Welding in the World (2024) 68:855–865
https://doi.org/10.1007/s40194-024-01710-y

RESEARCH PAPER

Impact of annotation quality on model performance of welding defect detection using deep learning

Jinhan Cui[1] · Baoxin Zhang[1] · Xiaopeng Wang[1] · Juntao Wu[1] · Jiajia Liu[1] · Yan Li[2] · Xiong Zhi[3] · Wenpin Zhang[3] · Xinghua Yu[1,2]

Received: 5 October 2023 / Accepted: 30 January 2024 / Published online: 28 February 2024

Abstract

The use of X-ray-based non-destructive testing (NDT) methods is widespread in the task of welding defect detection. Many scholars have turned to deep-learning computer vision models for defect detection in weld radiographic images in recent years. Before model training, annotating the collected image data is often necessary. We need to use annotation information to guide the model for effective learning. However, many researchers have been focused on developing better models or refining training strategies, often overlooking the quality of data annotation. This paper delved into the impact of eight types of low-quality annotations on the accuracy of object detection models. In comparison to accurate annotations, inaccuracies in the annotated locations significantly impact model performance, while errors in category annotations have a minor effect on model performance. Incorrect location affects both the recall and precision of the model, while incorrect categorization only impacts the precision of the model. Additionally, we observed that the extent of the impact of location errors is related to the detection accuracy of individual classes, with classes having higher original detection AP experiencing more substantial decreases in AP under location errors. Finally, we analyzed the influence of annotator habits on model performance. The study examines the effects of various types of low-quality annotations on model training and their impact on individual detection categories. Annotator habits lead to the left boundary of annotated boxes being less accurate than the right boundary, resulting in a greater impact of annotations biased to the left than those biased to the right. Based on experiments and analysis, we proposed annotation guidelines for weld defect detection tasks: prioritize the quality of location annotations over category accuracy and strive to include all objects, including those with ambiguous boundaries.

Keywords Non-destructive testing · Defect detection · Annotation quality

1 Introduction

Welding is a method of joining workpieces, typically involving the application of heat, pressure, or a combination of both, with or without the use of filler material. The welding process plays an indispensable role across various industries. However, it is essential to note that no welding technique can guarantee defect-free results. Defects can significantly impact various performance aspects of welded components, potentially posing safety risks. Hence, the timely and accurate detection of defects in welded workpieces is paramount. There are various methods for defect detection, with non-destructive testing (NDT) techniques based on radiographic technology being widely adopted. This approach involves obtaining images of weld seams through X-ray imaging and subsequently inspecting these images for defects. The traditional detection method relies on quality inspectors to

Recommended for publication by Commission V—NDT and Quality Assurance of Welded Products.

✉ Xinghua Yu
xyu@bit.edu.cn

[1] School of Materials Science & Engineering, Beijing Institute of Technology, 100081 Beijing, People's Republic of China

[2] Chongqing Innovation Center, Beijing Institute of Technology, 401147 Chongqing, People's Republic of China

[3] Chongqing Special Equipment Inspection and Research Institute, 401121 Chongqing, People's Republic of China

visually examine weld images and determine whether defects are present. With the continuous advancement of artificial intelligence (AI), detection models based on machine learning (ML) or deep learning (DL) are being developed. More and more scholars are applying visual models to the task of weld defect detection.

As early as 2000, Nacereddine Nafaâ [1] and colleagues applied neural networks to the task of defect detection in weld radiographic images. They verified the effectiveness of artificial neural networks (ANNs) for edge detection in radiographic images and employed supervised learning to classify defect images. On the transformed non-trained set, the classification accuracy reached 96%. In 2013, O. Zahran [2] and others enhanced images through contrast enhancement and filtering, then utilized artificial neural networks to match features of defects in weld radiographic images for automated defect identification. They evaluated the method's performance using 150 radiographic images, demonstrating the effectiveness of features extracted from the MUSIC and Eigenvector methods. In 2017, Boaretto [3] and colleagues used a multi-layer perceptron (MLP) for defect detection in double-wall double-image radiographic welding images, achieving an accuracy of 88.6% and an F-score of 87.5% on test data. With the development of computer vision networks and the explosion of deep learning, more scholars are exploring the application of convolutional neural networks (CNNs) in defect detection. In 2019, Wenhui Hou [4] and team employed deep convolutional neural networks to classify weld defect types. They extracted local regions from radiographic images as the dataset, addressing the imbalance through resampling, and achieved a top-performing model with 97.2% accuracy. Paolo Sassi [5] and others developed an intelligent system for detecting welding defects on a spray gun assembly line using a deep learning approach and transfer learning, achieving an accuracy of 97.22%. In the same year, Yanxi Zhang [6] and team developed a deep learning algorithm based on convolutional neural networks for detecting laser welding defects, reaching an accuracy of approximately 94% for three defect categories. Zhifen Zhang [7] and team implemented real-time defect detection in robot arc welding using convolutional neural networks, achieving an average classification accuracy of 99.38%. Compared to object classification tasks, object detection tasks require the additional challenge of providing the coordinates of the detected defects (usually in the form of bounding boxes). Dingming Yang [8] and colleagues applied the single-stage network YOLOv5 to steel pipe weld defect detection in 2021, achieving a detection speed of 0.12 s per image and a mAP@50 of 97.8%. In 2023, Chen Ji [9] and team integrated the SPAM attention mechanism into the two-stage network Faster R-CNN, achieving a mAP of 86.3% in pipeline defect detection tasks. Jianyong Wang [10] and others improved the Faster R-CNN model by adding the FPN pyramid structure, variable convolutional networks, and background suppression, resulting in a mAP@50 of 93.5% for five defect detection tasks, a 24.3% improvement over the original Faster R-CNN model.

