electronics

MDPI

Article

Value-Guided Adaptive Data Augmentation for Imbalanced Small Object Detection

Haipeng Wang [1,2], Chenhong Sui [3,*], Fuhao Jiang [3], Shuai Li [4], Hao Liu [5] and Ao Wang [3]

1 School of Computer Science, Harbin Institute of Technology, Harbin 150006, China
2 Institute of Information Fusion, Naval Aviation University, Yantai 246001, China
3 School of Physics and Electronic Information, Yantai University, Yantai 264005, China
4 School of Control Science and Engineering, Shandong University, Jinan 250061, China
5 The School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China
* Correspondence: sui6662015@ytu.edu.cn

Abstract: Data augmentation is considered a promising technique to resolve the imbalance of large and small objects. Unfortunately, most existing methods augment all small objects indiscriminately, regardless of their learnability and proportion. This tends to result in wasteful enlargement for many weak, low-information objects but under-augmentation for rare and learnable objects. To this end, we propose a value-guided adaptive data augmentation for scale- and proportion-imbalanced small object detection (ValCopy-Paste). Specifically, we first develop a non-learning object value criteria to determine whether one object should be expanded. Both scale-based learnability and quantity-based necessity are involved in this criteria. Then, the value distribution of objects in the dataset can be further constructed on the basis of the relevant object values. This helps to ensure that those uncommon, learnable objects that deserve enhancement are more likely to be enhanced. Additionally, we propose to enhance the data by pasting the sampled objects into relatively smooth portions of fresh background images, rather than arbitrary areas of any background images. This helps to boost data diversity while reducing the interference from complicated backgrounds. Evidently, our method does not require sophisticated training and just depends on the size and distribution of the objects in the dataset. Extensive experiments on MS COCO 2017 and PASCAL VOC 2012 demonstrate that our method achieves better performance than state-of-the-art methods.

Keywords: data augmentation; small object; imbalanced

Citation: Wang, H.; Sui, C.; Jiang, F.; Li, S.; Liu, H.; Wang, A. Value-Guided Adaptive Data Augmentation for Imbalanced Small Object Detection. *Electronics* **2024**, *13*, 1849. https://doi.org/10.3390/electronics13101849

Academic Editors: Stefanos Kollias and Manohar Das

Received: 21 September 2023
Revised: 5 May 2024
Accepted: 6 May 2024
Published: 9 May 2024

1. Introduction

As an essential step for traffic surveillance and maritime rescue, object detection has experienced tremendous progress [1–8]. This is not only due to the powerful representation ability of deep neural networks but it is also reliant on massive training data [9–12]. Unfortunately, most training data suffer from the heavily imbalanced ratio of large objects to small objects. Typically, there exhibits obvious quantity skewness in favor of large objects [13]. Then, small objects with inferior quantity proportion contribute less to the detection model, whereas the large ones with great quantity advantage are dominant [14,15]. As a result, the learned detection model is severely biased. It cannot perform well towards the small objects with low quantity proportion. However, in many practical tasks, detecting small objects with uneven distribution is a challenge that must be faced. For example, in terms of autonomous driving, the vehicles have to emphasize small objects to avoid traffic accidents. Respecting satellite and aerial remote sensing images, imbalanced small objects are a common presence that cannot be ignored. Therefore, small object detection has garnered increasing research interest and many efforts have resorted to improving the performance of imbalanced small object detection [16].

Concerning this, data augmentation is recognized as an extremely efficient technique [9,17–22] to improve the generalization ability of detection model. Data augmentation is less expensive and time-consuming than labeling larger-scale per-pixel annotations. A typical scheme is deep neural network-based augmentation [23], e.g., meta-learning [24–28], generative adversarial network [29–33], and reinforcement learning [34,35]. For example, the generative adversarial network is leveraged to produce useful pseudo-data in [15,30]. Regarding AutoAugment [36], reinforcement learning is employed to find the optimal policy combination for data augmentation. Despite their feasibility, they have to encounter the challenge of training stability and computational overhead [37].

In this connection, many efforts have been concentrated on augmentation methods involving no training, e.g., Cutout [38], Cutmix [39], Random Erasing [40], and Copy-Paste [41]. Regarding Random Erasing, it randomly selects a rectangle region in an image and erases its pixels with random values. In this connection, images with different levels of rectangle occlusion are yielded. Analogous to Random Erasing, Cutout aims at randomly masking out square regions of images during training. Obviously, both Random Erasing and Cutout help to reduce the risk of over-fitting. While, by overlaying a patch of either black pixels or random noise, discriminative information can be lost. Regarding this, Cutmix proposes to cut and paste patches of training images. Specifically, the ground truth labels are also mixed according to the area of the patches. Evidently, the amount and diversity of training images can be guaranteed. In terms of the above-mentioned methods, Copy-Paste augmentation is more advantageous as it is instance-oriented. Specifically, it expands the object instance rather than the circumscribed rectangular area containing that object. Then, with various levels of expansion, the diversity of training samples is enhanced, and the quantity imbalance is alleviated to a certain extent. Clearly, this helps to improve the generalization performance of the detection model. It is worth noting, however, that current copy and pasting-based methods tend to augment all small objects indiscriminately, regardless of their learnability or intelligibility. Consequently, this could cause invalid expansion for weak objects containing extremely low amounts of information, but under-augmentation for the learnable ones.

To this end, we propose a tri-sampling-based object-relevant value-guided explainable copy-paste policy (ValCopy-Paste) for scale and quantity imbalanced small object detection. As tiny objects involve limited identifiable information, they are difficult to learn. Therefore, apart from the balance of the number of samples in the training set, the detection performance is closely related to the object size. Inspired by this, we take both the size and quantity of object instances into consideration and further develop a criterion (i.e., indicator) for object-specific value (i.e., significance) characterization. Within this criteria, both scale-based learnability and quantity-based necessity are involved. According to the acquired object-relevant value, an object-specific value distribution is further established. Then, guided by this distribution, we can select the instances to be copied via the sampling technique. Evidently, this encourages those learnable and scarce objects to be augmented with a greater probability. Instead, invalid augmentation is suppressed. Moreover, for the sake of filtering the interference from complex contexts, we paste the copied instances to the relatively uniform regions of new scene images. This is profitable to break the spurious association between objects and backgrounds and improve the robustness of the detection model. Experimental results on MS COCO and PASCAL VOC2017 datasets show that, compared with state-or-the-art, our method exhibits obvious superiority for small object detection. Our main contributions can be summarized as follows.

