Benefits of Implementing Super Resolution to Video and Photo Evidences in Court Cases

Sai Pranav Theerthala *

March 22, 2023

Abstract

The world today has advanced in the areas of security by employing the use of security cameras. Security cameras enable us to monitor any given space and access previous recordings. However, most security cameras in large and populated areas can't identify individuals within the camera's scope correctly. This issue becomes especially prevalent in criminal court cases, where this lack of clarity in security cameras can potentially let criminals avoid punishment for their crimes. This paper analyzes the use of super-resolution models to enhance the clarity of standard security camera recordings and how this modified footage can be used as evidence in court cases.

1 Introduction

Although the concept of surveillance cameras has been prominent since the 1950s worldwide, security footage to identify and prosecute criminals is a relatively recent development. Until the turn of the century, surveillance cameras were only used to indicate a crime transpiring, allowing officers to apprehend the criminal in real-time. After the second attack on the World Trade Center, the public pushed for more preventive measures to ensure safety. This push for enhanced safety measures resulted in the creation of more powerful cameras that can identify facial features and distinct clothing patterns. The more powerful footage of CCTV cameras has enabled lawyers to prove clients' undeniable innocence or guilt, as the Department of Justice estimates that video evidence is involved in about 80 percent of crimes. The need for clear and indicative video evidence is more prevalent with such a high dependence upon security camera footage.

There are numerous models of cameras used in small and large areas. The two most commonly used CCTV cameras are bullet CCTV cameras and c-mount CCTV cameras. Bullet cameras are set in place and, depending upon the lens, can reach maximum visibility of 100 feet.

*Advised by: Dr. Parsa Akbari of Cambridge University

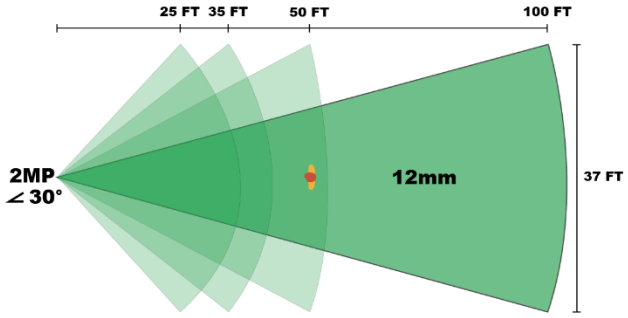

Figure 1: Bullet Camera visibility at 98 Degrees(Bowman 2012)[1].
Bullet camera peripheral visibility is highest when using 2.8mm camera, but has the shortest visibility distance

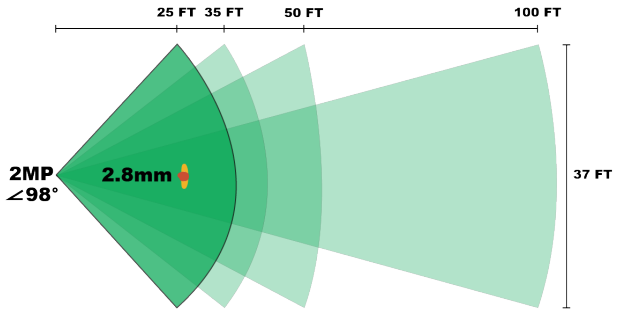

Figure 2: Bullet Camera visibility at 30 Degrees(Bowman 2012)[1]. Bullet camera peripheral visibility is lowest when using 12mm camera, but has the highest visibility distance

On the other hand, c-mount cameras can move 180 degrees in any direction but only have a visibility distance of 40 feet. These cameras can move in all directions but reach a maximum capture distance much smaller than bullet cameras.

