Deep learning on super resolution : A Survey

Yujin CHO, *Master student, University Paris-saclay*

***Abstract*—As high-resolution displays such as Ultra-High Definition (UHD) emerge and consumer demand for super-resolution increases, interest in research that can convert low-resolution images into high-resolution images is growing. Super resolution refers to the need to restore a low-resolution image to a high-resolution image. It is also called a regular inverse problem or an Ill-Posed problem in which there are multiple correct answers for a high-resolution image that is a target to be restored. There are many studies on converting to high resolution using a single image (Single Image Super Resolution-SISR) or research using multiple images (Multi Image Super Resolution-MISR). Mostly, research on SISR is dominated, and this survey introduces the latest technology status and trends in SISR algorithm research using deep learning. We will analyze representative networks and make appropriate comparisons from various perspectives. Finally, we conclude this review with some current challenges remained and future trends in SISR.**

I. Introduction

SINGLE Image super resolution(SISR) is a branch of computer vision research, a technology that generates high-resolution images(HR) from low-resolution images(SR). This field has made a many progress in the past 5 years by introducing deep learning algorithms. Prior to the application of deep learning, polynomial-based interpolation methods such as Bicubic interpolation, or local patch-based super-resolution techniques using linear mapping were widely studied. The interpolation method literally makes the image larger, and it can be confirmed that there is blur or deterioration in detail. A super-resolution algorithm that restores a high-resolution patch directly from a low-resolution patch that maps to a linear mapping. It is quite complex and produces complex high-resolution images with computational complexity. However, since this method is based on linear mapping, it is difficult to implement a complex and nonlinear model. In order to solve this problem, a deep learning method was introduced, and a convolutional neural network was introduced to show higher performance compared to existing algorithms. It learns a complex nonlinear relationship between low-resolution input and high-resolution output using a multi-layered network stacked in several layers. In 2014, Dong first applied deep learning to super resolution, which is called SRCNN [1]. This is a network with a simple three-layer fully convolutional network structure, and is an algorithm that improves the quality of an image magnified by bicubic interpolation. After that, Dong proposed the Fast Super Resolution Convolutional Neural Network (FSRCNN) [2], a network that increases the performance while reducing the number of filter parameters by reducing the weight of the network. By increasing the number of convolution layers and applying the deconvolution operation at the end, the amount of computation is greatly reduced and the accuracy (PSNR) is also improved. Although all early papers used deep learning, it can be seen that it is somewhat different from the direction that the more the number of convolution layers, the better the accuracy. A paper that proposed a deep structure by breaking down these barriers and designing a relatively deep network of 20 layers was published in 2016. [4] Inspired by ResNet, it can efficiently learn deeper networks through residual learning techniques. From 2017, the study focused on super-resolution algorithms using the Generative Adversarial Network(GANs). [8] After that, several studies have shown higher performance by suggesting various networks that are applied to with the introduction of the super-resolution algorithm. In this review, we would like to look at the current state and trend of the latest technology for performing a single image super-resolution algorithm based on deep learning. Finally, the trends and challenges are summarized by comparing different algorithmic techniques from different perspectives.

Y. Cho is with the Department of Electrical Engineering, University Paris-Saclay, France, e-mail: amycho92@naver.com

II. Related works

A. Convolutional Neural Networks

SRCNN

Super resolution is to recover Low resolution images to High resolution images, but there are many ways to recover in different way. It is called Regular Inverse Problem or ill-Posed Problem that there are no unique answers as shown in Figure 1. SRCNN(Super-Resolution Convolutional Network) [1] is the first convolutional neural network proposed by Dong to deal with Single Image Super Resolution(SISR) in 2014. At the time when CNN was just introduced, so only 3 convolutional layers were used without stacking hundreds of layers, and it showed higher performance values compared to methods without deep learning. (Figure 2). This study shows the possibility that deep learning can be applied to the super-resolution field as well. When constructing an architecture, the meaning of each convolutional layer is interpreted in terms of traditional super-resolution, and each layer is in charge of patch extraction, non-linear mapping, and reconstruction. From a given low resolution image, first convolution layer extracts feature maps. The Second layers map non linearly to high resolution patch problems. The last layer combines the prediction to produce high-resolution output (Figure 3).

FRCNN

FRCNN(Fast Super Resolution Convolutional Neural Networks) is a follow-up work published by the author of SRCNN. It points out that inefficient computation occurs in the process of convolution after interpolating LR images into HR images in SRCNN, and suggests a method to improve this. It uses the method of inserting the LR image entering the input into the convolution layer as it is, and finally, the HR

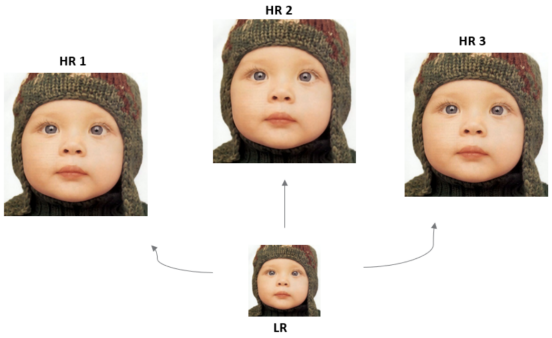

Figure 1: : Single Image Super Resolution(SISR) is to use a single image with low resolution and convert into high resolution. As described in the figure, there are no specific answers to recover into high resolution. It means an undefinable problem in which the only correct answer cannot exist, and this case is called a regular inverse problem or an ill-posed problem.