- Instead of extensive data-driven black-box training, we give a tri-sampling-based simple and explainable data augmentation framework for imbalanced small object detection. Specifically, we introduce the learnability and scarcity of data and formulate the quantity-based necessity and scale-based learnability to characterize object-relevant value without training. This is capable of reasonably reflecting which object instances need and deserve augmentation as well.

- Instead of extending all small object instances with equal probability, we leverage the distribution of the attained object value as guidance to sample out the objects to be augmented. Then, those valuable objects, which are learnable and scanty, will be expanded with high probability, and vice versa. Therefore, on the premise of ensuring the diversity of the expanded samples, invalid or unnecessary expansion was avoided.
- Aimed at averting the interference of extremely complex contexts, we paste the selected objects to the relatively uniform areas of new scene images. This considers both the diversity and low interference of contexts. In addition, after the objects are pasted onto the background, the spurious correlation between objects and the scenes is broken. This is beneficial to enhance the generalization and robustness. Experimental results demonstrate that compared with others, the proposed method exhibits obvious superiority.

2. Materials and Methods

2.1. Scale-Imbalanced Small Object Detection

For scale-imbalanced small object detection, one common strategy is to enhance the resolution of the feature maps through super-resolution techniques [42]. For example, a generative adversarial network is exploited to generate high-resolution feature maps for small object detections in [43]. Another popular strategy is to make full use of the feature maps from small objects. In general, this relies on effective network structures capable of fusing multi-scale features [44–47]. A typical example is the feature pyramid network (FPN) [44,48–50]. Additionally, to better integrate the high-level features with the low-level ones, various FPN-based variants are further developed [50]. Among them, NAS-FPN [51] is widely concerned since it can automatically learn the network connections. Despite the improvement of detection performance brought by network upgrades, more and more computing consumption is also an important problem that can not be ignored [52]. Therefore, it has attracted more attention to alleviate the problem of scale imbalance through simple data augmentation [53], so as to improve the detection accuracy for small objects.

2.2. Copy and Paste Based Data Augmentation

Another typical strategy to augment small objects is to use copy and paste [13]. Copy and paste is a simple and straightforward data augmentation technique to mitigate the scale imbalance [37]. For example, in [13], small objects are first oversampled and then pasted to any area that does not overlap with the objects in the original background image. Note, that the entire bounding box area containing the objects is expanded. As a result, the pseudo association between the objects and their contexts is susceptible to being learned, which severely affects the generalization ability of detectors. To this end, object instances are used for augmentation [54]. As the object instances are still pasted to the original background images, the diversity of scenes cannot be guaranteed. In this connection, Copy-Paste is developed, which advocates pasting the target into arbitrary background images [41]. Apparently, this is advantageous to avoid over-fitting and enhance the generalization of the detection model. Notably, since all object instances are augmented with the same probability, learnable and inadequate object instances may obtain insufficient augmentation. On the contrary, those who are difficult to learn or rich in quantity are excessively augmented. To address this issue, we propose an object-relevant value-guided copy-paste strategy (ValCopy-Paste) for scale-imbalanced small object detection.

3. Our Method

3.1. Problem Formulation

Note, that the severely unbalanced ratio between large and small objects has become one of the primary challenges for object detection. It is sensible to augment more small objects to mitigate the bias of the detection model through data augmentation. Consequently, typical copy-paste-based methods are suggested to supplement data for generalization enhancement.

Unfortunately, the majority of known methods augment all objects uniformly, regardless of their quantity or learnability. In the case of extremely weak or plentiful objects, this can result in a tremendous wasteful or useless expansion for objects that are exceedingly weak or quite abundant. Additionally, it is well known that deep learning has the trait of shortcut learning [55]. In other words, the model learns task-irrelevant shortcut features, which often exist in the training sets, but once the test set is out-of-distribution (OOD), the robustness of the model will be greatly reduced. For example, in the task of cow recognition, the model can learn some connection between "grass" and "cows" and then recognize the cows through "grass". Consequently, when the cows are moved to the beach, the model fails to recognize them. This is because the model recognizes the cows via the "grass". In this case, the "grass" is a shortcut for cows. As a result, due to the shortcut learning issue, the deep detection model may become heavily focused on background unrelated to the objects, drastically reducing the model's robustness. Therefore, to prevent learning of the fictitious semantic information of the objects, different background images should be utilized. Moreover, when the expanded small objects are placed in the complex area of the background images, the background texture easily interferes with the model's learning for the augmented small objects, as depicted in Figure 1.

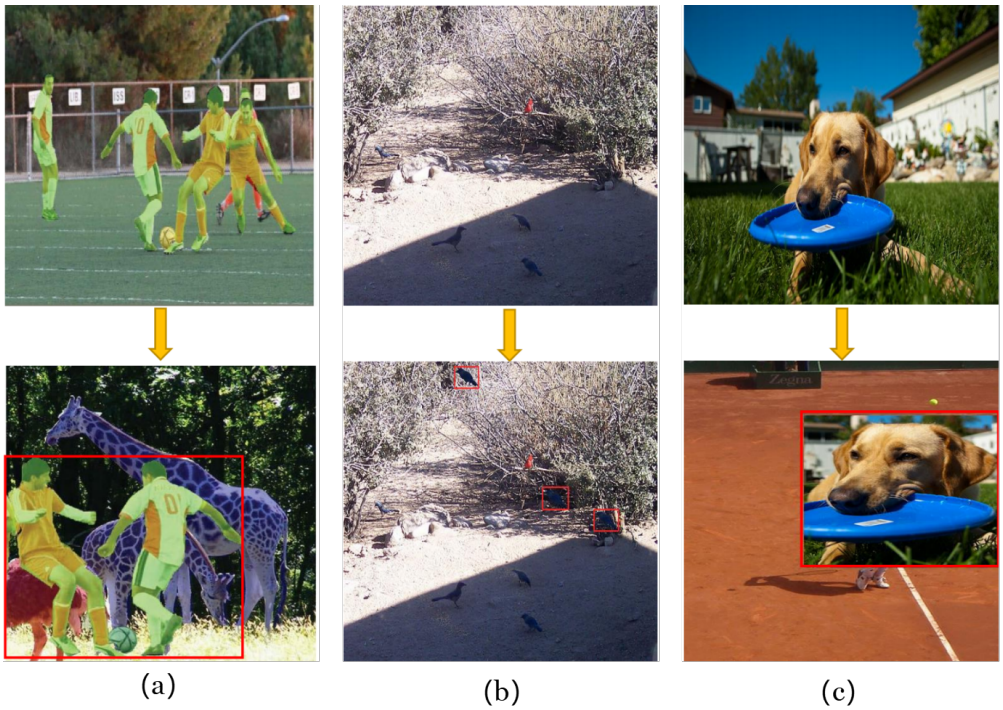


Figure 1. Display of Enhancement Results of Three Typical Copy-Paste Related Methods: (**a**) Cutout pastes the randomly cut patches to any areas of random background; (**b**) Augsmall pastes the randomly selected small objects to any areas of original background; (**c**) CopyPaste pastes the randomly selected objects to any areas of random background.