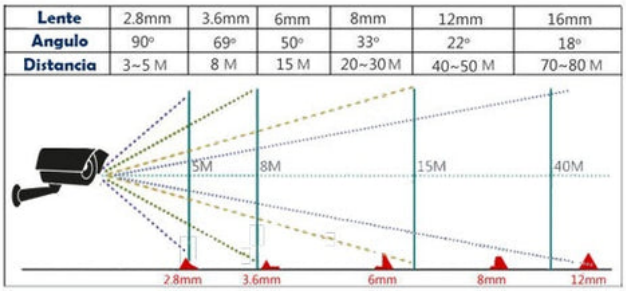

Figure 3: Camera visibility over distance(Chiu 2016)[2]. Camera visibility declines over distance regardless of camera mobility

These cameras have the potential to identify criminals within their range of vision accurately but can lose credibility due to a range of factors. One example is the footage of the Infamous Miami Cannibal Attack. This video was shared in 2019 when a remote camera for the Miami Herald building captured Ruby Eugene attacking Ronald Poppo. The police were alerted by a passing cyclist, who eventually arrived at the place and shot Ruby Eugene. However, upon further inspection of the surveillance footage, we can observe some commotion but can't identify either individual.

Figure 4: Camera Captures Cannibal Attack(Quigley 2012)[3] The distant camera catches movement but not any determining features

In addition to distance, another problem with powerful security cameras is that they can't produce quality footage when there is a lack of light. Given that security cameras can't create their own light source, they can only record black-and-white footage in dark settings.

Figure 5: Camera Visibility in Dark(KTVU 2022)[4] The lack of consistent light sources makes this camera footage is almost useless in identifying thieves

This image is security footage of an ice cream shop that captures robbers smashing the door to an ice cream shop and stealing from the store. Unfortunately, due to the lack of consistency in the visual lighting, the camera can't accurately capture any valuable identification details, such as the burglar's facial features or the car's plate number. These issues underscore the need for an advanced security system that enables cameras to magnify the details of the footage they capture, as these pieces of evidence can not prove the guilt of a criminal in court.

2 Methods

2.1 Model Idea:

In order to address this issue, we hypothesized a few possible solutions. At first, we planned to create a super-resolution model that could be applied to live video footage. This model would enhance the quality of real-time security footage and enable cops to analyze the monitored space more clearly. Upon further research, we realized that this model would require a high level of computing power to train, which would not be possible with the current system. As a result, we scaled the idea down to enhance the clarity of singular images.

2.2 Model Selection:

For this research topic and further applications, the super-resolution network needed to use as little storage as possible. After scouring the internet for appropriate models, we found the Enhanced Deep Residual Networks for Single Image Super-Resolution paper[5]. This paper proposed a neural network that omitted numerous intermediate layers to increase the speed of the overall super-resolution model.

2.3 Model Description:

This section describes the internal structure of the EDSR model used in this paper. We provide details about the intermediate layers removed from traditional super-resolution models, along with an analysis of the model's performance against other high-performing super-resolution models.

2.3.1 Residual Blocks:

Recently, residual networks [6, 7, 8] exhibited excellent performance in computer vision problems from low-level to high-level tasks. Although Ledig et al. [7] successfully applied the ResNet architecture to the super-resolution problem with SRResNet, the EDSR model improved the performance by employing a better ResNet structure.

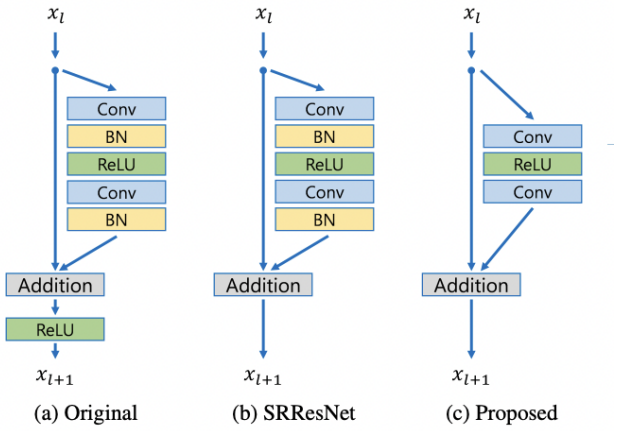

Figure 6: EDSR Model Analysis Comparison of residual block in original ResNet, SRResNet, and EDSR Paper.

