# **Improving SRGAN for Object Detection**

Ivan Abraham University of Illinois at Urbana-Champaign 1308 W Main Street itabrah2@illinois.edu Wyatt McAllister University of Illinois at Urbana-Champaign 1308 W Main Street

wmcalli2@illinois.edu

Justin Wasserman University of Illinois at Urbana-Champaign 1308 W Main Street

jbwasse2@illinois.edu

# 1. Summary

In our proposal we called for implementing the network introduced in [2] in the production version of PyTorch as a way to learn more about neural networks as well as the new PyTorch library. We hypothesized that the network could be made more robust by changing the down-sampling techniques it is exposed to during training. Recall, in the original paper the authors implemented a convolutional neural network which, when trained on sample images, was able to perform science-fiction like image enhancement as commonly shown in television. Termed, single image superresolution (SISR), this problem had received extensive attention in the community. As is well known, this is an illposed problem exacerbated by loss of high frequency information due to down-sampling. Their approach to solving the problem was based on a content loss measuring perceptual similarity in super resolved images, in combination with adversarial training to create Generative Adversarial Network for Super Resolution (SRGAN). We were able to get the network implemented in PyTorch; further we were able to test our hypothesis with positive results vis-a-vis an object detection metric.

We first studied the algorithm the authors made in [2]. The authors implemented the network in a very old version of TensorFlow  $(1.2 \sim 1.4)$ . We re-implemented it with latest release of the PyTorch library and ran it on our GPU. A very similar model in [6] had a more objective performance metric of object detection, while [2] has a subjective metric Mean Opinion Score (MOS), based on human viewers. To avoid the use of human opinions when testing our hypothesis, we used an object detection performance metric. Specifically, we compared the number of objects detected in super resolved images generated by the original network, and by a network trained with our modifications.

### 2. SRGAN

We first discuss the basic premise behind the SRGAN network. Many previous methods focused exclusively on pixel-wise reconstruction methods such as peak signal-to-noise ratio, which focus only on the image pixel data, and do not correlate with the image content perceived by humans. SRGAN instead focuses on a content loss,  $l_X^{SR}$  (Equation 1), capturing the difference between the features of the high resolution and generated images, where  $I^{HR}$  and  $I^{LR}$  are the high resolution and low resolution images,  $\phi$  is the feature map, and G is the generator of super resolution images.

$$l_X^{SR} = \phi\left(I^{HR}\right) - \phi\left(G\left(I^{LR}\right)\right) \tag{1}$$

During training the content loss  $\phi$  was implemented using pre-trained VGG16 and VGG19 networks available by default in the PyTorch library [?]. In the original paper the total loss for the generator was a combination of the content loss from (1), the pixel-wise mean square error (torch.nn.MSELoss()) and adversarial loss where the adversarial loss,  $l_G^{SR}$  is given as in (Equation 2). Here D is the discriminator, which outputs the probability that the generated image is a natural high resolution image.

$$l_G^{SR} = -\log D\left(G\left(I^{LR}\right)\right) \tag{2}$$

SRGAN consist of two parts: a convolutional feed forward network that learns features to create super-resolved images, and a discriminative network that functions as an adversary, and is trained to tell apart real and super resolution images. The novel approach in this paper is the use of a content loss, which associates local textures with patches of the image, and uses them to fill in detail in the high resolution image. The neural network finds a feature representation for the image, and determines textures for each patch based on these features. These textures are then combined and smoothed to create the final super resolution image.



Figure 1: Architecture of Generator and Discriminator Networks with corresponding filter size (k), feature number (n), and filter time step (s) indicated for each convolutional layer.

### 2.1. Discussion

The neural network architecture used in [6] and [2] presented several advantages over past work. The exclusive use of convolutional layers allowed a single model to be trained on an input image of any size. The use of a feed-forward architecture rather than a recurrent architecture improved efficiency by passing the image through only once to get the result. Rather than provide a bicubic upsampling of the low resolution (LR) image as input, the authors upsampled feature activations so that the image gets upsampled to the final dimensions of the high resolution (HR) image as it passes through the network. Additional convolutional layers were added for extra smoothing of the output image features. The use of residual blocks instead of stacked convolutional layers improved efficiency.

#### 3. Project Description

We used the object recognition method from [6], as implemented in [5] to benchmark the performance of SRGAN, from [2]. We used the Python Imaging Library (PIL) [?], with Nearest Neighbor, Linear, Bicubic, and Sinc downsampling methods to try and improve robustness. We then performed a second experiment where we removed the Linear and Nearest Neighbor interpolations, and updated the loss from VGG16 to VGG 19, which is known to have higher performance, and examined the affect on the performance for object detection.