In the realm of computer vision and object detection, supervised learning is the prevailing approach [11]. This involves creating annotated training and validation datasets where objects of interest are meticulously labelled. The quality of data annotation directly impacts the accuracy of object detection models—generally, lower annotation quality results in decreased model accuracy [12]. In the context of welding seam defect detection, it is worth noting that the task differs from more conventional object detection tasks, such as face [13] or vehicle detection [14]. Annotating objects in this domain demands annotators with a certain level of expertise. This need for expertise arises due to the unique characteristics of the objects of interest. In welding seam defect detection, the object boundaries, which primarily constitute defects, are often inherently ambiguous. Furthermore, these defects are relatively small, making their annotation challenging [15]. The ambiguity in object boundaries, coupled with the subjective nature of human annotation, can lead to variations and inconsistencies in the annotations. These factors contribute to the low quality of the annotations in the welding seam defect detection datasets. In the field of NDT, researchers continuously explore detection models with higher accuracy and faster speeds. However, no one has investigated the impact of data annotation quality on these models. We recognize that annotation quality directly correlates with the accuracy of detection models, making the establishment of a comprehensive set of annotation standards crucial. Therefore, we conducted a detailed study on how low-quality annotations can affect the accuracy of detection models.

2 Methods

2.1 Dataset

Our dataset comprises exclusively images obtained from the welding production line of small-diameter pipes. We employed an X-ray imaging technique to capture images of the weld seam regions on the small-diameter pipes. Specifically, during X-ray imaging, to prevent overlap of the circular weld seams on the images, we intentionally deviated the X-ray source from the vertical direction by a certain angle [16]. This approach is referred to as the X-ray oblique radiographic imaging method, as illustrated in Fig. 1. This method makes the annular weld seams appear as ellipses in the resulting images.

Our dataset comprises 31,323 images, all grayscale images depicting elliptical weld seams. The dataset

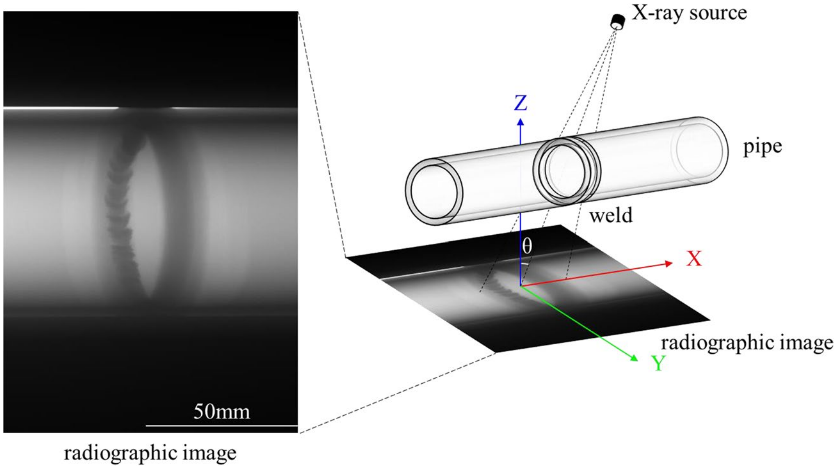


Fig. 1 X-ray oblique radiographic imaging method

encompasses nine distinct annotation categories, listed in descending order of quantity: weld, porosity, other defect, overlap, faulty formation, lack of fusion, penetration, undercut, and hollow. It is worth noting that the other defects primarily include fake defects.

Our work divided the dataset into training and validation datasets following a 9:1 ratio. Figure 2 illustrates the data distribution of different object categories in the training and validation datasets. It is important to emphasize that the images in the training and validation datasets remain consistent across all subsequent experiments.

2.2 Simulation of low-quality annotations

Based on the annotation scenarios and leveraging the dataset, we simulated eight instances of low-quality annotations. These low-quality annotation scenarios encompass left-shifted, right-shifted, upward-shifted, downward-shifted, and overestimated, underestimated, missing, and incorrect annotations, as depicted in Fig. 3. We categorize low-quality annotation situations into two types: location errors (the first six) and category errors (the last two).

Notably, the magnitude of annotation box shifts was set at 20% of the original dimensions (width or height). The probability of missing and incorrect annotations was also set at 20%. Though it is acknowledged that such significant errors and inaccuracies are typically not encountered in practical annotation processes, the 20% ratio was deemed reasonable to investigate the impact of various low-quality annotation scenarios. Importantly, these variations were applied exclusively to the training dataset, ensuring that the annotations in the validation dataset remained consistent and accurate across all circumstances.

2.3 Network

We employed the advanced YOLOv8 network as our experimental model in our experiments. YOLOv8, short for “You Only Look Once version 8,” represents the latest iteration in the YOLO series. It is renowned for its