As shown in Figure 6, the EDSR paper removes the batch normalization layers from our network as Nah et al.[9] presented in their image deblurring work. Since batch normalization layers eliminate range flexibility from networks by normalizing the features, it is better to remove them. Furthermore, GPU memory usage is also sufficiently reduced since the batch normalization layers consume the same amount of memory as the preceding convolutional layers. Our baseline model without a batch normalization layer saves approximately 40% of memory usage during training compared to SRResNet. Consequently, we can build a larger model with better performance than the conventional ResNet structure under limited computational resources.

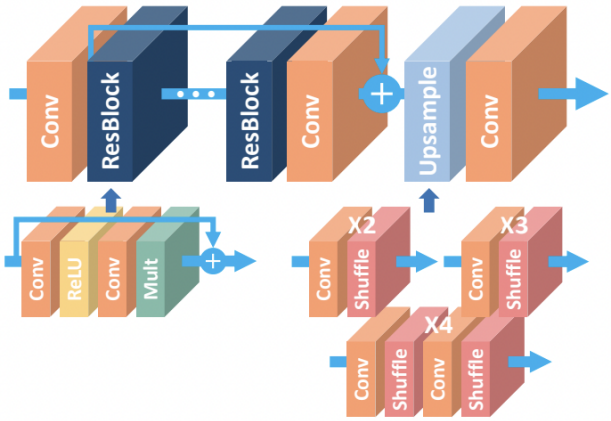

Figure 7: EDSR Model Representation The architecture of the proposed single-scale SR network (EDSR)

2.3.2 Single-Scale Model:

The EDSR model constructs the baseline (single-scale) model with the proposed residual blocks in Figure 6. The structure is similar to SRResNet, but the EDSR model does not have ReLU activation layers outside the residual blocks. Also, this baseline model does not have residual scaling layers because the EDSR model uses only 64 feature maps for each convolution layer. In our final single-scale model, the EDSR model expands the baseline model by setting B = 32 and F = 256 with a scaling factor of 0.1. The model architecture is displayed in Figure 7. When training the model for upsampling factors $\times 3$ and $\times 4$, the EDSR uses model parameters initialized with a pre-trained $\times 2$ network. This pre-training strategy accelerates the training and improves the final performance, as clearly demonstrated in Figure 8. For upscaling $\times 4$, if a pre-trained scale $\times 2$ model (blue line) is used, the training converges much faster than the one started from random initialization (green line).

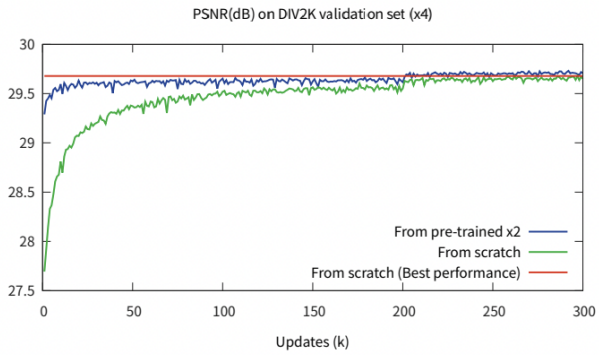

Figure 8: Model Pretraining Analysis Effect of using pre-trained $\times 2$ network for $\times 4$ model (EDSR). The red line indicates the best performance of the green line. Ten images are used for validation during training.

