2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10221948

NERD: NEURAL FIELD-BASED DEMOSAICKING

Tomáš Kerepecký [1,2] , *Filip Šroubek* [1] , *Adam Novozámský* [1] , *Jan Flusser* [1]

[1] Institute of Information Theory and Automation, The Czech Academy of Sciences, Czechia
[2]Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Czechia

ABSTRACT

We introduce NeRD, a new demosaicking method for generating full-color images from Bayer patterns. Our approach leverages advancements in neural fields to perform demosaicking by representing an image as a coordinate-based neural network with sine activation functions. The inputs to the network are spatial coordinates and a low-resolution Bayer pattern, while the outputs are the corresponding RGB values. An encoder network, which is a blend of ResNet and U-net, enhances the implicit neural representation of the image to improve its quality and ensure spatial consistency through prior learning. Our experimental results demonstrate that NeRD outperforms traditional and state-of-the-art CNN-based methods and significantly closes the gap to transformer-based methods.

Index Terms— Demosaicking, neural field, implicit neural representation.

1. INTRODUCTION

Raw data acquired by modern digital camera sensors is subject to various types of signal degradation, one of the most severe being the color filter array. To convert the raw data (Fig. 1a) into an image suitable for human visual perception (Fig. 1b), a demosaicking procedure is necessary [1].

Two main categories of image demosaicking exist: model-based and learning-based methods. Model-based methods, such as bilinear interpolation, Malvar [2], or Menon [3], are still widely used, but they fail to match the performance of recent deep learning-based approaches using deep convolutional networks (CNN) [4, 5, 6] or Swin Transformers [7].

Recently, Transformer networks have seen remarkable success in computer vision tasks and have become a state-of-the-art approach in demosaicking. However, a new paradigm in deep learning, Neural Fields (NF) [8], is gaining attention due to its comparable or superior performance in several computer vision tasks [8, 9, 10, 11, 12, 13, 14]. The basic idea behind NF is to represent data as the weights of a Multilayer Perceptron (MLP), known as implicit neural representation.

This work was supported in part by the Czech Science Foundation grant GA21-03921S, the *Praemium Academiae* awarded by the Czech Academy of Sciences, and the Fulbright commission under the Fulbright-Masaryk award.

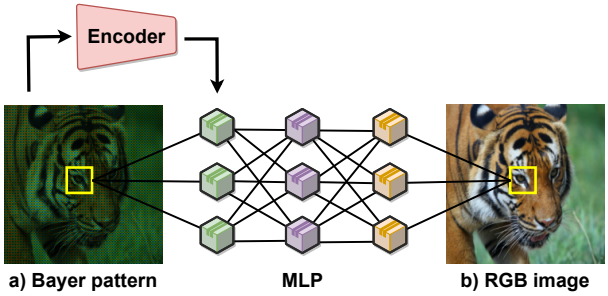


Fig. 1. An illustration of demosaicking using coordinate-based Multilayer Perceptron and local encoding technique.

NF has been applied in various domains and applications including Neural Radiance Fields (NeRF) [9] which achieved state-of-the-art results in representing complex 3D scenes. NeRV [11] encodes entire videos in neural networks. The Local Implicit Image Function (LIIF) [12] represents an image as a neural field capable of extrapolating to 30 times higher resolution. SIREN [13] uses a sinusoidal neural representation and demonstrates superiority over classical ReLU MLP in representing complex natural signals such as images.

Prior information from training data can be encoded into neural representation through conditioning (local or global) using methods such as concatenation, modulation of activation functions [15], or hypernetworks [14]. For example, CURE [10], a state-of-the-art method for video interpolation based on NF, uses an encoder to impose space-time consistency using local feature codes.

NF has also been used in image-to-image translation tasks such as superresolution, denoising, inpainting, and generative modeling [8]. However, to the best of our knowledge, no NF method has been proposed for demosaicking.

In this paper, we present NeRD, a novel approach for image demosaicking based on NF. The proposed method employs a joint ResNet and U-Net architecture to extract prior information from high-resolution ground-truth images and their corresponding Bayer patterns. This information is then used to condition the MLP using local feature encodings. The proposed approach offers a unique and innovative solution for image demosaicking.

