

Frequency-Aware Facial Image Shadow Removal through Skin Color and Texture Learning

Ling Zhang^{1,2}  , Wenyang Xie¹  , Chunxia Xiao^{†3} 

¹School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

²Hubei Key Laboratory of Intelligent Information Processing and Realtime Industrial System, Wuhan University of Science and Technology, Wuhan, China

³School of Computer Science, Wuhan University, Wuhan, China

zhling@wust.edu.cn, xiewenyang@wust.edu.cn, cxxiao@whu.edu.cn

Abstract

Existing facial image shadow removal methods predominantly rely on pre-extracted facial features. However, these methods often fail to capitalize on the full potential of these features, resorting to simplified utilization. Furthermore, they tend to overlook the importance of low-frequency information during the extraction of prior features, which can be easily compromised by noises. In our work, we propose a frequency-aware shadow removal network (FSRNet) for facial image shadow removal, which utilizes the skin color and texture information in the face to help recover illumination in shadow regions. Our FSRNet uses a frequency-domain image decomposition network to extract the low-frequency skin color map and high-frequency texture map from the face images, and applies a color-texture guided shadow removal network to produce final shadow removal result. Concretely, the designed Fourier sparse attention block (FSABlock) can transform images from the spatial domain to the frequency domain and help the network focus on the key information. We also introduce a skin color fusion module (CFModule) and a texture fusion module (TFModule) to enhance the understanding and utilization of color and texture features, promoting high-quality result without color distortion and detail blurring. Extensive experiments demonstrate the superiority of the proposed method. The code is available at <https://github.com/laoxie521/FSRNet>.

CCS Concepts

- Computing methodologies → Shadow removal; Facial image; Feature fusion; Frequency-aware;

1. Introduction

Due to the variations in lighting conditions and the uniqueness of facial structures, shadows often occur in facial images. The low brightness of shadow regions not only reduces the visibility and authenticity of the images, but also weakens the ability of crucial tasks such as face recognition [ABBR20, WY22, ZZLQ16], image restoration [DJBY20, LDR*22, LZZ*24], and facial image reconstruction [JP19, DTA*21, LZZ*24], affecting the accuracy and effectiveness of computer vision processing. Thus, effectively removing shadows from a facial image and recovering a clear image is a necessary and practical task.

Face image shadow removal is a complex and challenging task. Firstly, due to variations in lighting intensity and direction, the illumination in shadow regions varies significantly, increasing the difficulty of the process. Secondly, the color and illumination in shadow regions are significantly different from that in non-shadow regions. Ensuring the consistency of the appearance is also a challenge. In addition, human faces are rich in natural facial features

such as eyes, mouth, nose, and ears. The method must maintain the accuracy and naturalness of these facial features during shadow removal, preventing distortions and unnatural appearance.

Despite the significant progress in image shadow removal methods [WLY18, ZGZ22, GHL*23, GWY*23, WYW*22a, FZG*21, LCC20], there are still some obvious difficulties when dealing with face images. Human faces possess rich and intricate facial features, which become blurred or missing under shadows. Since natural image methods do not take into account these unique characteristics of face, they tend to suffer from detail distortion when processing face images [GWY*23]. Additionally, the skin color and texture of human faces are complex and delicate. These subtle variations may be difficult to be accurately processed in natural image methods, leading to skin color inconsistency and texture blurring, as shown in Figure 1(c).

Several face image shadow removal methods have been proposed [ZBT*20, ZCLX23, HXZC21, LHH*22]. Traditional methods often rely on heuristic algorithms such as illumination compensation and estimation [DH19, ZZMC18, HLL*18]. These methods often remove shadows by adjusting lighting conditions. However,

† Corresponding author.

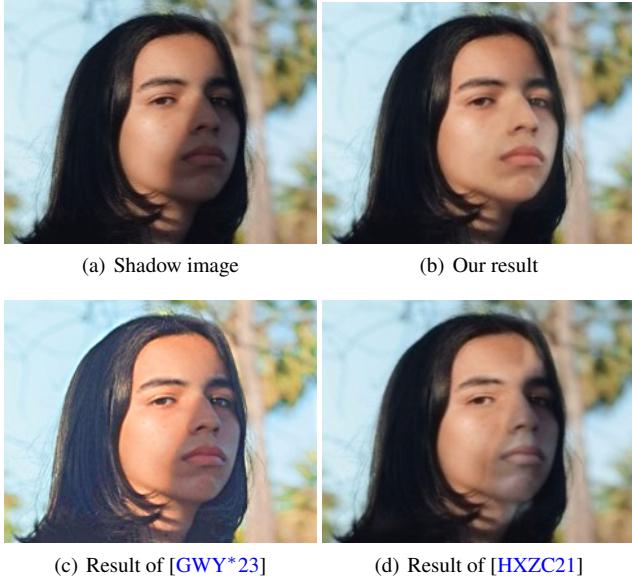


Figure 1: Facial image shadow removal. Results of [GWY*23] and [HXZC21] may cause skin color inconsistency and texture blurring, our method can produce more desirable result. The shadow image is sourced from the Internet.

since the light intensity on the face does not vary linearly, they often lead to the problem of inconsistency in the appearance between the shadow and non-shadow regions. In contrast, recent learning-based methods have shown good performance [HXZC21]. However, these methods are weakly correlated with face features. Their fusion of prior features is more homogeneous and does not fully utilize the relevant features of the face, limiting its effectiveness in guiding the image reconstruction process and resulting in unsatisfactory results. Moreover, these methods mainly focus on spatial domain processing, which tends to ignore some important detail information, leading to problems such as color distortion, as shown in Figure 1(d).

To address the above problems, we propose a frequency-aware shadow removal network (FSRNet) to remove shadows in the facial image. Figure 2 presents the framework of our FSRNet. First, we introduce a frequency-domain image decomposition network (FDecomposeNet) to decompose the image into a high-frequency part and a low-frequency part, producing a skin color map and a texture map that are unrelated to shadows. In particular, we introduce a fourier sparse attention block (FSABlock) to convert the image from the spatial domain to the frequency domain. With the sparse feature selection, our FSABlock can help the network reduce computational complexity while maintaining focus on the key information.

Then, we propose a color-texture guided shadow removal network (CTShadowNet), which utilizes the skin color and texture information in the face to help recover the illumination in the shadow regions. Specifically, we design a skin color fusion module (CFModule) to fuse image features with color features, helping the

network obtain global color feature information. To get the texture features, we introduce a texture fusion module (TFModule), which can help the network maintain better local consistency of the image.

In summary, our main contributions are as follows:

- We propose a frequency-aware shadow removal network (FSRNet) for facial image shadow removal. Our FSRNet can produce high-quality results with consistent appearance by using the skin color map and the texture map as auxiliary information.
- We introduce a fourier sparse attention block (FSABlock) to convert the image from the spatial domain to the frequency domain and focus on the key information. We also design a skin color fusion module (CFModule) and a texture fusion module (TFModule) to fuse effective color and texture features.
- Extensive experiments and evaluations verify the effectiveness and generalization ability of our FSRNet, outperforming state-of-the-art methods.

2. Related Work

2.1. Natural Image Shadow Removal

Natural image shadow removal methods can be mainly divided into two categories: traditional methods [JHK18, BT06, BDS*16, GDH11, XXZC13, FHLD05] and deep learning-based methods [LCC20, WLY18, CLZX21, ZGZ22, CPS20, GHL*23]. Traditional methods often use the physical priors to design the model, while learning-based methods learn the mapping relationship between shadow and non-shadow images through large-scale training datasets.

