

UNSUPERVISED REMOTE SENSING SUPER-RESOLUTION VIA MIGRATION IMAGE PRIOR

Jiaming Wang¹, Zhenfeng Shao¹, Tao Lu^{2*}, Xiao Huang³, Ruiqian Zhang¹, and Yu Wang²

¹ Wuhan University, ² Wuhan Institute of Technology, ³ University of Arkansas
 {wjmecho, shaozhenfeng, zhangruiqian}@whu.edu.cn, lutxyl@gmail.com
 xh010@uark.edu, wangyu949374585@gmail.com

ABSTRACT

Recently, satellites with high temporal resolution have fostered wide attention in various practical applications. Due to limitations of bandwidth and hardware cost, however, the spatial resolution of such satellites is considerably low, largely limiting their potentials in scenarios that require spatially explicit information. To improve image resolution, numerous approaches based on training low-high resolution pairs have been proposed to address the super-resolution (SR) task. Despite their success, however, low/high spatial resolution pairs are usually difficult to obtain in satellites with a high temporal resolution, making such approaches in SR impractical to use. In this paper, we proposed a new unsupervised learning framework, called “MIP”, which achieves SR tasks without low/high resolution image pairs. First, random noise maps are fed into a designed generative adversarial network (GAN) for reconstruction. Then, the proposed method converts the reference image to latent space as the migration image prior. Finally, we update the input noise via an implicit method, and further transfer the texture and structured information from the reference image. Extensive experimental results on the Draper dataset show that MIP achieves significant improvements over state-of-the-art methods both quantitatively and qualitatively. The proposed MIP is open-sourced at <https://github.com/jiaming-wang/MIP>.

Index Terms— Super-resolution, unsupervised learning, latent space, deep neural networks

1. INTRODUCTION

Recently, remote sensing satellites, which are especially appropriate for uninterrupted observing targets, have drawn widespread concerns in various practical applications. Continuously monitoring moving targets by high temporal resolution satellites, can expand the application range than satellites

with a static image, such as, the Jilin-1 and Zhuhai-1 OVS-1 A/B video satellites. It is common knowledge that the spatial resolution and spectral resolution are always a pair of contradictory for the optical remote sensor. Additionally, due to bandwidth and hardware cost limitations, the spatial resolution of high temporal resolution satellite images is decreased that cannot meet the demand of high precision applications. Therefore, improving the spatial resolution of satellite images with large compression ratios, has become an urgent issue in remote sensing applications.

Super-resolution (SR) aims to reconstruct the high spatial resolution (HSR) image from observed low spatial resolution (LSR) images [1], which breaks the limitations of the imaging system for the best cost/benefit ratio. In real-world remote sensing scenarios, the SR problems often have the following properties: 1) HSR and high temporal resolution (HSR-HTR) datasets are unavailable, 2) HSR and low temporal resolution images (HSR-LTR), which enjoy the same image content with LSR high temporal resolutions imageries (LSR-HTR), are easy to obtain as the reference, 3) there are obvious differences between the imaging environments of the LSR-HTR and HSR-LTR images, which makes it difficult to transfer texture directly.

Existing SR methods tend to generate HSR images/patches from the prior information provided by datasets. The development of machine learning promotes the progress of SR. Traditional supervised deep-learning-based SR methods obtain excellent performance by designing a network to extract deep features, *i.e.*, enhanced deep residual networks [2] and residual dense network [3]. The rationale of these algorithms can be summarized as follows: a deep model learns the mapping between the corrupted LSR images and HSR ones by a convolutional neural network (CNN), and the LSR images are degraded from their original versions. Although these methods are intended for obtaining deep features from image prior information, they have the main disadvantage: they require HSR training examples, which are limited by economic and technical reasons in remote sensing [4].

Under the above circumstances, unsupervised image SR has received more attention. To exploit the prior structure,

This work was supported in part by the National key R & D plan on strategic international scientific and technological innovation cooperation special project under Grant 2016YFE0202300, the National Natural Science Foundation of China under Grants 61671332, 41771452, 51708426, 41890820, 62072350, and 41771454, the Natural Science Fund of Hubei Province in China under Grant 2018CFA007, the Independent Research Projects of Wuhan University under Grant 2042018kf0250.

information in the residual domain, which effectively enhanced the high-frequency information. Jiang *et al.* [14] proposed a generative adversarial network (GAN) based edge-enhancement method that can generate clean and sharp details. Haut *et al.* [4] firstly proposed an unsupervised hour-glass model to super-resolved LSR remote sensing images from random noise. However, it is difficult to recover the high-frequency information of the image from the existing image prior.