3 Results

3.1 Data Preparation:

Once we found the appropriate model, we needed to find a large data set of low-quality security footage photos. Unfortunately, upon researching numerous outlets such as Kaggle, Data.world, and GitHub, there was no valid data set that enabled us to run the model as designed. To work around this problem, we selected "A data set for automatic violence detection in videos," which compiled numerous videos of crimes. We obtained one frame from each video at the same timestamp for this data set to create our data set. This data set was a collection of videos of numerous violent scenarios played out by a group of people from different angles in a room.

3.2 Test Details:

We ran a standard 80% of the data set through the model for training. All batches ran with four unique images for two epochs, as we found that higher epochs did not increase the accuracy and created the risk of overfitting. Every 20 batches, we would save the captured image for analyzing and improving the model. Every model ran on a kernel size of 3, an image size of 300, and a scale of 2. Every model ran through 140 steps, producing seven different information frames throughout the testing. These values remained static across all tests.

To enhance the model's effectiveness, we tested some variables in the original code proposed in the EDSR paper. The first variable we changed was the layers of the convolution network; we began with 210 layers and kept increasing the layers by 15 until we reached 300. Each variation took anywhere between 9 hours to 15 hours to complete training, with the number of layers increasing.

Throughout these 7 test cases, we hoped to observe a downward trend in the loss values as the layers increased.