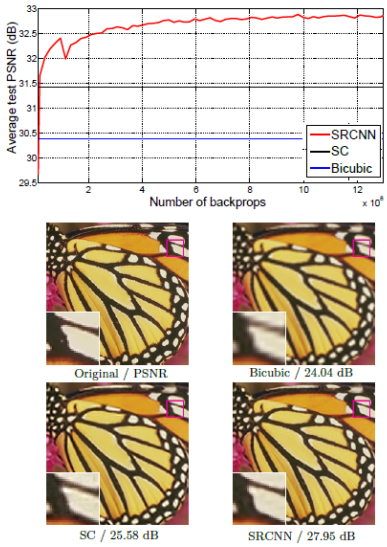

Figure 2: SRCNN results compared with example based methods.It shows SRCNN outperforms than other methods(Bicubic, SC(Sparse-Coding-based method). PSNR represents Peak Signal-to-Noise-Ratio.

image is made using a deconvolution operation that increases the horizontal and vertical size of the feature map. When the LR image is subjected to a convolution operation, the amount of calculation decreases in proportion to the square of the multiple. As the number of computations decreases, the number of convolution layers increases, and the PSNR can be taken care of. (Figure 3)

ESPCN

ESPCN(Efficient Sub-Pixel Convolutional Neural Network)is a method to increase the resolution of an image directly from LR to HR unlike other conventional method upscaling from LR. Here we apply Gaussian filter to the HR image to obtain LR image and perform mapping the LR image to HR image through the last sub-pixel layer. While introducing sub-pixel convolutoinal layer, the author points out the problem of the large amount of computation of deconvolution layer. Deconvolution is a method of applying convolution to the layer

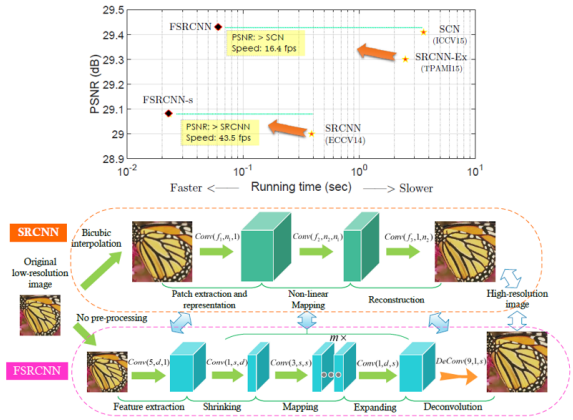

Figure 3: FRCNN architecture versus SRCNN architecture

after zero -padding and increasing the size of the feature map. For example if we want to upscale r times, the size of feature map is $r \times r$ at the last layer. By arranging feature maps in order we can get one channel of HR image. Through this method, the amount of computation can be reduced and better accuracy can be obtained.

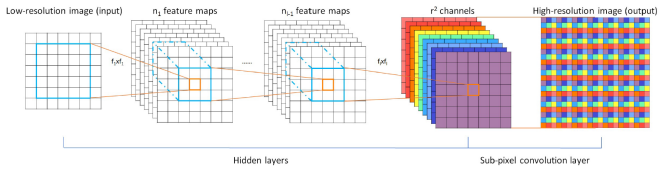

Figure 4: Sub-pixel convolution

B. Residual Networks

VDSR

In general, if deep learning model is designed with very deep network, the number of computation increases when the number of parameters increases. However, during deep network learning, gradient vanishing problems or gradient exploding problems may occur as it goes to the input layer in the backpropagation process, so that filter parameters are not properly learned. Simply designing deeply does not increase the performance of the network proportionately. To solve this problem, Kim proposed a new learning method called residual learning in 2016 and proposed VDSR(Very Deep Convolutional Networks), which can improve performance without causing problems even in deep networks. [4] Residual learning is a method of adding the LR image to the final output HR image and learning the difference value between the two images. In general, since the LR image and the output HR image are similar, the difference value is very small or 0, so the gradient vanishing or exploding problem could be solved. In this work, VGG [6] based convolutional networks are used with 20 layers. As described before, the residual learning method is used which add input images into the final output. Notably, gradient clipping is also performed to ensure good convergence by using a high learning rate at the beginning. The result of synthesizing these methods shows that the accuracy can be improved compared to the

existing methods, and learning can also be converged well. This work greatly influenced the super resolution research that were studied later.

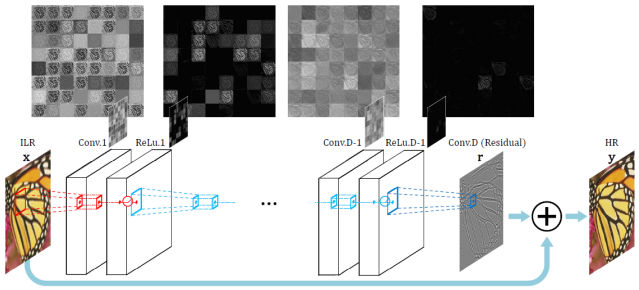

Figure 5: VDSR architecture. It cascades many pairs of layers repeatedly. An interpolated low-resolution(ILR) images goes through layers and transform into a high-resolution(HR) image. The network predicts a residual image and the addition of IRL and the residual gives the desired output. [4]

ESDR

Lim proposed EDSR(Enhanced Deep Convolutional Network) [7] network using residual blocks. It has more than 32 layers and the number of channels has increased more than four times compared to other existing networks, so the number of parameters has increased proportionately. All convolution have the same feature size with 3x3. (Figure 6) To learn a deep network stably, the author divides the network by residual block and designs the network so that filter parameters are more easily optimized using skip connection. To solve the problem that learning is difficult due to the case that the variance of the feature maps increases as feature maps are added in each residual block, a Multi-layer that multiplies a certain constant value after the CNN layer is added. Noteworthy is that all the Batch Normalization layers used in the existing Residual block have been removed. Batch Normalize reduced the flexibility of the network by normalizing the features. By removing this, the network could operate flexibly. It is said that it was possible to reduce memory usage. In order to increase the performance additionally, they reused pre-trained model which have trained to increase the resolution by 2 times to improve resolution by 3 and 4 times.

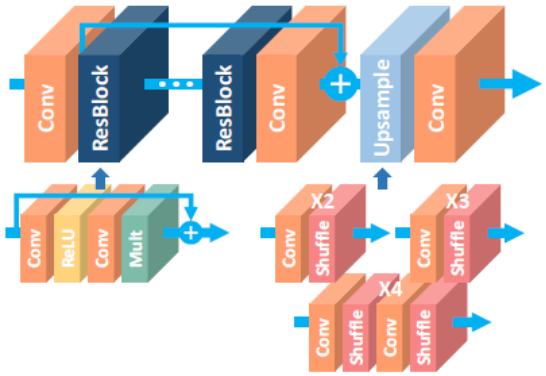

Figure 6: EDSR architecture. [7]

C. Generative Adversarial Networks

SRGAN

In general, when learning a super resolution network through deep learning, the MSE (Mean Square Error) loss function is used between the output HR image restored by inputting the input LR image to the network and the Ground Truth (GT) image. The restored output HR is learned with a high PSNR value, but sometimes produces a blurry output. (Figure 7) Therefore, a high PSNR value does not mean a clear image.

