UMKC
UNIVERSITY OF MISSOURI
KANSAS CITY

DCT-based residual network for NIR image colorization

Hongcheng Jiang[1], Paras Maharjan[1], Zhu Li[1], Gorge York[2]
[1]University of Missouri-Kansas City, USA,
[2]US Air Force Academy

Introduction

- In this paper, we propose the Discrete Cosine Transform (DCT)-based Residual Network (DCT-RCAN) for NIR image colorization. Specifically, the output of our network is four coefficients generated by the Two-Dimensional (2D) 4×4 DCT of RGB images, which reduces the training difficulty of our network by explicitly separating low-frequency and high-frequency details into four subgroups. We adopt the Residual in Residual (RIR) module as a basic module in our network, which can reduce the complexity of the model. Thus, our method can focus on more crucial underlying patterns in channel dimension in a lightweight manner. Extensive experiments validate that our DCT-RCAN is computationally efficient and demonstrate competitive results against state-of-the-art NIR image colorization methods.

Key Contributions

- 1. We use the four subgroups generated by the 4×4 DCT of RGB images as the output, so that the coarse contents and sharp details are separated explicitly during training. The size of subgroups is 1/4th of the original image, which alleviates the learning difficulty of our network without information loss. The DCT and its inverse operation are both invertible, leading to no information loss. Thus, our network can easily generate the RGB images via the inverse DCT.
- 2. We exploit the RCAN which includes shallow feature extraction and deep feature extraction. DCT can be seamlessly integrated to RCAN and add trivial computational cost. In addition, our DCT-RCAN can focus on more crucial underlying patterns in channel dimension in a lightweight manner.
- 3. Extensive experiments validate that our DCT-RCAN is computationally efficient and demonstrate competitive results against state-of-the-art NIR images colorization methods.

Dataset

- Our dataset comes solely from the Grand Challenge on NIR image colorization of IEEE VCIP 2020. The training and test set respectively contains 372 and 28 paired images (NIR and RGB images). There are also 1020 RGB images without paired NIR images that can be used optionally. The resolution of all images is 256×256. The Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Angular Error (AE) are used to evaluate the quality of translated RGB images.