2.2. Reference-based Image Super-Resolution

Different from single image super-resolution (SISR) methods, RefSR algorithms provide more accurate and realistic details, which are transferred from the reference image (the reference is similar to LSR one in content, but with different focal lengths and shot perspectives).

Considering the incomplete coupling of the LSR and reference image, some algorithms [6], [7] achieved great performance when they are tightly aligned. This means they only swap the information in the image level. In view of this, Zhang *et al.* [15] proposed a deep model and adopted local texture matching for long-distance dependency. Most recently, Yang *et al.* [16] introduced a more accurate way to search and transfer relevant textures from Ref to LSR images. SSEN [17] aligned the Ref and LSR images in the feature domain to capture similarity-aware. In general, these deep methods achieve better results than SISR methods.

However, the improvements of RefSR methods [18], [19] rely on lots of training images. At the same time, due to different synthetic bands, some satellite images used in this paper (Draper) show different visual characteristics. It is difficult to transform the reference image feature into the input image.

3. OUR METHOD

3.1. Problem Formulation

Focusing on the primary goal of SR, to recover the high-frequency information from LSR-HTR images $I^{LSR-HTR} \in \mathbb{R}^{C \times H \times W}$ and obtain an HSR-HTR version image $I^{HSR-HTR} \in \mathbb{R}^{C \times t \cdot H \times t \cdot W}$, the conventional formulation of SR methods is $I^{LSR-HTR} = D I^{HSR-HTR}$, where D denotes the down-sampling matrix, and t is the factor. Here, we assume that the paired HSR-LSR high-temporal training data are unavailable, which makes them with simulated paired data impractical. Nevertheless, we can obtain a set of HSR low-temporal reference images that can be used for unsupervised training. Rather than minimizing the error between the SR images and the ground truth in the supervised method, the proposed method is based on the reference image to the texture and content information. Therefore, the key issue of the proposed method is to explore a unified framework to fuse the information at different times.

The pipeline of MIP is summarized as Fig. 2. We denote $I^{Ref} \in \mathbb{R}^{C \times t \cdot H \times t \cdot W}$ the corresponding HSR refer-

ence image $I^{HSR-HTR}$. The random noise maps n_{init} is $C' \times t \cdot H \times t \cdot W$. The proposed method mainly consists of three parts: the generative network, the reference feature extraction network and the migration image prior model. First, we learn a mapping from noise maps to an HR image. Second, we adopt an encoder-decoder model to code and transform the reference image. In the end, we map the coded feature maps of the reference image into the latent space, and update the random noise n_{init} . Details are given in the following.

3.2. Image Generation

Different from GAN-based image generation tasks, SR requires the result as real, not just a high-quality image. If we directly apply image generation [20] or GAN-based SR [21] models, we need to up-sample the input, which will also cause the obvious checkerboard phenomenon in the unsupervised framework.

Given an input HSR-sized noise maps to generate an image. Hence, it can be formulated as,

$$I^{SR} = H(n_{init}), \quad (1)$$

where n_{init} is the noise maps, and $H(\cdot)$ denotes the function of the SR network in the proposed method. I^{SR} refers to the output of the SR network. In this paper, we adopt stacked skip models for reconstructing, as shown in Fig. 3. For the proposed skip model, the first extract the shallow feature maps and concatenate it with deep-level features. Another advantage of the skip model is that it can reduce the cost of calculation than the densely connected convolutional network [22].

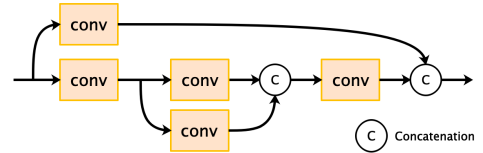


Fig. 3: Illustration of the skip model in the generative network.

3.3. Reference Feature Extraction

Traditional reference-based supervised approaches all try to design deep networks, and align the reference and SR images in the feature domain [15], [16]. Therefore, the performance of these algorithms is highly dependent on the pixel by pixel supervision. Considering the universal rigid transformation in satellite images, which as taken at different times, the images will have a great difference in shooting angle. In this paper, we advocate an encoder-decoder-based model to exploit the prior of the reference image.

