information

Article

Data Augmentation Method for Pedestrian Dress Recognition in Road Monitoring and Pedestrian Multiple Information Recognition Model

Huiyong Wang, Liang Guo, Ding Yang and Xiaoming Zhang *

School of Information Science and Engineering, Hebei University of Science and Technology, Yuxiang Street, Shijiazhuang 050018, China
* Correspondence: zhangxiaom@hebust.edu.cn

Abstract: Road intelligence monitoring is an inevitable trend of urban intelligence, and clothing information is the main factor to identify pedestrians. Therefore, this paper establishes a multi-information clothing recognition model and proposes a data augmentation method based on road monitoring. First, we use Mask R-CNN to detect the clothing category information in the monitoring; then, we transfer the mask to the k-means cluster to obtain the color and finally obtain the clothing color and category. However, the monitoring scene and dataset are quite different, so a data augmentation method suitable for road monitoring is designed to improve the recognition ability of small targets and occluded targets. The small target mAP (mean average precision) recognition ability is improved by 12.37% (from 30.37%). The method of this study can help find relevant passers-by in the actual monitoring scene, which is conducive to the intelligent development of the city.

Keywords: data augmentation; instance segmentation; deep learning

Citation: Wang, H.; Guo, L.; Yang, D.; Zhang, X. Data Augmentation Method for Pedestrian Dress Recognition in Road Monitoring and Pedestrian Multiple Information Recognition Model. *Information* **2023**, *14*, 125. https://doi.org/10.3390/info14020125

Academic Editor: Chuan-Ming Liu

Received: 14 December 2022
Revised: 25 January 2023
Accepted: 6 February 2023
Published: 15 February 2023

1. Introduction

Pedestrian recognition technology has shown high application value in the fields of the intelligent transportation system, intelligent security monitoring, and intelligent robot, and it has become one of the important research directions in the field of computer vision. However, due to different or too small target scales and the problem of target occlusion in crowded scenes, pedestrian recognition faces great challenges. The current recognition algorithm is divided into one-stage recognition and two-stage recognition. The one-stage recognition method does not need to obtain the suggestion frame stage but directly generates the category probability and position coordinate value of the object, and then, it directly obtains the final recognition result. Two-stage recognition focuses on finding the location of the target object, obtaining the suggestion box, and then classifying the suggestion box to find a more accurate location. Two-stage recognition can be more accurate, but the speed is slower. In monitoring, complex scenes are a challenging recognition situation. Due to the existence of NMS (non-maximum suppression), the traditional recognition framework makes highly overlapping objects difficult to identify. A low NMS threshold can cause pedestrian overlap, while a higher one brings in plenty of false positives [1]. Chu [2] introduces the concept of multi-instance prediction, predicts a group of instances for each suggestion box, and improves the recognition ability of overlapping objects. Wu [3] analyzed from the perspective of video in the past, using the information of the before and after frames to correct the information of the current frame. Zhang [4] used 3D point cloud technology and proposed generating temporal proposals with both current and past boxes. There is no need to explicitly correlate objects across frames, and the prediction quality is improved. Recently, the visual attention mechanism [5] has been shown to have strong performance in a series of visual tasks. Jin [6] added a multi-scale deformation self-attention mechanism to the one-stage detector, which improved the cross-level spatial

adaptive features. Zhang [7] considered the direction of coordinate points and added a coordinate attention mechanism to focus on the region of interest in the image by way of weight adjustment, which enhanced the feature extraction capability of the model.

At present, researchers prefer to change the model and loss function in the research direction of pedestrian recognition. This paper tries to solve the problem from the dataset source. According to the characteristics of monitoring scenarios, it uses data enhancement methods and sets reasonable ways to improve the identification results. In order to have more applicability, the recognition of people is changed into the recognition of clothing, and the corresponding personnel can be found according to the clothing information. The responsibility of data enhancement is to improve the generalization ability of the model, which is a key component of the initial stage of the training network. It has proven to be a key technology for solving a variety of challenging deep learning tasks, including object recognition [8], natural language understanding [9], text recognition [10] and semi-supervised learning [11]. Initially, researchers augmented the dataset with some basic geometric changes, such as cropping and flipping operations, to generate more images with data perturbations. However, in word recognition, operations such as rotation and flipping are not applicable, so different augmentation methods need to be designed according to the needs of different requirements of engineering. Now, more attention is paid to the data enhancement of auto-tuning parameters. However, the automated approach itself requires a lot of iteration and is time-consuming. The emergence of GAN (generative adversarial network) brings a different approach to data enhancement, with generator and discriminator optimizing in opposite directions to counter training. Brais [12] designed a downsampled GAN, which requires a large object dataset and a small object dataset. The large object dataset uses the generator to generate small fake samples, which are mixed into the small object dataset to fool the discriminator.

The popularity of online stores and the high transaction volume of the clothing market make clothing retrieval more and more important. Most buyers will filter by search type. The feature recognition of clothing images has become a hot research topic. Attractive clothing color details are making more and more people buy clothing in specified colors. It is also widely used in the security field. Face recognition provides face distinction and recognition functions for security services. However, in winter or cold periods, facial information may not be able to correctly compare the similarity with the sample due to headwear, scarves, and camera angles. In the case that facial information cannot be obtained normally, clothing features can be used to retrieve and identify specific target people, which can not only protect the privacy of residents but also identify the target they are looking for.

