Mathematical Biosciences and Engineering

MBE, 18(2): 986–999.
DOI: 10.3934/mbe.2021053
Received: 24 October 2020
Accepted: 21 December 2020
Published: 05 January 2021

http://www.aimspress.com/journal/MBE

Research article

How to construct low-altitude aerial image datasets for deep learning

Xin Shu[1], Xin Cheng[1], Shubin Xu[2], Yunfang Chen[1], Tinghuai Ma[3] and Wei Zhang[1,4,*]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] Cyberspace Security Research Institute, China Electronics Technology Group Corporation, Xiong'an New Area 071000, China
[3] School of Computer & Software, Nanjing University of information science & Technology, Nanjing 210044, China
[4] Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

* **Correspondence:** Email: zhangw@njupt.edu.cn.

Abstract: The combination of Unmanned Aerial Vehicle (UAV) technologies and computer vision makes UAV applications more and more popular. Computer vision tasks based on deep learning usually require a large amount of task-related data to train algorithms for specific tasks. Since the commonly used datasets are not designed for specific scenarios, in order to give UAVs stronger computer vision capabilities, large enough aerial image datasets are needed to be collected to meet the training requirements. In this paper, we take low-altitude aerial image object detection as an example to propose a framework to demonstrate how to construct datasets for specific tasks. Firstly, we introduce the existing low-altitude aerial images datasets and analyze the characteristics of low-altitude aerial images. On this basis, we put forward some suggestions on data collection of low-altitude aerial images. Then, we recommend several commonly used image annotation tools and crowdsourcing platforms for data annotation to generate labeled data for model training. In addition, in order to make up the shortage of data, we introduce data augmentation techniques, including traditional data augmentation and data augmentation based on oversampling and generative adversarial networks.

Keywords: UAVs; aerial image; datasets; deep learning; data augmentation

1. Introduction

The combination of Unmanned Aerial Vehicle (UAV) technologies and deep learning makes UAV applications more and more popular in various fields, such as surveillance, search and rescue, tree height and biomass estimation [1]. One of the ways to improve the performance of deep learning model is to increase the size of datasets [2]. In order to give UAVs automated computer vision capabilities, it is necessary to organize large enough low-altitude aerial image datasets that meet the requirements of algorithm training. There are numerous classical datasets in the field of computer vision, such as Visual Object Classes (VOC) [3], ImageNet [4], Microsoft Common Objects in Context (MS COCO) [5] and so on. These datasets play the major roles in training and evaluating the algorithms, but the images and videos taken by UAVs are obviously different from them. For example, due to the shooting height, the images taken by UAVs have a wide view, and the object size is much smaller than that in ordinary images. Therefore, the general computer vision datasets cannot be used directly in training and evaluating UAV computer vision models.

Researchers have developed a few aerial view datasets for UAV vision tasks. These datasets generally focus on object detection, object tracking and action/event recognition tasks. Datasets used for object detection include UAVDT [6], the DET subset and the VID subset of VisDrone [7]. Datasets used for object tracking include UAVDT, Stanford Drone [8], UAV123 [9] and the MOT subset and the SOT subset of VisDrone. Datasets used for action/event recognition include Okutama-Action [10], VIRAT [11] and UCLA Aerial Event dataset [12]. In addition, there are two datasets used for other tasks, i.e. Mini-drone video dataset [13] and CARPK dataset [14]. Mini-drone video is a dataset for privacy protection in UAV surveillance. In this dataset, human behaviors are divided into three categories: normal, suspicious and illicit behaviors. CARPK is a dataset used for car counting. We can obtain the above datasets directly from the Internet, or download them by filling in the questionnaire. However, the dataset constructed by Kamran et al. [15] for military vehicle detection from low-altitude aerial images is not released. We summarize the above available datasets and sort out the attributes of each dataset in Table 1. Du et al. [6] also did the collection and sorting of datasets. Unlike their work, our work focuses on datasets with aerial view.

The existing low-altitude aerial image datasets are aimed at specific computer vision tasks or specific scenarios, so these datasets cannot fully meet the training needs of other tasks. In order to complete a task, we need to organize a dataset that meets its requirements.

This paper makes the following contributions:

(i) Review three classical datasets in object detection and the existing low-altitude aerial image datasets;

(ii) Propose an overall framework for low-altitude aerial image dataset construction, including data collection, data annotation and dataset usage;

(iii) Suggest the existing data augmentation techniques to expand the original data.

2. Related work

2.1. Computer vision standard datasets

In order to evaluate the performance of computer vision models, researchers organized some large-scale computer vision datasets as the criteria to test the performance of models. This section

introduces three commonly used classical datasets: VOC, ImageNet and MS COCO. These three datasets are authoritative evaluations in computer vision field.

