ACTA INFOLOGICA 2020;4(2):139-154
dergipark.org.tr/acin

İSTANBUL UNIVERSITY PRESS

DOI: 10.26650/acin.765320

REVIEW ARTICLE

A Survey on Image Super-Resolution with Generative Adversarial Networks

Üretken Çekişmeli Ağlar ile Görsel Çözünürlük Artırımı Üzerine Bir Araştırma

Hürkal Hüsem[1], **Zeynep Orman**[2]

[1]Istanbul University, Cerrahpaşa, Computer Engineering Department, İstanbul, Turkey.
[2]Istanbul University, Cerrahpaşa, Computer Engineering Department, İstanbul, Turkey.

ORCID: H.H. 0000-0002-5414-6481; Z.O. 0000-0002-0205-4198

Corresponding author:
Hürkal Hüsem,
Istanbul University, Cerrahpaşa, Computer Engineering Department, İstanbul, Turkey.
Telephone: +90 212 217 50 67
E-mail address: iletisim@hurkal.com

Submitted: 07.07.2020
Revision Requested: 28.07.2020
Last Revision Received: 15.08.2020
Accepted: 17.08.2020

Citation: Hurkal, H., & Orman, Z. (2020). A survey on image super-resolution with generative adversarial networks. *Acta Infologica, 4*(2), 139-154. https://doi.org/10.26650/acin.765320

ABSTRACT

Super-resolution is a process to increase image dimensions with a specific upscaling factor while trying to preserve details that matche with the original high-resolution form. Super-resolution can be done with many techniques. But the most effective technique is the one that takes advantage of several neural network designs. Some network designs are more appropriate than others on the specific subject. This study focuses on super resolution studies using Generative Adversarial Network. Many studies use this neural network type to look at various topics such as artificial data production and making the data more meaningful. The key point of this neural network type is having two different sub-networks that try to defeat each other in order to make more realistic results. Performance metrics that measure the quality of a generated image, loss functions used in a neural network and research papers on super-resolution with Generative Adversarial Network are the main domains of this study.

Keywords: Image Super-Resolution, Generative Adversarial Networks, Resolution Enhancement

ÖZ

Çözünürlük artırımı (süper-çözünürlük) belirli bir artırım değeri ile görselin yüksek çözünürlükteki detaylarını korumaya çalışarak boyutlarını artırma işlemidir. Süper-çözünürlük birçok teknik ile gerçekleştirilebilir. Ancak bu konudaki en etkili teknikler çeşitli sinir ağı tasarımlarından yararlanan tekniklerdir. Bazı ağ tasarımları belirli konularda diğerlerine göre daha uygundur. Bu çalışma Üretken Çekişmeli Ağlar ile gerçekleştirilmiş çözünürlük yükseltme işlemlerine odaklanmıştır. Birçok çalışma yapar veri üretimi ve verinin daha anlamlı hale getirilmesi gibi çeşitli konularda bu yapay sinir ağı tipini kullanır. Bu yapay sinir ağı tipi ile yapay veri üretimi ve verinin daha anlamlı hale getirilmesi gibi alanlarda başarılı çalışmalar mevcuttur. Daha gerçekçi sonuçlar üretebilmesi için birbirini yenmeye çalışan iki alt ağdan oluşması bu ağ türünün kilit noktasıdır. Üretilen görselin kalitesini ölçen başarım ölçümleri, sinir ağında kullanılan yitim fonksiyonları ve Üretken Çekişmeli Ağ kullanarak çözünürlük artırımı üzerine çalışılmış araştırma makaleleri bu çalışmanın temel alanında yer almaktadır.

Anahtar kelimeler: Görsel Çözünürlük Artırımı, Üretken Çekişmeli Ağlar, Çözünürlük Geliştirme

1. INTRODUCTION

Resolution is a measure of pixel density within the specified unit. Higher resolution images provide more detail about the scene. In some domains, there is a strong need to increase the details on images to work on.

Super-resolution is a technique to enhance low-resolution images with minimum loss. This technique includes several processes such as denoising and deblurring (Protter, Elad, Takeda, & Milanfar, 2008). It is important for improving human understanding and getting higher accuracy values from computational tasks for the image. Higher resolution provides more details about the scene.

A survey on super-resolution gives detailed information about the history of the problems, domain, and algorithms (Nasrollahi & Moeslund, 2014). According to this survey, the first algorithm on super-resolution introduced the Fourier transform and the given solution was followed by many researchers (Gerchberg, 1974). The first hallucination solution solved with a neural network was applied to this problem area, to improve the resolution of fingerprint images (Mjolsness, 1985).

There are some traditional methods similar to super-resolution, such as interpolation. Interpolation is a similar technique with super-resolution but it shouldn't be confused because interpolation cannot restore high-frequency details (Gotoh & Okutomi, 2004). Interpolation includes several simple and easy to implement methods such as nearest-neighbor interpolation, bilinear interpolation, and bicubic interpolation; but they also show poor results in quality as shown in Figure 1. Therefore, there is a strong need for detail in discovering and data completion.