With the advancement of technology, deep learning has greatly improved the performance of recognition and classification tasks, and the development of clothing image datasets [13,14] has promoted the development of clothing feature acquisition. Niza [15] et al. proposed a segmentation mechanism based on RoI (Region of Interest), which uses human anatomy to divide the person in the picture into several parts and uses HOG (Histogram of Oriented Gradient) features for similarity comparison to filter similar images. However, when shopping, images are often strange. This method requires pictures of specific poses, so the search is limited and it is difficult to apply in real life. Hussain [16] et al. developed an intelligent formal clothing recognition system, which is based on scale-invariant features and artificial intelligence technology for unified clothing classification. The method used SIFT (Scale-Invariant Feature Transform) and SURF (Speeded Up Robust Feature) features to extract invariant features in the video and then inputs the features into KNN for training and optimization feature solving, which can judge the wearing situation of formal clothes in real time. The CenterNet [17] feature point recognition network of Alexey [18] takes the target as a point and uses the center point to regress some attributes such as size and 3D range. The algorithm is simple, fast, and end-to-end. This neural network architecture has only one stage, each image takes about 35 ms, and it has a high accuracy rate for clothing recognition tasks. The color of clothing as a non-biological feature is an important

piece of information that can easily classify people by color. Jiri [19] proposed a clothing color recognition method, which uses mean filtering to remove the background to extract the outline of the person and detects the extracted outline using the gradient descriptor algorithm based on the histogram. It selects 26 nodes as the feature points of human body parts and finally divides the entire outline into 26 parts and merges them. However, the algorithm still fails to remove the problem of insufficient background separation ability.

The key issue of this paper is how to make the clothing-based instance segmentation model improve the generalization ability of the model in the case of a small dataset and speed up the training speed in the case of a large data volume. Therefore, the data augmentation of clothing images is designed to increase the generalization ability and training speed of the model, and a framework for multi-information extraction of clothing is studied according to the characteristics of instance segmentation. It is worth noting that the previous clothing recognition task is to detect single information through clustering or deep learning models. However, this paper not only designs an instance segmentation data augmentation algorithm but also implements a multi-task clothing information recognition system. In addition to identifying the type of clothing in the image, color recognition can also be performed.

For the problems of small datasets and the low efficiency of training, we propose Mask-Mosaic and Mask-Mosaic++ data augmentation for clothing-based instance segmentation. The validity of the model is verified from three aspects: time effect comparison experiment, model switching adaptation experiment, and generalization effect experiment. In the acquisition of color information, in view of the lack of existing background removal, this paper designs a method combining instance segmentation and clustering to remove background through powerful instance segmentation and obtain color by k-means clustering used for foreground pixel information.

For the data augmentation problem in the field of clothing feature recognition, the main contributions of this paper are as follows.

- Mask-Mosaic and Mask-Mosaic++ data augmentation methods are proposed. Both of them involve data augmentation of mixed samples, the former can speed up the training speed of the network, and the latter can improve the generalization ability of the network. Both of these data augmentation methods combine four pictures into one picture by placing them in a certain form. In the end, the recognition ability of Mask-Mosaic++ in the case of small targets and occlusions is greatly improved.
- An integrated system of clothing type recognition and color recognition is proposed. By using the Mask R-CNN [20] instance segmentation model to predict the clothing type, the instance mask is input into the k-means clustering, and the main color information is obtained after k-means clustering of the color.

The remainder of this paper is organized as follows. Section 2 introduces the overall structure of the model, data enhancement methods, case segmentation model, network training parameters and model evaluation formula. Section 3 visualizes the effect of the model and verifies the performance of data enhancement. Finally, Section 4 concludes the paper and provides guidelines for future work.

1.1. Instance Segmentation

CNN (Convolutional Neural Network) has brought revolutionary changes to many fields of computer vision in recent years. R-CNN [21] made a breakthrough in using candidate regions to locate objects, changing the way of image recognition and image classification. However, the selective search algorithm is used to randomly generate a large number of candidate boxes, which leads to a slower running speed. In order to solve this problem, He Kaiming proposed Faster R- CNN [22], which uses a fully convolutional RPN (Region Proposal Network), which can greatly reduce the speed of recognition frame generation. In order to further improve the accuracy, He Kaiming added a branch of prediction mask on the basis of faster R-CNN, which can achieve pixel-level segmentation masks, with higher accuracy and a certain degree of slowdown. As the transformer [23]

achieved excellent results in the NLP (Natural Language Processing) task, it has also begun to migrate to use this method for recognition in vision. The Swin ransformer [5] builds hierarchical feature maps by incorporating deeper image patches, computes self-attention only within each local window, and has linear computational complexity over the input image size. Therefore, it can serve as a general backbone for image classification and dense recognition tasks.

1.2. Data Augmentation

Data augmentation [24] is mainly divided into three ways: basic geometric transformations [25,26], mixed sample [27–29], and data augmentation based on generative adversarial networks [30–32]. The data augmentation of mixed samples brings people different ideas. Mixup [30] is equivalent to the superposition of two faded images, but this faded method will change the distribution of data to a certain extent, and there is a negative increase in the ImageNet dataset. CutMix [28] is used to paste the rectangular frame of the target key area of one image to a random area of another image. Mosaic [29] randomly crops, scales, and stitches four images to generate a new composite image, which improves the robustness of the model for small object recognition. Data enhancement is also applicable to medical images; magnetic resonance imaging can use discrete wavelet and inverse discrete wavelet transform to enhance image quality [33].

1.3. Clothing Recognition Task

At present, there is not much research on the task of clothing recognition. This paper aims to accurately identify the target person from the picture according to the clothing characteristics of the pedestrian. Clothing mainly has three characters: category, color [33], and style [34]. In color recognition, it is usually necessary to use clustering to remove the background. With the development of neural networks, feature point recognition [16,17] becomes an effective way to remove background. In type recognition, the recognition is usually performed by means of deep learning. In the style judgment, the recognition is usually performed from the previous single image clothing separation to the current multimodal information style judgment method [35].