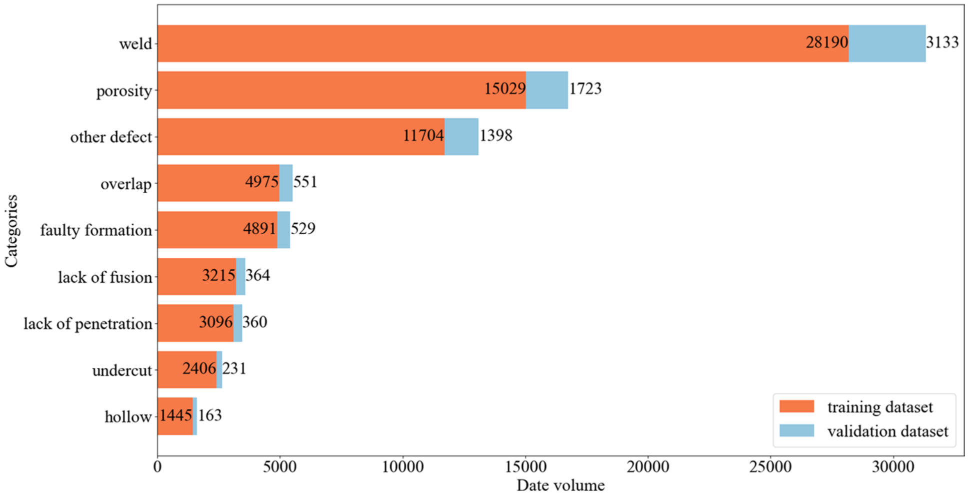


Fig. 2 Data quantities for each category in the training and validation datasets

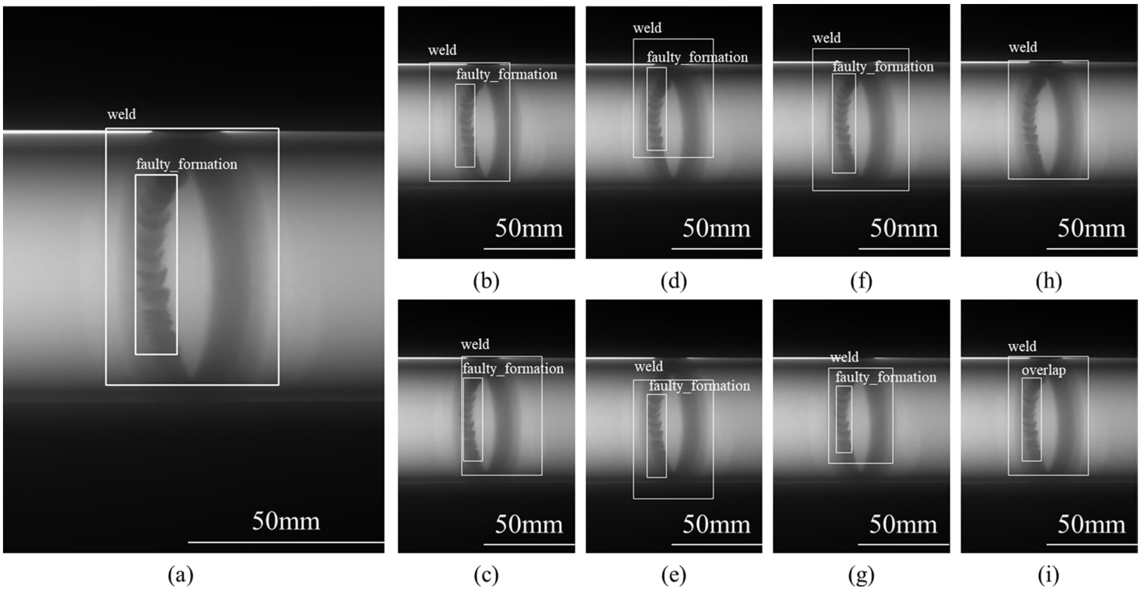


Fig. 3 Low-quality annotation scenarios. **a** Accurate annotations; **b** left-shifted annotations; **c** right-shifted annotations; **d** upward-shifted annotations; **e** downward-shifted annotations; **f** overestimated annotations; **g** underestimated annotations; **h** missing annotations; **i** incorrect annotations

real-time object detection capabilities and continuous evolution across multiple versions [17]. Building upon the successes of its predecessors, YOLOv8 integrates key architectural enhancements and advanced training strategies to achieve exceptional performance in object detection tasks.

One notable feature of YOLOv8 is its anchor-free design, which eliminates the need to preset a series of anchor boxes, making it particularly advantageous for detecting small objects [18]. In addition to architectural improvements, YOLOv8 benefits from advanced training techniques, including data augmentation [19], transfer learning [20], and robust optimization strategies [21]. These techniques contribute to the model's outstanding generalization and performance across diverse datasets and object categories.

The YOLOv8 model is available in five versions, denoted as n, s, m, l, and x, varying in parameter count. In our work, we utilized the n-version of the model. The primary hyperparameters were configured during the training process, as outlined in Table 1. It is worth noting that, to investigate the impact of left-shifted annotations and right-shifted annotations, we turned off the random horizontal flipping data augmentation strategy, which is typically enabled by default [22].

2.4 Loss function

The loss function in YOLOv8 comprises category classification loss and bounding box regression loss.

Table 1 Partial hyperparameter configuration

Hyperparameter (training)	Value	Hyperparameter (data augmentation)	Value
Epochs	100	Degrees	0.0
Batch size	16	Translate	0.1
Input image size	640	Perspective	0.0
Pretrained	true	Flipud	0.0
Amp	true	Fliplr	0.0
Initial learning rate	0.01	Mosaic	1.0
Final learning rate	0.01	Close mosaic epochs	10
Box loss weight	7.5	Mixup	0.0
Cls loss weight	0.5	Copy paste	0.0
Dfl loss weight	1.5		

For the classification loss, we utilize binary cross entropy (BCE) loss, which is an ordinary loss function in classification tasks [23].

As for the regression loss, we employ a combination of distribution focal loss (DFL) [24] and complete intersection over union (CIOU) loss [25].

In practice, the design of the DFL already considers the fuzziness and uncertainty in annotated boundaries, often caused by occlusions. However, in the context of welding seam defect detection, the fuzziness and uncertainty in boundary delineation primarily arise from the grayscale transition changes at defect boundaries. Introducing the DFL can mitigate the impact of boundary fuzziness.

The CIOU loss considers three critical parameters of the predicted and ground truth boxes: overlap area, center point distance, and aspect ratio. This comprehensive loss function ensures that the predicted boxes align more closely with the ground truth.

In YOLOv8, the total loss of the model is obtained by weighting and summing the individual component losses.

2.5 Metrics

In our work, we utilized mean average precision (mAP) and average precision (AP) to evaluate model performance.

