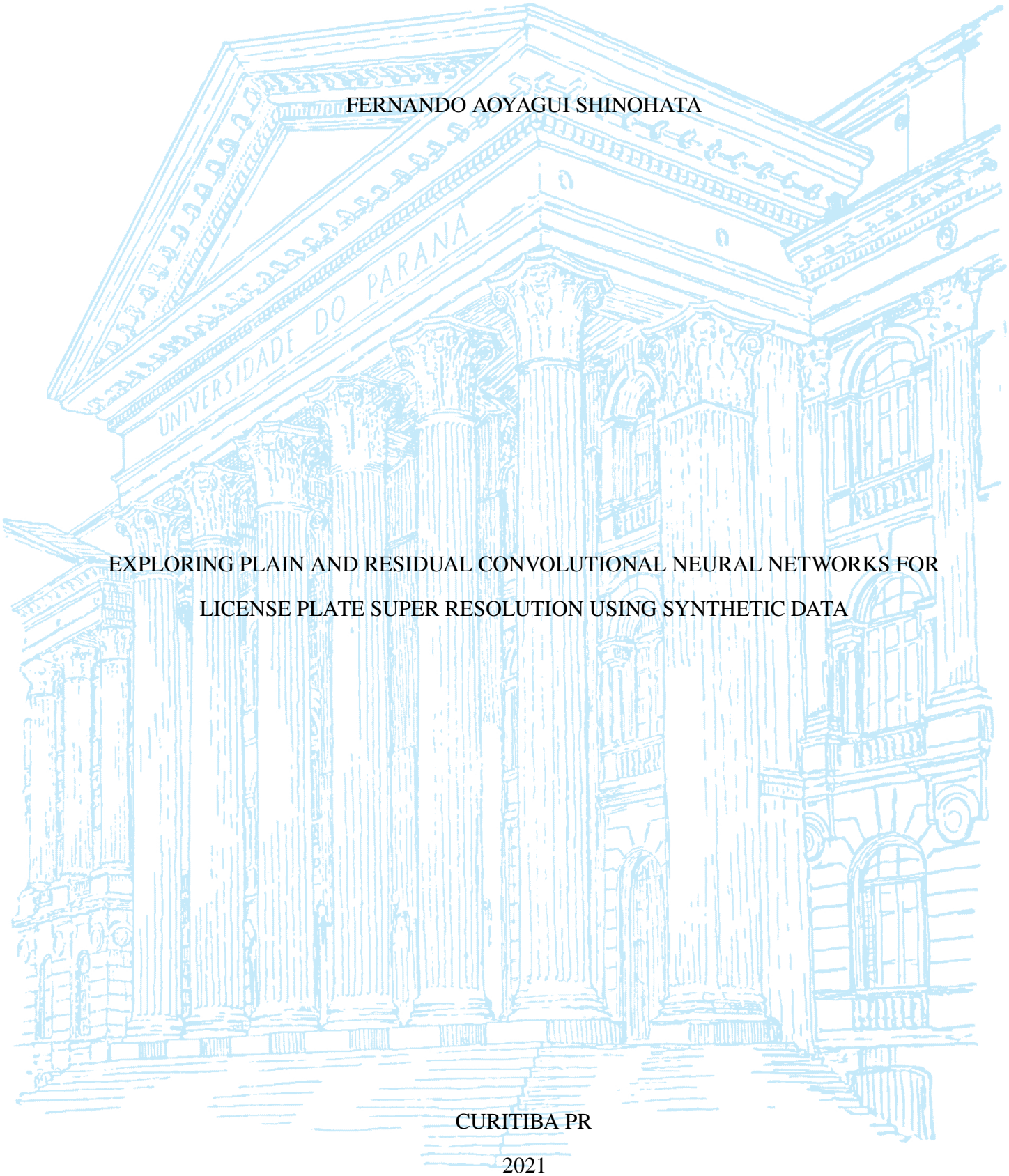
UNIVERSIDADE FEDERAL DO PARANÁ

FERNANDO AOYAGUI SHINOHATA



EXPLORING PLAIN AND RESIDUAL CONVOLUTIONAL NEURAL NETWORKS FOR LICENSE PLATE SUPER RESOLUTION USING SYNTHETIC DATA

CURITIBA PR

2021

FERNANDO AOYAGUI SHINOHATA

EXPLORING PLAIN AND RESIDUAL CONVOLUTIONAL NEURAL NETWORKS FOR LICENSE PLATE SUPER RESOLUTION USING SYNTHETIC DATA

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti Gomes.

CURITIBA PR

2021

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: **EXPLORING PLAIN AND RESIDUAL CONVOLUTIONAL NEURAL NETWORKS FOR LICENSE PLATE SUPER RESOLUTION USING SYNTHETIC DATA**

Autor(es):
GRR_20165388 None: FERNANDO AOYAGUI SHINOHATTA
GRR_____ Nome:_____
GRR_____ Nome:_____

Apresentação: Data: 25 / 03 / 2021 Hora: 14:00 Local: VIRTUAL-meet / DINF / UFPR

Orientador: DAVID MENOTTI GOMES
Membro 1: RAYSON BARTOSKI LAROCA DOS SANTOS
Membro 2: VALFRIDE WALLACE DO NASCIMENTO

(nome) (assinatura)

Prof. David Menotti Gomes
Dep. de Informática
Mat.: 204841 - UFPR

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)	32	32	32	32
Referência Bibliográfica	(00-10)	05	07	06	06
Formato	(00-05)	04	04	04	04
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)	13	15	14	14
Desenvolvimento do Assunto	(00-05)	05	05	05	05
Técnica de Apresentação	(00-03)	02	02	02	02
Uso do Tempo	(00-02)	02	02	02	02
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)	15	**********	**********	15
NOTA FINAL		**********	**********	**********	**80**

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

I dedicate this work to my family, for all the support they gave me so I could focus on my education.

