Multimodal Causal Reasoning for UAV Object Detection (Appendix)

Anonymous Author(s)
Affiliation
Address
email

A Illustration of Backdoor Adjustment

Backdoor adjustment is a core method of causal inference used to eliminate confounding bias. The key idea is to adjust a set of confounders Z that block all non-causal paths (ack door) between treatment X and outcome Y, allowing estimation of the causal effect X on Y.

A.1 Backdoor Adjustment Formula

If the variable set Z satisfies the backdoor criterion, the causal effect of X on Y(Average Causal Effect, ACE) can be estimated using:

$$P(Y = y|do(X)) = \sum_z P(Y = y|X = x, Z = z) \cdot P(Z = z) \tag{1}$$

where the Z must satisfy the backdoor criterion: 1) Blocks all backdoor paths: Z must block every path between X and Y that contains an arrow into X. 2) No new bias introduced: Z must not include any descendants of X.

A.2 Derivation Process.

The backdoor adjustment is derived using causal graph rules and probability theory, with the following key steps: **From intervention to Conditional Probability.** The intervention do(X=x) corresponds to removing all incoming edges to X in the causal graph and fixing $X = x$. In this intervention, the distribution of Y depends only on X and its parents. If Z satisfies the backdoor criterion, the post-intervention distribution can be expressed as:

$$P(Y|do(X = x)) = \sum_z P(Y|X = x, Z = z) \cdot P(Z = z|do(X) = x) \tag{2}$$

Since do(X=x) does not affect Z(because Z is not a descendant of X), we have $P(Z = z|do(X = x)) = P(Z = z)$,leading to:

$$P(Y|do(X = x)) = \sum_z P(Y|X = x, Z = z) \cdot P(Z = z). \tag{3}$$

A.3 Confounder Dictionary Construction.

Collecting confounder images in 218 different images and UAV conditions is extremely challenging. To address this, we fully leverage multi-modal knowledge by constructing and initializing the confounder dictionary using textual prompts. Specifically, we employ the large language model GPT [1] to generate descriptive texts for various confounders, such as "a photo of a car on a rainy day without occlusion from a rear view." The confounders include weather conditions (sunny, rainy, foggy,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: The 36 prompt templates used in our method, each describing a [CLS] token in various UAV imaging conditions including weather, occlusion, scale, and viewpoint.

#	Prompt Template
1	a [CLS] in a sunny scene with no occlusion, viewed from the front at a large scale.
2	a [CLS] in a sunny scene with no occlusion, viewed from the side at a medium scale.
3	a [CLS] in a sunny scene with no occlusion, viewed from the rear at a small scale.
4	a [CLS] in a sunny scene with partial occlusion, viewed from the top at a large scale.
5	a [CLS] in a sunny scene with partial occlusion, viewed from the front at a medium scale.
6	a [CLS] in a sunny scene with partial occlusion, viewed from the side at a small scale.
7	a [CLS] in a sunny scene with heavy occlusion, viewed from the rear at a large scale.
8	a [CLS] in a sunny scene with heavy occlusion, viewed from the top at a medium scale.
9	a [CLS] in a sunny scene with heavy occlusion, viewed from the front at a small scale.
10	a [CLS] in a rainy scene with no occlusion, viewed from the side at a large scale.
11	a [CLS] in a rainy scene with no occlusion, viewed from the rear at a medium scale.
12	a [CLS] in a rainy scene with no occlusion, viewed from the top at a small scale.
13	a [CLS] in a rainy scene with partial occlusion, viewed from the front at a large scale.
14	a [CLS] in a rainy scene with partial occlusion, viewed from the side at a medium scale.
15	a [CLS] in a rainy scene with partial occlusion, viewed from the rear at a small scale.
16	a [CLS] in a rainy scene with heavy occlusion, viewed from the top at a large scale.
17	a [CLS] in a rainy scene with heavy occlusion, viewed from the front at a medium scale.
18	a [CLS] in a rainy scene with heavy occlusion, viewed from the side at a small scale.
19	a [CLS] in a foggy scene with no occlusion, viewed from the rear at a large scale.
20	a [CLS] in a foggy scene with no occlusion, viewed from the top at a medium scale.
21	a [CLS] in a foggy scene with no occlusion, viewed from the front at a small scale.
22	a [CLS] in a foggy scene with partial occlusion, viewed from the side at a large scale.
23	a [CLS] in a foggy scene with partial occlusion, viewed from the rear at a medium scale.
24	a [CLS] in a foggy scene with partial occlusion, viewed from the top at a small scale.
25	a [CLS] in a foggy scene with heavy occlusion, viewed from the front at a large scale.
26	a [CLS] in a foggy scene with heavy occlusion, viewed from the side at a medium scale.
27	a [CLS] in a foggy scene with heavy occlusion, viewed from the rear at a small scale.
28	a [CLS] in a night scene with no occlusion, viewed from the top at a large scale.
29	a [CLS] in a night scene with no occlusion, viewed from the front at a medium scale.
30	a [CLS] in a night scene with no occlusion, viewed from the side at a small scale.
31	a [CLS] in a night scene with partial occlusion, viewed from the rear at a large scale.
32	a [CLS] in a night scene with partial occlusion, viewed from the top at a medium scale.
33	a [CLS] in a night scene with partial occlusion, viewed from the front at a small scale.
34	a [CLS] in a night scene with heavy occlusion, viewed from the side at a large scale.
35	a [CLS] in a night scene with heavy occlusion, viewed from the rear at a medium scale.
36	a [CLS] in a night scene with heavy occlusion, viewed from the top at a small scale.

