Resolution enhancement of traffic surveillance images with SRGANs

Authors: Javier López-Tello Morales, Guillermo Rubio Méndez,
Miguel Zabaleta Sarasa, Miguel Ángel De Moya Jiménez

May 9, 2022

1 Abstract

This project constitutes a theoretical description of a real-world problem in society and a methodology based on deep learning used to solve it. Namely, we will use an architecture based on SRGANs to enhance the resolution of surveillance images in traffic accidents. Additionally, we propose an implementation for this system and discuss its utilities and limitations.

2 Introduction

There are thousands of traffic surveillance cameras used by traffic authorities, being the majority low-quality and low-resolution cameras. This may reduce the ability to quantify the gravity of an incident and as a result, overvalue or undervalue the resources required to reduce or prevent damages. Image super-resolution (SR) refers to the process of recovering high-resolution images from low-resolution images, and may be applied to tackle this issue.

3 State of the art

Image Super-Resolution is a relevant image processing technique meant to increase the resolution of images in the field of computer vision [1]. Several SR methods have been developed with excellent performance employing diverse strategies, like edge-based approaches, statistical methods, patch-based methods, etc. The most relevant ones, considering SR applied to surveillance [2] are also learning-based [3], [4] and self-similarity-based methods [5], [6]. However, this field has experienced enormous advancements recently thanks to the use of Deep Learning methods, achieving state-of-the-art performance on various SR benchmarks.

The employed strategies in Deep Learning methods vary wildly, having different types of network architectures, loss functions, and learning principles and strategies [7], as can be seen in Figure 1, including improvements widely applied to other deep learning methods, like data augmentation and context-wise network fusion. Deep Learning SR methods were first based on Convolutional Neural Networks (CNN), which receive a down-sampled image as input and output a high-sampled image ([8], [9]), but the most recent and powerful are Generative Adversarial Nets (GAN) methods.

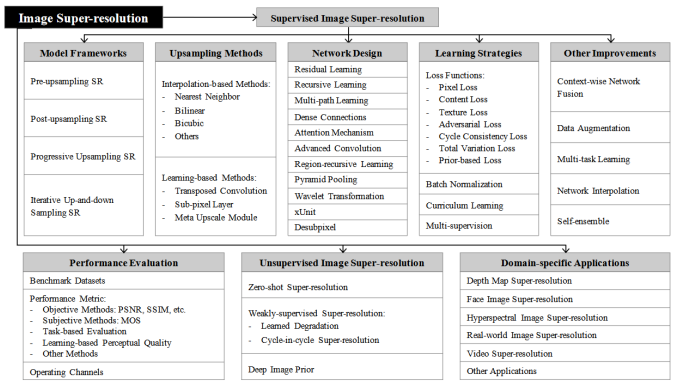

Figure 1: Deep Learning Super-Resolution methods and features [7]

GANs employ a generator and discriminator to be trained. The generator upsamples low-resolution images to super-resolution images, and the discriminator is used to distinguish the high-resolution images and back-propagate the GAN loss to train the generator and the discriminator. One of the most relevant methods is SRGAN [10], its impressive results are shown in Figure 4.

SR methods have applications in many image-related domains, like standard video enhancement, surveillance, medical diagnosis, earth-observation remote sensing, astronomical observation, and biometric information identification [1]. In addition, deep-learning methods have shown to work particularly well with domain-specific problems [7] (like depth map SR and face images SR), which is precisely the case of the problem presented in this project, traffic images from surveillance cameras.

4 Database

Nowadays, existing SR works mostly focus on supervised learning, i.e. use pairs of low resolution - high resolution (LR-HR) images for training the algorithms [7]. Nevertheless, due to the difficulty of collecting images of the same scene with different resolutions, low resolution images are often obtained by performing certain types of degradation methods, like the mainstream Bicubic Interpolation with anti-aliasing [11]. Furthermore, considering the type of images to be analyzed, using networks trained with the same image-type dataset will produce better results [12].

Today there are a wide variety of datasets available for image super-resolution, which vary widely in terms of image quantity, quality, resolution, and diversity. We propose the following datasets: BSD500 [13], Urban100 [14], OutdoorScene [15] and the huge dataset from ImageNet [16], containing a wide range of images from cities, vehicles, animals, urban and outdoor scenery, people, buildings, water, structures, plants, etc.