ACKNOWLEDGEMENTS

I thank all professors that I had contact with in university, for being the pillar and foundation to allow me to reach this point.

I thank my friends for all the fun moments we had together, making university not just a place you go to attend lessons.

I thank my family for all the support given to me in many occasions, allowing me to focus completely on my education.

A very special thanks to my advisor, David Menotti Gomes, for the astounding patience to lead me during this work, as well as everything taught to me prior to this day.

RESUMO

O objetivo deste trabalho é testar e comparar duas arquiteturas de redes neurais convolucionais bem conhecidas e utilizadas na tarefa de super resolver imagens de placas de veículos, também conhecida como *License Plate Super Resolution*. O que motiva este trabalho é o fato de Super Resolução ser um problema largamente explorado hoje em dia, com uma variedade de aplicações que incluem até mesmo ganho de performance em *video games*. Um cenário de aplicação é o já citado *License Plate Super Resolution*, onde o objeto-alvo a ser super resolvido é uma placa veicular. É muito comum que filmagens the câmeras de segurança possuam baixa qualidade, algumas vezes tornando a leitura dos caracteres de uma placa impossível. Ter uma solução de Super Resolução para este cenário é útil especialmente para aplicações forenses. A comparação entre ambas arquiteturas será feita utilizando duas configurações customizadas de rede, feitas para serem o mais próximo possível uma da outra. O treinamento também será feito sobre exatamente o mesmo conjunto de dados com exatamente o mesmo tempo de treinamento, felizmente permitindo uma comparação de performance justa. Com os resultados apresentados aqui, uma breve análise sobre cada arquitetura e seus problemas será dada.

Palavras-chave: Super resolução. Placas de veículos. Rede Neural Convolucional. Dados sintéticos.

ABSTRACT

The objective of this work is to test and compare two well known and used convolutional neural network architectures for the task of super resolving vehicle license plate images, also known as License Plate Super Resolution. What motivates this work is the fact that Super Resolution is a widely explored problem nowadays, with a variety of applications which include even performance gain in video games. One application scenario is the already cited License Plate Super Resolution, where the target object to be super resolved is a vehicle license plate. It is very common for security camera footage to have low quality, sometimes making it impossible to read the characters of the license plate. Having a Super Resolution solution for this scenario is useful especially for forensic applications. The comparison between both architectures will be done using two custom network configurations, made to be as close as possible to one another. The training will also be done over the exact same dataset for the exact same amount of training time, hopefully allowing a fair performance comparison. With the results presented here, a brief analysis over each architecture and their problems is given.

Keywords: Super Resolution. Vehicle license plates. Convolutional Neural Network. Synthetic data.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
SR	Super Resolution
LR	Low Resolution
HR	High Resolution
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
ReLU	Rectified Linear Unit
LP	License Plate

CONTENTS

1 INTRODUCTION

Super Resolution is a image enhancement technique where a low resolution image is upscaled (or reconstructed) to a higher resolution, with more details that may or may not be present in the low resolution version. Most recent works use convolutional neural network architectures to learn how to super resolve images for specific scenarios. This technique has a variety of applications, including performance boost for video games when using neural network solutions, which is done by the Deep Learning Super Sampling (DLSS) technology by NVIDIA (NVIDIA, 2020).

When the target object for Super Resolution is a license plate, there is the extra restriction of character readability. This scenario has been explored and tested by many works, since it is very common to have low quality footage of car traffic in which the LPs cannot be read by human eyes. Having a solution for this problem is especially useful for forensic applications.

1.1 GOALS AND MOTIVATION

The objective of this work is to test two network architectures that are commonly proposed for Super Resolution solutions: plain CNNs and residual CNNs. Since most works use custom architectures, there are always many differences that may cause unfair comparisons, which is why the networks built here are made to be as close as possible to one another.

Although there are well known problems about using each architecture for License Plate Super Resolution, the objective is to achieve a decent comparison between both, showing their differences, advantages and disadvantages.

1.2 DOCUMENT STRUCTURE

The next chapter, Background, gives a brief overview about generic super resolution and license plate super resolution, along with its challenges. A simple explanation about what are convolutional neural networks and their purpose is also briefly described.

In Chapter 3, Bibliography Review, pioneer methods for super resolution are discussed, along with a few mentions of recent works and their performance.

In Chapter 4, Methodology, the network architectures built for comparison are presented, and their training method, parameters, dataset and time are given.

In Chapter 5, Experiments, the results obtained after the training phase are shown, along with a comparison between the networks and some speculations about the differences in the output images.

2 BACKGROUND

2.1 SUPER RESOLUTION

Super Resolution refers to a technique that increases the resolution of images based on one or multiple samples. The main objective is to obtain an image with higher resolution, while also achieving better quality. For that reason, simple image rescaling algorithms like bicubic and bilinear interpolation are not classified as super resolution algorithms.

When a better quality is also desired, rescaling solutions are switched with recursive reconstruction, convolutional neural network and other methods that allow more details to be generated in the final high resolution sample.

Like image rescaling, works related to super resolving or "enhancing" images are not new. It is considered as an ill-posed problem, since low resolution (LR) images lack information about the high resolution (HR) counterpart.

The ill-posed part is closely related to a surjective function inversion problem, since multiple HR images end up being mapped (or undersampled) to the same LR image, because the HR set is bigger than the LR set. In this context, super resolving a LR image means selecting one of the HR images from the inversion of the undersampling function.

Of course, this is not a trivial problem. You need something to counter the lack of information in order to achieve good fidelity. One option is to use a set of similar LR images that contain the "area" that you wish to super resolve. This method is called *Multi-Frame Super Resolution*, and it is commonly used when you have low resolution video footage or many images of the same thing with slightly different angles. Some examples are the works of (Avrin e Dinstein, 1998; Zeng e Lu, 2012; Seibel et al., 2015; Li e Wang, 2017).