Network

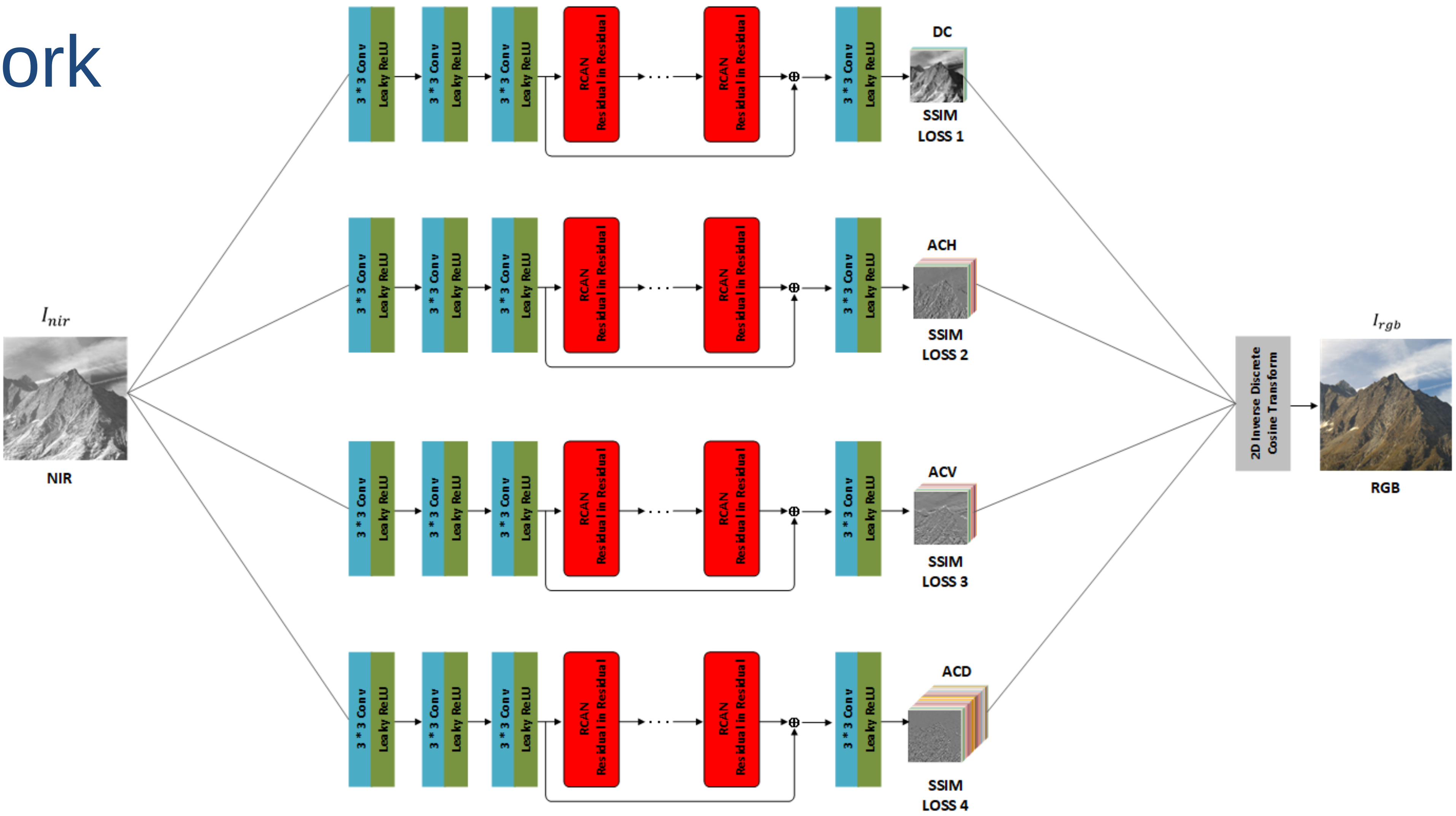


Figure 1. Proposed Network Structure

DCT

- The Discrete Cosine Transform (DCT) is one of many transforms that takes its input and transforms it into a linear combination of weighted basis functions. DCT basis functions for N = 4 can be seen in Figure 2.

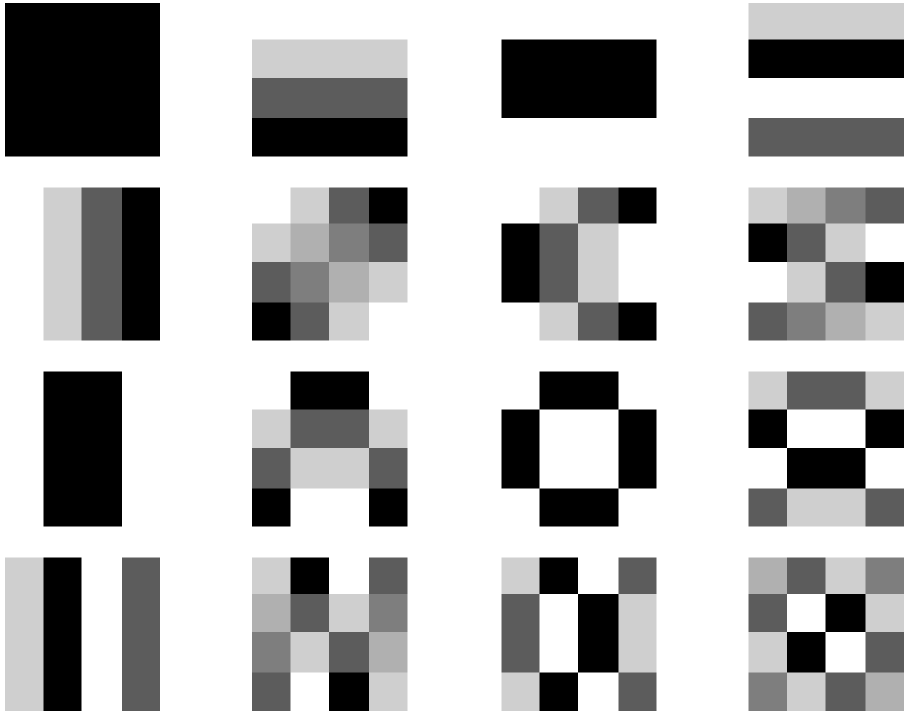


Figure 2. 4x4 DCT basis

DCT

- We used 4×4 DCT to decompose the compressed image and subsampled the resulting DCT image to create 16 sub-band images, each coefficient representing DC, AC1, AC2,..., AC15. DC coefficient represents the low-frequency component and has global information of the image. AC1 to AC15 represents high-frequency components of the image. Particularly, AC1, AC2, AC3 coefficients represent the horizontal subgroup (ACH), AC4, AC8, AC12 coefficient represent the vertical subgroup (ACV), and the rest of the AC coefficients represent the diagonal subgroup (ACD).