VOC dataset is a classical dataset in computer vision field, which has a good image quality and complete annotations. It can be used for tasks such as image classification, object detection, image segmentation, personnel layout and action recognition. The VOC dataset has been used in Pascal VOC Challenge, which began in 2005 and ended in 2012. The VOC2005 dataset contains only 4 classes of objects: bicycles, cars, motorcycles and people. It is only suitable for classification and detection tasks. In 2007, the dataset was expanded to 20 classes, which can be clustered into 4 categories: people, animals, vehicles and household objects. VOC2007 contains 9963 images with 24,640 annotated objects. VOC2012 is the final version. There are 11,530 images in the training set and the verification set in VOC2012, including 27,450 annotated objects.

ImageNet dataset was used in ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which began in 2010 and ended in 2017. The main challenges of the competition include object localization, object detection, object detection from video and scene classification. ImageNet dataset contains 1419k images that are divided into 21,841 categories.

MS COCO dataset was designed for the challenge of the same name. MS COCO is one of the most popular and authoritative competitions in computer vision field. MS COCO was labeled by Microsoft in 2014. It consists of images with daily scenes containing common objects. Many objects in this dataset can only be identified by context, due to their small size or ambiguous appearance. To push research in contextual reasoning, the annotation information of images not only includes category and location information but also semantic text description of images. However, MS COCO does not focus on aerial view.

2.2. Available low-altitude aerial image datasets

The images in the above datasets generally do not have aerial view, so they cannot be used to train UAV vision tasks. In order to meet different UAV task requirements, researchers have constructed a few datasets that have aerial view images. These datasets generally focus on object detection, object tracking and action/event recognition tasks. Next, we will introduce 10 aerial image datasets in Table 1.

UAVDT [6] is a dataset produced by the University of the Chinese Academy of Sciences, in which only vehicles are annotated. The videos are taken by DJI Inspire 2, with a resolution of 1080 × 540 pixels. About 80,000 frames in UAVDT dataset are annotated over 2700 vehicles. The dataset is divided into two parts. One part is used for single object tracking and the other part is used for object detection and multi-object tracking. The annotations of the two parts are different. For single object tracking tasks, 8 attributes are annotated for each sequence, i.e., Background Clutter, Camera Rotation, Object Rotation, Small Object, Illumination Variation, Scale Variation and Large Occlusion. For multi-object tracking tasks, three attributes are annotated for each sequence, i.e., Flying Altitude, Camera View and Weather Condition. For object detection tasks, the other 3 attributes are annotated, i.e., Vehicle category, Vehicle occlusion and out-of-view.

VisDrone dataset [7] is a dataset used for the VisDrone Challenge. It was collected by the AISKYEYE team at Lab of Machine Learning and Data Mining, Tianjin University, China. VisDrone dataset covers a wide range of aspects including location (taken from 14 different cities in China), environment (urban and country), objects (pedestrians, vehicles, bicycles, etc.), and density (sparse and

crowded scenes). The data is collected by various drone platforms, i.e., DJI Mavic, Phantom series. These data are used for four tasks: object detection in images, object detection in videos, single-object tracking and multi-object tracking. VisDrone2019, used for the VisDrone2019 Challenge, consists of 288 video clips formed by 261,908 frames and 10,209 static images. The maximal resolutions of video clips and static images are 3840 × 2160 and 2000 × 1500, respectively.

Table 1. Aerial image datasets**.**

Datasets	Scale	Contents	Camera altitude	Maximum resolution
UAVDT[6]	80 k images	Vehicles on the road	10~70 m	1080 × 540
VisDrone[7]	10 k images, 288 videos	People and vehicles in daily scenes	N/A	Image:2000 × 1500 Video:3840 × 2160
Stanford Drone[8]	N/A	People and vehicles on the road	80 m	1400 × 1904
UAV123[9]	110 k images	Cars, trucks, ships, people, etc.	Set1:5~25 m Set2, Set3: N/A	1280 × 720
Okutama-Action[10]	77 k images	People performing various actions	10~45 m	3840 × 2160
VIRAT2.0[11] ground subset	N/A	People and vehicles in natural scenes	N/A	1920 × 1080
VIRAT2.0 aerial subset	N/A	People and vehicles in natural scenes	N/A	640 × 480
UCLA Aerial Event[12]	27 videos	People in outdoor scenes	25 m	N/A
Mini-drone video[13]	N/A	People and vehicles in a parking lot	N/A	1920×1080
CARPK[14]	1448 images	Vehicles in parking lots	40 m	N/A

Stanford Drone dataset [8] is provided by Computational Vision and Geometry Laboratory, Stanford University for human motion trajectory prediction in crowded scenes. This dataset consists of eight different scenarios: gates, little, nexus, coupa, bookstore, deathCircle, quad and hyang. The agents include pedestrians, cyclists, skateboarders, cars, buses and golf carts. Different types of agents are labeled with bounding boxes of different colors. The dataset is large in scale and contains 60 video sequences. In these videos, most of the agents are cyclists and pedestrians.