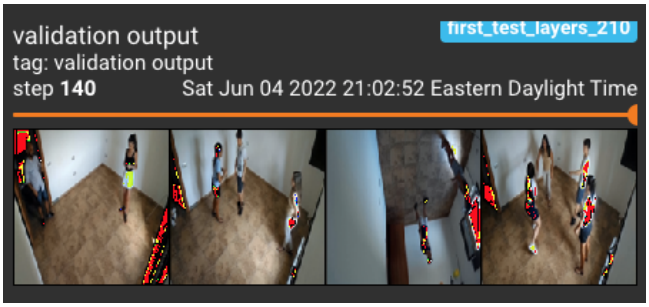

Figure 9: Validation Output for 210 layers in CNN

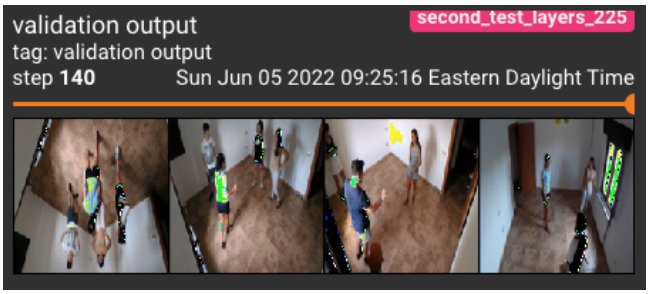

Figure 10: Validation Output for 225 layers in CNN

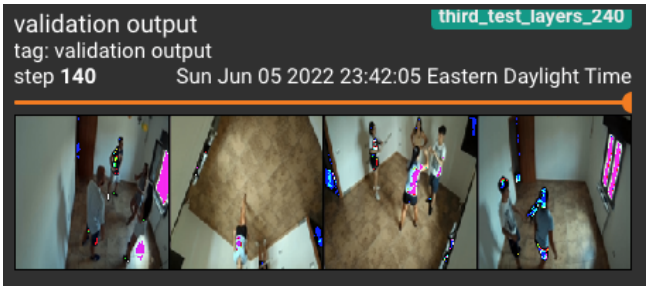

Figure 11: Validation Output for 240 layers in CNN

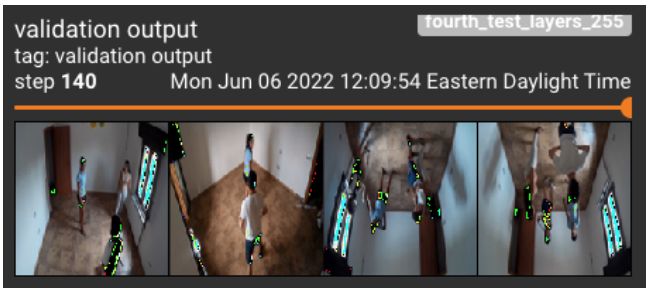

Figure 12: Validation Output for 255 layers in CNN

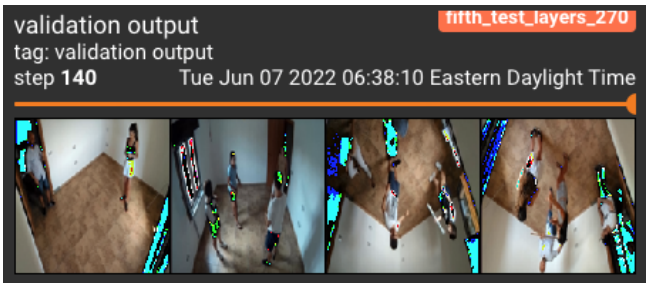

Figure 13: Validation Output for 270 layers in CNN

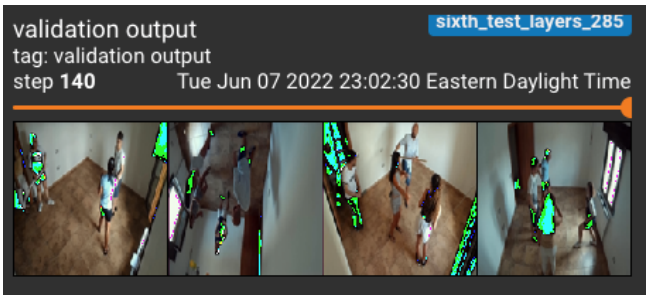

Figure 14: Validation Output for 285 layers in CNN

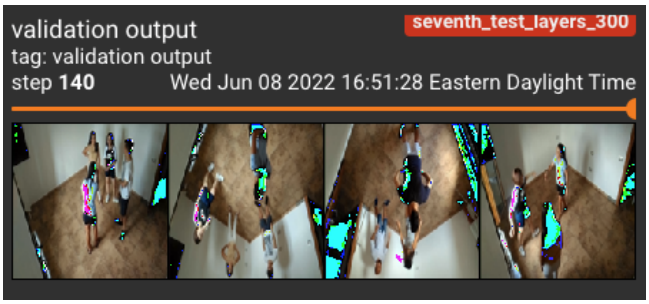

Figure 15: Validation Output for 300 layers in CNN

We also wanted to test the kernel size and image size of the super-resolution model. However, when we increased or decreased these values, the model ran into irreparable problems that would result in changes to the whole CNN.

3.3 Model Accuracy:

We plotted the accuracy of each model at every step to analyze the connection between the CNN layers and the super-resolution model's accuracy. Figure 16 demonstrates a common trend of lower validation loss as the number of steps.

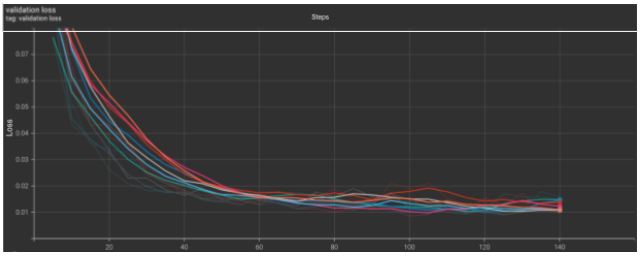

Figure 16: Validation Output Analysis of Validation loss of multiple layers of CNN

Upon further inspection of the results, however, we realized that changing the number of layers in the convolution network did not impact the model's accuracy. Figure 16 represents the validation loss at step 140 for each layer. Figure 16 does not present an evident trend as the number of layers increases, indicating that changing the layers in the CNN did not impact the model results.

