Downsampling GAN for Small Object Data Augmentation

Daniel Cores[1[0000−0002−5548−4837]], Víctor M. Brea[1[0000−0003−0078−0425]], Manuel Mucientes[1[0000−0003−1735−3585]], Lorenzo Seidenari[2[0000−0003−4816−0268]], and Alberto Del Bimbo[2[0000−0002−1052−8322]]

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
{daniel.cores,victor.brea,manuel.mucientes}@usc.es
Media Integration and Communication Center (MICC), University of Florence, Florence, Italy {lorenzo.seidenari,alberto.delbimbo}@unifi.it

Abstract. The limited visual information provided by small objects — under 32×32 pixels— makes small object detection a particularly challenging problem for current detectors. Moreover, standard datasets are biased towards large objects, limiting the variability of the training set for the small objects subset. Although new datasets specifically designed for small object detection have been recently released, the detection precision is still significantly lower than that of standard object detection. We propose a data augmentation method based on a Generative Adversarial Network (GAN) to increase the availability of small object samples at training time, boosting the performance of standard object detectors in this highly demanding subset. Our Downsampling GAN (DS-GAN) generates new small objects from larger ones, avoiding the unrealistic artifacts created by traditional resizing methods. The synthetically generated objects are inserted in the original dataset images in plausible positions without causing mismatches between foreground and background. The proposed method improves the $AP_s^{@[.5,.95]}$ and $AP_s^{@.5}$ of a standard object detector in the UVDT small subset by more than 4 and 10 points, respectively.

Keywords: object detection, GAN, data augmentation

1 Introduction

Object detection is a fundamental technique within computer vision, as identifying objects in images or videos is mandatory for image understanding. The accuracy of detectors has experienced a lot of progress year on year since the release of large training datasets and the continuous improvement of convolutional neural networks (CNNs) [7, 6].

Small object detection has emerged as a specific problem that has drawn the attention of the research community [13, 1, 11]. It plays a fundamental role in

many applications in which early detection is key, including self-driving cars or obstacle avoidance on unmanned aerial vehicles (UAVs). Also, solving problems such as satellite image analysis requires the identification of objects represented by just a few pixels on the input image. However, the detection precision of small objects remains a challenging problem, which makes current state-of-the-art models perform poorly in this field. Moreover, the small object subset remains underrepresented in standard public datasets such as MS COCO [14] or ImageNet [19], mainly focused on larger objects.

Previous work has proven the benefits of applying strong data augmentation to improve the precision of objects detectors [28]. Data augmentation techniques have also been extensively studied in the image classification field, achieving very promising results. Therefore, data augmentation has the potential of generally improving the object detection precision, specially in data-scarce scenarios. Thus, it may compensate for the lack of small object annotations in most use cases, avoiding the high costs of manually annotating new data.

The introduction of generative adversarial networks (GANs) [5], brings new opportunities for more robust data augmentation [9]. The adversarial training ensures that the generated images contain the same artifacts as those present in real world images. This is specially relevant for small object data augmentation, as traditional scaling methods produce unrealistic artifacts [2, 20].

For all these reasons, we define a new data augmentation framework based on GANs to insert synthetic small objects in existing video datasets to alleviate the fall in precision caused by the lack of objects. The hypothesis is that, synthetic small objects can be generated by a GAN taking as input a real larger object. The generator should create a small image, visually similar to the input, free of unrealistic artifacts that are typical from traditional re-scaling methods. The main contributions of this paper are:

- Downsampling GAN (DS-GAN), a generative adversarial network architecture that transforms high resolution images containing large objects into low-resolution images containing small objects.
- An insertion method able to place the synthetic generated objects into plausible positions of the video frames without causing mismatches between background and foreground.
- Extensive experiments on the public dataset UAVDT [4], analyzing the improvement of applying our data augmentation method in different data availability scenarios.

2 Related Work

Small object detection focuses on improving the detection precision of objects represented by just a few pixels, typically below 32×32 pixels [14]. Although the current trend in object detection is to design deeper models that can extract more semantic information [7], the limited visual information of small objects fades in the deeper layers. Therefore, specific solutions such as the Feature Pyramid Network (FPN) [13] or the Region Context Network (RCN) [1] are required in order

to replicate the same success achieved in general object detection. Moreover, the most popular datasets are unbalanced towards large objects [14, 19]. This issue is partially addressed by new datasets, specially those focused on videos recorded from UAVs onboard cameras including UAVDT [4] and VisDrone2019-VID [27]. Also, datasets specifically designed for small object detection evaluation have been released [1]. Despite efforts to develop new architectures and the compilation of more specific datasets, the detection precision achieved with small objects remains significantly behind of the results achieved with larger objects [16].

