Attention-based super-resolution GAN for drone image small object detection

Soorya Ram Simgekar(soorya2), Pratikshit Singh (ps71), Apoorva Udupa (audupa3)

Abstract

We present a small object detection model that uses a GAN(1)-based super-resolution as one of the steps for better detection results. Previous work on small object detection (2) shows the problems for such tasks, because tiny object only contains a few pixels in size, when used with only CNN models. Our work tries to build a method that can be used to generate better results. We will be using ESRGAN (3) as our reference architecture to enhance the image resolution. Along with this, we propose a two-step attention mechanism namely; spatial and channel, to further increase the Super Resolution Output. The Spatial unit will extract "where" the important information is located, while the Channel unit will extract "what" is meaningful from the image. We propose to use a GAN(1)-based architecture for Super Resolution instead of just Deep CNNs because we found that approaches using only CNNs fail to produce finer details after enhancement, which is of crucial importance for us to get better super-resolution results. The reason is that the objective function in these approaches is related to Mean Squared Reconstruction Error (4), which has high peak signal-to-noise ratios but lacks high-frequency details for finer patterns and textures. ESRGAN (3), tries to alleviate this problem by not only using the general Adversarial Loss of GANs(1) but also proposing an additional "content loss", which is based on perceptual similarity instead of pixel similarity. We propose combining it with the attention mechanism (5) because, given the large image dimensions and quantity, we wanted to maximize the training quality by improving the extracted image features while reducing the training time needed.

Keywords: GAN based Super resolution, Small object detection, Attention, YOLOv5

1. Introduction

Small object detection can be particularly challenging due to the limitations in images and the small size of the objects themselves, as stated in (6) and (2). Some of the key limitations of small object detection include:

- Low resolution: Drone/Satellite images may have a lower resolution than other types of imagery, making it difficult to detect small objects.
- Occlusion: Small objects may be occluded by other objects, making them difficult to detect.
- Scale variability: Small objects may vary in size, making it difficult for the object detection algorithm to recognize them.
- Background clutter: Images may contain a lot of background clutter, such as trees, buildings, or other objects, which can make it difficult to accurately detect small objects of interest.
- Small objects occupy a relatively small number of pixels in an image, they can be difficult to detect and differentiate from background noise or other small objects.

Techniques like super-resolution can help to improve the accuracy of small object detection. Super-resolution(7) is the process of generating a high-resolution image from a low-resolution input image. This can be accomplished using a generative adversarial network (GAN)(1). GANs consist of two parts: a generator network and a discriminator network. The generator network takes a low-resolution image as input and tries to generate a high-resolution version of the same image. The discriminator network, on the other hand, takes in a high-resolution image and tries to determine whether it is real or fake (generated by the generator network). The two networks are trained together, with the generator network trying to fool the discriminator network and the discriminator network trying to correctly identify real and fake images. Over time, the generator network learns to generate high-resolution images that are indistinguishable from real images, allowing it to perform super-resolution. Super-resolution can be used to improve the performance of object detection algorithms, particularly in the case of small objects. Because super-resolution generates a high-resolution version of an input image, it can make small objects appear larger and easier to detect. To use super-resolution for small object detection, the low-resolution input image is first passed through a super-resolution network to generate a high-resolution version of the image. The high-resolution image is then fed into an object detection algorithm, which can more easily identify small objects that were not visible in the low-resolution input image. Overall, super-resolution can be a useful tool for improving the accuracy of object detection algorithms, particularly when the objects of interest are small and difficult to detect in low-resolution images.

Along with this, a mechanism called attention is also employed by us, to focus on specific parts of the input image and generate those parts in greater detail. There are two branches of attention mechanisms(8) that we plan to use: channel attention

and spatial attention. Channel attention refers to the process of adjusting the importance of different channels (i.e., color channels in an image) in the generated image. For example, if the input image has a strong red color, the generator network may use channel attention to increase the importance of the red channel in the generated image to make it more realistic. Spatial attention, on the other hand, refers to the process of adjusting the importance of different spatial locations in the generated image. For example, if the input image has a sharp edge in a particular location, the generator network may use spatial attention to increase the detail in that location in the generated image to make it more realistic. In general, attention mechanisms in GANs help the generator network to focus on the most important parts of the input image and generate higher-quality images.