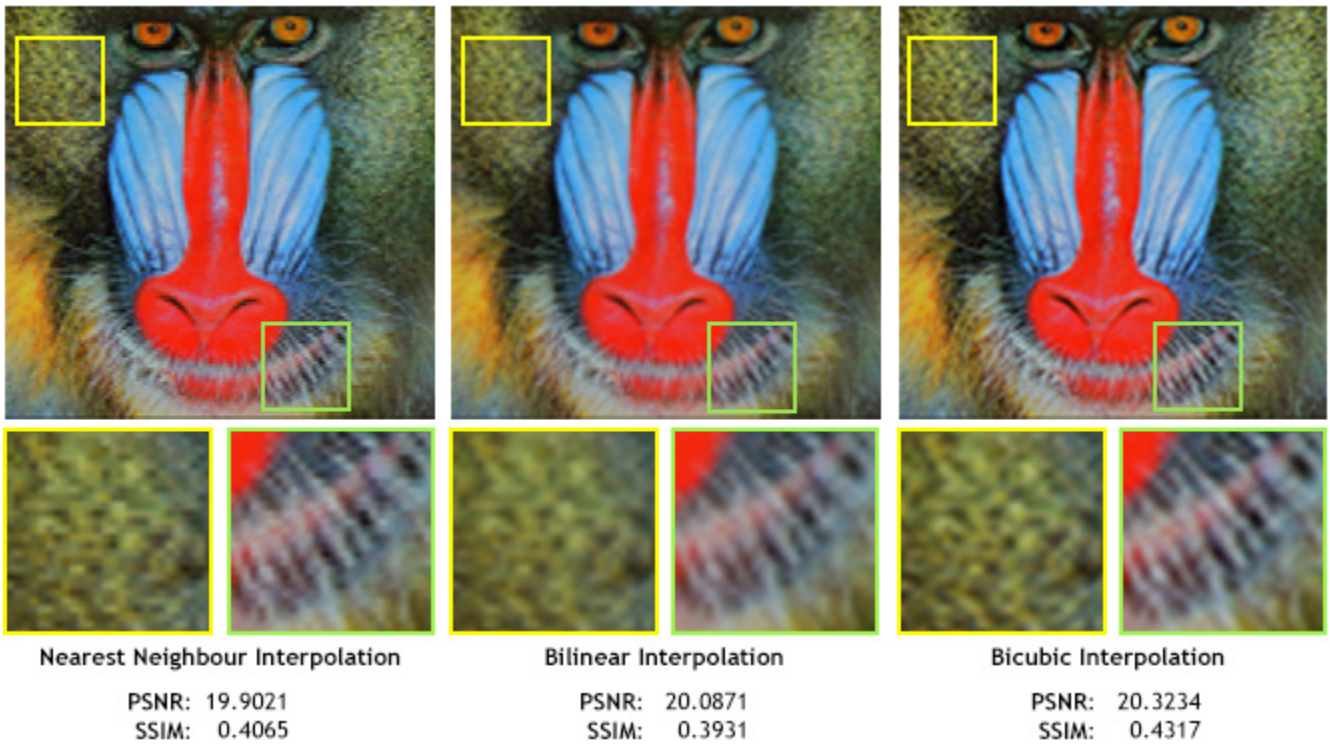

Figure 1. Quality comparisons of interpolation methods with "peak signal-to-noise ratio" (PSNR) and "structural similarity index" (SSIM)

Neural network-based studies, especially generative adversarial network (GAN) designs, overcome super-resolution problems. Visual details of state-of-the-art single-image super-resolution studies, SRGAN (Ledig, et al., 2016) and ESRGAN (Wang, et al., 2018) are shown in Figure 2 to highlight this necessity over the Set14 dataset baboon image. GAN-based studies significantly give better results than traditional methods. Somehow, ESRGAN's peak signal-to-noise ratio (PSNR) is reported 20.35 which is almost the same with bicubic interpolation as shown in Figure 1 despite bicubic interpolation's poor quality. This situation occurs with the familiarity of the used performance metric and neural network tuning strategy. It also discussed

why PSNR and similar performance metrics were not the best metrics for super-resolution comparisons (Zhang, Shao, Hu, & Gao, 2017).

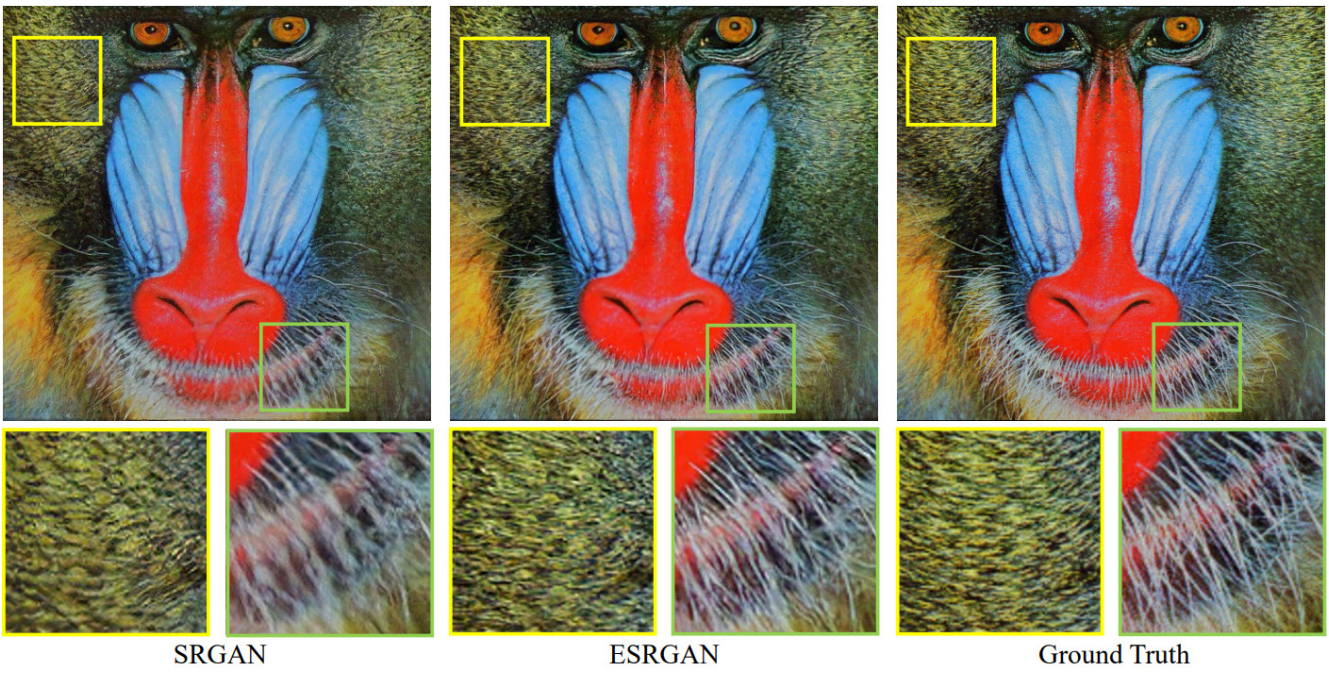

Figure 2. Visual details of state-of-the-art super-resolution methods, SRGAN and ESRGAN (Wang, et al., 2018).

GANs have attracted attention in recent years and breathed new life into existing approaches from machine learning (Goodfellow, et al., 2014). Thanks to the increased use of GAN, highly successful results can be achieved in many areas such as image processing, signal processing, and security.

The second section gives general information about GAN, the third section includes several studies that use GAN on super-resolution operations, the fourth section is about the performance metrics used in the overview studies, the fifth section is about loss functions used in overview studies, and the final section is about datasets used in overview studies.

2. GENERATIVE ADVERSARIAL NETWORKS (GAN)

GAN is a generative system in which two separate neural networks overcome one another with a competition principle. These two networks, called the generative and the discriminative, are operated simultaneously. While the generative network (*G*) aims to produce realistic artificial data, the discriminative network (*D*) tries to distinguish whether the data received is real or false.

After a competitive process, both networks specialized for their purposes and as a result, the GAN design realistically generates data. From this point of view, GAN is likened to a two-player mini-max game rather than an optimization problem. While the generative network aims to increase the error rate of the discriminative network, the discriminative network tries to reduce the failure probability by itself (Goodfellow, et al., 2014).