Data augmentation is commonly used to train more general models. Basic data augmentation techniques for computer vision usually comprise a series of simple concatenated transformations, such as image mirroring and object-centric cropping [21]. Following this basic approach, a straightforward solution to augment the number of small objects would be randomly inserting the objects in different positions or resizing large objects [11]. However, that does not increase object variability —the appearance of the object remains invariant— and the context may not be suitable for a specific object —e.g., a car in the sky. The former issue is addressed in AdaResampling [3] with a prior context map that helps to insert the objects according to their scale and position. Also, conventional resizing functions generate artifacts not present in real world images [2, 20]. As a more elaborated alternative, adversarial learning can generate realistic synthetic objects.

Adversarial learning consists of training two —or more— networks with contrasting objectives. A successful use case of adversarial network are GANs [5]. These models are composed of two networks that are trained in an adversarial process: the generator and the discriminator. The role of the generator is to generate fake images that fool the discriminator, while the discriminator is trained to differentiate synthetic from real images. Deep Convolutional GANs (DCGAN) [18] were popularized for the generation of synthetic images. Different variations of the original architecture have been proposed to solve a wide range of computer vision problems: image synthesis [25], image super-resolution [12], or image inpainting [24], among others.

PTGAN [22] addresses the classical problem of domain gap, transferring instances of persons from one dataset to another keeping the same size. CycleGAN [26] with additional constraints can be used without downsampling. PTGAN ignores the object positioning problem, not relating object appearance and the background in the target position. DetectorGAN [15] was proposed to perform image-to-image translation. This approach does not define an insertion method either, and it has not been tested for small objects.

Alternatively, we propose to augment the training set with synthetically generated small objects, downsampling large objects through GANs. Moreover, the proposed method includes a random component that allows it to generate multiple different synthetic small objects from a single large object. We also define an object insertion procedure, avoiding inconsistencies between the inserted object and the background in the target image.

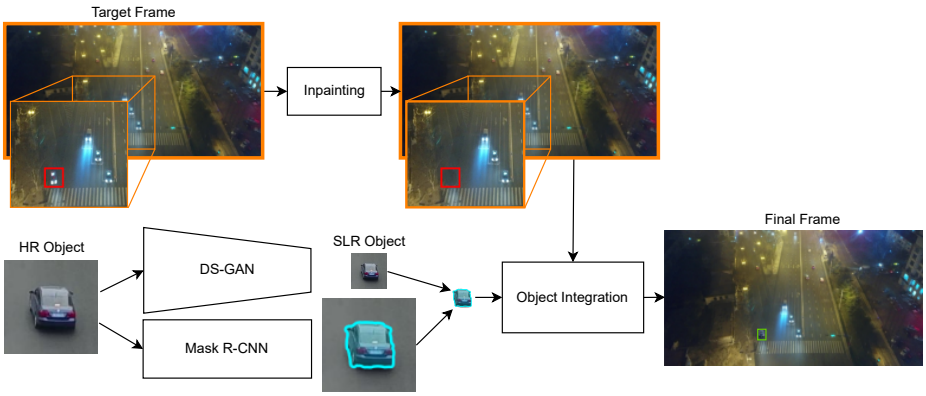

Fig. 1: Overall framework for small data augmentation. DS-GAN downsamples large HR objects, converting them into small SLR objects. The resulting SLR object is inserted on the target frame.

3 Method

The proposed data augmentation framework for small objects in video datasets is described in Figure 1. The pipeline consists of two fundamental stages: (i) small object generation and (ii) small object integration. In the first stage, real high-resolution (HR) objects and their context are transformed into synthetic low-resolution (SLR) objects. A segmentation mask is also calculated to precisely remove the context from the generated SLR image. As small object segmentation is a very challenging task, we propose to calculate this segmentation mask in the original HR image, and then scale it down to the size of the SLR object (Figure 1). Finally, SLR objects are inserted in positions where a real low-resolution (LR) object exists in the current input frame, or in previous or following frames. The position selector compares the direction and shape of the original HR object and the LR objects to select the optimal position for the corresponding SLR object. This method ensures that the background is adequate for the insertion of the new object. If necessary, an object inpainting method removes the object that will be replaced, as shown in Figure 1. The final augmented dataset facilitates the training stage of an object detector, improving the small object detection precision.

