Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 6195189, 11 pages
https://doi.org/10.1155/2020/6195189

Research Article

Conceptual Cognitive Modeling for Fine-Grained Annotation Quality Assessment of Object Detection Datasets

Lei Guo,[1] **Xinying Xu**,[2] **Gang Xie**,[2,3,4] **and Jerry Gao**[2,5]

[1]*College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China*
[2]*College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China*
[3]*School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China*
[4]*Shanxi Key Laboratory of Advanced Control and Intelligent Information System, Taiyuan University of Science and Technology, Taiyuan 030024, China*
[5]*Department of Computer Engineering, San Jose State University, San Jose, CA, USA*

Correspondence should be addressed to Gang Xie; xiegang@tyut.edu.cn and Jerry Gao; jerry.gao@sjsu.edu

Received 29 February 2020; Accepted 8 April 2020; Published 5 May 2020

Guest Editor: Jianbiao Zhang

In many supervised computer vision tasks such as object detection, manual annotation crowdsourcing platforms are widely used for acquiring large-scale labeled data. However, the annotation quality may suffer low quality that can severely affect the training of models. As a result, the evaluation of the annotations within the dataset is critical, yet it has seldom been addressed in object detection. In this paper, we present a fine-grained annotation quality assessment (FGAQA) framework for evaluating the quality of object detection datasets. First, we formulate a generic annotation quality assessment framework based on the core general-purpose data quality dimensions, using the bounding box and the label. Second, cognition theory in terms of hierarchy and continuity is utilized to refine the basic framework, including the consistency of the bounding box, completeness of the category, hierarchical accuracy of the label, and the consistency of the label. Comprehensive experiments on the two object detection datasets are used for performance evaluation. It is found that the ground truth annotations of the Urban Traffic Surveillance dataset have more quality issues than the ones of the PASCAL VOC 2007 detection dataset. The proposed FGAQA framework performs an effective fine-grained evaluation of the annotations, which is significant for quality assurance of annotations from crowdsourcing platforms and the subsequent model's training.

1. Introduction

In supervised learning, annotation quality plays a vital role in training and assessment of the models for several computer vision tasks such as object classification [1, 2], detection [3–6], and segmentation [7–9]. The training of object detection models relies on accurate and sufficient annotations. For large-scale object detection datasets, annotations are usually obtained through crowdsourcing platforms, which results from anonymous participants, and can be collected for efficiency [10–12]. However, due mainly to the untrained participants involved in the professional and time-consuming annotation tasks, this has inevitably led to subjective inconsistency and relatively low quality of the collected annotations. As a result, the annotation quality cannot be guaranteed, where the quality assessment of such annotations becomes a challenge in this context.

Annotation quality in object detection is a specialized-purpose data quality problem. Data quality has been widely studied since the 1980s [13]. According to [14], data quality can be defined as the degree to which a set of characteristics of data fulfills the requirements. Data with high quality should represent the real-world entities accurately in the structure and fit for their intended uses. Besides, data quality is of multidimensional characteristics. By reviewing the related literature [14–19], a core set of data quality

dimensions is defined, including the completeness, accuracy, and consistency. Moreover, there are a fair number of researches about annotation quality. Regarding the annotation quality in classification, accuracy is employed generally [20], not considering the hierarchy of categories. For annotation quality in object detection, quality is evaluated by Intersection-over-Union (IoU) [21]. IoU is the ratio of the intersection area of the ground truth and human annotation to the total area, only considering the quality of the bounding box [22]. There are few systematic researches about annotation quality of object detection. Consequently, we refer to general-purpose data quality and construct an annotation quality framework.

To date, there are relatively few works reported on this topic. This is only addressed from the perspectives of the object category and IoU [21]. However, a few general-purpose metrics can also be applied for annotation quality assessment. And we should perform annotation quality assessment from various aspects of the two attributes: bounding box and label.

Evaluation measures for object classification, detection, and segmentation could serve as a reference for annotation quality in object detection. Regarding flat object classification, precision and recall are employed to assess the performance [23–26]. As for hierarchical object classification, distance in the tree or the directed acyclic graph (DAG) is used to assess the performance [27–30]. The distance can treat the prediction errors differently. In terms of object detection, the mAP is usually employed [31–36], integrating precision, recall, and IOU. The mAP is calculated according to the predicted results and confidence scores. However, for annotations, reasonable confidence scores are hard to obtain. As a result, in this paper, we employ the metrics of precision and recall. Regarding object segmentation, evaluation measures can be categorized into three types: area-based measures, location-based measures, and combined measures [37–41]. These image segmentation measures pay more attention to the details and the intrinsic visual characteristics. Consequently, the idea of image segmentation evaluation is introduced into the annotation quality assessment framework.

In this paper, we propose a fine-grained framework for annotation quality assessment of object detection datasets, containing three dimensions: accuracy, completeness, and consistency. First, we construct the basic quality assessment framework based on the core general-purpose data quality (DQ) measurement, including accuracy and completeness, which considers the characteristics of annotation. For consistency, we find that it is difficult to give a strict definition. Further, the relationship of classes should be considered. Previous literature indicates that the cognition of humans is hierarchical in concept [42, 43] and consistent in space-time representations [44–46]. Inspired by these observations, the consistency of bounding box, completeness of category, hierarchical accuracy of label, and consistency of label are extracted as four additional elements for annotation quality assessment. The main contributions of this paper are as follows:

(1) We present a fine-grained annotation quality assessment (FGAQA) framework for evaluating the quality of object detection datasets. By analyzing the characteristics of the attributes of the bounding box and the corresponding label, the annotation quality contains three dimensions: accuracy, completeness, and consistency.

(2) To tackle the limitations of the basic quality assessment framework, we introduce the theory of cognitive perception to analyze the annotation quality and add four elements of annotation quality, including the consistency of bounding box, completeness of category, hierarchical accuracy of the label, and consistency of label. Specifically, the hierarchical accuracy of the label can treat annotation errors distinctively and softly.

(3) Comprehensive case studies on the Urban Traffic Surveillance (UTS) dataset and the PASCAL VOC 2007 detection dataset verify the effectiveness of the proposed annotation quality assessment framework. We find that the ground truth annotations of the UTS dataset have more quality issues, compared to the ones of the PASCAL VOC 2007 detection dataset.