Different from previous, which transform the local features in the reference image into the SR image, we employ spatial transformer networks (STN) [9] to improve the invariance of the affine transformation of a CNN network. It is

trained with learnable localisation and grid, as the affine transformation matrix. Then, the image sampling function is used to sample feature maps, and merge with them into a spatial transformer. In this work, we leverage the STN block for the transcoding process.

The spatial transformer network is defined as follows:

$$\begin{bmatrix} \mathbf{x}^{output} \\ \mathbf{y}^{output} \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{input} \\ \mathbf{y}^{input} \\ 1 \end{bmatrix}, \quad (2)$$

where $(\mathbf{x}^{input}, \mathbf{y}^{input})$ is the coordinates in the input feature maps, $(\mathbf{x}^{output}, \mathbf{y}^{output})$ is the coordinates in the output maps, and θ denotes the 2D transformation parameters. MIP consists of several STN blocks as shown in Fig. 2. MIP gradually aligns reference features in each scales. These blocks facilitates the method to transform feature maps for semantic aligning.

3.4. Image Prior Migration

We then investigate how to explore the prior in a reference image. Considering the weak supervision in this framework than traditional approaches, it is difficult to transform texture from the reference with a great difference. In the previous work [23], the authors introduce the influence of latent space on the generated results, which laid the solid foundation for fine image generation. Both target and attribute in the results can be mapped to noise vector. InfoGAN [24] decomposes the input noise vector into random noise and the latent code, which can target the structured features. With fixed noise, InfoGAN learns interpretable representations by manipulating latent code.

In this paper, we now propose a method for implicit updating: we convert the feature maps of the reference image into latent space, which carries structural information and code for texture generation. And then the code is used to generate the same semantic targets and attributes. It can be formulated as,

$$\mathbf{n} \leftarrow \Psi(\mathbf{n}_{init}, \mathbf{f}^{Ref}), \quad (3)$$

where \mathbf{f}^{Ref} is the feature maps of the reference image. In particular, the input noise is updated:

$$\mathbf{f}(x) = \frac{1}{\sqrt{2\pi} \text{std}(\mathbf{f}^{Ref})} \exp\left(-\frac{(x - \text{mean}(\mathbf{f}^{Ref}))^2}{2 \text{std}(\mathbf{f}^{Ref})^2}\right), \quad (4)$$

where $\text{std}(\cdot)$ denotes the standard deviation function, and $\text{mean}(\cdot)$ is the mean function. $\mathbf{f}(x)$ is the hidden space matrix generated by the Gaussian function, which conducts the migration image prior from the reference image. Then, the input can be viewed as the combination of the initialized noise and the latent code.

$$\mathbf{n}_{i+1} = \mathbf{n}_{init} + \alpha \cdot \mathbf{f}(x), \text{ and } \mathbf{n}_1 = \mathbf{n}_{init}, \quad (5)$$

where i is the number of iterations. The updated noise maps will be used for the input of the generation network. As a matter of experience, α is 0.03.

3.5. Loss Function

In the current literature, the goal of supervised SR is to generate an HSR image/patch from LSR one, and minimize the error in the HSR space. Considering the lack of ground truth, we downsample the SR image \mathbf{I}^{SR} by the Lanczos resampling [25] function as [4], and minimize the mean squared error (MSE) between it with the LSR-HTR image $\mathbf{I}^{LSR-HTR}$ in LSR domain. This process can be described as follows:

$$\begin{aligned} \mathcal{L}(\theta, S) &= \|\mathbf{I}^{LSR-HTR} - \text{down}(\mathbf{I}^{LSR-HTR})\|_2 \\ &= \|\mathbf{I}^{LSR-HTR} - \mathbf{I}^{LSR-HTR'}\|_2, \end{aligned} \quad (6)$$

where θ denotes the parameters in the proposed method, and S is the training data. $\mathbf{I}^{LSR-HTR'}$ denotes the LSR-sized version SR image. The Lanczos kernel can be described as follows:

$$L(x) = \frac{3 \sin(\pi x) \sin(\pi x/3)}{\pi^2 x^2}, \quad (7)$$

where x in the input pixel.