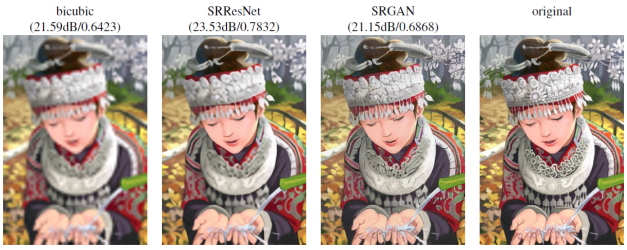

Figure 7: As the existing super resolutions are restored using MSE loss, the value of PSNR is high, but output is blurry. It shows the importance of introducing GANs optimized for a loss more sensitive to human perception.

In 2017, Ledig applied Generative Adversarial Networks (GANs) to the super resolution field. Also known as SRGAN(Super Resolution Using a Generative Adversarial Network) [8], it consists of a generator network that restores images and a discriminator that separates the ground truth (GT) and the output of the generator, and is used in style transfer instead of the existing MSE loss as well as GAN loss. Perceptual loss is proposed by using VGG loss as well. The experimental results suggest that the PSNR value decreases, but it can produce a more plausible result for the human eye, and the greater the multiple, the more effective it can be. Figure 8 represents architecture of generator and discriminator. Here sub-pixel method(Figure 4) is used at the end of the generator that increasing the number of pixels by combining the feature maps of the input image around.

$$\min_{\Theta_G} \max_{\Theta_D} E_{I^{\mathrm{HR}}\ p_{\mathrm{train}}(I^{\mathrm{HR}})} \left[\log D_{\Theta_D}\left(I^{\mathrm{HR}}\right)\right] + E_{I^{\mathrm{LR}}\ P_G(I^{\mathrm{LR}})} \left[\log\left(1 - D_{\Theta_D}\left(G_{\Theta_G}\left(I^{\mathrm{LR}}\right)\right)\right)\right] \quad (1)$$

$$I^{\mathrm{SR}} = I_X^{\mathrm{SR}} + 10^{-3} I_{\mathrm{Gen}}^{\mathrm{SR}} \quad (2)$$

$$I_{\mathrm{MSE}}^{\mathrm{SR}} = \frac{1}{r^2 \mathrm{WH}} \sum_{x=1}^{\mathrm{rW}} \sum_{y=1}^{\mathrm{rH}} \left(I_{x,y}^{H,R} - G_{\Theta_G}\left(I^{\mathrm{LR}}\right)_{x,y}\right)^2 \quad (3)$$

$$I_{\mathrm{VGG}\ /i \cdot j}^{\mathrm{SR}} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}\left(I^{\mathrm{HR}}\right)_{x,y} - \phi_{i,j}\left(G_{\Theta_G}\left(I^{\mathrm{LR}}\right)\right)_{x,y}\right)^2 \quad (4)$$

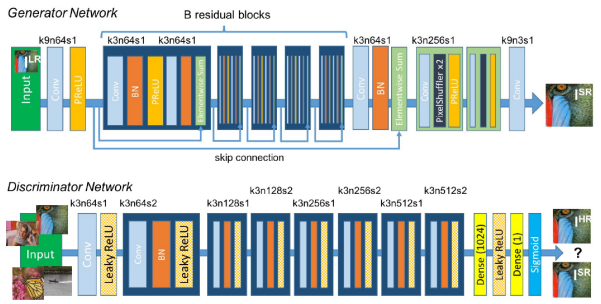

Figure 8: Architecture for SRGAN, generator and discriminator network with kernel size(k), number of feature maps(n) and stride(s) indicated for each convolutional layer.

ESRGAN

ESRGAN(Enhanced Super Resolution Using a Generative Adversarial Network) is proposed by Wang to enhance visual quality using key components of SRGAN network architecture. [9] In this work, Residual-in-Residual Dense Block(RDDB) without batch normalization is introduced. Figure 10 shows architecture of ESRGAN using RDDB. First, Batch Normalization(BN) is removed. By removing BN layers has proven to increase performanc and reduce computational complexity. Second, the Dense block, which is originated in DenseNet [10] replace the residual block to enhance the network. In SRGAN, like GAN in general, one generates an image and the other determines whether the image is real or fake. ESRGAN compares the generated image to a real image and tries to determine which is more real. This approaches forces algorithm to eventually generate details with sharper edges and more realistic textures. (Figure 9)

$$L_G = L_{\text{percep}} + \lambda L_G^{\text{Ra}} + \eta L_1 \tag{5}$$

$$L_1 = E_{x_i} \| G(x_i) - y \|_1 \tag{6}$$

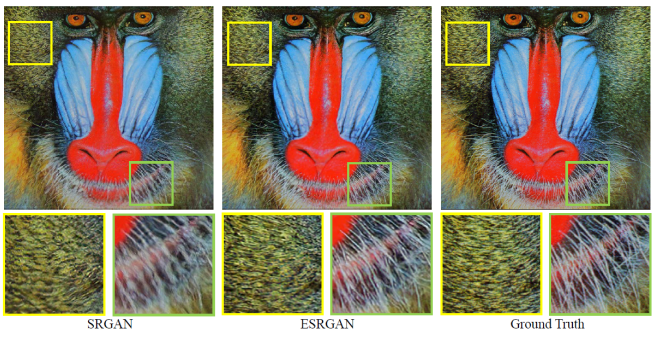

Figure 9: The super-resolution results by 4 times, ESRGAN outperforms SRGAN in details

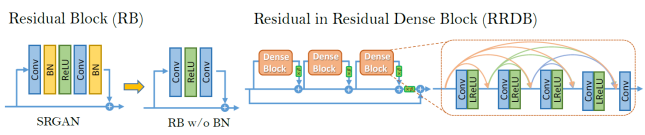

Figure 10: Architecture for ESRGAN

SinGAN

The goal of SinGAN(Learning a Generative Model from a Single Natural Image) [11] is to create an unconditional generative model that can learn the internal distribution of a single training image. The Model should learn fine features such as image arrangement, shape of the object, detail information and texture. For this, Shaham et al. used multi-scale GAN structure to generate images coarse-to-fine. Trinaing is similar to the existing GAN learning method. It consists of adversarial loss and reconstruction loss, and WGAN-GP loss as the adversarial loss. (Equation 7) Reconstruction loss is to learn in the direction of reducing the pixel difference between the image generated by the Generator and the GT (downsampled) image at that stage. (Equation 8)

$$\min_{G_n} \max_{D_n} L_{\text{adv}}(G_n, D_n) = \alpha L_{\text{rec}}(G_n) \tag{7}$$

$$L_{\text{rec}} = \left\| G_n\left(0, \left(x_{n+1}^{\text{rec}}\right) \uparrow^r\right) - x_n \right\|^2 \tag{8}$$

