Generating Custom Datasets with Multi Generative Adversarial Networks

Donghee Lee
Department of Electrical Engineering, University of Ulsan
93 Daehak-ro, Nam-gu, Ulsan 44610, Korea

Byeongwoo Kim
Department of Electrical Engineering, University of Ulsan
93 Daehak-ro, Nam-gu, Ulsan 44610, Korea

ABSTRACT
Object detection and data collection from custom targets suffer from certain problems, which inherently occur in deep learning networks owing to problems such as difficulty of collection and data bias. Therefore, in this study, we proposed the Multi-GAN framework for generating augmented datasets. This framework comprises two parts: the first part generates data that reflect various textures related to decep learning based on deep convolutional GAN (DCGAN) and Wasserstein GAN (WGAN) structures. The second part provides multiple resolutions based on super-resolution GAN (SRGAN). Here, this paper presents efficient dataset construction methods along with a conventional augmentation method called manipulation technique. Through the experiments, which were based on average precision, conducted on the collected and augmented datasets, the proposed frameworkdemonstrated to improve detection accuracy. Additionally, we confirmed that the multi-GAN framework is superior with respect to efficiency to data collection.

General Terms
Computer Vision

Keywords
Generative Adversarial Network(GAN), Object detection, Data Augmentation, Deep learning, YOLO

1. INTRODUCTION
Convolutional neural network (CNN)-based deep learning structures first appeared in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in the image recognition field and improved performance to surpass human cognitive abilities based on deep neural networks and improved structures [1]. Recent research in the object detection field such as YOLO [2] and SSD [3] ensure sufficient accurate and real-time detection. To improve low detection accuracies for small targets and slow detection speeds owing to a high computation load, latest deep learning structures and methods aim to improve the overall performance of such applications.

Constructing datasets is essential for training artificial neural networks using deep learning methods, such as CNNs. Depending on the size and quality of the dataset, considerable effort is required to achieve a desirable detection outcome. For example, MS COCO dataset has millions of training datasets, while the ImageNet dataset contains 1.2 million images in its training dataset. It comprises 1000 widely used categories. Building such datasets is significantly costly and labor intensive. To recognize special or personal objects that are not included in the shared dataset, a separate dataset must be newly constructed. This process requires manually writing label files that contain categories such as object collection and object information, and location content.

Image augmentation is executed by processing original images via image manipulation techniques, such as rotation, reversal, and random noise, which is in line with the previous studies in image augmentation. CutOut [5], CutMix [6], and Mosaic [7] represent the latest data augmentation studies; however, these prior studies have limitations in performance improvements because of having adopted traditional methods. Additionally, they cannot be a fundamental solution to the data collection problems. Consequently, an augmentation framework that allows efficient augmented dataset construction is presented in this study. Furthermore, the limited reproducibility of the conventional augmentation method is addressed without being biased toward a small number of collected data by generating a new image that reflects the features of the collected image based on the deep learning generative adversarial network (GAN) technique from the image detection field.

The main contribution of this research can be summarized as follows:

To enhance the ability to detect insufficient data, a diverse texture images dataset was constructed by applying unconditional GANs; Deep Convolutional GAN (DCGAN) [28] and Wasserstein GAN (WGAN) [29].

A dataset that can be used to learn multi-scale images was constructed by generating SR images by applying super-resolution GAN (SRGAN) [31] to low-resolution images generated through an unconditional GAN scheme.

After conducting deep learning based on the YOLO object detection algorithm, we conducted accuracy verification of an average precision (AP). The dataset was evaluated the effect of proposed augmentation on object detection as AP(@0.5) and AP(@0.75) and confirmed a 22% performance improvement on average over the original datasets.

2. RELATED WORK
2.1 Image augmentation
Constructing large datasets is necessary for improving cognitive performance. For deep learning, numerous datasets can provide accuracy and stable detection rates. However, there are limits to acquiring specific targets online. Most global enterprises are investing whether to proceed with direct or platform-based data collection. Moreover, many recent studies in the data augmentation filed aim to reduce labor and costs.

