User-Guided Instance-Level Data Augmentation and Detection-Aware Optimization Framework

Daohu Zhang, Dongxiang Fu*

School of Optical Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Corresponding Author: Dongxiang Fu (fudx@usst.edu.cn)

Abstract: Deep learning–based object detection models have achieved significant progress in industrial vision and robotic perception; however, their performance heavily depends on large-scale, high-quality an-notated data. To address the challenges of frequent emergence of new objects and the high cost of annotation, this paper proposes a user-guided instance-level data augmentation and detection-aware optimization framework for extremely few-shot scenarios. The proposed method leverages limited human–computer interaction to guide a segmentation model in extracting target instances and employs a mask quality evaluation mechanism to filter valid samples; meanwhile, semantic-consistency-aware instance-level copy-paste and adaptive illumination enhancement are combined to generate diverse training data. In addition, a detection-aware feedback mechanism utilizes model error information to guide data generation in a closed-loop manner, further improving robustness. Experimental results demonstrate that, under few-shot settings, the proposed method significantly outperforms conventional data augmentation strategies in terms of precision and recall while substantially reducing manual annotation costs, and its engineering feasibility and stability are further validated through real vision-guided robotic arm grasping experiments. Under the COCO 10-shot setting, the proposed method achieves a 15.2% improvement in mAP@50 and exhibits stronger robustness under complex illumination conditions.

Keywords: Object detection; Data augmentation; Few-shot learning; Instance segmentation; YOLOv8; Segment Anything Model

1. Introduction

In recent years, deep learning-based object detection algorithms have made significant progress in the field of computer vision. Among these, the one-stage detection models represented by the YOLO series [1,2] have been widely applied in industrial inspection, robotic grasping, and autonomous driving due to their efficiency and high detection accuracy [3,4]. However, such models typically rely on large-scale, accurately annotated datasets for training. In practical applications, obtaining a large number of annotated samples is often costly and time-consuming.

In robotic and industrial vision scenarios, the detection targets are characterized by diverse categories, frequent updates, and significant scene variations. The traditional approach of "collecting large amounts of data first, followed by offline annotation" fails to meet the demands of rapid deployment [5]. Few-Shot Object Detection (FSOD) aims to alleviate the problem of data scarcity, but existing methods typically rely on complex meta-learning frameworks or feature re-weighting strategies, which suffer from instability during training and limited generalization ability. For example, Meta R-CNN [6] adapts to new categories through meta-learning but is sensitive to the distribution of initial samples, while DeFRCN [7] introduces feature re-weighting but incurs high computational costs.

Data augmentation, as an important method to improve the model's generalization ability, is widely used in object detection tasks. Common methods include geometric transformations, color perturbations, and combination augmentation techniques such as Mosaic and MixUp [8]. However, these methods mainly perform transformations at the pixel or image level, making it difficult to introduce new semantic information. The instance-level copy-paste method [9], although able to increase target diversity to some extent, often disrupts the semantic consistency between the foreground and back-ground due to random pasting, which can introduce noisy samples and make model convergence more challenging.

With the development of Foundation Segmentation Models, methods represented by the Segment Anything Model (SAM)[10] have demonstrated powerful general segmentation capabilities, providing new possibilities for instance-level data generation. However, directly using segmentation results for detection training still presents challenges: First, the segmentation masks are not optimized for detection tasks and may include background noise; second, the authenticity and environmental consistency of the generated samples are difficult to guarantee.

To address the above-mentioned issues, this paper proposes a user-guided instance-level data augmentation and detection-aware optimization method. The method utilizes a small amount of user interaction to guide the segmentation model in extracting high-quality target instances, and evaluates and selects the segmentation results based on the requirements of the detection task. On this basis, instance-level copy-paste with semantic consistency constraints and adaptive illumination enhancement are applied to generate high-quality synthetic data. Additionally, a detection-aware feedback mechanism is introduced to reverse-guide the data generation process using the error information of the object detection model, forming a closed-loop optimization. It should be noted that this paper does not design a new few-shot object detection network structure but proposes a data generation and optimization framework for object detection tasks. This framework alleviates training difficulties under few-shot conditions at the data level through user guidance and detection-aware feedback mechanisms, demonstrating good generalization ability. It can be used as a plug-in to integrate with existing object detection models or few-shot detection methods.