In cases where you do not have access to a set of images containing the target object to be super resolved (*i.e.* you only have one image to extract the data from), there is the problem of how to fill the HR pixel grid. The technique that super resolve images this way is called *Single-Frame Super Resolution*. Some examples are the works of (Chuang et al., 2014; Zou et al., 2019; Zhang e Cai, 2020; Ren et al., 2019).

2.1.1 Multi-Frame Super Resolution

This kind of method is suitable for cases where you have access to either a sequence of images or video footage, with low difference between each element in the sequence. As the name says, it uses multiple, similar samples in order to reconstruct a single, higher resolution one.

There are many factors that can lead to the acquisition of low quality images or video footage such as hardware quality, interference during transmission, and lack of storage space for higher quality data.

Satellite data, for example, is transmitted wirelessly, and thus is much more vulnerable to signal corruption. Satellite imagery usually forms a sequence with overlapping areas, and these can be used to collect extra information about the image in order to generate another image with more details (*i.e.* a super resolved version). (Kim et al., 1990) explored this topic in detail, also considering noise in the image sequence.

In this scenario, if the satellite is taking photos with little or no zooming at all, the images taken will have almost nonexistent local motion between each sample in the sequence. Hence, using motion field estimation is not necessary considering that each image in the sequence was taken within a short time of each other.

Another common scenario is footage from security cameras. They usually use very low bitrates during recording in order to minimize storage usage, generating high resolution footage with very poor quality. In this case, super resolving the footage requires local motion to be considered, especially if the camera is recording a road or sidewalk.

This scenario is especially tricky because the low quality makes detection of object edges a very challenging problem. Without proper detection of the objects, estimating motion fields to compensate the images can lead to weird artifacts appearing in the super resolved version.

2.1.2 Single-Frame Super Resolution

As the name suggests, single-frame super resolution attempts to obtain a higher resolution image based on a single sample. Image rescaling algorithms solve this problem without generating new information, while neural network based methods use the knowledge of the network to fill the pixel grid, thus generating information that may not be present in the LR image.

2.2 CONVOLUTIONAL NEURAL NETWORK (CNN)

A CNN is a type of Artificial Neural Network (ANN) designed specifically for image processing. In other words, its structure assumes that the input will consists of images. The overall architecture of a CNN has three types of layers: convolutional layer, pooling layer and fully-connected layers. By stacking these layers you obtain a CNN architecture (O'Shea e Nash, 2015).

Like other ANNs, a Convolutional Neural Network does its hyperparameter tuning by using a loss function for backpropagation. Since the input is assumed to be a image, the usual choices for the loss function are the ones that can give the best evaluation about what is expected of the output image. For SR using single CNNs, some commonly used loss functions are Mean Absolute Error (MAE), Mean Square Error (MSE), Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR). There are also loss functions specialized for other architectures, like the Adversarial Loss, crafted for training a Generative Adversarial Network (GAN).

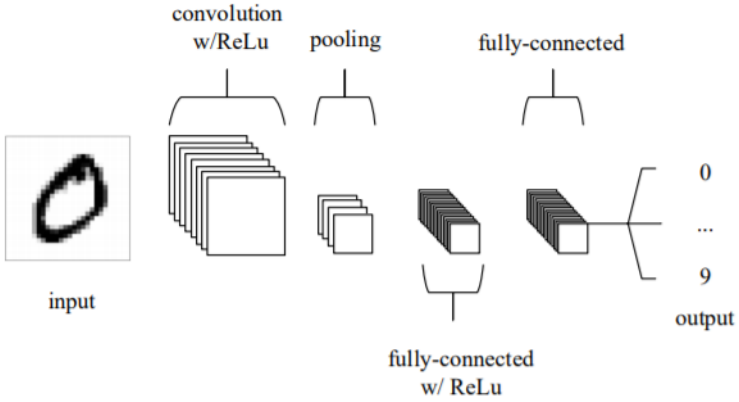


Figure 2.1: Example of a CNN with four layers (O'Shea e Nash, 2015).

2.2.1 The convolutional layer

This layer is the main "component" of a CNN. It uses learnable kernels followed by an activation function (commonly ReLU) to process the input into a 2D activation map. A **kernel** is a low dimensionality vector that is applied by "sliding" it along the input vector, as shown in Figure 2.2.

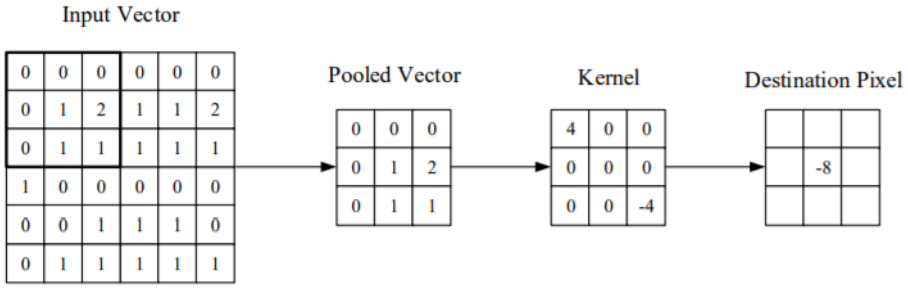


Figure 2.2: Visual representation of a convolutional layer. The kernel is placed on the input vector and then used to calculate a weighted sum of itself and the nearby pixels (O'Shea e Nash, 2015).

The result, as said before, will be a new vector whose values form what is called a 2D activation map. These maps store characteristics deemed important by the CNN during the training phase.

2.2.2 The pooling layer

This layer is responsible for aggregating the data from the activation map (obtained by the convolutional layer) and reducing its dimensionality, usually by using the MAX function or a normalization function like L1 or L2.

Since it reduces the complexity of the incoming input, we can say that this layer is destructive, and thus there is a need to be cautious about how much you want to simplify your data.

2.2.3 The fully-connected layer

It is a layer that contains neurons that are directly connected to the neurons in its adjacent layers, just like a regular ANN layer. Each neuron has a matrix of weights and bias that are adjusted during the training phase to transform the input to the expected output.