Guo et al. [GDH11] proposed a domain-based method that used the mean shift algorithm to segment the image and the minimum cut maximum flow algorithm to label shadow and non-shadow regions. This method reconstructs the non-shadow regions based on the differences in light intensity but struggles with complex texture details. Xiao et al. [XXZC13] developed a parameter-adaptive shadow removal algorithm based on texture matching, which effectively removes small-scale shadows but often encounters issues at shadow boundaries. Finlayson et al. [FHLD05] utilized gradient invariance, employing Poisson equations to construct the gradients and boundary conditions of shadow-free regions, and used the brightness outside the shadow boundary to restore illumination in shadowed areas. This method preserves the texture of shadowed regions well and has achieved significant results in shadow removal, though shadow boundary issues still persist.

In recent years, deep learning-based methods have achieved remarkable progress in shadow removal [LS19, JST21, LCC20, MXZP12, HFZ*19, WYW*22b, CPS20]. Wang et al. [WLY18] proposed ST-CGAN, a stacked conditional generative adversarial network framework for joint shadow detection and removal, using a discriminator to identify the relationship between shadow detection and removal. Fu et al. [FZG*21] introduced an over-exposed fusion shadow removal method, which cleverly combines over-exposed images and original shadow images through a learnable pixel weighting map. Mask-ShadowGAN [HJFH19] redefined cycle-consistency constraints to perform shadow removal on unpaired data. To fully utilize datasets, DHAN [CPS20] synthe-

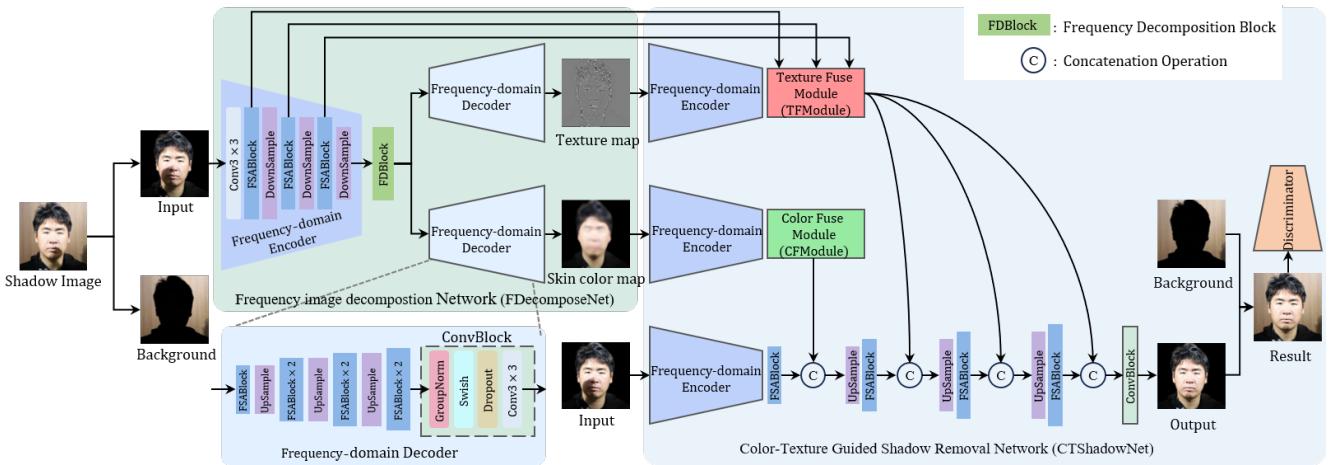


Figure 2: The framework of the proposed FSRNet. With the proposed fourier sparse attention block (FSABlock), we first use the FDecomposeNet to decompose the image into a high-frequency part and a low-frequency part, producing a skin color map and a texture map. Then, we use the CTShadowNet to remove shadows in the image. CFModule and TFMModule are used to fuse image features with color and texture features, help the network produce high-quality shadow removal results.

sized pseudo-shadow images and learned boundary-free pseudo-shadow images through a dual hierarchical aggregation network to reconstruct shadow-free images. SG-ShadowNet [WYW*22a] investigated the importance of normalization, learning style representations of non-shadow regions to harmonize shadow and non-shadow parts. Shadow-Former [GHL*23] introduced shadow interaction modules and attention mechanisms to integrate contextual information between shadow and non-shadow areas. ShadowDiffusion [GWW*23] used a diffusion model that learns the noise distribution by adding noise and then denoising, guiding the restoration of shadowed images with a refined shadow mask. DMTN [LWF*23] designed a single-stage decoupled multi-task network that jointly guides the target image through various tasks.

Although these methods are effective for shadow removal in natural images, they do not generalize well to shadow removal in face images due to the different characteristics between natural images and face images.

2.2. Facial Image Shadow Removal

Most existing facial image shadow removal methods use heuristic approaches to capture features of facial images. Due to the unique structure of faces, these algorithms can leverage facial characteristics to guide the image restoration process. Zhang [ZBT*20] employed a method of facial symmetry to guide the restoration of facial shadows. He et al. [HXZC21] proposed a progressive optimization strategy, using a pre-trained model to generate a facial shadow mask that guides the reconstruction of shadow-free images. However, this method is unstable in the shadow removal process. Zhang et al. [ZCLX23] employed a two-stage network to remove shadows by utilizing facial symmetry, generating facial optical flow to obtain a coarse shadow-free image. They then used a graph convolutional encoder to produce the final shadow-free image, which integrates with a feature modulation module. Liu et al. [LHH*22] used

grayscale shadow removal and recoloring techniques to remove facial shadows. These methods do not adequately address skin color consistency and detail preservation. To solve these problems, we extract skin color and texture features from the face to guide the image restoration process, resulting in more realistic images.

3. Methodology

In this paper, we propose a frequency-aware shadow removal network (FSRNet) to remove shadows in the facial image. Our FSRNet utilizes the skin color and texture information in the face to help recover the illumination in shadow regions of the face. Figure 2 presents the framework of our FSRNet, which consists of a frequency-domain image decomposition network (FDecomposeNet) and a color-texture guided shadow removal network (CTShadowNet). We first use FDecomposeNet to decompose the image into a high-frequency part (texture map) and a low-frequency part (skin color map). Then, under the guidance of the skin color map and the texture map, our CTShadowNet removes shadows in the facial image.

Note that, to avoid the noise and interference introduced by the background in the image, we only process the face part in our method. We use the face segmentation pre-trained model of Chen [CZL*22] to get a portrait mask (Figure 3(b)) for the original shadow image. We separate the portrait from the background through the portrait mask to get the input image containing only the person (Figure 3(c)).

