CutFreq: Cut-and-Swap Frequency Components for Low-Level Vision Augmentation

Hongyang Chen[1], **Kaisheng Ma**[2*]
[1]Xi'an Jiaotong University, Xi'an, China
[2]Tsinghua University, Beijing, China
chenhy@stu.xjtu.edu.cn, kaisheng@mail.tsinghua.edu.cn

Abstract

Low-level vision plays a crucial role in a wide range of imaging quality and image recognition applications. However, the limited size, quality, and diversity of datasets often pose significant challenges for low-level tasks. Data augmentation is the most effective and practical way of sample expansion, but the commonly used augmentation methods in high-level tasks have limited improvement in the low-level due to the boundary effects or the non-realistic context information. In this paper, we propose the Cut-and-Swap Frequency Components (*CutFreq*) method for low-level vision, which aims to preserve high-level representations with directionality and improve image synthesis quality. Observing the significant frequency domain differences between reconstructed images and real ones, in CutFreq, we propose to transform the input and real images separately in the frequency domain, then define two stages for the model training process, and finally swap the specified frequency bands respectively and inversely transform to generate augmented samples. The experimental results show the superior performance of CutFreq on five low-level vision tasks. Moreover, we demonstrate the effectiveness of CutFreq in the low-data regime. Code is available at https://github.com/DreamerCCC/CutFreq.

Introduction

Labeling data is a time-consuming and labor-intensive task, especially in low-level vision tasks, which have the characteristics of high overhead, low savings, and poor adaptability (Devillers, Vidrascu, and Lamel 2005). Whereas data augmentation (Yun et al. 2019; Zhang et al. 2017; Li et al. 2021; Yoo, Ahn, and Sohn 2020; Han et al. 2022) are techniques used to increase the amount of data by adding slightly modified copies of existing data or synthetic data newly created from existing data. In mainstream tasks (Trabucco et al. 2023; Liang, Liang, and Jia 2023; Zou et al. 2023), the augmented samples are always strongly correlated with the original samples. At the same time, data augmentation can bring regularization effects and reduce the structural risk of the model (Yun et al. 2019). In a certain way, the model is more immersed in observing the general patterns in the dataset, and some data irrelevant to the patterns are suppressed. Recent approaches (DeVries and Taylor 2017; Yoo, Ahn, and Sohn 2020) have focused on altering certain regions in an image to ensure that deep neural networks (DNNs) can emphasize more discriminative regions used to generate predictions. However, these methods are not always applicable to low-level vision, such as image super-resolution (Chen et al. 2022) and image restoration (*e.g.*, image deblurring (Tao et al. 2018), image dehazing (Engin, Genç, and Kemal Ekenel 2018)). Extensive experiments have shown that training with augmentation with warped spatial properties results in poor algorithm performance compared to simple data augmentation methods such as flipping and rotating (Yoo, Ahn, and Sohn 2020; Han et al. 2022). As of now, few data augmentation methods can be used universally for low-level vision. From the perspective of low-level vision, for data augmentation methods to increase samples and improve the model's generalization

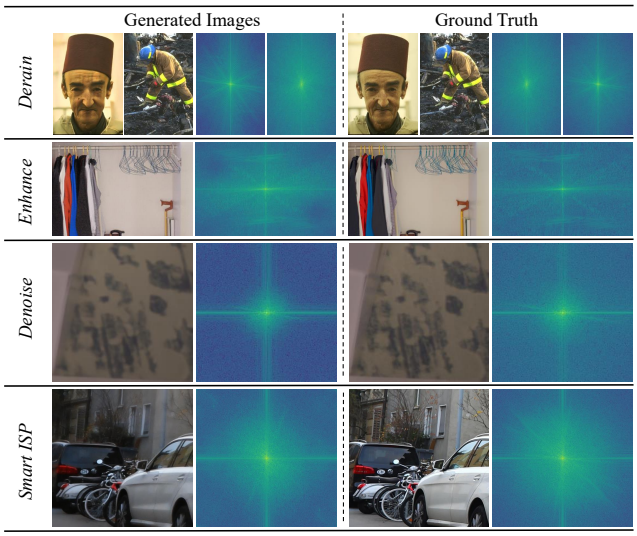


Figure 1: *Frequency domain gaps between the generated and the real images* by the state-of-the-art models in low-level vision tasks. Row 1: Restormer (Zamir et al. 2021) on Rain100H (Yang et al. 2017). Row 2: Restormer on LoL (Wei et al. 2018). Row 3: Restormer on SIDD (Abdelhamed, Lin, and Brown 2018). Row 4: LW-ISP (Chen and Ma 2022) on Zurich (Ignatov, Van Gool, and Timofte 2020).

* Corresponding author.

ability through priors, on the one hand, such priors should avoid introducing dissonances, such as additional blurring and color blocks. On the other hand, it can bring meaningful supervision to the model (Lee et al. 2021; Chen et al. 2021a), that is, directional changes in the high-dimensional feature domain of the image or other representations. To design a more efficient augmentation method, we observe two key points.

Observation 1: The successful application of instance normalization (IN) (Ulyanov, Vedaldi, and Lempitsky 2016) confirms the potential of instance-independent augmentation. This stands in contrast to the classic batch normalization (BN) (Ioffe and Szegedy 2015), which normalizes data within a mini-batch and is integral to recognition tasks such as image classification and object detection. IN, on the other hand, normalizes data exclusively across the height and width dimensions (H$\times$W) and is independent of other data instances, thus offering a distinct advantage for image generation tasks.