This is the part where the input is processed and classified. In most cases, you want to feed this layer the output of a pooling layer, as the complexity of the input will then be reduced, requiring less overall neurons to be trained. The greatly reduced amount of neurons required to train in this layer is the main difference between a CNN and an ANN. For super resolution, however, this layer is not used, since the desired output is an image with the same or higher resolution than the input, and not just a classification vector.

Attempting to use this layer for SR defeats the objective of all prior steps that reduces the complexity of the input, since the output dimensionality will still require more neurons to be allocated to allow the layer to properly calculate it.

2.2.4 About Training

Training a single CNN architecture is the same process as any other singular ANN. Singular because some architectures like Generative Adversarial Networks (GAN) use more than one independent networks that need to learn different things during training.

For generic Super Resolution, generally you just need a decent size image dataset with a good variety of environments. Each image on the dataset is labeled as the ground truth, and a lower resolution version is generated to use during the training phase. The lower resolution images are fed to the CNN, and its output images are then compared with the ground truth

versions. The same process is applied for "specialized" Super Resolution. Only the training dataset is changed to one that contains only the objects that you wish your network to learn.

One advantage of training a CNN for Super Resolution is the easiness of applying data augmentation. Changing brightness, rotation, saturation and other image aspects can be done quickly, and it is a good way to artificially increase your dataset size without compromising training time.

2.3 RESIDUAL CONVOLUTIONAL NEURAL NETWORK

Also known as *ResNet*, a residual CNN consists of a non-linear layer architecture. Instead of one layer L_i receiving the output of the previous layer L_{i-1} and sending its own output to the next layer L_{i+1}, a residual network introduces *skip connections*, which are connections from a layer L_i to another layer L_k, where $k > i + 1$. In other words, the output of a layer is sent to its adjacent layer and also to other layers further down the network (He et al., 2015).

This architecture was proposed to mitigate the problem of vanishing gradients, and its benefits include better generalisation capabilities and overall better accuracy for deep neural networks. It is very common to see this architecture being used to solve SR problems due to its good generalisation.

2.4 SUPER RESOLUTION OF VEHICLE LICENSE PLATES

On Section 2.1 an overall explanation about generic Super Resolution was given. To recapitulate, when the objective is to achieve a HR image with better quality and detail, one of the factors that have to be considered is the fidelity of the resulting image with the ground truth.

To measure fidelity, many methods use measuring algorithms like Mean Absolute Error (MAE) or Structural Similarity (SSIM). Although these techniques do suffice when you want your HR image to look like the ground truth, unfortunately that is not the case for vehicle license plates, due to the presence and importance of its characters.

When you need to apply SR for any kind of text, using MAE or SSIM to train your network will not give you the fidelity needed when the image has a poor quality. If the network is not able to discern which character is displayed, it will "merge" characters resulting in overlappings (Zhang e Cai, 2020). (see Figure 2.3). This happens because doing so gives a lower error than choosing a single character to put in the output plate image. In other words, the MAE and SSIM do not enforce character fidelity and readability enough, and thus the network will not learn to preserve their structure correctly.

Figure 2.3: Example of the output image of a SR-CNN. Notice how the network decided to overlap the characters that appears to be a "P" and an "L". Image taken from the GitHub page of (Zhang e Cai, 2020).

In order to avoid allowing the network to "cheat" by using overlapped characters or using shapes that only resembles a character, extra restrictions need to be applied through the loss function.

One example is using a GAN architecture, where the discriminator network learns to recognize real license plate images. In this case, if the generator network attempts to cheat by using overlapped characters, the discriminator network will not recognize the plate as real.

For single network architectures, the loss function must be able to discern by itself if the license plate contains invalid characters. This is a very challenging problem, as it requires it being capable of recognizing characters reliably. With a GAN, this problem is solved by using a loss function that mixes a structural/perceptual evaluation of the output from the generator network with the evaluation of the discriminator network (Zhang e Cai, 2020).

2.4.1 Usability of License Plate Super Resolution systems

Having a solution that is capable of super resolving license plates with low resolution and quality has a simple, direct usability: it can be used to improved already existing and deployed ALPR solutions.

On a different note, some works attempt to create an *Automatic License Plate Recognition* system that is able to read low resolution images directly. One such example is the work of (Gonçalves et al., 2019), which attempted to apply a multi-task learning method to recognize license plate images with low resolution. While this approach is also viable, it requires the introduction of a brand new solution for this specific scenario.

2.5 CONCLUSION

Super Resolution of License Plates shares an interesting restriction with generic text Super Resolution problem. Not only the fidelity, but the character readability must also be considered in any evaluation of an algorithm. If the image has good detail but the characters are not readable or do not make sense, then the super resolved image becomes useless.

As discussed, common methods used today involve training CNN architectures (single-network or GAN) in an attempt to guarantee character readability in the license plate, since crafting a "traditional" mathematical approach is much more challenging due to the need of the method being able to consider character structure during the reconstruction process.

3 BIBLIOGRAPHY REVIEW

3.1 OVERVIEW ABOUT SUPER RESOLUTION

Most pioneer methods crafted for image super resolution were based on existing methods for denoising and "cleaning" of wave signals. The published works of (Kim et al., 1990; Bose et al., 1993) are one of the earliest to attempt applying signal denoising and cleaning techniques to images. Their method works with a series of copies of a image, where each copy can have its own noise, degradation and global offset. The images are first converted to elements of the wavenumber domain, where frequency domain analysis is applied along with a recursive least square reconstruction algorithm to obtain the high resolution sample.

These mathematical reconstruction methods usually require multiple, slightly different samples of the target image to add details in the HR output. If only a single image is given, there is not much information that can be recovered to recreate lost details, which is the scenario of image rescaling problems.

For cases where you have an image sequence, it is possible to apply local motion estimation and create compensated frames, and these frames can be used to reconstruct a single HR frame recursively (Avrin e Dinstein, 1998).

