CSCI 1430 Final Project Report: Explict MAP Optimization for Photo-Realistic Super Resolution

Anonymous: Nick Huang, Shijie Mao, Rohit Mohanty, Dustin Wu.
Brown University

Abstract

We present a novel approach to super-resolution using generative methods: Super Resolution MAP (SRMAP). Generative modelling grants our network the ability to generate multiple super-resolution outputs per input, unlike previous approaches that use a deterministic pixel-wise loss. The key insight comes in three parts: 1) explicitly modelling image restoration under the MAP (maximum a posteriori) framework; 2) using adversarial loss as the regularization term of the MAP framework to represent the natural image prior; and 3) using noise injection to model the stochastic behavior of the ill-posed nature of super-resolution reconstruction.

1. Introduction

Super-resolution imaging is the task of increasing the resolution, or concentration of detail, within an image. This problem is increasingly relevant in the modern digital age: the rapid growth of media sharing means that services like Youtube must rely on compression algorithms that make storing content affordable, but result in poorer image quality.

While ongoing efforts are continuously advancing the super-resolution state-of-the-art, the problem is inherently difficult to solve because it is under-specified: a given grainy image might map onto one of any number of possible outputs, and given the large domain of natural images, there isn't a set procedure for super-resolution that can apply generally.

To meet this challenge, researchers have turned to deep learning. However, deep learning super-resolution outputs often appear excessively smooth because the network fails to fully infer the missing information. The core issue is that super-resolution networks are usually trained with a deterministic pixel-wise loss: the network is trained to produce a high-resolution output image given a low-resolution input image, and is penalized directly on the difference between its output and the ground-truth. We believe that this objective is at odds with allowing multiple, higher-quality solutions, and by constraining the model it is forced to return an average.

Our answer to this dilemma is to use generative modelling. The model's objective is no longer to minimize the difference between its output and the ground-truth; rather, it aims to generate images that 1) resemble the low-resolution input when downsampled and 2) appear to belong in a set of natural images as determined by a discriminator network. This means that our model generates outputs stochastically rather than deterministically, allowing for the sampling of multiple possible output super-resolution image from a single input.

2. Related Work

We drew initial inspiration from Lim et al. [5], who present the enhanced deep residual network (EDSR) an end-to-end super-resolution network that uses residual blocks, a series of weight layers with skip connections that add the input directly to the output from a number of layers down. This has the effect of ensuring that information is retained, thereby helping to mitigate the vanishing gradient problem. Our generative models adapt the usage of residual blocks, but build upon the EDSR architecture in ways that will be discussed below. However, as previously mentioned, EDSR uses a constrained deterministic objective (Eq. 1):

$$\mathrm{L1}(F(y), x), \tag{1}$$

where y is the input low-resolution image, x is the ground-truth high-resolution image, $F(y)$ is the model's outputted super-resolution image, and the L1 loss quantifies the difference between the two images. We say that this objective is deterministic because it motivates the model to replicate the ground-truth, resulting in identical behavior given identical input.

Using generative models to produce alternative training objectives is not a new idea. An example of one such effort is Ledig et al. [4], who present a generative adversarial network (GAN) for super-resolution (SRGAN) that also uses residual blocks. One half of the GAN is the generator, which generates super-resolution images given low-resolution counterparts and has essentially identical functionality to an end-to-end network like EDSR. The other half is the discriminator, which is tasked with determining whether a given super-resolution image produced by the generator, or the corresponding high-resolution image from the training dataset,

appears more natural. If it predicts that the generated image is more natural, then it has been fooled, and its loss increases while the generator's decreases. The authors combine this adversarial loss with a deterministic objective, resulting in (Eq. 2):

$$\text{L1}(F(y), x) + \alpha R(F(y)), \tag{2}$$

where $R(F(y))$ is the adversarial loss on the generated super-resolution image, and α is a constant determining the relative contribution of adversarial loss. Note that this loss term still contains the constrained L1 term.

In order to address super-resolution with stochastic generation, we draw inspiration from Karras et al. [2]'s A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN). The authors present a generator architecture that they not only show improves generated image quality, but also allows for finer control over the generation itself. As an example of the effectiveness of this approach, the authors demonstrate that when the latent vectors of two human face images are combined and fed to the network, the result is a face that hybridizes the two inputs to a startling degree: the output face melds the skin tone and hair color of one face with the face shape and hairstyle of the other face. As described below, we took inspiration from the notion of noise injection to introduce stochasticity into our network's outputs.