Observation 2: A significant frequency-domain gap still exists between the reconstruction results of the current state-of-the-art (SOTA) models (Chen and Ma 2022; Zamir et al. 2021) and the real images. We demonstrate that the difference between the reconstructed image of the SOTA models of different tasks of low-level vision and the real image in the frequency domain is still discernible (as shown in Figure 1), and whether it is high-frequency or low-frequency components. It is reflected in different network structures, including CNN and transformer.

From the above observations, we suppose that multi-instance augmentation strategies, such as CutMix (Yun et al. 2019), Mixup (Zhang et al. 2017), and MoEx (Li et al. 2021), may not be the most suitable for low-level vision tasks. Likewise, considering avoiding the introduction of discordant pixels or features, we focus on single-instance fine-grained augmentation of the input sample. In this paper, we introduce Cut-and-Swap Frequency Components (*CutFreq*) in low-level vision, a general data augmentation method that provides directional information for model convergence and representation learning. In order to guide the model in learning the required frequency domain component information during the training process, we start by presenting and summarizing the convergence behavior of the model on the frequency domain components for different low-level vision tasks. Based on this analysis, we categorize these tasks into two groups: those that exhibit sustained convergence of high-frequency components and those that exhibit fast convergence of high-frequency components (as shown in Figure 4). CutFreq adopts the standard discrete wavelet transform (Edwards 1991) to decompose the input and the real image into frequency domain representation (LL, HL, LL, and HH bands) to complete the "*Cut*" operation. Inspired by the frequency domain component convergence experiment, CutFreq augments the input in two training stages to complete the "*Swap*" operation: (a) Fast convergence. Swap the specified frequency components of the real image with the input image so that the model can focus on learning other components. (b) Performance approximation. The different frequency components of the real image are exchanged with the input image respectively to realize the biased learning of the input image. To the best of our knowledge, CutFreq is the first augmentation method that can be applied to different tasks in low-level vision.

Abundant experiments on five different low-level vision tasks demonstrate the effectiveness of our method both quantitatively and qualitatively. Compared to mainstream augmentation methods (Cutout (DeVries and Taylor 2017), Cutmix (Yun et al. 2019), Mixup (Zhang et al. 2017), *etc.*) and existing methods in low-level vision (Cutblur (Yoo, Ahn, and Sohn 2020) and Copy-blend (Shyam et al. 2021)), CutFreq achieves the best performance. In the smart ISP task, our method brings about 0.17 dB gain. CutFreq achieves new state-of-the-art results in image denoising, deraining, deblurring, and enhancement tasks. Moreover, we validate the potential of our method in the low-data regime.

To sum up, our contribution is three-fold:

1. We demonstrate the prevalence of frequency-domain gaps in low-level vision (both qualitatively and quantitatively) and use frequency-band normalized $\mathcal{L}_1$ distance to measure model convergence for different frequency components.
2. We design the CutFreq augmentation method to guide the model to learn the required frequency domain component information during the training process.
3. Quantitative and qualitative experiment results on five low-level vision tasks demonstrate the effectiveness of our method. Furthermore, We investigate the superiority of CutFreq in the low-data regime.

Related Work

High-level Augmentation

Data augmentation (DA) is an effective strategy for training neural networks. In recognition tasks, DA can increase label information (such as Mixup (Zhang et al. 2017), CutMix (Yun et al. 2019)) or force the model to learn some local details (such as Cutout (DeVries and Taylor 2017)). PA-AUG (Choi, Song, and Kwak 2020) exploits the delicate structure information of the point cloud, divides the objects in the point cloud into multiple parts, and then randomly applies five existing augmentation methods. TransMix (Chen et al. 2021a) is oriented toward the transformer structure and mixes the labels according to the attention map of the vision transformer, thereby reducing the gap between the input space and the label space caused by Mixup. MoEx (Li et al. 2021) adopts PONO normalization for feature enhancement between different samples and is orthogonal to mainstream methods such as Cutout and CutMix. YOCO (Han et al. 2022) cuts one image into two pieces and performs DA individually within each piece. DA-Fusion (Trabucco et al. 2023) edits images with a pre-trained diffusion model to change their semantics, generating new images belonging to other semantic attributes. In the context of adversarial training, CropShift (Li and Spratling 2023) has demonstrated that an appropriate DA method can improve the accuracy and robustness of adversarial training. Researchers have found