UAV123 [9] is a dataset for single object tracking, which consists of 123 video sequences with more than 110k frames. UAV123 dataset can be divided into three subsets. Set1 contains 103 sequences, and an unmanned aerial vehicle (DJI S1000) takes the video, with altitudes ranging from 5 m to 25 m. Set2 consists of 12 sequences, captured by a boardcam (with no image stabilization) mounted at a small, low-cost UAV. Due to the limitation of video transmission bandwidth, these sequences have low quality and resolution, and contain reasonable noise. Set3 contains eight sequences, which are synthetic videos captured by UAV simulators. Set1 is provided at 30 FPS. The annotation was done manually at 10 FPS and then linearly interpolated to 30 FPS.

Okutama-Action dataset [10] is a dataset for action recognition which is captured from UAVs (DJI Phantom 4) at a baseball field in Okutama, Japan. Okutama-Action dataset contains 43 video sequences with 77k frames in 4k resolution. Compared with the previous datasets, Okutama-Action

has a significant increase in sequence length, averaging 60 seconds per sequence. In each video, up to 9 actors sequentially perform a diverse set of actions. Due to the long length of the single video, this dataset can also be used for object detection. This dataset contains not only tracking labels for object tracking and single action labels for action recognition, but also multi-action labels for action recognition. In multi-action labels, one person can be annotated more than one label, such as the label of a person standing on the phone may be "standing, calling". This dataset can also be used for pedestrian detection tasks if all the information representing actions in the labels is artificially blocked. It should be noted that the actions in the video are performed by actors, that is to say, these videos are not taken by people in natural scenes, and the video background is relatively clean, so Okutama-Action has no advantages in naturalness and reality.

VIRAT2.0 dataset [11] supported by DARPA is also a dataset for action recognition. In April 2011, the first version of VIRAT dataset was released, which contains an annotated training subset and an unannotated test subset. Then in October 2011, VIRAT2.0 Ground Video Subset (captured by cameras stationed at the top of buildings) was released. There are total 12 event types annotated in Ground Video subset. These videos were captured from 11 different outdoor scenes. In January 2012, VIRAT2.0 Aerial Video Subset (captured by aerial vehicles) was released. VIRAT2.0 includes diverse types of human actions and human-vehicle interactions, with a large number of examples (>30) per action class. These videos were collected in natural scenes showing people performing normal actions in standard contexts. The backgrounds are uncontrolled and cluttered, and there are frequent incidental movers. These video sequences which include different camera angles and resolutions are shot at several different sites, and the actions are performed by different people. Compared with the existing action recognition dataset, VIRAT is realistic, natural and challenging in resolving power, background clutter, scene diversity and human activity categories. So far, the Aerial Video Subset has not been annotated.

UCLA Aerial Event dataset [12] is a dataset used for joint inference of groups, events and human roles in aerial videos. The videos are captured by a GoPro stationed on a low-cost hex-rotor. During the shooting, the hex-rotor was flying at an altitude of 25 meters. Typically, the size of a person is only 15 × 15 pixels in a frame. There are 27 video sequences in the dataset, ranging in length from 2 to 5 minutes. These videos are captured at a park where the terrain is interesting: hiking routes, parking lots, camping sites, picnic areas with shelters, restrooms, tables, trash bins and BBQ ovens. The annotation in the dataset includes individuals, objects, groups, events, human roles and goals (destinations). There are 12 events, 18 human roles and 12 object categories. Detecting/tracking humans and objects in the videos can recognize some events recognized, such as BBQ, queuing, exchanging objects, loading/unloading.

Mini-drone video dataset [13] is used to explore privacy protection in UAV surveillance. It was established by the Multimedia Signal Processing Group, EPFL. Safety issues need to be carefully considered when holding major events. Considering the difficulty of establishing a complete surveillance system, drone-based surveillance is particularly advantageous. However, researchers have observed that UAV surveillance may affect visual privacy. To analyze these surveillance devices, Bonetto et al. established the Mini-drone video dataset. This dataset consists of 38 different contents captured in 1920 × 1080 resolution and the duration of each content is 16 to 24 seconds. The videos were taken in a parking lot with the mini-drone Phantom 2 Vision+. The behaviors in this dataset can be clustered into three categories: normal, suspicious, and illicit behaviors. Normal behaviors include walking, getting in their cars and parking. In suspicious content, there is no priori

wrong, but people act in a questionable way. Illicit behaviors include mis-parking vehicles, stealing items and cars, or fighting. Mini-drone video dataset was annotated using the open source ViPER-GT tool, and the annotations include body silhouette, facial region, accessories, vehicles, license plates, and video capture. This dataset was originally used for privacy protection research, but considering the content of the video sequences, it is possible to extend this dataset to other Computer Vision tasks such as detection and action recognition by re-annotation.