In this paper, we proposed a GAN-based super-resolution method based on a deep fusion network for small object detection on the COWC dataset. We combined super-resolution model ESRGAN(3) architecture with channel and spatial attention mechanism and used YOLOv5 for small object detection. (9)

2. Motivation

The reason why our problem statement is significant is that it has many real-world applications in sectors like surveillance and security. In these cases, using a drone, an easy and fast method is required to detect objects of interest quickly. The motivation of our work is, therefore, to improve the object detection capabilities for drone images using our attention-enhanced GAN architecture of super-resolution. Along with this, we want to explore the use of channel and spatial attention on a GAN architecture with the goal to improve the quality of super-resolution with a decreased training time. The project aligns with the course because we're applying various concepts from the course, such as attention and generative adversarial networks. More precisely the super-resolution part aligns well with the course because we'll be working on a conditional generative adversarial network to generate the enhanced images.

3. Related Work

Similar work on SRGAN(5) and ESRGAN(3) implementations were reviewed for the super-resolution model for the first step of our proposed method. SRGAN produces visually similar high-resolution images and is robust to distortions commonly found in low-resolution images. However, it can have unpleasant artifacts. ESRGAN enhances the visual quality and provides sharper edges and visually pleasing results through improvements to the SRGAN architecture, adversarial and perceptual loss, and residual-residual dense block (RRDB). However, it requires careful tuning and may be sensitive to hyperparameters and longer training time than other super-resolution methods.

The Remote Sensing Image Scene Classification Based on an Enhanced Attention Module(8) paper proposes an Enhanced Attention Mechanism (EAM) to improve object detection in remote sensing images with complex backgrounds and small objects. The attention mechanism works by weighting input features based on a "query" vector to focus on relevant parts of the input. The proposed model includes multiple layers of CNNs, adding computation to the model.

3.1. Super Resolution

Traditional super-resolution methods use deterministic models to generate high-resolution images from low-resolution inputs. In contrast, SRDiff(10) utilizes diffusion probabilistic models to generate super-resolved images.

The authors of the paper(10) propose a new method for single image super-resolution, which is the task of increasing the resolution of a low-resolution image to create a higher-resolution version that preserves important image details. The proposed method, called SRDiff, uses diffusion probabilistic models to generate super-resolved images. The diffusion-based upsampling involves iteratively applying a diffusion process to increase the resolution of a low-resolution input image shown in Figure 1. The process starts with a low-resolution image and gradually introduces noise while increasing the resolution. The noise is added to the image in a way that is proportional to the gradient of the image so that edges and other high-frequency features are preserved. The noise level is gradually reduced as the resolution of the image increases, and process is repeated until the desired resolution is achieved.

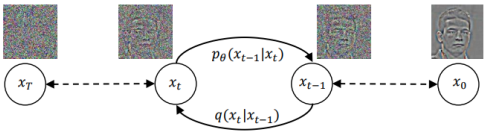


Figure 1: Overview of two processes in SRDiff. The diffusion process is from right to left and the reverse process is from left to right. θ in pθ denotes the learnable components including conditional noise predictor and low-resolution encoder in SRDiff.(10)

Once upsampling of the diffusion model is complete, the final image is denoised to remove the noise added using a trainable denoising network. The denoising network is further applied to the upsampled image to remove the noise and produce the final super-resolved image. The authors evaluate the method on several standard datasets, like Set5, Set14, and Urban100. The results from above were compared to state-of-the-art methods and demonstrated that their method achieves better performance in terms of both quantitative metrics (like peak signal-to-noise ratio and similarity in structural index) and visual quality.