The noise setting injected at each stage for learning is aimed at reducing the pixel difference of the image, therefore fixed noise is injected only in the first stage(N stage) and no noise is injected in the other stages. Figure 11 shows the architecture of network. At each step, the Generator generates an image using noise and the resulting image generated in the previous step as inputs, and the Discriminator at that step is trained to distinguish the down-sampled GT from the generated image. As an exception, the first step (bottom) creates an image using only noise. At the beginning of step, while learning to create down-sampled GT, it focuses on the coarse and global features, and the higher it is, the more fine features are created.

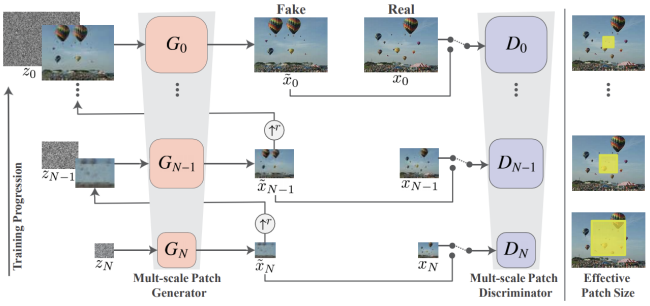

Figure 11: Architecture for SinGAN. x_0 is training image, As going down one step, it is down-sampled by r times.

D. Recent studies

SRFBN

Through SR methods for various models such as CNN, Residual Network, and GANs, the development of SR is steadily progressing. Li et el proposed SRFBN (Image Super Resolution Feedback Network) [13] which is SR method using a feedback network. It satisfies the feedback manner by using the hidden states of the RNN. [12] The feedback system has two conditions, the first is iterativeness, and the second is to change the output path of the system. In this network, RNN is used, and the loss for each iteration is tied so that the hidden state can send high level information. Also, LR image is provided for each iteration. Figure 12 shows the architecture of the

system. It contains 3 parts LR feature extraction block(LRFB), Feedback Block(FB) and Reconstruction block(RB). Global residual skip connection allows to recover residual images of input LR image. The FB contains sequentially a group of G projections with dense skip connections. Each projection group in which the HR function can be projected onto the LR function mainly contains an up-sample operation and a down-sample operation. (Figure 13)

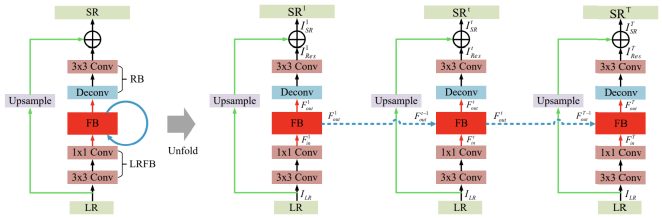

Figure 12: Architecture for SRFBN. It divides in 3 parts : LR feature extraction block(LRFB), Feedback block(FB), Reconstruction block(RB). Each block of weight are shared and there are global residual skip connection.

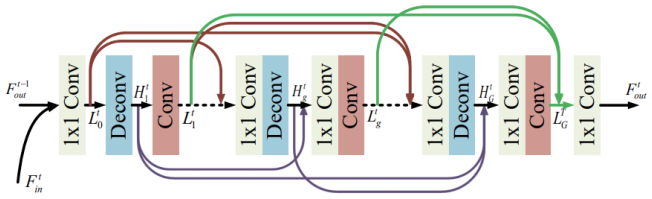

Figure 13: FB(Feedback) receives input which are concatenated between F_{out}^{t-1} and F_{in}^{t}. It has G numbers of projection(dense skip connection) group. Each projection group contains upsample, downsample operation.

PULSE

Menon et al proposed a novel super resolution algorithm called PULSE(Photo Upsamling via Latent Space Exploration) [14] which gives high-resolution and realistic images. This method is an entirely self-supervised not confined to a specific degradation operator used using training, not requiring training on databases of LR-HR images pairs for supervised learning. Traditional supervised networks (CNN for example) uses Mean Square Error(MSE) between generated Super-Resolved(SR) images and Ground Truth(GT) images. However, this approach has been noted to neglect perceptually relevant details critical to photorealism in HR images, such as texture. They proposed method that generates images using pre-trained generative model approximating the distribution of natural images under consideration. PULSE traverses the high-resolution natural image manifold, searching for images that downscale to the original LR image. Loss guides exploration through the latent space of generative model by downscaling loss. The low-resolution input image is denoted by I_{LR}. The aim is to learn a conditional generating function G that when applied to I_{LR}, yields a high resolution super-resolved image I_{SR}.

$$I_{\text{SR}} := \text{SR}\left(I_{\text{LR}}\right) \tag{9}$$

We can get the best recover of I_{HR} given I_{SR}. It approaches to reduce the problem to an optimization task which minimizes Equation 10. M is the natural image manifold in $R^{N\times M}$ and P is a probability distribution over M describing the likelihood of an image appearing in dataset. R is the set of images that downscale correctly. The optimal I_{SR} is a weighted pixelwise average of the set of high resolution images that downscales properly.

$$L := \|I_{\text{HR}} - I_{\text{SR}}\|_p^p \tag{10}$$

($\|\cdot\|$ denotes some l^p norm)

$$\|\text{DS}\left(I_{\text{SR}}\right) - I_{\text{LR}}\|_p \leq \epsilon \tag{11}$$

$$\int_{M\cap R} \|I_{\text{HR}} - I_{\text{SR}}\|_p^p \ \text{dP}\left(I_{\text{HR}}\right) \tag{12}$$

$$I_{\text{SR}} = \int_{M\cap R} I_{\text{HR}} \ \text{dP}\left(I_{\text{HR}}\right) \tag{13}$$

$R_\epsilon \subset R^{N\times M}$ is the set of images downscale properly.

$$R_\epsilon = \left\{I \in R^{N\times M} : \|\text{DS}\left(I\right) - I_{\text{LR}}\|_p^p \leq \epsilon\right\} \tag{14}$$