The mAP is a valuable metric, considering the model's performance across various categories and providing an overall performance assessment. Higher mAP values typically indicate superior performance in multi-category object detection. The AP is a commonly used metric for evaluating the performance of object detection models. AP is primarily employed to measure the accuracy of a model across specific object categories. In object detection tasks, it is customary to report both category-specific AP and the overall mAP to understand the model's capabilities comprehensively. The calculation process for mAP and AP is as follows:

1. Prediction result sorting and filtering: Firstly, the model's detection results are sorted in descending order for each object category based on confidence scores. This means predictions with higher confidence scores are placed at the front. Non-maximum suppression (NMS) removes overlapping and lower-scoring prediction boxes, retaining only the highest-scoring detection boxes [26].
2. True positives (TP) and false positives (FP): Each detection result is categorized as either a TP or an FP based on its match with the ground truth labels, as shown in Table 2. If a model's detection result has an intersection over union (IOU) with the ground truth labels greater than a predetermined IOU threshold (typically 0.5) and has a confidence score high enough to meet a set threshold (typically 0.5), it is considered a TP. Otherwise, it is considered an FP. The calculation of IOU is illustrated in Fig. 4 and expressed by Formula 1, where S represents the area.

Table 2 The explanation of TP, FP, FN, and TN

		Ground truth	
		Positive	Negative
Predict	Positive	TP	FP
	Negative	FN	TN

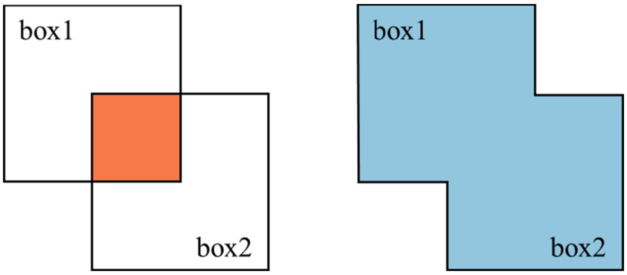


Fig. 4 Explanation of IOU calculation

$$IOU = \frac{S(box1 \cap box2)}{S(box1 \cup box2)} = \frac{S(box1 \cap box2)}{S(box1) + S(box2) - S(box1 \cap box2)} \tag{1}$$

3. Precision-recall curve: Next, based on different confidence thresholds, the precision and recall are calculated for each threshold. Precision indicates how many of the model's predictions are correct, while recall indicates how many true objects the model successfully detects. The formulas for calculating precision and recall are provided in Eq. 2 [27].

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{2}$$

4. Calculating AP and mAP: AP is the area under the precision-recall curve, as shown in Fig. 5. Typically, AP is computed by integrating the precision-recall curve. The specific calculation method may vary depending on different standards, but a common approach is calculating the sum of the areas of all small rectangles under the curve. Each small rectangle's height is the difference between two adjacent recall values, and its width corresponds to the associated precision value. For multi-category detection tasks, AP is calculated for each category. Then the mean of these AP values is computed to obtain mAP, as shown in Eq. 3, where

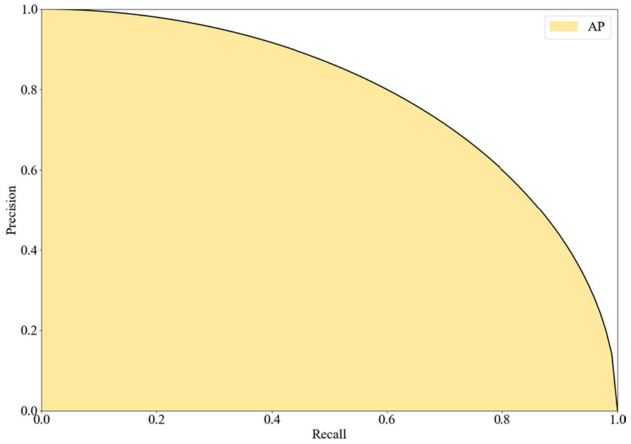


Fig. 5 Explanation of AP calculation

k represents the number of categories, and AP_i represents the AP for the category i.

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{3}$$

It is worth noting that AP and mAP are often labelled with the IOU threshold used. mAP@50 and mAP@50:5:95 are two commonly used mAP metrics. The former indicates that the IOU threshold for determining a positive sample is 0.5. The latter involves setting this threshold to a range of values, such as 0.5, 0.55, 0.6, …, 0.95 (with a step of 0.05), calculating the corresponding mAP for each threshold, and then averaging them.

3 Results and discussion

3.1 Impact of low-quality annotations on model performance

The influence of eight low-quality annotations on the model mAP is illustrated in Fig. 6. Through analysis, it becomes evident that left-shifted annotations, right-shifted annotations, upward-shifted annotations, downward-shifted annotations, overestimated annotations, and underestimated annotations significantly impact model performance. Conversely, missing and incorrect annotations have a relatively minor impact, with missing annotations showing the least influence, almost negligible.

These eight types of low-quality annotations can be categorized into two groups: location errors (the first six) and category errors (the last two), representing inaccuracies in the annotation location and annotation category, respectively. Figure 6 demonstrates that the impact of location errors far outweighs category errors.

Furthermore, we conducted tests to evaluate the AP of the model under these eight low-quality annotation conditions for single-class predictions, as shown in Fig. 7. Single-class predictions exclusively predict the location of objects in the image without predicting their categories.

In the case of single-class prediction, the impact of low-quality annotations follows a consistent pattern with regular predictions, where the influence of location errors is significantly more significant than that of category errors.

To provide a more accurate analysis of the impact of low-quality annotations on the model's decision-making process, we also compared the mean recall and mean precision of the model under various annotation scenarios, as shown in Fig. 8.