The rest of this paper is organized as follows. In Section 2, the proposed cognitive-driven FGAQA framework is presented in detail. Section 3 discusses experiments as two case studies on the UTS and PASCAL VOC datasets. Finally, concluding remarks and future work are given in Section 4.

2. Annotation Quality Assessment Framework

A novel annotation quality assessment framework in object detection is given in this section, which is shown in Figure 1. The annotation has two attributes: bounding box and label. Annotation quality depends on its characteristics. For the bounding box, the size, location, and quantity could have some quality issues. Regarding the label, there may exist the quality problems of value and quantity. And the annotation quality serves reference for the training of the object detection model. Therefore, we define the quality dimensions according to the quality problems and the use of annotation. Inspired by some existing work [14–19], the dimensions of completeness, accuracy, and consistency are selected as the core set of the data quality dimensions. By considering the theory of cognitive perception, we redefine some elements based on annotation characteristics. As a result, a fine-grained annotation quality assessment framework is proposed, as shown in Figure 1. The framework is constructed from the views of the bounding box and label. Regarding the quality of the bounding box, completeness, accuracy, and consistency are defined. The completeness of the bounding box can be divided into the completeness of the bounding box's quantity and the completeness of the bounding box's size. In terms of the quality of the label, we define completeness, accuracy, and consistency. The completeness of the label consists of the completeness of the bounding box's label and the completeness of the category. The accuracy of

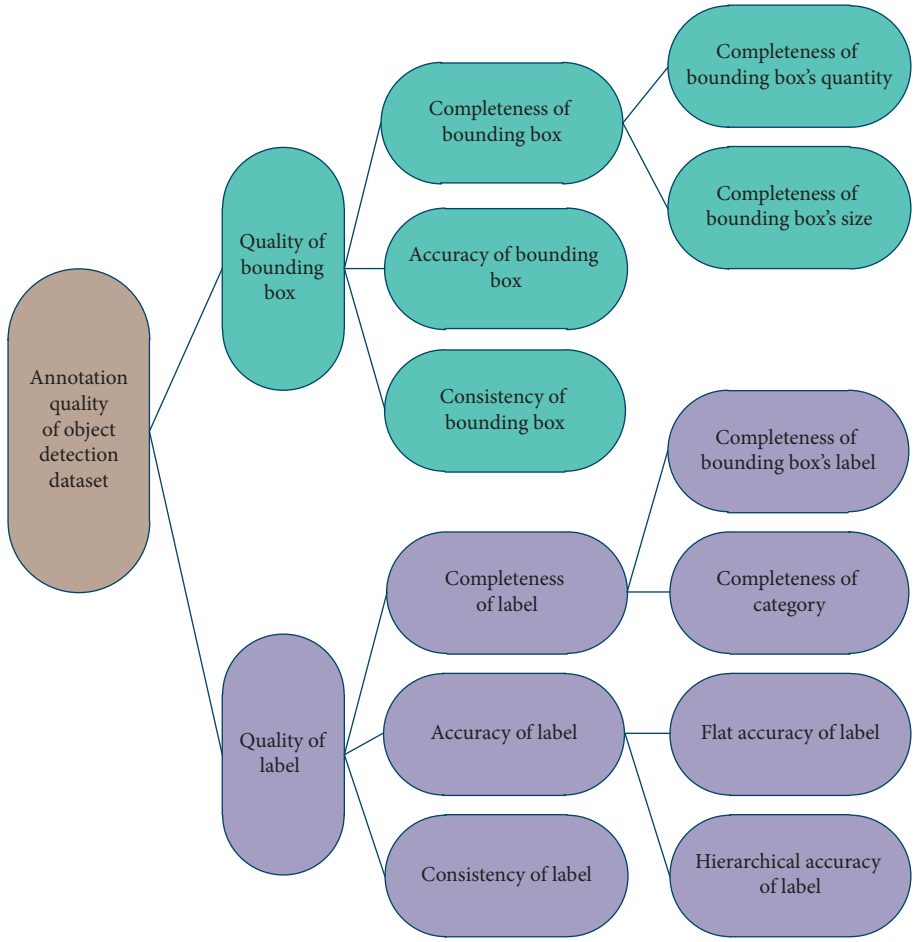

Figure 1: Annotation quality evaluation framework.

the label contains flat and hierarchical accuracy. And most of these dimensions are computed for every object and are averaged for an image and the total dataset.

2.1. Annotation Quality of Bounding Box's Quantity

2.1.1. Completeness of Bounding Box. The dimension can be defined as the extent to which bounding boxes are of sufficient quantity and coverage degree for the object. The dimension of completeness focuses on the null values. As for the completeness of the bounding box's quantity, the null values correspond to unannotated objects. In an object detection dataset, small objects are often be neglected. During the modeling process of object detection, the unannotated objects would be regarded as background. For the completeness of the bounding box's size, the null values correspond to the uncovered areas of the bounding boxes.

(1) Completeness of bounding box's quantity: for image i, completeness of bounding box's quantity is a metric that can be defined as follows:

$$\mathrm{CB}_i^{\mathrm{Quantity}} = \frac{n_i^{\mathrm{Hu}}}{n_i}, \tag{1}$$

where n_i is the true object number and n_i^{Hu} is the number of human annotations, namely, the number of bounding boxes. For the dataset, $\mathrm{CB}^{\mathrm{Quantity}}$ is

$$\mathrm{CB}^{\mathrm{Quantity}} = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{CB}_i^{\mathrm{Quantity}}}{N}, \tag{2}$$

where N is the number of images in the dataset.