CARPK dataset [14] is used for vehicle counting, which is produced by National Taiwan University. The data is taken by DJI PHANTOM 3 Pro with a shooting height about 40m in four different parking lots in Taipei, including National Taiwan University, in front of Chiang Kai-shek Memorial Hall, behind Chiang Kai-shek Memorial Hall and Taipei Zoo. The CARPK dataset contains 89,777 car information, in which the maximum number of vehicles in a single scene is 188. The dataset has four different scenarios and uses bounding boxes to locate the object accurately. Although the CARPK is a dataset for counting, it is possible to extend it to object detection tasks.

3. Construction of low-altitude aerial image datasets

3.1. Data collection

Mobile cameras equipped on UAVs have a unique perspective. The images taken by these cameras are significantly different from those taken by ordinary cameras. The characteristics of low-altitude aerial images must be taken into account when constructing a low-altitude aerial image dataset. Generally speaking, low-altitude aerial images have the following characteristics.

1) Small objects. Due to the high shooting height, the objects in the images are usually very small, which is a main character of UAV-view images and brings difficulties to object detection. At present, there is no clear definition of the object size or the proportion of small objects in the whole picture. In VIRAT dataset supported by DARPA, the human height is 10–200 pixels or the height proportion of human in in video is 2–20% [11]. In DAC-SDC dataset [16], the size of many small objects is 1–2% of the images. However, in standard datasets, the size of objects is usually larger. Such as, in VOC dataset, the average size of objects is 20% of the images.

2) Objects high-density. UAV cameras have a wide view that leads to a large number of objects [6].

3) Wide view. Since the camera is far away from the ground, aerial images generally have a wide field of vision.

4) Overlooking view. Unlike general images, aerial images have an overlooking view, which is also a challenge for UAV vision tasks.

5) Light conditions and shading. Aerial image datasets usually contain attributes such as light conditions and occlusion. Therefore, it is necessary to take light conditions and occlusion into account when collecting aerial images.

Usually, the images used to construct the image datasets can be obtained from the internet. The data of VOC dataset comes from Flickr [3]. The data of UCF series datasets, HMDB-51 dataset and AVA dataset are all from YouTube. And some images of the Military Vehicle Detection dataset [15] published by Kamran are also from YouTube. Generally, web crawlers can help to get pictures from the internet quickly. However, the low-altitude aerial image datasets have specific requirements on the images and videos, so it is difficult to collect enough data from the network. In order to construct

a low-altitude aerial image dataset, it is inevitable to use UAVs to capture images or videos. Due to the different task requirements for each dataset, the shooting scene and image content also exist differences. Okutama-Action, for example, was shot at a baseball field in Okutama with a clean background. The VisDrone dataset was gathered in natural scenes in 14 cities, making it more natural and realistic. When shooting aerial images, the height and the tilt angle of the camera and the light conditions must be taken into account according to the task requirements. In addition, a few datasets consist of non-real images. For example, the Set3 of UAV123 dataset contains 8 synthetic sequences captured by UAV simulators. In the Military Vehicle Detection datasets, 11,733 toy vehicle images are generated from RC Military Toy YouTube Channel.

3.2. Data annotation

The common image annotation tools include LabelImg, LabelMe and YOLO-Mark. LabelImg and LabelMe both originated from CSAIL, MIT. They are all written in Python and use Qt as their graphical interface. LabelImg can annotate objects with rectangular bounding boxes. It can be used to make datasets needed by Faster R-CNN, YOLO, SSD and other object detection networks. Its annotation format can be YOLO format or VOC format. LabelMe can annotate images in various shapes, including polygons, circles, rectangles, lines and points. LabelMe can produce datasets for a variety of vision tasks, including instance segmentation, semantic segmentation, object detection and image classification. YOLO-Mark can produce YOLO format dataset for object detection task. It is designed for YOLO series network specially by YOLO team. In addition, there are other image annotation tools, such as Sloth, Annotorious and RectLabel for object detection, Pixel Annotation Tool, Semantic Segmentation Editor and Image Annotation Tool for image segmentation.

The common video annotation tools are VATIC [17] and VoTT. VATIC is an open source video annotation tool for computer vision research that crowdsources work to Amazon's Mechanical Turk (Mturk). For an input video, VATIC can automatically extract the annotation tasks and integrate with Mturk. After all the frames of a video are annotated, the annotated frames can be synthesized into a complete video using FFmpeg. VATIC can export the annotation tool in Pascal VOC format. Some aerial datasets, such as UAVDT and Okutama-Action, are labeled with VATIC. VoTT is a visual annotation tool released by Microsoft. It is developed based on JavaScript and can run across Windows and Linux platforms. It can annotate both images and videos. VoTT also has the function of target tracking. It uses a tracking algorithm to assist computers to track and annotate objects in videos. It can export annotated data in various formats such as CNTK, VOC and YOLO. In addition, it also provides a trained Faster R-CNN model, which can be used to automatically annotate before manual correction. In addition, there are some other video annotation tools, such as video-labeler and CVAT.