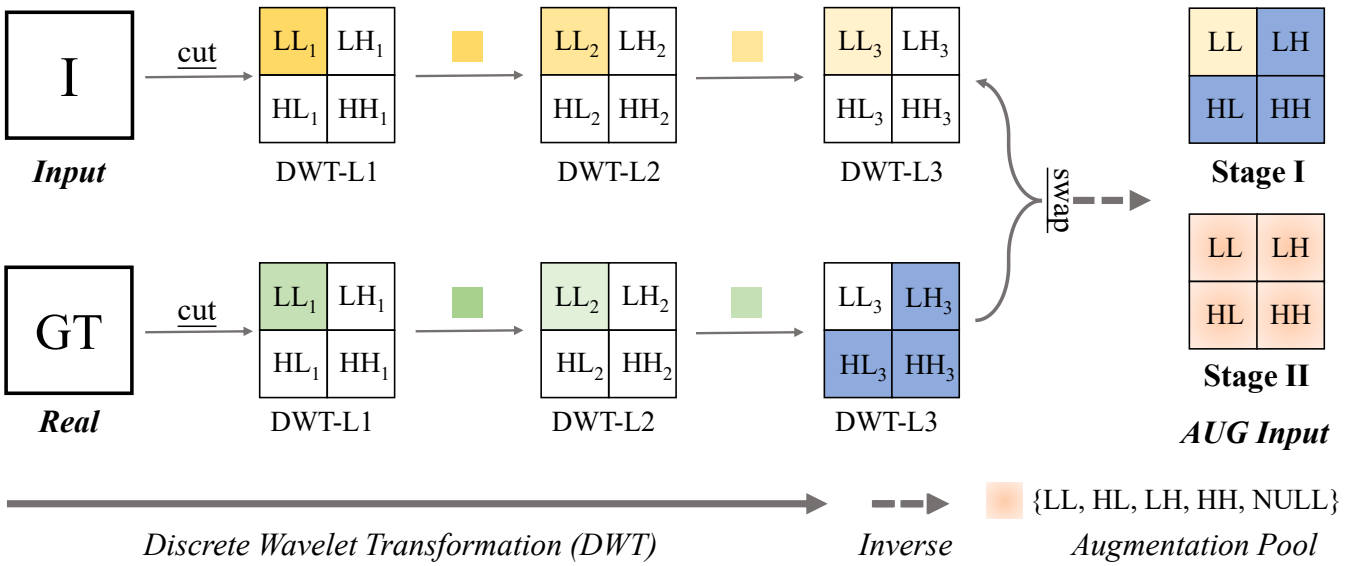


Figure 2: *The overview of CutFreq.* "*Cut*": Transform the input and real image into four frequency bands using discrete wavelet transformation (L_{th} represents the number of decompositions). "*Swap*": Define two training stages and design specific augmentation strategies (Stage I: fast convergence. Stage II: performance approximation.). The *Aug Input* in the figure represents the reconstruction guidance of LW-ISP (Chen and Ma 2022) in image denoising. Finally, the augmented samples are generated by inverse wavelet transformation.

that even the optimal hyperparameters for regularization after DA can lead to catastrophic performance degradation for minority classes (Balestriero, Bottou, and LeCun 2022).

Low-level Augmentation

In the low-level vision, Copy-blend (Shyam et al. 2021) crops patches of different sizes and shapes from the ground truth and adds them to the input so that the network can learn the position and degree of pixel restoration. CutBlur (Yoo, Ahn, and Sohn 2020) cuts low-resolution patches and pastes them into the corresponding high-resolution regions, allowing the model to adaptively decide how well to restore the image. However, this strategy is only specific to super-resolution tasks. CutDepth (Ishii and Yamashita 2021) performs a mix of input and output for depth estimation tasks, which is an extension of Mixup (Zhang et al. 2017) on depth estimation. DAMix (Chang, Sung, and Lin 2021) generates haze images via ground truth and a set of combined scales, from the source domain to the target domain, to better adapt to the domain shift. APA (Jiang et al. 2021a) encourages fair competition between the generator and discriminator by having the generator augment the real data distribution.

Frequency Domain Processing

When denoising or other low-level visual operations are performed directly in the frequency domain, although the image texture details can be easily protected, the smoothing of the image edge information is also easy to occur(Knaus and Zwicker 2013). Unlike style transfer, low-level augmentation does not change the original RAW color point or the color map of the RGB input to the desired output. StyleAugment (Chun and Park 2021) changes the predefined style samples to obtain styles of different samples in the mini-batch instantly. Strategies to reconstruct the output with high frequency details have emerged in image inpainting, but only through super-resolution networks to change the input size (Kim et al. 2020). There are already methods to utilize frequency information to enhance the image generation in image translation, and FDIT (Cai et al. 2021) introduces a training function based on the frequency domain.

Considering that low-level tasks are not suitable for hard-coded transformation, we aim to design a general augmentation method suitable for low-level from the perspective of the frequency domain.

Methodology

In this section, we first present a reformulation of the discrete wavelet transformation and then demonstrate the convergence process of the model for different frequency components in low-level vision. Furthermore, we introduce our proposed CutFreq for low-level vision, as shown in Figure 2.

Wavelet Analysis

In the realm of frequency analysis, wavelet transformation distinguishes itself from other techniques, such as Fourier analysis. This distinction arises from its ability to capture not only frequency information but also spatial information within a signal. This dual capability renders wavelet transformation particularly effective in low-level vision, as emphasized by Stephane (Stephane 1999). The process begins with a base function ψ and creates a set of dilations and shifts of ψ known as $\mathcal{X}(\psi)$.

$$\mathcal{X}(\psi) = \left\{ \psi_{pq} = 2^{-p/2} \psi \left(2^{-p} x - q \right) \mid p, q \in Z \right\}, \quad (1)$$

where ψ denotes the orthogonal wavelet. The Discrete Wavelet Transformation (DWT) deconstructs images into increasingly smaller and simpler segments. This decomposition is instrumental in simplifying the analysis and manipulation of images. Utilizing DWT, an image is partitioned into four frequency bands: LL representing the low-frequency band and HL, LH, and HH signifying the high-frequency bands. The LL band, obtained from the initial level of DWT decomposition, can be further subjected to DWT to yield second-level sub-components, namely $LL2, HL2, LH2, HH2$. If we denote DWT as $\Psi(\cdot)$, then the high and low-frequency bands for an image x are expressed as $\Psi^{\mathrm{HL}}(x)$, $\Psi^{\mathrm{LH}}(x)$, $\Psi^{\mathrm{HH}}(x)$, and $\Psi^{\mathrm{LL}}(x)$, respectively. The inverse process of DWT, known as IDWT (Graps 1995), is defined as follows:

$$\begin{aligned}\mathcal{W}_\varphi[j,k] =& h_\varphi[-n] * \mathcal{W}_\varphi[j+1,n] \\ &+ h_\theta[-n] * \mathcal{W}_\theta[j+1,n] \mid_{n=\frac{k}{2}, k\leq 0}.\end{aligned} \tag{2}$$