#### 4. Description of Network

Let k, n and s denote filter (or kernel) size, number of features detected by each layer, and the time step for each filter. Then the architecture used by the authors, [2], modified from the implementation in [3], is given as in Figure 1. All kernels were  $3 \times 3$ . We used 16 residual blocks. Conv stands for a convolutional layer, and (P/I)ReLU stands for (Parametric/leaky) rectified linear unit, BN stands for batch normalization,  $\oplus$  denotes elementwise sum.

For the object detection metric we used the third iteration of a pre-trained neural network called YoLo (You Only Look Once!), trained on the Common Objects in Context (COCO) Dataset [?]. This is a state of the art real-time object detection network that has been validated on many major data-sets including images and video. We refer the reader to [5] for details on the network's architecture.

#### 5. Methods and Experiments

First, we implemented a convolutional neural net to super resolve images, before adding the generative adversarial net, which give SRGAN its high performance. We then used the object detector in [5] as metric to compare performance for the proposed modification to training. We finally benchmarked our implementation, on the aforementioned object recognition metric to see if training the network on images downsampled with different methods improved the performance under this metric with positive results.



Object Detection Error Vs. Interpolation Scheme

Figure 2: Object Detection Error Results

# 6. Resources Used

We ran the experiments on a NVIDIA GeForce GTX 1070 GPU on Mr. Wasserman's PC. To train SRGAN, we utilized the Visual Object Classes (VOC) 2012 Dataset, which consists of over seventeen thousand images, labeled with object classes [?].

#### 7. Implementation

The production version of PyTorch used gave several advantages over the original TensorFlow implementation. In the implementation from [2], the VGG network and its API had to be written into the network, whereas Py-Torch already provided pre-trained models as we have already noted. The use of PyTorch made it much easier to implement the content loss of (1) using VGG models. Further, one could easily swap VGG16 model for the VGG19 model by simply changing a few lines. The module torch.utils.data.dataset.DataSet, when sub-classed properly, requires the definition of the internal objects \_\_getitem\_\_() and \_\_len\_\_(). This allows commands to be more "Pythonic" vis-a-vis generator and list comprehensions. Moreover, the pre-processing steps to be done on the image (crop, transform etc.) could be built into the \_\_\_\_\_() method so that this does not need to be handled separately. Finally, PyTorch also provided the related module torchvision.transforms, which contains most common image transformations one uses. By converting the loaded data to a Python Imaging Library (PIL) image format, many transformations could be composed together easily and conversions from images to PyTorch tensors to PIL Images could all be handled automatically. This supports the folk-news that PyTorch is indeed a more Pythonic framework for neural networks.

## 8. Results

#### 8.1. Object Detection Metric

For the object detection metric [5], the original SRGAN Network fails for several objects in many images, such as the one in Figures 4b and 5b. As shown in Figures 4d and 4f, the network trained with the use of multiple downsampling schemes, which we call Multi-Blur, outperforms the original SRGAN network. When the Multi-Blur network is retrained without the Linear and Nearest Neighbor interpolations, which are suboptimal for upsampling, the performance drops again, as shown in Figures 5d and 5f. This fact, along with the fact this new scheme included the VGG19 loss instead of the VGG16 loss, which is known to perform worse without additional multi-scale training tricks [4], suggests that the loss of some texture information leads to the increased performance. In essence, a loss of some texture content creates increased robustness of object detection to the downsampling method used.

As shown in Figure 2, the training scheme Multi-Blur was tested on the VOC 2012 data set, composed of images with pre-labeled object classes, and achieved superior performance in terms of object detection error for all interpolation schemes tested. The Multi-Blur (VGG19) scheme does not show the same improvement without the Linear and Nearest Neighbor interpolations. The reason for this is hypothesized to be the fact that the learned weights for the generative network for the super-resolved images reflect all of the interpolation schemes. This results in a smoothing out of textures across the entire image. This smoothing causes a lower variation in the texture information within objects, improving object detection across all interpolation schemes, at the cost of reducing photo-realism.



Figure 3: Super Resolution Comparison

#### 9. Discussion of Multi-Blur Performance

As seen in Figure 3, our network training scheme, which we call Multi - Blur, reduces image artifacts in homogeneous image regions, at the expense of smoothing textures within objects. We see from Figures 4a, 4c, and 4e, it also does well with fine grained features such as text. As shown in Figure 4, the Multi-Blur training scheme produces readable text for this excerpt of Moby-Dick, while the original network fails to do so. From Figures 5a, 5c, and 5e, we see Multi-Blur (VGG19) offers a slight improvement for text, despite only having a slightly higher classification accuracy.