In another related work, the authors of the paper(11) use an approach that involves a generator network that generates high-resolution images from low-resolution inputs and a discriminator network that differentiates real images from generated images. The generator is trained to generate high-quality images in order to fool the discriminator, which is trained to precisely distinguish between real images to that of generated images. The authors introduce a new loss function that combines both

L1 and SSIM losses to considerably improve image quality. Results from the above combination show that the method proposed outperforms state-of-the-art methods in terms of quantitative metrics and the visual quality of reconstruction. The method has the potential to improve the clinical diagnosis and treatment of cardiac diseases by providing high-quality images for medical analysis.

The authors of the paper(12) propose a new method for parallelly performing facial landmark localization and super-resolution of low-resolution faces in random poses using Generative Adversarial Networks (GANs). The proposed method is called Super-FAN, it mainly consists of two parts as shown in Figure 2: a network of facial landmark localization and a super-resolution network. The facial landmark localization network considers a low-resolution face image as input and predicts the 2D coordinates of around 68 facial landmarks. The super-resolution network takes the same low-resolution image of a face as input and outputs a high-resolution face image.

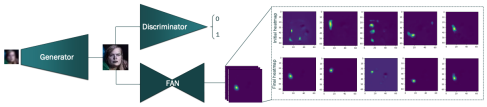


Figure 2: The proposed Super-FAN architecture consists of three connected networks: the first network is a newly proposed Super-resolution network. The second network is a WGAN-based discriminator used to distinguish between the original HR image and super-resolved The third network is FAN, a face alignment network for localizing the facial landmarks on the super-resolved facial image and improving super-resolution through a newly-introduced heatmap loss(12).

The super-FAN model is a GAN-based method that achieves parallelly facial landmark localization and super-resolution of low-resolution face in arbitrary poses. A new dataset LRFA is created by downsampling high-resolution face images from existing datasets and applying random rotations and translations to simulate arbitrary poses. The Super-FAN model is then trained with a landmark localization loss and a super-resolution loss additive to adversarial loss. The model achieves state-of-the-art performance on benchmark datasets and is robust to variations in pose and lighting conditions. The architecture comparison with SR-ResNet could be seen in Figure 3

4. Proposed Approach

As part of this implementation, we have explored various super-resolution models like SRCNN, and EDSR and implemented an ESRGAN (3) from scratch in Python using Keras and TensorFlow libraries taking the architecture described in the paper as a reference. The architecture in general first downsamples the low-resolution images in a way that the information is represented in dense vectors. Later while upsampling this dense representation the model learns how to represent it in a high-resolution way.

To improve the accuracy of the super-resolution model, we've included an Attention mechanism(8) in the residual

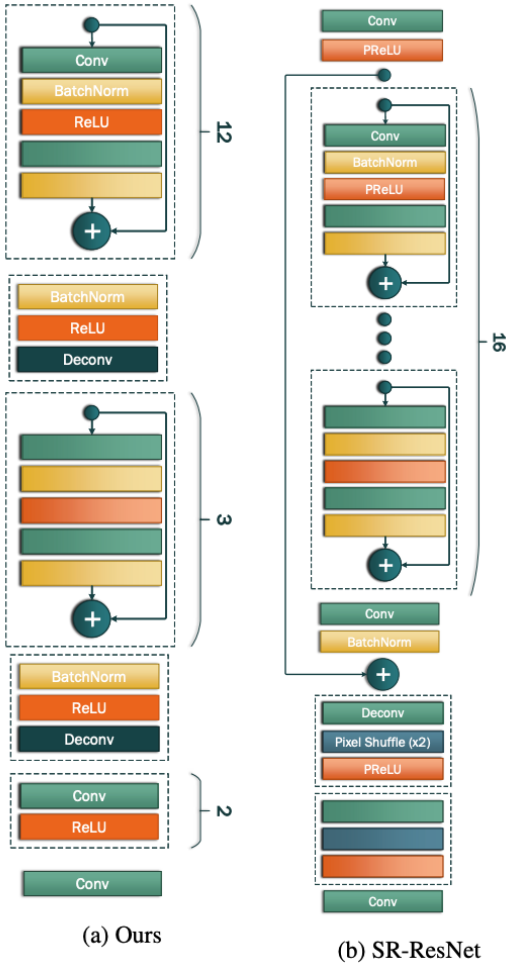


Figure 3: A comparison between the proposed superresolution architecture (left) and the one described in the paper.(12)

path of the ESRGAN. The attention consists of two sub-architectures namely, "Spatial" and "Channel". The Spatial unit will extract "where" the significant information in the image is located, while the Channel unit will extract "what" is meaningful from the image.