(2) Completeness of bounding box's size: the completeness of the bounding box's size is a pixel-count-based metric and can be defined as follows. For the j^{th} object in image i, the metric is

$$\mathrm{CB}_{ij}^{\mathrm{Size}} = \frac{S_{ij}^{\mathrm{Int}}}{S_{ij}^{\mathrm{Obj}}}, \tag{3}$$

where S_{ij}^{Int} is the intersection area of the object and bounding box, and S_{ij}^{Obj} is the area of the object. For image i, $\mathrm{CB}_i^{\mathrm{Size}}$ is

$$\mathrm{CB}_i^{\mathrm{Size}} = \frac{\sum_{j=1,\ldots,n_i^{\mathrm{Hu}}}^{n_i^{\mathrm{Hu}}} \mathrm{CB}_{ij}^{\mathrm{Size}}}{n_i^{\mathrm{Hu}}}. \tag{4}$$

For the dataset, $\mathrm{CB}^{\mathrm{Size}}$ is

$$\mathrm{CB}^{\mathrm{Size}} = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{CB}_i^{\mathrm{Size}}}{N}. \tag{5}$$

2.1.2. Accuracy of Bounding Box. The dimension is intended to measure the closeness of the bounding box to the object. When the accuracy is low, the bounding box contains too much background affecting the distinction between the object and the background. For the bounding box of j^{th} object in image i, the accuracy is

$$\mathrm{Acc}\, B_{ij} = \frac{S_{ij}^{\mathrm{Int}}}{S_{ij}^{\mathrm{BB}}}, \tag{6}$$

where S_{ij}^{BB} is the area of the bounding box. In image i, the accuracy is

$$\mathrm{Acc}\, B_i = \frac{\sum_{j=1,\ldots,n_i^{\mathrm{Hu}}}^{n_i^{\mathrm{Hu}}} \mathrm{Acc}\, \mathrm{B}_{ij}}{n_i^{\mathrm{Hu}}}. \tag{7}$$

For a dataset, the accuracy can be given as follows:

$$\mathrm{Acc}\, B = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{Acc}\, B_i}{\sum_{i=1,\ldots,N}^{N} n_i^{\mathrm{Hu}}}. \tag{8}$$

2.1.3. Consistency of Bounding Box. The dimension focuses on the violation of spatiotemporal continuity of size and location. In crowdsourcing platforms, bounding boxes in adjacent frames may be drawn by different workers. As a result, they could conflict in size and location. Faced with the case, we can perform a quality assessment of the consistency of the bounding box during the corresponding postprocessing. Afterward, the annotations would satisfy the constraints. Concretely, for example, if an object moves toward the camera parallelly, the constraints are as follows:

$$\begin{cases} x_{\mathrm{center}}^{\mathrm{previous}} \approx x_{\mathrm{center}}^{\mathrm{current}} \approx x_{\mathrm{center}}^{\mathrm{next}}, \\ y_{\mathrm{center}}^{\mathrm{previous}} \le y_{\mathrm{center}}^{\mathrm{current}} \le y_{\mathrm{center}}^{\mathrm{next}}, \\ w^{\mathrm{previous}} \le w^{\mathrm{current}} \le w^{\mathrm{next}}, \\ h^{\mathrm{previous}} \le h^{\mathrm{current}} \le h^{\mathrm{next}}, \end{cases} \tag{9}$$

where x_{center} and y_{center} are the coordinates for the center of the bounding box, and w and h are the width and height of the bounding box. When the j^{th} object in image i satisfies the constraints, the metric $\mathrm{Con}\, B_{ij} = 1$. Otherwise, $\mathrm{Con}\, B_{ij} = 0$. For image i, the consistency is

$$\mathrm{Con}\, B_i = \frac{\sum_{j=1,\ldots,n_i^{\mathrm{Hu}}}^{n_i^{\mathrm{Hu}}} \mathrm{Con}\, B_{ij}}{n_i^{\mathrm{Hu}}}. \tag{10}$$

For the dataset, Con$B$ is

$$\mathrm{Con}\, B = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{Con}\, B_i}{N}. \tag{11}$$

2.2. Annotation Quality of Label

2.2.1. Completeness of Label. The dimension can be split into two types. The completeness of the bounding box's label is employed to measure if each box has a label. The completeness of category describes the completeness for the category's quantity from the aspect of computational learning theory. In the common benchmarks for object detection, there exist minority categories. For a category, if the metric does not meet the requirement, the detection accuracy would be affected.

(1) Completeness of bounding box's label: for image i, the completeness is

$$\mathrm{CL}_i = \frac{n_i^{\mathrm{Label}}}{n_i^{\mathrm{Hu}}}, \tag{12}$$

where n_i^{Label} is the number of labels. For a dataset, the metric is

$$\mathrm{CL} = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{CL}_i}{N}. \tag{13}$$

(2) Completeness of category: the completeness of category is a metric that measures whether the number of samples can meet the training for the object detection model. As for a dataset, the classes are usually organized in a semantic hierarchy tree. Regarding a leaf node, if it meets the condition $n^{\mathrm{leaf}} > n^{\mathrm{lowbound}}$, the completeness is 1. Otherwise, the completeness is 0. For a parent node, the completeness is

$$\mathrm{CC}_{\mathrm{parent}}^{\mathrm{Label}} = \frac{\sum_{k=1,\ldots,n^{\mathrm{child}}}^{n^{\mathrm{child}}} CC_k^{\mathrm{Label}}}{n^{\mathrm{child}}}, \tag{14}$$

where n^{child} is the number of the corresponding child nodes. As a result, we can have the completeness of the category for a dataset.

2.2.2. Accuracy of Label. The dimension is employed to measure the closeness of the human and ground truth annotations. Regarding a dataset collected by a crowdsourcing annotation platform, the label noise is the most common error and has a direct influence on the training of the object detection model. The dimension has two elements: flat accuracy and hierarchical accuracy. The flat accuracy of the label is the usual element. However, the label space is often hierarchical. The hierarchical element can treat annotation errors distinctively and is the foundation of the utilization of annotation errors. As a result, we introduce these two kinds of elements for label accuracy evaluation.

(1) Flat accuracy of label: the flat accuracy of the label includes two metrics: precision and recall. The precision and recall of class t are

$$\begin{aligned} P_t &= \frac{\mathrm{tp}_t}{\mathrm{tp}_t + \mathrm{fp}_t}, \\ R_t &= \frac{\mathrm{tp}_t}{n_t^{\mathrm{GTr}}}, \end{aligned} \tag{15}$$

where n_t^{GTr} is the number of ground truth annotations for class t, and tp_t and fp_t are the numbers of true

positive objects and false-positive objects, respectively. For a dataset, precision can be calculated as follows:

$$P = \frac{\sum_{t=1,\ldots,M}^{M} P_t}{M}, \tag{16}$$

which treats each class equally. And similarly, the recall is obtained.