Motivated by this, this paper inherits the advantages of copy-paste-based methods and further develops an object-relevant value-guided adaptive augmentation method (ValCopy-Paste) (Algorithm 1). Respecting small objects, both the scale and quantity are exploited to measure the learnability and necessity for augmentation. For the background, the randomness and local smoothness are considered. Figure 2 gives the visual augmented data by our method. Table 1 lists the main notations of involved variables.

Algorithm 1 ValCopy-Paste Algorithm

1: Compute the object value based on Equation (1)
2: Calculate the object value distribution based on Equation (2)
3: Obtain the object instances to be augmented through sampling
4: Obtain the uniform areas of each image in the training set based on their variances
5: Acquire the background images from the training set via randomly sampling
6: Select the smooth areas of background images via randomly sampling
7: Generate the augmented images by putting the selected small object instances to the smooth regions of the background images
8: **return** augmented images

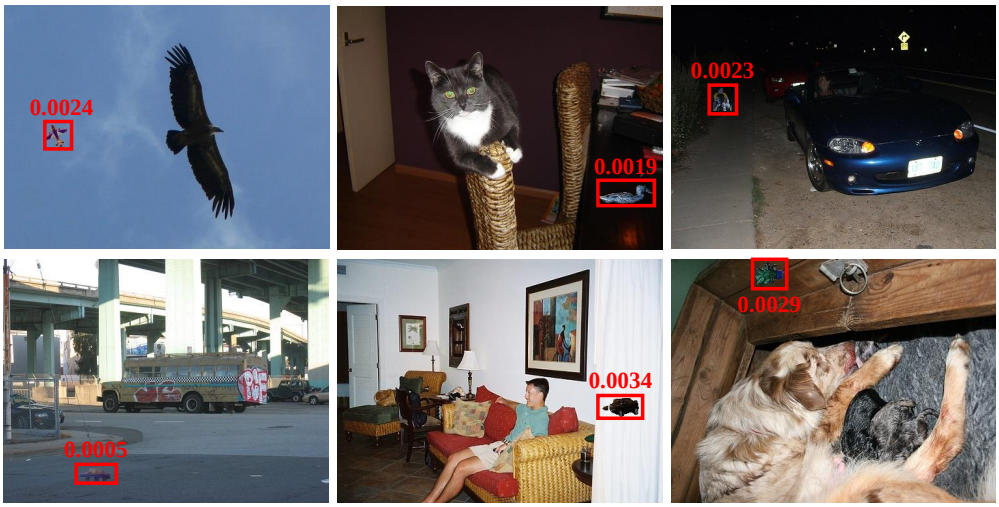


Figure 2. Examples of augmented data by our method. The instances in the rectangle are the augmented small objects through our method, whereas the values above denote the objects's values.

Table 1. Notations.

N	Total number of small object instances
M	Total number of background images
o_i	The ith small object instance
$\boldsymbol{B_j}$	The jth background image
$\boldsymbol{B_j^k}$	The kth flat region from $\boldsymbol{B_j}$
V_i	The value of o_i
N_i	Total number of small objects in the same class as o_i
$\boldsymbol{a_i}$	Area occupied by the bounding box of o_i

3.2. Establishment of Object Value Criteria

It is worth noting that object instances with an area less than 26×26 are tiny. They have extremely low information and are hard to identify. On the contrary, objects with an area greater than 96×96 are large. Between 26×26 and 26×26, the object with an area around 64×64 is a medium one. Concerning this, we divide the object scale into four intervals, which are $(0, 26)$, $[26, 64)$, $[64, 96)$, $[96, \infty)$, respectively. Here, the scale represents the root mean square of the area occupied by the bounding box. For example, assume o_i is the ith small object instance, $\boldsymbol{a_i}$ is the area owned by the bounding box of o_i, then the scale of o_i is defined as $\sqrt{\boldsymbol{a_i}}$.

In general, the larger the size, the more learnable the objects. For small objects with a scale of $(0, 26)$, as quite limited information is encompassed, they are usually difficult to learn, let alone detect. Therefore, it may be futile to enhance small objects with a scale ranging in $(0, 26)$. Specifically, for ease of averting useless augmentation, we do not augment small-scale objects with a scale ranging in $(0, 26)$. It can be observed that objects with a scale greater than 96, have noticeable advantages in both quantity and detection

accuracy. This indicates that they do not need to be enhanced as well. Regarding $(64, 96]$, there are relatively sufficient training data. This means that augmentation for $(0, 26]$ is usually futile, while for $(64, 96]$, enhancement is not so necessary. Inspired by this fact, the object scale interval for enhancement is set as $(26, 64)$ in our method. Moreover, owing to the strong correlation between learnability and instance size, we introduce scale-based learnability to identify whether an object deserves augmentation.

Additionally, due to the noticeably proportional advantage of large objects, they tend to play a leading role in training the detection models. Consequently, the learned model is biased and is inclined to detect large objects effectively rather than small ones. In other words, when the number of small objects in a scene is more than that of the large objects, small objects will likely take the dominant position instead of large objects, which will lead to a decrease in the accuracy of the model. From this perspective, it can be inferred that the object detection performance is closely related to the number of training data as well, as shown in Figure 3. This suggests that expanding the small objects with a low proportion is more necessary and meaningful, and vice versa.