Nowadays, most methods revolve around the usage of deep artificial neural networks, especially the convolutional neural network architecture. By providing sufficient data to train a proposed architecture, neural networks often give results that are as good or better than existing non-ANN algorithms, and most of the time, with less execution time (for evaluations, not considering training time).

Still, many non-ANN works of the field also provide good results. For example, (Zou et al., 2019) proposed a Sparse Coding Super Resolution method that builds semantic dictionaries to super-resolve license plate characters that were previously segmented.

In the scenario of multi-frame SR, (Seibel et al., 2015) proposed a geometric K-nearest neighbor method to super-resolve a low resolution image sequence. It makes use of motion field estimation to compensate local motion within each frame, aligning all frames with subpixel accuracy in order to define the value of each pixel in the HR grid.

For CNN approaches, (Li e Wang, 2017) proposed a solution for multi-frame super resolution that also makes use of motion compensation, along with a simple *ResNet* architecture. Given a target central frame, the algorithm takes a few frames that comes before and after the target, applies motion compensation based on the central frame and finally feeds these frames as input to the network.

3.2 LICENSE PLATE SUPER RESOLUTION

When targeting vehicle license plates for Super Resolution, as mentioned on Section 2.4, the new restriction of character readability must be considered when evaluating a proposed solution. If the license plate has good quality but the characters cannot be read, then the proposed solution is useless.

Considering character readability, there are a few approaches that are used. One of them is simple character super resolution, where the license plate is generally ignored. In this case, only the characters of the plate are focused as SR targets. (Chuang et al., 2014) proposed an ANN architecture that receives as input a character and outputs its super-resolved version. Given

Figure 3.1: Showcase of the capabilities of the model proposed by (Zhang e Cai, 2020). GT stands for Ground Truth.

a vehicle image, the method first extracts the license plate, segments its characters and finally feeds the segmented characters to the ANN architecture. Using a character SR approach has the advantage of having a smaller image to deal with during training, which can speed up the process significantly.

Another approach is to apply SR to a cropped license plate. In this case, you have the problem of variability of LP layouts, including color and the occasional presence of figures, flags, and other details that may interfere in the super resolution process and cause appearance of artifacts in the output. One such method is the one proposed by (Zhang e Cai, 2020), in which a residual GAN architecture is trained to super-resolve license plate images with an extremely low resolution, as shown in Figure 3.1.

3.3 DATABASE AVAILABILITY

There are many datasets created for training Automatic License Plate Recognition (ALPR) systems. Unfortunately, most of them are not specialized for super resolution, so in many cases you cannot obtain a ground truth license plate image with a decently high resolution and quality to use for training.

You can see detailed information about a few datasets on Table 3.1 and how to access each on Table 3.2. As said above, most of these datasets do not have close up, high resolution images of the license plates. The smallest distance that the photos are taken average to about two meters (see CCPD dataset).

3.4 CONCLUSION

Most recent works use convolutional neural network architectures to super resolve images and videos. For license plates, the preferred architecture is the GAN architecture, due to its discriminator network being able to prevent the generator network from "cheating" by using overlapped characters when generating the output image, as well as enforcing it to keep the license plate structure.

In the context of super resolving an entire license plate (which is often what clients desire), the variability of colors, layouts and details of LPs makes this problem extremely hard to generalise, since there is not many datasets with high resolution footage available. Hence, acquiring a good amount of training data is, by itself, a challenge.

Table 3.1: General information about some ALPR datasets. The ones with "Reported" written above means that the dataset was not downloaded for analysis, and thus the data shown is what is reported in the dataset webpage.

Dataset	# Images	Size	Resolution	Format	Details
OPEN-ALPR	489 vehicles 751 license plates 1240 total	395MB (ZIP)	Very varied	PNG JPG	Benchmark dataset for OpenALPR. Contains BR, EU and US LPs. Most images are human readable. One image per vehicle/license plate. One vehicle/license plate per image. Mostly labeled.
(Reported) SSIG-ALPR (Gonçalves et al., 2018)	6660 total	35GB	1920x1080	PNG	Brazilian LP dataset.
(Reported) SSIG-SEGPLATE (Gonçalves et al., 2016)	-	8.6GB	1920x1080	PNG	Brazilian LP dataset. Only images of already segmented LP characters.
(Reported) UFPR-ALPR (Laroca et al., 2018)	4500 total	9.7GB	1920x1080	PNG	Brazilian LP dataset. 3 different cameras used, 1500 images taken with each one. Labeled.
(Reported) VeRi-776 (Liu et al., 2016; Liu et al., 2016, 2018)	50000+ total	-	-	-	Labeled. 776 different vehicles. Proposed for vehicle re-identification purposes.
Brno University HDR (Špaňhel et al., 2017)	652 total	64MB (ZIP)	Very varied	PNG	Czech cropped LP dataset. Most images are human readable. Images of already cropped LPs. Varied angles. One image per license plate. Not labeled.
Brno University ReId (Špaňhel et al., 2017)	185903 total	1.8GB (ZIP)	Slightly varied around 130x40	PNG	Czech cropped LP dataset. Most images are human readable. Images of already cropped LPs. Varied angles. 6-30 frames per license plate. Not labeled.
Rear View	510 total	32MB (ZIP)	640x480	JPG	Croatian LP dataset. Most images are human readable. Only rear view images. Small angle shifts. Not labeled. One image per vehicle.
IRCP (Kasaei et al., 2009; Kasaei e Kasaei, 2011)	220 total	82MB (ZIP)	640x480 800x600 1024x768	JPG	Iranian LP dataset. Most images are human readable. Frontal and reat view of vehicles. Different perspectives, lighting and distances. Not labeled.
Chinese City Parking Dataset (CCPD) (Xu et al., 2018)	283037 total	12GB (.tar.bz2)	720x1160	JPG	Chinese LP dataset. Most images are human readable. One vehicle per image. One image per vehicle. Varied lighting, weather and tilt. Completely labeled. Labels have additional information about tilt degree, brightness, blur, etc

Table 3.2: Details of how to access the datasets.