Large-scale deep networks need huge training datasets. In order to give UAVs computer vision capability, we need to produce large enough aerial view image datasets. Data annotation is very arduous, especially for a large dataset. The UAVDT dataset has 840,000 annotation boxes, which annotate by more than 10 professionals using VATIC. The whole annotation work lasted for two months [6].

One way to solve this problem is to release the annotation task on crowdsourcing platform. There are many available image annotation platforms, such as MTurk. Sorokin et al. show that image annotations can be effectively outsourced to MTurk. Doing so has produced annotations in quite

large numbers relatively cheaply. These annotations are of good quality and can be checked and controlled. Many annotations of large image datasets are annotated on MTurk, such as ImageNet, MS COCO, LabelMe. In addition, the Ground Video Subset of the VIRAT2.0 is also annotated in MTurk. Except MTurk, there are other image annotation platforms, such as Jingdong Zhongzhi, Baidu Zhongce, Figure Eight, MicroWorkers etc.

Another possible solution is to use the hybrid annotation. The hybrid annotation method relies on the trained object detector. (Atomic Visual Actions) AVA dataset is a dataset for action recognition with 1.58M action labels. Because of the huge workload of labeling tasks, the mixed labeling method was adopted [18]. Firstly, Faster-RCNN person detector is used to generate a set of initial bounding boxes. Then, an annotation tool is used to annotate the remaining bounding boxes missing from the detector. In the next manual annotation process, the detector only missed 5% of the bounding boxes, which indicated that the hybrid annotation method was feasible. This hybrid method ensures the accuracy of annotations while minimizing the workload and time cost of manual annotations. Object detector has proved to be useful in the construction of AVA dataset. It is possible that this method can also be applied to build object detection datasets.

3.3. Dataset partition strategy

Datasets are generally divided into three parts: training set for model training, verification set for validation and test set for model testing. Generally, testing set accounts for about 20% of the total size of dataset, and training set plus validation set accounts for about 80% of the total size of dataset.

In VOC2005, the training-validation set provided 422 images containing 1215 segmentation objects, and the test set contains 210 images with 607 objects. In VOC2012, the training-validation set contains 11,530 images, but the test set has not been released. MS COCO dataset, released in 2015, contains 165,482 training images, 81,208 validation images and 81,434 test images (about 50% for training, 25% for validation and 25% for testing) [5].

When capturing aerial images or videos, the same scene may be taken by more than one UAVs. In order to ensure the objectivity of the test results, it is necessary to ensure that different sequences of the same scene are located in the same subset. In Okutama-Action dataset, the train-validation set consists of 33 video sequences, accounting for 77% of the total dataset. And the test set consists of 10 video sequences, accounting for 23% of the total dataset [10]. It is noteworthy that Okutama-Action dataset contains 22 different scenes, of which 21 scenes were shot simultaneously by two UAVs which have different perspectives. Therefore, in order to ensure that the test set is completely unseen to the model, the sequences captured from the same scene must be in the same set when splitting the dataset.

4. Application of data augmentation

Low-altitude aerial view datasets are limited in the data source, which may result in a small size of the datasets. Data augmentation can effectively expand the scale of training data and enhance the generalization performance of the model. Traditional data augmentation techniques make some simple operations on original training data to generate new images, such as flipping, clipping, color space transformation, adding reasonable noise, etc. These data augmentation methods are easy to implement. Recently, due to its good performance in image generation, Generative Adversarial Nets

(GAN) have attracted more and more attention. It can generate images according to specific requirements or even a text description. Some works show that oversampling and GAN can improve the performance of small object detection [19, 20].

4.1. Traditional data augmentation

Geometric transformation. Geometric transformation is the earliest data augmentation technique [21]. Compared with GAN augmentation method, the computation cost of geometric transformation is extremely low. The dataset can be expanded *n* times by geometric transformation. Geometric transformations commonly used include flipping, rotation, cropping, distorting, scaling, translation, etc. Flipping and rotation are the simplest augmentation methods. Flipping has been proven to be useful on CIFAR-10 and ImageNet. But Shorten et al. [18] believed that there is a "security" issue, that is, the use of flipping and rotation on datasets involving text recognition may lead to label changes. Aerial images sometimes involve license plates, so we have to take the issue of label changes into account when using flipping and rotation techniques. Researchers have validated the effectiveness of geometric transformations. Chatfield et al. [22] discussed two kinds of data augmentation techniques on VOC datasets. The first strategy is flipping augmentation, and the second strategy is C+F, which is the combination of clipping and flipping. The results show that flipping can only bring a slightly improvement compared with no augmentation, while the improvement is about 2–3% using C+F augmentation. Mash et al. [23] evaluated clipping, rotation, scaling, occlusion and the combination of these methods on a dataset for aircraft classification tasks. The results showed that the combination of occlusion and clipping had the best performance and the test set classification accuracy was improved by 9.1%. Taylor et al. [24] evaluated three geometric transformation methods of flipping, rotation and clipping, and found that the clipping transformation had the best performance.