3. Method

The first component of our project was devising an architecture for the generator network that is responsible for producing super-resolution images. Our generator architecture is inspired by RCAN [6] (another implementation of generative super-resolution) and the generator of StyleGAN2 [3] (an improved version of StyleGAN). The architecture is pictured in figure 1, and is described in detail in the appendix.

Model parameters are outlined in table 1 below, and are elaborated more in the appendix.

We trained this network with L1 loss (Eq. 3) in order to establish that the baseline performance of our network mirrored that of other networks trained with deterministic objectives. The dataset we used is DIV2K, a 1000-image dataset of high-resolution images. The training pipeline was adapted from the code base of EDSR. On each epoch, the pipeline loads a batch of 64 images, training the network with a randomly cropped, horizontally-flipped, and 90-degree-rotated 48x48 patch of each low resolution image and the corresponding high-resolution image.

Parameter	Generator
Stem Width	256
Feature Widths	[512, 256, 128]
Blocks Per Stage	16 x 4
Compression Factor	4
Receptive Field	3

Table 1. Hyperparameters for the network architectures.

$$L1 = \|x - G(y)\|_1 \tag{3}$$

Next, in order to augment our generator with the ability to generate outputs stochastically, we updated the generator to include noise injection. This involves adding noise injection layers in between the convolutional layers and the activation layers of the network. Each noise injection layer simply elementwise-adds a tensor of noise sampled from a standard normal distribution, scaled by a tensor of learned weights. This allows us to explore the super resolution space.

Our discriminator is similar to the StyleGAN2 discriminator, also consisting of convolutional layers and residual connections. One new feature is anti-aliasing, an operation normally used for reducing sharp, jagged edges in an image. This grants it the ability to more finely pick out flaws in our generator's output images. The architecture is pictured in the figure 1, hyperparameters in table 1 and is described in detail in the appendix.

On each epoch we divide the batch of 64 images into two batchs of 32 images, and trained the generator with half the data, and the discriminator was trained with both the generated fake samples and the other 32 real samples.

To train the GAN, we replaced L1 loss with a conjunction of two loss terms according to the MAP (maximum a posteriori) framework, which comes from Bayesian statistics and is used to estimate an unknown quantity given a prior distribution.

The first loss term is the fidelity loss 4:

$$L_{\text{Fidelity}} = \|y - G(y)_{\downarrow_{\text{bicubic}}}\|_1 \tag{4}$$

The second term is the adversarial loss 5:

$$L^G_{\text{Adversarial}} = \sup_{G:\mathcal{Y}\to\mathcal{X}} \mathbb{E}_{\substack{x\sim\text{HR}\\ y\sim\text{LR}}}[f(C(G(y)) - C(x))] \tag{5}$$

Our objective function 6 is derived by summing these terms and weighing the fidelity loss.

$$L_G = \alpha L_{\text{Fidelity}} + L^G_{\text{Adversarial}} \tag{6}$$

The discriminator must be trained separate; we do so with the critic loss 7:

$$L_C = \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\substack{x\sim\text{HR}\\ y\sim\text{LR}}}[f(C(x) - C(G(y)))] + \frac{\gamma}{2}\,\mathbb{E}_{x\sim\text{HR}}\left[\|\nabla_x C(x)\|^2\right] \tag{7}$$

Explanations for each of these functions are provided in the appendix.

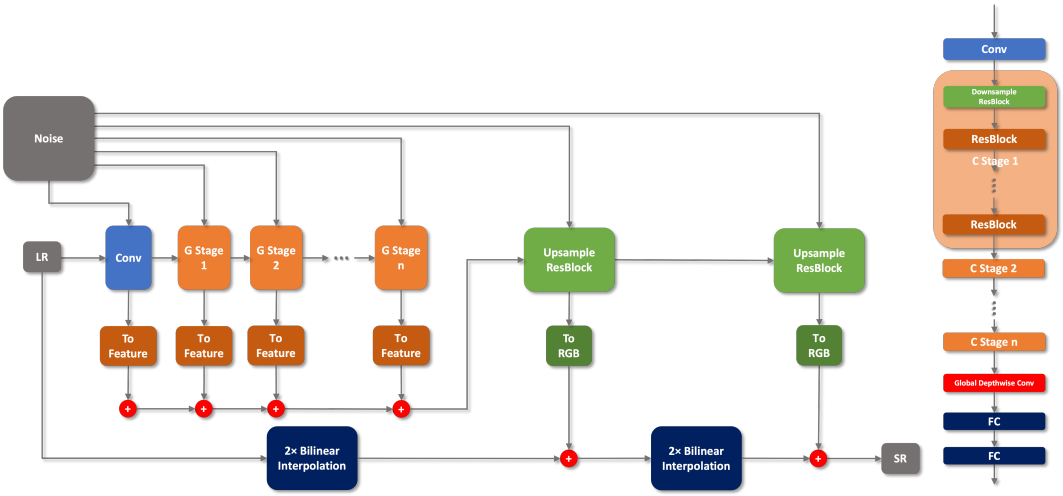

Figure 1. Architecture. *Left:* Generator. *Right:* Discriminator.