After the super-resolution model is trained, we pass all the low-resolution images (64x64), through the trained model, which up-scales it to (256x256). These images are then saved in a folder, which is used to train the YoloV5(9) model for detecting small objects, i.e., cars.

4.1. Dataset

We explored various satellite and aerial datasets and selected the Cars Overhead with Context dataset (COWCD) for our project. An example of the images in COWC dataset could be seeing in Figure 4. We will resize the data from (256, 256) to (64, 64) and back to (256, 256) using bilinear interpolation to create a low-resolution image. The original dataset is available at https://github.com/LLNL/cowc. The dataset consists of;

- Data from overhead at 15 cm per pixel resolution at ground (all data is EO).

Figure 4: Example image from the COWC dataset

- Data from six distinct locations: Toronto Canada, Selwyn New Zealand, Potsdam and Vaihingen Germany, Columbus Ohio, and Utah United States.
- 32,716 unique annotated cars. 58,247 unique negative examples.
- The intentional selection of hard negative examples.
- An established baseline for detection and counting tasks.
- Extra testing scenes for use after validation.

We have created the required dataset for GAN based super-resolution model, by first downscaling and rescaling it back to produce the desired low-resolution effect. This process is visualized in Figure 5.

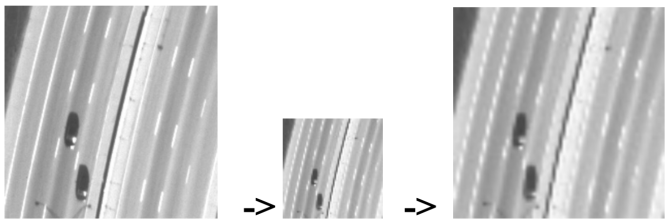


Figure 5: Process of creating high-res and low-res data for training

4.2. Image Super-Resolution

We will refer to ESRGAN (3) and use it as our base architecture using Keras in Python. The choice of the GAN-based model over only CNN is due to the ability of GAN-based super-resolution models to enhance finer details of the subject, which is crucial for small object detection from drone images. Residual blocks and dense connections will be used in the implementation.

For this model of super-resolution, the concept of perceptual loss is derived from (13). Here, the perceptual loss is defined on the activation layers of a pre-trained deep network where the distance between two activated features is minimized. VGG19 is the preferred pre-trained deep network of choice. However, this approach has two drawbacks. First, the activated features are very sparse, especially after a very deep network, which provides weak supervision and leads to inferior performance. Second, using features after activation causes inconsistent reconstructed brightness compared to the ground-truth image. To overcome these drawbacks, we propose to refer to the strategy proposed by the authors of (3) which uses features before the activation layers of VGG19. This approach will provide better supervision and result in more consistent reconstructed brightness compared to the ground-truth image.

The complete loss of the model will therefore include, perceptual loss (L_{percep}), content loss (L_1), and adversarial loss (L_G^{Ra}), as shown in Eq (1).

$$L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_1 \tag{1}$$

In Eq (1),

- L_{percep} = Distance between features from VGG19 of generated image and ground truth.
- L_G^{Ra} is the adversarial loss for the generator, which is given by, $L_G^{Ra} = -\mathbb{E}_{x_r}\left[\log\left(1 - D_{Ra}\left(x_r, x_f\right)\right)\right] - \mathbb{E}_{x_f}\left[\log\left(D_{Ra}\left(x_f, x_r\right)\right)\right]$. Here $x_f = G(x_i)$ and x_i stands for the input LR image.
- $L_1 = E_{xi}\left\|G\left(x_i\right) - y\right\|_1$ is the content loss that evaluates the 1-norm distance between recovered image $G(x_i)$ and the ground truth y.
- λ, η are hyperparameters to achieve a balance between noise and blur.