2. Model

The clothing recognition framework includes two parts: clothing type recognition and color recognition, as shown in Figure 1. Mask-Mosaic++ data augmentation using deepfashion2 [36] corrected data in type recognition is followed by Mask R-CNN training. The trained model splits the mask image of the target and uses k-means clustering to extract the main color. Finally, the extracted main color and color label is subjected to SVM (support vector machine) linear training to obtain a model for clothing color recognition.

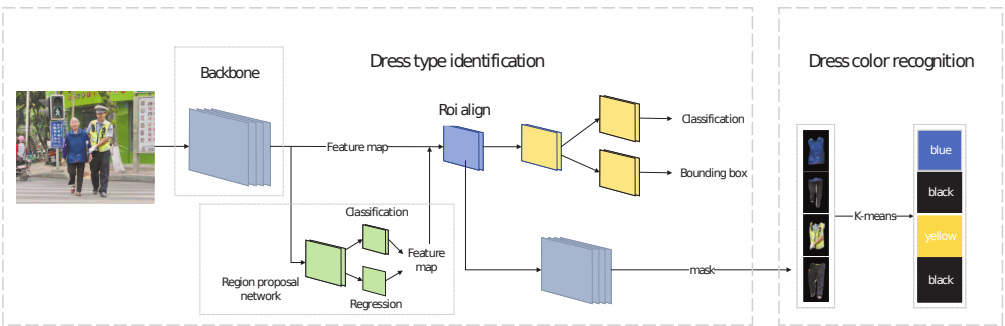

Figure 1. Clothing Multi-information Recognition Framework.

2.1. Mask-Mosaic++ Data Augmentation

In actual clothing recognition, the targets are often too small, and the number of large targets in the datasets is relatively large, which cannot fit the reality well, and the huge

amount of data will also bring difficulties for training. As shown in Figure 2, we first propose Mask-Mosaic instance segmentation data augmentation that can speed up training, and its principle is similar to the Mosaic [27] target recognition data augmentation method in yolov4. Four training images are mixed, and the size is changed to 512 pixels by 512 pixels (512 × 512). The four images are randomly flipped, randomly cropped, scaled, and stitched together. This data augmentation will be verified in subsequent experiments to speed up the pre-training of the model. The probability of a random flip is 0.5. In random clipping, if the bounding box area of the original target is less than 0.7 times the original, the original target will be deleted.

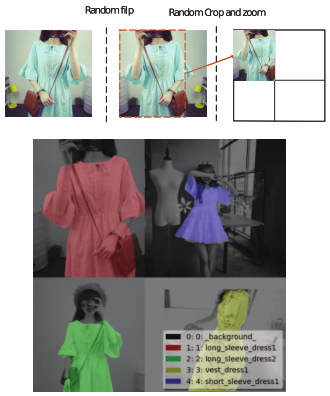

Figure 2. Mask-Mosaic++ data augmentation.

The ultimate purpose of data augmentation is to improve the robustness of the model, and speeding up the training speed is not the original intention of data augmentation. Improved Mask-Mosaic++ data augmentation is proposed through a modification of the strategy. The Mask-Mosaic data augmentation that is shown in Figure 3 selects four distinct images each time, and the number of datasets is compressed to one-quarter of the original. The improved Mask-Mosaic++ still selects four images at a time, but the images selected each time only replace one image for fusion instead of replacing four images, so the size of the transformed dataset is three less than the original dataset. We added the original datasets for training, so the amount of data is about twice the original, and the training strategy is used to regenerate the datasets every 10 rounds for training again.

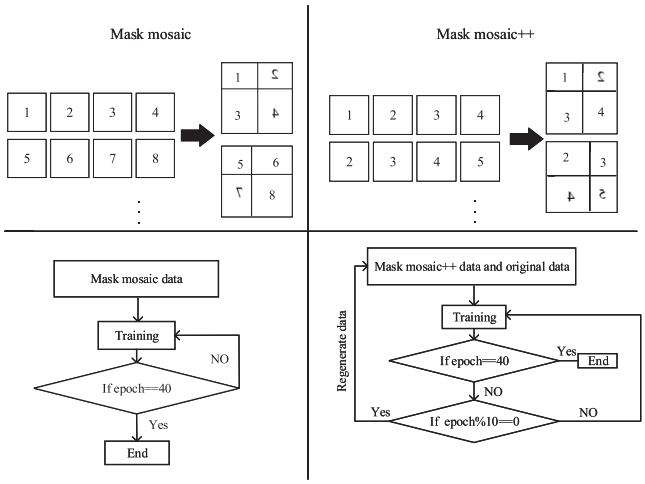

Figure 3. Data Augmentation Strategy.

2.2. Mask R-CNN

Mask R-CNN was proposed in 2017. This network can obtain high-quality image segmentation while accurately detecting objects. The network uses the Resnet series net-

work as the backbone network to perform image feature extraction at different levels. Mask R-CNN uses the backbone network to compress the feature image twice, three times, four times, and five times for the construction of the FPN (feature pyramid network) [37] to achieve multiple fusion. It input the fused image into the RPN (region recommendation network) to obtain candidate regions that may contain the target object, that is, the suggestion box [22]. Then, it uses the suggestion box to intercept the effective feature layer, obtains the local feature layer, unifies the intercepted results to a certain size [20], and then locates, classifies and segments the target.