(2) Hierarchical accuracy of label: the element also has two metrics. The metrics of class t are

$$\begin{aligned} \mathrm{HP}_t &= \frac{\sum_{k=1,\ldots,n_i^{\mathrm{Hu}}}^{n_i^{\mathrm{Hu}}} \left(\left|\mathrm{ans}(C_k) \cap \mathrm{ans}(C_k')\right| / \left|\mathrm{ans}(C_k')\right|\right)^{1/p}}{n_i^{\mathrm{Hu}}}, \\ \mathrm{HR}_t &= \frac{\sum_{k=1,\ldots,n_i^{\mathrm{GTr}}}^{n_i^{\mathrm{GTr}}} \left(\left|\mathrm{ans}(C_k) \cap \mathrm{ans}(C_k')\right| / \left|\mathrm{ans}(C_k)\right|\right)^{1/p}}{n_i^{\mathrm{GTr}}}, \end{aligned} \tag{17}$$

where n_i^{Hu} and n_i^{GTr} are the corresponding numbers of human and ground truth annotations, C_k and C_k' denote the ground truth and human annotation labels, and ans(C) is the operation for computing ancestors for class C, $p > 0$. Then, via macroaveraging the metrics for all classes, the hierarchical precision and recall can be calculated.

2.2.3. Consistency of Label. Similar to the consistency of the bounding box, consistency of label concentrates on the confliction of spatiotemporal continuity of label. In the crowdsourcing platform, the labels in the adjacent frames often conflict due to the existence of low-level workers. If the label of an object is consonant with the labels in the previous and next frames, the metric $\mathrm{Con}\,L_{\mathrm{object}}$ is 1; otherwise, $\mathrm{Con}\,L_{\mathrm{object}}$ is 0. For image i, the consistency is

$$\mathrm{Con}\,L_i = \frac{\sum_{j=1,\ldots,n_i^{\mathrm{Label}}}^{n_i^{\mathrm{Label}}} \mathrm{Con}\,L_{ij}}{n_i^{\mathrm{Label}}}. \tag{18}$$

For the dataset, $\mathrm{Con}\,L$ is

$$\mathrm{Con}\,L = \frac{\sum_{i=1,\ldots,N}^{N} \mathrm{Con}\,L_i}{N}. \tag{19}$$

3. Case Study

To verify the effectiveness of the quality framework, two case studies are conducted based on the UTS dataset [47] and PASCAL VOC 2007 detection dataset [48]. UTS dataset is a video dataset with varying illumination conditions and viewpoints. PASCAL VOC 2007 dataset is an image dataset and contains twenty categories. Note that a few dimensions of the quality assessment framework are not fit for the dataset. To acquire the annotations, we let a group of students fulfill the annotation work. Generally, ground truth annotations are employed as golden standard annotations. However, in the evaluation process, we find that, to a certain extent, the ground truth annotations have quality problems, especially for the UTS dataset. Consequently, ground truth annotations are evaluated, where human annotations are regarded as "ground truth annotations." Additionally, to verify the completeness of category, the relationship between this metric and detection performance is studied by conducting object detection experiments.

3.1. Case Study for UTS Dataset. In this case study, the UTS dataset is utilized for verification. To reduce the amount of annotation labor, four shots are selected, and we annotate an image for every four or five images. Finally, the numbers of images in the four shots are 75, 120, 100, and 120 with 1166, 686, 639, and 919 objects, respectively. The evaluation is presented from the aspects of an image and a dataset. We find that the ground truth annotations have quality problems, especially for the completeness of the bounding box's quantity and the flat recall of the label.

3.1.1. Annotation Quality of an Image. For the clarity of the description of annotation quality, an image is selected for evaluation, which is given in Figure 2. The semantic hierarchy tree we defined is presented in Figure 3. The quality evaluation results for an image are given in Table 1. The accuracy of the bounding box for each object is shown in Figure 4.

Now, the analysis is given below. According to Table 1, the flat precision of hatchback is 0.25. However, it is because of the quality problems of ground truth annotations. Reviewing the annotations, we find that there are two small unannotated objects as shown in Figure 2. Hierarchical measures can reflect the relation of the classes. For instance, hierarchical precision for the hatchback is 0.42, while the flat precision is 0.25. Further, the consistency of the label is less than 1. It shows that there are inconsistent labels with the labels in adjacent frames. In Table 1, four metrics are equal to 1, reflecting that there is no error from these aspects.

3.1.2. Annotation Quality of Human and Ground Truth Annotations. Afterward, we show the annotation quality of the UTS dataset for the human and ground truth annotations. The annotation accuracies of the label are given in Tables 2 and 3. The completeness of the category of the ground truth annotations for each class and the original vehicle dataset is given in Figure 3, where the threshold is set

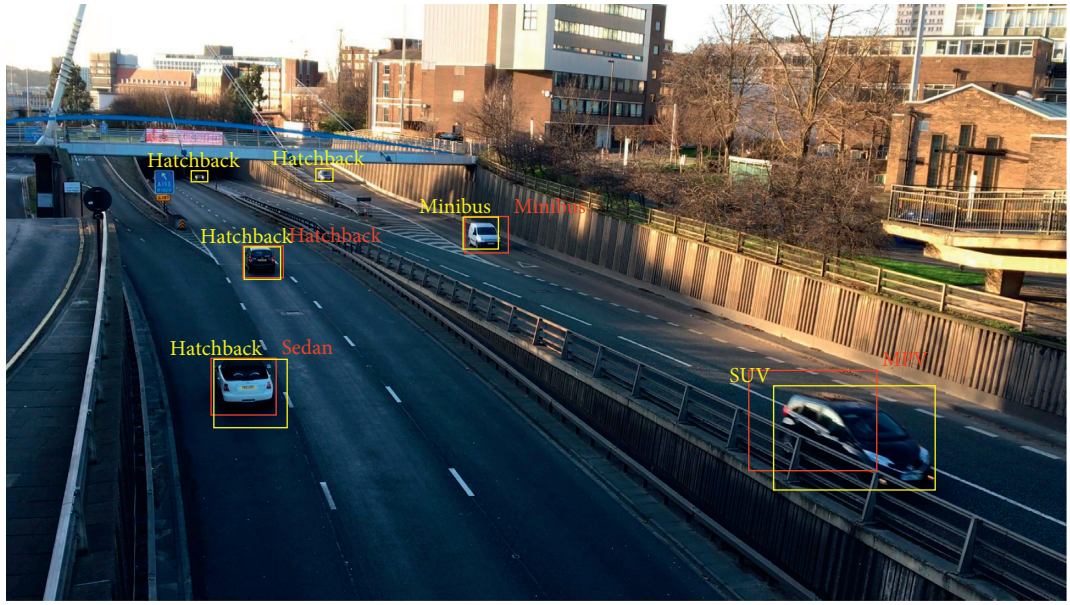

Figure 2: Human and ground truth annotations from the UTS dataset (ground truth and human annotations are shown in red and yellow, respectively).