Image manipulation has been widely used for image processing by accessing images in pixel units. It is used to reflect image transformations caused by weather, shooting angle, image quality, and movement of the camera. Through

image rotation and reversal manipulation techniques, the effect of conveying unexpected external factors may be remedied by applying noise to increase adaptation in various environments. This is done through image rotation, kernel filtering, and random noise that can convey the information of the target rotation caused by the change in the camera angle owing to the target conditions, movement of the object, or slope. Additionally, the color transformation for detection considers various colors, depending on the target characteristics. Adapting a conventionally used manipulation technique as a simple augmentation technique is mainly being studied in relation to data processing and synthesis. Recently, CutOut improved the performance by cutting off an arbitrary area of an existing image and filling it with zeros. Based on CutOut, CutMix presented a combination method that cuts off an arbitrary area and replaces it with another image [5][6]. Moreover, Mosaic improved the performance through generating a single image by combining four existing images at random ratios [7]. However, such augmentation techniques based on existing data could not provide significant improvements in cognitive and location accuracy and were relatively inefficient when applied to small amounts of datasets. Therefore, we proposed a novel framework based on deep learning GAN schemes to build an efficient dataset.

2.2 Object detection

Algorithms must be selected according to their deep learning applications based on established datasets. Particularly, there are two algorithms used in object detection: one-stage and two-stage algorithms. Typical one-stage algorithms are YOLO, SSD [8], RetinaNet [9], and EfficientNet [10]; two-stage algorithms are R-CNN [4], fast R-CNN [11], faster R-CNN [12], Mask R-FCNN [13], and RepPoints [14]. R-CNN notably comprises a stage for presenting thousands of proposals from images and a classification stage for detecting objects with classifiers, based on the first stage. Despite its high accuracy, R-CNN suffers from long processing times owing to the vast amount of computation required. Therefore, a one-stage algorithm base is mainly applied in systems with a rapid response demand, such as robots and vehicles. This is because rapid response and accuracy that corresponds to reasonable user requests are possible. Subsequently, current algorithm research aims to ensure real-time response and infer high performance. Among them, deep learning is conducted based on the YOLO algorithm series, which achieves good performance in speed and accuracy.

YOLOv3 & YOLOv3-tiny: YOLOv3 [15] achieved a relatively high speed and accuracy in performance, forming an efficient layer over other detectors while increasing the depth of the backbone used for feature extraction. Previously comprising 19 backbone layers, YOLOv3 is configured through a 53-layer hierarchical structure. Thus, YOLOv3 is slightly slower than YOLOv2 but it provides high accuracy. Furthermore, it comprises three detectors with different scales through the residual skip connection and up-sampling. This helps solve the low accuracy problem during small object detection in the previous version. The convolutional module consists of conv. 1×1 and five conv. 3×3 kernels. Additionally, considering low-performance embedded systems owing to environmental and cost limitations, a so-called YOLOv3-tiny, which efficiently improves speed and accuracy, consisting of 23 layers and a thin CNN has been trained and analyzed.

YOLOv4: While YOLOv4 [16] has no innovative structural changes compared to the previous version YOLOv3, it applies the latest deep learning techniques to optimize performance. In the activation function section, $f(x) = xtanh(\ln(1 + e^x))$ of Mish [17] was mainly adopted, and DropBlock [18] was applied in the regularization section with dropping out interrelated dense areas instead of randomly dropping out activations. In the data augmentation section, Mosaic and Self-Adversarial Training [19] were applied. The CSP module reconstructed the backbone based on multi-input weighted residual connections that were used in BiFPN and cross stage partial (CSP) connections to build CSPdarknet53 [20]. The architecture was constructed by adding SPP [21], PAN [22], and SAM [23] blocks into CSPdarknet53. Furthermore, the latest training methods, namely, CmBN [24], cosine annealing scheduler [25], DIoU [26], optimal hyperparameters, and random training shapes were utilized. This achieved similar speed and improved accuracy compared to YOLOv3.