Dataset	Link	License/Agreement	How to Obtain
OPEN-ALPR	https://github.com/openalpr/benchmarks	*GNU Affero General Public License v3.0*	Open
SSIG-ALPR (Gonçalves et al., 2018)	http://www.smartsenselab.dcc.ufmg.br/ssig-alpr-database	-	Request Requires signed agreement
SSIG-SEGPLATE (Gonçalves et al., 2016)	http://www.smartsenselab.dcc.ufmg.br/ssig-segplate-database/	-	Request Requires signed agreement
UFPR-ALPR (Laroca et al., 2018)	https://web.inf.ufpr.br/vri/databases/ufpr-alpr/	-	Request Requires signed agreement
VeRi-776 (Liu et al., 2016; Liu et al., 2016, 2018)	https://vehiclereid.github.io/VeRi/	Non-Commercial Use No Unauthorized Share	Request
Brno University HDR (Špaňhel et al., 2017)	https://medusa.fit.vutbr.cz/traffic/research-topics/general-traffic-analysis/holistic-recognition-of-low-quality-license-plates-by-cnn-using-track-annotated-data-iwt4s-avss-2017/	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International	Open
Brno University ReId (Špaňhel et al., 2017)	https://medusa.fit.vutbr.cz/traffic/research-topics/general-traffic-analysis/holistic-recognition-of-low-quality-license-plates-by-cnn-using-track-annotated-data-iwt4s-avss-2017/	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International	Open
Rear View	http://www.zemris.fer.hr/projects/LicensePlates/english/results.shtml	-	Open
IRCP (Kasaei et al., 2009; Kasaei e Kasaei, 2011)	https://github.com/SeyedHamidreza/car_plate_dataset	*GNU General Public License v3.0*	Open
CCPD (Xu et al., 2018)	https://github.com/detectRecog/CCPD	*MIT License*	Open

4 METHODOLOGY

In this section are presented all the details about which CNN architectures will be used and how they are configured and trained for the comparison stage. The networks are programmed using the *Tensorflow* python library.

4.1 NETWORK ARCHITECTURES

For this work, two CNN architectures will be tested: a simple (or plain) CNN and a residual CNN. As explained, a residual network is often more capable of generalising its input data during the training phase, so the expectation is that it should give an overall better result than a plain CNN given different scenarios.

Both architectures were made to be as similar as possible of one another. The plain CNN has 16 convolutional blocks, whilst the residual CNN has 17. A convolutional block is a sequence of a 2D convolutional layer followed by a batch normalization layer and a LeakyReLU activation layer. In addition to these convolutional blocks is a final 2D convolutional layer to produce the output image. A visual representation of both networks can be seen in Figure 4.2.

The residual CNN has 17 convolutional blocks because the first one is used as a feature extractor. These features are stored and added to the output of last layer of the network, just before the upscaling block.

4.1.1 Layer configuration

All of the convolutional blocks outputs feature vectors with 128 channels that are generated by a 3x3 kernel, using a 1 by 1 stride. The LeakyReLU activation layers use an alpha value (negative slope of ReLU) of 0.2. The initial weights are set to the default setting of the Tensorflow library.

4.2 TRAINING DATA

Due to the lack of HR images of brazilian license plates, the networks will be trained with synthetic LPs created with a custom and simple generator. In total, 8 LP templates are generated, as shown in Figure 4.1. All samples are stored in JPEG format. For data augmentation, the samples are preprocessed with random contrast, brightness and JPEG quality. To randomize the downscaled image pixels more to avoid having the networks learning only a few pixel patterns, the HR images are downscaled using, at random, one of the following algorithms provided by the Tensorflow library:

- resize by area;
- bicubic interpolation;
- bilinear interpolation;
- gaussian interpolation
- Lanczos interpolation;
- Mitchell-Netravali cubic interpolation;

- nearest-neighbor interpolation;

Figure 4.1: Random samples generated for each license plate template.

4.3 TRAINING THE NETWORKS

To allow a fair comparison between both networks, along with their architectural similarity, both will be trained using the same dataset for an equal amount of epochs. The training will be done with the Adam optimizer using a 0.004 learning rate, and VGG content loss function (Simonyan e Zisserman, 2015; Zhang e Cai, 2020).

The dataset contains 64000 synthetic license plate images, which are chosen at random. The training phase uses batches of 8 images, and runs for 300 steps per epoch, for 900 epochs. Each LR image has a height of 8 pixels and width of 25 pixels, and their HR counterpart has a height of 64 pixels and width of 200 pixels.

Although the training time is short, since the LR images are very small, the networks can rapidly learn the pixel patterns and fit the training dataset. It is expected that the plain CNN will "suffer" from overfitting more than the residual CNN, which includes the overlapped characters effect. To validate this, a variety of synthetic license plates that are not in the training dataset will be used. For curiosity purposes, a real license plate downscaled image will also be evaluated by the networks in an attempt to see whether or not the synthetic LPs are good enough to train networks for real environments.

4.4 CONCLUSION

Two CNN architectures (plain and residual) will be tested and compared using synthetic license plate images generated by a custom generator. Both architectures are made to be as close as possible and will be trained by the same amount of time in order to allow a fair comparison of the results.

The expected results include the residual CNN showing a better generalisation with less apparent overlapped characters, whereas the plain CNN is expected to show a clearer overfitting problem, with more overlappings for similar characters like 0, 8, 6 and 9.