4. EXPERIMENTS

4.1. Datasets

The Draper dataset¹ is a publicly available benchmark for remote sensing image ordering in southern California, including 324 scenarios with 5 images in each scenario. The photographs were captured from a plane as a reasonable facsimile for satellite images, which were taken at different times. The HR image size is $3,099 \times 2,329$ pixels. We randomly select two sets of images (five images in a group) from this dataset, and name them “Day 1”, “Day 2”, “Day 3”, “Day 4”, and “Day 5”. It is noteworthy that this is not necessarily a consecutive time. We select 115 LSR version images from “Day 5”, and corresponding ones in “Day 4” as the reference images.

4.2. Implementation Details

All the models presented in this paper are trained with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 8$. Each mini-batch contains one noise map with size 192×192 and the Ref patches with size 192×192 . We initialize learning rate to $1e - 4$. We set the spectral bands of noise $C' = 32$. These experiments run at a desktop with two NVIDIA GTX 2080Ti GPUs and 3.60 GHz Intel Core i7-7820X CPU, 32GB memory. We implement the proposed method using PyTorch 1.1.0 library with Python 3.5.6 under Ubuntu 18.04, CUDA 10.1, and CUDNN 7.5 systems. We train the model over 10000 iterations, until it converges.

¹<https://www.kaggle.com/c/draper-satellite-image-chronology/data>

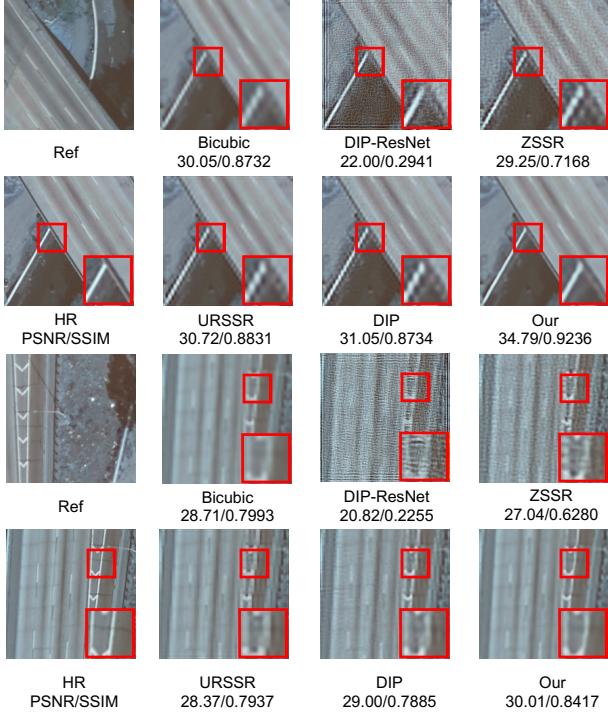


Fig. 4: Visual comparison among different SR methods on draper dataset with scale factor $\times 4$. We report the PSNR (dB), and SSIM results of the competing methods. The proposed method achieves state-of-the-art performance.

Evaluation measures. Four widely used image quality assessment indices are employed to evaluate the performance, including peak signal to noise ratio (PSNR), structural similarity (SSIM), visual information fidelity (VIF) [26], and erreur relative globale adimensionnelle de synth se (ERGAS).

4.3. Comparison with Unsupervised Methods

We compare the results of the proposed method with those of state-of-the-art unsupervised SR methods, DIP [5]², URSSR [4], and ZSSR [27]³, among which URSSR [4] is considered to achieve state-of-the-art performance in remote sensing image SR. DIP [5] has achieved the state-of-the-art visual quality, even if compared with the supervised SR algorithm. All experiments are performed with $\times 4$. For a fair comparison, all methods are trained with only the input image.

Table 1 shows the average performance of the PSNR, SSIM, VIF, and ERGAS results of competing methods with $\times 4$ on the draper dataset. Clearly, the proposed MIP framework outperforms all other competing methods. On average, the PSNR and SSIM values of the proposed MIP framework for upsampling factor $t = 4$ are 0.92/1.26 dB and 0.0274/0.0293 higher than the second-best method, respec-

²<https://github.com/DmitryUlyanov/deep-image-prior>

³<https://github.com/assafshocher/ZSSR>

Table 1: Average quantitative comparisons of different approaches with scale factor $\times 4$.