In Figure 6, we can see the generator architecture of our GAN-based super-resolution model which is a modification of the ESRGAN generator.

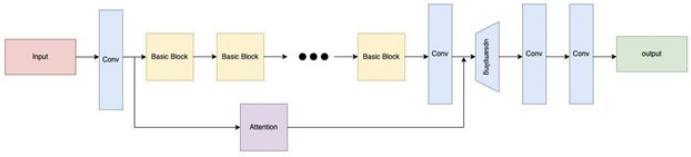


Figure 6: Architecture of Super Resolution model

Initially, we multiplied the residual blocks of ESRGANs with 0.2, as stated in the ESRGAN. But this resulted in a high perceptual loss of around 85%, making the output completely black. Therefore, our modifications are as follows, resulting in a reduced perceptual loss of only around 12%. The list of changes that we came up with when compared to the original ESRGAN are;

- Removed residual scaling of 0.2
- Removed residuals in the residual block and replaced it with a simple feed-forward block. We've used the below as our residual block. There is a total of 16 such residual blocks in the generator.

The Generator architecture including the attention could be seen "HERE" and the Discriminator architecture could be seen "HERE".

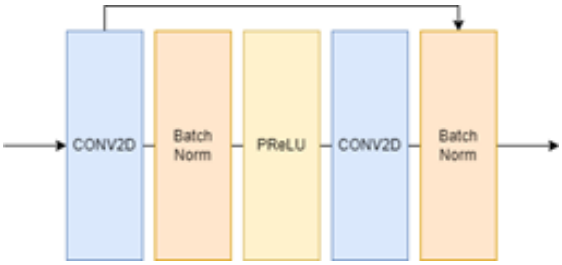


Figure 7: Modified residual block

In Figure 7, we can see the updated basic block based on the above-mentioned points.

For the discriminator, we have used a perpetual loss function to better train the model as stated in ESRGAN (3) paper. For this vgg19 architecture was used with "ImageNet" weights to extract features from images. These features are used by SoftMax at the end of the discriminator that predicts whether the generated image and true image are the same or not. The Discriminator architecture could be visualized in Figure 8.