3.1 Downsampling GAN

A simple bilinear interpolation or nearest neighbor method suffices to downsample the original image containing a real HR object by a factor r. The output image contains an SLR object that can be inserted in new valid positions, augmenting the original dataset. However, as the experiments in Section 4 show, this naive approach creates artifacts that make the output image not suitable for data augmentation. Therefore, we propose a generative adversarial network

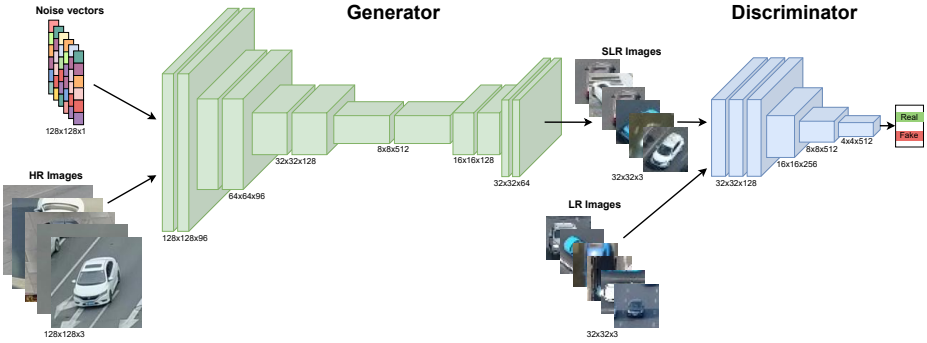

Fig. 2: DS-GAN architecture overview.

for image downsampling (DS-GAN) that transforms HR objects into SLR objects. The adversarial loss ensures that the generated SLR object has the same feature distribution as actual LR objects from the original dataset. Thus, the objective of the generator is to fool the discriminator, making SLR objects indistinguishable from LR objects.

The main challenge of designing a generative model for image downsampling in this context is the lack of the corresponding LR object for each HR object, making it an unpaired problem. The model must generate SLR objects similar to the corresponding HR input and follow the same feature distribution as real LR objects. DS-GAN receives images with size $W \times H \times C$ containing an HR object and produces images with size $\frac{W}{r} \times \frac{H}{r} \times C$ containing an SLR object that meets these requirements. Thus, the training set for DS-GAN contains real large (HR) objects and real small (LR) objects.

Figure 2 describes the architecture of DS-GAN in detail. This architecture consists of two main components: a generator and a discriminator network. The generator is implemented following an encoder-decoder model with six groups, each group with two residual blocks of the same dimension with pre-activation and batch normalization [8]. This network downscales the input image by a factor of 4, achieved by applying a pooling layer with a 2×2 kernel after each of the first four groups and a 2× up-sample deconvolution layer after each of the last two groups. The discriminator is also composed of the same residual blocks as the generator —without batch normalization— followed by a fully connected layer and a sigmoid function. In the discriminator case, there are six residual blocks with two downsampling pooling layers after each of the last two blocks.

The generator network (G) receives as input an image with an HR object and a noise vector (z), and the output is an image 4× smaller than the input ($r = 4$) containing an SLR object. The noise vector z follows a normal distribution and includes a random component to the input that allows the generation of multiple SLR objects from the same HR object. Both, the generator network G and the discriminator network D are alternatively optimized following the methodology proposed in [5]. As the goal of G is to generate SLR objects based on

the appearance of the HR input object, the objective function for the adversarial loss is defined as hinge loss [17]:

$$l_{adv}^{D} = \mathop{\mathbb{E}}_{s\sim\mathbb{P}_{LR}}[\min(0, 1 - D(s))] + \mathop{\mathbb{E}}_{\hat{s}\sim\mathbb{P}_{G}}[\min(0, 1 + D(\hat{s}))], \tag{1}$$

where $\mathbb{P}_{LR}$ is the LR subset distribution and $\mathbb{P}_G$ is the generator distribution to be learned through the alternative optimization. $\mathbb{P}_G$ is defined by $\hat{s} = G(b, z) \mid b \in \mathbb{P}_{HR}$, where $\mathbb{P}_{HR}$ is the HR subset. This equation defines a training goal for G that consists of generating SLR images that are hard to distinguish from LR images by D. Hence, the resulting images of the generator are suitable for data augmentation as D —that was trained to differentiate SLR images from LR images— cannot identify any pattern in the synthetic generated objects.

The loss function $\mathcal{L}$ for G is defined as:

$$\mathcal{L} = l_{pixel} + \lambda l_{adv}^{G}, \tag{2}$$

where l_{adv}^{G} is the adversarial loss, l_{pixel} is the L_2 pixel loss, and λ is a hyperparameter that balances the influence of each component in the final loss.