We could also use the Div2k dataset [17], which provides pairs of LR-HR images; or more specific datasets for our problem, like the one created in this paper [18] or the Traffic-Net dataset, with traffic images from accidents, dense traffic, fire and sparse traffic. Additionally, image datasets could be created selecting specific frames from traffic videos recorded with surveillance cameras.

5 Proposed system

5.1 Model description

The system we propose is based on SRGANs. As was detailed before, SRGANs are formed by a generator network which takes a low-resolution (LR) image and outputs an estimation of its high-resolution (HR) counterpart (named as super-resolution image), and a discriminator network, which is in charge of discerning between a real HR image and a generated SR image. This image shows the architecture of the generator and discriminator:

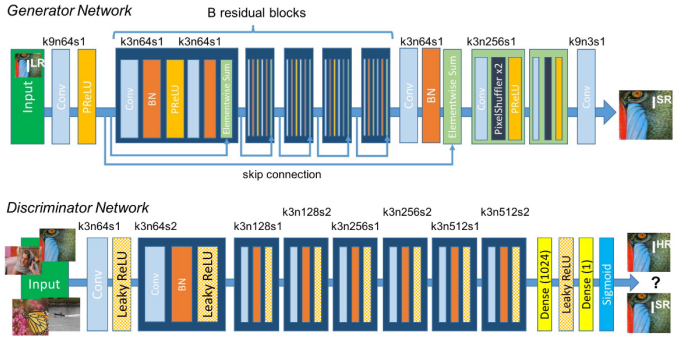

Figure 2: SRGAN architecture [10]

In particular, the generator network is composed of B residual blocks with identical configuration (each block is made out of two convolutional layers with 3x3x64 kernels, a batch-norm layer and a ParametricReLU as activation function), and two sub-pixel convolution layers, which increase the resolution of the input image.

On the other hand, the discriminator contains eight convolutional layers with kernels increasing in size by a factor of two (from 3x3x64 to 3x3x512). It also uses batch-norm after applying each

convolution, and Leaky ReLU as activation function. After these convolutions, two dense layers and a final sigmoid activation function follow, in order to obtain a probability for sample classification.

The usual loss function used on other supervised SR algorithms is the mean squared error (MSE). Optimizing for this function does not provide great results as it tends to find generated images where **overall**, the difference between the estimated image and the ground truth in terms of its pixel values, is small. This results in generated images with too much smoothing and thus lower resolution.

A key novelty from this model is that it incorporates a specific loss function which simultaneously trains the generator to fool the discriminator and to reconstruct SR images. This named perceptual loss function is defined as follows:

$$l^{SR} = \underbrace{\underbrace{l_{\mathrm{X}}^{SR}}_{\text{content loss}} + \underbrace{10^{-3} l_{Gen}^{SR}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}}$$

Figure 3: Perceptual loss [10]

The content loss is calculated measuring the average distance between the feature maps applied on the generated SR image and the HR image. This way, the network is able to focus on particular patches of the image which result in a perceptual similarity (as opposed to the similarity in pixel space provided by minimizing the MSE). The results provided by this methodology and ones which use MSE is illustrated in the following image:

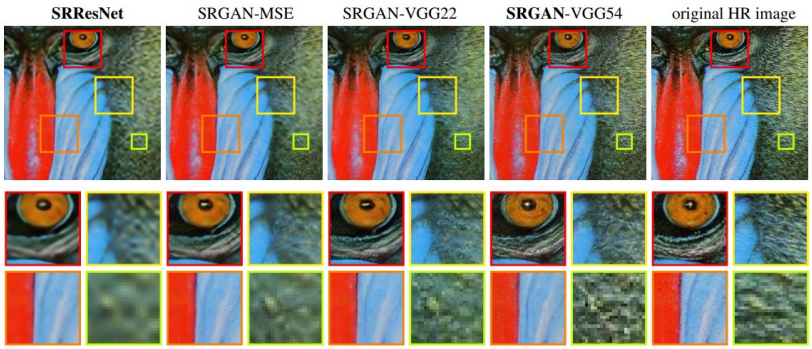

Figure 4: Comparison using MSE vs Perceptual loss [10]

As we can see, even when the MSE is combined with the adversarial loss (second picture), the level of blurriness is higher than in the third and fourth images, which results in less convincing images as far as being more similar to the original image.

5.2 System's implementation

The code that provides for this model can be found in the following link. We would first test the performance of the model with the mentioned databases, trained using Google Colab. After that, we would send our results to an institution which may be interested in them, such as the "Directorate-General for Traffic" (DGT). Once they validated our model, we would begin with the implementation in production.