Method	PSNR \uparrow	SSIM \uparrow	VIF \uparrow	ERGAS \downarrow
Bicubic	28.79	0.7910	0.4018	1.6029
DIP-ResNet	14.36	0.1964	0.0812	7.9412
ZSSR	28.99	0.7487	0.3194	1.5508
URSSR	29.29	0.8095	0.3996	1.5215
DIP	29.63	0.8114	0.3869	1.4530
Ours	30.55	0.8388	0.4453	1.3266

tively.

Several subjective results with upsampling factors $t = 4$ are illustrated in Fig. 4. From the visual reconstruction results, we can see that DIP-ResNet [5] and ZSSR [27] achieve not only shape edge (high-frequency information), but also a lot of noise. DIP [5] and URSSR [4], which are designed for the unsupervised SR, fail to generate stable and touching detail information. We think this is mainly due to the limitation of the unsupervised SR task. Our method produces sharper edges and finer details than the other methods.

5. CONCLUSION

In this paper, we introduce an unsupervised reference-based image SR method termed Migration Image Prior (MIP). In particular, in order to solve the problem of missing high spatial resolution data of high temporal resolution images, we carefully design an end-to-end framework to fully exploit the available high spatial resolution image as the reference. In addition, we adopt a novel way to update input noise, which is used to generate a corresponding high-resolution image. In this way, we encode the reference image into the latent space as the migration image prior, and update noise maps to obtain stable results. Experimental results on the public dataset demonstrate that our method achieves state-of-the-art performance quantitatively and qualitatively.

6. REFERENCES

- [1] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.
- [2] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1132–1140.
- [3] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

- [4] Juan Mario Haut, Ruben Fernandez-Beltran, Mercedes E. Paoletti, Javier Plaza, Antonio Plaza, and Filiberto Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6792–6810, 2018.
- [5] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov, "Deep image prior," in *Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
- [6] Yuwang Wang, Yebin Liu, Wolfgang Heidrich, and Qionghai Dai, "The light field attachment: Turning a dsr into a light field camera using a low budget camera ring," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 10, pp. 2357–2364, 2017.
- [7] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *European Conference on Computer Vision*, 2018, pp. 87–104.
- [8] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and Garrison W Cottrell, "Event-specific image importance," in *Computer Vision and Pattern Recognition*, 2016, pp. 4810–4819.
- [9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *Neural Information Processing Systems*, 2015, pp. 2017–2025.
- [10] M.T. Merino and J. Nunez, "Super-resolution of remotely sensed images with variable-pixel linear reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1446–1457, 2007.
- [11] Shuyuan Yang, Fenghua Sun, Min Wang, Zhizhou Liu, and Licheng Jiao, "Novel super resolution restoration of remote sensing images based on compressive sensing and example patches-aided dictionary learning," in *International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping*. IEEE, 2011, pp. 1–6.
- [12] Yimin Luo, Liguang Zhou, Shu Wang, and Zhongyuan Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2398–2402, 2017.
- [13] Tao Lu, Jiaming Wang, Yanduo Zhang, Zhongyuan Wang, and Junjun Jiang, "Satellite image super-resolution via multi-scale residual deep neural network," *Remote Sensing*, vol. 11, no. 13, pp. 1588, 2019.
- [14] Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019.
- [15] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi, "Image super-resolution by neural texture transfer," in *Computer Vision and Pattern Recognition*, 2019, pp. 7982–7991.
- [16] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, "Learning texture transformer network for image super-resolution," in *Computer Vision and Pattern Recognition*, 2020, pp. 5791–5800.
- [17] Gyumin Shim, Jinsun Park, and In So Kweon, "Robust reference-based super-resolution with similarity-aware deformable convolution," in *Computer Vision and Pattern Recognition*, 2020, pp. 8425–8434.
- [18] Shunta Maeda, "Unpaired image super-resolution using pseudo-supervision," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 291–300.
- [19] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [21] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Computer Vision and Pattern Recognition*, 2017, pp. 105–114.
- [22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [23] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [24] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Neural Information Processing Systems*, 2016, pp. 2172–2180.
- [25] Ken Turkowski, "Filters for common resampling tasks," in *Graphics gems*. Academic Press Professional, Inc., 1990, pp. 147–165.
- [26] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [27] Michal Irani Assaf Shocher, Nadav Cohen, "zero-shot" super-resolution using deep internal learning," in *Computer Vision and Pattern Recognition*, June 2018.