# **10.** Conclusions

This training scheme improves the SRGAN network performance for object detection. This improvement comes at the expense of smoothing out image textures. While less photorealistic, the SR images have less image artifacts, and do well on images with regular features, such as text.

## **11. Possible Extensions**

The authors of SRGAN comment that super-resolution of text images is hard because the loss function used tends to 'hallucinate' texture details which may corrupt text content. One could investigate whether this is an artifact of the network architecture, or that of the loss used from [7]. The substitution of *more appropriate* loss function suited to text may allow the same network to perform well on SR of textual images. This may possibly be done by training modifying the VGG network to train on text images.

A possible extension to this work is to utilize optimal character recognition (OCR) in conjunction with the adversarial network so the adversary compares OCR generated text from SR image (i.e. SR-text) to OCR generated text from the HR image in an effort to force the generative network to learn parameters that are suited for text-image SR. A version of this idea has appeared at the 2018 International Conference on Pattern Recognition [1].

# References

- A. Lat and C. Jawahar. Enhancing OCR Accuracy with Super Resolution. In 2019 International Conference on Pattern Recognition (ICPR). IEEE, jul 2018.
- [2] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 105–114. IEEE, jul 2017.
- [3] leftthomas on GitHub. Pytorch Implementation of SRGAN. https://github.com/leftthomas/SRGAN, 2018.
- [4] Medium. A Review of VGGNet The 1st Runner Up for The Image Classification Task, in ILSVRC. https://medium.com /coinmonks/paper-review-of- vggnet-1st-runner-up-of -ilsvlc-2014-image-classification-d02355543a11, 2014.
- [5] ryoppippi and ayooshkathuria. An Implementation of the YOLO v3 Object Detection Algorithm. https://github.com/ ayooshkathuria/pytorch-yolo-v3, 2018.
- [6] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. *Computer Vision (ICCV), 2017 IEEE International Conference on*, dec 2016.
- [7] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv, sep 2014.

# 12. Appendix: Additional Examples

Figure 4 shows additional examples.



(a) Object Detection: Original Image



(c) Object Detection: SR Image - Multi-Blur



(e) Object Detection: SR Image - SRGAN

Herman Meirsile
It would by waves and cooled will content them but the or- ply bream which for the human means that if the hand hi- pervisons were out of again of terring under the shady less of limit. Loss it the croweds of yood wavehouses will not water-gazen there. suffices National States and the states and the states of the states of the states of the Carcinnoise and the states of the states of the states of form Corleans Host to Car- ter and the states and states and of them the states and the states and states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the correst states and the states and the states and the states of the states and the states and the states and the states of the states and the states and the states and the states of the states and the states and the states and the states of the states and the states of the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and the states and th
What do you such "build like logs, stretch and ar- elect seminals all accound the erasesmoth start, south, and town, atand thousands upon west. Yet here they all units, thousands of mortal mer fixed Tell m, do so the magnetic in ocean revertes. Some lean- virtue of the needles of the ing against the splites some compasses of all those ships at-
sould upon the piter-backs, tract them titther? some looking over the bal. Once mees, say you are in warks of ships from China; the country, in some high land some high adds in the rigging, or lakes. Take almost any path as if striving to gat a still better you please, and ten to one it seaward peerg Bot these are all carriery our down in a dak, and
in an animity, of weak any's pict. leaves you there by a face in up in this and plaster—which to the aream. There is magic is clinched to deals. How there is offen the picture of the clinched to deals. How there is of much be pinqued in his deep- thel A are the green field green ext ervecters—and that mus what do they here? one has he will infillably lead you
crowds, pacing straight for the to water, if water there be in all water, and sceningly bound that region. Shaddy one over be for a dive. Strangel Nothing athirst in the great American

(b) Text: Original Image

	Norman Michelle
1. Loomings	a washadi yaran asi could will convert down but on he beams, which is its hears turns much limit of the italia persons users out of affer of turing moder the data bits of band. Limit is the commod or much washams, and and use against films. It is the turn of the second of the observation of the second of t
when the standard based and the standard starts are to the transmission of the start of the star	We work at a second se
anotes da resu el de es engries. Nelle ataré e dans, corres very menty the anni industa suveral du occani with ana. There new is your incidea ethy of the Markathone, bibled overal by whereas an latitus aleb proving mode-constraines arranged is with her next Right and left, the atoms take your sentement. In extreme downtown is the battery, where that node mode	bin of its global management on restorm-stand stars stars. What do like have? on a lisk keys with low drag applic, But loads have room more and be will stallably load yes crowle, paraging its glob for the twent, if water fracts by a Bi water, and severagily board that region. Should you ever be for a drive Stangel Nothing athein in the great American