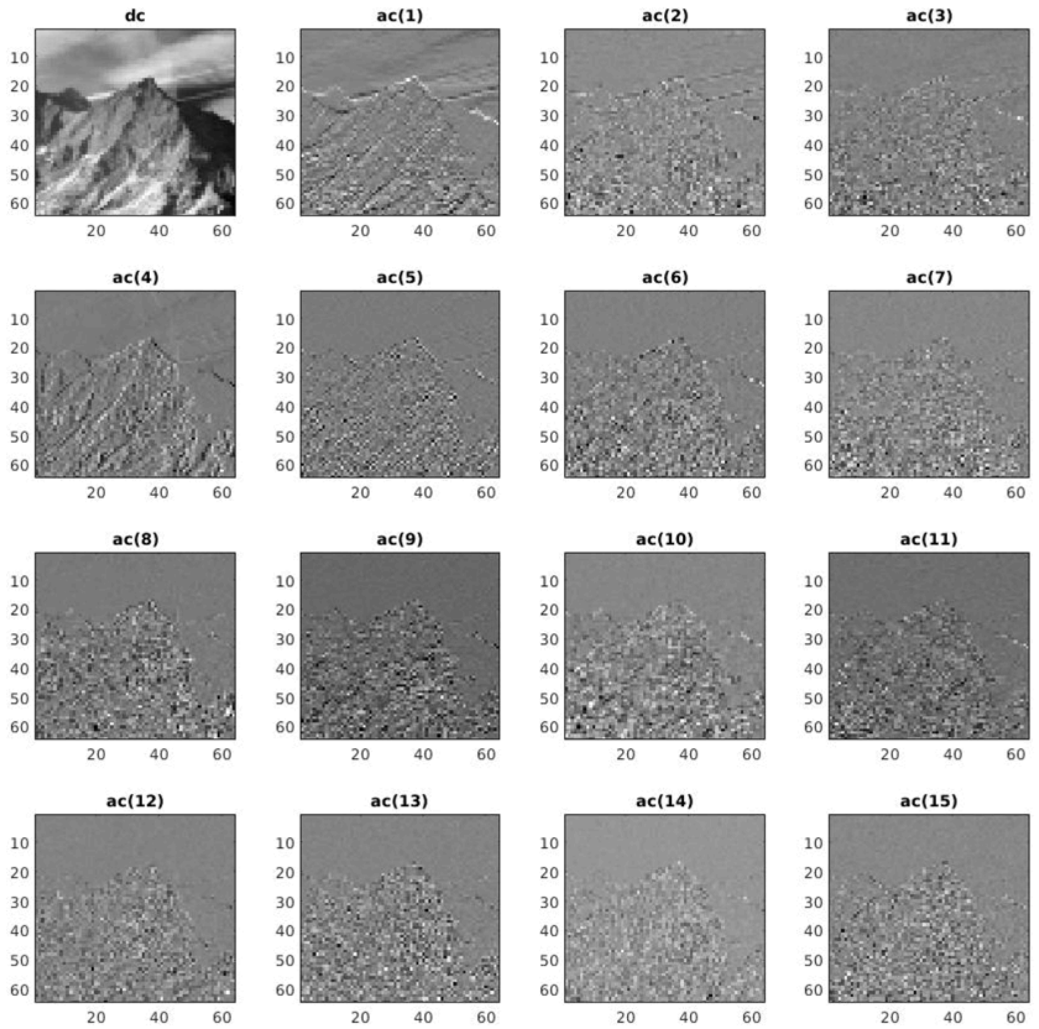