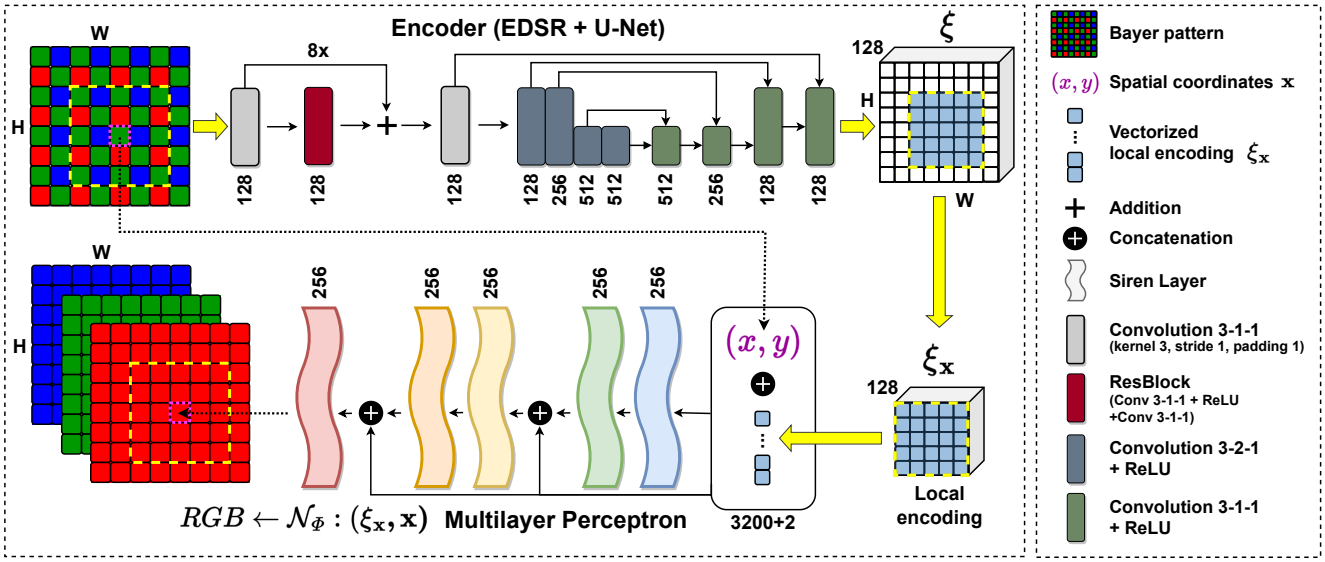


Fig. 2. The overall architecture of NeRD. Encoder consisting of 8 residual blocks and U-net architecture generates encoding ξ for the input Bayer pattern. Numbers below each layer in the encoder represent the number of output channels. Spatial coordinates $\mathbf{x} = (x, y)$ concatenated with the corresponding local encoding vector $\xi_{\mathbf{x}}$ are transformed into RGB value using a multilayer perceptron with 5 hidden layers each with 256 output channels, siren activation functions, and two skip connections.

2. PROPOSED METHOD

NeRD converts spatial coordinates and local encodings into RGB values. The local encodings are generated by an encoder that integrates consistency priors in NeRD. The overall architecture of NeRD is depicted in Fig. 2.

The core of NeRD is a fully connected feedforward network $\mathcal{N}_\Phi : (\xi_{\mathbf{x}}, \mathbf{x}) \rightarrow \mathbf{n}$ with 5 hidden layers, each with 256 output channels and sine activation functions. Φ denotes the network weights. The input is a spatial coordinate $\mathbf{x} = (x, y) \in \mathbb{R}^2$ and local encoding vector $\xi_{\mathbf{x}}$. The output is a single RGB value $\mathbf{n} = (r, g, b) \in \mathbb{R}^3$. The SIREN architecture [13] was chosen for its ability to model signals with greater precision compared to MLPs with ReLU. There are two skip connections that concatenate the input vector with the output of the second and fourth hidden layers.

Using the MLP without local encoding $\xi_{\mathbf{x}}$ leads to suboptimal demosaicking results due to the insufficient information contained in the training image. This is demonstrated by the result in Fig. 3-NeRD.0, where the reconstructed image is the output of the SIREN model trained only on original input Bayer pattern in self-supervised manner. The lack of spatial consistency in these results highlights the need for additional prior information in the form of spatial encoding, which is why we utilize an encoder.

The encoder provides local feature codes $\xi_{\mathbf{x}}$ for a given coordinate $\mathbf{x}$ and its architecture is shown in the first row of Fig. 2. The Bayer pattern is processed through a combined network that incorporates 8 residual blocks (using the EDSR architecture [16]) and 4 downsampling and 4 upsampling layers (U-Net architecture [17]) connected by multiple skip connections. The result is a global feature encoding $H \times W \times 128$, where H and W denote the height and width of the initial Bayer pattern in pixels. The local encoding $\xi_{\mathbf{x}}$ is extracted from the global encoding as a 5×5 region centered at $\mathbf{x}$, which is then flattened into a 3200-dimensional feature vector. The architecture of the encoder is adopted from [10].