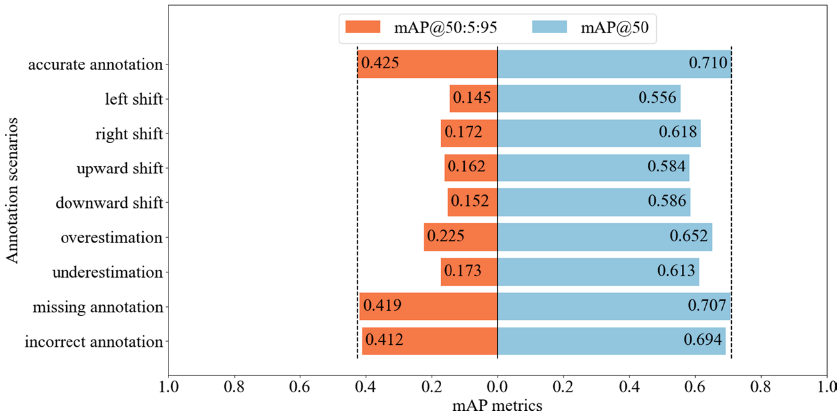


Fig. 6 Model mAP under various annotation scenarios

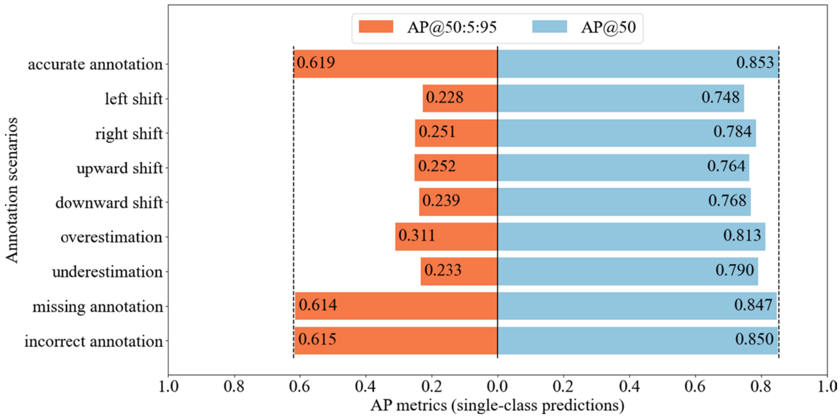
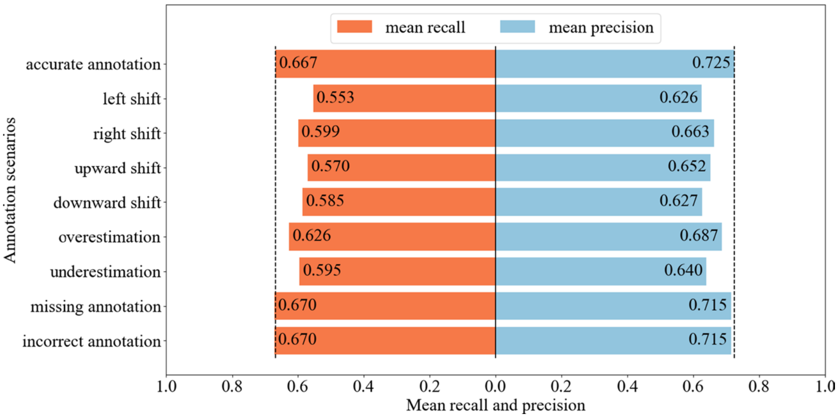


Fig. 7 Model single-class prediction AP under various annotation scenarios

Fig. 8 Model mean recall and precision under various annotation scenarios



It can be observed that location errors not only reduce the model's precision but also lower the recall, with both metrics showing a similar reduction magnitude. On the other hand, category errors mainly affect the model's precision, while recall remains unaffected.

To provide deeper insights, we selected the 6th feature layer of the model's backbone and generated grad-CAMs for one defective instance, as depicted in Fig. 9 [28].

Through the grad-CAMs, it is evident that location errors impact the model's feature extraction capability significantly, whereas category errors have a relatively minor effect. When the annotation locations are inaccurate, the training data provides erroneous object location information to the network. This leads to the network learning only partial features during learning object features and introduces substantial background information. This severely hampers the network's feature extraction ability. In cases of category errors, where the object location information provided to the network is accurate, the network can effectively learn object features, and its feature extraction capability remains unaltered, except during the classification phase. More specifically, inaccurate location affects both the calculation of classification and regression losses in the network, while inaccurate categorization only impacts the calculation of

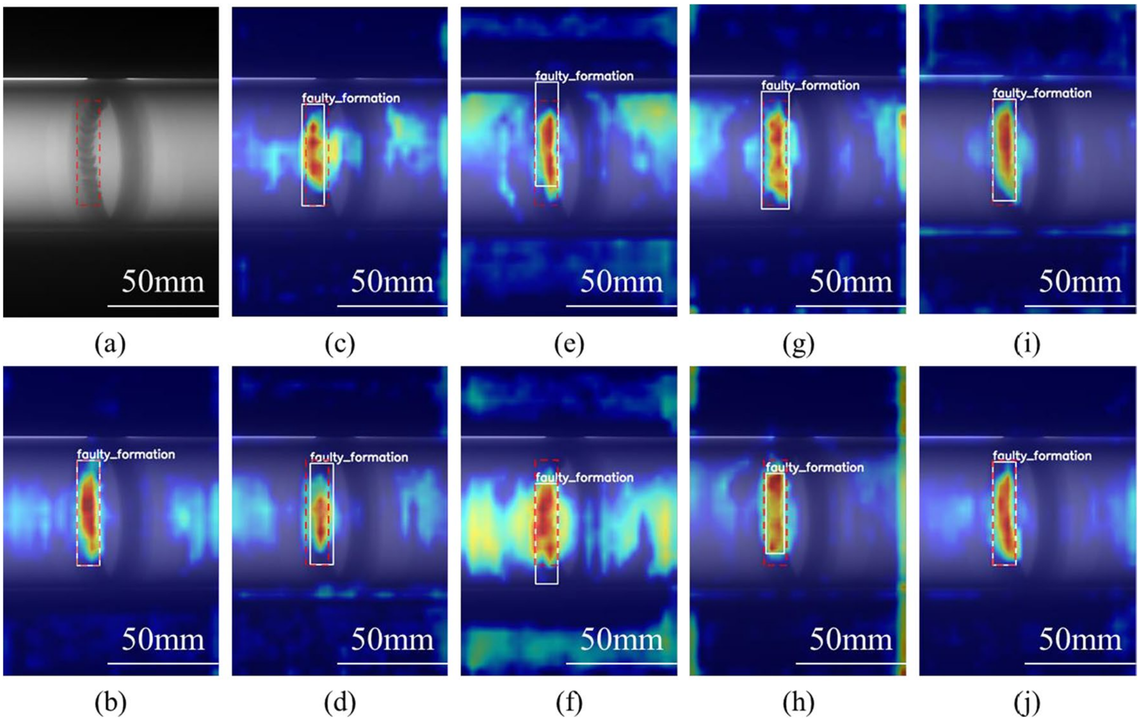


Fig. 9 Grad-CAMs for various annotation scenarios (6th feature layer). **a** Original image; **b** accurate annotations; **c** left-shifted annotations; **d** right-shifted annotations; **e** upward-shifted annotations; **f** downward-shifted annotations; **g** overestimated annotations; **h** underestimated annotations; **i** missing annotations; **j** incorrect annotations