3. MULTI-GAN FRAMEWORK OF CUSTOM DATASET

3.1 Overview

Prior augmentation methods for processed images, such as image manipulation, improve the accuracy of similar images during inference; however, this can cause data bias in small datasets. In addition, the augmentation method struggles with detecting unexpected targets. However, for deep learning-based object detection applications, the prediction of data that has never been encountered before through learning on specific datasets is important. Thus, an image augmentation framework based on an unconditional GAN structure has been proposed to alleviate the limitation in the number of existing images collected and cost burden of direct data collection.

First, GAN [27] was newly released in 2014 in the field of machine learning. A GAN comprises a generator that generates images and a discriminator that evaluates the generator performance in an adversarial position. It is a model designed to gradually improve the performance of each part through training and reach a result, at which the original and newly generated images can no longer be distinguished. For thediscriminator learning, the existing image sample x is set to D(x) = 1, and the sample made with input z drawn from a random noise distribution from the generator is set to D(G(z)) = 0. This is called a minimax problem, which is depicted in Equation (1).

$$min_G\, max_D\, V(D,G) = E_{x\sim p_{data}\,(x)}[logD(x)] + Ez\sim pz(z)[log(1-DG(z))] \quad (1)$$

The development of deep learning structures and learning methods based on the GAN concept, and their applications in replaces fully connected layers of a GAN with convolutional connected layers, and WGAN [29], which converts the existing similarity measurement methods based on Jensen-Shannon (JS) and Kullback-Leibler (KL) divergence to EM distance, facilitate learning and stabilization. They are simultaneously developed with CycleGAN [30], which offers a method of replacing the domain of a certain image with that of another image, and SRGAN [31], which generates high-resolution images in the field of resolution restoration.

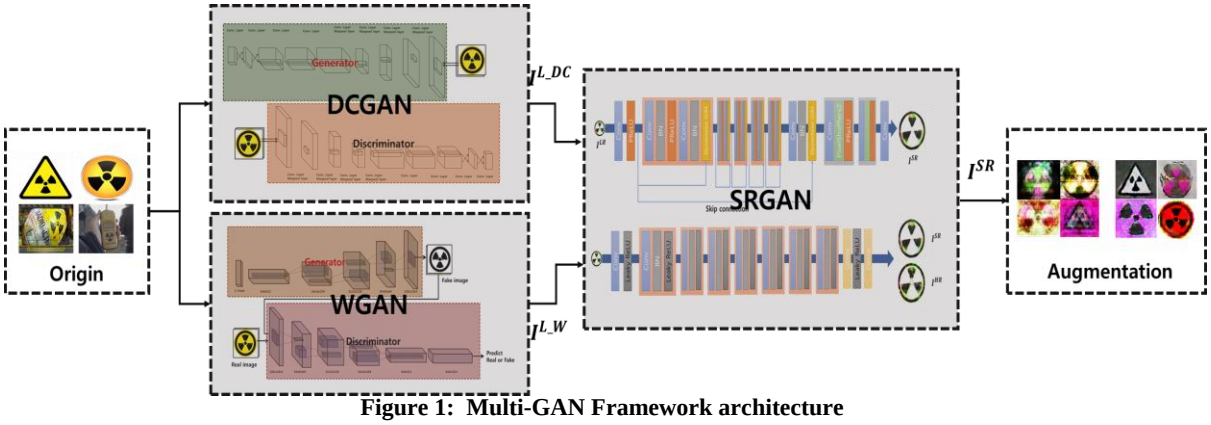

Figure 1: Multi-GAN Framework architecture

While deep learning structures and learning methods based on the GAN concept are integrated into various fields, their applications in the object detection field remain lacking. Therefore, to present specialized dataset construction methods in the object detection field, this study aims to integrate several GAN structures to perform data augmentation and build an unbiased and efficiently optimized dataset.