The main contributions of this paper are as follows:

1) A user-guided instance-level data augmentation framework is proposed, which generates high-quality object detection training data with minimal human interaction;

2) A mask quality evaluation strategy for object detection tasks is designed, effectively improving the reliability of instance-level augmented data;

3) Semantic-consistency-aware copy-paste and adaptive illumination enhancement are introduced to reduce noise in synthetic data;

4) A detection-aware feedback mechanism is proposed to achieve collaborative optimization of data generation and detection performance.

2. Related Work

2.1 Few-Shot Object Detection

Few-shot object detection aims to achieve effective detection under conditions with extremely few annotated samples. Existing methods primarily include meta-learning-based methods and feature re-weighting-based methods. For example, FsDet [11] uses a meta-learning framework to adapt to new categories, but the training process is complex and sensitive to noise. Methods based on automated data augmentation, such as AutoAugment [12], improve model performance by searching for augmentation strategies. However, their design is mainly targeted at large-scale data scenarios and struggles to perform well under few-shot conditions. In recent years, some works have attempted to combine data augmentation to enhance FSOD performance, but these are often limited to image-level transformations and fail to effectively introduce new instances.

2.2 Data Augmentation in Object Detection

Traditional data augmentation methods mainly include geometric transformations (e.g., flipping, rotation) and illumination perturbations (e.g., brightness adjustment). In recent years, instance-level copy-paste methods have been introduced for object detection tasks. For example, Ghiasi et al. [9] proposed Copy-Paste, which enhances data diversity by randomly pasting instances. However, the random strategy tends to disrupt semantic consistency, affecting the model's convergence. Some improvements, such as Context-Aware Copy-Paste [13], introduce background matching, but they do not consider the specific requirements of detection tasks and lack a feedback mechanism.

2.3 Foundation Segmentation Models

Foundation segmentation models, such as SAM, demonstrate good generalization ability and can generate high-quality instance masks through prompts (e.g., bounding boxes). However, their output is not designed for object detection tasks, and direct use may introduce noise, limiting their effectiveness in detection applications. Some works, such as SAM-Adapter [14], attempt to adapt to downstream tasks but do not integrate with data augmentation.

3. Method

3.1 Overall Framework

The overall process of the framework proposed in this paper is shown in Figure 1. The system takes a small number of images provided by the user as input and uses user interaction to guide the segmentation model in extracting target instances. After instance filtering and enhancement, training data compliant with the YOLOv8[15] standard are automatically generated. The data generation strategy is continuously optimized under the guidance of the detection-aware feedback mechanism. Figure 1 illustrates the flowchart of the framework, including the user interaction module, instance extraction, quality evaluation, data augmentation, and feedback loop.

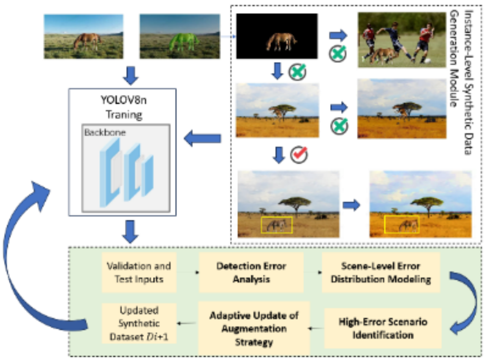

YOLOV8n
Traning
Backbone
Instance-Level Synthetic Data Generation Module
Validation and Test Inputs
Detection Error Analysis
Scene-Level Error Distribution Modeling
High-Error Scenario Identification
Adaptive Update of Augmentation Strategy
Updated Synthetic Dataset Di+1

Figure 1: Overall workflow of the user-guided instance-level data augmentation and detection-aware optimization framework.