The adversarial loss l_{adv}^{G} is defined on the basis of the probabilities of the discriminator as:

$$l_{adv}^{G} = - \mathop{\mathbb{E}}_{b\sim\mathbb{P}_{HR}}[D(G(b, z))], \tag{3}$$

where $\mathbb{P}_{HR}$ is the HR subset and z is the random noise vector. By including the LR subset to calculate the adversarial loss, we force the SLR objects to contain real-world artifacts. Thus, this adversarial loss is computed in an unpaired way.

The l_{pixel} loss is implemented as a L_2 distance between the input HR and the output SLR images:

$$l_{pixel} = \frac{r^2}{WH} \sum_{i=1}^{\frac{W}{r}} \sum_{j=1}^{\frac{H}{r}} (AvgP(b)_{i,j} - G(b, z)_{i,j}) \mid b \in \mathbb{P}_{HR}, \tag{4}$$

where W and H is the input HR size, r represents the downsampling factor and $AvgP$ is an average pooling function that transforms the HR input to the output $G(b, z)$ resolution. Different to the adversarial loss, l_{pixel} is calculated in a paired way between the SLR object and the corresponding HR object, downsampled by the average pooling. Adding this term to the loss calculation ensures that the appearance of the generated SLR image is similar to the original HR object. Finally, in addition, to solve the stabilization of the discriminator training we normalize its weights by the spectral normalization technique [17].

4 Experiments

In this section, we evaluate the benefits of augmenting the training set with synthetic small objects generated by DS-GAN. A state-of-the-art object detector is optimized with different training sets to assess the detection improvement on the small objects subset.

4.1 Experimental setting

We selected the car category of the UAVDT dataset [4] to evaluate the performance of our system. This dataset provides 23,829 training frames and 16,580 test frames belonging to 30 and 20 videos respectively, with a resolution of $\approx$ 1,024 $\times$ 540 pixels. Following previous work [14], we consider small objects those with an area smaller than 32×32 pixels. Due to the high redundancy between consecutive video frames, the training set contains only 10% of the original frames.

For the construction of the HR subset, we include objects with an area between 48×48 and 128×128 pixels. To keep a fixed input dimension of 128×128, we include more context in smaller objects. As the generator of DS-GAN has a final stride 4$\times$, output downsampled images have a size of 32×32 pixels. The training HR subset for DS-GAN also includes annotations from the Visdrone dataset that meet the same requirements. This dataset, as well as the UVDT dataset, contains urban footage recorded from a UAV onboard camera. Therefore, images from both datasets are very similar. Overall, the HR subset for the DS-GAN training contains 5,731 objects while the LR subset contains 5,226 objects. The actual number of real LR objects is higher, but we simulate a data scarcity scenario by selecting only 25% of the available videos. The test set contains 316,055 car instances with 274,438 small objects.

For the evaluation of the object detector in the UAVDT test set, we use the standard Average Precision metrics defined by MS COCO. These include the $AP^{@.5}$, in which the overlap between an object detection and the corresponding ground truth must be greater than 0.5, and the $AP^{@[.5,.95]}$, which averages the AP for overlap thresholds from 0.5 to 0.95 with increments of 0.05. To effectively assess the improvement of the object detector applying the proposed data augmentation method, we report the AP_s, i.e., the AP for the small subset.

4.2 Implementation details

DS-GAN is trained for 1,000 epochs with an update ratio 1:1 between the discriminator and the generator. The optimizer is Adam [10] with $\beta_1 = 0$ and $\beta_2 = 0.9$ and an initial learning rate of 1e-4, with two reductions by a factor of 10 during the training phase. The hyperparameter λ in Equation 2 is set to 0.01 setting the influence of the adversarial loss l^G_{adv} two orders of magnitude higher than the pixel loss l_{pixel}. The training data is augmented by applying random image flipping and the noise vector (z) is randomly sampled from a normal distribution for each HR input image.

For image inpainting we apply DeepFill to remove the original object whenever it is necessary to insert the new SLR object. The DeepFill training process on the UAVDT dataset uses the hyperparameters defined by [23], setting $\tau = 40$. The selected object detector is the Faster R-CNN framework with a Feature Pyramid Network (FPN) [13] as it has proven to be robust against multiple scale objects, obtaining very competitive results in the small object subsets.

Data augmentation	$AP_s^{@.5}$	$AP_s^{@[.5,.95]}$
LR	39.0	17.6
LR + Interp.	38.1	16.5
LR + SLR	46.3	20.1
LR + SLR×6	50.9	22.5

Table 1: Results of FPN on the small object detection testing subset of UAVDT training only with 25% of the UAVDT training videos to simulate data scarcity for small objects.