The Choice of Wavelet Basis Functions. Unlike in the case of the discrete Fourier transform (DFT), there are several sets of basis functions used in the DWT. When applied to computer vision, the Haar and Daubechies (db) basis functions are the most commonly used. In general, the selection of wavelet bases depends on properties such as support length, symmetry, and regularity. In wavelet analysis, the Haar wavelet is the most commonly used orthogonal wavelet function with compact support. On the other hand, Daubechies wavelet has reasonable regularity and can achieve better frequency band division, but it has a weaker time domain support and significantly increases the computational complexity. Furthermore, except for $N+1$, the Daubechies wavelet is asymmetric, which can lead to phase distortion. For CutFreq, which requires discrete wavelet and inverse transform operations, the symmetry of the wavelet function is crucial. Therefore, the Haar basis function is the preferred choice.

Preliminary Analysis

As stated in Sec. *Introduction*, we propose the frequency-band normalized $\mathcal{L}_1$ distance (FBND) to measure model convergence for different frequency components. Given a set of training samples $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and their corresponding real images $\mathcal{J}(x_1, x_2, \ldots, x_n)$. On the task of low-level vision, since the prediction result $f(x_i)$ is the value of pixels instead of the categorical probability distribution, FBND can be conveniently used to judge the convergence state of frequency components, which can be formulated as:

$$\begin{aligned}\mathcal{FBND} = \frac{1}{n}\sum_{i}^{n} \left\| \Psi^t \circ f(x_i) - \Psi^t \circ \mathcal{J}(x_i) \right\|_1, \\ t \in LL, HL, LH, HH.\end{aligned} \tag{3}$$

Different tasks exhibit unique frequency patterns. By computing and averaging the frequency domain differences between degraded images and their ground truth across the dataset, we illustrate the distinct frequency domain patterns for various tasks in Figure 3. We observe that this reflects the general frequency features to be learned within the task

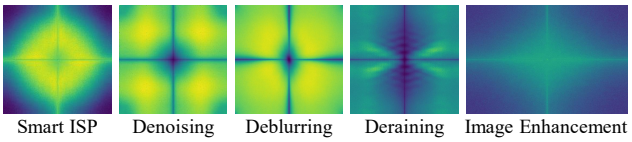


Figure 3: *Frequency patterns of low-level vision tasks*. Different degraded tasks correspond to a specific frequency nature and have obvious discrimination.

datasets. For instance, low-frequency information emerges as a crucial frequency component that the task focuses on in low-light image enhancement.

The convergence behavior of frequency components is closely related to the specific task. Figure 4 presents the quantitative results of LW-ISP (Chen and Ma 2022) on image denoising (SIDD (Abdelhamed, Lin, and Brown 2018)) and image enhancement (LoL (Wei et al. 2018)). We observe the following: (a) By combining the information from Figure 1 and Figure 4, we can see that the distance between the reconstructed image of the current model and the real image in the frequency domain is still distinct, and can be quantified by our FBND. (b) Image denoising models do not learn high-frequency components as much as low-frequency components, although high-frequency bands are inherently sparse. In contrast, learning low frequencies in image enhancement is a significant characteristic in the latter half of model training.

Upon computing the FBND after each epoch of standard training, we derive two insights across five low-level vision tasks: (1) Different tasks demonstrate varied convergence trends in frequency components. (2) These reconstruction tasks can be broadly categorized into two groups: high-frequency sustained convergence (*image denoising, image deblurring, and image deraining*) and high-frequency fast convergence (*smart ISP, image enhancement*). For instance, in image enhancement, the model rapidly learns high-frequency components primarily during the initial phase of the training.

CutFreq: Cut-and-Swap Frequency Components

High-level augmentation methods encourage underlying algorithms to focus on multiple discriminative features. However, as these techniques destroy the spatial relationship with neighboring regions, performance can deteriorate when used for low-level tasks, where textural consistency between recovered and its neighboring regions is vital to ensure effective performance. To this end, we introduced CutFreq by defining two operations.

Definition 1: Cut. *Given a training sample I, the ground truth can be denoted as J. Each input can be decomposed into four bands ($\Psi^{\mathrm{LL}}(I)$, $\Psi^{\mathrm{HL}}(I)$, $\Psi^{\mathrm{LH}}(I)$ and $\Psi^{\mathrm{HH}}(I)$) by discrete wavelet transformation $\Psi(\cdot)$. The LL band can be further degraded into four corresponding bands by $\Psi(\cdot)$.*

As defined above, CutFreq adopts the standard discrete wavelet transform (Edwards 1991) to decompose the input and the real image into frequency domain representation. Inspired by the frequency component convergence experi-

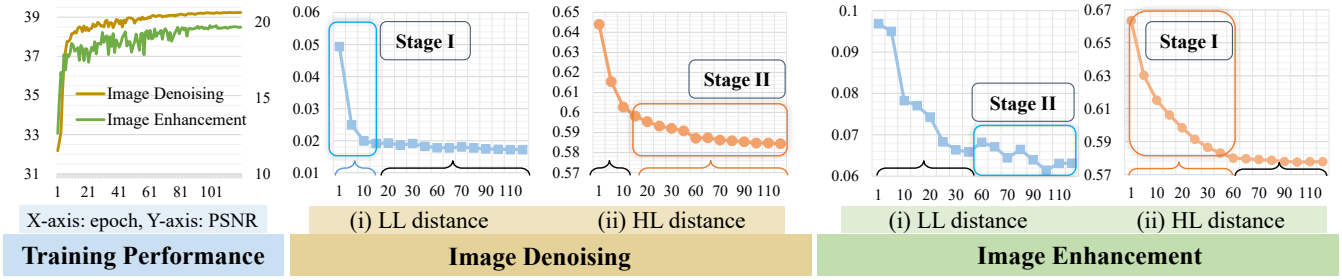


Figure 4: *Convergence process of each frequency band*. (Left panel) Performance curves of LW-ISP (Chen and Ma 2022) during training on image denoising and image enhancement. (Middle panel) DWT decomposes each sample into four bands, including LL, LH, HL, and HH. Only HL is placed due to limited space, measured by our FBND. X-axis: epochs. Y-axis: averaged frequency distance value. (Right panel) The same experiment on the image enhancement.

ment, CutFreq augments the input in two stages.