nighttime), occlusion levels (none, partial, heavy), and viewing perspectives (front, side, rear, top). In this way, we systematically generate linguistic priors for confounder modeling, thus providing rich semantic support for downstream tasks, as shown in Table 1.

B More Experiment Results

Evaluation metrics. To evaluate the detection performance of our proposed enhanced model, we use several metrics: AP, AP50 and AP75 [1, 4]. The following parameters are utilized: TP (true positives), FP (false positives), and FN (false negatives). Intersection over Union (IoU) measures the overlap between the predicted bounding box and the ground truth box. Precision is defined as the ratio of true positive predictions to the total number of detected samples, calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall represents the ratio of true positive predictions to the total number of actual positive samples, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

The average precision (AP) is the area under the precision-recall curve, computed by:

$$AP = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{6}$$

Mean average precision (mAP) is obtained by averaging the AP values across all sample categories to measure the model's performance across all categories:

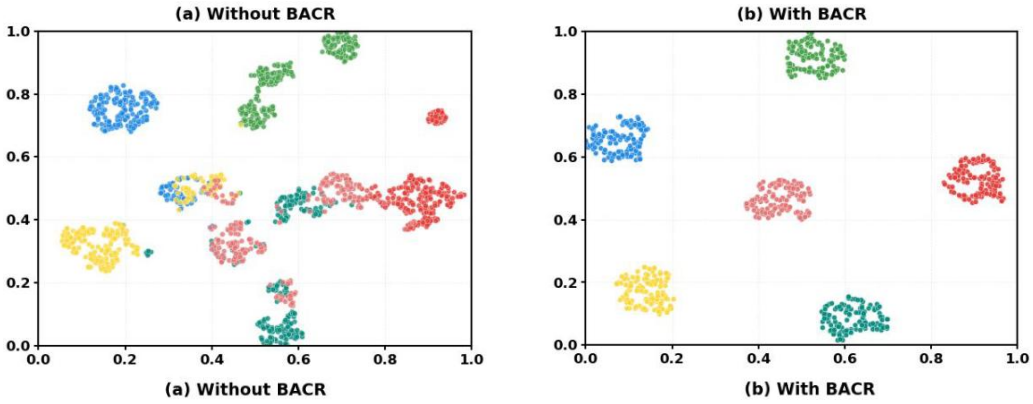


Figure 1: Visualization of t-SNE with and without BACR module.