3.1. Frequency-domain Image Decomposition Network

Existing methods for face image shadow removal mainly focus on spatial domain processing, but these methods often tend to overlook some important detail information, leading to issues such as

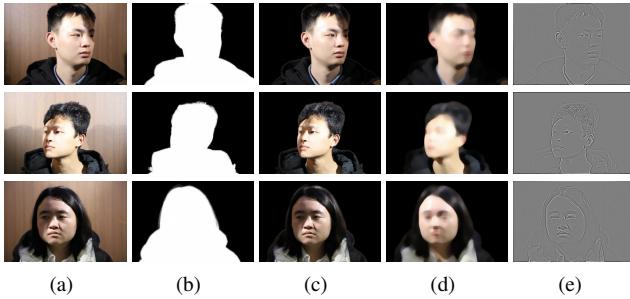


Figure 3: Skin color map and texture map. (a) is the original shadow images. (b) is portrait masks. (c) is portrait shadow images. (d) is the skin color map and (e) is the texture map. The images are sourced from FSTD [ZCLX23].

color distortion and detail blurring. To obtain better color and texture information, we design a frequency-domain image decomposition network (FDecomposeNet) to extract low-frequency and high-frequency information from the image, resulting in a shadow-free facial skin color map (Figure 3(d)) and a shadow-free texture map (Figure 3(e)). During training, we apply the Laplacian filter [APH*14] to the shadow-free ground truth in the training dataset to compute a texture map as supervised data for the high-frequency part. For the low-frequency part, we extract a face skin color map from the shadow-free ground truth to serve as supervised data. Specifically, we first use the YCrCb color domain of the face image to extract color information. Then, we employ a smoothing operation and a blur operation to weaken facial details to get a face skin color map as the label.

In FDecomposeNet, we use a frequency-domain encoder to extract the image features. Next, we apply two frequency-domain decoders to extract the skin color map and the texture map respectively. To better accomplish the decomposition, we propose a fourier sparse attention block (FSABlock) to transform images from the spatial domain to the frequency domain. The frequency-domain encoder contains a 3×3 convolution and three FSABlock+DownSample. The frequency-domain decoder contains a FSABlock+UpSample, two FSABlock+FSABlock+UpSample and a FSABlock+FSABlock+ConvBlock. The ConvBlock applies a group normalization, a swish activation, Dropout, and a convolution to restore the number of feature channels, ensuring the consistency of input and output channels. Between the encoder and decoders, we construct a frequency-domain decomposition block (FDBlock) to perform feature division on frequency domain.

Fourier Sparse Attention Block. Traditional attention mechanisms mostly utilize dense matrices to model the dependency between different positions in image processing. This strategy leads to the inclusion of some unnecessary information in the computation process, which increases the computational complexity. In addition, traditional attention mechanisms are relatively insensitive to images in the frequency domain, often tending to ignore important low-frequency information in images.

To overcome these limitations, we introduce a fourier sparse attention block (FSABlock). It converts the image from the spatial

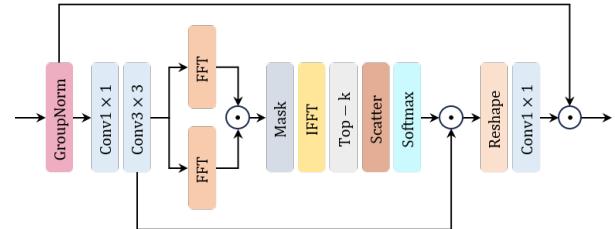


Figure 4: The architecture of our fourier sparse attention block (FSABlock).

domain to the frequency domain and extracts global features of the image. In our FSABlock, we use sparse matrices to replace traditional dense matrices. In this way, we are able to selectively focus on the key information in the image while ignoring the unimportant or redundant parts. This greatly reduces computational complexity while maintaining focus on the key information.

Figure 4 illustrates the architecture of the proposed FSABlock. We first normalize the input features F_1 to get features F_2 . Then, we use a 1×1 convolution to adjust the number of feature channels and a 3×3 convolution is employed to extract spatial context features F_3 . We split features F_3 into patches and reshape them to obtain query Q , key K , and value V , respectively. We perform fast fourier transform (FFT) on Q and K , and get F_Q and F_K respectively. We perform element-wise product multiply on F_Q and F_K to get an attention matrix $F_{spacial}$. Next, we use an inverse fast fourier transform (IFFT) to obtain the attention matrix A based on the frequency domain.

To reduce the computational effort, we use the Top-k sparsity matrix [WWW*22] to preserve the important components of attention and remove the useless information. Here, k is an adjustable parameter that dynamically controls the size of sparsity. We can replace the indicator that satisfies less than k to make the dense matrix into a sparse matrix. We utilize a scatter operation to fill the mask with the pixel values in the matrix that satisfy the condition. Next, we apply softmax to normalize the sparse matrix to obtain the attention weight F_W . We perform element-wise product multiply on V and F_W to get the sparse context features F_4 . Finally, after a reshape operation, we utilize a 1×1 convolution to obtain the final sparse attention features F_{spase} .

Frequency-domain Decomposition Block. We introduce a frequency-domain decomposition block (FDBlock) to perform feature division at the bottleneck layer on the features. Figure 5 shows the architecture of FDBlock.

We first use a LayerNormalization layer to normalize the input feature D and enhance the features. Then, we employ a FSABlock to filter out unnecessary features and noise. The obtained features are transformed from the spatial domain to the frequency domain using a fast fourier transform (FFT). Next, we use a low-pass filter to extract low-frequency features and a high-pass filter to extract high-frequency features. The extracted low-frequency and high-frequency features are subsequently transformed to the spatial domain using the inverse fast fourier transform (IFFT). Finally, we use a 3×3 convolution layer and a Relu activation function to further

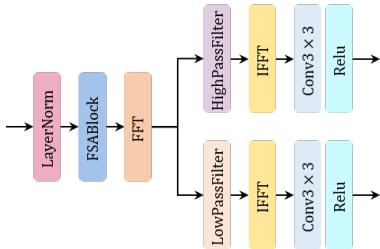


Figure 5: The structure of our frequency-domain decomposition block (FDBlock).

enhance the feature expression. The decomposed high-frequency and low-frequency features are fed into the two frequency-domain decoders, respectively.

3.2. Color-Texture Guided Shadow Removal Network

Face images usually have complex texture and skin color information, which are crucial for image processing tasks. However, current shadow removal methods for face images often suffer from challenges such as color distortion and detail blurring. As aforementioned, the low-frequency skin color information contains the brightness and color distribution of the image, while the high-frequency texture information captures and preserves the texture structure of the face. Therefore, we propose a color-texture guided shadow removal network (CTShadowNet) that utilizes the extracted skin tone map and texture map as auxiliary information. Our CTShadowNet is an encoder-decoder network with a discriminator. We employ Markov discriminator [IZZE17] as our discriminator.

Due to insufficient consideration of the correlation between features, simply concatenating features may obscure or lose some useful features. Additionally, direct concatenating features may increase the number of network parameters, thus elevating the risk of overfitting. Thus, we design a skin color fusion module (CFModule) and a texture fusion module (TFModule) that aim to better fuse image features with skin color and texture features. The fused features can aid the network in recovering illumination in the shadow regions of the face while preserving the natural appearance and texture details.

Our CTShadowNet contains three steps, as shown in Figure 2. First, we utilize three frequency-domain encoders to extract the frequency-domain features of the skin color map, texture map, and the input image separately. We denote them as F_{color} , $F_{texture}$ and F_{image} . Then, we fuse the image features with texture features and color features using TFModule and CFModule. Afterward, the fused features along with the original image features are fed into a color-texture fusion decoder for feature decoding and reconstruction of a shadow-free face image. The color-texture fusion decoder contains three FSABlock+UpSample and a FSABlock+ConvBlock.

The fused features from CFModule are concatenated with features from the first FSABlock in the color-texture fusion decoder. The output features from TFModule at each scale are concatenated

with the features from the subsequent three FSABlocks, respectively, according to their corresponding scales. This concatenation strategy ensures that features from different scales and levels of abstraction are properly aligned and combined, allowing the network to utilize both local and global contextual information. The concatenated features are then fed into further layers for processing and refinement, ultimately contributing to the generation of a high-quality shadow-free face image.