3.2 Framework Architecture

An augmentation framework was designed by fusing three GAN structures: DCGAN, WGAN (which are two unconditional GAN), and SRGAN (which is applied in high-resolution restoration). First, DCGAN and WGAN have been implemented to learn the features of the collected datasets. New images are generated after placing the trained generators in the front of the framework for dataset construction. Each structure adopted different learning methods and loss functions, which caused image generations to vary in texture and form. The detailed features of the structure are as follows.

DCGAN [28]: There was a serious issue of poor stability during the initial release of GAN training. The generated sample (using GAN) has no quantitative scale, making it difficult to evaluate the model performance. The DCGAN released in 2016 applied the CNN structure as opposed to the conventional a fully connected neural network structure of a GAN. Additionally, the following structure was adopted for a stable learning model. The pooling layers applied in the existing discriminator were replaced with strided convolutions, and the feature map size was expanded using fractional-strided convolutions in the generator. The fully connected hidden layer was also removed, and the generator active functions tanh and ReLU were used as reported in Equation (2). The discriminator used a leakyReLU function. The structure was optimized experimentally on various structures. Based on this structure, subsequent GAN papers have been published.

$$ReLU: f(x) = \begin{cases} x\ (x \geq 0) \\ 0\ (x < 0) \end{cases} \quad (2)$$

$$LeakyReLU: f(x) = \begin{cases} x\ (x \geq 0) \\ \alpha x\ (x < 0) \end{cases} \quad (3)$$

WGAN [29]: It is difficult to maintain the balance between the discriminator and generator in DCGAN applications. One major problem is the dropping occurrence after the completion of learning. This is mainly attributed to the discriminator failing to sufficiently discern and enable the model to learn its optimum point. To address these challenges, WGAN offers two main solutions. The discriminator conventionally uses a sigmoid function to distinguish between truth and false, resulting in a product with a prediction probability value. Instead, WGAN performs learning by measuring the distance between the probability distributions with Earth-Mover (EM) distance using the newly defined critic, as shown in Equation (4). Conventionally used JS and KL divergences measure the distance strictly, and consequently, the result often is not continuous. Thus, EM distance better delivers gradients and enables critics and generators to learn their optimal points. The EM distance aims to estimate the expected distance value between x and y, among the joint probability distribution of the two probability distributions, to be minimal. The EM distance, as the name suggests, indicates the cost of moving sand piles; it is the minimum cost of a pile being transferred from one distribution to another. The learning process is relatively stable in this case because it can produce results of linear features when continuous compared to the existing JS divergence method. The EM distance estimates the smallest expectation value between x and y among the joint probability distribution $\pi(P_r, P_g)$ of two probability distributions, as shown in the expression below.

$$W(P_r, P_g) = inf_{\gamma \to \pi(P_r, P_g)} E_{(x,y)}[\|x - y\|] \quad (4)$$

SRGAN [31]: The augmented image was generated by applying the DCGAN and WGAN to perform image augmentation based on the fixed input image of the same size. The image size for learning is expected to be large as the image must be in high resolution; thus, a longer generation time is required as the image size increases. SRGAN is used in this situation to additionally provide high-resolution augmented images to the object detection algorithm. SRGAN was created for super-resolution (SR) applications, which converts low-resolution images to high-resolution images. It was intended to compensate for the lack of high-frequency details, which is a problem described in previous SR research. Previously, PSNR indicators were used based on the MSE loss function to evaluate the SR method, as shown in Equations (5) and (6). PSRR represents the power of noise from the maximum power that a signal can have, which is mainly used in evaluating image quality loss information in video or image loss compression.

$$MSE = \frac{1}{XY} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} e(x,y)^2 \quad (5)$$

$$I_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \quad (6)$$

While the MSE loss is suitable for obtaining high performance from PSNR evaluation indicators, the minimization of MSE results in excessive smoothness or poor

visual effects. Therefore, a new loss function has been proposed and evaluated,shown in Equation (7), focusing on perceptual similarity. MOStesting indicated better performance in comparison to other techniques.