We found that training our GAN from scratch did not result in any signs of convergence. We found that pre-training our network on just the L1 loss before transitioning to our objective loss addressed this. After 10 hours of pre-training, Network training took place on a cloud-accessed Tesla V100 over 200 epochs and 5 days of continuous training. For 1 out of 10 batches, we save the model's output as an image so that we can examine the quality of the super-resolution. This was crucial, since given the fact that the generator and discriminator are actively competing against one another, we cannot use their losses as the sole indicator of convergence progression; the quality of output images was our primary means of doing so.

4. Results

Figure 2. Super-resolution Demonstration. *First:* Input low-resolution image. *Second:* Ground-truth high-resolution image. *Third:* Output from L1 loss network. *Fourth:* Output from final network.

Our SRMAP network is capable of generating 4x super-resolution images that look have a higher level of detail than an L1 loss model, depicted in Figure 2.

We found that, as with other super-resolution methods, performance was not perfect. In particular, while the network was capable of recovering unstructured features such as hair or fur, such as the penguin in Figures 1 and 2, its flaws become apparent when recovering structured features, particularly the windows and angled surfaces of buildings. Still, the quality of the results stands as quite impressive

Network	LPIPS	LR PSNR
SRMAP (Ours)	0.153	58.40
Generator w/ L1 (Ours)	0.267	52.80
RRDB	0.253	49.20
ESRGAN	0.124	39.03
NCSR	0.119	50.75

Table 2. Evaluation scores for our and other networks.

compared to other implementations in academia, as seen in table 2. PSNR is a measure of fidelity (higher = better), and LPIPs is a measure of the subjective quality of an image, as determined by a deep network (lower = better).

Compared to the L1-only network, which is similar to deterministic networks like EDSR, our final network was able to recover more fine details, such as the texture of the penguin's beak in Figure 2. This is a demonstration of our earlier claim that deterministic training objectives constrain the network, thereby causing its outputs to be overly smooth. With our relaxed objective, the network is granted enough flexibility to generate finer details. However, this also means that our network also has more flexibility to make mistakes, as observed in the artifacts in the penguin's hair.

Figure 3. Stochasticity. *First:* High-resolution image. *Second:* Corresponding stochasticity regions marked with white pixels.

There was one component of our project that did not meet our expectations: the network's ability to explore the super-resolution space. We found that even with different random seeds, the model's looked perceptually similar. Figure 4 shows the regions where the most difference was observed across ten generated images, brighter regions mean more difference. This visualization indicates that while our model successfully generates different images in the super-resolution space, it does not explore this space very widely. We believe that this is because the model was not trained sufficiently long enough to learn that weighing the noise more heavily could be advantageous for the quality of its output.

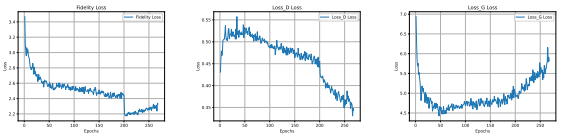

Figure 4. Loss plots. *Left:* Fidelity. *Middle:* Generator. *Right:* Discriminator.

Limited training time is one limiting factor of our network performance. As shown in Figure 4, discriminator loss was still in the process of decreasing rapidly at the time that we stopped training. This means that the discriminator was still in the process of improving its ability to distinguish between real and fake images. In response, the generator's loss was increasing. GANs in academia are often trained upwards of a month. The downward spike in loss was caused by halving the learning rate at 200 epochs, and we suspect that a similar spike would occur had we run 400 epochs.

4.1. Technical Discussion

One tradeoff inherent to our and other super-resolution methods is that, in practice, generative models do not have infinite capacity. Our GAN is many times smaller than the training set, which means that it cannot perfectly represent that natural image prior. Therefore, when asked to recover fine details, it is bound to inherit artifacts caused imperfect memorization. The difference is how such imperfection emerges in the output images. In deterministic networks, the imperfection emerges in the form of excessive smoothness due to a lack of detail. In our approach, the imperfection emerges as noise in the finer details of reconstruction. The up-side is that this noise is sometimes correct because it coincides with the fine detail of the ground-truth. However, at other times the noise does not match up with the ground-truth, which arguably depreciates the quality of the image compared to that of a deterministic network.