Definition 2: Swap. *Given a training process $\mathcal{P}$, fast convergence $\mathcal{P}_e$ and performance approximation $\mathcal{P}_l$ are defined as two learning stages of the model. In $\mathcal{P}_e$, the operation is selected based on the specific task: replacing either the high frequency ($\Psi^{\mathrm{HL,LH,HH}}(J) \rightarrow \Psi^{\mathrm{HL,LH,HH}}(I)$) or low frequency ($\Psi^{\mathrm{LL}}(J) \rightarrow \Psi^{\mathrm{LL}}(I)$) bands of J with the corresponding frequency bands from I. In $\mathcal{P}_l$, select a frequency band t from J to replace the corresponding band ($t \in LL, HL, LH, HH$, and $NULL$). Finally, perform the IDWT $\hat{\Psi}(\cdot)$.*

Inspired by the analysis in Sec. *Preliminary Analysis*, we customize augmentation strategies for the two types of reconstruction tasks identified. The essence lies in swapping frequency bands of varying granularities. *Stage I-Fast convergence*: This stage focuses on guiding the model to more efficiently learn key features by adjusting specific frequency bands of the input image. *Stage II-Performance approximation*: Different frequency bands of the real image are alternately exchanged with the input image. Note that the exchange of *NULL* bands (band discarding) should avoid involving low-frequency components, as the image reconstruction process fundamentally relies on the integrity of the image's information. Moreover, during the real bands exchange, selective fusion with the original input is advisable, as this facilitates the introduction of a suitable amount of detail and texture information while preserving the primary structure of the image.

Experiment

Experiment Setup

We evaluate the proposed CutFreq on benchmark datasets and experimental settings for five low-level vision tasks: (a) smart ISP, (b) image denoising, (c) image deraining, (d) image deblurring, and (e) image enhancement. For smart ISP, we evaluate our method on Zurich RAW to RGB (Ignatov, Van Gool, and Timofte 2020) (Zurich for short) dataset. In tables, the best and the second-best quality scores of the evaluated methods are **highlighted** and underlined.

Baselines and Evaluation. We use three low-level vision models: LW-ISP (Chen and Ma 2022), HINet (Chen et al. 2021b) and Restormer (Zamir et al. 2021). These models have varying numbers of parameters and network structures (CNN and transformer). For fair comparisons, every model is trained from scratch using the authors' official code unless mentioned otherwise. Please note that we do not use multi-stage algorithms, such as heterogeneous knowledge distillation, to ensure fairness.

Implementation Details. The experiments are implemented with PyTorch 1.2.0 on RTX NVIDIA 2080Ti and PH402 SKU 200 with 12G memory GPUs. In experiments, PSNR (Huynh-Thu and Ghanbari 2008) and SSIM (Wang et al. 2004) are used to evaluate the image quality.

Study on Different Augmentation Methods

Considering the difference in the output form of the generation tasks and the recognition tasks, the previous high-level methods cannot achieve satisfying results in the low-level vision. (a) High-level visual methods: Mixing by combining different image content information changes the color and structural information of the image (CutMix, Mixup, CutMixup). Cropping the selected area of the sample will result in the loss of spatial information (Cutout). (b) Traditional methods: such as RGB permutation and blend, do not cause severe spatial distortion but bring about the effect of a sharp transformation of the structure. (c) Low-level vision methods: CutBlur is more suitable for super-resolution tasks, requiring more urgent modeling of local and global relationships between pixels. We choose LW-ISP (Chen and Ma 2022) to compare augmentation methods in smart ISP and image denoising. Table 1 and Table 2 show the substantial improvement of our CutFreq and demonstrate the feasibility of data augmentation in the frequency domain.

Study on Different Low-level Tasks

Smart ISP. The ISP pipeline derives the smart ISP (Schwartz, Giryes, and Bronstein 2018; Ratnasingam 2019; Ignatov, Van Gool, and Timofte 2020) from the existing DL-based methods. Although LW-ISP achieves superior performance, it suffers from color distortion and local details missing. The results of our method (+0.17 dB, Table 1) are more in line with the real characteristics and can generate better details.

Image Denoising and Image Deblurring. We train

Method	PSNR (dB)	MSSSIM
baseline	21.37	0.8591
Mixup (Zhang et al. 2017)	21.36 (*-0.01*)	0.8605
CutMix (Yun et al. 2019)	21.38 (*+0.01*)	**0.8628**
CutMixup (Yoo, Ahn, and Sohn 2020)	21.45 (*+0.08*)	0.8614
Cutout (DeVries and Taylor 2017)	21.36 (*-0.01*)	0.8601
CutBlur (Yoo, Ahn, and Sohn 2020)	19.87 (*-1.50*)	0.8461
Copy-blend (Shyam et al. 2021)	19.87 (*-1.50*)	0.8442
Blend (Prados Gutiérrez et al. 2013)	20.89 (*-0.48*)	0.8557
RGB perm. (Lee, Hwang, and Shin 2020)	21.01 (*-0.36*)	0.8557
CutFreq	**21.54 (*+0.17*)**	0.8624

Table 1: *Smart ISP*: Comparison of different augmentation methods on Zurich (Ignatov, Van Gool, and Timofte 2020).