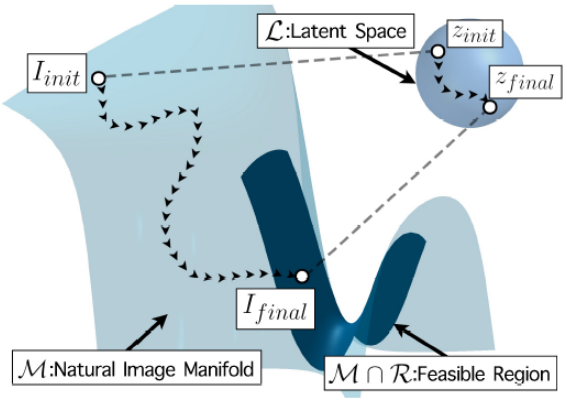

Figure 14: Pulse algorithm. While traveling from z_{init} to z_{final} in the latent space L, it travels from $I_{\text{init}} \in M$ to $I_{\text{final}} \in M \cap R$

TTSR

TTSR(Texture Transformer Network for Image Super-Resolution) [15] is a method proposed by Yang et el in 2020. LR and Reference images are formulated as queries and key in a transformer respectively. Recent studies have been made by reference based image super resolution slightly different with SISR. RefSR can obtain more accurate details from the Ref image compared to SISR. It transfers HR textures from a given Refernece image to produce visually better. They propose learnable texture extractor that parameters will be updated during end to end training. There is an embedding module to computing the relevance between LR and Ref image. Extracted features are fomulated from LR and Ref image as the query and key in a transformer to obtain hard-attention and a soft-attention map. Hard attention module and soft attention model transfer and fuse HR features from Ref image into LR features extracted from backbone through the attention maps.

The texture transformer search and transfer relevant textures from Ref to LR images. Figure 15 represents architecture of method. LR, LR $\uparrow$ and Ref are the input image, 4×4 bicubic upsampled input image and the reference image respectively. It sequentially apply bicubic down sample and upsample. There are 4 parts in the texture transformer such as learnable texture extractor(LTE) : It update parameters during end-to-end training to encoruage joint feature learning across LR and Ref image (Euation 15, 16, 17), the relevance embedding module(RE) : It aims to embed the relevance between LR and Ref image by estimating the similiarity between LR and Ref image by estimating the similarity between Q and K, the hard attention module for feature transfer(HA) : It transfers the HR texture features V from the Ref image. It only transfer features from the most relevant position in V for each query q_i to prevent blurring effect. and the soft-attention module for feature synthesis(SA) : It synthesize features from the transfered HR textur features T and the LR images F of the LR image from a DNN Backbone.

$$Q = \mathrm{LTE}\left(\mathrm{LR}\uparrow\right) \tag{15}$$

$$K = \mathrm{LTE}\left(\mathrm{Ref}\downarrow\uparrow\right) \tag{16}$$

$$V = \mathrm{LTE}\left(\mathrm{Ref}\right) \tag{17}$$

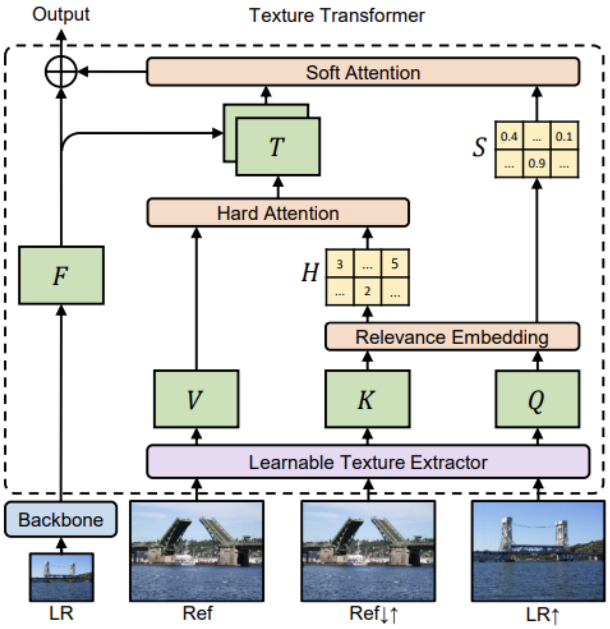

Figure 15: Texture transformer. Q(texture features extracted from an up-sampled LR images), K(down/up-sampled reference image) and V(original reference image). H and S are hard/soft attention map calculated from relevance embedding. $F$is the LR features extracted from DNN backbone, T is transferred texture features

MZSR

Usually, in order to train the model in the learning stage, LR-HR pair is made by down-sampling the original image. Then, resolution restoration (SR) is performed based on the information trained in the test step. Like this, down sampling is necessary to make an LR-HR pair. It shows performance only for specific coniditions such as bicubic down sampling, and shows low performance when downsampling to Gaussian. The MZSR(Meta Transfer Learning for Zero Shot Super Resolution) method [16] is a method that improves flexibility and enables rapid application to each task by using the Meta-transfer learning technique because the existing CNN-based method has high performance only in limited conditions. While overcoming the limitations of CNN, it uses external sample of SISR. In the training stage, down sample first as an external sample, then create and train HR and LR pairs, and go through the process of predicting LR as HR based on this learned model. The model extract information itself and construct HR-LR pairs for training and use original image to predict. The author points out it takes many time for training and hard to use for other images. In the large-scale training stage, it is possible to learn representations that are commonly used from various images. High performance is achieved by taking feature values from natural images and utilizing them. Equation 18 shows that training is proceeded by minimizing $L1$ loss between HR image and bicubic-LR image. The model parameters Θ are optimized to achieve minimal test error of Θ with respect to Θ_i.

$$L^D\left(\Theta\right) = E_{D\ \left(I_{\mathrm{HR}}, I_{\mathrm{LR}}^{\mathrm{bic}}\right)}\left[\left\|I_{\mathrm{HR}} - f_\Theta\left(I_{\mathrm{LR}}^{\mathrm{bic}}\right)\right\|_1\right] \tag{18}$$

Figure 16 represents Meta-Transfer learning. It is called training for training to make learning quickly with each specific task in order to be able to learn well later. It uses MAML(Model-Agnostic Meta-Learning)method to search the most sensitive initial point in various kernel condition. Covariance matrix is also used for kernel distribution. We train meta-learner and update model parameter through task-level loss. Optimization proceeds in the direction to minimize the test error. In the meta-test stage, when one image is entered, the meta-learning weight value is quickly updated for each kernel. The SR image is returned through this process.