4.2. Societal Discussion

An overview of some of the ethical concerns, and our responses, are provided below.

1. "Malicious users who gain access to surveillance camera recordings could apply the high-resolution technique on the video to see the faces of the people in the video clearly and even identify them, which is a significant intrusion of privacy." **Response:** While the notion of being able to reveal hidden details in low-resolution images is understandably worrisome, super-resolution imaging is highly unlikely to be of any use in the recovery of faces, because a low-resolution face image could map onto a wide range of similar-looking yet distinct faces. This is means that it is virtually impossible to recover the exact face from a low-resolution image..

2. "The datasets (DIV 2K and DIV 8K) only contains around 3000 images in total. How could you make sure such a small dataset would produce unbiased result for objects of different categories?" **Response:** Model bias is certainly a valid concern for any deep learning application, but because we want the model to generate images that look "natural" to us, a certain amount of bias is important for our model to do well. Furthermore, while the datasets may at first appear to contain a relatively small number of images, we train the network on randomly-sampled small 48x48 patches of the images, which are typically 2000x1000 pixels, so our dataset is actually sufficient for the task.

3. "What if someone uses the software to reconstruct thumbnail images they find online in order to avoiding paying for copyright? Should there be any restriction to the input image source to make sure the images are legal?" **Response:** Copyright issues are indeed quite prevalent in the digital age, and regularly threaten the livelihoods of artists and designers, but it is for this reason that copyright is already taken seriously and accounted for via policies like the Digital Millennium Copyright Act (DMCA) that are heavily enforced on media sharing sites like Youtube. In addition, one solution is to bundle any distribution of our network with a content identification algorithm, which would flag any images that are recognized as being under copyright; such a system is implemented on Youtube.

4. "The high-resolution technique can help government authorities track down refugees fleeing the country if the authorities apply the high-resolution technique to the face and iris from a border crossing surveillance video. Similarly, government authorities can use this technique to track down illegal immigrants. If the model produces incorrect results, it could lead to unlawful arrests" **Response:** As discussed above, the recovery of exact facial features is virtually impossible to achieve with our and current state-of-the-art super-resolution imaging techniques, but the fact that governments might

be able to utilize super-resolution via other means to take unjust actions against its citizens is a valid concern. Since emergent technologies in the digital age are often discovered by different researchers in different places near-simultaneously, ultimately the best thing to do is to allow the technology to be used for good to counteract that harm that it might inevitably cause.

5. "The high-resolution technique may infringe upon the right to be forgotten. If I dig up someone else's old, low-resolution video showing something they did ten years ago and use the high-res technique to figure out what they did in the video in the past, that may not be ethical." **Response:** This is certainly a possibility; even if the face of the individual cannot be identified, super-resolution might still be powerful enough to somehow violate privacy rights if the subject of the image is known prior. Ultimately, the source of the issue is how the low-resolution image/video was distributed in the first place: to ensure that sensitive content is dealt with before distribution, implementations of our super-resolution network could be bundled with detection algorithms that flag potentially sensitive content to organizations/authorities, and/or send a warning to the user; Apple devices use such a system.

5. Conclusion

We have presented a novel method for using a relaxed, adversarial objective function for performing image super-resolution. In spite of limited training time, our network is capable of generating realistic super-resolution images at 4x magnification can generate finer details than networks trained with deterministic methods. Finally, the use of noise injection allows our network to explore the space of super-resolution images, although not to the extent that we had hoped. Future work will involve further training and potentially modifications to the architecture that promote heavier influence of noise injection.

References

[1] Alexia Jolicoeur-Martineau. On relativistic f-divergences. In *International Conference on Machine Learning*, pages 4931–4939. PMLR, 2020. 6

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. 2018. 2

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. *Proc. CVPR*, 2020. 2

[4] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1

[5] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1

[6] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. 2018. 2

Appendix

Generator Architecture Details: The generator consists of three primary components: the generator stages, the bilinear interpolation, and the upsampling blocks.

There are 4 generator stages, and each stage consists of 16 generator blocks. Each block consists of convolutional layer, followed by noise injection, a nonlinear activation, another convoluiton, a residual connection from the input, and another activation. Each generator stage has an input size (Stem Width) of 256 units, and has an output size (Feature Width) of 512 units. The outputs from each generator stage are combined via residual connections to mitigate gradient explosion.