3.2 User-Guided Instance Segmentation

The user only needs to specify the target region in a small number of images through bounding box selection or clicking. The system utilizes SAM to automatically generate target instance masks, as shown in Figure 2. Specifically, the user provides bounding box prompts, and SAM generates the masks based on the ViT-H encoder. This approach significantly reduces the cost of manual annotation while ensuring the semantic accuracy of the target instances. Compared to traditional instance seg-mentation models, the Segment Anything Model (SAM) has stronger generalization ability on unseen categories, a characteristic that is particularly important under few-shot conditions. This allows the user to obtain high-quality instance masks with minimal interaction. Assuming the user provides K images (K=3~10), with each interaction taking less than 1 minute per image, the total annotation cost is much lower than traditional bounding box annotations.

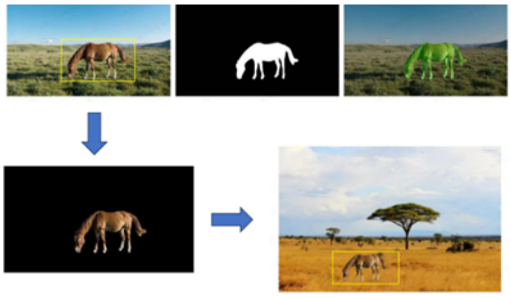


Figure 2: Target Instance Segmentation Result.

3.3 Mask Quality Evaluation for Detection Tasks

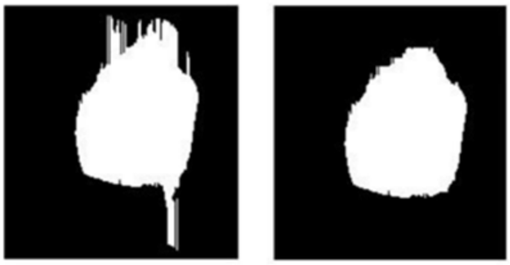
To ensure that the instance masks obtained by SAM segmentation are suitable for subsequent object detection training, this paper introduces a mask quality evaluation function designed for detection tasks. During the data synthesis phase, instances are automatically filtered to avoid interference from low-quality segmentations in the learning process of the detection model. Given a binary mask M, its overall quality score is defined as:

$$Q = \alpha S_{sam} + \beta C_{edge} + \gamma A_{ratio} \quad (1)$$

Here, $S_{sam} \in [0,1]$ represents the segmentation confidence output by the SAM model, which is used to measure the reliability of the instance at the semantic level; C_{edge} represents the geometric continuity of the mask edges, calculated using the Sobel operator to compute the mask gradient and is defined as:

$$C_{\text{edge}} = 1 - \frac{\sum |\nabla M|}{|\partial M|} \quad (2)$$

Where ∂M is the set of boundary pixels of the mask, which is used to suppress fragmented or noisy segmentation results; $A_{\text{ratio}} = \frac{\sum M}{H \times W}$ represents the ratio of the foreground region to the entire image, which is used to constrain the instance scale's reasonableness, preventing very small or very large objects from introducing bias into detection training. The weight parameters are set as $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$, and are tuned using the validation set. Only when Q>0.7is the instance mask retained for subsequent instance pasting and data generation. Figure 3 provides a comparison example between high-quality and low-quality masks.



(a) Low-Quality Segmentation Masks （b）High-Quality Segmentation Masks

Figure 3: Example Comparison Between High-Quality and Low-Quality Masks.

3.4 Instance-Level Data Augmentation with Semantic Consistency

To avoid introducing semantic conflicts in the synthetic data through random copy-pasting, this paper introduces a semantic consistency constraint during the instance-level data augmentation phase. DINOv2 [16] is used as the visual semantic feature encoder. DINOv2 learns stable high-level semantic representations on large-scale unsupervised data, effectively capturing the semantic relationship between foreground instances and background scenes, and it does not rely on category annotations, making it suitable for data synthesis tasks in open-ended scenarios.