$$I^{SR}_{VGG\ /\ i.j} = \frac{1}{WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I^{HR}_{x,y} - G_{\theta_G}(I^{LR})_{x,y})^2 \quad (7)$$

Therefore, as shown in Figure 1, the images of various textures were generated based on DCGAN and WGAN. Subsequently, Multi-GAN framework, which generates high-resolution images by restoring SR images using SRGAN to provide multi-scale images, was proposed. Through the proposed framework, the images of various textures and images of multi-scale resolution are provided to build a dataset that is unbiased and has a size of various scales. Through experiments, we attempted to verify the performance of the efficient dataset with the proposed framework-based images and image manipulation.

4. EXPERIMENTAL RESULTS

4.1 Training Setup

Dataset: The images of radioactive display panels and valves for special environments were collected using data crawling for collecting online data. Additionally, data were acquired through video extraction and image conversion from online video-broadcasting platforms, such as YouTube. However, these data are limited and uniform. In the actual environment encountered during inference, with such data structures, it is difficult to detect targets that are observed in various aspects. During this process, we conducted the crawled image selection and target location labeling operations within the corresponding image. After completing the data collection and preprocessing steps, generating images through collected image manipulation and using the proposed framework to overcome the lack of data or data bias are recommended.

Repetition in learning was executed to generate images through deep learning, and as a result, detailed parameters of DCGAN and WGAN were set up as follows: Epochs were 1000, batch size was 64, learning rate of the generator was 0.00004, Adam initial decay was 0.5 and 0.999, and image size was 128×128. Figure 2 shows images generated with the generator optimized by learning, based on DCGAN (left) and WGAN (right). Observe that each GAN technique produced different image textures. To restore the low-resolution images to high-resolution, deep learning scheme with SRGAN proceeded based on the collected images to create a model, and a four-fold resolution restoration was conducted for all GAN augmented images.

Next, various image processing methods were applied through the image manipulation technique. To form a blurry image, kernel filtering was applied to allow input images to pass through a kernel using 11×11 pixel size. No larger size filters were applied owing to the high probability of eliminating the object properties. Random noise is used to provide special effects by adding random values to the pixels of each image. An image was generated by applying a Gaussian noise, which adds random values within a given range of the overall pixel value. Image generation was performed by arbitrarily controlling the color space, which transforms the brightness, saturation, and color of the image. Transforming features such as brightness, contrast, saturation, and hues enables building a robust dataset despite various target colors under different environmental conditions. Consequently, we also generated images by applications of manipulation like rotation and inversion.

Figure 2: Generated images using DCGAN

Figure 3: Generated images using WGAN

The total size of the dataset was 110,450 images, comprising 11,976 images collected through web crawling and image extraction from videos, 5208 images obtained through color transform, 1063 augmented images through DCGAN, 1390 augmented images through WGAN, 2453 images generated through SRGAN, and 93,568 images generated through image manipulation.

Object detection:Deep learning methods based on YOLOv4, YOLOv3, and YOLOv3-tiny are executed as follows using the built dataset. Learning was conducted individually for the three datasets: The first dataset composed of collected images; second composed of collected and GAN generated images; and third composed of collected images, GAN generated images, and images generated by manipulation. In addition, the hyper-parameter settings for deep learning were as follows: learning steps were 12,000; step decay learning rate was initially 0.01; 0.1 was multiplied at 80% and 90% learning step levels; momentum was 0.949; weight decay was 0.0005; batch size was 64, and mini batch size was set to 16 or 8. The constructed dataset was divided into training and testing datasets, and the training model was verified by randomly dividing the training and testing sets at a ratio of 9:1.