classification loss. Therefore, location errors' impact is more significant than category errors.

In summary, the eight types of low-quality annotations, ranked by their impact from most to least significant, are as follows: left-shifted annotations, upward-shifted annotations, downward-shifted annotations, right-shifted annotations, underestimated annotations, overestimated annotations, incorrect annotations, and missing annotations. Notably, upward-shifted and downward-shifted annotations result in similar decreases in model performance, as do right-shifted annotations and underestimation.

3.2 The impact of low-quality annotations on model AP for different categories

The influence of eight low-quality annotations on the AP for each category is illustrated in Fig. 10. To provide a more intuitive representation of the relationship between model AP and categories, we have plotted the decrease in AP for each category under various low-quality annotation scenarios against the original AP, as shown in Fig. 11.

Evidently, for annotations with inaccurate locations, the reduction in AP is approximately linearly proportional to the original AP. In other words, as the original detection AP for one category increases, the AP reduction under low-quality annotation conditions also increases. On the other hand, for annotations with incorrect categories, the reduction in AP is independent of the original AP and remains relatively constant.

3.3 The impact of annotation habits on model performance

We also observed an interesting phenomenon where the impact of left-shifted annotations was significantly more significant than that of right-shifted annotations. We hypothesize that this is due to the annotation habits of the annotators. As shown in Fig. 12, when people

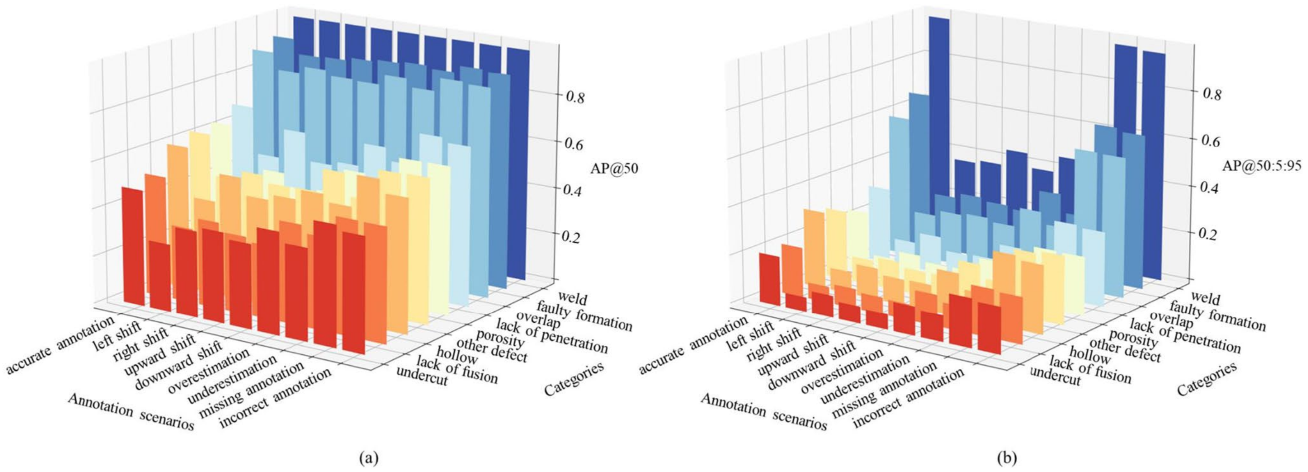


Fig. 10 AP for different categories under various annotation scenarios

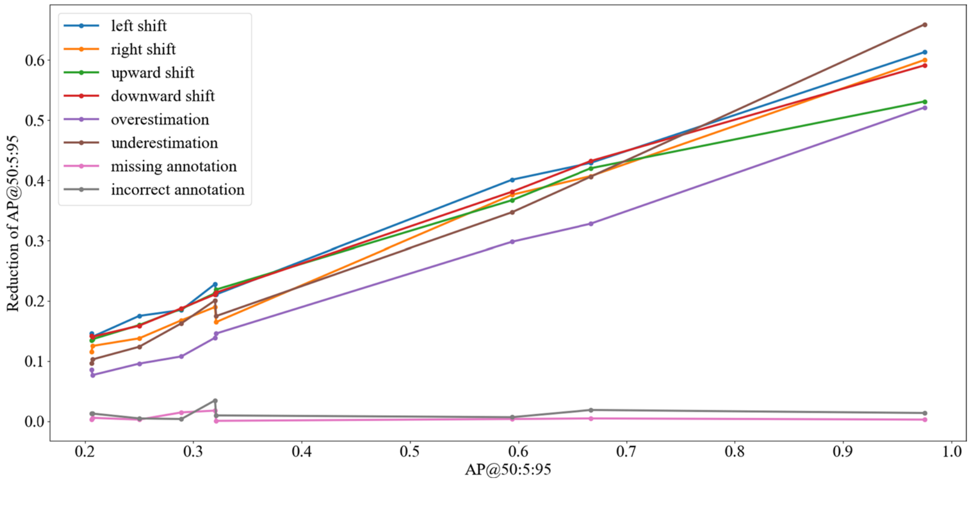


Fig. 11 Relationship between AP reduction and original annotation AP under various annotation scenarios

annotate rectangular boxes, they often start by clicking the rectangle's top-left corner and then click the bottom-right corner to define the entire rectangle. It means that annotators, without horizontal and vertical lines as references when clicking the first point, subconsciously click in the upper-left region of the true object box to ensure that the rectangle fully encloses the object. However, when clicking the second point, there are horizontal and vertical lines as references, enabling annotators to mark the lower and right boundaries of the object accurately. In summary, the annotators' habits influence the annotation boxes, resulting in lower annotation quality for the upper and left boundaries compared to the lower and right boundaries.