Specifically, given the foreground instance image I_{fg} and its corresponding mask M, as well as the background image I_{bg}, the semantic feature vectors are extracted for each:

$$f_{fg} = \text{DINOv2}(I_{fg} \odot M)$$
$$f_{bg} = \text{DINOv2}\big(I_{bg} \odot (1 - M)\big) \tag{3}$$

Subsequently, the cosine similarity between the two is calculated:

$$\text{Sim}(f_{fg}, f_{bg}) = \cos(f_{fg}, f_{bg}) \tag{4}$$

The instance-level copy-paste operation is allowed only when the semantic similarity satisfies $\text{Sim}(f_{fg}, f_{bg}) > \theta$. In this paper, the threshold is set to $\theta = 0.6$, a value obtained through empirical tuning on the validation set. This constraint effectively suppresses the generation of semantically mismatched samples, making the synthetic data more contextually aligned with the real distribution, thereby reducing the interference of noisy samples in the training of the detection model. Figure 4 presents a comparative example of the semantic consistency constraint.

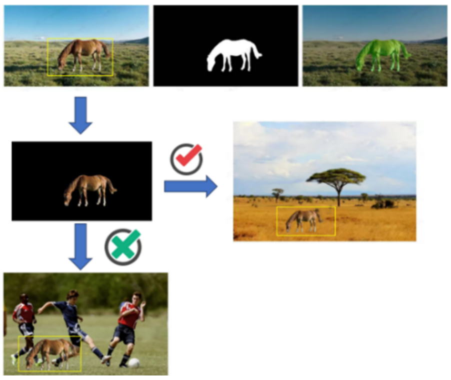


Figure 4: Comparison with and Without Semantic Contextual Consistency Constraints.

3.5 Adaptive Illumination Enhancement

In response to the complex and variable lighting conditions in real-world applications, this paper introduces an adaptive illumination enhancement strategy during the instance-level data synthesis stage to improve the robustness of the detection model under different lighting environments. Specifically, perturbations are applied to the foreground instances across three dimensions: brightness, contrast, and color. The strategy integrates Gamma correction and HSV color space enhancement.

In terms of brightness and contrast adjustment, a Gamma correction operator is introduced, with its parameter γ sampled from a uniform distribution:

$$\gamma \sim \text{Uniform}(0.5,1.5) \tag{5}$$

This simulates different lighting conditions such as overexposure and underexposure. For color perturbation, the image is converted to the HSV space, and random perturbations are applied to each channel:

$$H \pm 10^{\circ}, S \pm 20\%, V \pm 30\% \tag{6}$$

Unlike traditional random augmentation, this paper adapts the perturbation parameters based on the overall illumination distribution of the background image. This ensures that the enhanced foreground instances maintain consistency in brightness and color statistical properties with the background environment, thereby reducing synthetic artifacts. This strategy effectively expands the coverage of the training data in the illumination dimension, making the generated samples more aligned with real-world environmental distributions. Figure 5 presents example results of the adaptive illumination enhancement.

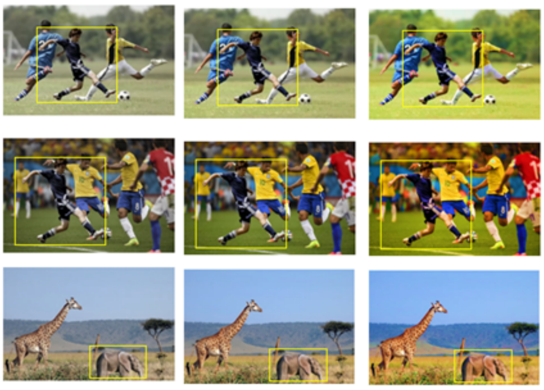


Figure 5: Results of Adaptive Illumination Enhancement.

3.6 Detection-Aware Feedback Optimization

After the initial synthetic dataset is constructed, this paper further introduces a detection-aware feedback optimization mechanism. By explicitly utilizing the error distribution of the object detection model under different scene conditions, the mechanism provides reverse guidance for the dynamic adjustment of the data generation strategy. This enables the synergistic optimization of data generation and detection performance.

Specifically, the initial synthetic dataset D_0 is used to train the YOLOv8 detection model, and its detection errors $E$are computed on the validation set, including IoU loss and confidence error. To obtain more targeted feedback information, the detection errors are categorized and statistically analyzed based on scene attributes, such as target scale ranges, target location distributions, and lighting conditions. This allows the construction of a scene-aware error distribution.