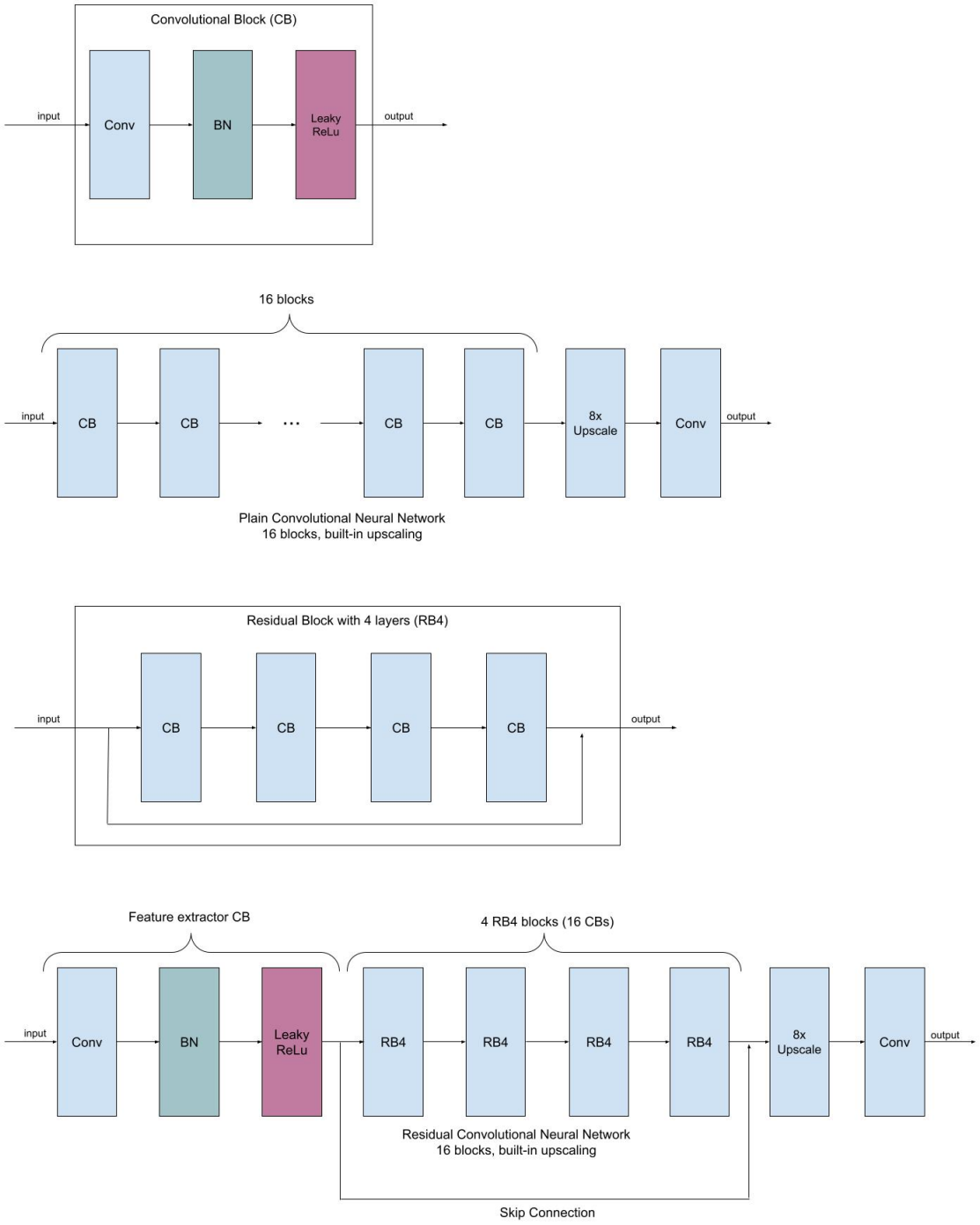


Figure 4.2: Architectures used for the networks. The residual versions have a skip connection for each residual block and an extra connection that sends the first feature vector to the upscaling layer.

5 EXPERIMENTS

5.1 ENVIRONMENT

All experiments explored in the next sections are done using a testing dataset with 8000 synthetic license plates, with none of them appearing in the training dataset. For this chapter, the networks are named *SRConvNet* (plain CNN) and *SRResNet* (residual CNN).

5.2 IMAGE QUALITY COMPARISON

In this section, a simple image quality comparison was done, without considering character readability. For this task, the following loss metrics were used: Mean Absolute Error (MAE), Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) and VGG loss (Simonyan e Zisserman, 2015). The implementation of the VGG content loss is taken from (Zhang e Cai, 2020).

In general, both networks achieved good fidelity for each of the 8 license plate templates. The Table 5.1 shows the results obtained for both networks. The values for each loss is the mean of all 8000 evaluations.

Table 5.1: Image quality evaluation results for both networks. Best values are marked with bold text.

Network	MAE	MSE	PSNR	SSIM	VGG
SRResNet	0.0212	0.0016	**28.5283**	**0.9595**	0.0530
SRConvNet	**0.0204**	**0.0015**	28.4423	0.9583	**0.0508**

As shown in the table, both networks achieved very close averages. The plain CNN had a surprinsingly better performance when it comes to replicating the overall structure of the license plates. The speculation as to why the plain CNN is better at replicating the ground truth image is because it can simply copy the pixel values it learned during training rather than calculating the proper residue that will result in the ground truth plate.

5.3 CHARACTER READABILITY

As the section name suggests, here will be presented a showcase of the overall character readability problems of each network. From what was already told in Section 4.3, the results are expected to include character overlappings, because the VGG loss function used for training does not enforce character readability enough. Also, it was expected that the plain CNN would end up using more character overlappings than the residual CNN. This fact can be observed when comparing the Figures 5.1 and 5.2. Note that the predictions from the SRConvNet present much heavier overlappings.

It can also be observed that the synthetic LPs do not represent real world scenarios with enough fidelity. The last sample shown in the Figures 5.1 and 5.2 show how poorly the training dataset is, when considering that the networks will be used in real world scenarios.

One interesting fact is that since none of the networks saw the real license plate example, they instead change the output to the nearest template that they did see during the training phase.