(d) Text: SR Image - Multi-Blur

	Heemen Mabelle
1. Loomings	is stabled by warve, and coded will content them has the by browne, which is fore learn iteration later of the limits to limit. Look it the crawho of yorder variablemass will re- water-gaters there. Crossmethelists the cry of simplify where they post a dwarry Sabbah infermion. by can without failing to a Go frame Gaters Hole (no. 2) where they post a dwarry Sabbah infermion.
all me tahenaol. Some years ago-never mind how long precisely-shaving little or no money in my parse, and nothing particular to interest me on shore, I thought I	by Writeball, rasefiseard, they come from hance and al Writeball, rasefiseard, they come from hance and al Writeball, account the comes-meth, each confi, an effect semisolis all account the comes-meth, each confi, an town, stand thoesands apon west. Yet here they all units of the semi-semison of the semi-semi-semi-semi-semi- teball.
would sail also a fitted and use the watery part of the workd. If it	in ocean reverts. Some lean-
a way I have of driving off the sphere and regularing the circula-	virtue of the meetine of the meetine of the
tion. Whenever 1 ind reposit graving grin about the mostly	tag against the splace some compassion of all those ships at
whenever it is a damp, drizzly Nosember in my analy whenever 1	seated upon the part-backs; tract them thather?
and mysell involuntiarity passing before colum warehouses, and	some tooking owr the bas-
bringing up the near of every fuseral 1 meet; and especially	works of ships from China; the country; in some high lan-
whenever my hypeo get such an upper hand of me, that it re-	some high aloft in the rigging, of lakes. Take almost any path
stepping into the street, and methodically knocking people's hats	seaward peep. But these are all carries you down in a dale, an
off-then, I account it high time to get to an as soon as I can.	landsmore of week days pent, leaves year there by a pool to
This is my substitute for pixel and ball. With a philosophical	up in lath and plaster—tied to the stream. There is magic is
floarish Cato throws himself upon his sword; I quietly take to	counters, mailed to benches, it. Let the most absent-minder
the ship. There is nothing surprising in this. If they but knew it,	clinched to desic. How then is of men be plunged in his deep
almost all men in their degree, some time or other, cherish very	that Are the green fields gene? out reventes-stand that may
nearly the same feelings towards the occurs with me.	What do they here? on his logs, set his feet a going
There now is your insular city of the Manhattons, belied round	That look? here come more and he will infallibly load ye
by seharnes as Indian tiles by coral reefs—commerce surrounds	crowds, pacing straight for the to water, if water there be in al
it with her surf. Right and left, the streets take you waterward. As	water, and seemingly bound that region. Should you ever b

(f) Text: SR Image - SRGAN

Figure 4: Examples with Multi - Blur. Top row shows original image. Bottom row shows output of SRGAN when given LR images downsampled with algorithm different than what it trained on. Middle row shows our algorithm, which is more robust to downsampling technique.

Figure 5 shows additional examples.



(a) Object Detection: Original Image



(c) Object Detection: SR Image - Multi-Blur (VGG19)



(e) Object Detection: SR Image - SRGAN

	Herman Melville
1. Loomings	It washed by waves, and cooled will contact thom but the cri- ly brance which a fee hours - more limit of the land. Ici- previous were out of split of terring under the shark let of land. Lock is the cruwel of your burger wavehouses will not water gazen hore. Sum of your wavehouses will not crease the start of the split of the split of the provi- a dramy Subhut literation. We can without falling in And Griffon Toolens likels too:
In the half-hand long years age-nerve mind how long function-through their sea means in two pressus and the half-hand pressing their sea means in two pressus and the half-hand pressing the mind hand the half-hand pressing grant how the half-hand pressing the half-hand pressing grant how the half-hand pressing the half-hand pressing the half-hand pressing grant how the half-hand pressing grant how the half-hand pressing grant how the hand pressing the half-hand pressing the half-hand pressing grant how the hand pressing the half-hand pressing the half-hand pressing the half-hand pressing hand pressing has been hand pressing the half-hand pressing hand pressing has pressing the half-hand pressing hand pressing has pressing hand pressing has pressing hand pressing has pressing hand hand pressing has pressing hand hand pressing hand hand hand pressing hand hand hand pressing hand hand hand pressing hand hand hand hand pressing hand hand hand pressing hand hand hand pressing hand hand hand pressing hand hand hand hand hand hand hand hand	entre Sip, and from theore, them-laques. Inlution: all by Withing, northered, the concer from Lane and al- What drymater.1-Notable large, means and al- What drymater.1-Notable large, means and al- tors, main discussing, support with the relation of lating, theorem and the splits, means and the splits of large dis- tion constructions. Since lane, "turn of the median of the large against the splits, means compares of all doord hips at some large over the last." One means, stryma are in works of datage from Chine, the compry, in some high land and minimizing our data last. The lands are path in fairtung up at all Blance, the construction of the path indication of the largest of the last last last last last handmann of which last last last last last last last last
ı.	a
(b) Text: Or	iginal Image
	Hernse Mebelle
1. Loomings	It would by versus all coded will content then but its ar- by brones, thick is to be here. There there it the link kar- provises were out of algo of string and/or the add by los of had. Los all is the corrected or product would have all Concentrations for the string of the string o
Call my bilanced Some water and prover stead how loans	What do you see?-Posted like leve, streets and av-