In the weld defect detection task, most object categories have an aspect ratio (width-to-height ratio) less than 1. This implies that during the training process, the model has a lower tolerance for deviations in the left–right direction than in the up-down direction. This is why the impact of upward-shifted annotations and downward-shifted annotations remains similar.

We horizontally flipped all training data to validate the above hypotheses and then applied left and right shifts to the annotations. We continued investigating the impact of left-shifted and right-shifted annotations on model performance, as shown in Fig. 13.

After applying the annotations to the flipped dataset, it can be observed that the impact of right-shifted

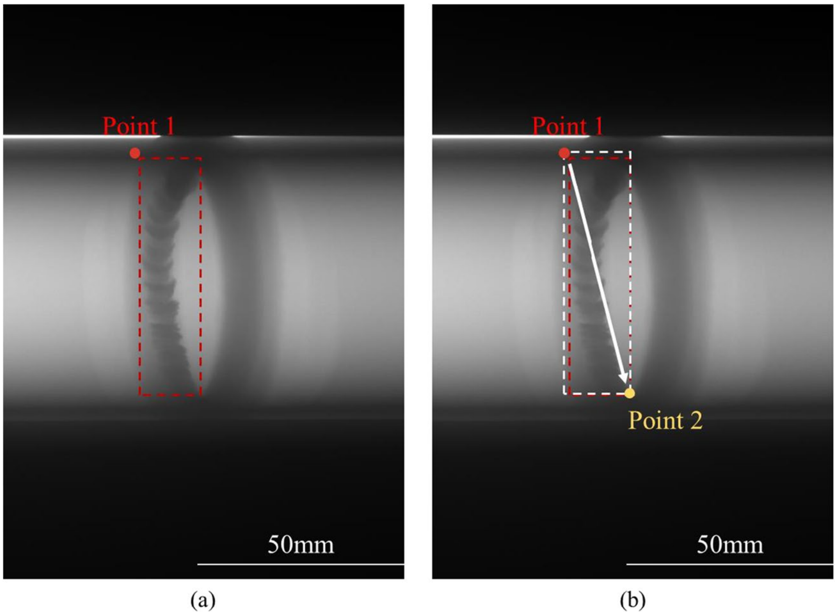


Fig. 12 Typical annotation habits of annotators. The red point represents the first annotated point, the yellow point represents the second annotated point, the white dashed rectangle represents the annotated bounding box, and the red dashed rectangle represents the actual defect boundary box

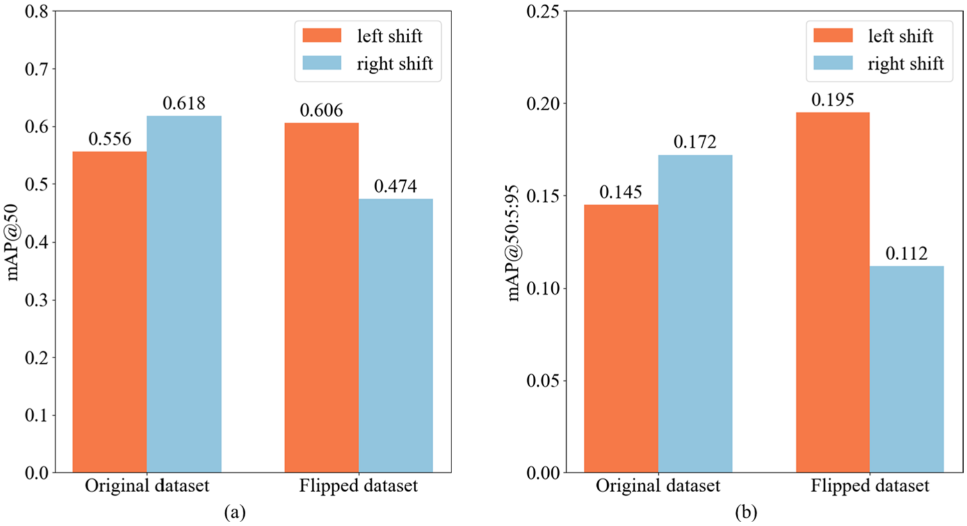


Fig. 13 Model mAP of left-shifted and right-shifted annotations on original and horizontally flipped datasets

annotations is more significant than that of left-shifted annotations. This suggests that the original annotations' left boundary relative to the right boundary needs to be more accurate, likely due to the annotation habits of annotators.

4 Conclusion

In our study on the weld detection task, we investigated the impact of eight low-quality annotation scenarios on computer vision models. The specific conclusions are as follows:

1. Inaccurate location of annotations has a more significant impact on the model than inaccurate categorization. When ranking the eight low-quality annotation scenarios based on their impact on the model's mAP, the order of impact is as follows: left-shifted annotations, upward-shifted annotations, downward-shifted annotations, right-shifted annotations, underestimated annotations, overestimated annotations, incorrect annotations, and missing annotations. Notably, upward-shifted and downward-shifted annotations lead to a similar decrease in model performance, while right-shifted and underestimated annotations also result in a similar decrease.
2. Inaccurate location affects both the recall and precision of the model, while inaccurate categorization only impacts the precision of the model.
3. For scenarios with inaccurate location of annotations, the decrease in AP is approximately linearly correlated with the original AP. However, for scenarios with inaccurate categorization of annotations, the decrease in AP is unrelated to the original AP.
4. Human annotation habits contribute to the inaccuracy of the left boundaries of annotation boxes relative to the right boundaries, causing a more significant impact from left-shifted annotations than right-shifted annotations.