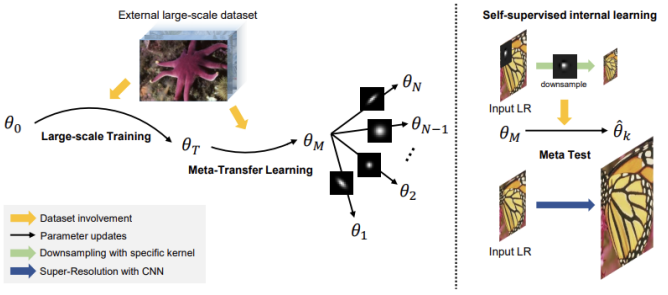

Figure 16: The overall scheme of MZSR. During meta-transfer learning, external dataset is used. Large-scailing training, Meta-transfer learning and Meta test. It shows N tasks for simplicity. For the test phase, self supervision is used for learning internal information.

RFDN

Residual feature distillation network(RFDN) is an architecture which won in AIM2020. RFDN is inspired by two recent works IMDN [18] and RFANet [19]. The author highlights the heavy computation of CNN-based methods. To solve this problem Liu et el proposed various fast and lightweight CNN models. The feature dillation connection(FDC) acts equivalents to the channel splitting operation while being more lightweight and flexible. RFDN uses multiple feature

distillation connection to learn more discriminative feature representations. A shallow residual block (SRB) is the main block of RFDN which makes network take advantage from residual learning. Figure 17 represents the architecture of the model. IDMN had a good performance in terms of both PSNR and interference time and won the first place in the AIM2019 for constrained image super resolution challenge. However, the number of parameters of IDMN is more than most of lightweight SR models. Liu et el gives more comprehensive analysis of the feature distillation connection(FDC) which is more lightweight and flexible than IDM. (Figure 18).

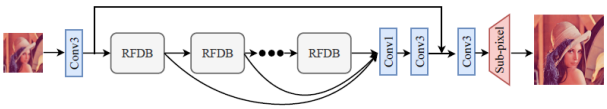

Figure 17: The architecture of RFDN. It achieves good performance with much fewer parameters than competitors.

RFDN is much more light weight by using the feature dillation connections (FDCs) than IDFM. They also proposed a shallow residual block(SRB) that uses blocks of RFDN to improve SR performance. It has a residual learning without extra parameters.

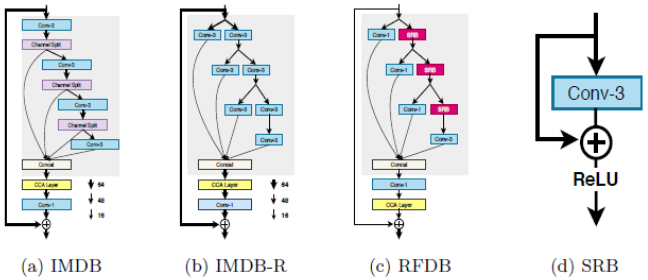

Figure 18: (a) IMDB: the original information multi-distillation block. (b) IMDB-R: rethinking of the IMDB. (c) RFDB: residual feature distillation block. (d) SRB: shallow residual block. [17]

BSRGAN

It is known that SISR methods would not perform well if degradation model deviates from in real images. Kai et el highlights the problem that SISR methods are mainly designed from bicubic degradation. Recently there have been progressed for Gaussian degradation too, however these methods are not fit for the most real images such as JPEG compressed ones. They propose a new practical degradation model using deep learning method. There are two steps to design degradation model to synthesize LR images for training. First, we need to make the blur, downsampling and noise more practical. Blurring is done by convolution between isotropic and anisotropic Gaussian kernels from both HR space ad LR space. For the downsamping, it is done by nearist, bilinear, bicubic and down-up-sampling. For the noise, we use gaussian noise, JPEG compression noise, and processed camera sensor noise. Second step is to do degradation shuffle, instead of using the commmonly-used blur/downsampling/noise addition pipline they adpot randomly shuffled degradations to synthesize LR images. [20] In this work, they focus designing degradation model for deep blind DNN model. To evaluate the model, they used ES-RGAN super resolver and then applied it to super-resolve both synthetic and real images with diverse degradation.

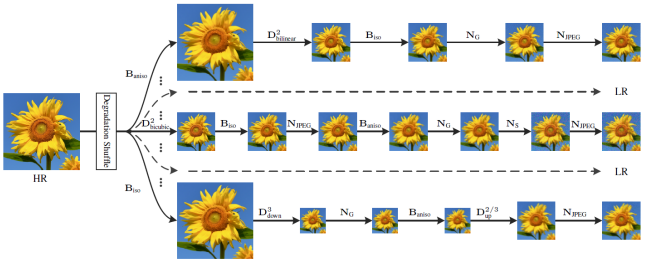

Figure 19: Proposed schematic of the degradation model [20]. HR images are randomly shuffled degradation sequence(B_{iso}, B_{aniso}, D^2, N_G, N_{JPEG}, N_s are performed. D^s represents the downsampling operation with scale factor s.

FKP

Liang et el proposed FKP(Flow-based Kernel Prior) method to apply in super resolution problem. Kernel estimation is one of the important problems for SR. They express the most of existing works assume that blur kernel is fixed and down-sampled by bicubic method which is not fitting with the real world images. They propose to deal with unkown blur kernel by comparing exist methods Double-DIP [22] and KernelGAN [23]that are based on kernel estimations. Figure 20 represents the schematic illustration of FKP. It consists of several batch normalization layers, permutation layers and affine transforamation layers that allows to capture the kernel distribution by learning an invertible mapping between the kernel space and the latent space. It optimizes the loss of the kernel in un-supervised way. The advantage of this method is that they have fewer parameters than existing methods [22][23]. Moreover, it is more stable in convergence by initializing a reasonable kernel of bijection function. It outpefroms better kernel estimations compared to 2 methods. FKP can be easily incorporated into existing kernel estimation methods as DIP-FKP and KernelGAN-FKP. We know that general classical degradation model of image SR can be formulated as Equation 19. where $(x \otimes k)$ is a convolution between x and blur kernel k. $\downarrow_s$ is for downsampling with scale factor s. According to Maximum A Posteriori frame work (MAP), it is an optimization problem as Equation 20. The probability of kernel can be calculated using Equation 21 where $\frac{\partial f_\Theta(k)}{\partial k}$ is the Jacobian of f_Θ at k. Finally, Θ can be obtained by miminizing negative log-likelihood(NLL) loss.