Method	PSNR (dB)	SSIM
baseline	39.1966	0.9162
Mixup (Zhang et al. 2017)	39.2180 (*+0.021*)	0.9162
CutMix (Yun et al. 2019)	39.1445 (*-0.052*)	0.9153
CutMixup (Yoo, Ahn, and Sohn 2020)	39.1492 (*-0.047*)	0.9153
Cutout (DeVries and Taylor 2017)	39.2440 (*+0.047*)	0.9164
CutBlur (Yoo, Ahn, and Sohn 2020)	38.9014 (*-0.295*)	0.9133
Copy-blend (Shyam et al. 2021)	39.1977 (*+0.001*)	0.9163
Blend (Prados Gutiérrez et al. 2013)	37.2671 (*-1.930*)	0.8949
RGB perm. (Lee, Hwang, and Shin 2020)	39.0681 (*-0.129*)	0.9146
CutFreq	**39.2904 (*+0.094*)**	**0.9168**

Table 2: *Image Denoising*: Comparison of augmentation methods on SIDD (Abdelhamed, Lin, and Brown 2018).

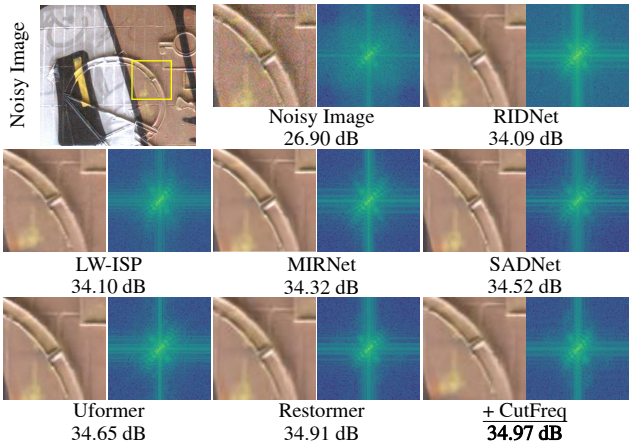


Figure 5: *Image denoising results on DND* (Plotz and Roth 2017). We apply CutFreq on Restormer (Zamir et al. 2021). The frequency map of the clipped region is also shown.

HINet (Chen et al. 2021b) and Restormer (Zamir et al. 2021) on the training set of SIDD (Abdelhamed, Lin, and Brown 2018) and directly evaluate it on the test images of both SIDD and DND (Plotz and Roth 2017) datasets. Notably, on the SIDD dataset, our CutFreq obtains PSNR gains of 0.09 dB and 0.06 dB over the previous CNN method HINet and transformer model Restormer, respectively. As shown in Figure 5, our method generates denoised images that are natural and vivid in appearance and have better global and local contrast. We evaluate image deblurring on GoPro (Nah, Hyun Kim, and Mu Lee 2017) and HIDE (Shen et al. 2019). When averaged across all datasets, our method obtains a performance boost of 0.11 dB over the Restormer.

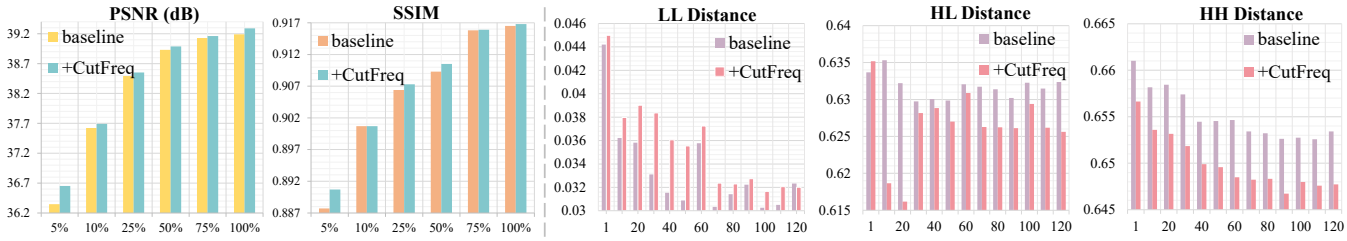
Image Deraining and Image Enhancement. CutFreq achieves consistent performance gain on five deraining datasets (Test100, Rain100H, Rain100L, Test2800, Test1200). Compared to Restormer (Zamir et al. 2021), CutFreq achieves 0.06 dB improvement when average across all datasets. We also achieve quite competitive results on the low-light LoL dataset (Wei et al. 2018). On the basis of LW-ISP, CutFreq can improve the PSNR by 0.13 dB.

Ablation Study

Augmentation Setup. We consider that for data augmentation in low-level vision, the training objectives have already reached pixel-level complexity, and it may not be necessary to augment every single sample. By comparing the complete augmentation after 5 epochs (-0.044 dB), the complete augmentation after half training (-0.050 dB), and the probability augmentation (+0.0214 dB[1]), we find that the performance of probability augmentation is the best, and other methods may even affect the normal training process.