4.2 Evaluate accuracy

The evaluation indicator was analyzed based on AP, which is primarily used in the object detection field. mAP indicates the mean AP value of each class. We divided this into the augmented dataset and deep learning object detection types and evaluated the performance. AP(@0.5) indicates the accuracy at which the IoU value is greater than or equal to 0.5. AP(@0.75) indicates an accuracy greater than or equal to IoU 0.75. This not only indicates the accuracy of the object class but also the accuracy that considers the location accuracy within an image. he results extracted from training conducted through the YOLOv4 structure are reported in Table 1. The accuracy of the AP was reported for each dataset on the left.

Table 1. AP accuracy with YOLOv4

Dataset					Accuracy	
Origin	DCGAN	WGAN	SRGAN	Manipulations	AP(@0.5)	AP(@0.75)
√					63.21	42.36
√	√				66.26 (+3.05)	47.74 (+5.38)
√		√			66.50 (+3.29)	48.59 (+6.23)
√	√	√	√		69.95 (+6.74)	50.88 (+8.52)
√	√	√	√	√	87.59 (+24.38)	76 (+33.64)

Table 2. AP accuracy with YOLOv3

Dataset					Accuracy	
Origin	DCGAN	WGAN	SRGAN	Manipulations	AP(@0.5)	AP(@0.75)
√					61.75	37.25
√	√				63.83 (+2.08)	45.56 (+8.31)
√		√			64.18 (+2.43)	45.73 (+8.48)
√	√	√	√		68.06 (+6.31)	46.37 (+10.12)
√	√	√	√	√	85.31 (+23.56)	66.45 (+29.8)

AP(@0.5) has an accuracy of 63.21% when training was conducted with the existing online collected dataset. In comparison, when training was conducted with the dataset comprising the collected images and DCGAN and WGAN that underwent learning through color transformed images, the accuracy compared to the existing dataset improved by approximately 3% in DCGAN and WGAN, and 6% in the dataset including SRGAN. When the models that were trained with a dataset based on the proposed solution were compared, the accuracy was improved by 24.38%. The performance has improved in AP(@0.75) as well. The accuracy obtained when applying YOLOv3 is reported in Table 2. Based on the existing collected dataset, approximately 61% mAP(@0.5) was achieved, and the dataset generated based on DCGAN and WGAN resulted in a 2% accuracy improvement. In addition, the SRGAN-based dataset achieved a performance improvement of approximately 6%. Compared to the models that were trained with a dataset based on the proposed solution, an accuracy greater than or equal to 23% was obtained. A higher-performing cognitive accuracy was also confirmed based on mAP(@0.75). Overall, YOLOv3 tends to be less accurate compared to YOLOv4.

The augmented dataset proposed in this study was used with all three types ofobject detection architecture and went through performance verification. The model that was used with the three deep learning structures shows a relatively greater increase in recognition accuracy when manipulation is added rather than GAN-generated images. Such improvement in performance is because of a significant increase in the existing collected datasets to 9642, in the images generated by the GAN algorithm to 4906, in the images generated by manipulation to 93,568, and in the final augmented datasets to 117,515. Because the number of images generated from manipulation is relatively higher, of course the increase was higher.

As shown in Table 1, the dataset with image manipulation seems to have increased the accuracy rapidly. However, the images generated with GANs are not infinite images that can be generated in proportion to the number of origin images. The dataset generated with image manipulation can create tens to tens of thousands images while varying hyperparameters related to random rotation, color information, and noise. However, the generated image is very similar to the collected image. In this case, overfitting problems may occur in a small number of collected images. Therefore, we created the number of imageswith GAN framework and then created the images through image manipulation.

Based on YOLOv4, the growth rates per one GAN generation image and one manipulation image are respectively 0.0013% and 0.00018%. It can be seen that the accuracy of the GAN generation image per unit is improved by about 10 times more than that of the manipulated image per unit.the quality of the

Table 3. AP accuracy with YOLOv3-tiny

Dataset					Accuracy	
Origin	DCGAN	WGAN	SRGAN	Manipulations	AP(@0.5)	AP(@0.75)
√					50.28	12.91
√	√				54.74 (+4.66)	13.95 (+1.04)
√		√			53.49 (+3.21)	14.08 (+1.17)
√	√	√	√		56.97 (+6.69)	14.19 (+1.28)
√	√	√	√	√	67.96 (+17.68)	15.59 (+2.68)

GAN-generated image is better than that of the manipulated image. It means that the probability of recognizing a new type of target is higher.