Based on the error analysis results, for the scene categories with higher detection errors, the instance-level data augmentation parameters are dynamically adjusted as follows:

1) Scale Adjustment: Increase the sampling probability for higher error scale ranges, with the instance paste scale range set to $\text{scale} \sim [0.5,2.0]$;

2) Location Adjustment: Suppress sampling of paste positions near the image boundaries to reduce the interference of boundary truncation in detection training;

3) Illumination Adjustment: Prioritize enhancing the sample generation ratio under high error lighting conditions, in combination with the adaptive illumination enhancement strategy.

Using the updated data generation strategy, a new round of synthetic dataset D_i is generated, and the detection model is retrained or fine-tuned. The above process is iterated in a closed-loop for 3 – 5 rounds, until the detection error distribution stabilizes. The specific procedure is shown in Algorithm 1. This detection-aware feedback mechanism effectively guides the data distribution to adaptively focus on the "hard areas" of the detection model, thereby continuously improving the model's detection robustness in complex scenes.

Algorithm 1：Detection-Aware Feedback Optimization

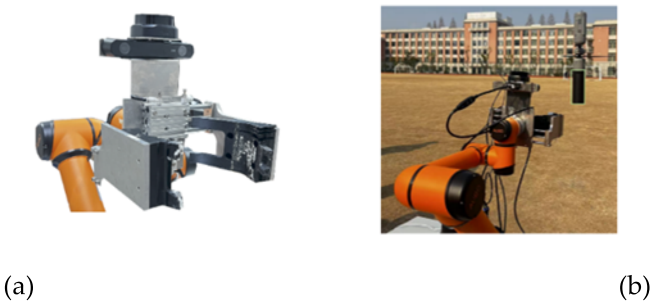
```
Input:
Initial synthetic dataset $D_0$, detection model YOLOv8
Output:
Optimized detection model
    1. Train the YOLOv8 model using the dataset $D_0$
    2. Evaluate the model performance on the validation set and compute the
       detection errors $E$(IoU loss, confidence error)
    3. Classify and statistically analyze the errors $E$based on scene attributes
       (scale, location, lighting)
    4. For high error scenes, adjust the data augmentation parameter distributions
       (scale, location, lighting)
    5. Generate a new dataset $D_i$based on the updated strategy
    6. Retrain or fine-tune the detection model using $D_i$
    7. If the error distribution has not converged, return to step 2; otherwise, finish
```

4. Experiment

4.1 Experimental Setup

To validate the data usability and engineering effectiveness of the user-guided instance-level data augmentation and detection-aware optimization methods proposed in this paper, the augmented dataset generated by this method is directly applied to the robotic arm visual grasping system. System-level validation is then conducted in a real drone grasping scenario.

The experimental platform consists of a six-degree-of-freedom AUBO-i5 robotic arm, with the grasping target being a coaxial drone hovering in mid-air. The system uses a single ZED2i depth camera as the visual sensor, mounted above the end effector of the robotic arm, as shown in Figure 6, for target detection and grasping guidance.



(a) (b)

Figure 6: (a) Schematic of the sensor installation; (b) Grasping system with the target UAV and sensors

In the system architecture, YOLOv8n is deployed as the object detection network on the industrial computer. The user-guided instance-level data augmentation method proposed in this paper is used to construct the training dataset for the drone target, generating training samples that can be used for actual grasping with minimal manual annotation. The detection network is trained solely on the augmented data generated in this paper, and its output is directly used for robotic arm grasping decision-making. Unlike traditional methods that validate detection performance solely on offline datasets, this paper focuses on the usability of the generated dataset in real robot closed-loop tasks.

4.2 Dataset Construction

In real robot applications, the rapid deployment of new targets is often limited by high annotation costs and the availability of limited samples. To address this issue, based on the drone targets used in the experiments, this paper applies the user-guided instance-level data augmentation method to expand a small set of real drone images, constructing a training dataset suitable for robotic grasping tasks. Examples of the augmented target instances are shown in Figure 7. Specifically, the user only needs to provide interactive annotations on a few original images, and the system will automatically extract high-quality target instances. These instances are then filtered using the semantic consistency constraint and the detection-aware feedback mechanism. Subsequently, using instance-level copy-paste and adaptive illumination enhancement strategies, the target instances are pasted onto different backgrounds and lighting conditions to create a training dataset suitable for real-world grasping scenarios.