Color space transformation. Simple color space transformations include isolating a single color channel such as R, G, or B. An image can be transformed into its representation in one color channel by isolating that matrix and adding 2 zero matrices from the other color channels [21]. Unlike geometric transformation, color space transformation changes the colors of the images instead of the shape of objects. That is to say, the relative position of objects and the position of bounding boxes in images are not changed. Therefore, color space transformation will not lead to label changes in UAV object detection and tracking tasks.

Noise injection. When there are a lot of useless features in the training datasets that are not helpful to train models, it may lead to over-fitting of the models. Appropriate noise addition to these datasets can enhance the generalization performances of models and has little impact on the accuracy. Barea et al. [25] had demonstrated the effectiveness of noise injection techniques on 9 benchmark datasets that are taken from the UCI database and PROBEN1 benchmark set.

In addition to the above methods, there are other data augmentation methods, including pixel erasure, changing image brightness [26], sharpness, definition and contrast. Normally, traditional data augmentation can be implemented by simple algorithms, and the costs of calculations are low.

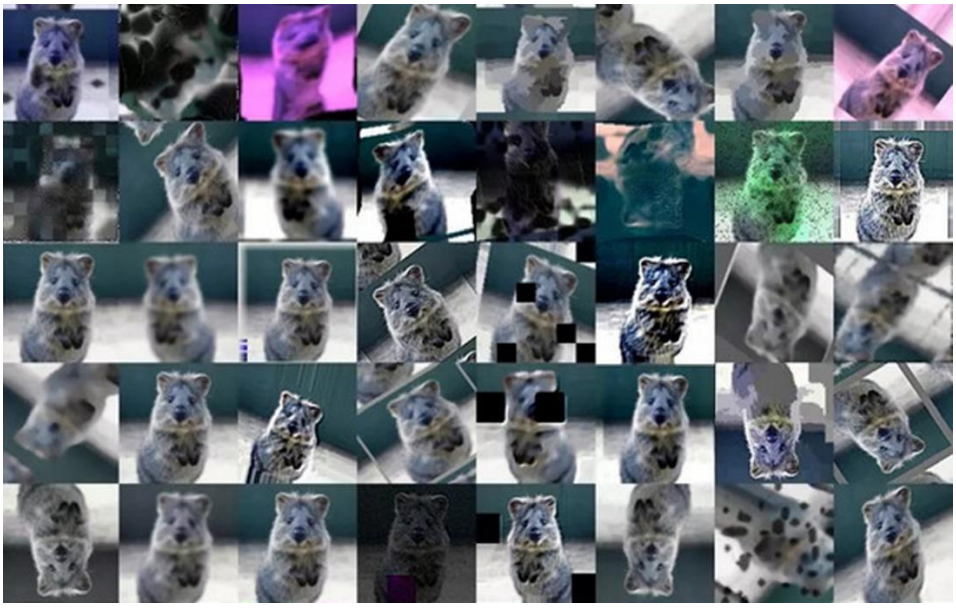

Figure 1. Examples of traditional data augmentation [26].

4.2. Data augmentation based on oversampling techniques

We may encounter the problem of imbalance of sample categories when constructing low-altitude aerial image datasets. This problem can be solved by sampling strategies. This section introduces three oversampling techniques: SMOTE, sample pairing and mixup.

1) SMOTE. SMOTE (Synthetic Minority Over-sampling Technique) [27] is a minority over-sampling method, which synthesizes samples through a synthesis technique to achieve the balance of data categories. SMOTE generates synthetic samples for categories with fewer samples by operating in "feature space" rather than "data space". Firstly, the difference between the feature vector and its nearest neighbor is taken. Then, multiply the difference by a random number between 0 and 1 and add it to the feature vector considered, which will result in a random selection of a point along the line between two specific features. Repetition of the above steps can balance the number of large and small samples. Unlike general oversampling techniques, SMOTE can avoid duplicate samples in the process of sample synthesis.

2) Sample pairing. In 2018 Inoue [28] proposed a sample pairing data expansion technique, which creates a new sample from an image by superimposing another image randomly selected from the training data (i.e., simply taking the average of two images for each pixel), and uses the label of the first sample as the correct label for the mixed sample. This method has been proved to significantly improve the accuracy of image classification tasks on ILSVRC 2012, CIFAR-10, CIFAR-100 and other datasets.

3) Mixup. In 2018, Zhang et al. [29] proposed a method called mixup, which is a new method of data augmentation. This method uses linear interpolation to generate extended data. The formula is as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \tilde{y} = \lambda y_i + (1-\lambda)y_j \quad (1)$$

Where (x_i, y_i) and (x_j, y_j) are two samples randomly extracted from training data, and $\lambda \in [0,1]$.

Experiments show that mixup can improve the accuracy of image classification tasks on CIFAR-10, CIFAR-100 and ImageNet-2012.