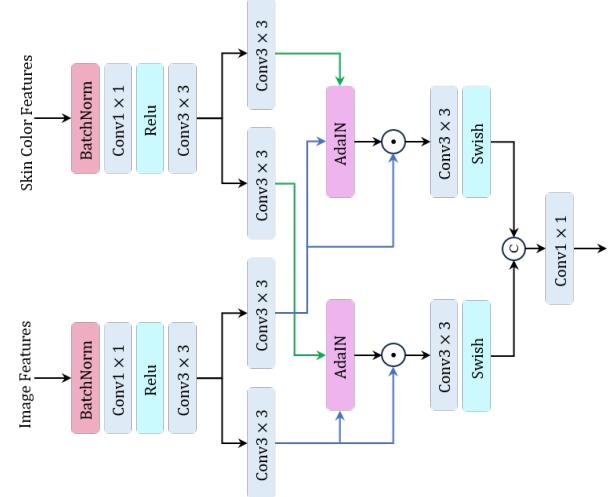


Figure 6: The structure of our skin color fusion module (CFModule).

Skin Color Fusion Module. The skin color map serves as a crucial component in the network, providing global color feature information of the image. Through the skin color fusion module (CFModule), we can precisely adjust the colors of the shadow regions based on the color information from the surrounding areas of the face, making the generated image appear more natural and realistic.

Figure 6 shows the architecture of the proposed CFModule. First, we use BatchNorm to normalize color features F_{color} and image features F_{image} separately. Then, we use a 1×1 convolution, a ReLU activation function, and a 3×3 deep convolution to obtain structural feature representations of the two normalized features, denoted as S_{color} and S_{image} . Next, we perform a convolution on S_{color} to obtain features S_1 and S_2 . Similarly, we perform a convolution on S_{image} to obtain features S_3 and S_4 . We use the AdaIN module to transfer the color of feature S_1 to S_3 and the color of feature S_2 to S_4 , obtaining the transferred features $S_{transfer1}$ and $S_{transfer2}$. Subsequently, we perform element-wise product multiply on $S_{transfer1}$ and S_3 to get feature S_5 and on $S_{transfer2}$ and S_4 to get feature S_6 . Both S_5 and S_6 are further enhanced for expressiveness through a 3×3 convolution and a Swish activation function. Finally, we integrate the enhanced two features using a concatenation operation and utilize a 1×1 convolution layer to restore the original number of channels, outputting the final color fused features F_{color} .

To further enhance the feature representation capabilities of the network, we specifically integrate the color fused features F_{color} at the lower levels of the color-texture fusion decoder. Features at this

level tend to be global and abstract. By introducing color features at this level, our CTShadowNet can more effectively capture global features related to skin color, significantly improving its resilience to color variations and potential noise. This design not only improves the accuracy of the model in skin color processing but also enhances the overall quality and naturalness of the result.

Texture Fusion Module. During the process of image restoration, ordinary upsampling methods often lead to the loss of image details. While skip connections can preserve contextual information to a certain extent, they still fall short of restoring the texture of shadow regions. Therefore, we propose a texture fusion module (TFModule) to combine image features and texture features more perfectly. Effective texture features can help the network to recover the texture information more accurately while removing shadows in the image.

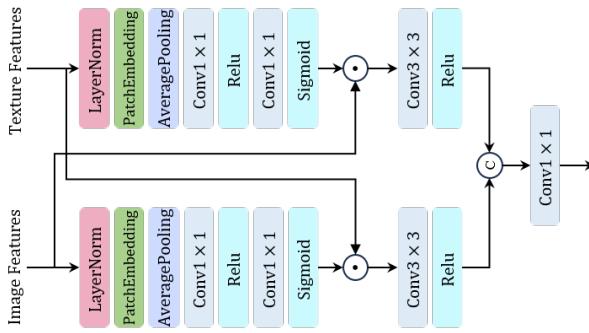


Figure 7: The structure of our texture fusion module (TFModule).

Figure 7 illustrates the structure of the proposed TFModule. First, we employ Layer Normalization (LayerNorm) to normalize the texture features F_{texture} and the image features F_{image} , ensuring stability during data propagation. Subsequently, we utilize PatchEmbedding to divide these two features into multiple small patches and convert them into lower-dimensional vector representations, reducing computational costs while capturing local information. We input the two sets of feature vectors into a channel attention unit to obtain their channel attention weights, which are A_{texture} and A_{image} . These attention weights can help the model focus on channels that contribute more to the task. The channel attention unit consists of a global average pooling, a 1×1 convolution, a Relu activation function, and a 1×1 convolution.

Next, we utilize the sigmoid activation function to convert A_{texture} and A_{image} into probability distributions P_{texture} and P_{image} . We perform element-wise multiplication on F_{image} and A_{texture} to obtain the fusion feature $F_{\text{fuse}1}$. Performing the same operation on F_{texture} and A_{image} obtains feature $F_{\text{fuse}2}$. Following that, we employ a 3×3 convolution and a Relu activation function both on $F_{\text{fuse}1}$ and $F_{\text{fuse}2}$ to extract spatial features, capturing local information in the images. Finally, we concatenate the two spatial features and use a 1×1 convolution to restore the original number of channels, outputting the final texture-fused feature. This fused feature contains rich information from both the input image and the texture, emphasizing key channels and spatial information for the shadow removal task.

To better capture image details and structures, we introduce texture features into the later layers of the color-texture fusion decoder. These features are rich in information about the structures and details of the image, enriching the feature representation of the decoder and helping the network maintain better local consistency of the image. Thus, we can effectively address issues such as the loss of detail and texture distortion that may occur during shadow removal.

3.3. Loss Function

The loss function we utilize for network optimization consist of four components: frequency-domain reconstruction loss $L_{\text{frequency}}$, appearance consistency loss $L_{\text{appearance}}$, structural consistency loss $L_{\text{structure}}$ and adversarial loss L_{adv} . The total loss function is expressed as:

$$L_{\text{loss}} = L_{\text{frequency}} + \lambda_1 L_{\text{appearance}} + \lambda_2 L_{\text{structure}} + \lambda_3 L_{\text{adv}}, \quad (1)$$

where λ_1 , λ_2 and λ_3 are the weight parameters.

Frequency-domain reconstruction loss is used to constrain FDecomposeNet to generate desirable facial skin color map I_{color} and texture map I_{texture} , which is described as:

$$L_{\text{frequency}} = \|I_{\text{color}} - I_{\text{color}}^{\text{gt}}\| + \|I_{\text{texture}} - I_{\text{texture}}^{\text{gt}}\|_1, \quad (2)$$

where $I_{\text{color}}^{\text{gt}}$ is the supervised data for the high-frequency part, and $I_{\text{texture}}^{\text{gt}}$ is the supervised data for the low-frequency part.

Appearance consistency loss is used to ensure the authenticity of the shadow removal result I_{free} generated by FSRNet. We use the L_1 distance between I_{free} and shadow-free ground truth I_{gt} to evaluate the data loss, that is,

$$L_{\text{appearance}} = \|I_{\text{free}} - I_{\text{gt}}\|_1, \quad (3)$$

Structural consistency loss is used to evaluate the structural loss of the shadow removal result I_{free} and the shadow-free ground truth I_{gt} , which is calculated as,

$$L_{\text{structure}} = \|VGG(I_{\text{free}}) - VGG(I_{\text{gt}})\|_2^2, \quad (4)$$

where $VGG()$ is the feature extractor of the pre-trained VGG19 model.