The set obtained, hence yours ago—never straid here long processing—howing life or no money in my perior, and yours and an interest in the set of the set of the set of the yours of an interest in the set of the set of the set of the set that with the set of the se	Concentration the ory of a single between the proof- s theory behaviors. How conversions labeling is a de- creases by, and from theses. How conversions labeling is and the single sector of the single sector of the single sec- lection of the single sector of the single sector of the description of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector the single sector of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector of the single sector the single sector of the single sector of the single sector research sector of the single sector sector of the single sector of the single sector of the single sector sector of the single sector of the single sector of the single sector of the single sector sector of the single sector of the single sector of the single sector of the single sector of the single sector with the defined between the single sector of the single sector of the single sector with the single sector of the single se
by wharves as Indian idee by coral reefs—conneecce surrounds	crowie, pacing stratight for the to water, if water there be in all
it with her surf. Right and left, the streets take you waterward. Its	water, and seemingly bound that region. Should you ever be
surreme downtown is the buttery, where that noble mole	for a direc. Strangel Nothing athinst in the prost American

(d) Text: SR Image - Multi-Blur (VGG19)

1. Loomings . Constraints and the service of a give of a service of the service of a service of a service of a service of the service of a service		Herman Medvelle
Support of the start of the sta	1. Loomings	is washed by waves, and cooled will content them but the en- by Pereson, which a few beams termed hand the lanck los- provision were cost of algo? of waves and the second second lance is a second second second second second second name-game them. The Second Second Second Second Concentrabilish the effort of the second second second or form Cost in the second second second second second cost of the second second second second second second cost of the second second second second second second cost of the second s
These roots is your tradie (cry of the Machathene, belied recend) Tai Joshi here corear mass and he will institliby load yes by wharres as Lahati usiby your an effect meansature arreamed. It with here rard Right and left, the interest take year surgered for a dire. Standie for the two which, if we first here is the correst downshows to the bettery, where this node node for a dire. Stangie Neeking, addit the period. Here the	We into schemals, losse your age-accor mind here long records—burstle for ex reasons in trapera, and your data all adva, tittle and see the samey part of the workli. It is a start that all the advances in the start the trans- titue. We see even if in the days of the advances in the start of the start the start of the start of the start. It is a start of the start of the start of the start is the start of the short of the start of the start of the start of the start of short of the start of the start of the start of the start of the short of the start of the start of the start of the start of the short of the start of the start of the start of the start of the short of the start of the start of the start of the start of the short of the start of the start of the start of the start of the start of the start of the start of the start of the	What do you cell-broad like large, neces and an distance trade of anomal for more meeting, and an other broad for a section of more meeting and the section of the broad for a section of the section of the negative flag profile the section of the section of the negative large profile the section of the section of the negative large profile the section of the section of the negative large profile the section of the section of the negative large profile the negative l
	These face is your instant CRy of the MaintainCool, Genton (Genta) by scharters as a findiant takeb by conflection-contractors unremeable it with her surf. Right and left, the streets take you swaterward. Yo surreme downtown is the battery, where that noble mole	and issues here come mode that he was intransity and you crowells, pacing straight for the to water, if water there be in the water, and seemingly bound that region. Should you ever be for a dive. Strangel Nething athint in the great American

(f) Text: SR Image - SRGAN

Figure 5: Examples with Multi - Blur (VGG19) with Linear and Nearest Neighbor Interpolations Removed. In this case the middle row is only slightly better than the bottom row showing that inclusion sub-optimal interpolation algorithm during training make for a more robust network.