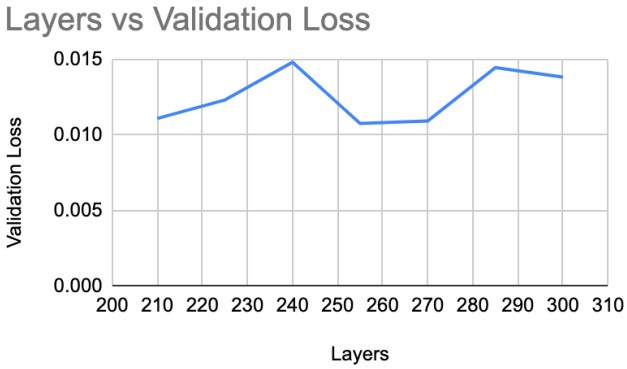

Figure 17: Validation Output for 300 layers in CNN Analysis of Final Validation loss for each CNN

When we analyzed the images, the super-resolution model could identify the individuals in each frame and any assault item they carried. In addition, the resulting images' specific details were more apparent when maximized, and the people's features became much more visible. Being able to recognize the critical aspects of a picture and increase its quality enables the creation of detailed images that make a difference in criminal court cases.

4 Discussion

The research presented in this paper suggests new insight into the possibility of using AI to aid in court cases. The model could identify critical details in the images, including the people involved and any weapons they were holding. More powerful super-resolution models in the future can build on these results to create higher-definition images of low-quality security footage.

There are two main requirements to uphold any statement, referred to as burdens of proof(Wex 2022)[10]. First, in civil court cases, any statement can be proved with a preponderance of the evidence, which means there is a greater than 50% chance that the claim is valid(Wex 2022)[11]. On the other hand, a statement in criminal court cases can only be proven true if established beyond a reasonable doubt. This distinction in the burden of proof limits the effectiveness of the research presented in this paper and with AI-modified footage in general. Security camera images or video data modified by super-resolution models have the potential to present sufficient evidence in civil court cases. Still, they fall short of establishing a claim beyond doubt because an altered image is not enough to prove guilt. This limitation is partly due to the public mistrust of artificial intelligence and how models can be changed to achieve a particular result if they are not appropriately trained. For AI to make a difference in everyday court cases, super-resolution models must build their reputation as systems that operate without bias.

Despite these current setbacks, there is optimism for the future. AI is improving because companies worldwide recognize the undeniable need for it. If super-resolution keeps advancing, we will reach the day when every criminal on any type of camera will be caught and prosecuted for their crimes.

References

[BL17] et al. Bee Lim, Sanghyun Son. Computer vision foundation. *https://arxiv.org/abs/1707.02921*, 2017.

[Bow22] Justin Bowman. Arcdyn security camera field of view (fov) tool - tools - resources. *arcdyn*, 2022.

[Chi16] Camille Chiu. What's the difference between fixed lens and varifocal lens cameras? *101audiovideoinc*, 2016.

[CL17] et al. Christian Ledig, Lucas Theis. Photo-realistic single image super-resolution using a generative adversarial network. *Computer Vision Foundation*, 2017.

[JK16] Kyoung Mu Lee Jiwon Kim, Jung Know Lee. Accurate image super-resolution using very deep convolutional networks. *Institute of Electrical and Electronics Engineers*, 2016.

[KH15] et al. Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. *Computer Vision Foundation*, 2015.

[Qui12] Rachel Quigley. Revealed: New footage shows how eugene came upon his victim ronald poppo by chance as he was strolling along naked, then stripped him and ate his face off. *DailyMail*, 2012.

[SN18] Kyoung Mu Lee Seungjun Nah, Tae Hyun Kim. Deep multi-scale convolutional neural network for dynamic scene deblurring. *Computer Vision Foundation*, 2018.

[Sta22] KTVU Staff. Oakland ice cream shop destroyed after robbers smash car through store. *KTVU*, 2022.

[Tea22a] Wex Team. Burden of proof. *Cornell Legal Information Institute*, 2022.

[Tea22b] Wex Team. Preponderance of the evidence. *Cornell Legal Information Institute*, 2022.