Adversarial loss is used for the discriminator to determine whether the generated result is real or fake, which is calculated as,

$$L_{\text{adv}} = \mathbb{E}_{(I, I_{\text{free}}, I_{\text{gt}})} [\log(D(I_{\text{gt}})) + \log(1 - D(I))], \quad (5)$$

where D is the discriminator, and I is the shadow image.

4. Experiments

4.1. Implementation Detail

Our network is implemented using Pytorch, which is trained on NVIDIA GeForce RTX3090. Our FSRNet is trained using the Adam optimizer with 300 epochs. The decay rate beta is set to (0.5, 0.999). The initial learning rate is set to 0.0003. In our experiments, the weight parameters λ_1 , λ_2 and λ_3 are set to 5, 1 and 0.01, respectively.

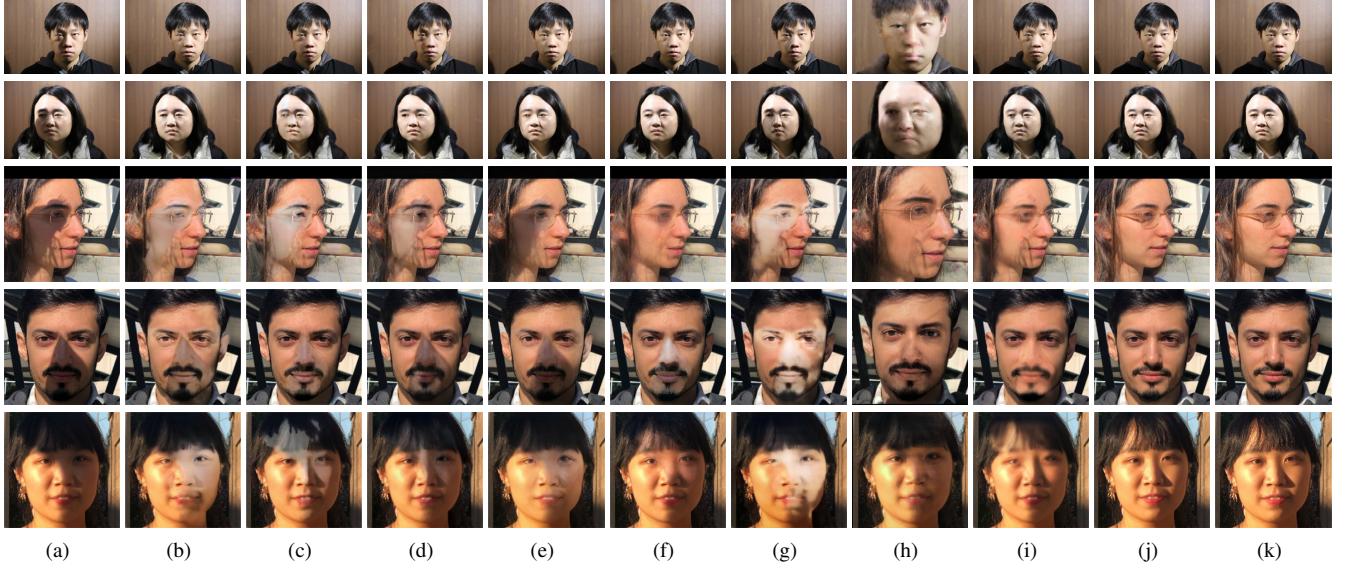


Figure 8: Visual comparison among state-of-the-art shadow removal results: (a) input images, (b) Auto-Exposure Fusion [FZG*21], (c) Spa-Former [ZGZ22], (d) Style-Guided [WYW*22a], (e) DMTN-Net [LWF*23], (f) Shadow-Former [GHL*23], (g) He et al. [HXZC21], (h) Liu et al. [LHH*22], (i) Zhang et al. [ZCLX23], (j) our FSRNet, (k) ground truth images. The images are sourced from FSTD [ZCLX23] and Zhang [ZBT*20].

Table 1: Quantitative comparisons of shadow removal on FSTD [ZCLX23] and Zhang [ZBT*20] datasets in terms of RMSE, PSNR, and SSIM. All the learning-based methods are trained on FSD+ dataset. \uparrow means the larger the better while \downarrow means the smaller the better.

Methods	Venue/Year	FSTD			Zhang		
		PSNR \uparrow	SSIM \uparrow	RMSE \downarrow	PSNR \uparrow	SSIM \uparrow	RMSE \downarrow
Fu et al. [FZG*21]	CVPR/2021	31.682	0.965	8.743	19.385	0.802	30.128
Spa-Former [ZGZ22]	CVPR/2021	29.631	0.954	11.849	26.710	0.892	14.317
SG-ShadowNet [WYW*22a]	ECCV/2022	32.869	0.973	7.683	23.458	0.872	21.472
DMTN-Net [LWF*23]	TMM/2023	32.054	0.967	8.942	26.973	0.894	15.457
Shadow-Former [GHL*23]	AAAI/2023	36.475	0.972	6.812	30.279	0.904	11.678
ShadowDiffusion [GWY*23]	CVPR/2023	37.475	0.976	5.595	32.367	0.918	10.032
Zhang [ZBT*20]	SIGGRAPH/2020	26.386	0.894	24.728	23.816	0.782	29.834
He et al. [HXZC21]	CVPR/2021	24.541	0.931	18.545	21.870	0.816	28.438
Liu et al. [LHH*22]	ECCV/2022	22.652	0.842	27.139	19.427	0.742	31.289
Zhang et al. [ZCLX23]	PG/2023	36.423	0.982	5.389	29.775	0.931	9.901
our FSRNet	PG/2024	38.024	0.984	4.698	34.356	0.936	7.360

4.2. Comparison with State-of-the-arts

Dataset. We use the dataset FSD+ as the training data for our method, which contains two parts. One is the FSD dataset [ZCLX23], which consists of 2,800 pairs of face shadow and shadow-free images. The other is a dataset constructed by Zhang et al. [ZCLX23], including 1,612 pairs of face shadow and shadow-free images. We use two test datasets to evaluate our FSRNet. One is FSTD [ZCLX23], containing 964 pairs of face images. The other test dataset is proposed by Zhang [ZBT*20], which contains 100 pairs of images.

Metrics. We use the root mean square error (RMSE), the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) between the shadow removal result and the ground

truth shadow-free image to evaluate the performance of our FSRNet.

Quantitative Comparison. To validate the effectiveness of our FSRNet, we compare our results with state-of-the-art shadow removal methods, including six natural shadow removal methods [FZG*21, ZGZ22, WYW*22a, LWF*23, GHL*23, GWY*23] and four facial shadow removal methods [ZBT*20, HXZC21, LHH*22, ZCLX23]. For a fair comparison, we train all learning-based methods on the FSD+ dataset using the same hardware. Table 1 concludes the comparison results using three metrics. From the table, we can observe that our method achieves the best values for all metrics among all comparison methods, confirming the effectiveness of our approach.