Bilinear interpolation is a non-machine-learning method of upscaling the input image. Two 2x magnification bilinear interpolation layers magnify the input image to 4x, and these are connected with residual connections to the generator stage output to ensure that a reference to the input is kept throughout the network's forward pass.

The upsample blocks implement up-sampling, which is essentially the reverse of a convolution operation that takes input data and uses learned kernels to produce a larger output. There are two upsample blocks, and they have output sizes (Feature Widths) of 256 and 128 units.

The compression factor is equal to the magnification factor, and is used to module the number of output channels for each generator block. The receptive field specifies the kernel size of the convolutional layers of the generator blocks, and we use exclusively 3x3 kernels.

Discriminator Architecture Details: The discriminator consists of discriminator stages, followed by a global depthwise convolution and then fully connected layers.

Each discriminator stage consists of a downsampling block that compresses the data representation, followed by residual blocks that have identical structure to the generator blocks except without noise injection.

Global depthwise convolution essentially reduces teh dimensionality of the data, thereby preparing it for the fully connected layers.

There is nothing special about the fully connected layers; they serve as a critic network that takes the data

representation prepared by the previous parts of the network and produce the discriminator's output critic score.

Additional Architecture Details: The implementations for our model architectures are found in the src/model/NetworksV2.py file in our repository.

Fidelity Loss Details: As stated in the Method section, fidelity loss is defined as:

$$L_{\text{Fidelity}} = \|y - G(y)_{\downarrow_{\text{bicubic}}}\|_1 ,$$

where y is the low-resolution image, and $G(y)_{\downarrow_{\text{bicubic}}}$ indicates bicubic downsampling, which is an operation for reducing the resolution of an image. This means that the fidelity loss motivates the network to produce an output that, when downsampled, resembles the input.

Adversarial Loss Details: As stated in the Method section, adversarial loss is defined as:

$$L^G_{\text{Adversarial}} = \sup_{G:\mathcal{Y}\to\mathcal{X}} \mathbb{E}_{\substack{x\sim\text{HR}\\ y\sim\text{LR}}}[f(C(G(y)) - C(x))],$$

where y is the low-resolution image, $\sup_{G:\mathcal{Y}\to\mathcal{X}} \mathbb{E}$ means to maximize the expectation of the inner term by modulating the generator G, f is a concave function used to set an upper bound on the inner term, C is the discriminator (aka critic) network, $G(y)$ is the super-resolution image generated by the generator, and x is the ground-truth high-resolution image. This means that the adversarial loss motivates the generator network to produce fake images get as much as an improvement as possible over the ground-truth, as evaluated by the critic network.

Critic Loss Details: As stated in the Method section, critic loss is defined as

$$L_C = \sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\substack{x\sim\text{HR}\\ y\sim\text{LR}}}[f(C(x) - C(G(y)))] + \frac{\gamma}{2}\,\mathbb{E}_{x\sim\text{HR}}\left[\|\nabla_x C(x)\|^2\right]$$

$\sup_{C:\mathcal{X}\to\mathbb{R}} \mathbb{E}$ means to maximize the expectation of the inner term by modulating the discriminator C. f is again a concave function. $\|\nabla_x C(x)\|^2$ is known as the gradient penalty, which is commonly used to train GANs to improve training stability. It enforces the gradients to have unit norm, which mitigate the likelihood for vanishing and exploding gradients.

Additional Loss Details: [1] has theoretically proven that $L^G_{\text{Adversarial}}$ and L_C in conjunction define a divergence, which is to say that it measures the difference between the distribution of the generator and the distribution of natural images. Therefore, we can be confident that objectives push generator G to produce results which lie in the set of natural images.

The implementations for our loss functions are found in the src/loss/LossV2.py file in our repository.

Team contributions

Please describe in one paragraph per team member what each of you contributed to the project.

Nick Huang Devised and implemented network architectures by incorporating his highly technical background and prior experience with generative models. Helped tune model hyperparameters, and also outlined initial structure of the poster. Acted as a team leader who defined the direction of our project.

Shijie Mao Found an affordable cloud gpu option, maintained payment, and ran our experiments on the cloud and updated us on training progress. Helped tune model hyperparameters, and also helped document our progress on project reports and drafted initial graphics for the network architecture.

Rohit Mohanty Provided feedback on the direction we were taking the project, poster, and report. Documented our progress on project reports, and provided revisions for poster design and final report. Drew up the final network architecture diagrams used in this report and our poster.

Dustin Wu Worked with and adapted the EDSR code base so that we could use it to train our models, produce image outputs, and save model checkpoints. Helped Shijie Mao with running experiments, outlined structure of the final project report and responses to societal discussion, and added graphics to the poster and report.