Figure 5.1: Predictions of the SRConvNet for a few random samples of each license plate template. The last two samples are real LPs, not present in the training dataset.

Figure 5.2: Predictions of the SRResNet for a few random samples of each license plate template. The last two samples are real LPs, not present in the training dataset.

5.4 CONCLUSION

The experiments show that both CNN architectures can replicate structural details of license plates with very high image fidelity. However, when it comes to text fidelity and character readability, since the loss function does not enforce this factor enough, both networks decide to use overlappings when they cannot decide on which character to place. This happens especially for similar numbers like 0, 6, 8 and 9.

The performance comparison over character readability also shows that the residual CNN has much better generalisation capabilities. You can observe it by the fact that the character which the network is most confident to be the correct one is "drawn" with much stronger colors, while the overlappings of other characters remain opaque. The plain CNN does not use stronger colors based on its prediction, resulting in almost unreadable characters (*e.g.* the first character of the MRV-5646 plate).

Also, testing the networks with a simple, frontal and clear picture of a "real world" license plate shows that the synthetic ones are not good enough to substitute the usage of actual, real footage for training them for real world applications.

6 CONCLUSION

In this work, two well known Convolutional Neural Network architectures (plain CNN and residual CNN) were tested for the task of License Plate Super Resolution. Due to the lack of real world footage of brazilian LPs, synthetic license plates were used for training instead.

One residual CNN was previously tested in another work, showing the importance of having a loss function that prevents the network from "cheating" by using overlapped characters. The article in question solved the problem by using a GAN architecture, which was not included in this work.

The results over the synthetic dataset show that using a plain CNN architecture for character reconstruction without proper enforcement does not generate reliable characters, having the presence of a heavy amount of overlappings. For the residual CNN, the overlappings are also present, but are much more subtle. The result is a readable character with the presence of very opaque, different characters overlapping it.

REFERENCES

Avrin, V. e Dinstein, I. (1998). Local motion estimation and resolution enhancement of video sequences. Em *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, volume 1, páginas 539–541 vol.1.

Bose, N. K., Kim, H. C. e Valenzuela, H. M. (1993). Recursive total least squares algorithm for image reconstruction from noisy, undersampled frames. *Multidimensional Systems and Signal Processing*, 4(3):253–268.

Chuang, C., Tsai, L., Deng, M., Hsieh, J. e Fan, K. (2014). Vehicle licence plate recognition using super-resolution technique. páginas 411–416.

Gonçalves, G., Diniz, M., Laroca, R., Menotti, D. e Schwartz, W. (2019). *Multi-task Learning for Low-Resolution License Plate Recognition*, páginas 251–261.

Gonçalves, G. R., da Silva, S. P. G., Menotti, D. e Schwartz, W. R. (2016). Benchmark for license plate character segmentation. *Journal of Electronic Imaging*, 25(5):1–5.

Gonçalves, G. R., Diniz, M. A., Laroca, R., Menotti, D. e Schwartz, W. R. (2018). Real-time automatic license plate recognition through deep multi-task networks. páginas 1–8.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Deep residual learning for image recognition.

Kasaei, H. e Kasaei, M. (2011). Extraction and recognition of the vehicle license plate for passing under outside environment. páginas 234–237.

Kasaei, H., Kasaei, M. e Monadjemi, S. (2009). A novel morphological method for detection and recognition of vehicle license plates. *American Journal of Applied Sciences*, 6(12):2066–2070.

Kim, S. P., Bose, N. K. e Valenzuela, H. M. (1990). Recursive reconstruction of high resolution images from noisy undersampled multiframes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(6):1013–1027.

Laroca, R., Severo, E., Zanlorensi, L. A., Oliveira, L. S., Gonçalves, G. R., Schwartz, W. R. e Menotti, D. (2018). A robust real-time automatic license plate recognition based on the YOLO detector. páginas 1–10.

Li, D. e Wang, Z. (2017). Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 3(4):749–762.

Liu, X., Liu, W., Ma, H. e Fu, H. (2016). Large-scale vehicle re-identification in urban surveillance videos. páginas 1–6.

Liu, X., Liu, W., Mei, T. e Ma, H. (2016). A deep learning-based approach to progressive vehicle re-identification for urban surveillance. 9906:869–884.

Liu, X., Liu, W., Mei, T. e Ma, H. (2018). Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658.

NVIDIA (2020). Nvidia dlss 2.0: A big leap in ai rendering. https://www.nvidia.com/en-us/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/. Acessado em 13/03/2021.

O'Shea, K. e Nash, R. (2015). An introduction to convolutional neural networks. *CoRR*, abs/1511.08458.

Ren, H., El-khamy, M. e Lee, J. (2019). Dn-resnet: Efficient deep residual network for image denoising. Em Jawahar, C., Li, H., Mori, G. e Schindler, K., editores, *Computer Vision – ACCV 2018*, páginas 215–230, Cham. Springer International Publishing.

Seibel, H., Goldenstein, S. e Rocha, A. (2015). Fast and effective geometric k-nearest neighbors multi-frame super-resolution. páginas 103–110.

Simonyan, K. e Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Xu, Z., Yang, W., Meng, A., Lu, N. e Huang, H. (2018). Towards end-to-end license plate detection and recognition: A large dataset and baseline. páginas 255–271.

Zeng, W. e Lu, X. (2012). A generalized damrf image modeling for superresolution of license plates. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):828–837.

Zhang, Z. e Cai, C. (2020). License plate enhancement - from tv shows to reality. https://github.com/zzxvictor/License-super-resolution. Acessado em 12/03/2021.

Zou, Y., Wang, Y., Guan, W. e Wang, W. (2019). Semantic super-resolution for extremely low-resolution vehicle license plate. páginas 3772–3776.

Špaňhel, J., Sochor, J., Juránek, R., Herout, A., Maršík, L. e Zemčík, P. (2017). Holistic recognition of low quality license plates by cnn using track annotated data. páginas 1–6.