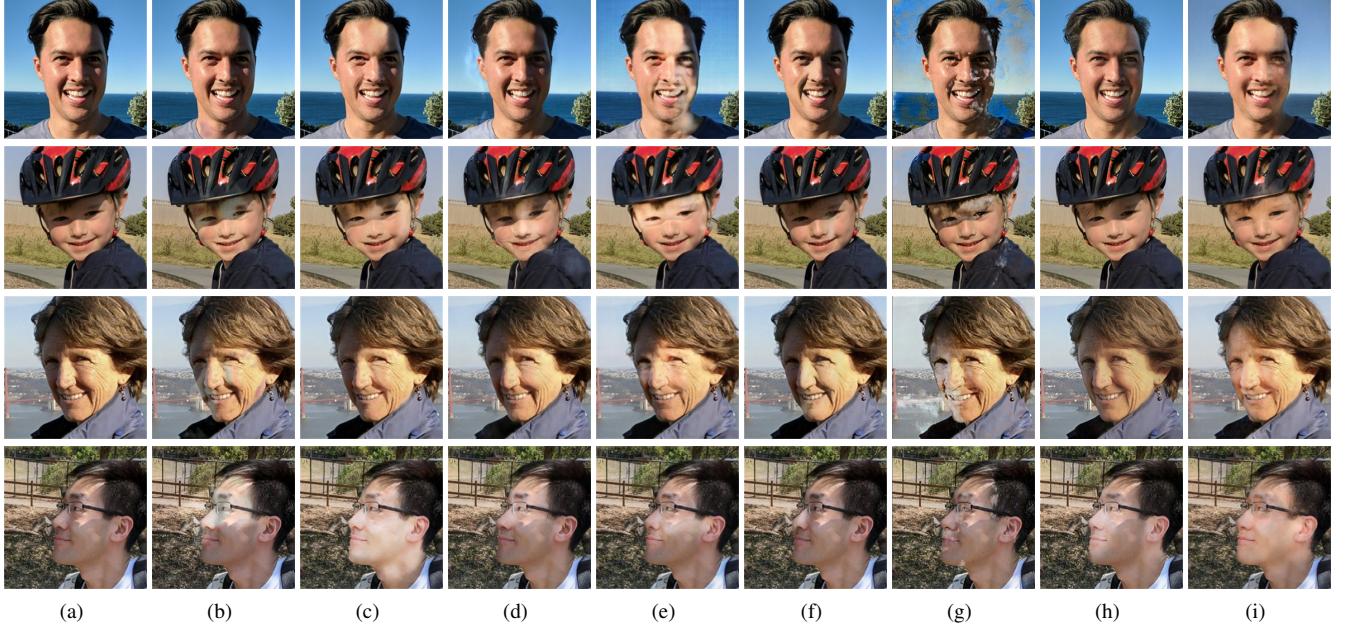


Figure 9: Visual comparison among state-of-the-art shadow removal results: (a) input images, (b) Auto-Exposure Fusion [FZG*21], (c) Spa-Former [ZGZ22], (d) Style-Guided [WYW*22a], (e) DMTN-Net [LWF*23], (f) Shadow-Former [GHL*23], (g) He et al. [HXZC21], (h) Zhang et al. [ZCLX23], (i) our FSRNet. The images are sourced from Zhang [ZBT*20].

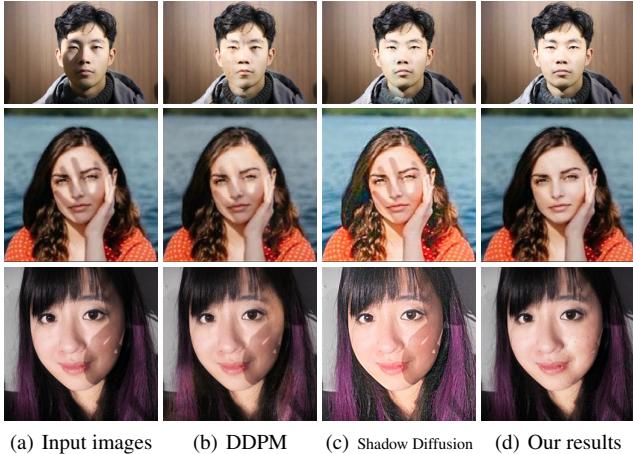


Figure 10: Compared with the diffusion model. (b) and (c) are shadow removal results produced by DDPM [LDR*22], ShadowDiffusion [GWY*23], (d) is our results. Row 1 is from the FSTD [ZCLX23], and Rows 2 and 3 are from the Internet.

Visual Comparison. To demonstrate the superiority of our method, we provide some visual results of facial image shadow removal, as shown in Figure 8. It can be observed that, Fu et al. [FZG*21] exhibit overexposure of the face, resulting in color distortion, as shown in Figure 8(b). Spa-Former [ZGZ22] fails to remove all shadows for complex facial images, as shown in Figure 8(c). SG-ShadowNet [WYW*22a] is capable of removing shad-

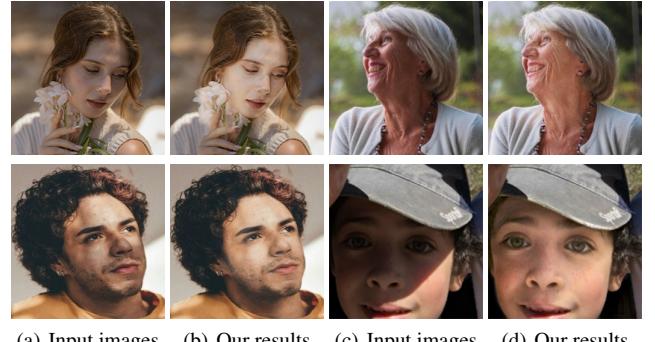


Figure 11: Comparison of face images in complex scenes. The input images are sourced from the Internet.

ows from faces but introduces artifacts around shadow boundaries, as shown in Figure 8(d). DMTN-Net [LWF*23] suffers from the loss of details in the results, as shown in Figure 8(e). Shadow-Former [GHL*23] can get artifacts with shadow artifacts, as shown in Figure 8(f). He et al. [HXZC21] are insensitive to environmental lighting, resulting in significant differences in skin color, as shown in Figure 8(g). Liu et al. [LHH*22] struggle with complex shadows and exhibit unstable performance, as shown in Figure 8(h). Zhang et al. [ZCLX23] cannot preserve facial skin color and texture information well, as shown in Figure 8(i). In contrast, our FSRNet effectively removes shadows in the images, maintaining consistent appearance without color distortion and loss of detail, as shown in Figure 8(j). Our results are similar to the ground truth images.