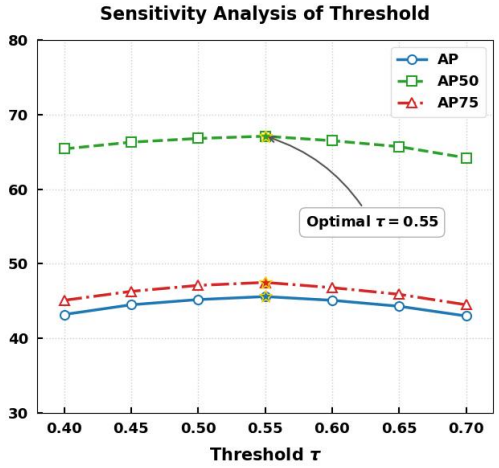


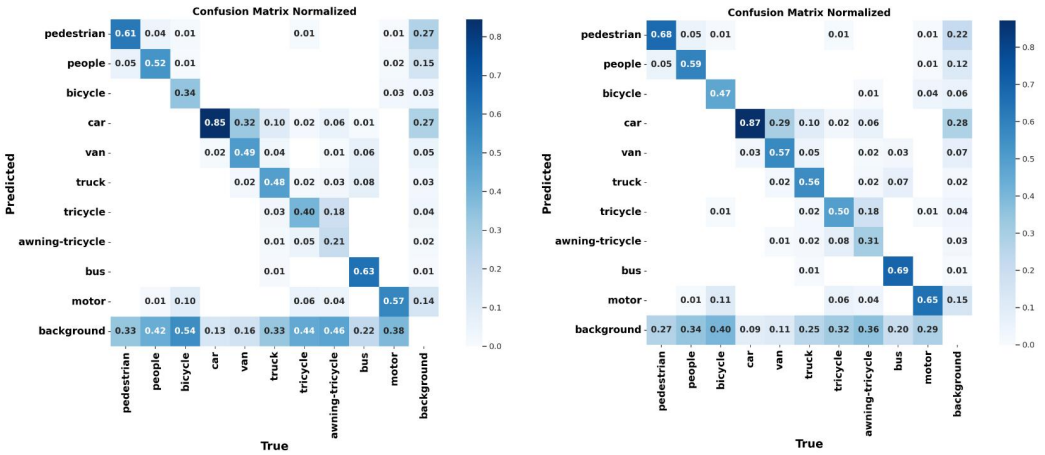
Figure 2: Sensitivity analysis of threshold τ.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{7}$$

Here, AP_i represents the AP value for category i, and N is the number of categories in the training dataset (in this paper, $N = 10$). AP50 denotes the average precision when the IoU threshold is set to 0.5, while AP75 represents the average precision over IoU thresholds to 0.75.

t-SNE visualization on the effect of BACR module. To validate the effectiveness of the BACR module, we visualize the category-wise features on the VisDrone dataset using t-SNE [5], as shown in Fig. 1. In the figure, the points with different colors represent the features with different object categories. It can be observed that without the BACR module, the features of the same category are distributed more loosely, and some categories are easily confused. This is mainly due to the inconsistent appearance of UAV-captured objects under varying imaging conditions and distances. In contrast, with the BACR module applied, the features of the same category become more compact and coherent, while the features of different categories are more distinguishable. These results demonstrate that the BACR module effectively enhances intra-class compactness and inter-class separability, leading to more robust and discriminative feature representations for UAV-based object detection.

Sensitive analysis. In our method, there are not many parameters. The only adjustable parameter is τ in Eq.(5). To investigate the impact of different values τ on detection performance, we conducted a parameter sensitivity analysis as shown in Figure 2. We systematically tested τ values within the



(a) YOLOv8 (b) MCR-UOD

Figure 3: (a) Confusion matrix plot of YOLOv8; (b) confusion matrix plot of our model.

range [0.4, 0.7]. The results show that both excessively small and large τ values lead to reduced accuracy. Overly small τ values introduce more false negatives by including inaccurate regions, while overly large τ values produce false positives by missing valid detection areas. The optimal performance is achieved at τ=0.55, which is consequently selected for our experiments. At the same time, we can also observe that as the τ value changes, the performance changes are also flat. It confirms that the value of τ is not very sensitive to performance.

Confusion matrix. From Fig. 3, it can be seen that the diagonal region of the confusion matrix for MCR-UOD is darker in color compared to YOLOv8, indicating that our proposed method has improved the model's ability to correctly predict object categories. This improvement is particularly notable when detecting smaller objects, such as bicycles, tricycles, and awning-tricycles, where our method outperforms YOLOv8. Although there are still some missed detections for these smaller objects in complex backgrounds, our method significantly reduces the proportion of objects misclassified as background compared to YOLOv8. Bicycles, tricycles, and awning-tricycles often appear in dense or occluded environments, making detection in complex backgrounds challenging. Our method improves the feature extraction ability and classification mechanisms of the model, leading to better detection performance and reduced missed detection rates for these small objects. Although the percentage of correctly predicted small objects still needs improvement, our method shows a notable advancement in performance over the traditional YOLOv8 model in complex scenarios.

Visualization of feature maps. The heatmap visualization of feature maps, shown in Fig. 4, highlights the superior performance of the MCR-UOD method compared to YOLOv8, SPAR [3] and UFPMP [2]. The MCR-UOD heatmaps demonstrate more precise and concentrated activation areas, especially for small objects. This indicates a more refined understanding and localization of critical features in the image. In contrast, UFPMP and SPAR, the previous state-of-the-art methods, while effective, show less focus on these smaller targets. This suggests that MCR-UOD is particularly adapted to capture essential information, leading to enhanced detection and classification performance, especially in scenarios involving small objects.