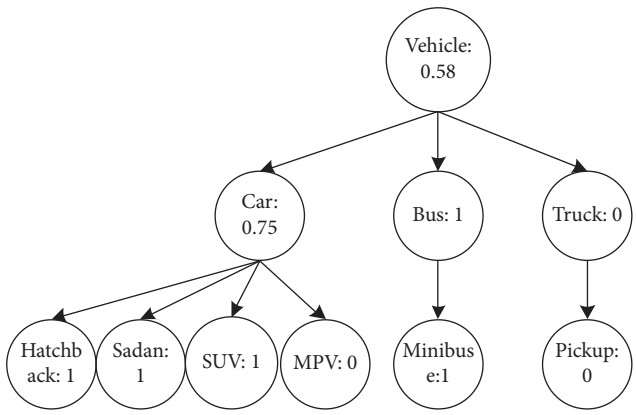

Figure 3: Semantic hierarchy tree and completeness of category of ground truth annotations for the original UTS training dataset.

Table 1: Results of other quality dimensions for an image.

Annotation quality dimension	Value
Completeness of bounding box's quantity	1
Completeness of bounding box's size	0.64
Accuracy of bounding box	0.67
Consistency of bounding box	1
Completeness of bounding box's label	1
Flat precision/recall of label	–/0.5
Hierarchical precision/recall of label	0.69/0.79
Flat precision/recall of hatchback	0.25/1
Hierarchical precision/recall of hatchback	0.42/0.83
Consistency of label	0.84

to 1000. The results of other quality dimensions are presented in Table 4.

The quality of human annotations is analyzed first. According to Tables 2 and 4, the overall annotation quality of the bounding box is good, while the annotation quality of the label is relatively poor. Accordingly, it can be inferred that the label's annotation is a more difficult task. In particular, for SUV and MPV, the accuracy and recall are too low. The hierarchical accuracy is higher than the flat accuracy, treating errors distinctively. According to Table 4, compared with other dimensions, the consistency of the label is lower on account of the own property.

The quality of ground truth annotations is evaluated here. According to Tables 2–4, the completeness of bounding box's quantity, flat and hierarchical recall of label, and consistency of label for ground truth annotations are lower than those for human annotations. When reviewing ground truth annotations, we find that ground truth annotations neglect some small and incomplete objects. But these small and incomplete objects can be annotated properly by experience. There are more inconsistent labels in ground truth annotations than in human annotations. Figure 3 shows that the completeness of category for MPV and pickup is 0, as the corresponding category's quantities do not reach the threshold. Generally, the quality problem exists in the ground truth annotations. Therefore, it is significant to perform a quality assessment in the process of annotation and ground truth inference.

3.1.3. Relationship between the Completeness of Category and Detection Performance. For the sake of exploring the relationship between the completeness of category and detection performance, the following experiment is conducted, which implies the effectiveness of the dimension. The object detection experiment on the UTS dataset is performed on the original dataset and downsampled dataset. As for downsampling, we just select images for every two images. The detection algorithm we use is Faster RCNN [3]. Table 5 presents the corresponding result.

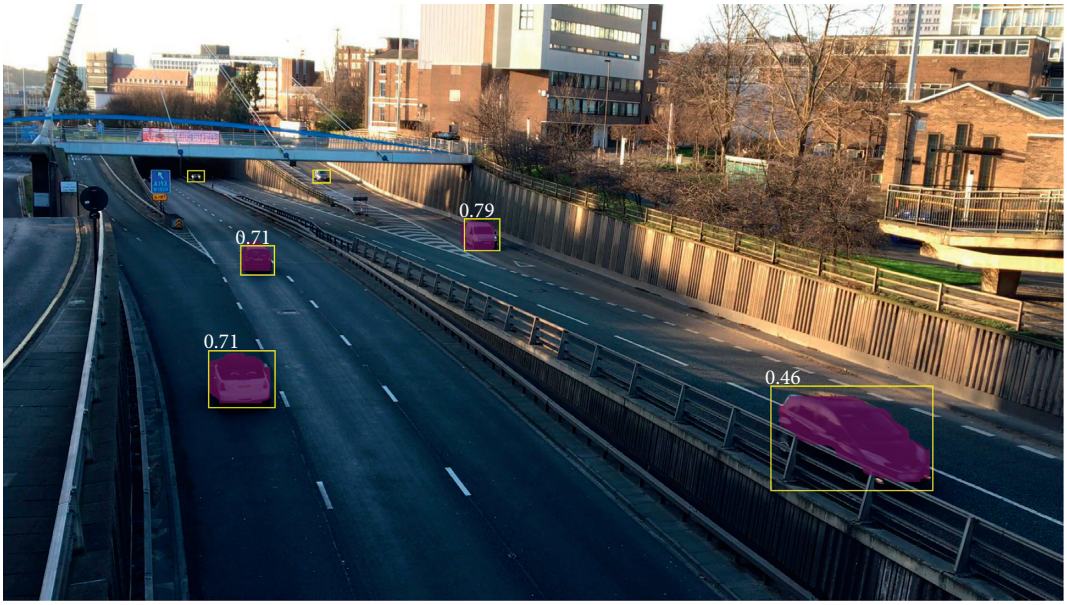

Figure 4: Accuracy of the bounding box in an image (note that two small objects are missed by the image instance segmentation algorithm.).

Table 2: Annotation accuracy of human annotations for the downsampled UTS dataset.