The final RGB image is produced by independently retrieving the RGB pixel values from NeRD at the coordinates specified by the input Bayer pattern.

3. EXPERIMENT

We numerically validated NeRD on standard image datasets. Experiments also include an ablation study highlighting the key components of the proposed architecture and comparisons with state-of-the-art methods.

3.1. Dataset and Evaluation Metrics

A training set was created by combining multiple high-resolution datasets, such as DIV2K [18], Flickr2K [16], and OST [19], resulting in a total of 12 000 images. During each epoch, 10 000 randomly cropped patches of size 200×200 and corresponding Bayer patterns (GBRG) were generated. The Kodak and McM [20] datasets were used for testing.

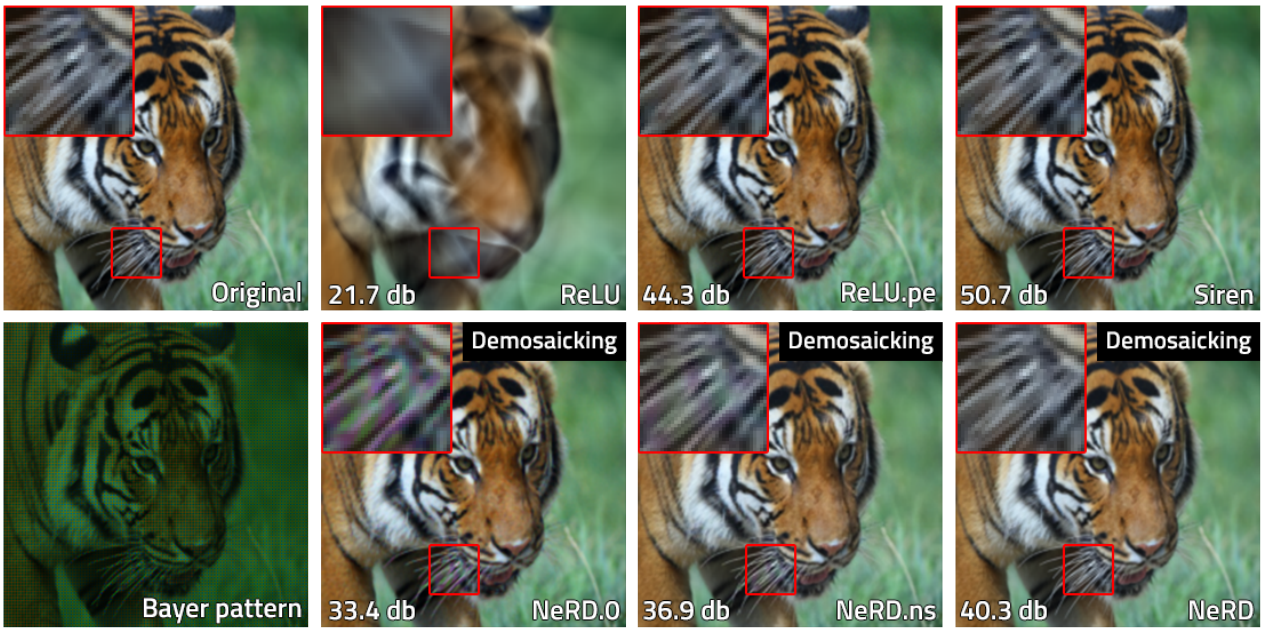


Fig. 3. The ablation study of NeRD. The original image is from DIV2K dataset. "ReLU" and "Siren" models show the implicit neural representation of the original image using MLP with ReLU and sine activation functions, respectively. These models were trained in a self-supervised manner to fit the original image. "ReLU.pe" stands for "ReLU" model with additional positional encoding in the form of Fourier feature mapping. "NeRD.0" model is identical to "Siren" model but is only trained using the input Bayer pattern. "NeRD" is the proposed demosaicking method, while "NeRD.ns" represents the proposed architecture without skip connections in the MLP. Each image is labeled with its PSNR value with respect to the original image.

The evaluation was performed using Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM).