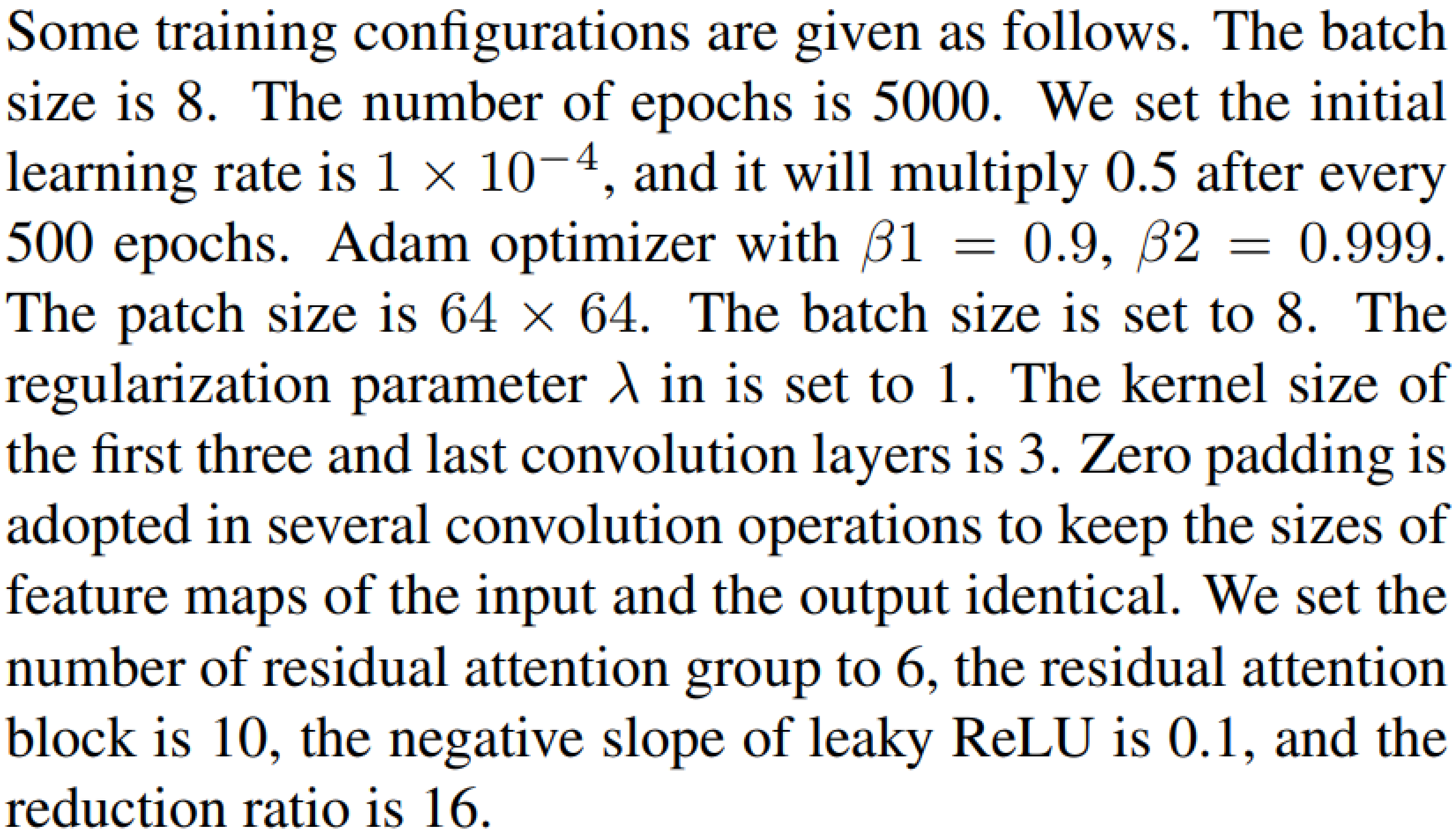