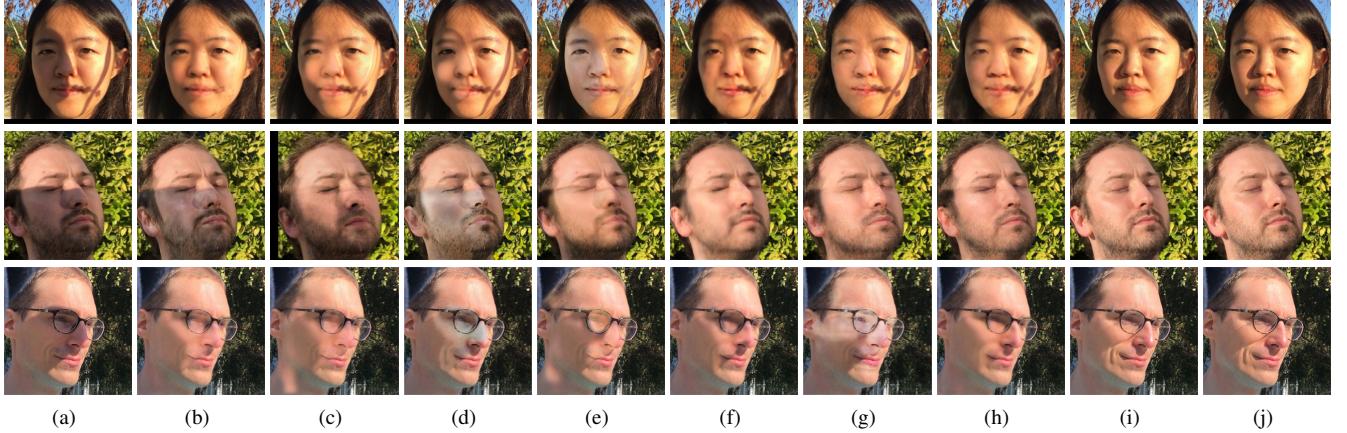


Figure 12: Visual comparison for ablation study: (a) input images, (b) $FSRNet_1$, (c) $FSRNet_2$, (d) $FSRNet_3$, (e) $FSRNet_4$, (f) $FSRNet_5$, (g) $FSRNet_6$, (h) $FSRNet_7$, (i) our $FSRNet$ and (j) ground truth images. The images are sourced from Zhang [ZBT*20].

To further validate the robustness and generalization ability of our network in image shadow removal, Figure 9 presents some shadow removal results on facial images from real-world scenarios. These include challenging cases such as facial skin tone variations, inconsistent lighting, and heavy shadows on the face. Apparently, our results look more realistic and natural, ensuring sufficient facial details. This demonstrates the good robustness and generalization ability of our method.

To further demonstrate the superiority of our method, we compared our method with two methods based on diffusion model. Figure 10 provides the shadow removal results. It can be seen, DDPM [LDR*22] fails to preserve the details of the image effectively, as shown in Figure 10(b). ShadowDiffusion [GWY*23] exhibits noticeable differences in skin tones, as shown in Figure 10(c). Comparatively, Our method effectively removes shadows in the image and achieves a more natural and realistic appearance.

Moreover, our method is capable of effectively handling facial shadow areas in complex scenes, as shown in Figure 11. Our excellent visual results demonstrate the robustness of our method.

User study. We conduct a user study to evaluate the visual performance of our method and some state-of-the-art shadow removal methods. We have prepared 120 sets of shadow removal images. Each group includes eight different shadow removal results produced by our $FSRNet$, Fu [FZG*21], Spa-Former [ZGZ22], SG-ShadowNet [WYW*22a], DMTN-Net [LWF*23], Shadow-Former [GHL*23], He et al. [HXZC21] and Liu et al. [LHH*22]. We randomly select 120 volunteers. For each volunteer, we randomly provide them with 20 image sets. Volunteers are asked to choose the best shadow-free image for each group. Statistics of all results, we found that 19.12% of the shadow removal images generated by $FSRNet$ are selected as the best shadow-free images, while 10.66%, 10.12%, 11.14%, 13.65%, 12.62%, 10.64% and 12.05% of the results are selected by Fu [FZG*21], Spa-Former [ZGZ22], SG-ShadowNet [WYW*22a], DMTN-Net [LWF*23], Shadow-Former [GHL*23], He et al. [HXZC21] and Liu et al. [LHH*22]. Compared to other methods, our method gets the best result, which

demonstrates that the shadow removal images obtained by our method are more visually satisfactory.

Table 2: Quantitative results of ablation study on FSTD [ZCLX23] and Zhang [ZBT*20] datasets using PSNR, SSIM, and RMSE.

Methods	FSTD			Zhang		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
$FSRNet_1$	29.420	0.956	11.009	26.505	0.867	23.710
$FSRNet_2$	32.704	0.965	7.765	23.162	0.883	20.947
$FSRNet_3$	36.475	0.968	6.812	30.279	0.894	15.678
$FSRNet_4$	35.472	0.976	5.242	22.973	0.904	11.457
$FSRNet_5$	36.683	0.970	6.025	25.317	0.915	10.726
$FSRNet_6$	37.430	0.980	5.036	32.480	0.925	8.059
$FSRNet_7$	37.683	0.978	5.025	32.317	0.918	9.726
$FSRNet$	38.024	0.984	4.698	34.356	0.936	7.360

4.3. Ablation Study

To evaluate the performance of different components used in $FSRNet$, we perform ablation experiments by disabling or modifying a specific component. We design seven variants:

- (1) $FSRNet_1$: Replace FSABlock with the vanilla self-attention model [VSP*17];
- (2) $FSRNet_2$: $FSRNet$ without both skin color and texture information as the auxiliary information;
- (3) $FSRNet_3$: $FSRNet$ without skin color information as the auxiliary information;
- (4) $FSRNet_4$: $FSRNet$ without texture information as the auxiliary information;
- (5) $FSRNet_5$: $FSRNet$ without TFMModule and CFModule, texture and color features from the encoders are directly connected to the color-texture fusion decoder;
- (6) $FSRNet_6$: $FSRNet$ without TFMModule, and texture features from the encoder are directly connected to the color-texture fusion decoder;
- (7) $FSRNet_7$: $FSRNet$ without CFModule, and color features from the encoder are directly connected to the color-texture fusion decoder.

We train seven variants on FSD+ and evaluate the results on two test datasets. Table 2 summarizes the quantitative results. As can be seen from the table, (1) all the components designed in our method can improve the performance of our FSRNet; (2) skin color fusion module and texture fusion module can help improve the performance of the method; (3) our fourier sparse attention block is effective. We also provide some visualization results in Figure 12, from which we can find that our FSRNet with all components produces more realistic shadow removal results.

4.4. Limitation

Our FSRNet can effectively remove shadows in face images. However, when the shadow is very dark, some high-frequency information on the face may be lost, such as beards and hair. Skin color information cannot be fully utilized, resulting in blurred details and poor appearance, as shown in Figure 13.

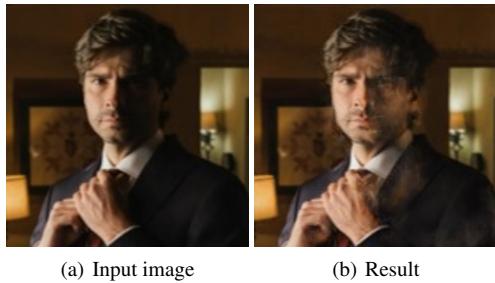


Figure 13: Limitation. The image is sourced from the Internet.

5. Conclusion

In this paper, we propose a frequency-aware shadow removal network (FSRNet) for facial image shadow removal, which contains a frequency-domain image decomposition network (FDecomposeNet) and a color-texture guided shadow removal network (CTShadowNet). We first use FDecomposeNet to extract the low-frequency skin color map and high-frequency texture map from the face images. Then, with the color and texture features as auxiliary information, CTShadowNet can produce the final shadow removal result. Concretely, the designed FSABlock can transform images from the spatial domain to the frequency domain, helping the network focus on the key information. CTShadowNet uses CFModule and TFModule to fuse image features with skin color and texture features, promoting high-quality results without color distortion detail blurring. The extensive experiments validate the superiority of our FSRNet.

Acknowledgments

This work is supported by NSFC (No.61902286, No.62372336). It is also supported by Nature Science Foundation of Hubei Province (No.2023AFB615).