As shown in Figure 4, Mask R-CNN is an instance segmentation model proposed by He [13], which is mainly divided into three parts: backbone network, region proposal network, and RolAlign. RolAlign is a method of generating an Rol (Region of Interest), and Rol is the proposed region for the original image. The original image extracts a specific map through the backbone network, inputs the feature map into the region proposal network, obtains the candidate region, and then inputs the feature map and the candidate region into RolAlign to obtain a fixed feature map. The output results predict the category, bounding box, and mask of the image. Due to the feature pyramid structure, there are five feature maps of different sizes, and objects of different sizes are assigned to different feature maps. In the observation of the dataset, it is found that the target of the data is too large, and the target in the application is too small, so the training of the small target should be increased. It can be seen from formula 1 that the preselected frame is assigned to a feature map of the feature map P2-P6, depending on the width and height of the target. Therefore, it will be reduced to obtain the target of the small object so that the target will be allocated to the small feature map, and the training of the small target will be increased to improve the recognition ability of the small target. The experimental analysis and application are calculated as shown in Equation (1).

$$K = \left[K_0 + \log_2(\sqrt{wh}/256)\right] \tag{1}$$

w and h correspond to the width and height of the preselected box, respectively;
K represents the level at which the preselected box belongs to the feature layer;
K_0 generally takes 4.

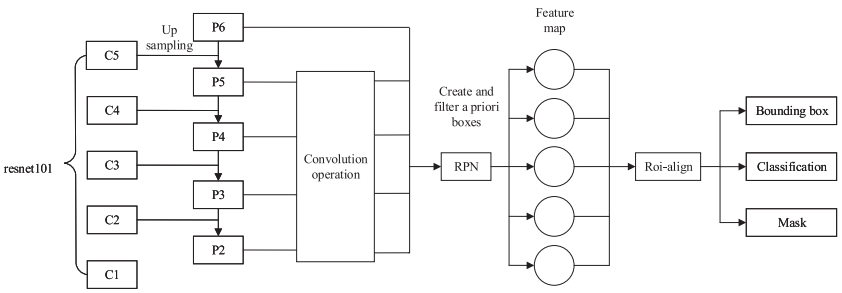

Figure 4. Mask R-CNN framework.

Mask R-CNN has three branches: prediction of classification loss of target category, prediction of boundary box loss of target boundary box, and prediction of mask prediction loss of target contour. Equation (2) is the total loss function, which is the sum of the three.

$$L = L_{cls} + L_{box} + L_{mask} \tag{2}$$

L_{cls} represents the loss of classification branch;
L_{box} represents the loss of the location branch, which is the boundary frame regression loss;
L_{mask} represents the loss of the split branch, which is the loss of mask.

Equation (3) is the calculation of classification and location loss. The first half of Equation (3) represents classification loss, and the second half represents location loss or boundary box regression loss.

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3}$$

$$L_{cls} = -[p_i^* \ln(p_i) + (1 - p_i^*) \ln(1 - p_i)] \tag{4}$$

$$L_{\text{reg}}(t_i, t_i^*) = \sum_i \text{smooth}_{L_1}(t_i, t_i^*) \tag{5}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| \mid 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{6}$$

P_i^* represents the probability that the ith anchor is predicted to be a true frame;
λ is a constant, generally 10 in the experiment;
t_i represents the boundary box regression parameter of the ith anchor predicted;
t_i^* represents the regression parameter of the boundary box of the ith anchor corresponding to the real label;
N_{cls} represents the number of samples; and
N_{reg} represents the number of anchor positions.

When training Mask R-CNN, it is also necessary to label the mask: that is, mark the outline of the object. L_{mask} is defined as the average binary cross entropy loss function, as shown in Equation (7). This function will classify each pixel and use the sigmoid function to perform the secondary classification to determine whether it is the category.

$$L_{\text{mask}} = -\sum_y y \ln(1 - \hat{y}) + (1 - y) \ln(1 - \hat{y}) \tag{7}$$

y represents the true value after binarization; and
$\hat{y}$ represents the predicted value after binarization.

2.3. k-Means Clustering Algorithm

The k-means algorithm is a typical representative of the partition clustering algorithm. In essence, this algorithm is based on the average value of objects in the cluster. To achieve global optimization, partition-based clustering requires all possible partitions to be exhausted. The processing process of the algorithm is as follows:

(1) Given a set of sample values X, randomly select k sample points as the center of k clusters. (2) For each data object, calculate the distance between the object and the center of k clusters and allocate the remaining samples to a cluster according to the minimum distance principle. (3) Calculate the arithmetic mean E of all data points in the cluster, as shown in Equation (8). (4) Repeat two and three times until Equation (9) is established.

$$E = \sum_{i=1}^{K} \sum_{X \in C_i} \left| X - \overline{X_i} \right|^2 \tag{8}$$

$$|E_2 - E_1| < \varepsilon \tag{9}$$

K is the total number of clusters, and $\overline{X_i}$ is the average value of cluster C_i.
ε is a small number, E_1 and E_2, respectively, represent the results of the previous and second iteration of Equation (9).

2.4. Network Training

We conduct all the experiments on NVIDIA GTX 1070Ti with the tensorflow1.15.0 and Keras2.1.5 frameworks. The size of the input training network image is 512 × 512 pixels, and the non-maximum suppression value of the suggested box is 0.7. The number

of boxes before is 1000, the resnet101 network is used as the backbone network, and the ratio of positive and negative samples is 1:3. In terms of optimizer selection, Adam can have faster gradient descent, and SGD (Stochastic Gradient Descent) is relatively more inclined to tuning. The mixed use of Adam and SGD was initially used, and the test found that the mixed use is not as effective as the direct use of SGD. Based on this, an optimizer in the form of SGD+Momentum is used to iterate for 40 epochs with a learning rate of 3×10^{-4}. Usually, the evaluation standard of data augmentation is based on the results of Box AP and Mask AP. In this experiment, Box AP is used for verification. It is better to use a pre-trained model in the case of small data [36], and for the initialization parameters of the backbone network, the coco dataset is used to pre-train the model.