In this experiment, the user provided 30–100 annotation points to generate approximately 1,000 synthetic training images. The entire process took about 15 minutes, with the user interaction time being less than 10 minutes, significantly reducing the labor costs compared to traditional manual annotation methods.

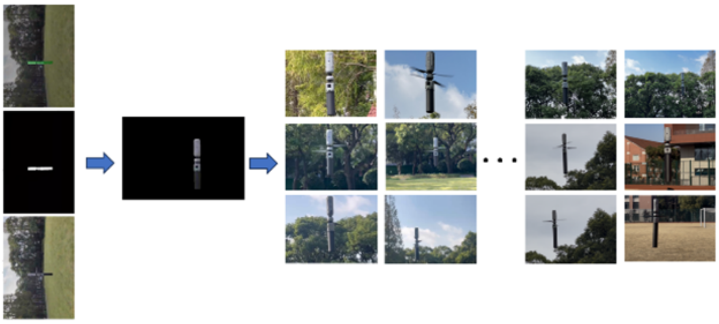


Figure 7: Examples of Augmented Target Instances for Robotic Grasping Tasks.

4.3 Quantitative Results Analysis

To validate the basic effectiveness of the user-guided instance-level data augmentation and detection-aware optimization method proposed in this paper for object detection tasks, a quantitative evaluation is first performed under the novel classes division protocol of the COCO dataset [17] (with base class pretraining and novel class for evaluation). The evaluation is conducted in 1-shot, 3-shot, 5-shot, and 10-shot scenarios. The comparison methods include random Copy-Paste [9], Mosaic [8], AutoAugment [18], and the representative few-shot detection method DeFRCN [7]. All methods are based on the YOLOv8n backbone network and are trained under consistent settings (200 epochs, batch size = 16, AdamW optimizer). The evaluation metrics used are mAP@50, mAP@50:95, and Recall.

As shown in Table 1, under different few-shot settings, the method proposed in this paper achieves the best or second-best detection performance on the COCO novel classes, especially under extremely low sample conditions such as 1-shot and 3-shot, where it demonstrates a more significant advantage in terms of mAP@50 and Recall metrics.

In addition, this paper constructs a test image set containing 100 images under complex lighting conditions (such as backlighting, local shadows, and strong reflections) to further evaluate the method's robustness and generalization ability in real-world scenarios. YOLOv8n is selected as the experimental backbone network, primarily due to its wide application in industrial vision and embedded systems, as well as its high inference efficiency. It should be noted that the proposed method is independent of the specific detection network architecture and can theoretically be extended to other scale models of YOLOv8 or two-stage detectors. Relevant experiments will be further validated in future work.

Table 1: Comparison of Detection Performance on COCO Novel Classes under Different Few-Shot Settings.

Method	**1-shot mAP@50**	**3-shot mAP@50**	**5-shot mAP@50**	**10-shot mAP@50**	**mAP@50:95 (10-shot)**
Random Copy-Paste	18.6	28.3	34.9	41.2	19.6
Mosaic	19.8	29.6	36.1	42.6	20.3
AutoAugment	20.5	30.4	36.8	43.1	20.9
DeFRCN	22.7	32.8	39.5	45.9	22.8
Ours	25.1	35.2	41.3	47.2	24.1

4.4 Ablation Study

To validate the contribution of each module to detection performance, ablation experiments are conducted under the COCO 1-shot and 3-shot settings, sequentially removing the semantic consistency, illumination enhancement, and detection-aware feedback modules. Table 2 presents the performance of each ablation version in terms of mAP@50, mAP@50:95, and Recall metrics

Table 2: Ablation Study Results.