The value function *V(D,G)* representing the mini-max game which is shown in (1) as an equation. While *G* wants to minimize *V*, *D* wants to maximize it. *D* refers to the differentiable function of discriminative network calculated by a multi-layer perceptron while *G* is generative. In the given equation, *pz* is expressed as the distribution generated over the *x* which is generated data.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

2.1. Generative Sub-network (*G*)

The purpose of *G* is to increase the likelihood that the discriminative network will fail by generating better data. *G* takes noise vector (*Z*) as input sampled from Gaussian or uniform distributions. If there was no update rule over *z*, the generative network would produce only noisy data. Generated data is easy to distinguish for *D* in the early epochs of training because noisy data is not good enough yet. *G* attempts to minimize the log(1-D(G(z))) value to descend its gradient so that it generates more realistic synthetic data (Goodfellow, et al., 2014).

The generative network tries to generate data that makes the discriminative network think it is real. In other words, if the discriminative network returns that $D(x) = 1$ for generated data, so this data marked as "real" and the discriminative network is deceived.

2.2. Discriminative Sub-network (*D*)

The discriminative network (*D*) always tries to find out whether the data received comes from the original data or not. This makes *D* a binary classifier. *D(x)* evaluates the likelihood of *x* coming from real data rather than *pg*. *D* is responsible for the proper labeling of the actual data and the generated data from *G*, therefore there is training to increase the probability of *D*. Proper labeling of *D* means the separation of real and artificial data correctly. This sub-network measures the difference in how generated data differs from the original one (Goodfellow, et al., 2014). Despite the fact that *D* and *G* are opponents, *D* tells *G* how to make real-like data.

GAN architecture is shown in Figure 3. The data that comes from the generative network is presented in a mixed manner with the actual data to the discriminative network. A loss value is calculated as a result of the predictive result of the discriminative network.

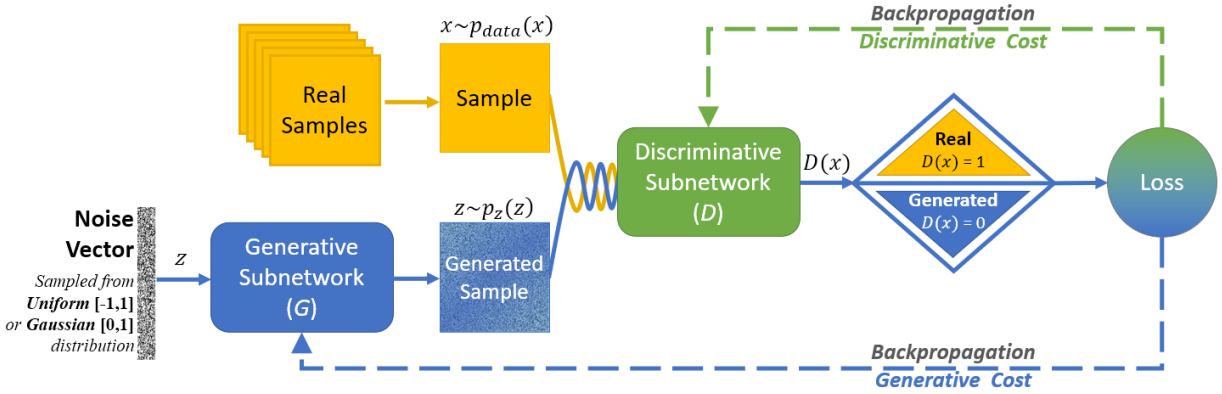

Figure 3. GAN architecture.

Small changes in inputs in the deep neural networks may cause serious changes in output values (Goodfellow, Shlens, & Szegedy, 2015). It emphasizes that the values applied in the same direction in the low-value bits of the input values may cause a conscious orientation at the output. The major reason for this is the frequent use of a linear or "linearized" activation function in the many neural networks.

Another generative neural network type similar to GAN is Variational Autoencoders (VAE). VAE is also used for generating data according to the input in an unsupervised manner. VAE can be divided into two parts as encoder and decoder. Encoder tries to reduce dimensions with a bottleneck in latent space, thus data is transformed into mean and standard deviation vectors. This process is a kind of learning loss compression algorithm. The last section of VAE, which is called decoder, generates new data according to these vectors and specified probability distribution (Kingma & Welling, 2014; Kingma & Welling, 2019). VAEs are useful for creating similar data to the input for focusing on explicit information, but GANs focus on implicit information so that they can create new data that is not available yet. This situation makes GANs useful for realistic data generation and completing missing parts of the data.

In super-resolution problems, VAEs tend to generate blurry results because of the bottleneck. Important and necessary details on an image fail to encode and decode during this process. GAN architecture was designed to be unsupervised like VAEs and can be a very advantageous approach to increase the amount of data needed by producing artificial data in deep learning applications that need a large dataset.

3. OVERVIEW

Super-resolution studies aim to produce higher resolution and quality images than low-resolution images. But also, some studies benefit from super-resolution as well are explained in another sub-section.

3.1. Single-Image Super-Resolution (SISR) GAN Studies

This section includes GAN based super-resolution studies on context-free or context-aware images such as the human face, text, traffic signs, satellite, etc. All studies in this section are SISR studies.

Inspired by many super-resolution studies using GAN, "a generative adversarial network (GAN) for image super-resolution (SR)" (SRGAN) applied a realistic super-resolution process in the quality of photo-shooting to each image (Ledig, et al., 2016). Perceptual loss and content loss functions were used together instead of pixel-based similarity. SRGAN also benefited from the deep residual network and achieved higher mean opinion score (MOS) than the state-of-the-art techniques in the literature.

Four times and eight times magnifications were applied with the style transfer and resolution upgrade approach (Johnson, Alahi, & Fei-Fei, 2016). Semantic analysis was used to increase success in both processes. The proposed system consists of two parts; an image transformation network and a loss network which is a convolutional neural network. The basis of this selection is that the semantic and perceptual information that the loss function wants to calculate can be easily coded with the convolutional neural network.

Perceptual GAN is used to solve the small object detection problem which is traffic sign detection (Li, et al., 2017). There are cases where the boundary or silhouette is certain, but when interpreted due to its small size, incorrect results will likely occur. To develop a better object recognition application, the resolution upgrade process is applied to the small size objects.

In order to overcome the limitations of pixel-based loss methods, loss functions have been designed for both the generative and the discriminative network. High SSIM results in 4x and 8x upscaling were achieved with the super-resolution perceptual generative adversarial network (SRPGAN) (Wu, Duan, Liu, & Sun, 2017).

Face Conditional Generative Adversarial Network (FCGAN), named neural network design, is applied to human faces for enhancement (Bin, Weihai, Xingming, & Chun-Liang, 2017). Four times scaling was performed with this network structure. There was no need for any preprocessing such as alignment and semantic information input. Also, this model was not affected by accessories such as hats and glasses.