Additionally, given the considerable time required for building datasets for specific object detection applications, significant efficiency is required to reduce the time necessary for dataset construction. It took half a day to collect images online through web crawling and extract images from videos. Labeling the indicated object information within an image required an entire day. When an optimal model is created by learning collected images using DCGAN, WGAN, and SRGAN deep learning structures, it takes two to seven days per one deep learning structure of one object. Depending on the GPU server performance, the learning period varies. Generating a new image with all GAN learning models that underwent such a learning process requires no longer than one day. Moreover, it was possible to generate images with certain manipulation techniques applied, such as image rotation and blur, within six hours. Given that years of effort was required to build the existing shared dataset, the dataset building process performed in this study may be useful in actual industrial environment applications because it required a maximum of one month and one person.

The accuracy obtained when applying YOLOv3-tiny is reported in Table 3. The training was executed after classifying dataset similar to YOLOv3 and YOLOv4. While the overall cognitive accuracy is relatively low because of the relatively shallow neural network structure, the growth rate for each dataset in AP(@0.5) increased, similar to previous cases. However, significant performance degradation was observed in the AP(@0.75). Its performance improvement was minimal despite the increase in the dataset size. This performance degradation may be because of structure simplification considering the speed aspect.

By analyzing the cause of improvement from the results given in the three tables above, we determined that the dataset of custom objects is not sufficient in size. Therefore, improving the accuracy with an increase in the relevant data is logical. However, depending on the quality of each data, the impact of accuracy differs. As evident in all tables, GAN generation techniques exhibit a relatively high improvement rate of accuracy per image compared to the image manipulation techniques. It can be evaluated that the quality of the generated image from the deep learning-based GAN technique is higher than that from the manipulation technique. The images generated based on the manipulation technique are obtained from the original image, resulting in the reproducibility being biased toward a small number of dataset because the images are largely influenced by the original image. Based on the experiment results, this limitation can be overcome by the GAN augmentation technique presented in this study.

5. CONCLUSION

In this study, the solution was proposed for the augmentation dataset construction that compensates for the deficiency evident in custom datasets. To complement the lacking dataset collection, a dataset was constructed via the multi-GAN augmentation framework composed of image manipulation (rotation, blur, random noise, etc.) and deep learningbased DCGAN, WGAN, and SRGAN schemes. Deep learning was executed for each dataset and object detection algorithm; the performance of the augmented dataset on the mAP basis was improved by an average of 22% based on three YOLO models, evaluated through the collected dataset and accuracy comparison. Consequently, reproducibility biased toward a small number of datasets, which is a limitation of the manipulation technique that was resolved by applying the multi-GAN augmentation framework proposed in this study. Furthermore, simultaneous application of DCGAN and WGAN enabled the generation of an image that reflects different textures, and a dataset that provides images of various scales was constructed through SRGAN. A specific method was presented for building an unbiased dataset by efficiently quantifying numerous images through various augmentation techniques, which reduces the time and costs required for the collection of images.

6. ACKNOWLEDGMENTS

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003) and the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) through the OEM demonstration cluster construction project to support autonomous vehicle parts suppliers. [Project Number: P0014572]

7. REFERENCES

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E.

Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. NIPS, 2012, pp. 1097-1105,

[2] Redmon, Joseph, Santosh Divvala, Ross Girshick and Ali Farhadi, "You only look once: Unified, real-time object detection," in Proc. CVPR, 2016, pp. 779-788, 10.1109/CVPR.2016.91.

[3] Wei Liu et al, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, Cham, 2016, pp. 21-37, 10.1007/978-3-319-46448-0_2.

[4] R. Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. CVPR, 2014, pp. 580-587, 10.1109/CVPR.2014.81.