Modules	**1-shot mAP@50**	**1-shot mAP@50:95**	**1-shot Recall**	**3-shot mAP@50**	**3-shot mAP@50:95**	**3-shot Recall**
All Modules（Ours）	**25.1**	**12.9**	**40.6**	**35.2**	**19.8**	**53.7**
Remove Semantic Consistency	22.3	11.1	37.2	32.4	17.5	50.1
Remove Illumination Enhancement	24.0	12.0	38.5	34.1	18.7	52.0
Remove Detection-Aware Feedback	23.1	11.5	37.8	33.5	18.3	51.4

The ablation study in Table 2 shows that semantic consistency, illumination enhancement, and the detection-aware feedback loop all contribute significantly to performance improvement. Among them, semantic consistency has the greatest impact on mAP in low-shot scenarios, illumination enhancement improves the model's robustness under complex lighting conditions, and the feedback loop further optimizes the data distribution, making the detection results more stable.

4.5 Real Robotic Arm Drone Grasping Experiment

To directly validate the usability of the user-guided instance-level data augmentation dataset in real robot tasks, drone grasping experiments are conducted on the AUBO-i5 six-degree-of-freedom robotic arm platform. All experiments use the same YOLOv8n network architecture, with only the augmentation strategy for the training data being changed, ensuring that the differences in grasping performance are solely due to the data construction method. The comparison methods include: no data augmentation, random Copy-Paste, Mosaic, AutoAugment, and the method proposed in this paper. The experiments are performed in two typical lighting conditions: standard lighting and low lighting, with the latter including complex scenarios such as backlighting, shadows, and uneven lighting. Each method is tested through multiple independent grasping trials under different lighting conditions, and the grasping success rate and detection failure rate are recorded.

The experimental results are shown in Table 3, with some successful grasping examples illustrated in Figure 8. Under standard lighting conditions, all methods are able to complete the grasping task, but the method proposed in this paper achieves the highest success rate. In low-light scenarios, models trained without augmentation or with generic augmentation show a significant decline in performance, while the method proposed in this paper still maintains stable performance, with a grasping success rate improved by approximately 10–15% compared to other methods.

The above results indicate that the user-guided instance-level data augmentation method proposed in this paper not only improves detection performance but also generates data that can directly support real robotic arm grasping tasks, demonstrating excellent engineering usability and practical application value.

Table 3: Grasping Performance Comparison under Different Data Augmentation Methods.

Method	**Standard Lighting Grasping Success Rate (%)**	**Low Lighting Grasping Success Rate (%)**	**Detection Failure Rate (%)**
No Data Augmentation	82.4	63.1	18.7
Random Copy-Paste	85.6	68.9	15.4
Mosaic	87.2	70.3	13.9
AutoAugment	88.1	72.0	12.8
Ours	**91.5**	**84.2**	**7.6**

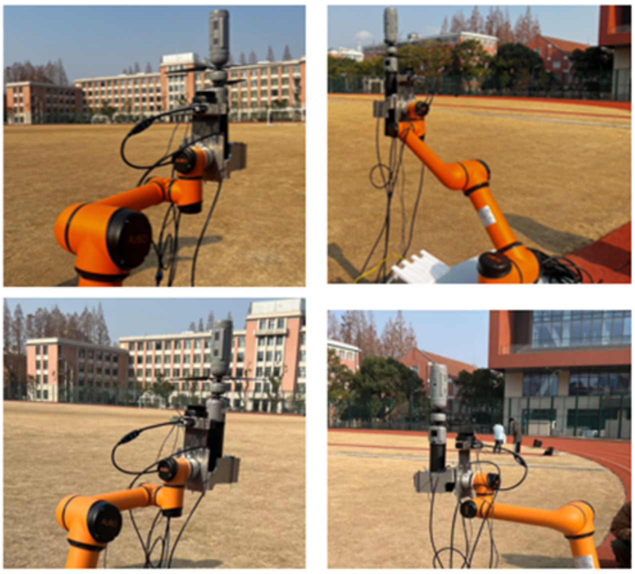


Figure 8: Examples of successful grasping under different illumination conditions.

4.6 Efficiency Analysis

The method proposed in this paper generates 1,000 synthetic images in just 15 minutes, with the user interaction time being less than 10 minutes. Compared to traditional fully manual bounding box annotation (using the LabelMe tool), this method reduces the annotation cost by approximately 92% in the COCO 10-shot scenario (the traditional method requires annotating around 5,000 bounding boxes, while this method only requires the user to select 30–100 annotation points).