Figure 3. 4x4 DCT basis

RCAN

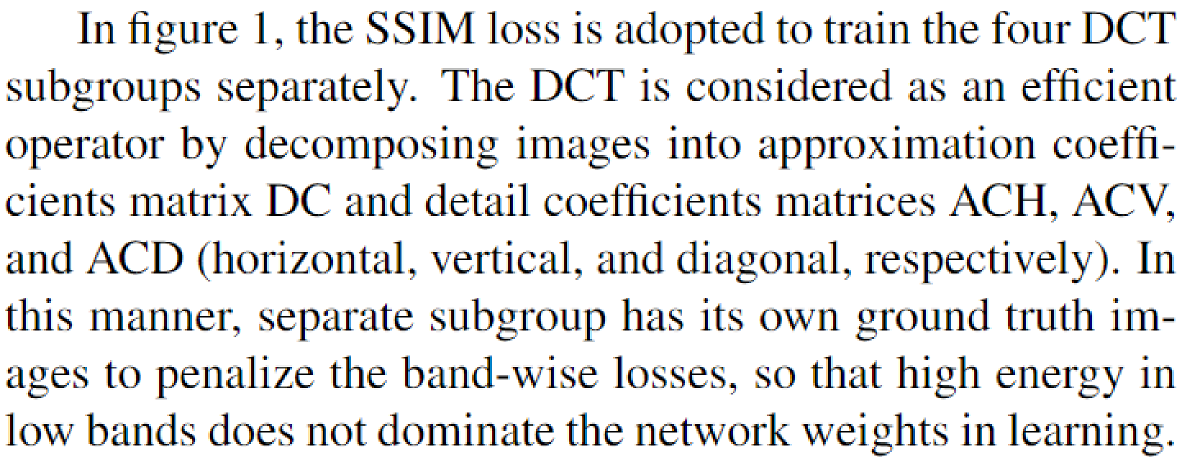
- The RCAN includes shallow feature extraction and deep feature extraction. For our network, three convolutional layers (Conv) are used to extract the shallow feature. The RIR module is used for getting deep features, which consists of several residual groups with long skip connections. Each residual group contains some residual blocks with short skip connections. Meanwhile, RIR module allows abundant low-frequency information to be bypassed through multiple skip connections, making the main network focus on learning high-frequency information. Furthermore, a channel attention mechanism is adopted, which can be viewed as a collection of the local descriptors, whose statistics contribute to express the whole image.