3.2. Training Configuration

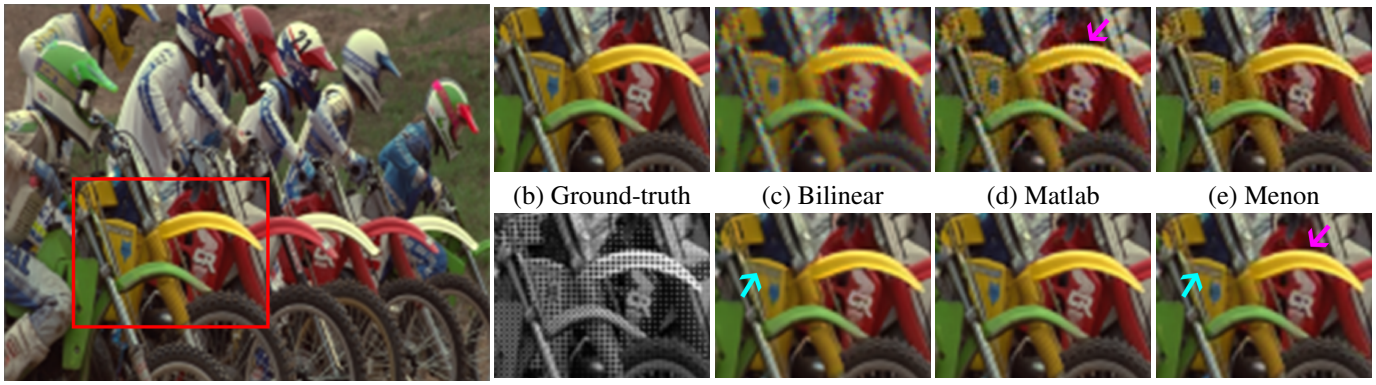
The training was conducted using an Nvidia A100 GPU. The NeRD model was optimized using the Mean Squared Error loss function, and the Adam optimizer was used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to 0.0001, and a step decay was applied, reducing the learning rate by 0.95 every epoch consisting of 10 000 iterations. The patch size was set to 200×200 and the batch size was 5.

3.3. Ablation Study

MLP and activation functions. RGB images can be represented as the weights of a fully connected feedforward neural network. This representation is achieved by training an MLP in a self-supervised manner to fit the original image. However, the usage of standard ReLU activation functions in MLPs produces unsatisfactory results, as shown in Fig. 3-ReLU. To significantly improve reconstruction, Fourier feature mapping of input spatial coordinates can be used (see Fig. 3-ReLU.pe). This technique is referred to as "positional encoding". Nonetheless, an even better outcome can be achieved by replacing ReLU with sine functions, also known as SIRENs. They demonstrate the capability of MLPs as image decoders and hold promise for demosaicking applications. SIREN architecture has the capacity to model RGB images with great precision. As demonstrated in Fig. 3-Siren, the SIREN with 5 hidden layers, each with 256 neurons, achieved a PSNR of 50.7 dB when trained for just 1000 iterations to fit the original image.

Encoder. The naive approach of decoding RGB images from Bayer patterns using SIREN architecture fails as it loses two-thirds of the original information, as shown in Fig. 3-NeRD.0. To improve the demosaicking capability of the MLP, prior information must be incorporated through an encoder. This encoder learns prior information across various training image pairs and conditions the MLP with local encodings. The effectiveness of the encoder is demonstrated in Fig. 3-NeRD, which shows the results of demosaicking using the NeRD architecture described in Sec. 2.

Skip Connections. The integration of encoding into the MLP can be achieved through various methods. However, methods such as modulation of activation functions or the use of hypernetworks present challenges in terms of parallelization. Hence, we utilized a method of concatenation, where the



(a) Kodak:kodim05.png (cropped) (b) Ground-truth (c) Bilinear (d) Matlab (e) Menon (f) Raw (g) DeepDemosaick (h) RSTCANet (i) NeRD

Fig. 4. A visual comparison of NeRD and the current state-of-the-art methods on an example from the Kodak dataset. The visual differences are highlighted by close-ups, which correspond to the red box in the original image. Although NeRD exhibits slightly inferior visual performance compared to RSTCANet, it outperforms traditional methods in terms of reconstruction accuracy (indicated by the magenta arrow) and avoids over-smoothing details, as seen with the DeepDemosaick method (indicated by the cyan arrow).

coordinates and feature vectors are combined at the input and later concatenation of the input with the second and fourth hidden layers is performed using skip connections. The significance of incorporating skip connections into the MLP is illustrated in Fig. 3-NeRD.ns (no-skip). This figure demonstrates a degradation in both the quality of the reconstruction and the PSNR value when these connections are omitted.