2.5. Model Assessment

The *mAP* index is used for the evaluation of data enhancement results. *mAP* is defined as the average of the curve areas of precision (*P*) and recall (*R*) for each category. Generally speaking, when *R* is very low, *P* is very high. When more objects are recognized, more errors are recognized, and *P* may decrease. Finally, when all objects are recalled, that is, r is 1. At this time, the area of the curve from 0 to 1 is the *AP* value of this category. *P*, *R*, *AP* and *mAP* are calculated by Equation (10), Equation (11), Equation (12) and Equation (13), respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$AP = \int_0^1 P_{(R)} dR \tag{12}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{13}$$

TP, *FP* and *FN* are true positive, false negative and false positive, respectively, *C* is the number of categories, each category has an *AP* value, and *P(R)* is the *P* value when *R* is a certain value.

3. Results

As shown in the left of Figure 5, with the idea of solving the difficulty of identifying complex scenes, a two-stage detector was used. Trading precision for speed, it is only about four frames, but in a surveillance scenario, recognition can be set to three to four frames (GTX 1650 GPU) per second. The most important thing is that it can be selectively identified according to the color and category of clothing, and it can select specific clothing information to find relevant people, as shown in the right of Figure 5.

Figure 5. On the **left** is the resulting figure identifying clothing categories and colors at the same time, and on the **right** is the category set to recognize trousers.

3.1. Dataset

The data enhancement is based on the fact that the data volume is relatively small, so only a portion of the deepfashion2 [38] dataset was selected, as shown in Table 1. Moreover, it is difficult to label the instance segmentation dataset. As shown in Figure 6, it is inevitable that there will be multi-labeled and incorrectly labeled images, which will affect the training and need to use Labelme to re-label the mask.

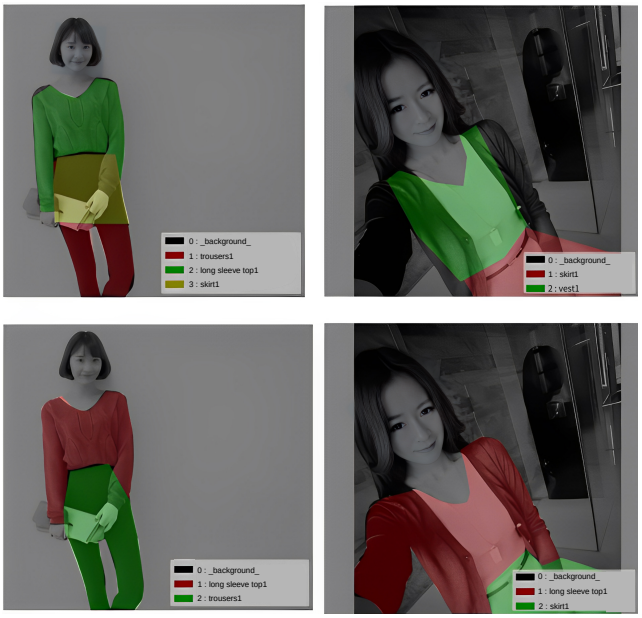

Figure 6. Adjusting the wrong (**top**) clothing tag to the correct one (**bottom**).

Table 1. Dataset.

Dataset	Train	Valid	Test						
			degree of occlusion			object size			
Deepfashion2 [38] (a little)	1044	456	small 269	moderate 265	serious 193	small 191	medium 253	large 192	normal test 418

3.2. Data Augmentation Experiment

In this section, we will show that Mask-Mosaic++ is an effective data augmentation method and provide the reasons. Table 2 gives the comparison results of mAP in the case of original data and data augmentation training. It can be seen that from the degree of occlusion and the size of the target, two cases are performed to verify the effectiveness of this data augmentation. It can be seen that in the case of a small amount of data, Mask-Mosaic++ has an average improvement of 6.17% compared with the model trained on the original data in different degrees of occlusion. In the case of smaller instance sizes, there is a significant increase of 12.37%.

The reason can be explained. Most of the images of the original data have an area larger than 512 × 512, and the size of the four placement areas of the image in the data augmentation is about 205 × 205∼307 × 307 pixels. Therefore, many pictures have to be reduced to more than twice the original size. In the section on the clothing recognition model, the value of K obtained from Equation (1) is reduced by 1, and the number of feature maps trained with small fields of view is increased accordingly. It can be seen from the Mask R-CNN frame diagram that in the training process, as shown in P2–P5, the smaller the feature map, the more detailed things can be trained, and it has a certain feedback effect on the network area with a large field of view. Therefore, a model trained in this way will effectively increase the ability to recognize small objects. Mask-Mosaic++ has a certain degree of improvement compared to the original results in various occlusion

situations, mainly because it generates some occlusion-like data through random clipping, which directly increases the model's anti-occlusion ability. Use the valid dataset, we try to replace different backbone networks and train these backbone models on the image size of 512 × 512. As shown in Table 3, the applicability of Mask-Mosaic++ is still very good, while the performance of Mask-Mosaic has dropped significantly. From the mAP change graph in Figure 7, it can be seen that Mask-Mosaic cannot be trained effectively, so I want to check the change of Mask-Mosaic generalization ability by observing val_loss.

Table 2. Instance segmentation mAP for different augmentation strategies.