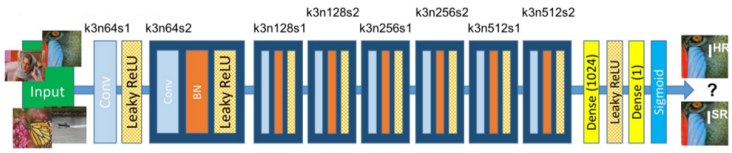


Figure 8: Discriminator Architecture

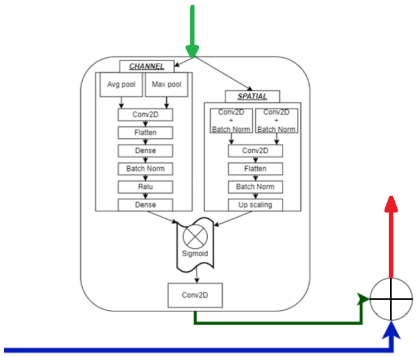


Figure 9: Attention Architecture

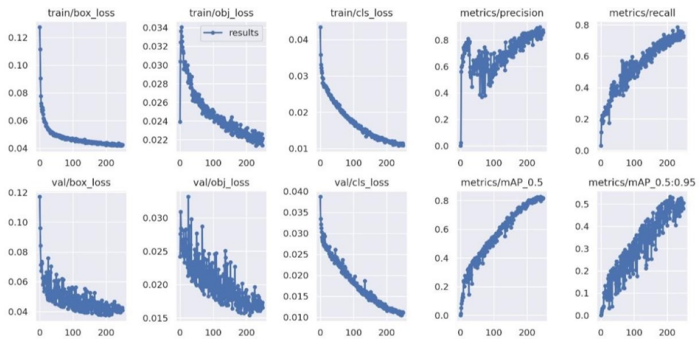


Figure 10: YOLOv5 training graphs for super-resolution images

4.3. Attention Mechanism

The SRGAN (5) is referenced to improve the accuracy of the super-resolution model by using channel and spatial attention. For example, Channel attention adjusts the importance of different channels in the generated image, while spatial attention adjusts the importance of different spatial locations. These techniques are used to make the generated image more realistic by emphasizing important features in the input image. The architecture of the attention is described in Figure 9. In Figure 9, the green arrow on the top represents the data that we get from the immediate adjacent block, which in this case is from the Conv layer just before the Basic Blocks. The output from this attention is then combined with the data from the Conv layer after the last basic block and is given as the input to the upsampling block.

5. Results

In this section, we'll be discussing the various results that we've achieved from our GAN-based Super-Resolution model to detect small objects.

5.1. YOLOv5 Model Training & Testing

Out of 1000 super-resolved images, 70% are used for training the YOLOv5 model(9), and the remaining 30%, 20% is used for validation and 10% for testing. The graphs shown in Figure 10 show the trends in loss values for box_loss, object_loss, class_loss, mAP, and Recall for both training and validation data. We could clearly observe that the loss is decreasing in a way that is expected while training for boxes, objects, and classes. On the other hand, we can also see that performance metrics like precision and recall are increasing with epochs.

5.2. Super Resolution

Coming to the results of Super Resolution, we observe from Table 1 that the loss we get from only the ESRGAN is around 16% but when we combine attention to it, it drops to 5%. We can observe though, that the training time is almost the same with only a slight reduction. Looking at the generated super-resolution images in Figure 11 from both ESRGAN and ESRGAN+Attention, we observe that our model generates better images in terms of overall quality of image super-resolution.

Model	Perceptual Loss
original ESRGAN	~85%
modified ESRGAN	~16%
modified ESRGAN+Attention	~5%

Table 1: Perceptual loss comparison between only ESRGAN and ESRGAN+Attention

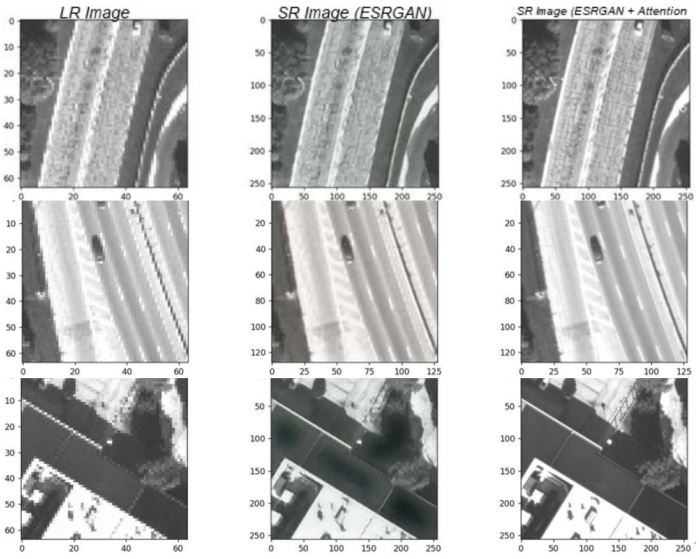


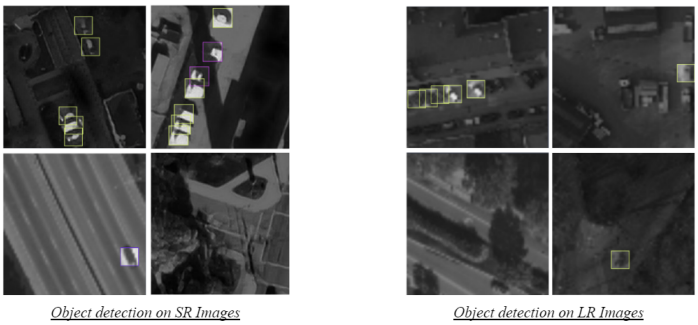
Figure 11: Super Resolution Image Comparison