Experimental Setup

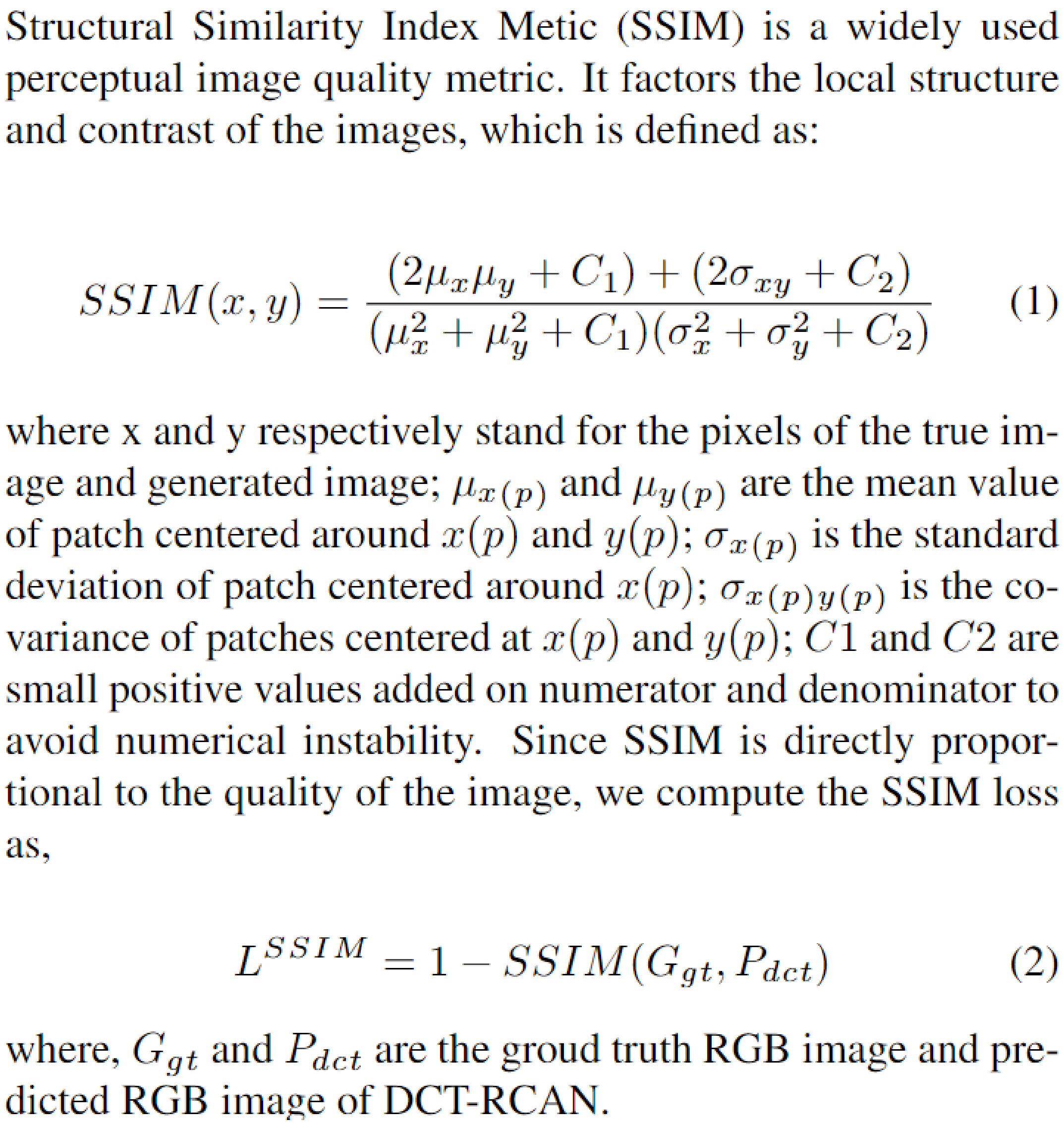


Some training configurations are given as follows. The batch size is 8. The number of epochs is 5000. We set the initial learning rate is 1×10^{-4}, and it will multiply 0.5 after every 500 epochs. Adam optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$. The patch size is 64×64. The batch size is set to 8. The regularization parameter λ in is set to 1. The kernel size of the first three and last convolution layers is 3. Zero padding is adopted in several convolution operations to keep the sizes of feature maps of the input and the output identical. We set the number of residual attention group to 6, the residual attention block is 10, the negative slope of leaky ReLU is 0.1, and the reduction ratio is 16.

Subgroup Recovery Loss Functions



In figure 1, the SSIM loss is adopted to train the four DCT subgroups separately. The DCT is considered as an efficient operator by decomposing images into approximation coefficients matrix DC and detail coefficients matrices ACH, ACV, and ACD (horizontal, vertical, and diagonal, respectively). In this manner, separate subgroup has its own ground truth images to penalize the band-wise losses, so that high energy in low bands does not dominate the network weights in learning.



Structural Similarity Index Metic (SSIM) is a widely used perceptual image quality metric. It factors the local structure and contrast of the images, which is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \qquad (1)$$

where x and y respectively stand for the pixels of the true image and generated image; $\mu_{x(p)}$ and $\mu_{y(p)}$ are the mean value of patch centered around $x(p)$ and $y(p)$; $\sigma_{x(p)}$ is the standard deviation of patch centered around $x(p)$; $\sigma_{x(p)y(p)}$ is the covariance of patches centered at $x(p)$ and $y(p)$; $C1$ and $C2$ are small positive values added on numerator and denominator to avoid numerical instability. Since SSIM is directly proportional to the quality of the image, we compute the SSIM loss as,

$$L^{SSIM} = 1 - SSIM(G_{gt}, P_{dct}) \qquad (2)$$

where, G_{gt} and P_{dct} are the groud truth RGB image and predicted RGB image of DCT-RCAN.

Performance Evaluation

- PSNR, SSIM, and AE are used to evaluate the quality of translated RGB images and the performance of the proposed network. Higher PSNR and SSIM values indicate that more important features, including richer textures and more vivid color information, are preserved in the generated results. In addition, the smaller the AE value, the more consistent the color information of the objects in the generated results. In these three quantitative indexes, translated RGB images are referred at as predicted images and corresponding RGB images are treated as ground truth images.

Experimental Results

We analyze the DCT-RCAN on the performance. The results of these three metrics on the same dataset are shown in Table 2. We observe that the network with DCT grately improves the performance of SSIM, AE, and PSNR, because the high frequency texture details and low frequency features are explicitly separated to four subgroups, which reduces the network training difficulty.