Augmentation Strategies. Our method involves stage-wise swapping of frequency bands between the input and ground truth images in the frequency domain. The convergence trends of the model on different tasks inspired fine-grained adjustments to CutFreq. To comprehensively demonstrate the learning bias of the reconstructed model to frequency bands during training, we conduct ablation experiments using a range of frequency bands (LW-ISP on denoising). (a) The frequency information of different decomposition times (denoted as J_n) is used as the cutting standard instead of different frequency bands. We package the remaining high-frequency information (LL, LH and HH) after each decomposition into a group as the content of "*Swap*". The results prove that this strategy is slightly worse than band swapping (J_2: -0.02 dB, J_3: -0.01 dB). (b) Under the strategy of maintaining the first stage, we replace another frequency band every certain number of training epochs (10 epochs and 20 epochs). Experiments show that this strategy does not allow the model to concentrate on learning the unreplaced frequency bands (-0.12 dB, -0.07 dB). The distinction between LH, HL and HH frequency bands is not obvious. (c) What are the implications if we overlook the settings of the two stages? Focusing solely on replacing either low-frequency components or high-frequency components throughout the training cycle (J_3) yields distinct results: for the former, we achieve 39.23 dB (+0.04 dB), and for the latter, we record 39.08 dB (-0.11 dB). This observation is in line with the typical characteristics of image denoising. It suggests that the model tends to reach better convergence when provided with more targeted and precise information.

[1]Taking Mixup on image denoising as an example, baseline: 39.1966 dB.



(a) Low-data Regime (Denoising - SIDD) (b) Frequency Band Distances (FBND) (Smart ISP - Zurich)

Figure 6: *Study on the low-data regime and demonstration of frequency convergence trends*. (a) X-axis: the percentage of training data. Y-axis: PSNR/SSIM values. (b) X-axis: epochs. Y-axis: the frequency-band normalized $\mathcal{L}_1$ distance (FBND).

(d) During "*Swap*", we incorporate a *NULL* option, signifying the discarding of specific frequency components in input samples. Randomly dropping certain high-frequency bands during training aids in better convergence on high-frequency components, resulting in performance improvements of 0.08 dB and 0.03 dB in smart ISP and image enhancement, respectively, compared to the baseline. The frequency discrepancies are depicted in Figure 6.

Hyperparameters. In developing CutFreq, we observe that model performance is affected by hyper-parameters. Specifically: (a) Decomposition times. In the image denoising of LW-ISP, for example, CutFreq with J_2 (+0.06 dB) and J_4 (+0.03 dB) slightly outperforms J_1 (-0.10 dB). (b) Basis function. Our ablation studies on DWT's basis functions reveal the Haar basis functions' superiority over Daubechies. Using Daubechies, performance consistently remains below the baseline, be it J_1 (-0.11 dB), J_2 (-0.03 dB), or J_3 (+0.01 dB). We speculate this might stem from the asymmetry of Daubechies wavelets, potentially introducing noise that deviates from the original distribution in the enhanced image.

High-low Frequency Decomposition. Traditional frequency decomposition methods, like the discrete Fourier transform (DFT), are adept at segregating high-frequency and low-frequency components. They primarily aim to reduce image distortion originating from impurity components by converting the image from the spatial to the frequency domain. However, DWT stands out as it concurrently captures spatial and frequency information. We conduct ablation study replacing DWT with DFT in our CutFreq. Under the LW-ISP setting, DFT performs 0.10 dB worse than DWT in image enhancement (0.03 dB higher than baseline).

Discussion

To begin with, we discuss the interpretation of our CutFreq. Regarding the understanding of the frequency domain, our method is suitable for observing the convergence trend of the model for different components. By using "*Cut*" and "*Swap*", the model can identify the general pattern of data in a specific frequency component, which can provide a certain degree of regularization.

Low-data Regime. It is generally known that a large model benefits more from augmentation than a small model does (Yoo, Ahn, and Sohn 2020). Meanwhile, in order to verify the superiority of our method in the low-data regime, we choose the lightweight model (LW-ISP) to verify the gain effect of CutFreq. Figure 6 demonstrates the consistent superiority of our method under the different sizes of training data. We investigate the model performance while decreasing the data size (100%, 75%, 50%, 25%, 10%, 5%) for training. With the reduction, the performance differential between the baseline and our approach remains consistent.

Spectral Loss. In our exploration of spectral loss as a method to foster convergence in the frequency domain, we explore an initial evaluation of established frequency supervision approaches. This includes the focal frequency loss (*FFL*)(Jiang et al. 2021b) and the wavelet loss (*WL*)(Zhang et al. 2022). In our LW-ISP experiments tailored for image denoising, the PSNR values achieved by *FFL* and *WL* are recorded as 39.2028 dB (+0.0062) and 39.2411 dB (+0.0445), respectively. Similarly, for smart ISP, *FFL* yields a PSNR of 21.304 dB (-0.066), while *WL* registers 21.208 dB (-0.162). The integration of our observations from Figure 3 and Figure 4 with the conceptualization of spectral loss is a direction we believe merits future exploration.

Limitations and Future Work. In the realm of low-level vision, we propose CutFreq as an augmentation strategy. However, its effectiveness may not be universally optimal, given that deep neural networks typically do not learn frequency domain information explicitly. This limitation opens up intriguing possibilities for explicitly incorporating frequency signal learning through approaches like contrastive learning or tailored loss design. These approaches are poised to assist models in discerning *when* and learning *which* frequency components are crucial for reconstructing samples.