5. Discussion

Although the method proposed in this paper achieves significant performance improvements in few-shot object detection tasks, it still has certain limitations in some extreme scenarios. For example, when the target is in a complex background or is severely occluded, the instance segmentation model may make errors in determining the target's boundaries, which can affect the quality and effectiveness of the subsequent instance-level data synthesis. In future work, the introduction of depth estimation or 3D geometric information could be used to constrain the spatial structure of the target, further enhancing the realism and diversity of the generated instances in complex scenes.

Moreover, the user-guided instance-level data augmentation and detection-aware feedback framework proposed in this paper exhibits good model independence and scalability, making it transferable to other object detection architectures (e.g., DETR [19]) and applicable to real-time robotic perception systems. Given that the experiments in this paper are primarily based on the YOLOv8n backbone network, further validation of the method's generalizability and robustness will be carried out on detection models of different scales and structures.

6. Conclusion

This paper presents a user-guided instance-level data augmentation and detection-aware optimization method, which effectively alleviates the issues of data scarcity and high annotation costs in object detection. Experimental results show that the method significantly improves detection performance in few-shot scenarios while maintaining high efficiency, demonstrating its practical value and scalability in industrial applications and real-time environments.

References

[1] Hussain M. Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. IEEE access. 2024;12:42816-33.

[2] Vijayakumar A, Vairavasundaram S. Yolo-based object detection models: A review and its applications. Multimedia Tools and Applications. 2024;83(35):83535-74.

[3] Wei J, As'arry A, Rezali KAM, Yusoff MZM, Ma H, Zhang K. A Review of YOLO Algorithm and Its Applications in Autonomous Driving Object Detection. IEEE Access. 2025.

[4] Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies. 2024;12(2):15.

[5] Brust C-A, Käding C, Denzler J. Active learning for deep object detection.

[6] Wu X, Sahoo D, Hoi S. Meta-rcnn: Meta learning for few-shot object detection. Proceedings of the 28th ACM international conference on multimedia2020. p. 1679-87.

[7] Qiao L, Zhao Y, Li Z, Qiu X, Wu J, Zhang C. Defrcn: Decoupled faster r-cnn for few-shot object detection. Proceedings of the IEEE/CVF international conference on computer vision2021. p. 8681-90.

[8] Dadboud F, Patel V, Mehta V, Bolic M, Mantegh I. Single-stage uav detection and classification with yolov5: Mosaic data augmentation and panet. 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS): IEEE; 2021. p. 1-8.

[9] Ghiasi G, Cui Y, Srinivas A, Qian R, Lin T-Y, Cubuk ED, et al. Simple copy-paste is a strong data augmentation method for instance segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2021. p. 2918-28.

[10] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. Proceedings of the IEEE/CVF international conference on computer vision2023. p. 4015-26.

[11] Liu C, He Y, Zhang X, Wang Y, Dong Z, Hong H. CS-FSDet: A Few-Shot SAR Target Detection Method for Cross-Sensor Scenarios. Remote Sensing. 2025;17(16):2841.

[12] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. Autoaugment: Learning augmentation strategies from

data. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition2019. p. 113-23.

[13] Guo Q, Wang S, Chang C, Rambach J. CACP: Context-Aware Copy-Paste to Enrich Image Content for Data Augmentation. Proceedings of the Computer Vision and Pattern Recognition Conference2025. p. 5177-86.

[14] Chen T, Zhu L, Deng C, Cao R, Wang Y, Zhang S, et al. Sam-adapter: Adapting segment anything in underperformed scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision2023. p. 3367-75.

[15] Sohan M, Sai Ram T, Rami Reddy CV. A review on yolov8 and its advancements. International Conference on Data Intelligence and Cognitive Informatics: Springer; 2024. p. 529-45.

[16] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193. 2023.

[17] Jain S, Dash S, Deorari R. Object detection using coco dataset. 2022 International Conference on Cyber Resilience (ICCR): IEEE; 2022. p. 1-4.

[18] Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. Autoaugment: Learning augmentation policies from data.

[19] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection.