROV A., KONG N., GOKA H., PARK K., LEMPITSKY V.: Resolution-robust large mask inpainting with fourier convolutions. 2149–2159. 2
- [SMB*20] SITZMANN V., MARTEL J., BERGMAN A., LINDELL D., WETZSTEIN G.: Implicit neural representations with periodic activation functions. *Advances in neural information processing systems* 33 (2020), 7462–7473. 4
- [SRA99] SOPENA J. M., ROMERO E., ALQUEZAR R.: Neural networks with periodic and monotonic activation functions: a comparative study in classification problems. 3
- [TSM*20] TANCİK M., SRINIVASAN P., MILDENHALL B., FRIDOVICH-KEIL S., RAGHAVAN N., SINGHAL U., RAMAMOORTHI R., BARRON J., NG R.: Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems* 33 (2020), 7537–7547. 3
- [TTZ*22] TU Z., TALEBI H., ZHANG H., YANG F., MILANFAR P., BOVIK A., LI Y.: Maxim: Multi-axis mlp for image processing. 5769–5780. 1, 2, 5, 6
- [WCB*22] WANG Z., CUN X., BAO J., ZHOU W., LIU J., LI H.: Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 17683–17693. 6
- [WLC02] WONG K.-W., LEUNG C.-S., CHANG S.-J.: Handwritten digit recognition using multilayer feedforward neural networks with periodic and monotonic activation functions. In *2002 International Conference on Pattern Recognition* (2002), vol. 3, IEEE, pp. 106–109. 3
- [WLSL20] WANG L.-W., LIU Z.-S., SIU W.-C., LUN D. P.: Lightning network for low-light image enhancement. *IEEE Transactions on Image Processing* 29 (2020), 7984–7996. 2
- [WWY*22] WANG Y., WAN R., YANG W., LI H., CHAU L.-P., KOT A.: Low-light image enhancement with normalizing flow. 2604–2612. 2, 5, 6
- [WWYL18] WEI C., WANG W., YANG W., LIU J.: Deep retinex decomposition for low-light enhancement. *arXiv.org* (aug 2018). [arXiv:1808.04560v1](https://arxiv.org/abs/1808.04560v1). 2, 5, 6
- [WWZ*22] WU W., WENG J., ZHANG P., WANG X., YANG W., JIANG J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5901–5910. 6

- [WZF*19a] WANG R., ZHANG Q., FU C.-W., SHEN X., ZHENG W.-S., JIA J.: Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 6849–6857. 2
- [WZF*19b] WANG R., ZHANG Q., FU C.-W., SHEN X., ZHENG W.-S., JIA J.: Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 6849–6857. 6
- [XYYL20] XU K., YANG X., YIN B., LAU R. W.: Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 2281–2290. 6
- [YLLT16] YANG S., LUO P., LOY C.-C., TANG X.: Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 5525–5533. 7
- [YS20] YANG Y., SOATTO S.: Fda: Fourier domain adaptation for semantic segmentation. *arXiv.org* (apr 2020). [arXiv:2004.05498v1](https://arxiv.org/abs/2004.05498v1). 2
- [YWF*20] YANG W., WANG S., FANG Y., WANG Y., LIU J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 3063–3072. 6
- [YWH*21] YANG W., WANG W., HUANG H., WANG S., LIU J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing* 30 (2021), 2072–2086. 6
- [YYR*19] YUAN Y., YANG W., REN W., LIU J., SCHEIRER W. J., WANG Z.: Ug2+ track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments. *arXiv.org* (apr 2019). [arXiv:1904.04474v4](https://arxiv.org/abs/1904.04474v4). 2, 5, 6, 7
- [ZAK*20] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H., SHAO L.: Learning enriched features for real image restoration and enhancement. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16* (2020), Springer, pp. 492–511. 1, 2, 6
- [ZAK*22] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 5728–5739. 5, 6
- [ZCDLP17] ZHANG H., CISSE M., DAUPHIN Y. N., LOPEZ-PAZ D.: mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017). 5
- [ZCL*20] ZENG H., CAI J., LI L., CAO Z., ZHANG L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2058–2073. 6
- [ZGM*21] ZHANG Y., GUO X., MA J., LIU W., ZHANG J.: Beyond brightening low-light images. *International Journal of Computer Vision* 129 (2021), 1013–1037. 1, 6
- [ZUK*19] ZHUMEKENOV A., UTEULIYEVA M., KABDOLOV O., TAKHANOV R., ASSYLBEKOV Z., CASTRO A. J.: Fourier neural networks: A comparative study. *arXiv.org* (feb 2019). [arXiv:1902.03011v1](https://arxiv.org/abs/1902.03011v1). 3
- [ZZG19a] ZHANG Y., ZHANG J., GUO X.: Kindling the darkness: A practical low-light image enhancer. *arXiv.org* (may 2019). [arXiv:1905.04161v1](https://arxiv.org/abs/1905.04161v1). 2
- [ZZG19b] ZHANG Y., ZHANG J., GUO X.: Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia* (2019), pp. 1632–1640. 6