Conclusion

Inspired by the characteristic difference between BN and IN, we perform single-sample fine-grained augmentation. In this paper, we validate the prevalence of frequency-domain gaps in low-level vision qualitatively and quantitatively and use frequency-band normalized $\mathcal{L}_1$ distance to measure model convergence for different frequency components. Furthermore, We introduce the CutFreq for low-level vision, which aims to preserve high-level representations with directionality and improve image synthesis quality. Experiments on five low-level vision tasks demonstrate the effectiveness of our method. We also demonstrate the superior performance of the method in the low-data regime.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB2804103), the Key Research and Development Program of Shaanxi (2021ZDLGY01-05), the National Natural Science Foundation of China (20211710187, 31970972), Tsinghua University Dushi Program, Tsinghua University Talent Program, Institute for Interdisciplinary Information Core Technology (IIISCT) and Ant Group through CCF-Ant Research Fund.

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1692–1700.

Balestriero, R.; Bottou, L.; and LeCun, Y. 2022. The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*.

Cai, M.; Zhang, H.; Huang, H.; Geng, Q.; Li, Y.; and Huang, G. 2021. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13930–13940.

Chang, C.-M.; Sung, C.-S.; and Lin, T.-N. 2021. DAMix: Density-Aware Data Augmentation for Unsupervised Domain Adaptation on Single Image Dehazing. *arXiv preprint arXiv:2109.12544*.

Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R. E.; and Zhu, C. 2022. Real-world single image super-resolution: A brief review. *Information Fusion*, 79: 124–145.

Chen, H.; and Ma, K. 2022. LW-ISP: A Lightweight Model with ISP and Deep Learning. *arXiv preprint arXiv:2210.03904*.

Chen, J.-N.; Sun, S.; He, J.; Torr, P.; Yuille, A.; and Bai, S. 2021a. TransMix: Attend to Mix for Vision Transformers. *arXiv preprint arXiv:2111.09833*.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021b. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192.

Choi, J.; Song, Y.; and Kwak, N. 2020. Part-aware data augmentation for 3d object detection in point cloud. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3391–3397. IEEE.

Chun, S.; and Park, S. 2021. StyleAugment: Learning Texture De-biased Representations by Style Augmentation without Pre-defined Textures. *arXiv preprint arXiv:2108.10549*.

Devillers, L.; Vidrascu, L.; and Lamel, L. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4): 407–422.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Edwards, T. 1991. Discrete wavelet transforms: Theory and implementation. *Universidad de*, 28–35.

Engin, D.; Genç, A.; and Kemal Ekenel, H. 2018. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 825–833.

Graps, A. 1995. An introduction to wavelets. *IEEE computational science and engineering*, 2(2): 50–61.

Han, J.; Fang, P.; Li, W.; Hong, J.; Armin, M. A.; Reid, I.; Petersson, L.; and Li, H. 2022. You Only Cut Once: Boosting Data Augmentation with a Single Cut. *arXiv preprint arXiv:2201.12078*.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Ignatov, A.; Van Gool, L.; and Timofte, R. 2020. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 536–537.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. PMLR.

Ishii, Y.; and Yamashita, T. 2021. CutDepth: Edge-aware Data Augmentation in Depth Estimation. *arXiv preprint arXiv:2107.07684*.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021a. Deceive D: Adaptive Pseudo Augmentation for GAN Training with Limited Data. *Advances in Neural Information Processing Systems*, 34.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021b. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13919–13929.

Kim, S. Y.; Aberman, K.; Kanazawa, N.; Garg, R.; Wadhwa, N.; Chang, H.; Karnad, N.; Kim, M.; and Liba, O. 2020. Zoom-to-Inpaint: Image Inpainting with High-Frequency Details. *arXiv preprint arXiv:2012.09401*.

Knaus, C.; and Zwicker, M. 2013. Dual-domain image denoising. In *2013 IEEE International Conference on Image Processing*, 440–444. IEEE.

Lee, H.; Hwang, S. J.; and Shin, J. 2020. Self-supervised label augmentation via input transformations. In *International Conference on Machine Learning*, 5714–5724. PMLR.

Lee, J.; Kim, E.; Lee, J.; Lee, J.; and Choo, J. 2021. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34.

Li, B.; Wu, F.; Lim, S.-N.; Belongie, S.; and Weinberger, K. Q. 2021. On feature normalization and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12383–12392.

Li, L.; and Spratling, M. 2023. Data Augmentation Alone Can Improve Adversarial Training. *arXiv preprint arXiv:2301.09879*.

Liang, W.; Liang, Y.; and Jia, J. 2023. MiAMix: Enhancing Image Classification through a Multi-stage Augmented

Mixied Sample Data Augmentation Method. *arXiv preprint arXiv:2308.02804*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Plotz, T.; and Roth, S. 2017. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1586–1595.

Prados Gutiérrez, R.; et al. 2013. Image blending techniques and their application in underwater mosaicing.

Ratnasingam, S. 2019. Deep camera: A fully convolutional neural network for image signal processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Schwartz, E.; Giryes, R.; and Bronstein, A. M. 2018. DeepISP: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2): 912–923.

Shen, Z.; Wang, W.; Lu, X.; Shen, J.; Ling, H.; Xu, T.; and Shao, L. 2019. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5572–5581.

Shyam, P.; Sengar, S. S.; Yoon, K.-J.; and Kim, K.-S. 2021. Evaluating copy-blend augmentation for low level vision tasks. *arXiv preprint arXiv:2103.05889*.

Stephane, M. 1999. A wavelet tour of signal processing.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8174–8182.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1357–1366.

Yoo, J.; Ahn, N.; and Sohn, K.-A. 2020. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8375–8384.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2021. Restormer: Efficient Transformer for High-Resolution Image Restoration. *arXiv preprint arXiv:2111.09881*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L.; Chen, X.; Tu, X.; Wan, P.; Xu, N.; and Ma, K. 2022. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12464–12474.

Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; and Ye, J. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE*.