A GAN based residual neural network was designed using a 4x upscaling factor (Zhang, Shao, Hu, & Gao, 2017). In addition to the changes in the convolutional content loss and adversarial loss functions, several operations were performed during the training of the network and pre-processing the data. Especially, with the "mean opinion score" (MOS) criteria was found to be prominent among similar studies in the literature. It was criticized that in super-resolution studies, many models use mean squared error, which generates a higher PSNR signal but does not give a strong perceptual result.

Small and blurred human faces in visuals were made more detailed (Bai, Zhang, Ding, & Ghanem, 2018). Unlike previous studies in the literature, up-sampling and refinement sub-networks were used together. Another innovation was the ability to distinguish between real and generated data acquired in the discriminative network, as well as the ability to distinguish whether the area sampled in the relevant visual if it is a face. The dataset in this study was not developed to directly study human faces. Therefore, it has become necessary to do this in the discriminative network.

Thanks to the Transferred GAN (TGAN), a combination of the transfer-learning approach and the abandonment of the batch normalization process, super-resolution is applied to satellite images (Ma, Pan, Guo, & Lei, 2018). The batch normalization

increased the computational time despite the increase in performance in other image-processing tasks, but there was not enough remarkable effect on super-resolution applications.

To further improve image quality of SRGAN networks, both the adversarial and perceptual loss functions and network enhancements were introduced in Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) (Wang, et al., 2018). The normalization process was abandoned like Ma et al. (2018). The most important point of the study was the method called Residual-in-Residual Dense Block (RRDB), the patterns in natural visuals studied were much better than SRGAN. Thus, this work placed first at the PIRM2018-SR Challenge (region 3) (Blau, Mechrez, Timofte, Michaeli, & Zelnik-Manor, 2018) with the best perceptual index.

EnhanceNet uses both perceptual and texture matching lossand focuses on texture details in images (Sajjadi, Scholkopf, & Hirsch, 2017). Architecture was designed with a fully convolutional network and residual learning blocks.

SRFeat also used long-skipped 16 residual block connections in the generative network to produce high-frequency structural features (Park, Son, Cho, Hong, & Lee, 2018). SRFeat has two discriminative networks; one has an adversarial loss and the other has pixel-wise loss functions. ImageNet dataset is used for pre-training while DIV2K dataset is used for fine-tuning.

Learning to super-resolve both text and face images within a single model, it aimed to produce solutions with a common network design called Multi-class GAN (MCGAN) instead of separate networks for each class (Xu, et al., 2017). Although there was only one generative network in design, there were as many discriminative networks as the number of classes. The discriminative networks were updated simultaneously. MCGAN used feature matching loss (23) that extracts features dynamically from the discriminative network instead of getting from a fixed Visual Geometry Group (VGG) network.

A different perspective for training images provided to GANs that stands in front of the standard downscaling operation which is a bilinear interpolation (Bulat, Yang, & Tzimiropoulos, 2018). But it was discovered that this method was not good enough for real-world low-resolution images because of factors such as blur, compression artifacts, sensor noise, etc. To overcome this problem, a high-to-low generation network was designed for the degradation process before low-to-high evaluation.

TextSR is focused on text-image super-resolution and uses “text perceptual loss” inspired by perceptual loss (Wang, et al., 2019). TestSR uses ASTER (Shi, et al., 2018) as a base recognition network and also focused on text correction and ASTER is not a super-resolution study. But, with the help of image super-resolution, text recognition was taken one step further and generating text images was advanced to a better level.

3.2. Multi-Image or Video-based Super-Resolution GAN Studies

License plate number recognition was another challenging task in computer vision. Domain Priori GAN (DP-GAN) recovered license plate numbers, even unrecognizable by humans, from various viewpoints of multiple surveillance cameras with the help of other components in progressive vehicle search (Liu W. , Liu, Ma, & Cheng, 2017). The vehicle search was used for recognizing vehicles in combination with the plate numbers as well. DP-GAN has the capability of aligning plate numbers also. None of the performance metrics were included in our study. Human evaluation was used for recognizing numbers to measure system performance. At this point, MOS seems similar but not the same exactly.

A temporally coherent generative model was designed for fluid flow super-resolution (Xie, Franz, Chu, & Thuerey, 2018). This study was the first on GAN. It was realized on a four-dimensional dataset with two discriminative networks, one focuses on space and the other on temporal aspects. The loss function used in this study is a novel adversarial loss function that evaluated the temporal coherence of the outputs.

4. PERFORMANCE METRICS

Performance metrics used to compare the studies and network optimization metrics used by each network to produce better results are given below.

The performance metric was used to calculate how effective the proposed method was. Similar studies can be compared using these metrics. The matching of performance metric and studies are shown in Table 1.

Table 1

Performance metrics and studies in which they are used

Performance Metrics	Studies
Mean Opinion Score (MOS)	(Ledig, et al., 2016) (Zhang, Shao, Hu, & Gao, 2017) (Wang, et al., 2018)
Peak Signal-to-Noise Ratio (PSNR)	(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Wu, Duan, Liu, & Sun, 2017) (Bin, Weihai, Xingming, & Chun-Liang, 2017) (Zhang, Shao, Hu, & Gao, 2017) (Sajjadi, Scholkopf, & Hirsch, 2017) (Ma, Pan, Guo, & Lei, 2018) (Wang, et al., 2018) (Park, Son, Cho, Hong, & Lee, 2018) (Xu, et al., 2017) (Bulat, Yang, & Tzimiropoulos, 2018) (Wang, et al., 2019)
Structural Similarity Index (SSIM)	(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Wu, Duan, Liu, & Sun, 2017) (Zhang, Shao, Hu, & Gao, 2017) (Xu, et al., 2017) (Ma, Pan, Guo, & Lei, 2018) (Wang, et al., 2018) (Park, Son, Cho, Hong, & Lee, 2018) (Wang, et al., 2019)
Dark Channel Ratio (DCR)	(Xu, et al., 2017)
Confusion Matrix (Accuracy)	(Li, et al., 2017) (Bai, Zhang, Ding, & Ghanem, 2018)

4.1. Mean Opinion Score (MOS)

It is a subjective assessment average with score values numbered from one to five. Usually used in the telecommunications industry to measure user experience and averages user ratings (ITU-T, 2006). Similarly, it is a metric made with the average of the scores given by the individuals in the super-resolution studies. Therefore, a subjective judgment is made. A higher MOS value means that images are more similar.

4.2. Peak Signal-to-Noise Ratio (PSNR)

It is the ratio between the maximum possible signal strength and the noise distortion that affects image quality. Correlated with the average of squared errors in each pixel (Dosselmann & Yang, 2005). A higher PSNR value indicates that the two images are more similar.