	Resnet101(512)	**Degree of Occlusion**			**Target Size**		
		Small	**Moderate**	**Serious**	**Small**	**Medium**	**Large**
$AP^{iou=0.50}_{box}$	Original	34.89	32.33	31.11	30.37	35.75	34.59
$AP^{iou=0.50}_{box}$	Mask-Mosaic	32.60	32.63	25.55	36.05	30.35	26.89
$AP^{iou=0.50}_{box}$	Mask-Mosaic++	**39.8 1**	**39.90**	**37.13**	**42.74**	**39.65**	**38.52**
$AP^{iou=0.75}_{box}$	Original	25.62	23.43	17.62	21.67	24.35	22.56
$AP^{iou=0.75}_{box}$	Mask-Mosaic	22.73	23.34	15.23	25.36	20.57	16.39
$AP^{iou=0.75}_{box}$	Mask-Mosaic++	**29.63**	**29.47**	**26.35**	**30.28**	**28.37**	**28.88**
AP_{box}	Original	17.26	13.68	13.33	12.19	16.58	14.36
AP_{box}	Mask-Mosaic	14.75	14.26	10.38	16.48	11.78	8.37
AP_{box}	Mask-Mosaic++	**20.75**	**19.89**	**18.38**	**21.45**	**17.45**	**18.23**

Table 3. Validation using two backbone networks on the common testing set.

Model	**#Params**	$\mathbf{AP^{iou=0.75}_{box}}$
Resnet101 FPN (512)	250 M	33.49
w/Mask-Mosaic	250 M	29.38
w/Mask-Mosaic++	250 M	**37.65**
Resnet51 FPN (512)	175 M	24.82
w/Mask-Mosaic	175 M	12.34
w/Mask-Mosaic++	175 M	**27.49**

As shown in Figure 8, the change diagram of val_loss and train_loss, train_loss is a normal decline, only Mask-Mosaic++ can effectively decline in the change of val_loss, val_loss in the original tends to remain unchanged, and val_loss of Mask-Mosaic has an increased trend, indicating data overfitting. These findings indicate that only Mask-Mosaic++ can effectively improve the generalization ability of the model in the case of a small number of clothing datasets.

Although the Mask-Mosaic data augmentation experiment has a negative effect on the generalization ability, due to the advantages of the Mask-Mosaic datasets enrichment, it may speed up the training speed, so we tried a speed comparison experiment in this paper. Since the time consumed by saving the model is quite different from the training time, it is found through experiments that the size of the dataset is almost proportional to the training time. We take one epoch as the time benchmark for Mask-Mosaic training: that is, Mask-Mosaic, Mask-Mosaic++, and original training for one epoch consume the time of 1, 8, and 4, respectively. It can be seen from Figure 9 that Mask-Mosaic can speed up the pre-training speed. When the datasets are huge, Mask-Mosaic can be used to improve the pre-training effect and then use the original datasets or Mask-Mosaic++ datasets for training. In the case of a large instance segmentation dataset, Mask-Mosaic can be used to speed up the training in the early stage, and Mask-Mosaic++ or the original dataset can be used for replacement in the later stage.

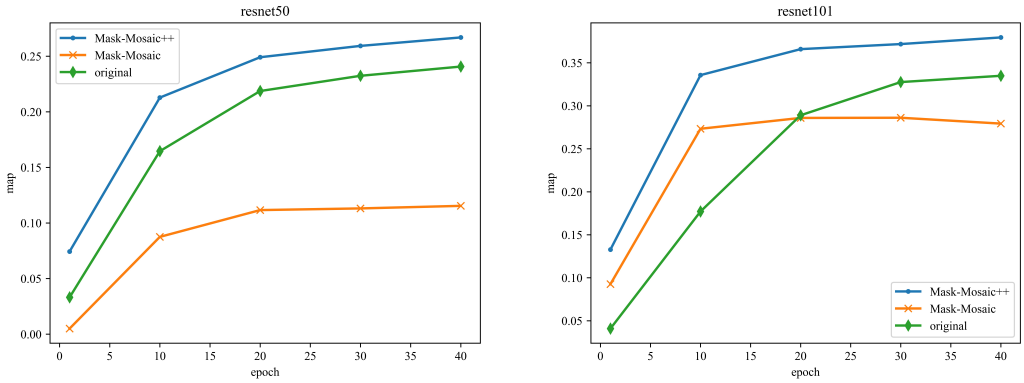

Figure 7. mAP changes of different backbone networks.