4.3. Data augmentation based on GAN

In 2014, Goodfellow et al. [30] proposed the concept of Generative Adversarial Nets. A GAN consists of two parts: the generator *G* and the discriminator *D*. The goal of *G* is to generate a sample close to real data according to prior distribution to fool the discriminator, while the goal of *D* is to separate the sample generated by *G* from the real sample as far as possible. This framework corresponds to a minimax two-player game. The game process can be expressed as follows:

$$\min_G \max_D (D, G) = E_{x \sim p_{data}(x)}[logD(x)] + E_{z \sim p_z(z)}[log(1 - D(G(z)))] \tag{2}$$

GAN has powerful generating ability, especially in the field of image generation. So it is a promising application to enhance data by using generative adversarial network.

Goodfellow et al. carried out experiments on MNIST, Toronto Face Database (TFD) and CIFAR-10. Although they did not claim that the samples generated by GAN were better than those generated by existing methods, they believed that these samples could at least be compared with better generation models. It has the same competitiveness, and GAN has more potential.

Since the concept of GAN was first proposed in 2014, GAN has attracted wide attention, and there are many variants of GAN now. Conditional Generative Adversarial Nets (CGAN) [31] can train networks with different categories of images and control generators to generate an image with specific categories. CGAN inputs a one-hot vector *y* as an additional vector of random noise vector *z* to generators and discriminators to control the categories of generated images, where *y* can be any type of auxiliary information, such as class labels or data from other patterns.

Cycle-Consistent Adversarial Networks (Cycle GAN) [32] has good performances in style transfer. It can capture the special features of an image set and learn how to transfer these features to other image sets. The training images generated by Cycle GAN have different color, density and light conditions from the original image. Liu et al. used three object detection models, SSD, YOLOv3 and Faster R-CNN, to carry out experiments on a brain slices microscopic dataset. The results show that Cycle GAN data augmentation can effectively improve the detection performance of the three detection models [20].

Style-Based Generative Adversarial Networks (Style GAN) [33] allows more linear and less entanglement representation of different change factors. So Style GAN could control image synthesis by modifying the style in a specific proportion. Style GAN can be used for style transfer and has achieved good performance in face image generation.

GAN can also be combined with other techniques. Wei et al. [34] combined foreground-background separation model with a GAN and proposed a data augmentation method based on the foreground-background separation model to enhance the performance of object detection in DSSD model. This method does not change the main network of DSSD but only adds a GAN to the training process of DSSD. The whole network consists of two training stages. In the first stage, the foreground-background separation model is realized and the object detection model is pre-trained. In the second stage, data augmentation is used to assist the object detection training.

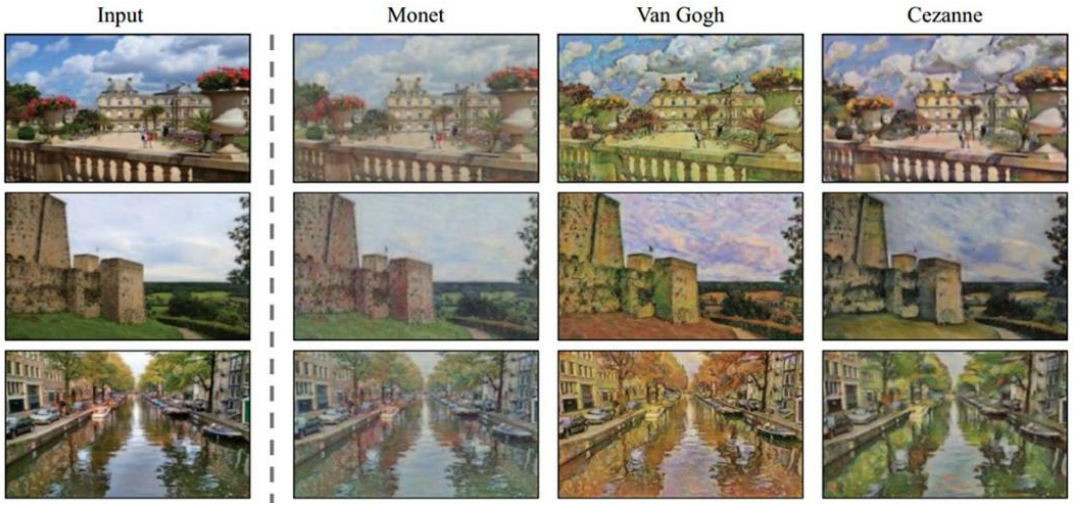

Figure 2. Using a collection of paintings of a famous artist, learn to render a user's photograph into their style [32].

5. Conclusions

In this paper, taking UAVs computer vision capability as an example, we propose an overall framework for dataset construction for the training of deep learning, including data collection, data annotation and dataset usage. Our work may provide a guidance on how to construct an appropriate dataset for a deep learning model. In addition, for UAV vision task, a huge challenge is that it is difficult to obtain enough appearance information due to the small size of the object. Therefore, exploring more effective object detection methods for small objects is an important way to promote UAV vision technologies in the future.