PSNR, which measures the quality of compression in image compression processes, is still widely used due to its simple structure, especially in video images. It is one of the most reliable methods for comparing super-resolution images. This method is commonly used in analog systems and provides value in decibels (Huynh-Thu & Ghanbari, 2008).

First, the mean squared error (MSE) for PSNR is calculated as in equation (2) (PSNR, 2020). *m* and *n* represent the rows and columns in the input image. *I* is the input image, *K* is the image as a result that comes from the super-resolution process. The total square error is calculated for each pixel of the two images.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \quad (2)$$

The result obtained with (2) will be used in (3). *MAXI* represents the highest possible pixel value in the input image.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{3}$$

MAXI directly depends on the bit depth of the image. For images with a bit depth of *b*, the *MAXI* value can be obtained as in (4).

$$L = MAX_I = 2^b - 1 \tag{4}$$

4.3. Structural Similarity Index (SSIM)

It is used to measure the structural similarity between two images. Calculated by taking into account the characteristics of both images such as brightness, contrast, and structure (Wang, Bovik, Sheikh, & Simoncelli, 2004). A higher SSIM value means more similar images.

Detects changes between two images by focusing on luminance, contrast, and structure properties of images. The luminance equation is shown in (5), contrast in (6), and structure in (7).

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{7}$$

C1, *C2*, and *C3* are used in the equations in (8) to get rid of the uncertainty in case of $\mu_x^2+\mu_y^2$ result converges to zero. *K1* and *K2* constants are fixed numbers are less than 1. *L* can be calculated as *MAXI*, which represents the bit depth in the PSNR calculation expressed in (4).

$$C_1 = (K_1 L)^2$$

$$C_2 = (K_2 L)^2 \tag{8}$$

$$C_3 = C_2 \,/\, 2$$

Finally, SSIM is the weighted average of luminance, contrast and structure values as shown in equation (9).

$$SSIM(x,y) = [l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma] \tag{9}$$

In order to generalize the formula, by accepting the values a, β and γ equal to 1, the equation can be reduced to the state in (10). In this case, μ denotes the mean of the image while σ is standard deviation. So, σ_{xy} is the covariance of the *x* and *y* images.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

4.4. Dark Channel Ratio (DCR)

Used to express clarity and sharpness between two images. It is not a widely used technique in super-resolution studies. It was used to calculate how much the blurred images were clarified in blurred face images (Xu, et al., 2017). The DCR equation is defined in (11). x is the input image and x_{gt} is the ground-truth of x. $\varphi(x)$ is the dark channel of x. ε is set to 10^{-8} to avoid division by zero.

$$DCR(x, x_{gt}) = \frac{f_L(\varphi(x))}{f_L(\varphi(x)) + \varepsilon} \tag{11}$$

5. LOSS FUNCTIONS

In neural network design, some loss functions are used. Although some changes have been made on these loss functions, the actual loss function, which it is based on, has been highlighted in terms of meaningful pairing. The studies and optimization metrics are shown in Table 2, but studies use adversarial loss formulated as GAN formula in (1) not shown in this table.

Table 2

Super-resolution aimed loss functions and studies in which they are used

Loss Functions	Studies
Perceptual Loss	(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Li, et al., 2017) (Liu W. , Liu, Ma, & Cheng, 2017) (Wu, Duan, Liu, & Sun, 2017) (Sajjadi, Scholkopf, & Hirsch, 2017) (Wang, et al., 2018) (Park, Son, Cho, Hong, & Lee, 2018) (Wang, et al., 2019)
Pixel-wise Loss	(Xu, et al., 2017) (Bin, Weihai, Xingming, & Chun-Liang, 2017) (Sajjadi, Scholkopf, & Hirsch, 2017) (Zhang, Shao, Hu, & Gao, 2017) (Liu W. , Liu, Ma, & Cheng, 2017) (Bai, Zhang, Ding, & Ghanem, 2018) (Ma, Pan, Guo, & Lei, 2018) (Bulat, Yang, & Tzimiropoulos, 2018)
Charbonnier Loss	(Wu, Duan, Liu, & Sun, 2017)
Feature Matching Loss	(Xu, et al., 2017) (Park, Son, Cho, Hong, & Lee, 2018) (Bulat, Yang, & Tzimiropoulos, 2018)
Texture Matching Loss	(Sajjadi, Scholkopf, & Hirsch, 2017) (Xie, Franz, Chu, & Thuerey, 2018)

5.1. Perceptual Loss Function

Systems that perform performance evaluation per pixel are not able to make a fair assessment. With one-pixel shift, calculating a significant difference between the original image and the generated image, even if it contains all of the visual features created in its original form (Johnson, Alahi, & Fei-Fei, 2016). During the training phase, using perceptual loss focused on higher-level features rather than pixel data. Therefore, perceptual loss deals with error in property space rather than pixel space. However, despite the increase in success, with the emergence of the optimization problem, the processing time is prolonged.

The perceptual loss calculated by a loss network collects all the squared errors and averages them. In order to calculate feature and style cost in a loss network (ϕ), two different equations (13) and (16) are used. The weight values (W) in the network producing super-resolution is minimized by the stochastic gradient descent method (Johnson, Alahi, & Fei-Fei, 2016). W is shown as (12).

$$W^* = arg \min_W E_{x,\{y_i\}} \left[\sum_{i=1} \lambda_i l_i(f_w(x), y_i) \right] \tag{12}$$

Each loss function calculates a single value. While calculating this value, the difference between the original image and the generated image is examined. The loss functions are deep convolutional neural networks. However, there is no need to calculate style loss for single-image super-resolution (Johnson, Alahi, & Fei-Fei, 2016).

In the feature reconstruction loss function, the feature differences calculated by the loss network are obtained. For the *x* image, the value generated in the *j*-th layer is represented by $\phi_j(x)$. If the *j*-th layer is a convolutional neural network, a feature map of *Cj* x *Hj* x *Wj* is obtained. Here, *Cj* defines the color depth, *Hj* is the height of the image, and *Wj* is the width of the image, each in the *j*-th layer. The output from the *x* image in a network trained according to *W* weights is expressed as $\hat{y} = f_W(x)$. The desired image *y* is obtained by feature loss (13) between these two images (Johnson, Alahi, & Fei-Fei, 2016).

$$l_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \left\| \phi_j(\hat{y}) - \phi_j(y) \right\|_2^2 \tag{13}$$

Perceptual loss focuses on the structural features of the image. Differences in color, pattern, and shape are excluded.

There is no need for style loss calculation for super-resolution. However, in combination with style reconstruction loss, it makes a significant difference. Equation (16) is used for style reconstruction loss.