In summary, we recommend that annotators prioritize the quality of annotation location over category accuracy. Additionally, annotations should strive to encompass all object information, including ambiguous boundaries, to the greatest extent possible.

Funding This work was supported by the Natural Science Foundation of China (NSFC No. U2141216).

Data Availability Data are available under request to corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Nafaa, N.; Redouane, D.; Amar, B (2000) Weld defect extraction and classification in radiographic testing based artificial neural networks. In Proceedings of the 15th World Conference on Non Destructive Testing, Rome, Italy, p 15–21
2. Zahran O, Kasban H, El-Kordy M, El-Samie FEA (2013) Automatic weld defect identification from radiographic images. NDT E Int 57:26–35. https://doi.org/10.1016/j.ndteint.2012.11.005
3. Boaretto N, Centeno TM (2017) Automated detection of welding defects in pipelines from radiographic images DWDI. NDT E Int 86:7–13. https://doi.org/10.1016/j.ndteint.2016.11.003
4. Hou W, Wei Y, Jin Y, Zhu C (2019) Deep features based on a DCNN model for classifying imbalanced weld flaw types. Measurement 131:482–489. https://doi.org/10.1016/j.measurement.2018.09.011
5. Sassi P, Tripicchio P, Avizzano CA (2019) A smart monitoring system for automatic welding defect detection. IEEE Trans Industr Electron 66:9641–9650. https://doi.org/10.1109/TIE.2019.2896165
6. Zhang Y, You D, Gao X, Zhang N, Gao PP (2019) Welding defects detection based on deep learning with multiple optical sensors during disk laser welding of thick plates. J Manuf Syst 51:87–94. https://doi.org/10.1016/j.jmsy.2019.02.004
7. Zhang Z, Wen G, Chen S (2019) Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding. J Manuf Process 45:208–216. https://doi.org/10.1016/j.jmapro.2019.06.023
8. Yang D, Cui Y, Yu Z, Yuan H (2021) Deep learning based steel pipe weld defect detection. Appl Artif Intell 35:1237–1249. https://doi.org/10.1080/08839514.2021.1975391
9. Ji C, Wang H, Li H (2023) Defects detection in weld joints based on visual attention and deep learning. NDT E Int 133:102764. https://doi.org/10.1016/j.ndteint.2022.102764
10. Wang J, Mu C, Mu S, Zhu R, Yu H (2023) Welding seam detection and location: deep learning network-based approach. Int J Press Vessels Pip 202:104893. https://doi.org/10.1016/j.ijpvp.2023.104893
11. Cunningham P, Cord M, Delany SJ (2008) Supervised learning. Springer, Berlin Heidelberg
12. Ma J, Ushiku Y, Sagara M (2022) The effect of improving annotation quality on object detection datasets: a preliminary study. IEEE/CVF Conf Comput Vision Pattern Recog Workshop (CVPRW) 2022:4849–4858. https://doi.org/10.1109/CVPRW56347.2022.00532
13. Kumar A, Kaur A, Kumar M (2019) Face detection techniques: a review. Artif Intell Rev 52:927–948. https://doi.org/10.1007/s10462-018-9650-2
14. Mukhtar A, Xia L, Tang TB (2015) Vehicle detection techniques for collision avoidance systems: a review. IEEE Trans Intell Transp Syst 16:2318–2338. https://doi.org/10.1109/TITS.2015.2409109
15. Zou Z, Chen K, Shi Z, Guo Y, Ye J (2023) Object detection in 20 years: a survey. Proc IEEE 111:257–276. https://doi.org/10.1109/JPROC.2023.3238524
16. Zhang B, Wang X, Cui J, Wu J, Wang X, Li Y, Li J, Tan Y, Chen X, Wu W, Yu X (2023) Welding defects classification by weakly supervised semantic segmentation. NDT and E Int 138:102899. https://doi.org/10.1016/j.ndteint.2023.102899

17. Terven J, Córdova-Esparza D-M, Romero-González J-A (2023) A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extraction 5:1680–1716. https://doi.org/10.3390/make5040083
18. Tian Z, Shen C, Chen H, He T (2022) FCOS: a simple and strong anchor-free object detector. IEEE Trans Pattern Anal Mach Intell 44:1922–1933. https://doi.org/10.1109/TPAMI.2020.3032166
19. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. Journal of big data 6:1–48. https://doi.org/10.1186/s40537-019-0197-0
20. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big data 3:1–40. https://doi.org/10.48550/arXiv.1804.06353
21. Zhao Z-Q, Zheng P, S-t Xu, Wu X (2019) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30:3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865
22. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Netw Vis Recognit 11:1–8. https://doi.org/10.48550/arXiv.1712.04621
23. Ruby U and V Yendapalli (2020) Binary cross entropy with deep learning technique for image classification. Int J Adv Trends Comput Sci Eng 9: https://doi.org/10.30534/ijatcse/2020/175942020
24. Li X, Wang W, Hu X, Li J, Tang J, Yang J (2021) Generalized focal loss V2: learning reliable localization quality estimation for dense object detection. IEEE. https://doi.org/10.1109/CVPR46437.2021.01146
25. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell 34:12993–13000. https://doi.org/10.1609/aaai.v34i07.6999
26. Neubeck A, Van Gool L (2006) Efficient non-maximum suppression. 18th Int Conf Pattern Recog (ICPR'06) 3:850–855. https://doi.org/10.1109/ICPR.2006.479
27. Sokolova M, N Japkowicz and S Szpakowicz (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Australas Joint Conf Artif Intell 1015–1021. https://doi.org/10.1007/11941439_114
28. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2016) Grad-CAM: Why did you say that? https://doi.org/10.48550/arXiv.1611.07450

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.