Class	Flat accuracy of label		Hierarchical accuracy of label	
	Precision	Recall	Precision	Recall
Hatchback	0.79 ± 0.02	0.55 ± 0.14	0.92 ± 0.01	0.81 ± 0.07
Sedan	0.58 ± 0.08	0.78 ± 0.12	0.86 ± 0.03	0.88 ± 0.07
Minibus	0.93 ± 0.10	0.57 ± 0.33	0.95 ± 0.06	0.70 ± 0.22
SUV	0.20 ± 0.06	0.27 ± 0.10	0.69 ± 0.03	0.74 ± 0.04
MPV	0.19 ± 0.14	0.26 ± 0.14	0.62 ± 0.14	0.72 ± 0.08
Pickup	0.57 ± 0.41	1 ± 0	0.72 ± 0.27	1 ± 0
On average	0.55 ± 0.11	0.57 ± 0.07	0.79 ± 0.07	0.81 ± 0.05

Table 3: Annotation accuracy of ground truth annotations for the downsampled UTS dataset.

Class	Flat accuracy of label		Hierarchical accuracy of label	
	Precision	Recall	Precision	Recall
Hatchback	0.57 ± 0.13	0.57 ± 0.04	0.85 ± 0.06	0.67 ± 0.05
Sedan	0.79 ± 0.12	0.44 ± 0.04	0.91 ± 0.08	0.65 ± 0.02
Minibus	0.58 ± 0.34	0.77 ± 0.17	0.72 ± 0.23	0.79 ± 0.17
SUV	0.27 ± 0.10	0.15 ± 0.05	0.75 ± 0.04	0.53 ± 0.09
MPV	0.26 ± 0.14	0.17 ± 0.15	0.72 ± 0.08	0.51 ± 0.18
Pickup	1 ± 0	0.28 ± 0.19	1 ± 0	0.28 ± 0.19
On average	0.58 ± 0.07	0.40 ± 0.07	0.83 ± 0.04	0.57 ± 0.07

Table 4: Results of other quality dimensions for the downsampled UTS dataset.

Annotation quality dimension	Human annotations	Ground truth annotations
Completeness of bounding box's quantity	0.98 ± 0.02	0.75 ± 0.05
Completeness of bounding box's size	0.96 ± 0.02	0.99
Consistency of bounding box	0.977 ± 0.002	0.96
Completeness of bounding box's label	0.52 ± 0.11	0.58
Consistency of label	0.86 ± 0.01	0.71

Table 5: Comparison of detection results based on the original training dataset and downsampled dataset.

Class	Object number in the training dataset	mAP (original)	mAP (downsampled)
Hatchback	12165	0.669	0.744
Sedan	5484	0.573	0.565
Minibus	3220	0.663	0.601
SUV	1761	0.560	0.576
MPV	898	0.154	0.142
Pickup	263	0.020	0.0001
On average	3965.2	0.440	0.438

According to Table 5, we argue that the detection result is closely related to the completeness of category. Overall, for the complete class whose training samples' quantity is over 1000, the corresponding mAP is high, while the detection mAPs of other classes are quite low. However, for SUV in the downsampled dataset, the quantity is about 880. The detection performance is still acceptable. It is due to its salient visual feature. Thus, the threshold varies with the class. Additionally, for the incomplete class, the performance declines with downsampling.

3.2. Case Study for PASCAL VOC 2007 Detection Dataset. In the case study, PASCAL VOC 2007 detection dataset is utilized for verification. To save labor, we select twenty images for each class as annotation samples. Finally, a random-selected dataset containing 353 images is obtained. The PASCAL VOC 2007 dataset is an image dataset. Consequently, a few quality dimensions are not fit for the dataset.

3.2.1. Annotation Quality for Human and Ground Truth Annotation. The quality of human and ground truth annotations for the PASCAL VOC 2007 dataset is given below. Accuracies of the label for the human and ground truth annotations are given in Tables 6 and 7. The semantic hierarchy tree and completeness of category quantity are given in Figure 5, where the threshold is set as 400. The results of other quality dimensions are provided in Table 8.

According to Tables 6 and 8, we can see that the human annotation quality for the dataset is good overall. However, the accuracies of the chair, potted plant, and dining table are relatively poor. For instance, the average flat recall for the potted plant is 0.54. This is because the potted plant is small and tends to be neglected. And for the other dimensions of human annotations, quality is relatively reliable.

Afterward, we evaluate the annotation quality of ground truth annotations. According to Tables 6–8, we find that the quality of ground truth annotations is slightly worse than that of human annotations. Specifically, the completeness of the bounding box's quantity and the flat recall of the label are relatively low. These dimensions indicate that there are more unannotated objects. As there are not enough images in the random-selected dataset, we calculate the completeness of category according to the original training set. The total completeness of category is 0.62, as 38% of the classes do not have enough samples.

Table 6: Annotation accuracy of human annotations for the selected images of the PASCAL 2007 dataset (the average values are computed for the twenty classes).

Class	Flat accuracy of label		Hierarchical accuracy of label	
	Precision	Recall	Precision	Recall
Person	0.98 ± 0.01	0.92 ± 0.05	0.99 ± 0.01	0.92 ± 0.04
Car	0.99 ± 0.02	0.94 ± 0.04	0.99 ± 0.01	0.94 ± 0.03
Chair	0.96 ± 0.03	0.74 ± 0.08	0.98 ± 0.01	0.82 ± 0.07
Bottle	0.98 ± 0.01	0.81 ± 0.08	0.99 ± 0.01	0.82 ± 0.07
Potted plant	1 ± 0	0.54 ± 0.31	1 ± 0	0.57 ± 0.29
Cow	0.99 ± 0.01	0.95 ± 0.01	0.997 ± 0.005	0.96 ± 0.01
Dining table	0.75 ± 0.16	0.59 ± 0.18	0.91 ± 0.06	0.64 ± 0.15
Bus	1 ± 0	0.94 ± 0.04	1 ± 0	0.96 ± 0.03
On average	0.96 ± 0.01	0.89 ± 0.04	0.983 ± 0.005	0.90 ± 0.04

Table 7: Annotation accuracy of ground truth annotations for the selected images of the PASCAL 2007 dataset (the average values are computed for the twenty classes).