References

- [ABBR20] AMMOUR B., BOUBCHIR L., BOUDEN T., RAMDANI M.: Face-iris multimodal biometric identification system. *Electronics* 9, 1 (2020), 85. 1
- [APH*14] AUBRY M., PARIS S., HASINOFF S. W., KAUTZ J., DURAND F.: Fast local laplacian filters: Theory and applications. *ACM Transactions on Graphics (TOG)* 33, 5 (2014), 1–14. 4
- [BDS*16] BAKO S., DARABI S., SHECHTMAN E., WANG J., SUNKAVALLI K., SEN P.: Removing shadows from images of documents. In *Asian Conference on Computer Vision* (2016), Springer, pp. 173–183. 2
- [BT06] BROWN M. S., TSOI Y.-C.: Geometric and shading correction for images of printed materials using boundary. *IEEE Transactions on Image Processing* 15, 6 (2006), 1544–1554. 2
- [CLZX21] CHEN Z., LONG C., ZHANG L., XIAO C.: Canet: A context-aware network for shadow removal. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 4743–4752. 2
- [CPS20] CUN X., PUN C.-M., SHI C.: Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 10680–10687. 2
- [CZL*22] CHEN X., ZHU Y., LI Y., FU B., SUN L., SHAN Y., LIU S.: Robust human matting via semantic guidance. In *Proceedings of the Asian Conference on Computer Vision* (2022), pp. 2984–2999. 3
- [DH19] DU L., HU H.: Nuclear norm based adapted occlusion dictionary learning for face recognition with occlusion and illumination changes. *Neurocomputing* 340 (2019), 133–144. 1
- [DJBY20] DIN N. U., JAVED K., BAE S., YI J.: A novel gan-based network for unmasking of masked face. *IEEE Access* 8 (2020), 44276–44287. 1
- [DTA*21] DIB A., THEBAULT C., AHN J., GOSELIN P.-H., THEOBALT C., CHEVALLIER L.: Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12819–12829. 1
- [FHLD05] FINLAYSON G. D., HORDLEY S. D., LU C., DREW M. S.: On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence* 28, 1 (2005), 59–68. 2
- [FZG*21] FU L., ZHOU C., GUO Q., JUEFEI-XU F., YU H., FENG W., LIU Y., WANG S.: Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 10571–10580. 1, 2, 7, 8, 9
- [GDH11] GUO R., DAI Q., HOIEM D.: Single-image shadow detection and removal using paired regions. In *CVPR 2011* (2011), IEEE, pp. 2033–2040. 2
- [GHL*23] GUO L., HUANG S., LIU D., CHENG H., WEN B.: Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650* (2023). 1, 2, 3, 7, 8, 9
- [GWY*23] GUO L., WANG C., YANG W., HUANG S., WANG Y., PFISTER H., WEN B.: Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 14049–14058. 1, 2, 3, 7, 8, 9
- [HFZ*19] HU X., FU C.-W., ZHU L., QIN J., HENG P.-A.: Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence* 42, 11 (2019), 2795–2808. 2
- [HJFH19] HU X., JIANG Y., FU C.-W., HENG P.-A.: Mask-shadowgan: Learning to remove shadows from unpaired data. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 2472–2481. 2
- [HLL*18] HU C.-H., LU X.-B., LIU P., JING X.-Y., YUE D.: Single

- sample face recognition under varying illumination via qrcp decomposition. *IEEE Transactions on Image Processing* 28, 5 (2018), 2624–2638. 1
- [HXZC21] HE Y., XING Y., ZHANG T., CHEN Q.: Unsupervised portrait shadow removal via generative priors. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 236–244. 1, 2, 3, 7, 8, 9
- [IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1125–1134. 5
- [JHK18] JUNG S., HASAN M. A., KIM C.: Water-filling: An efficient algorithm for digitized document shadow removal. In *Asian Conference on Computer Vision* (2018), Springer, pp. 398–414. 2
- [JP19] JO Y., PARK J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 1745–1753. 1
- [JST21] JIN Y., SHARMA A., TAN R. T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5027–5036. 2
- [LCC20] LIN Y.-H., CHEN W.-C., CHUANG Y.-Y.: Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12905–12914. 1, 2
- [LDR*22] LUGMAYR A., DANELLIAN M., ROMERO A., YU F., TIMOFTE R., VAN GOOL L.: Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 11461–11471. 1, 8, 9
- [LHH*22] LIU Y., HOU A. Z., HUANG X., REN L., LIU X.: Blind removal of facial foreign shadows. In *BMVC* (2022), p. 88. 1, 3, 7, 8, 9
- [LS19] LE H., SAMARAS D.: Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8578–8587. 2
- [LWF*23] LIU J., WANG Q., FAN H., LI W., QU L., TANG Y.: A decoupled multi-task network for shadow removal. *IEEE Transactions on Multimedia* (2023). 3, 7, 8, 9
- [LZZ*24] LIU Z., ZHAO Y., ZHAN S., LIU Y., CHEN R., HE Y.: Pcdnf: Revisiting learning-based point cloud denoising via joint normal filtering. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (2024), 5419–5436. 1
- [MXZP12] MENG G., XIANG S., ZHENG N., PAN C.: Nonparametric illumination correction for scanned document images via convex hulls. *IEEE transactions on pattern analysis and machine intelligence* 35, 7 (2012), 1730–1743. 2
- [VSP*17] VASWANI A., SHAZER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.: Attention is all you need. *Advances in neural information processing systems* 30 (2017). 9
- [WLY18] WANG J., LI X., YANG J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 1788–1797. 1, 2
- [WWW*22] WANG P., WANG X., WANG F., LIN M., CHANG S., LI H., JIN R.: Kvt: k-nn attention for boosting vision transformers. In *European conference on computer vision* (2022), Springer, pp. 285–302. 4
- [WY22] WANG H., YAN W. Q.: Face detection and recognition from distance based on deep learning. In *Aiding Forensic Investigation Through Deep Learning and Machine Learning Frameworks*. IGI Global, 2022, pp. 144–160. 1
- [WYW*22a] WAN J., YIN H., WU Z., WU X., LIU Y., WANG S.: Style-guided shadow removal. In *European Conference on Computer Vision* (2022), Springer, pp. 361–378. 1, 3, 7, 8, 9
- [WYW*22b] WAN J., YIN H., WU Z., WU X., LIU Z., WANG S.: Cr-former: A cross-region transformer for shadow removal. *arXiv preprint arXiv:2207.01600* (2022). 2
- [XXZC13] XIAO C., XIAO D., ZHANG L., CHEN L.: Efficient shadow removal using subregion matching illumination transfer. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 421–430. 2
- [ZBT*20] ZHANG X., BARRON J. T., TSAI Y.-T., PANDEY R., ZHANG X., NG R., JACOBS D. E.: Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 78–1. 1, 3, 7, 8, 9
- [ZCLX23] ZHANG L., CHEN B., LIU Z., XIAO C.: Facial image shadow removal via graph-based feature fusion. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, p. e14944. 1, 3, 4, 7, 8, 9
- [ZGZ22] ZHANG X. F., GU C. C., ZHU S. Y.: Spa-former: Transformer image shadow detection and removal via spatial attention. *arXiv preprint arXiv:2206.10910* (2022). 1, 2, 7, 8, 9
- [ZZLQ16] ZHANG K., ZHANG Z., LI Z., QIAO Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503. 1
- [ZZMC18] ZHANG W., ZHAO X., MORVAN J.-M., CHEN L.: Improving shadow suppression for illumination robust face recognition. *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 611–624. 1