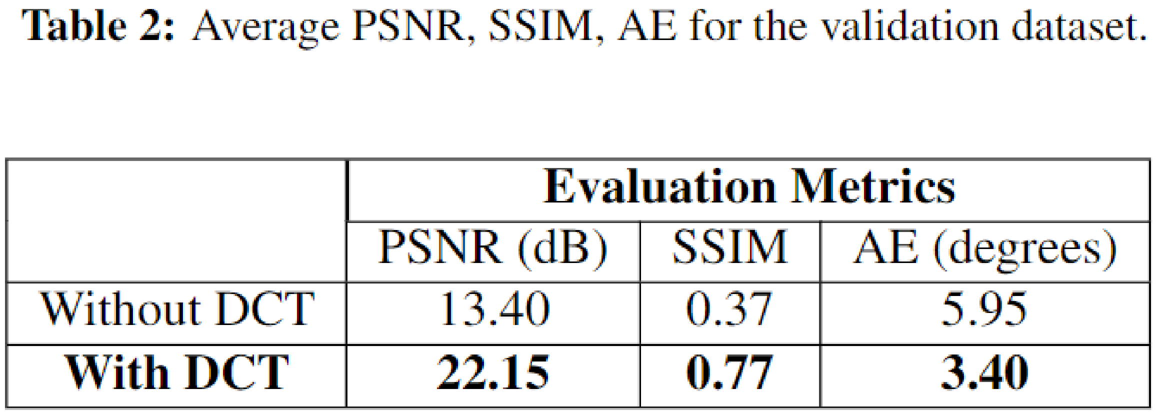


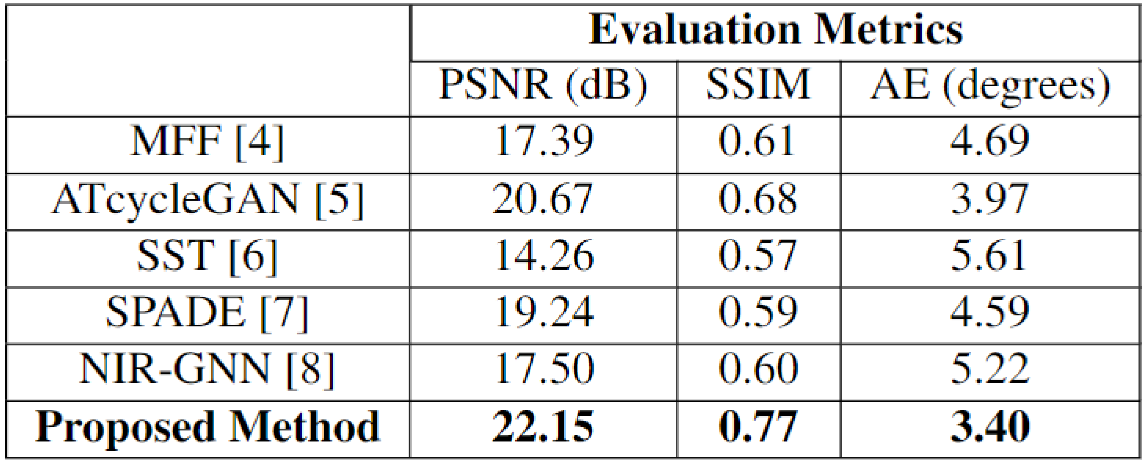
Table 2: Average PSNR, SSIM, AE for the validation dataset.

	Evaluation Metrics		
	PSNR (dB)	SSIM	AE (degrees)
Without DCT	13.40	0.37	5.95
With DCT	**22.15**	**0.77**	**3.40**

Experimental Results

It follows from Table 1 that PSNR and SSIM of our network are the biggest, and AE is the smallest in comparison with other five networks. In addition, the increased value about the PSNR and SSIM criterions achieve to 1.48 and 0.09 respectively, and the reduced value of AE criterion reach to 0.57 compared with the state-of-art result.

Table 1: Average PSNR, SSIM, AE for the validation dataset.



	Evaluation Metrics		
	PSNR (dB)	SSIM	AE (degrees)
MFF [4]	17.39	0.61	4.69
ATcycleGAN [5]	20.67	0.68	3.97
SST [6]	14.26	0.57	5.61
SPADE [7]	19.24	0.59	4.59
NIR-GNN [8]	17.50	0.60	5.22
Proposed Method	**22.15**	**0.77**	**3.40**

Experimental Results

- Visual comparison among different methods on validation dataset. In general, all methods can generate color images. However, the images generated by SST show that most of the colors are white, which cannot correctly display the real image. The results generated by ATCycleGAN have large areas of blur and a certain degree of color distortion. The images produced by NIR-GAN lack texture and color distortion. MFF and SPADE generate better results, but still fail to map the color information in the RGB domain correctly. Overall, our method generates results that are closer to the RGB ground truth, and the texture information from the NIR domain has been well transferred to RGB domain, color information is vivid and semantically correct.