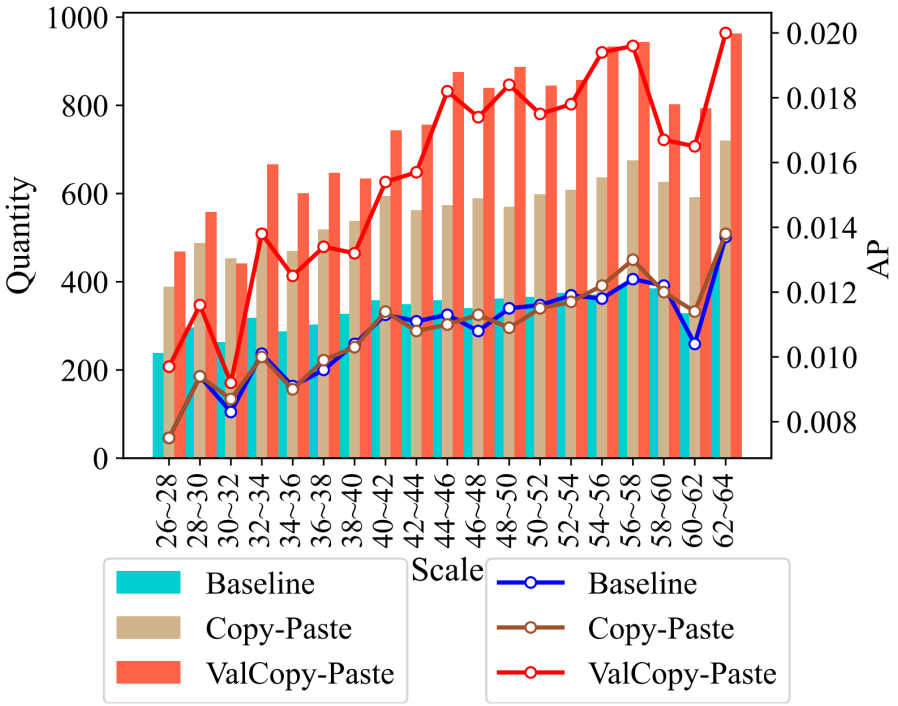


Figure 3. Comparisons of quantity and detection performance for small objects (i.e., size ranging from [26 to 64) before and after augmentation on the PASCAL VOC 2012 dataset. The bar graph shows the number of objects with different sizes, while the curve depicts their corresponding detection accuracy. Notably, Copy-Paste randomly increases the objects of each scale indiscriminately. Despite that the number of small objects is increased, the detection accuracy of small objects does not change significantly. For ValCopy-Paste, it adopts object-relevant value-guided sampling to ensure that learnable objects are augmented. Thus, the detection accuracy is significantly improved with the increase in the number of small objects in the training set.

Enlightened by this, we first exploit the quantity-based necessity to evaluate whether an object instance needs augmentation. Then, by integrating both scale-based learnability and quantity-based necessity, we establish object value criteria for data augmentation. Specifically, the criteria V_i corresponding to o_i can be formulated by

$$V_i = \begin{cases} e^{\frac{\sqrt{a_i}}{N_i}}, & \sqrt{a_i} \in [26, 64) \\ 0, & otherwise. \end{cases} \quad (1)$$

where N_i represents the total number of small objects in the same class as o_i.

There are two main reasons for the establishment of the criteria V_i. First, it is worth noting that the larger the object instance a_i, the more informative and learnable. Regarding this, the corresponding value v_i should increase with the increase in area a_i. Second, for the object instance with an area of less than 26, it is found that they are difficult to present visual representations for target recognition, resulting in a sharp decrease in the learnability of detectable models, which can even be ignored. Therefore, the value v_i of a_i less than 26 is set to 0. Taking into account the above two reasons we construct the value criteria as Equation (1), which can adaptively adjust with the size and number of the objects.

Apparently, V_i is a comprehensive index, which not only considers whether the instance is worth expanding but also whether it is necessary to expand. Equation (1) reflects that with a larger object value, the corresponding object is more worthy of augmentation from the perspective of both quantity and intelligibility. Then, whether an object instance is worth augmenting is directly proportional to its instance value. Therefore, the object-relevant value can be directly used to supervise the data augmentation.

3.3. Learnability and Scarcity Based Object Value Distribution

To enhance the diversity and randomness of augmented data, we do not directly utilize the object-relevant value. Instead, we further establish the value distribution. Based on Equation (1), the value distribution of objects can be attained through the maximum normalization. Concretely, the value probability corresponding to o_i is defined as

$$P_i = \frac{V_i}{\max(V)}, \tag{2}$$

where $\boldsymbol{V} = [V_1, V_2, \ldots, V_N]^T \in \mathbb{R}^{N\times 1}$, N denotes the total number of small object instances.

Based on the attained object instance value distribution, we can sample out objects to be copied. To be specific, due to the efficiency and effectiveness of the Walker–Vose alias method [56], it is utilized to attain the objects in this paper. Subsequently, the attained objects are copied for pasting in the next subsection.

3.4. Tri-Sampling-Based Generation of Augmented Training Images

To generate the augmented training images, we introduce a tri-sampling-based method. First, C instances are randomly selected from the training set based on their value distribution and then copied. This enables scarce and worth learning small goals to be enhanced with a greater probability. Second, background images are randomly selected from the training set. This is helpful to guarantee the diversity of backgrounds and avoid overfitting the detection model. Second, Third, uniform areas are randomly selected from the uniform area pools from the chosen background images. This can effectively avoid interference from complex backgrounds, thereby enabling the network to better learn the characteristics of small objects. Specifically, note that for background images extremely complex contexts cause serious interference to small objects, which makes the detection model hardly focus on the pasted small objects. This indicates that for the posted location, a too-complex context is not a reasonable choice. Otherwise, even after data augmentation, it is still difficult to achieve satisfactory detection performance for small objects. To alleviate this issue, it is sensible to force the instances to paste onto any relatively smooth regions.

Based on the above analysis, we will detail how to paste small object instances into relatively smooth areas. As variance is simple and easy to obtain, we adopt it to measure the degree of regional smoothness in this paper. It is worth noting that on the one hand, too large a threshold makes small object instances pasted onto complex scenes, which is unfavorable to small object learning and detection. On the other hand, too small a threshold can affect the diversity of the contexts, which can lead to overfitting of the model. In this connection, the relatively smooth areas obtain relatively flat regions in this paper. Then, C instances are selected and put into arbitrary uniform areas, which are from randomly chosen background images.

Figure 2 gives six instances of augmented data. In this figure, the objects in the red rectangle are expanded into small ones. Meanwhile, the value above denotes the value of the augmented small object.

4. Experiments

4.1. Dataset and Comparison Methods

To evaluate the effectiveness of our method, comparative experiments with five representative data augmentation methods, including Cutout [38], GridMask [57], the augmentation for small object detection [13] denoted by Augsmall, Cutmix [39], and CopyPaste [41] is conducted on PASCAL VOC 2012 [58] and MS COCO 2017 [59] datasets.

Following a conventional design, we use ∼118k images for training and 5k images for validation on the COCO2017 dataset. For VOC2012, 5717 and 5823 images were used for training and validation, respectively.