Class	Flat accuracy of label		Hierarchical accuracy of label	
	Precision	Recall	Precision	Recall
Person	0.97 ± 0.01	0.79 ± 0.11	0.98 ± 0.01	0.81 ± 0.1
Car	0.94 ± 0.02	0.82 ± 0.08	0.96 ± 0.01	0.83 ± 0.08
Chair	0.90 ± 0.02	0.81 ± 0.09	0.96 ± 0.01	0.84 ± 0.08
Bottle	1 ± 0	0.74 ± 0.10	1 ± 0	0.81 ± 0.09
Potted plant	0.98 ± 0.02	0.77 ± 0.02	0.99 ± 0.01	0.78 ± 0.03
Cow	0.99 ± 0.01	0.81 ± 0.10	1 ± 0	0.83 ± 0.10
Dining table	0.68 ± 0.16	0.63 ± 0.11	0.87 ± 0.06	0.65 ± 0.09
Bus	0.91 ± 0.09	0.89 ± 0.08	0.94 ± 0.05	0.90 ± 0.07
On average	0.94 ± 0.02	0.84 ± 0.05	0.97 ± 0.01	0.87 ± 0.05

3.2.2. Relationship between the Completeness of Category and Detection Performance. To explore the relationship between the completeness of category and detection performance, an experiment is conducted in the same way as the previous section. We conduct object detection experiments on the original dataset and downsampled dataset of which the sampling ratio is 0.5. And the major classes of person, car, and chair are not downsampled. Table 9 presents the detection results, where classes are in descending order of quantity of training samples.

According to Table 9, on the whole, the detection performance declines after the dataset is downsampled. For the majority classes of person, car, and chair, there are no obvious declines of mAPs, as we do not make downsampling on these classes. As for the minority classes, mAPs for the bottle and potted plant decline a lot, which can be regarded

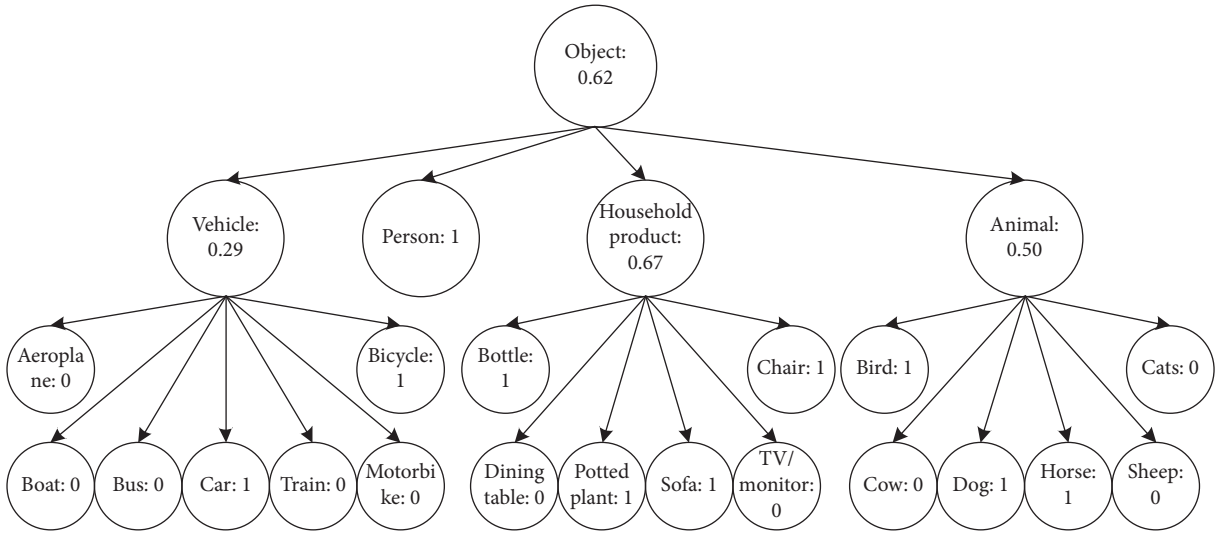

Figure 5: Semantic hierarchy tree and completeness of category for original PASCAL VOC 2007 training dataset.

Table 8: Results of other quality dimensions for the selected images and its original training dataset of the PASCAL VOC 2007 dataset.

Annotation quality dimension	Human annotations	Ground truth annotations
Completeness of bounding box's quantity	0.90 ± 0.04	0.88 ± 0.05
Completeness of bounding box's size	0.84 ± 0.02	0.85
Completeness of bounding box's label	0.9991 ± 0.0008	1

Table 9: Comparison of detection results based on the original training dataset and downsampled dataset (the average values are computed for the twenty classes).

Class	Object number in the training dataset	mAP (original)	mAP (downsampled)
Person	5447	0.779	0.778
Car	1644	0.831	0.807
Chair	1432	0.520	0.511
Bottle	634	0.576	0.519
Potted plant	625	0.459	0.376
Cow	356	0.767	0.721
Dining table	310	0.682	0.671
Bus	272	0.772	0.776
On average	783.1	0.714	0.678

as hard classes. But mAPs for the other classes of the minority are relatively high and change little, which should be regarded as easy classes. The hard classes are usually of small scale and have nonsalient visual features, hindering the learning of the object detection model. Therefore, the threshold for hard classes is relatively high. In the future process of constructing a dataset, the training samples' quantity for hard classes should be added.

4. Conclusion

Annotation quality is essential for the object detection model's training. In this paper, conceptual cognitive modeling for fine-grained annotation quality assessment is proposed. The annotation quality is calculated from the perspectives of the bounding box and label. To begin with, a generic framework based on general-purpose data quality dimensions is constructed from two aspects: the bounding box and the class label. This framework is used to assess the completeness and accuracy from the corresponding aspects. Nonetheless, the basic framework has limitations in assessing the consistency, the category's quantity, and the annotation errors. Thereupon, the cognitive theory is introduced, and we add the corresponding elements, including consistency of bounding box, hierarchical accuracy of label, consistency of label, and completeness of category. Case studies on the Urban Traffic Surveillance dataset and PASCAL VOC 2007 detection dataset indicate the validity of the framework. Currently, the annotation quality framework is constructed in an ideal condition. Future research is required to consider more practical factors.