Acknowledgments

This work is supported by National Key R&D Program of China (No. 2019YFB2101700).

Conflict of Interests

The author declares no conflict of interests.

References

1. J. M. Peña, A. I. Castro, J. Torres–Sánchez, D. Andújar, C. S. Martín, J. Dorado, et al., Estimating tree height and biomass of a poplar plantation with image-based UAV technology, *AIMS Agric. Food*, **3** (2018), 313–326.
2. S. Chen, Y. Zhang, Y. Zhang, J. Yu, Y. Zhu, Embedded system for road damage detection by deep convolutional neural network, *Math. Biosci. Eng.*, **16** (2019), 7982–7994.
3. M. Everingham, L. V. Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, *Int. J. Comput. Vision*, **88** (2010), 303–338.

4. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei, ImageNet: A large–scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255.
5. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft coco: Common objects in context, in *European Conference on Computer Vision*, Springer, (2014), 740–755.
6. D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, et al., The unmanned aerial vehicle benchmark: Object detection and tracking, in *European Conference on Computer Vision*, Springer, (2018), 375–391.
7. P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, Vision meets drones: A challenge, preprint, arXiv:1804.07437.
8. A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: Human trajectory understanding in crowded scenes, in *European Conference on Computer Vision*, Springer, (2016), 549–565.
9. M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for UAV tracking, in *European Conference on Computer Vision*, Springer, (2016), 445–461.
10. M. Barekatain, M. Marti, H. Shih, S. Murray, K. Nakayama, Y. Matsuo, et al., Okutama-action: An aerial view video dataset for concurrent human action detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2017), 2153–2160.
11. S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, et al., A large–scale benchmark dataset for event recognition in surveillance video, in *2011 IEEE Conference on Computer Vision and Pattern Recognition*, (2011), 3153–3160.
12. T. Shu, D. Xie, B. Rothrock, S. Todorovic, S. C. Zhu, Joint inference of groups, events and human roles in aerial videos, in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 4576–4584.
13. M. Bonetto, P. Korshunov, G. Ramponi, T. Ebrahimi, Privacy in mini–drone based video surveillance, in *2015 IEEE International Conference on Automatic Face Gesture Recognition*, (2015), 1–6.
14. M. Hsieh, Y. Lin, W. H. Hsu, Drone–based object counting by spatially regularized regional proposal network, in *2017 IEEE International Conference on Computer Vision*, (2017), 4165–4173.
15. F. Kamran, M. Shahzad, F. Shafait, Automated military vehicle detection from low-altitude aerial images, in *2018 Digital Image Computing: Techniques and Applications*, (2018), 1–8.
16. X. Xu, X. Zhang, B. Yu, X. S. Hu, C. Rowen, J. Hu, et al., DAC-SDC low power object detection challenge for UAV applications, preprint, arXiv:1809.00110.
17. C. Vondrick, D. Patterson, D. Ramanan, Efficiently scaling up crowdsourced video annotation, *Int. J. Comput. Vision*, **101** (2013), 184–204.
18. C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, et al., Ava: A video dataset of spatio–temporally localized atomic visual actions, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 6047–6056.
19. M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, in *9th International Conference on Advances in Computing and Information Technology*, 2019.
20. W. Liu, L. Cheng, D. Meng, Brain slices microscopic detection using simplified SSD with Cycle–GAN data augmentation, in *International Conference on Neural Information Processing*, Springer, (2018), 454–463.

21. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data*, **6** (2019), 1–48.
22. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, preprint, arXiv:1405.3531.
23. R. Mash, B. Borghetti, J. Pecarina, Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks, in *International Symposium on Visual Computing*, Springer, (2016), 113–122.
24. L. Taylor, G. Nitschke, Improving deep learning using generic data augmentation, preprint, arXiv:1708.06020.
25. F. J. Morenobarea, F. Strazzera, J. M. Jerez, D. Urda, L. Franco, Forward noise adjustment scheme for data augmentation, in *2018 IEEE Symposium Series on Computational Intelligence*, (2018), 728–734.
26. L. Hu, *The Quest for Machine Learning*, 1st edition, Posts and Telecommunications Press, Beijing, 2018.
27. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over–sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357.
28. H. Inoue, Data augmentation by pairing samples for images classification, preprint, arXiv:1801.02929.
29. H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, preprint, arXiv:1710.09412.
30. I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, *Adv. Neural Inf. Process. Syst.*, **27** (2014), 2672–2680.
31. U. Shaham, Y. Yamada, S. Negahban, Conditional generative adversarial nets, preprint, arXiv:1411.1784.
32. J. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *2017 IEEE International Conference on Computer Vision*, (2017), 2242–2251.
33. T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in *2019 IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 4401–4410.
34. W. Jiang, N. Ying, Improve object detection by data enhancement based on generative adversarial nets, preprint, arXiv:1903.01716.

AIMS Press