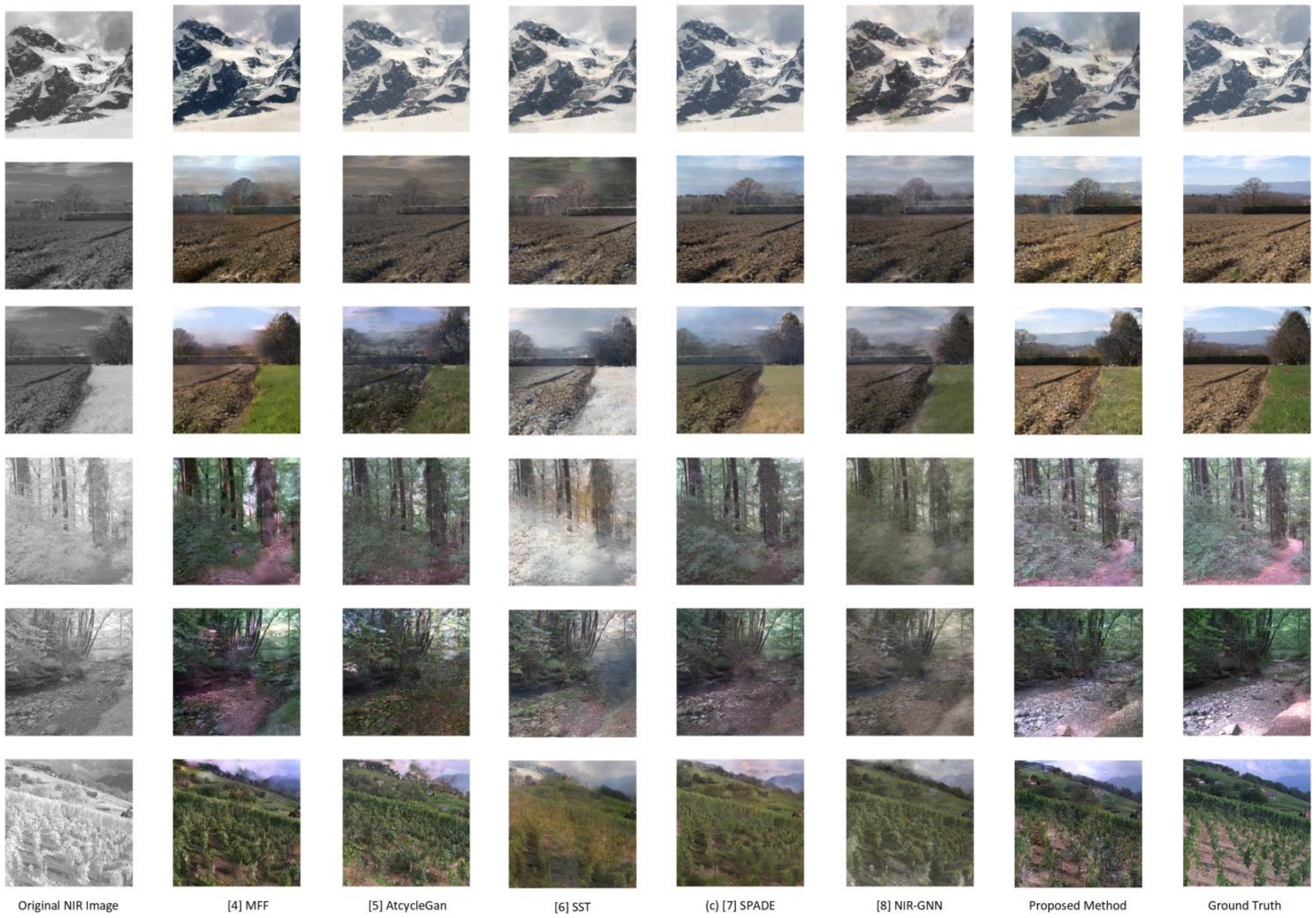


Figure 4. Comparison of the colorized NIR images

Conclusion

- We propose a DCT-based residual network (DCT-RCAN) for NIR image colorization. Instead of stacking large-scale models with a large amount of computation and enormous parameters, we adopt lightweight but an effective module RIR to improve the efficiency of NIR image colorization without sacrificing much performance. We use the 4×4 DCT to explicitly separate low-frequency and high-frequency details from one image to four subgroups during training. Thus, the learning difficulty of our network can be mitigated. Extensive experiments show that our DCT-RCAN requires relatively fewer parameters and generates competitive results against state-of-the-art methods quantitatively and qualitatively.