Data Availability

The Urban Traffic Surveillance dataset and PASCAL VOC 2007 detection dataset used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Key Research and Development Plan of Shanxi Province (Nos. 201703D111027 and 201703D111023), Shanxi International Cooperation Project (No. 201803D421039), and Natural Science Foundation of Shanxi Province (No. 201801D121144).

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, MIT Press, Cambridge, MA, USA, 2012.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, MIT Press, Cambridge, MA, USA, 2015.

[4] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2014.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision,*, pp. 2980–2988, Venice, Italy, October 2017.

[6] Z. Fang, J. Ren, S. Marshall et al., "Triple loss for hard face detection," *Neurocomputing*, 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, October 2017.

[9] X. Xu, G. Li, G. Xie, J. Ren, and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, vol. 2019, Article ID 9180391, 12 pages, 2019.

[10] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[11] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "Crowdforge: crowdsourcing complex work," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology,*, pp. 43–52, Santa Barbara, CA, USA, October 2011.

[12] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, "Turkit: tools for iterative tasks on mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 29-30, Washington DC, USA, 2009.

[13] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Management Science*, vol. 31, no. 2, pp. 150–162, 1985.

[14] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[15] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–52, 2009.

[16] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the meaningfulness of "big data quality" (invited paper)," *Data Science and Engineering*, vol. 1, no. 1, pp. 6–20, 2016.

[17] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Future Generation Computer Systems*, vol. 89, pp. 548–562, 2018.

[18] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018.

[19] R. Zhang, M. Indulska, and S. Sadiq, "Discovering data quality problems," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 575–593, 2019.

[20] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614–622, Las Vegas, NV, USA, August 2008.

[21] S. Vittayakorn and J. Hays, "Quality assessment for crowdsourced object annotations," in *Proceedings of the British Machine Vision Conference*, pp. 1–11, Dundee, UK, August 2011.

[22] D. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," in *Proceedings 15th International Conference on Pattern Recognition (ICPR-2000)*, pp. 167–170, IEEE, Barcelona, Spain, 2000.

[23] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[25] W. Feng, W. Huang, and J. Ren, "Class imbalance ensemble learning based on the margin theory," *Applied Sciences*, vol. 8, no. 5, p. 815, 2018.

[26] J. Jiang, J. Kohler, C. Williams et al., "Live: an integrated production and feedback system for intelligent and interactive tv broadcasting," *IEEE Transactions on Broadcasting*, vol. 57, no. 3, pp. 646–661, 2011.

[27] S. Kiritchenko, S. Matwin, and F. Famili, "Functional annotation of genes using hierarchical text categorization," in *Proceedings of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology*, pp. 1–6, Detroit, MI, USA, 2005.

[28] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos, "Evaluation measures for hierarchical classification: a unified view and novel approaches," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 820–865, 2015.

[29] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *International Conference on Machine Learning*, pp. 5075–5084, Stockholm, Sweden, July 2018.

[30] J.-Y. Park and J.-H. Kim, "Incremental class learning for hierarchical classification," *IEEE Transactions on Cybernetics*, vol. 50, no. 1, pp. 178–189, 2018.

[31] C. Gu, J. J. Lim, P. arbeláez, and J. malik, "Recognition using regions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 1030–1037, IEEE, June2009.

[32] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International*

Workshop on Multimedia Information Retrieval, pp. 321–330, New York, NY, USA, 2006.

[33] C. Zhao, X. Li, J. Ren, and S. Marshall, "Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery," *International Journal of Remote Sensing*, vol. 34, no. 24, pp. 8669–8684, 2013.

[34] J. Han, D. Zhang, C. Gong, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2014.

[35] Y. Xi, J. Zheng, X. Li, X. Xu, J. Ren, and G. Xie, "SR-POD: sample rotation based on principal-axis orientation distribution for data augmentation in deep object detection," *Cognitive Systems Research*, vol. 52, pp. 144–154, 2018.

[36] Z. Wang, J. Ren, D. Zhang, M. Sun, and J. Jiang, "A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos," *Neurocomputing*, vol. 287, pp. 68–83, 2018.

[37] N. Clinton, A. Holt, J. Scarborough, L. Yan, and P. Gong, "Accuracy assessment measures for object-based image segmentation goodness," *Photogrammetric Engineering & Remote Sensing*, vol. 76, no. 3, pp. 289–299, 2010.

[38] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, p. 34, IEEE, San Diego, CA, USA, 2005.

[39] G. Sun, A. Zhang, J. Ren et al., "Gravitation-based edge detection in hyperspectral images," *Remote Sensing*, vol. 9, no. 6, p. 592, 2017.

[40] J. Ren, J. Jiang, D. Wang, D. Wang, and S. S. Ipson, "Fusion of intensity and inter-component chromatic difference for effective and robust colour edge detection," *IET Image Processing*, vol. 4, no. 4, pp. 294–301, 2010.

[41] X. Xie, G. Xie, X. Xu, L. Cui, and J. Ren, "Automatic image segmentation with superpixels and image-level labels," *IEEE Access*, vol. 7, pp. 10999–11009, 2019.

[42] J. M. Mandler and L. McDonough, "Concept formation in infancy," *Cognitive Development*, vol. 8, no. 3, pp. 291–318, 1993.

[43] J. L. McClelland and T. T. Rogers, "The parallel distributed processing approach to semantic cognition," *Nature Reviews Neuroscience*, vol. 4, no. 4, pp. 310–322, 2003.

[44] D. Casasanto, O. Fotakopoulou, and L. Boroditsky, "Space and time in the child's mind: evidence for a cross-dimensional asymmetry," *Cognitive Science*, vol. 34, no. 3, pp. 387–405, 2010.

[45] Y. Yan, J. Ren, G. Sun et al., "Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognition*, vol. 79, pp. 65–78, 2018.

[46] Y. Yan, J. Ren, H. Zhao et al., "Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos," *Cognitive Computation*, vol. 10, no. 1, pp. 94–104, 2018.

[47] Yi Zhou, Li Liu, L. Shao, and M. Mellor, "DAVE: a unified framework for fast vehicle detection and annotation," in *Proceedings of the European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 278–293, October 2016.

[48] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.