Model	Training Time
ESRGAN	~1 hour
ESRGAN+Attention	~45 minutes

Table 2: Training time comparison between only ESRGAN and ESRGAN+Attention

5.3. object detection

Coming to the object detection results from Figure 12, we can see that the mAP score of yolov5, when using Super Resolution images is around 88%, and when low-resolution images are directly used, the maP is around 83%. We can also see that when using LR images directly, the detection quality is not that good where often a majority of the objects are not detected at all. Another interesting result we observe is that when using direct Low Resolution for object detection we get a lower mAP of 82.7% and when we use our super-resolution images to detect vehicles we get an mAP score of 88.4%. These results could be understood more clearly by Figure 13m Figure 14 and Figure 15. For comparison, we will consider two state-of-the-art small object detectors namely, PP-YOLOE-PLUS and Cascade R-CNN + NWD. The mAP50 score comparison can be seen in Table 3.



Object detection on SR Images *Object detection on LR Images*

Figure 12: Comparison of object detection with SR images and LR images

Model	mAP50
Our Model (ESRGAN+Attention)	88.4%
PP-YOLOE-Plus	85.7%
Cascade R-CNN + NWD	62.3%

Table 3: Comparison of mAP50 scores of our model, PP-YOLOE-Plus and Cascade R-CNN + New Wasserstein Distance

Figure 13: Precession for validation and test dataset using Low-Resolution images

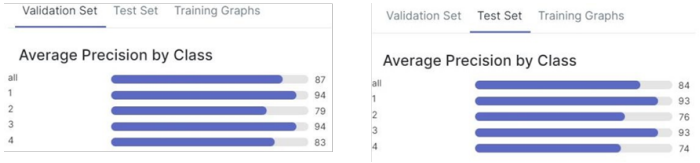


Figure 14: Precession for validation and test dataset using High-Resolution images

For Low-Resolution Images *For Super-Resolution images*

Figure 15: Accuracy, Precision metric comparison

6. Conclusion

We are proposing a GAN-based super-resolution step before object detection. Specifically, we propose to use the ESRGAN architecture to enhance the image resolution and a two-step attention mechanism (spatial and channel) to further improve the output. We prefer to use a GAN-based architecture for super-resolution instead of Deep CNNs because previous approaches have failed to produce finer details after enhancement. Finally, we will combine the ESRGAN super-resolution step with a two-step attention mechanism. Here we hypothesize the following when compared to other small object detection algorithms like PP-YOLOE-Plus, Cascade R-CNN + (Normalized Wassertian Distance), and TPH-YOLOv5;

- The additional Attention on ESRGAN will reduce the training time and improve the quality of super-resolution output.
- The enhanced super-resolution step will improve the overall object detection accuracy.

We validate the above assumptions based on the results shared above. Finally, we want to further improve the system by using other super-resolution methods such as Diffusion, to further improve the quality of super-resolution images. Along with this we also want to create an end-to-end architecture that does super-resolution and object detection together.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[2] Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. A normalized gaussian wasserstein distance for tiny object detection, 2022.

[3] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018.

[4] A. Liff. Mean square reconstruction error. *IEEE Transactions on Automatic Control*, 10(3):370–371, 1965.

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, 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. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017.

[6] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, and Baohua Lai. Pp-yoloe: An evolved version of yolo, 2022.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015.

[8] Zhicheng Zhao, Jiaqi Li, Ze Luo, Jian Li, and Can Chen. Remote sensing image scene classification based on an enhanced attention module. *IEEE Geoscience and Remote Sensing Letters*, 18(11):1926–1930, 2021.

[9] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios, 2021.

[10] Haoing Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models, 2021.

[11] Ming Zhao, Yang Wei, and Kevin K.L.Wong. A generative adversarial network technique for high-quality super-resolution reconstruction of cardiac magnetic resonance images, 2022.

[12] Georgios Tzimiropoulos Adrian Bulat. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans, 2018.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.