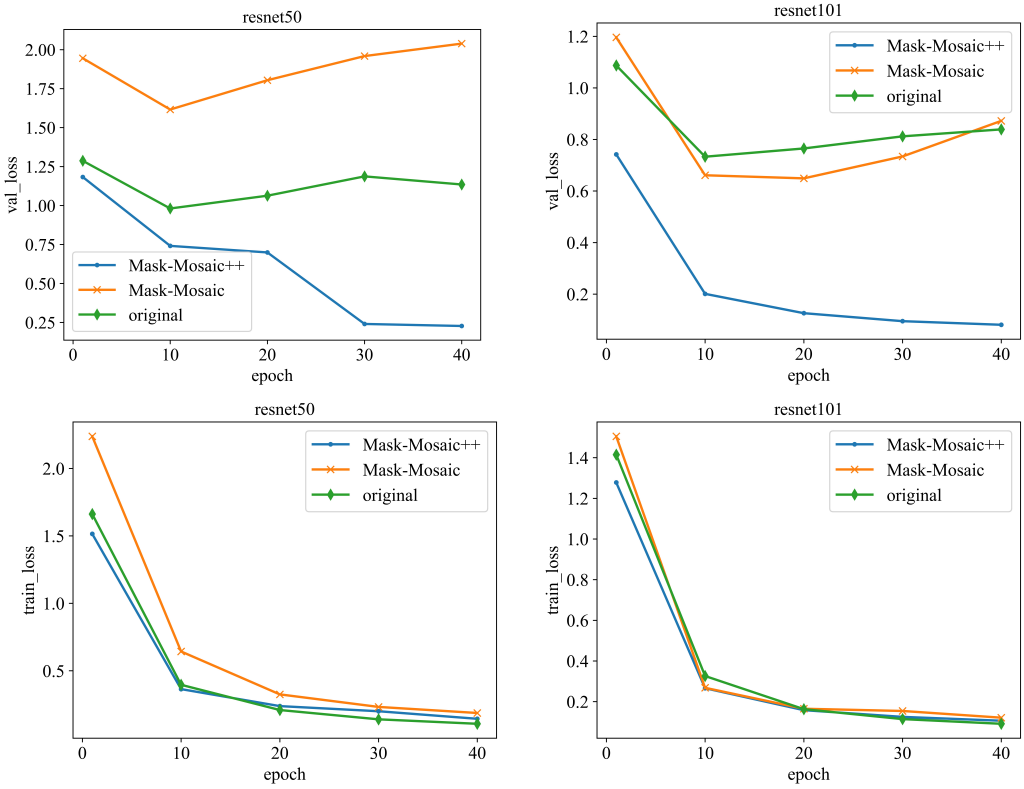

Figure 8. Comparison of val_loss and train_loss.

Finally, different one-stage detectors were used for verification. As shown in the Table 4, Mask-Mosaic++ data enhancement is feasible. In YOLOv5n and YOLOCT ++ [39], the two most advanced detectors, mAP improved by 1.36% and 1.47%, respectively.

Table 4. Validation using two backbone networks on the common testing set.

Model	$AP^{iou=0.50}_{box}$	$AP^{iou=0.50}_{mask}$
YOLOv5n + Resnet101	31.79	30.36
YOLOv5n + Mask-Mosaic++	**33.15**	**31.65**
YOLOCT++ + Resnet101	32.58	31.22
YOLOCT++ + Mask-Mosaic++	**34.05**	**32.52**

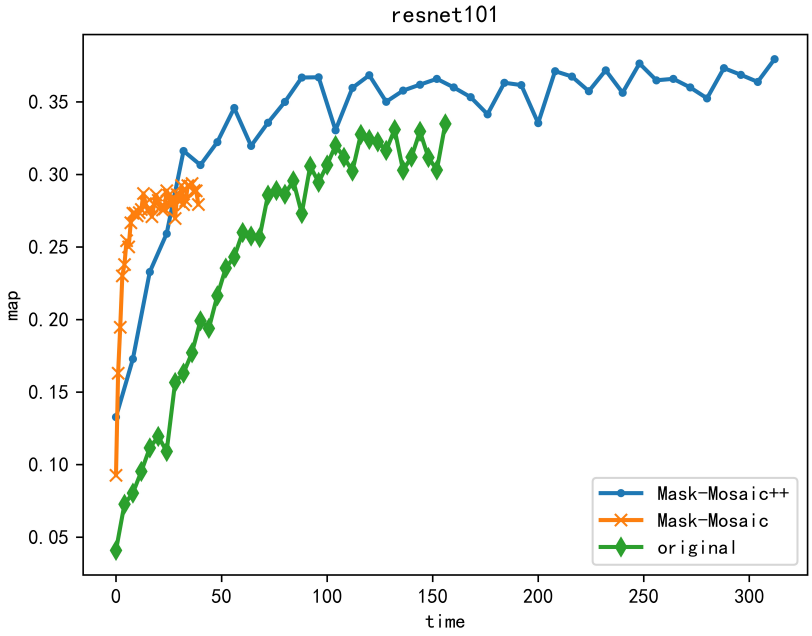

Figure 9. Time effect comparison.

3.3. Clothing Color Recognition

The effect of color clustering in different spaces is different. In addition, even though the foreground is clean, monochrome outfits are rare. For example, if the clothes are mostly black and k is 1, the result might not be black. Therefore, the k values of different color spaces and k-means are compared experimentally.

Color recognition is carried out on the basis of instance segmentation. Regardless of whether the instance segmentation classification is correct, we only look at the accuracy of the color at the final corresponding position. The mask of the instance segmentation is the outline of the object, which is 26 pixels by 26 pixels. By mapping the mask graph to the original graph, it collects the pixels of the target, and these pixels are the data source of k-means clustering. The number of pixels varies according to the target size. We add the color attribute to the deepfashion2 dataset, using only 400 images, in which the ratio of training and testing is 8:2. We can see in Table 5 that the best result is 76.66% after converting the mask of the image to HSV space and using the SVM linear classifier to train it. It can be seen that the correct increase of the k value can eliminate the erroneous influence of the secondary color to a certain extent.

Table 5. The influence of k value of k-means on accuracy.

	1	**2**	**3**	**4**
RGB_image + RGB_k-means [1]	73.33%	75%	75%	73.33%
RGB_image + HSV_k-means	71.66%	73.33%	73.33%	75%
HSV_image + RGB_k-means	63.33%	65%	62%	65%
HSV_image + HSV_k-means	71.66%	75%	76.66%	73.33%

[1] RGB_image means to convert the mask of the image to RGB value. RGB_k-means means to convert the masked values into RGB format. If the mask itself is RGB, no transformation is required.

4. Discussion

At present, there are few kinds of research on instance segmentation data augmentation, and different augmentation methods are suitable for different practical applications. To this end, this paper proposes a data augmentation algorithm Mask-Mosaic++ for clothing instance segmentation. Experimental results show that this data enhancement method can increase the generalization ability of the model and prevent the model from overfitting.