$$y = (x \otimes k) \downarrow_s + n \tag{19}$$

$$x^*, k^* = \text{argmin} \ \| y - (x \otimes k) \downarrow_s \|^2 + \lambda \Phi(x) + \lambda \Omega(k) \tag{20}$$

$$p_K(k) = p_z(f_\Theta(k)) \left| \det \left(\frac{\partial f_\Theta(k)}{\partial k} \right) \right| \tag{21}$$

$$L(k, \Theta) = -\log p_z(f_\Theta(k)) - \sum_{n=1}^{N} \log \left| \det \left(\frac{\partial f_\Theta^n}{\partial h^{n-1}} \right) \right| \tag{22}$$

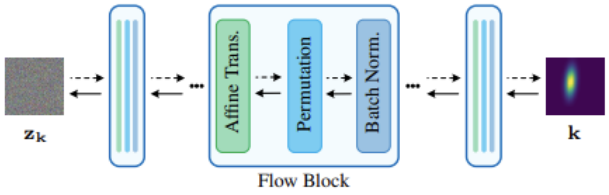

Figure 20: Proposed schematic of the flow-based kernel prior(FKP) network. It learns an invertible mapping between kernel k and the latent varialbe z_k. The flow blocks consist of batch normalization, permutation and affine transformation.

III. Metrics for evaluation

As Image Super resolution research have been progressed in recent few years, metrics for evaluation has been evolved. Early works used Mean Square Error (MSE : Equation 24) and PSNR(Peak Signal to Noise Ration: Equation 23) as the main evaluation indicators. PSNR is used to evaluate pixel loss information of generated/compressed image. SSIM(Structural Similarity Index Map) is also widely used for evaluating the quality of the images. It evaluates the quality of luminance, contrast, structure points of view. Generally the higher values of PSNR and SSIM, the better but it is not for all cases. Sometimes the picture with high PSNR and SSIM looks more blurry to human eyes. Therefore, they newly introduced MOS(Mean Opinion Score) in a SRGAN paper. Usually GAN based method like SRGAN has a lower performance in measuring distortion but it has a better performance in terms of indicators that reflects human subjective satisfaction.(Figure 21). In SRGAN paper, they introduced evaluation indicator of MOS(Mean opinion score) for human perceptual loss.(Equation 26). It is calculated as the arithematic mean over single rating performed by human subjects for a given stimulus in a subjective quality evaluation test, where R are individual ratings for a given stimulus by N subjects. There is a trade-off relationship between distortion measure and perceptual measure [24]. The lower distortion of an algorithm, the more its distribution deviate from the statistics of natural scenes. Therefore, research in the direction should be improving both indicators at the same time to yield good performance. After that, It has been studied that the LPIPS(Learned Perceptual Image Patch Similarity) measurement method is suitable for calculating perceptual similarity. It evaluates the distance between image patches. Higher means further/more different and Lower means more similar. It has been found that deep network activation work well as a perceptual similarity metric. LPIPS is also known to show a similar tendency to SSIM, but it is insufficient to discriminate photorealistic images, so we need to take in consider. Furthermore, the perception index(PI) is also used for challenge on perceptual image super resolution. It combines the no-reference image quality measures of NIQE [26] and Ma et al. [27] as Equation 27. The lower perceptual index indicates better perceptual quality. In a summary, PSNR, SSIM, MOS, LPIPS and PI are the most used indicators for the image super resolution challenge competition every year.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{23}$$

$$\mathrm{MSE} = \frac{\sum_{M,N}\left[I_1(m,n) - I_2(m,n)\right]^2}{M \cdot N} \tag{24}$$

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{25}$$

$$\mathrm{MOS} = \frac{\sum_{n=1}^{N} R_n}{N} \tag{26}$$

$$\mathrm{PI} = \frac{1}{2}\left((10 - \mathrm{Ma}) + \mathrm{NIQE}\right) \tag{27}$$

	SRResNet-		SRGAN-		
Set5	MSE	VGG22	MSE	VGG22	VGG54
PSNR	32.05	30.51	30.64	29.84	29.40
SSIM	0.9019	0.8803	0.8701	0.8468	0.8472
MOS	3.37	3.46	3.77	3.78	3.58
Set14					
PSNR	28.49	27.19	26.92	26.44	26.02
SSIM	0.8184	0.7807	0.7611	0.7518	0.7397
MOS	2.98	3.15*	3.43	3.57	3.72*

Figure 21: Performance of different loss functions for SRResNet and the adversarial networks on Set5 and Set14 benchmark data. MOS values are rated from 1(bad quality) to 5(excellent quality), the higher the better.

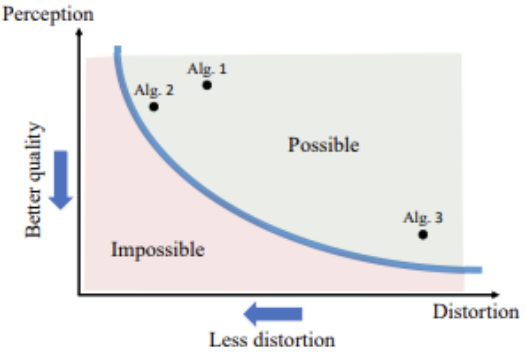

Figure 22: The perception-distortion tradeoff. Image restoration algorithms can be characterized by their average distortion and by the perceptual quality of the images they produce.

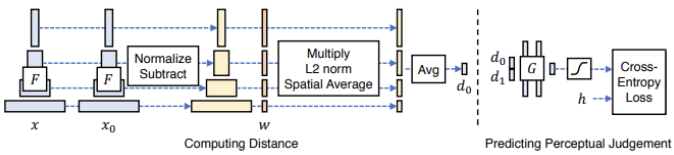

Figure 23: Computing distance using LPIPS metric. From a given network(F), distance(d_0) can be calculate between two patches x and x_0. After calculaing L2 distance, it averages across spatial dimensions and across all layers. A network(G) is trained to predict perceptual judgement(h) from distance pair(d_0,d_1).