Firstly, it is necessary to define gram matrix. $G_j^{\phi}(x)$ is a matrix of size *Cj* x *Cj* which is defined as shown in (14).

$$G_j^{\phi}(x)_{c,c\square} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{\mathrm{h,w,c}\square} \tag{14}$$

To make a more efficient calculation, if we put the matrix $\phi_j(x)$ into the form $\psi = C_j \; x \; H_j W_j$, the function $G_j^{\phi}(x)$ is written as (15).

$$G_j^{\phi}(x) = \frac{\psi \psi^T}{C_j H_j W_j} \tag{15}$$

Thus, the style reconstruction loss is calculated as in (16) by the square Frobenius norm of the difference of defined gram matrices.

$$l_{style}^{\phi,j}(\hat{y}, y) = \left\| G_j^{\phi}(\hat{y}) - G_j^{\phi}(y) \right\|_F^2 \tag{16}$$

5.2. Pixel-wise Loss (Mean Squared Error)

It is the metric used to calculate errors in pixel space. The total error of each matching pixel is obtained and used as shown in (17). This expression is the normalized Euclidean distance of two images of the same size. This method is only possible if we have a ground truth image that can be compared (Johnson, Alahi, & Fei-Fei, 2016).

$$l_{pixel}(\hat{y}, y) = \frac{\|\hat{y} - y\|_2^2}{CHW} \tag{17}$$

In equation (18), another content loss function is also defined as convolutional content loss (Zhang, Shao, Hu, & Gao, 2017). $G_{\theta_G}(I^{LR})$ is the high-resolution output image. This equation is also used to combine with other formulas to improve the performance. Zhang et al. (2017) used (18) in combination with cross-entropy while Liu et al. (2017) used adversarial loss.

$$l_X^{LR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_X^{LR} - Conv(G_{\theta_G}(I^{LR})_{x,y}))^2 \tag{18}$$

Pixel-based loss functions are far from producing an efficient result for the human eye but may give better results over PSNR and SSIM because they are focused on one-to-one mapping.

5.3. Charbonnier Loss

To ensure the correctness of low-frequency details on result image, Charbonnier loss is used as a content loss function by Wu et al. (2017). The Charbonnier loss is defined in (19).

$$l_y(y, \hat{y}) = E_{z,y\sim p_{data}(z,y)}(\rho(y - G(z))) \quad (19)$$

y is denoted as ground truth image and *G(z)* is the constructed image. Here, the Charbonnier penalty function is needed to be defined as shown in (20).

$$\rho(x) = \sqrt{x^2 + \varepsilon^2} \quad (20)$$

5.4. Feature Matching Loss

It is an error calculation method developed based on the features in the image on the idea that pixel-based study does not give good results. The feature matching loss (21) is defined and then adopted to original GAN formulation in (1) to recover more realistic details (Xu, et al., 2017).

$$l_{feat\_match} = \frac{1}{N}\sum_{i=1}^{N} \left\| \phi_\theta^l \left(G_\omega(y^i) \right) - \phi_\theta^l(x^i) \right\|^2 \quad (21)$$

ϕ_θ^l is the feature response to *x* at the *l*-th layer.

5.5. Texture Matching Loss

Source and generated image should be the same style to apply matching loss (Sajjadi, Scholkopf, & Hirsch, 2017). Gram matrix that defines correlations between different feature channels is shown in (22).

$$G(F) = FF^T \quad (22)$$

This Gram matrix is used in (23).

$$l_{texture} = \left\| G\left(\phi(I_{est})\right) - G\left(\phi(I_{HR})\right) \right\|_2^2 \quad (23)$$

With this loss method, generating more realistic results are possible. Xie et al. (2018) used a mathematically similar loss function to texture matching loss.

6. DATASETS

The datasets used in super-resolution and matching of the studies are shown in Table 3. SET5 and SET14 datasets include low and high-resolution images (Bevilacqua, Roumy, Guillemot, & Alberi-Morel, 2012). BSDS dataset has different variations such as BSDS100, BSDS200, BSDS300, and BSDS500 and they are created for image segmentation and boundary detection (The Berkeley Segmentation Dataset and Benchmark, 2019). Tsinghua-Tencent includes traffic signs inside a vehicle-view (Traffic-Sign Detection and Classification in the Wild, 2019). The Caltech benchmark dataset includes vehicle-view videos for detecting pedestrians (Caltech Pedestrian Detection Benchmark, 2019). The Manga109 dataset includes commercially made Japanese manga images between the 1970s and 2010s (Dataset, 2019). The T91 is created for training neural networks with high-resolution images (Kaggle - T91 Image Dataset, 2019). The General100 has uncompressed 100 BMP formatted images in good quality (Dong, Loy, & Tang, 2016). The Dataset of Hradis included scientific papers for text-deblurring (Hradiš, Kotera, Zemcık, & Šroubek, 2015). The CelebA dataset included more than 200K celebrity images with annotations (Large-scale CelebFaces Attributes (CelebA) Dataset, 2019). The Wider Face was created for face benchmark from a publicly available WIDER dataset (Yang, Luo, Loy, & Tang, 2016). The LS3D-W is a 3D facial landmark dataset (Bulat & Tzimiropoulos, 2017). The DIV2K was proposed for benchmarking on single-image super-resolution which includes 1000 images at 2K resolution and the test set is not publicly available (Agustsson & Timofte, 2017). The UC Merced dataset included remote sensing images (UC Merced Land Use Dataset, 2019). The Flickr2K was collected from the Flickr website and consisted of 2650 images at 2K resolution (Timofte, Agustsson, Van Gool, Yang, & Zhang, 2017). The OST was collected from search engines containing over ten thousand images (Wang, Yu, Dong, & Change Loy, 2018). The Urban100 included real human-made structures containing 100 images (Huang, Singh, & Ahuja, 2015). The ImageNet has over 14 million images built in a hierarchical structure for object recognition tasks (Deng, et al., 2009).

Ten different text-based datasets are included as shown in Table 3 (Wang, et al., 2019). Synth90k (Jaderberg, Simonyan, Vedaldi, & Zisserman, 2015) and SynthText (Gupta, Vedaldi, & Zisserman, 2016) are synthetic text datasets. IIT5k-Words includes texts with lexicons (Mishra, Alahari, & Jawahar, 2012). Street View Text is collected from Google Street View (Wang, Babenko, & Belongie, 2011). ICDAR 2003 has cropped words that have non-alphanumeric characters or less than three characters with lexicons (Lucas, et al., 2005). The ICDAR 2013 is an advanced form of ICDAR 2003 with no lexicons (Karatzas, et al., 2013). The ICDAR 2015 contained irregular texts within bounding boxes (Karatzas, et al., 2015). The SVT-Perspective was a benchmark dataset for recognizing perspective texts (Phan, Shivakumara, Tian, & Tan, 2013). The CUTE80 contained curved texts (Risnumawan, Shivakumara, Chan, & Tan, 2014). The VeRi dataset was created for vehicle re-identification but used as license plate number recognizing and this dataset has 50,000 images of 776 vehicles from 20 different cameras (Liu X. , Liu, Mei, & Ma, 2016).