The recognition ability of small target objects and occluded objects is improved. It also provides a simple expansion idea for splitting datasets for instances with small amount of data. Although the performance of Mask-Mosaic is not very good, it can be seen from Figure 9 that Mask-Mosaic can accelerate the early training speed. Continuing training on the Mask-Mosaic++ dataset can speed up the pre-training. In terms of application value, this paper proposes a clothing recognition method that combines clothing type recognition and color recognition. Experiments show that Mask R-CNN can extract the foreground of clothes, which can solve the problem of insufficient background separation in complex cases. The combination of the two tasks makes up for the problem of insufficient background removal in complex scenes and provides a new idea for information acquisition. This method can be directly optimized by replacing the model and clustering algorithm. Target tracking applications can be extended in the future. Tracking people in blue, for example, also provides an alternative way of thinking about goal-tracking apps.

Author Contributions: Conceptualization, H.W.; methodology, H.W.; software, H.W.;validation, H.W. and L.G.; formal analysis, L.G.; investigation, D.Y.; resources, H.W.; writing—original draft preparation, D.Y.; visualization, L.G.; writing—review and editing, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Project of the Hebei Education Department (grant no. ZD2021048), the Natural Science Foundation of the Hebei Province (grant no. F2022208002), and Hebei Province Special research and development plan project (grant no. SJMYF2022X13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the second author. The data has not been made public because of privacy concerns.

Acknowledgments: Thanks for the support of the School of Information Science and Engineering of Hebei University of Science and Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, X.; Ge, Z.; Jie, Z.; Yoshie, O. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 10750–10759.
2. Chu, X.; Zheng, A.; Zhang, X.; Sun, J. Detection in crowded scenes: One proposal, multiple predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 12214–12223.
3. Wu, J.; Zhou, C.; Yang, M.; Zhang, Q.; Li, Y.; Yuan, J. Temporal-context enhanced detection of heavily occluded pedestrians. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 13430–13439.
4. Zhang, Z.; Gao, J.; Mao, J.; Liu, Y.; Anguelov, D.; Li, C. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11346–11355.
5. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10012–10022.
6. Yuan, J.; Panagiotis, B.; Stathaki, T. Effectiveness of Vision Transformer for Fast and Accurate Single-Stage Pedestrian Detection. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10-16 December 2023.
7. Zhang, Y.; Zhou, A.; Zhao, F.; Wu, H. A lightweight vehicle-pedestrian detection algorithm based on attention mechanism in traffic scenarios. *Sensors* **2022**, *22*, 8480. [CrossRef] [PubMed]
8. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.Y.; Shlens, J.; Le, Q.V. Learning data augmentation strategies for object detection. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 566–583.
9. Zhou, K.; Zhao, W.X.; Wang, S.; Zhang, F.; Wu, W.; Wen, J.R. Virtual data augmentation: A robust and general framework for fine-tuning pre-trained models. *arXiv* **2021**, arXiv:2109.05793.
10. Luo, C.; Zhu, Y.; Jin, L.; Wang, Y. Learn to augment: Joint data augmentation and network optimization for text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 13746–13755.

11. Yuan, J.; Liu, Y.; Shen, C.; Wang, Z.; Li, H. A Simple Baseline for Semi-supervised Semantic Segmentation with Strong Data Augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 8229–8238.
12. Bosquet, B.; Cores, D.; Seidenari, L.; Brea, V.M.; Mucientes, M.; Del Bimbo, A. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognit.* **2023**, *133*, 108998. [CrossRef]
13. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1096–1104.
14. Zheng, S.; Yang, F.; Kiapour, M.H.; Piramuthu, R. Modanet: A large-scale street fashion dataset with polygon annotations. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1670–1678.
15. Aulia, N.; Arnia, F.; Munadi, K. HOG of Region of Interest for Improving Clothing Retrieval Performance. In Proceedings of the 2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Banda Aceh, Indonesia, 22–24 August 2019; pp. 7–12.
16. Hussain, T.; Ahmad, M.; Ali, S.; Khan, S.; Rahman, A.; Haider, A. An Intelligent Dress Uniform Identification System. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 30–31 January 2019; pp. 1–4.
17. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
18. Sidnev, A.; Trushkov, A.; Kazakov, M.; Korolev, I.; Sorokin, V. Deepmark: One-shot clothing detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
19. Prinosil, J. Clothing Color Based De-Identification. In Proceedings of the 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, 4–6 July 2018; pp. 1–5.
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
22. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
24. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
25. 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.
26. 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, 4–8 January 2022; pp. 2574–2583.
27. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
28. 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.
29. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
30. Zhang, X.; Wang, Z.; Liu, D.; Lin, Q.; Ling, Q. Deep adversarial data augmentation for extremely low data regimes. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 15–28. [CrossRef]
31. Mansourifar, H.; Chen, L.; Shi, W. Virtual big data for GAN based data augmentation. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 1478–1487.
32. Kora Venu, S.; Ravula, S. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest X-ray images. *Future Internet* **2020**, *13*, 8. [CrossRef]
33. Algabri, R.; Choi, M.T. Deep-learning-based indoor human following of mobile robot using color feature. *Sensors* **2020**, *20*, 2699. [CrossRef] [PubMed]
34. Patel, C.; Liao, Z.; Pons-Moll, G. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 7365–7375.
35. Hidayati, S.C.; Goh, T.W.; Chan, J.S.G.; Hsu, C.C.; See, J.; Wong, L.K.; Hua, K.L.; Tsao, Y.; Cheng, W.H. Dress with style: Learning style from joint deep embedding of clothing styles and body shapes. *IEEE Trans. Multimed.* **2020**, *23*, 365–377. [CrossRef]
36. Zoph, B.; Ghiasi, G.; Lin, T.Y.; Cui, Y.; Liu, H.; Cubuk, E.D.; Le, Q. Rethinking pre-training and self-training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3833–3845.
37. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

38. Ge, Y.; Zhang, R.; Wang, X.; Tang, X.; Luo, P. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2019; pp. 5337–5345.
39. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact++: Better real-time instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1108–1121. [CrossRef]