IV. State of the art

The field of image super-resolution applying deep learning has leaped in the past five years. This paper mainly focuses on the Single Image Super-Resolution (SISR) field. In the early stages, a CNN architecture was mainly used based on a supervised neural network, but LR to HR mapping, which is a non-linear mapping, is learned by making a pair of LR-HR. Focusing on the fact that only increasing the CNN layer does not result in performance improvement, the residual network was introduced, which brings a great performance improvement. The use of residual structures allowed for the training of larger networks. After that, many researches applying GANs was conducted, and a study on SR that seemed more plausible to the human eye was conducted, not just improving performance indicators such as PSNR and SSIM. It is to introduce a perceptual loss and create a high-resolution image that looks better to the human eye. With these two general trends, the first is to optimize the pixel-wise average between SR and HR, and the second is to focus on perceptual quality. Since then, various studies have been performed, and a method still using GANs or a method using a new architecture such as Transformer (TTSR) has been introduced in addition to CNN. In this way, researches showing performance improvement by proposing an efficient model architecture is continuing, and at the same time, attempts to reduce parameters to reduce the computational amount are continuing. In addition, the unsupervised method (MZSR) was introduced in the field of super-resolution, where supervised was dominant so far, and the self-supervised method (PULSE) was introduced, resulting in a more remarkable development. In addition, if the existing SR model performed well only in images that were downsampled by bicubic downsampling or by a specific method, recent research fields approach a more intrinsic problem. For example, research is being conducted on how to make a practical image degradation model achieve good results even for images that do not know how downsampling was performed, such as real images (like JPEG). Likewise, studies on kernel estimation that can estimate the unknown blur kernel are continuing.

V. challenges

There are some challenges to be solved in the future. First, deep learning-based algorithms have a high computational load and require many parameters. Therefore. Additional methods such as model compression and optimization are needed to apply in the real world. Second, in addition to the bicubic downsampling method, the biggest task is to study a model that can perform well for images downsampled by the unknown method. Third, metrics such as MOS and LPIPS have been introduced for human perceptual loss assessment, but need to take a lot of effort and are non-reproducible. And currently, there are no unified and admitted evaluation metrics for SR quality. Having a unified evaluation metric should be also done for future works. Finally, although recently unsupervised research is underway, it is difficult to collect images with different resolutions on the scheme. The unsupervised method will allow us to train dataset without paired LR-HR images, it will be a promising direction for future research.

VI. Conclusion

In this paper presents a briew survey of deep learning algorithms on SISR. As customers desires the higher quality of contents according to increasing high resolution display market, It has been remarked of importance to get the better resolution images. Super resolution is a field that has been steadily researched, but it has made great progress as deep learning is applied. We present the recent studies and summarize the state of the art. However, there are some challenges left in 4 aspects(Computation optimization, Practical image degradation model, Unified evalutaion metrics and Unsupervised method). With these challenges, the results will be more promising in the future.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in Proceedings of the European Conference on Computer Vision, 2014.

[2] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in Proceedings of the European Conference on Computer Vision, 2016.

[3] "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-pixel Convolutional Neural Network", 2016.

[4] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[5] M. Rad, B. Bozorgtabar, U. Marti, and M.Basler, "SROBB: Targeted Perceptual Loss for Single Image Super-Resolution", in ICCV 2019.

[6] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, 2014.

[7] B. Lim, S. Son, H. Kim, S. Nah, and K.M. Lee, "Enhanced deep residual networks for single image super-resolution." IEEE Conference on Computer Vision and Pattern Recognition in CVPR ,2017.

[8] C. Ledig, L. Theis, F. Husz´ar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[9] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," in ECCV Workshop, 2018.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in CVPR, 2017.

[11] T. Shaham, T. Dekel, and T. Michaeli,"Learning a Generative Model from a Single Natural Image", in CVPR, 2019.

[12] Amir R. Zamir, Te-Lin Wu, Lin Sun, William B. Shen, Bertram E. Shi, Jitendra Malik, and Silvio Savarese, "Feedback networks" in CVPR, 2017.

[13] Z. Li, J. Yang, Z.Liu, X. Yang, G. Jeon and W. Wu, "Feedback Network for Image Super-Resolution", 2019.

[14] S.Menon, A.Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models", in CPVR, 2020.

[15] F. Yang, H. Yang, J. Fu, H. Lu, and B. GUo, "Learning Texture Transformer Network for Image Super-Resolution" in CPVR, 2020.

[16] S. Jae woong, C. Sunwoo, and C. Nam Ik, "Meta-Transfer Learning for Zero-Shot Super-Resolution", in CPVR 2020.

[17] J.Liu, J.Tang, and G.Wu, "Residual Feature Distillation Network for Lightweight Image Super-Resolution", 2020

[18] Hui, Z. Gao, X. Yang, Y. Wang, X, "Lightweight image super-resolution with information multi-distillation network.", inACM MM, 2019.

[19] Liu, J. Zhang, W. Tang, Y. Tang, J. Wu,and G, "Residual feature aggregation net- work for image super-resolution.", in CPVR, 2020.

[20] Z. Kai, L. Jingyun, G. Luc, and T. Radu, "Designing a Practical Degradation Model for Deep Blind Image Super-Resolution", in CPVR, 2021.

[21] L. Jingyun, Z.Kai, G. Luc, and T. Radu, "Flow-based Kernel Prior with Application to Blind Super-Resolution", in CPVR, 2021.

[22] Y. Gandelsman, A. Shocher, and M. Irani, "Double-dip Unsupervised image decomposition via coupled deep-image-priors." , in CPVR 2019.

[23] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan." , in NIPS 2019.

[24] Y. Blau, T. Michaeli , "The Perception-Distortion Tradeoff", in CPVR 2018.

[25] Y.Blau, R. Mechrez, R.Timofte, TomerMichaeli, and L.Zelnik-Manor. "The 2018 pirm challenge on perceptual image super-resolution", in ECCV, 2018.

[26] Mittal, A. Soundararajan, R. Bovik, and A.C, "Making a "completely blind" image quality analyzer.", IEEE, 2013.

[27] Ma, C. Yang, C.Y Yang, X. Yang, and M.H, "Learning a no-reference quality metric for single-image super-resolution." in 2017.

[28] Y. Wenming, Z. Xuechen, T. Yapeng, W. Wei, X. J.H,and L. Qingmin, "Deep Learning for Single Image Super-Resolution:A Brief Review", in 2019.