4.3 Results

Table 1 reports the $AP_s^{@.5}$ and $AP_s^{@[.5,.95]}$ training the object detector with different training sets. The first row of the table (*LR*) represents the baseline, in which the detector is only trained with real objects extracted from the same 25% of videos as the DS-GAN training. Then, we conduct a series of experiments applying different data augmentation techniques. (*LR + Interp.*) expands the training set, duplicating the images and replacing the LR objects in those images with SLR objects. In this case, these SLR objects are generated by downsampling original HR objects through bilinear interpolation. Analogously, in the *LR + SLR* setting, LR objects are replaced in the duplicated images by SLR objects generated by DS-GAN. These two experiments include as many SLR objects as LR in the original training set. The last experiment (*LR + SLR*×n) explores the results of inserting n times the number of LR objects. Figure 3 shows a set of real HR objects and a set of SLR objects generated by DS-GAN.

(a) HR Objects: 128 × 128 pixel

(b) SLR Objects: 32 × 32 pixel

Fig. 3: Real large objects, input to DS-GAN (HR objects) and synthetic small objects generated by DS-GAN (SLR objects).

Results from Table 1 are in line with previous work [2, 20], proving that traditional re-scaling methods generate not visible artifacts that hinder the training process. On the other hand, DS-GAN produces useful extra training data that can be leveraged by the object detector to improve the detection precision. Duplicating the training set with SLR objects leads to an improvement of 7.3%

$AP_s^{@.5}$ and 2.5% $AP_s^{@[.5,.95]}$. Increasing the size of the training set up to 6× the number of objects leads to an improvement of 11.9% and 4.9% for $AP_s^{@.5}$ and $AP_s^{@[.5,.95]}$ respectively.

5 Conclusions

We have proposed a new generative model to augment the training set of a video dataset with synthetic generated small objects, i.e., objects under 32×32 pixels. This is crucial as the availability of small objects is limited in most object detection datasets. Contrary to small objects generated through interpolation methods, the output of DS-GAN is valuable to significantly improve the performance of an object detector by expanding the training set. DS-GAN is designed on the basis of state-of-the-art super-resolution techniques applied to generate low-resolution objects from high-resolution objects. The effectiveness of the proposed data augmentation pipeline is specially significant in data scarce scenarios, improving the detection precision of small objects by a large margin.

Acknowledgements

This research was partially funded by the Spanish Ministerio de Ciencia e Innovación (grant number PID2020-112623GB-I00), and the Galician Consellería de Cultura, Educación e Universidade (grant numbers ED431C 2018/29, ED431C 2021/048, ED431G 2019/04). These grants are co-funded by the European Regional Development Fund (ERDF). This paper was supported by European Union's Horizon 2020 research and innovation programme under grant number 951911 - AI4Media

References

1. Bosquet, B., Mucientes, M., Brea, V.M.: STDnet: Exploiting high resolution feature maps for small object detection. Eng. App. Artif. Intell. **91**, 103615 (2020)
2. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: ECCV. pp. 185–200 (2018)
3. Chen, C., Zhang, Y., Lv, Q., Wei, S., Wang, X., Sun, X., Dong, J.: RRNet: A hybrid detector for object detection in drone-captured images. In: ICCV. Workshops (2019)
4. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: ECCV. pp. 370–386 (2018)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. pp. 2961–2969 (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. pp. 630–645 (2016)
9. Ke, X., Zou, J., Niu, Y.: End-to-end automatic image annotation based on deep cnn and multi-label data augmentation. IEEE Trans. Multimedia **21**(8), 2093–2106 (2019). https://doi.org/10.1109/TMM.2019.2895511
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K.: Augmentation for small object detection. arXiv preprint arXiv:1902.07296 (2019)
12. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. pp. 4681–4690 (2017)
13. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
15. Liu, L., Muelly, M., Deng, J., Pfister, T., Li, L.: Generative modeling for small-data object detection. In: ICCV. pp. 6073–6081 (2019)
16. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. Int. J. Comput. Vis. **128**, 261–318 (2020)
17. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
18. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
20. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: CVPR. pp. 3118–3126 (2018)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
22. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR. pp. 79–88 (2018)
23. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR. pp. 5505–5514 (2018)
24. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV. pp. 4471–4480 (2019)
25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)
26. Zhu, J., Park, T., Isola, P., Efros, A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)
27. Zhu, P., et al.: VisDrone-VID2019: The vision meets drone object detection in video challenge results. In: IEEE Int. Conf. Comput. Vis. Workshops (2019)
28. Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V.: Learning data augmentation strategies for object detection. In: ECCV. pp. 556–583 (2020)