[5] Terrance DeVries and Graham W Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, arXiv preprint arXiv:1708.04552.

[6] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Cheo and Youngjoon Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features," in Proc. ICCV, 2019, pp. 6023-6032, 10.1109/ICCV.2019.00612.

[7] Wang Hao and Song Zhili, "Improved Mosaic: Algorithms for more Complex Images," In: Journal of Physics: Conference Series. IOP Publishing, 2020, pp. 012094, 10.1088/1742-6596/1684/1/012094.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollár, "Focal loss for dense object detection," in Proc. ICCV, 2017, pp. 2980-2988, 10.1109/ICCV.2017.324.

[9] Mingxing Tan and Quoc V Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. the IEEE international conference on computer vision, 2017, pp. 2980-2988, arXiv:1905.11946.

[10] R. Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. CVPR, 2014, pp. 580-587, 10.1109/CVPR.2014.81.

[11] R. Girshick, "Fast R-CNN," in Proc. the IEEE conference on computer vision and pattern recognition, 2015, arXiv:1504.08083.

[12] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, 2015, pp. 91-99, 10.1109/TPAMI.2016.2577031.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick, "Mask R-CNN," in Proc. ICCV, 2017, pp. 2961-2069, arXiv:1703.06870.

[14] Saining Xie, Ross Girshick, Piotr Dollár and Kaiming He, "Aggregated residual transformations for deep neural networks," in Proc. CVPR, 2017, pp. 1492-1500, 10.1109/CVPR.2017.634.

[15] Joseph Redmon and Ali Farhadi, "YOLOv3: An incremental improvement," 2018, arXiv preprint arXiv:1804.02767.

[16] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, arXiv preprint arXiv:2004.10934.

[17] Diganta Misra, "Mish: A self regularized nonmonotonic neural activation function," 2019, arXiv preprint arXiv:1908.08681.

[18] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le, "DropBlock: A regularization method for convolutional networks," 2018, arXiv preprint arXiv:1810.12890.

[19] Chia-Jung Chou, Jui-Ting Chien and Hwann-Tzong Chen, "Self adversarial Training for Human Pose Estimation," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 17-30, arXiv:1707.02439.

[20] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh and I-Hau Yeh, "CSPNet: A new backbone that can enhance learning capability of cnn," in Proc. CVPR, 2020, pp. 390-391, 10.1109/CVPRW50498.2020.00203.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE transactions on pattern analysis and machine intelligence, 2015, 37.9: 1904-1916, 10.1109/TPAMI.2015.2389824.

[22] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi and Jiaya Jia, "Path aggregation network for instance segmentation," in Proc. CVPR, 2018, pp. 8759–8768, 10.1109/CVPR.2018.00913.

[23] Sanghyun Woo, Jongchan Park , Joon-Young Lee and In So Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.

[24] Zhuliang Yao, Yue Cao, Shuxin Zheng, Gao Huang and Stephen Lin, "Cross-iteration batch normalization,"arXiv preprint arXiv:2002.05712, 2020, 10.1109/CVPR46437.2021.01215.

[25] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, arXiv preprint arXiv:1608.03983.

[26] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye and Dongwei Ren, "Distance-IoU Loss: Faster and better learning for bounding box regression," in Proc. the AAAI Conference on Artificial Intelligence, 2020, pp. 12993-13000, arXiv:1911.08287.

[27] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, "Generative adversarial nets," Advances in neural information processing systems, 2014, 27, https://doi.org/10.1145/3422622.

[28] Alec Radford, Luke Metz and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," 2015, arXiv preprint arXiv:1511.06434.

[29] Martin Arjovsky, Soumith Chintala and Leon Bottou, "Wasserstein GAN," 2017, arXiv preprint arXiv:1701.07875.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros, "Unpaired Image-to-Image Translation using

Cycle-Consistent Adversarial Networks," 2017, arXiv present arXiv:1703.10593.

[31] 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," in Proc. the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681-4690, 10.1109/CVPR.2017.19.