Datasets that are directly used for super-resolution tasks are included in Table *3* and Table *4*.

Table 3

Datasets and their domains used in super-resolution tasks using GAN

Datasets		**Studies**	**Text**	**Face**	**Object Recognition**	**Traffic Signs**	**Pedestrian Detection**	**Drawing**	**Urban**	**Training for Super-resolution**	**Image segmentation**	**Remote Sensing**	**Image Restoration**	**Annotated images**	**High res included**
SET5		(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Sajjadi, Scholkopf, & Hirsch, 2017) (Park, Son, Cho, Hong, & Lee, 2018)								✓					
SET14		(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Sajjadi, Scholkopf, & Hirsch, 2017) (Park, Son, Cho, Hong, & Lee, 2018)								✓					
BSDS	100	(Ledig, et al., 2016) (Johnson, Alahi, & Fei-Fei, 2016) (Sajjadi, Scholkopf, & Hirsch, 2017) (Wang, et al., 2018)									✓				
	200	(Wu, Duan, Liu, & Sun, 2017)									✓				
	300	(Ledig, et al., 2016) (Park, Son, Cho, Hong, & Lee, 2018)									✓				
Tsinghua-Tencent 100K		(Li, et al., 2017)				✓									
Caltech benchmark		(Li, et al., 2017)					✓								
Manga109		(Li, et al., 2017)						✓							
T91		(Wu, Duan, Liu, & Sun, 2017)								✓					
General100		(Wu, Duan, Liu, & Sun, 2017)								✓					
CelebA		(Bin, Weihai, Xingming, & Chun-Liang, 2017) (Zhang, Shao, Hu, & Gao, 2017) (Xu, et al., 2017)		✓										✓	

Wider Face		(Bai, Zhang, Ding, & Ghanem, 2018)		✓										
DIV2K		(Ma, Pan, Guo, & Lei, 2018) (Wang, et al., 2018) (Park, Son, Cho, Hong, & Lee, 2018)										✓		✓
UC Merced (Remote Sensing Dataset)		(Ma, Pan, Guo, & Lei, 2018)									✓			
Flicker2K		(Wang, et al., 2018)											✓	✓
OST (Outdoor Scene Training)		(Wang, et al., 2018)							✓					✓
Urban100		(Sajjadi, Scholkopf, & Hirsch, 2017) (Wang, et al., 2018)						✓						✓
ImageNet		(Park, Son, Cho, Hong, & Lee, 2018)			✓									
Dataset of Hradis		(Xu, et al., 2017)	✓											
Synth90k		(Wang, et al., 2019)	✓											
SynthText		(Wang, et al., 2019)	✓											
IIIT5k-Words		(Wang, et al., 2019)	✓											
Street View Text		(Wang, et al., 2019)	✓											
ICDAR	2003	(Wang, et al., 2019)	✓											
	2013	(Wang, et al., 2019)	✓											
	2015	(Wang, et al., 2019)	✓											
SVT-Perspective		(Wang, et al., 2019)	✓											
CUTE80		(Wang, et al., 2019)	✓											
VeRi		(Liu X. , Liu, Mei, & Ma, 2016)	✓											
LS3D-W		(Bulat, Yang, & Tzimiropoulos, 2018)		✓										
Four-Dimensional Fluid Dataset		(Xie, Franz, Chu, & Thuerey, 2018)										✓		
Number of Unique Studies			**3**	**5**	**1**	**1**	**1**	**1**	**2**	**6**	**6**	**1**	**4**	**4**	**4**

According to information in Table 3, different dataset counts are shown in Table 4. Goodfellow, et al. (2014) is not a super-resolution study, so not included. The average working dataset count per study is 2.81.

Table 4

Studies and the number of different datasets used for super-resolution

Studies	Dataset Count
(Ledig, et al., 2016)	4
(Johnson, Alahi, & Fei-Fei, 2016)	3
(Liu X. , Liu, Mei, & Ma, 2016)	1
(Li, et al., 2017)	3
(Wu, Duan, Liu, & Sun, 2017)	3
(Bin, Weihai, Xingming, & Chun-Liang, 2017)	1
(Zhang, Shao, Hu, & Gao, 2017)	1
(Sajjadi, Scholkopf, & Hirsch, 2017)	4
(Xu, et al., 2017)	1
(Bai, Zhang, Ding, & Ghanem, 2018)	1
(Ma, Pan, Guo, & Lei, 2018)	2
(Wang, et al., 2018)	5
(Park, Son, Cho, Hong, & Lee, 2018)	5
(Bulat, Yang, & Tzimiropoulos, 2018)	1
(Xie, Franz, Chu, & Thuerey, 2018)	1
(Wang, et al., 2019)	9
Average:	**2.81**

7. CONCLUSIONS

In the literature, there was not any surveys on GAN based super-resolution studies. This study aimed to overcome this shortcoming.

The popularity of GAN is increasing day by day. The most powerful part of the GAN is having two simultaneous networks trying to overcome each other. This process is likened to mini-max games and at the end of the process, more realistic data can be generated. Finally,, the most important point on GAN is the loss functions used in it. Thanks to the successful implementation of loss functions, the neural network generates more realistic results.

In addition to producing artificial data, GAN has been used in different working areas and has many variations with its applications to reduce noise in the data and clean up the data containing noise. GAN studies on various datasets are the biggest indicators of this situation. On the basis of good results of artificial neural networks, there is a need for training with a lot of data to make the pattern more meaningful. For example, when super-resolution is applied using GAN in faces that are not clearly apparent in photographs, a solution such as increasing the limited data can be considered.

Super-resolution is a process of the increment of details in images. GAN based super-resolution studies produced more detailed results than other traditional interpolation methods and state-of-the-art neural network-based studies. But some performance metrics cannot measure the quality perceptually. Most common performance metrics like PSNR and SSIM may be higher even if the quality is poor. The source of this kind of problem is about the loss functions used in neural networks. If the loss function in any neural network is similar to the performance metric, the neural network learns to increase its performance score, but the result may not as good as the score. It seems that MOS (mean opinion score) is more useful and accurate as a super-resolution performance metric.

Hakem Değerlendirmesi: Dış bağımsız.
Çıkar Çatışması: Yazarlar çıkar çatışması bildirmemiştir.
Finansal Destek: Yazarlar bu çalışma için finansal destek almadığını beyan etmiştir.

Peer-review: Externally peer-reviewed.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

References

Agustsson, E., & Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*(126-135).

Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Finding tiny faces in the wild with generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Bevilacqua, M., Roumy, A., Guillemot, C., & Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *Proceedings of the 23rd British Machine Vision Conference (BMVC).*

Bin, H., Weihai, C., Xingming, W., & Chun-Liang, L. (2017). *High-quality face image generated with conditional boundary equilibrium generative adversarial networks.* Pattern Recognition Letters.

Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., & Zelnik-Manor, L. (2018). The 2018 pirm challenge on perceptual image super-resolution. *Proceedings of the European Conference on Computer Vision (ECCV).*

Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *International Conference on Computer Vision.*

Bulat, A., Yang, J., & Tzimiropoulos, G. (2018). To learn image super-resolution, use a gan to learn how to do image degradation first. *Proceedings of the European conference on computer vision (ECCV)*, (pp. 185-200).

Caltech Pedestrian Detection Benchmark. (2019, 12 23). Retrieved from http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

Dataset, M. -J. (2019, 12 23). Retrieved from http://www.manga109.org/en/

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition.*

Dong, C., Loy, C., & Tang, X. (2016, 12 23). Accelerating the Super-Resolution Convolutional Neural Network. *European Conference on Computer Vision (ECCV).*

Dosselmann, R., & Yang, X. D. (2005). Existing and emerging image quality metrics. *Canadian Conference on Electrical and Computer Engineering.*

Gerchberg, R. W. (1974). Super-resolution through error energy reduction. *Optica Acta: International Journal of Optics, 21*(9), 709-720.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations.*

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 2672-2680.

Gotoh, T., & Okutomi, M. (2004). Direct super-resolution and registration using raw CFA images. *Computer Vision and Pattern Recognition (CVPR)*, *2*.

Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. *IEEE Conference on Computer Vision and Pattern Recognition.*

Hradiš, M., Kotera, J., Zemcık, P., & Šroubek, F. (2015). Convolutional neural networks for direct text deblurring. *Proceedings of BMVC, 10*(2).

Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. *44(13)*, 800-801. Electronics letters.

ITU-T. (2006). *Rec. P.10: Vocabulary for performance and quality of service.*

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2015). Deep structured output learning for unconstrained text recognition. *International Conference on Learning Representations (ICLR).*

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European conference on computer vision.*

Kaggle - T91 Image Dataset. (2019, 12 23). Retrieved from https://www.kaggle.com/ll01dm/t91-image-dataset

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., & Shafait, F. e. (2015). ICDAR 2015 competition on robust reading. *International Conference on Document Analysis and Recognition (ICDAR).*

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., & Mestre, S. R. (2013). ICDAR 2013 robust reading competition. *International Conference on Document Analysis and Recognition.*

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning. Representations.*

Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning, 12*(4), 307-392.

Large-scale CelebFaces Attributes (CelebA) Dataset. (2019, 12 23). Retrieved from http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., & Shi, W. (2016). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4681-4690).

Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Liu, W., Liu, X., Ma, H., & Cheng, P. (2017). Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN. *Proceedings of the 25th ACM international conference on Multimedia*, (pp. 1618-1626).

Liu, X., Liu, W., Mei, T., & Ma, H. (2016). A deep learning-based approach to progressive vehicle re-identification for urban surveillance. *European conference on computer vision*, (pp. 869-884).

Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., . . . Lin, X. (2005). ICDAR 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR), 2*(105-122), 7.

Ma, W., Pan, Z., Guo, J., & Lei, B. (2018). Super-resolution of remote sensing images based on transferred generative adversarial network. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium.*

Mishra, A., Alahari, K., & Jawahar, C. V. (2012). Top-down and bottom-up cues for scene text recognition. *IEEE Conference on Computer Vision and Pattern Recognition.*

Mjolsness, E. (1985). *Neural networks, pattern recognition, and fingerprint hallucination.* Diss. California Institute of Technology.

Nasrollahi, K., & Moeslund, T. B. (2014). Super-resolution: a comprehensive survey. *Machine vision and applications, 25*(6), 1423-1468.

Park, S. J., Son, H., Cho, S., Hong, K. S., & Lee, S. (2018). Srfeat: Single image super-resolution with feature discrimination. *European Conference on Computer Vision (ECCV).*

Phan, T., Shivakumara, P., Tian, S., & Tan, C. (2013). Recognizing text with perspective distortion in natural scenes. *International Conference on Computer Vision.*

Protter, M., Elad, M., Takeda, H., & Milanfar, P. (2008). Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on image processing, 18*(1), 36-51.

PSNR. (2020, 7 6). Retrieved 11 23, 2019, from MathWorks: https://www.mathworks.com/help/vision/ref/psnr.html

Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications, 18*(8027-8048), 41.

Sajjadi, M. S., Scholkopf, B., & Hirsch, M. (2017). Enhancenet: Single image super-resolution through automated texture synthesis. *International Conference on Computer Vision (ICCV).*

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., & Bai, X. (2018). Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence, 41*(9), 2035-2048.

The Berkeley Segmentation Dataset and Benchmark. (2019, 12 23). Retrieved from https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

Timofte, R., Agustsson, E., Van Gool, L., Yang, M. H., & Zhang, L. (2017). Ntire 2017 challenge on single image super-resolution: Methods and results. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 114-125.

Traffic-Sign Detection and Classification in the Wild. (2019, 12 23). Retrieved from https://cg.cs.tsinghua.edu.cn/traffic-sign/

UC Merced Land Use Dataset. (2019, 12 23). Retrieved from http://weegee.vision.ucmerced.edu/datasets/landuse.html

Wang, K., Babenko, B., & Belongie, S. (2011). End-to-end scene text recognition. *International Conference on Computer Vision.*

Wang, W., Xie, E., Sun, P., Wang, W., Tian, L., Shen, C., & Luo, P. (2019). *TextSR: Content-Aware Text Super-Resolution Guided by Recognition.* arXiv preprint.

Wang, X., Yu, K., Dong, C., & Change Loy, C. (2018). Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., & Change Loy, C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer Vision.*

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing, 13*(4), 600-612.

Wu, B., Duan, H., Liu, Z., & Sun, G. (2017). *Srpgan: Perceptual generative adversarial network for single image super resolution.* arXiv preprint.

Xie, Y., Franz, E., Chu, M., & Thuerey, N. (2018). tempoGAN: A temporally coherent, volumetric gan for super-resolution fluid flow. *ACM Transactions on Graphics (TOG), 37*(4), 1-15.

Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., & Yang, M. H. (2017). Learning to super-resolve blurry face and text images. *Proceedings of the IEEE International Conference on Computer Vision.*

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A Face Detection Benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Zhang, D., Shao, J., Hu, G., & Gao, L. (2017). Sharp and real image super-resolution using generative adversarial network. *International Conference on Neural Information Processing*, (pp. 217-226).