Precision-Confidence curve. Fig. 5(left) presents the Precision-Confidence (PC) curves for the MCR-UOD method, the baseline YOLOv8 and the state-of-the-arts SPAR and UFPMP. The MCR-UOD curve consistently demonstrates high precision across various confidence thresholds, indicating its effectiveness in reducing false positives. In contrast, UFPMP and SPAR exhibit more variability, reflecting less precision stability with changing confidence levels. The smooth and upward trend of the MCR-UOD curve highlights its superior performance and robustness, maintaining a high true positive rate as confidence increases. This comparison underscores the effectiveness of MCR-UOD in balancing precision and confidence.

Precision-Recall curve. Fig. 5(right) presents a comparative analysis of Precision-Recall (PR) curves for the MCR-UOD method, YOLOv8, SPAR and UFPMP. The PR curves clearly illustrate

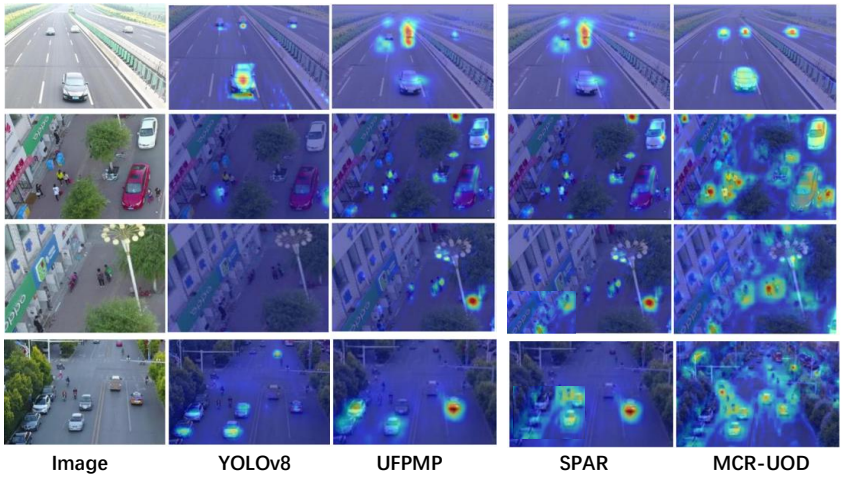


Figure 4: Visualization of feature maps.

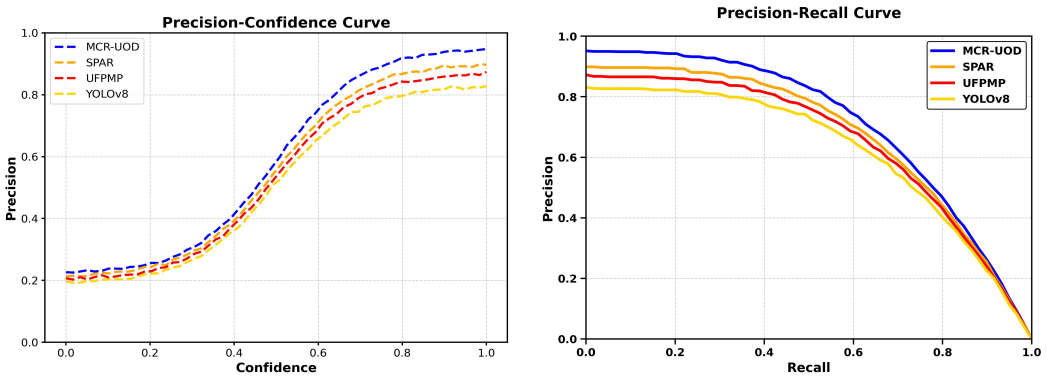


Figure 5: Comparisons of Precision-Confidence and Precision-Recall curves between MCR-UOD and other SOTA methods.

the performance of each model at different recall levels. Our MCR-UOD method consistently demonstrates superior precision compared to YOLOv8, SPAR and UFPMP at various recall rates. This indicates that the MCR-UOD method is more effective in minimizing false positives while maintaining high recall performance. In particular, the PR curve for MCR-UOD is higher than those of other methods in the recall spectrum, reflecting its improved accuracy and robustness in object detection. The area under the PR curve (AP) for MCR-UOD is significantly larger than that of YOLOv8, SPAR and UFPMP, further validating the effectiveness of our method. This improvement in AP underscores MCR-UOD's ability to achieve better precision and recall balance, particularly in detecting small objects and handling imbalanced datasets. Overall, the comparison reveals that MCR-UOD not only surpasses YOLOv8, SPAR and UFPMP in precision but also offers a more reliable detection performance. This indicates that the proposed MCR-UOD method provides substantial enhancement in object detection capabilities, making it more suitable for practical applications where high precision and recall are critical.