To do this, we would gather a much larger set of historical videos from such institution, and train the model for a certain time period on the AWS cloud, until we can confirm that the model is ready to put in use. Then, the model would be deployed on the AWS server and connected to the institution's software infrastructure.

This way, once a traffic accident is recorded and the original video is sent to the DGT, the frames of the video will be inputed to the AWS model deployment, and the super-resolution output would be sent back to the institution. In order to achieve a fast response of emergency, it is crucial that the time this whole process takes is very short. Therefore, the AWS service configuration should be of the highest performance possible in terms of speed of network.

As for the system's monitoring, it would be quite useful to have a monthly report of the system provided by the emergency professionals, which would describe in what ways did the high quality video help (or not), and how it could be improved. Additional updates in the implementation and training of the model would be made according to these reports.

6 Discussion: utilities and limitations

6.1 Utilities

There are thousands of cameras placed along public roads that are used to identify traffic accidents. Most of these cameras are of low quality and the images they produce are of low resolution.

This low quality means that certain details in the images cannot be properly appreciated, so that some road traffic accidents cannot be identified. traffic accidents cannot be correctly identified.

To solve this we can make use of SRGAN which allows to improve the quality of images captured by roadside surveillance cameras to detect traffic accidents. This allows to increase the accuracy with which traffic accidents are identified.

6.2 Limitations

There are certain limitations that we have to take into account when using this system. These are some of them:

- Not all videos can be scaled up. If the images in the video do not have a minimum quality that allows certain characteristics of the image to be appreciated, it is not possible to scale the quality of the image. It also happens that unseen data may be mapped out of the subspace, leading to poor results
- We use supervised learning in our system, this method does not correct noise and causes that images with a lot of noise may not be scaled correctly.
- The GANs are systems that are very difficult to train, it causes that they are not scalable. We can´t use it in a massive environment in which we improve the quality of millions of images at the same time in a distributed system.
- It may happen that the systems enter into a mode collapse in which the generator could collapse to a single point producing that our output will be only one sample or a small set of samples over and over again. In this case our system will scale always the same set of samples not taking into account the overall picture.

References

[1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016. DOI: https://doi.org/10.1016/j.sigpro.2016.05.002.

[2] M. A. Farooq, A. A. Khan, A. Ahmad, and R. H. Raza, "Effectiveness of state-of-the-art super resolution algorithms in surveillance environment," in *Digital Interaction and Machine Intelligence*, Springer International Publishing, 2021, pp. 79–88. DOI: 10.1007/978-3-030-74728-2_8.

[3] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[4] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[5] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.

[6] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th international conference on computer vision*, IEEE, 2009, pp. 349–356.

[7] Z. Wang, J. Chen, and S. C. H. Hoi, *Deep learning for image super-resolution: A survey*, 2019. DOI: 10.48550/ARXIV.1902.06068.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015.

[9] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[10] C. Ledig *et al.*, *Photo-realistic single image super-resolution using a generative adversarial network*, 2016. DOI: 10.48550/ARXIV.1609.04802.

[11] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981. DOI: 10.1109/TASSP.1981.1163711.

[12] N. Takano and G. Alaghband, *Srgan: Training dataset matters*, 2019. DOI: 10.48550/ARXIV.1903.09922. [Online]. Available: https://arxiv.org/abs/1903.09922.

[13] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011. DOI: 10.1109/TPAMI.2010.161. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2010.161.

[14] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206. DOI: 10.1109/CVPR.2015.7299156.

[15] X. Wang, K. Yu, C. Dong, and C. C. Loy, *Recovering realistic texture in image super-resolution by deep spatial feature transform*, 2018. DOI: 10.48550/ARXIV.1804.02815. [Online]. Available: https://arxiv.org/abs/1804.02815.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

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

[18] S. Ghisler, J. Sánchez-Soriano, and S. B. Rosende, *Traffic images captured from UAVs for use in training Machine Vision Algorithms for traffic management*, version 1, Please cite: Bemposta Rosende, S.; Ghisler, S.; Fernández-Andrés, J.; Sánchez-Soriano, J. Dataset: Traffic Images Captured from UAVs for Use in Training Machine Vision Algorithms for Traffic Management. Data 2022, 7, 53. https://doi.org/10.3390/data7050053, Zenodo, Dec. 2021. DOI: 10.5281/zenodo.5776219. [Online]. Available: https://doi.org/10.5281/zenodo.5776219.