4.2. Baseline and Evaluation Metrics

Baseline: Note, that SSD [60] and Faster RCNN [61] are typical one-stage and two-stage object detection methods, respectively. Therefore, in this paper, we employ SSD and Faster RCNN as our baseline models with ResNet-50 [62] as the backbone network. Both of them adopt the official implementation version of PyTorch. To be specific, the implementation of Faster RCNN please refer to https://github.com/pytorch/vision/tree/main/torchvision/models/detection (accessed on 5 May 2024). The implementation of SSD please refer to https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/Detection/SSD (accessed on 5 May 2024). All models are first initialized using ImageNet pre-trained weights, and then finely tuned on PASCAL VOC 2012 and COCO2017 datasets. To be fair, we removed all data augmentation methods during the data load process.

Evaluation Metrics: To better assess the effectiveness of our method on small object detection under limited table width, we utilize Average Precision (AP) values involving object scale, i.e., AP^{IoU}_{scale}, instead of the overall detection performance. Here, IoU refers to the overlap rate between the generated candidate bound and the original ground truth bound. For simplicity, the IoU used in AP^{IoU}_{scale} is expressed as $IoU = IoU \times 100\%$.

4.3. Experimental Setting

On the PASCAL VOC 2012 dataset, we set the learning rate, batch size, and epoch to 0.0003, 16, and 50 for SSD, respectively. Regarding Faster RCNN, they are set as 0.003, 8, and 30, respectively. On the COCO2017 dataset, the learning rate, batch size, and epoch for SSD are set as 0.001, 16, and 50, respectively. In terms of Faster RCNN, they are set as 0.02, 8, and 30, respectively. During training, the SGD optimizer is utilized with the momentum set to 0.9 and weight decay set to 0.0005. Additionally, for both SSD and Faster RCNN, we adopt the popular warm-up strategy to adjust the learning rate (lr). Concretely, in the first four epochs, the learning rate increases to 10 times the initial.

4.4. Parameter Impact Analysis

C refers to the number of instances pasted to one background image. Then, the larger the C, the more small objects will be expanded. To investigate the impact of different *C*, relevant experiments are performed on VOC2012. Figure 4 plots the curve of detection accuracy changing with C. From Figure 4 we can see that when C is 1, both SSD and Faster RCNN achieve the highest AP. However, with the increase in C, there is an obvious downward trend. Tables 2 and 3 further list the relationship between C and AP obtained through Faster RCNN under different learning rates. It can be found that even for different learning rates, C = 1 always corresponds to the highest AP. This is consistent with the analysis of C in [13]. The main reason for this is that with the increase in C, the proportion of small objects will be significantly increased as well. This incurs that the trained model is

not mainly dominated by large objects, which leads to the deterioration of the detection accuracy towards large objects with a large proportion.

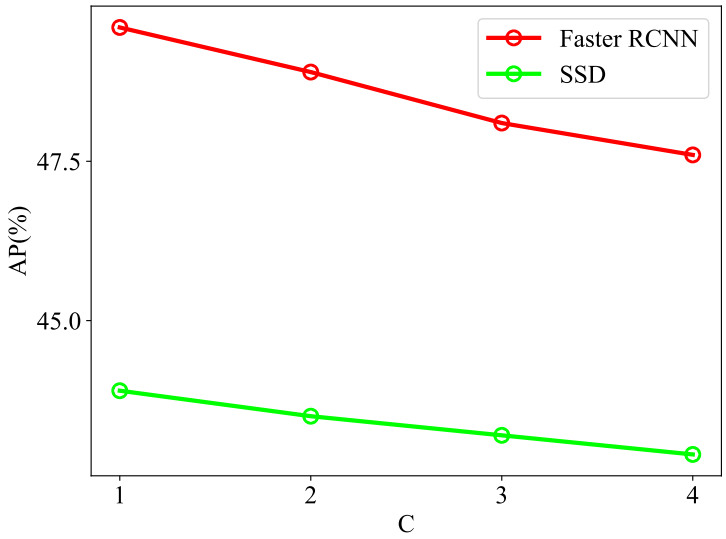
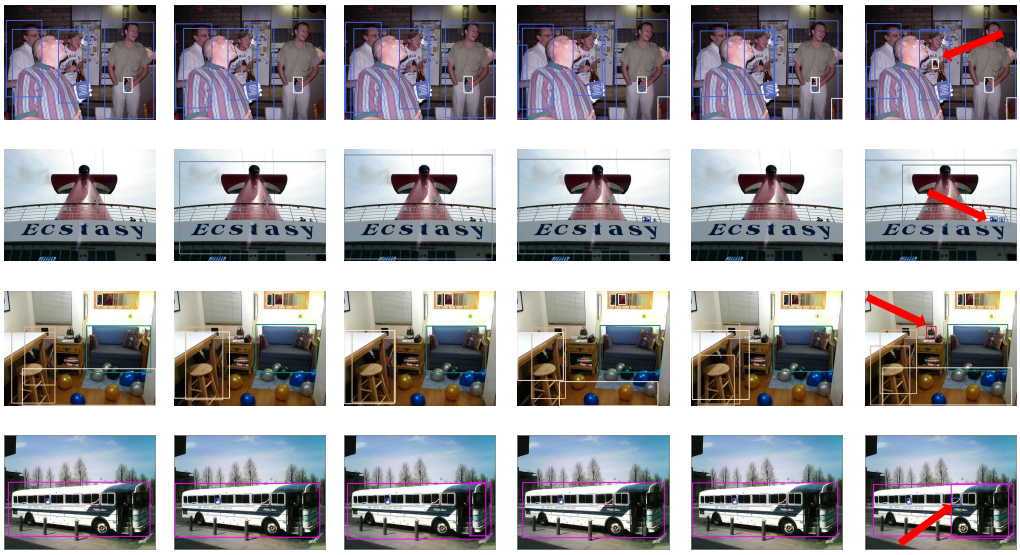


Figure 4. Curves of C vs. AP with both SSD and Faster RCNN on PASCAL VOC 2012 dataset. Evidently, the optimal choice of C is 1. The main reason for this is that the larger the C, the greater the change in data set distribution. Then, the learned model will not be dominated by large objects, which can lead to a decrease in the detection performance for large objects with high proportions.

In the first row of Figure 5, only our method successfully detects the small object, i.e., the cup, as pointed to by the red arrow in the last image. In the second row of Figure 5, the red arrow in the last image points to two persons with a small scale. Still, only our method successfully detects the two persons. Concerning other methods, GridMask detects one person, while others fail. In the third row of Figure 5, our method could find the pen holder in the middle of the image, but others failed.



(a) CopyPaste (b) Augsmall (c) Cutout (d) GridMask (e) Cutmix (f) Ours

Figure 5. Visualization of results for different methods on the PASCAL VOC 2012 dataset.

Table 2. SSD results based on detection results of different C and learning rate (lr) on PASCAL VOC 2012 dataset.

AP (%)		Lr						
		0.002	**0.003**	**0.004**	**0.005**	**0.006**	**0.007**	**0.008**
C	1	**43.6**	**43.9**	**43.6**	**43.2**	**43.2**	**42.6**	**42.4**
	2	43.5	43.5	43.2	42.9	42.3	42.1	41.7
	3	43.3	43.2	42.6	42.1	41.6	41.6	41.4
	4	42.9	42.9	42.5	41.9	41.4	41.2	40.7

Table 3. Faster-RCNN based detection results with different C and learning rate (lr) on PASCAL VOC 2012 dataset.

AP (%)		Lr						
		0.002	**0.003**	**0.004**	**0.005**	**0.006**	**0.007**	**0.008**
C	1	49.0	**49.6**	**49.3**	48.9	**49.5**	**49.6**	**49.2**
	2	**49.1**	48.9	48.5	**49.2**	49.3	48.8	48.6
	3	48.1	48.1	48.1	48.6	48.6	48.6	48.2
	4	47.8	47.6	48.3	48.0	47.6	47.4	46.5

4.5. Ablation Experiments

To investigate the effectiveness of each component, i.e., object value guided copy and paste to a relatively smooth area, ablation experiments are conducted. The experimental results are shown in Table 4.

Table 4. Ablation experimental result on PASCAL VOC 2012 dataset.

Strategy	$\mathbf{AP_S}$	$\mathbf{AP_M}$	$\mathbf{AP_L}$	$\mathbf{AP_{(0,26)}}$	$\mathbf{AP_{[26,64)}}$	$\mathbf{AP_{[64,96)}}$	$\mathbf{AP_{[96,\infty)}}$	$\mathbf{AP^{50}_{(0,26)}}$	$\mathbf{AP^{50}_{[26,64)}}$	$\mathbf{AP^{50}_{[64,96)}}$	$\mathbf{AP^{50}_{[96,\infty)}}$	$\mathbf{AP^{75}_{(0,26)}}$	$\mathbf{AP^{75}_{[26,64)}}$	$\mathbf{AP^{75}_{[64,96)}}$	$\mathbf{AP^{75}_{[96,\infty)}}$
SSD+Baseline	9.4	27.7	**51.1**	5.5	**22.8**	31.7	**51.1**	**13.7**	**42.4**	56.6	**79.9**	3.4	**23.0**	31.6	**56.7**
SSD+RI+FA	6.9	22.6	47.4	3.7	18.4	26.2	47.4	10.1	36.3	47.1	76.8	1.2	16.9	25.5	51.1
SSD+VGI+RA	7.0	22.4	46.0	4.7	18.4	25.9	46.0	11.3	36.0	48.2	75.2	3.2	15.8	24.0	48.9
SSD+VGI+FA (Ours)	**10.4**	**27.9**	50.9	**5.9**	22.6	**32.2**	50.9	13.5	42.3	**57.0**	**79.9**	**3.8**	21.4	**32.5**	55.9
Faster-RCNN+Baseline	20.9	40.1	**55.4**	**17.8**	35.6	43.8	**55.4**	**36.4**	60.3	70.8	**84.6**	14.4	36.2	47.2	62.1
Faster-RCNN+RI+FA	21.8	39.4	54.9	16.7	35.4	42.4	54.9	34.6	59.2	68.1	83.7	13.6	37.7	45.7	61.2
Faster-RCNN+VGI+RA	22.3	38.8	55.3	16.8	35.6	42.2	55.3	33.7	58.9	68.5	83.6	14.3	36.8	45.4	**62.0**
Faster-RCNN+VGI+FA (Ours)	**22.4**	**40.3**	53.9	17.5	**36.1**	**44.5**	53.9	36.3	**62.0**	**71.0**	83.5	**14.5**	**39.1**	**48.5**	59.9

In Table 4, "RI (RandomInstance)" refers to selecting the instances to be copied at random, whereas "VGI (ValueGuidedInstance)" represents the case that the instances are samples based on its value distribution. In addition, "RA (RandomAreas)" refers to where the copied object instances are pasted to arbitrary regions, while "FA (FlatAreas)" means that the copied instances are pasted to relatively flat regions. In the first three columns of Table 4, we show the accuracy results for S, M, and L objects divided by the previous method. The next few columns show that we re-partition the object interval according to the object size.

As demonstrated in Table 4, whether the SSD or Faster-RCNN strategy, our approach achieves the highest accuracy in small object detection (AP_S). However, the overall small object detection accuracy of the SSD method in small objects is not as good as that of the Faster-RCNN method. The reason is that SSD shows strong performance on large objects across feature extractors such as VGG, MobileNet, and ResNet, while the accuracy of small object detection is extremely low [63]. What is more, since we augment the small objects, the number of small objects in a picture increases which can cause the number of small objects to dominate, further compromising the accuracy of the model trained using the SSD

method. However, while we have augmentation for small objects, the loss of accuracy is very small, for example, for $AP^{50}_{[26,64)}$ we are only 0.1 below the baseline. Different from the SSD method, our method combined with the Faster-RCNN strategy can significantly improve the precision of small objects. For example, for $AP^{75}_{[26,64)}$, our detection accuracy is improved by 2.9, 1.4, and 2.3, respectively, compared with the other three methods.

4.6. Comparison

Comparative experimental results on PASCAL VOC 2012 dataset: To assess the effectiveness of our method on the PASCAL VOC 2012 dataset, comparative experimental results obtained by Faster RCNN are listed in Table 5. It is obvious that the five comparison methods perform absolutely differently under diverse detection metrics. Notably, our method always achieves the top 1 accuracy for the small objects with scale in $(26, 64]$ in terms of various metrics. This further indicates the effectiveness and robustness of our method.

Table 5. Result of object detection on the PASCAL VOC 2012 dataset.We achieve the best small object accuracy on Faster-RCNN-50-FPN.

Method	AP_S	AP_M	AP_L	$AP_{(0,26)}$	$AP_{[26,64)}$	$AP_{[64,96)}$	$AP_{[96,\infty)}$	$AP^{50}_{(0,26)}$	$AP^{50}_{[26,64)}$	$AP^{50}_{[64,96)}$	$AP^{50}_{[96,\infty)}$	$AP^{75}_{(0,26)}$	$AP^{75}_{[26,64)}$	$AP^{75}_{[64,96)}$	$AP^{75}_{[96,\infty)}$
Baseline [61]	20.9	40.1	**55.4**	**17.8**	35.6	43.8	**55.4**	**36.4**	60.3	70.8	**84.6**	14.4	36.2	47.2	62.1
None copy and paste families															
Cutout [38]	19.6	39.6	55.2	16.2	35.1	43.0	55.2	33.7	60.4	69.7	84.5	13.2	37.1	47.3	**62.2**
GridMask [57]	21.1	40.0	55.0	17.4	35.0	44.1	55.0	35.3	59.5	70.4	84.6	12.8	38.6	**48.8**	61.7
Cutmix [39]	19.9	39.4	53.8	17.1	34.1	43.7	53.8	34.6	58.5	**71.0**	83.7	13.5	36.5	46.8	60.5
Copy and paste families															
Augsmall [13]	21.1	39.9	54.3	15.8	34.7	43.8	54.3	34.1	59.9	**71.0**	84.0	12.1	36.4	47.1	60.5
CopyPaste [41]	21.9	38.3	52.3	16.2	33.6	42.3	52.3	36.1	58.8	69.0	82.6	13.8	34.0	45.5	58.2
Ours	**22.4**	**40.3**	53.9	17.5	**36.1**	**44.5**	53.9	36.3	**62.0**	**71.0**	83.5	**14.5**	**39.1**	48.5	59.9

As shown in Table 5, our method is well-performing on the scale of $(26, 64]$. This is because scale-based learnability encourages our method to augment the objects with scale in this range. Instead of extending all object instances uniformly, our method focuses on expanding small object instances more worthy of enhancement. This can avoid wasteful enlargement for weak, low-information tiny objects or large objects with large amounts. Therefore, it is possible that the detection performance for large or tiny objects without getting augmented may not be as good as other methods. However, our method usually outperforms others for objects with scale in $(26, 64]$ as depicted in Table 5. For example, our $AP_{(26,64]}$ is greater than that of CopyPaste by 2.5%. Regarding $AP^{50}_{(26,64]}$, our method can outperform CopyPaste by 3.2%. Specifically, our $AP^{75}_{(26,64]}$ surpasses that of CopyPaste by over 5%. Moreover, our method can achieve top-1 detection performance for objects of scale in $(0, 26]$ as well.

Comparative experimental results on COCO2017 dataset: Comparative experiments on the COCO 2017 dataset are also conducted. Table 6 presents the detection accuracy of various methods based on Faster RCNN. From Table 6 we could find that Augsmall is capable of reaching top-1 accuracy for $AP^{75}_{(0,26]}$ and AP_S. This is because Augsmall focuses on augmenting the small objects with scale in $(0, 26]$. Unfortunately, its $AP^{50}_{(0,26]}$ and $AP_{(0,26]}$ are inferior to ours. Besides, our method achieves the top-1 accuracy on a scale of $(0, 26]$ almost under all IoUs. Furthermore, our method also can outperform the others on a scale of $(0, 26]$.

Note, that our method mainly concentrates on imbalanced small object detection instead of the large ones. Therefore, after augmenting small objects through our method, the situation where the model is dominated by large objects will be alleviated to some extent, as shown in [16]. However, the detection performance for small objects can be

enhanced. These experimental results demonstrate the superiority over others regarding small object detection.

Table 6. Result of object detection on the COCO dataset. We achieve the best small object accuracy on Faster-RCNN-50-FPN.And in the medium object also has a better performance.

Method	AP_S	AP_M	AP_L	$AP_{(0,26)}$	$AP_{[26,64)}$	$AP_{[64,96)}$	$AP_{[96,\infty)}$	$AP^{50}_{(0,26)}$	$AP^{50}_{[26,64)}$	$AP^{50}_{[64,96)}$	$AP^{50}_{[96,\infty)}$	$AP^{75}_{(0,26)}$	$AP^{75}_{[26,64)}$	$AP^{75}_{[64,96)}$	$AP^{75}_{[96,\infty)}$
Baseline [61]	10.9	28.9	**40.4**	9.2	24.1	**34.7**	**40.4**	17.9	40.3	55.6	62.4	8.5	25.8	**37.3**	43.4
None copy and paste families															
Cutout [38]	10.9	27.6	41.0	9.3	22.9	33.6	41.0	17.8	38.8	55.1	63.2	8.7	24.1	35.1	**43.7**
GridMask [57]	11.1	27.8	39.0	9.3	23.2	33.6	39.0	18.0	39.7	54.6	61.0	8.4	23.9	35.7	41.9
Cutmix [39]	11.0	27.5	37.4	9.3	23.3	32.9	37.4	18.0	39.5	53.3	59.1	8.8	24.2	35.4	39.9
Copy and paste families															
Augsmall [13]	**11.3**	28.2	40.2	9.2	23.6	33.8	40.3	17.7	39.6	53.9	62.3	**9.0**	25.0	36.3	43.3
CopyPaste [41]	9.7	26.5	37.2	8.6	21.7	32.2	37.2	17.1	37.6	52.8	59.1	8.0	22.4	34.4	39.7
Ours	11.2	**29.2**	39.9	**10.2**	**24.4**	34.3	39.9	**20.2**	**41.7**	**57.4**	**63.4**	**9.0**	**25.7**	36.2	43.5

5. Conclusions

We present a value-guided adaptive augmentation method (ValCopy-Paste) for scale-imbalanced small object detection. In our method, we introduce scale-based learnability and quantity-based necessity to depict the object-relevant value and then construct the instance value distribution. This is advantageous to identify which object instances need and deserve augmentation as well. Benefiting from the guidance of the instance value distribution we can suppress the unnecessary and worthless augmentation, but encourage the imperative and invaluable one. Additionally, aimed at avoiding the interference of over-complex contexts, we propose to paste the small object instances to any flat areas of new background images. In this connection, we can achieve a trade-off between the diversity and low interference of contexts.

Comparative experiments are conducted with several representative data augmentation methods on MS COCO 2017 and PASCAL VOC 2012 datasets. Throughout the experiments, we can find that our ValCopy-Paste exhibits obvious superiority over others. As is well known, detecting and tracking imbalanced small targets in remote sensing images is a challenging problem. In the future, we will leverage the idea of ValCopy-Paste to improve the performance of remote sensing image-related detection or tracking tasks.

Author Contributions: Idea, writing and resources, H.W.; Methodology, supervision, and writing, C.S.; Software and validation, F.J.; Editing, A.W.; Formal analysis and editing, S.L.; Resources and supervision, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported in part by the National Natural Science Foundation of China (Grants No. 61601397, 62076249, 42241109, and 42271350), the National Key Research and Development Program of China under Grant 2022YFB3903401. We should thank Microsoft and the computer vision group at the University of Oxford, UK for providing the VOC and COCO datasets, respectively.

Data Availability Statement: The datasets are publicly available datasets.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kar, O.F.; Yeo, T.; Atanov, A.; Zamir, A. 3D Common Corruptions and Data Augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18963–18974.
2. Zhang, Z.; Qiao, S.; Xie, C.; Shen, W.; Wang, B.; Yuille, A.L. Single-shot object detection with enriched semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5813–5821.
3. Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; Xu, C. Distilling Object Detectors via Decoupled Features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2154–2164.

4. Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; Sun, J. OTA: Optimal Transport Assignment for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 303–312.
5. Joseph, K.J.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards Open World Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5830–5840.
6. Ma, Y.; Liu, S.; Li, Z.; Sun, J. IQDet: Instance-Wise Quality Distribution Sampling for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1717–1725.
7. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
8. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
9. Shen, R.; Bubeck, S.; Gunasekar, S. Data augmentation as feature manipulation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 19773–19808.
10. Leng, Z.; Cheng, S.; Caine, B.; Wang, W.; Zhang, X.; Shlens, J.; Tan, M.; Anguelov, D. PseudoAugment: Learning to Use Unlabeled Data for Data Augmentation in Point Clouds. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 555–572.
11. Yang, Y.; Zhang, X.; Guan, Q.; Lin, Y. Making Invisible Visible: Data-Driven Seismic Inversion with Spatio-Temporally Constrained Data Augmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
12. Atienza, R. Improving Model Generalization by Agreement of Learned Representations from Data Augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 372–381.
13. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
14. Bai, Y.; Yang, Y.; Zhang, W.; Mei, T. Directional Self-Supervised Learning for Heavy Image Augmentations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16692–16701.
15. Li, W.; Chen, J.; Cao, J.; Ma, C.; Wang, J.; Cui, X.; Chen, P. EID-GAN: Generative Adversarial Nets for Extremely Imbalanced Data Augmentation. *IEEE Trans. Ind. Inform.* **2022**, *19*, 3208–3218. [CrossRef]
16. Lim, J.S.; Astrid, M.; Yoon, H.J.; Lee, S.I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 181–186.
17. Van Dyk, D.A.; Meng, X.L. The art of data augmentation. *J. Comput. Graph. Stat.* **2001**, *10*, 1–50. [CrossRef]
18. Rebuffi, S.A.; Gowal, S.; Calian, D.A.; Stimberg, F.; Wiles, O.; Mann, T.A. Data augmentation can improve robustness. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29935–29948.
19. Lim, S.; Kim, I.; Kim, T.; Kim, C.; Kim, S. Fast autoaugment. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 6665–6675.
20. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
21. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
22. Cui, W.; Yan, S. Isotonic Data Augmentation for Knowledge Distillation. *arXiv* **2021**, arXiv:2107.01412.
23. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance Problems in Object Detection: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3388–3415. [CrossRef]
24. Hataya, R.; Zdenek, J.; Yoshizoe, K.; Nakayama, H. Meta approach to data augmentation optimization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2574–2583.
25. Ni, R.; Goldblum, M.; Sharaf, A.; Kong, K.; Goldstein, T. Data augmentation for meta-learning. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8152–8161.
26. Liu, J.; Chao, F.; Lin, C.M. Task augmentation by rotating for meta-learning. *arXiv* **2020**, arXiv:2003.00804.
27. Qin, Y.; Zhao, C.; Wang, Z.; Xing, J.; Wan, J.; Lei, Z. Representation based and Attention augmented Meta learning. *arXiv* **2018**, arXiv:1811.07545.
28. Yao, H.; Huang, L.K.; Zhang, L.; Wei, Y.; Tian, L.; Zou, J.; Huang, J. Improving generalization in meta-learning via task augmentation. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11887–11897.
29. Bird, J.J.; Barnes, C.M.; Manso, L.J.; Ekárt, A.; Faria, D.R. Fruit quality and defect image classification with conditional GAN data augmentation. *Sci. Hortic.* **2022**, *293*, 110684. [CrossRef]
30. Lim, S.K.; Loo, Y.; Tran, N.T.; Cheung, N.M.; Roig, G.; Elovici, Y. Doping: Generative data augmentation for unsupervised anomaly detection with gan. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 1122–1127.
31. Ntelemis, F.; Jin, Y.; Thomas, S.A. Image Clustering Using an Augmented Generative Adversarial Network and Information Maximization. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 7461–7474. [CrossRef] [PubMed]
32. Wang, W.; Chai, Y.; Cui, T.; Wang, C.; Zhang, B.; Li, Y.; An, Y. Restrained Generative Adversarial Network against Overfitting in Numeric Data Augmentation. *arXiv* **2020**, arXiv:2010.13549.
33. Golovneva, O.; Peris, C. Generative Adversarial Networks for Annotated Data Augmentation in Data Sparse NLU. *arXiv* **2020**, arXiv:2012.05302.

34. Hansen, N.; Su, H.; Wang, X. Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34, pp. 3680–3693.
35. Kostrikov, I.; Yarats, D.; Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv* **2020**, arXiv:2004.13649.
36. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 113–123.
37. Dwibedi, D.; Misra, I.; Hebert, M. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1301–1310.
38. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
39. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
40. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.
41. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2917–2927.
42. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1222–1230.
43. Noh, J.; Bae, W.; Lee, W.; Seo, J.; Kim, G. Better to Follow, Follow to Be Better: Towards Precise Supervision of Feature Super-Resolution for Small Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9724–9733.
44. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
45. Cui, L.; Ma, R.; Lv, P.; Jiang, X.; Gao, Z.; Zhou, B.; Xu, M. MDSSD: Multi-scale deconvolutional single shot detector for small objects. *arXiv* **2018**, arXiv:1805.07009.
46. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
47. Cao, G.; Xie, X.; Yang, W.; Liao, Q.; Shi, G.; Wu, J. Feature-fused SSD: Fast detection for small objects. In Proceedings of the Ninth International Conference on Graphic and Image Processing (ICGIP 2017), Qingdao, China, 14–16 October 2017; Volume 10615, p. 106151E.
48. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
49. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12595–12604.
50. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
51. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038. [CrossRef]
52. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
53. Dvornik, N.; Mairal, J.; Schmid, C. Modeling visual context is key to augmenting object detection datasets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 364–380.
54. Fang, H.S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.L.; Lu, C. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 682–691.
55. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
56. Walker, A.J. New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electron. Lett.* **1974**, *10*, 127–128. [CrossRef]
57. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Gridmask data augmentation. *arXiv* **2020**, arXiv:2001.04086.
58. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
59. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollar, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* **2015**, arXiv:1504.00325.
60. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2016**, arXiv:1512.02325.
61. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]

62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
63. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7310–7311.