3.4. Comparison With Existing Methods

The evaluation of the proposed NeRD demosaicking algorithm was performed on the McM and Kodak datasets, which were resized and cropped to 200×200 px. A comparison of NeRD with traditional demosaicking algorithms and state-of-the-art methods is presented in Table 1 in terms of average PSNR and SSIM values calculated from the demosaicked images. The results show that NeRD outperforms traditional methods and the CNN-based DeepDemosaick [4], but falls slightly behind the transformer-based RSTCANet [7].

A visual comparison of the demosaicked images is presented in Fig. 4. The figure highlights differences between NeRD and the other methods and provides insights into their performance. One notable characteristic of NeRD is that it avoids over-smoothing details, unlike the DeepDemosaick [4] method, as indicated by the cyan arrow in the Fig. 4g. Furthermore, NeRD outperforms traditional methods in terms of preserving fine details and avoiding unpleasant artifacts, as indicated by the magenta arrow in the Fig. 4d.

Table 1. Average PSNR/SSIM obtained by NeRD and the current state-of-the-art methods on the McM* and Kodak* datasets (*resized and cropped to 200×200 px). **Bold** and underline highlights the highest and second highest values, respectively. Note the superior results of NeRD over the CNN-based and traditional methods. Only RSTCANet, which is based on transformers, has slightly higher scores.

Method	McM* [20] PSNR/SSIM	Kodak* PSNR/SSIM
Bilinear	27.15/0.912	28.01/0.894
Matlab (Malvar) [2]	30.54/0.923	33.52/0.957
Menon [3]	31.40/0.918	35.20/0.968
DeepDemosaick [4]	33.31/0.942	37.76/0.976
RSTCANet [7]	**37.77/0.978**	**40.84/0.988**
NeRD	36.18/0.969	39.07/0.984

4. CONCLUSION

This paper presents a novel demosaicking algorithm, NeRD, that leverages the recent class of techniques known as Neural Fields. The ablation study results emphasize the significance of incorporating an encoder and skip connections within the MLP, which results in significant improvement over traditional techniques and outperforms the CNN-based DeepDemosaick method in preserving fine details while avoiding undesirable artifacts. Although NeRD shows slightly lower visual performance compared to the transformer-based RSTCANet, it still demonstrates remarkable accuracy in terms of reconstruction. Future research can focus on enhancing NeRD through fine-tuning using input Bayer pattern-specific loss functions and integrating Transformer networks or ConvNeXt into the encoder. In addition, expanding the training set by more diverse datasets can improve the prior. Albeit NeRD may not attain the performance level of Transformer-based demosaicking, our contribution broadens the range of domains where Neural Fields can be applied.

5. REFERENCES

[1] Daniele Menon and Giancarlo Calvagno, "Color image demosaicking: An overview," *Signal Processing: Image Communication*, vol. 26, no. 8-9, pp. 518–533, 2011.

[2] Henrique S Malvar, Li-wei He, and Ross Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 3, pp. iii–485.

[3] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno, "Demosaicing with directional filtering and a posteriori decision," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 132–141, 2006.

[4] Filippos Kokkinos and Stamatios Lefkimmiatis, "Iterative joint image demosaicking and denoising using a residual denoising network," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 4177–4188, 2019.

[5] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 6, pp. 1–12, 2016.

[6] Tomáš Kerepecky and Filip Šroubek, "D3net: Joint demosaicking, deblurring and deringing," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1–8.

[7] Wenzhu Xing and Karen Egiazarian, "Residual swin transformer channel attention network for image demosaicing," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[8] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*. Wiley Online Library, 2022, vol. 41, pp. 641–676.

[9] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[10] Wentao Shangguan, Yu Sun, Weijie Gan, and Ulugbek S Kamilov, "Learning cross-video neural representations for high-quality frame interpolation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 2022, pp. 511–528.

[11] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21557–21568, 2021.

[12] Yinbo Chen, Sifei Liu, and Xiaolong Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8628–8638.

[13] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[14] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny, "Adversarial generation of continuous images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10753–10764.

[15] Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum, "From data to functa: Your data point is a function and you should treat it like one," *arXiv preprint arXiv:2201.12204*, 2022.

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[18] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[19] Chao Dong Xintao Wang, Ke Yu and Chen Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li, "Color demosaicking by local directional interpolation and nonlocal adaptive thresholding," *Journal of Electronic imaging*, vol. 20, no. 2, pp. 023016–023016, 2011.