Statistical verification. To further validate the performance advantage of our proposed MCR-UOD framework, we conducted a statistical significance test against SPAR using the Wilcoxon signed-rank test, as shown in Tabel 2. This test, widely used for paired comparison without assuming data normality, allows us to assess whether the observed improvements are statistically meaningful. We perform the analysis on the UAVDT dataset using two key evaluation metrics: AP50 and AP75. The computed p-values are reported in the corresponding table. Notably, both p-values are well below the 0.1 significance level, providing strong evidence that the performance gains of MCR-UOD

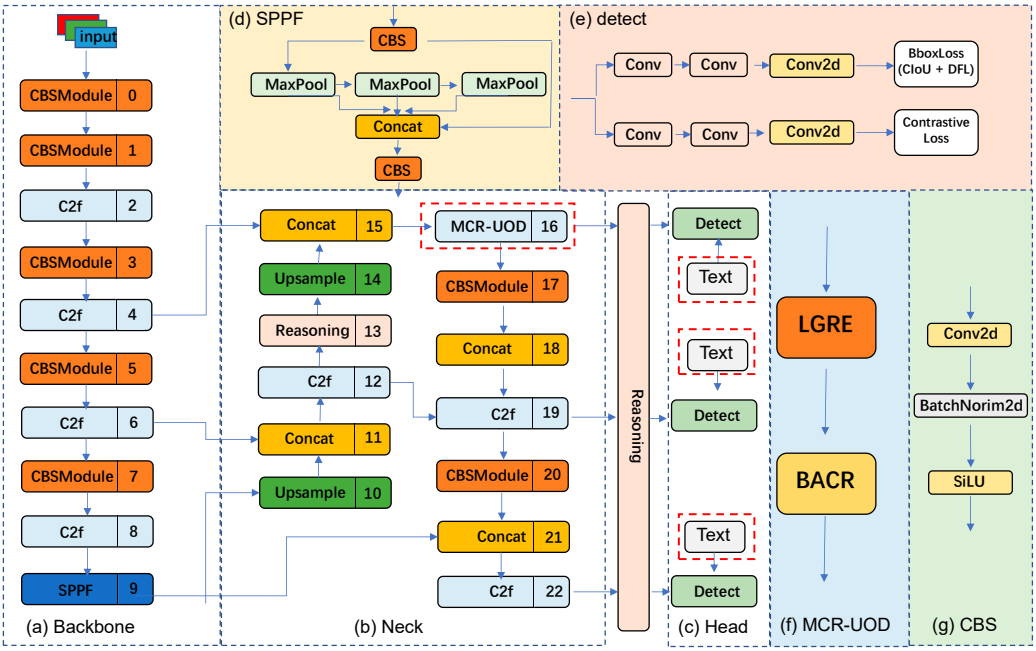


Figure 6: The network structure of YOLOv8 with MCR-UOD. The w (width) and r (ratio) are parameters used to represent the size of the feature map. The size of the model can be controlled by setting the values of w and r to meet the needs of different application scenarios.

Table 2: Statistical significance (p-values) of performance differences between MCR-UOD and SPAR.

Metric	SPAR (Mean ± Std)	MCR-UOD (Mean ± Std)	p-value
AP50	43.90 ± 0.25	44.70 ± 0.31	0.018
AP75	34.70 ± 0.28	35.60 ± 0.34	0.015

over SPAR are not due to random variation. These findings confirm the robustness and consistent superiority of our causal reasoning approach to enhance UAV-based object detection.

C Model Architecture with YOLOv8

We implemented the proposed MCR-UOD method based on the YOLOv8 detection framework. The overall architecture is illustrated in Figure 6. YOLOv8 adopts a modern and streamlined structure composed of a backbone, neck, and detection head, offering improvements in both detection accuracy and speed over previous YOLO versions such as YOLOv5 and YOLOv7.

In our implementation, we retain the original backbone of YOLOv8 and focus on modifying the neck and detection head to incorporate our MCR-UOD strategy. As highlighted in the red box in Figure 6, we replace the last C2f module processing the low-level feature map before the head with a customized version. Specifically, the input feature C_1 is passed through two newly designed modules: LGRE and BACR. The output of this process, denoted as C_1^n, is then fed into the detection head.

Furthermore, we replace the original classification head with a contrastive head based on text embeddings, as shown in the figure. This change enables the model to perform self-prompted open-set recognition by leveraging text-based semantic information, allowing its potential generalization to unseen object categories.

D Limitations

Our method uses multimodal knowledge and causal reasoning to improve object detection on UAV imagery. Although it shows promising results, there are limitations. First, relying on CLIP for semantic guidance limits performance due to its representational capacity, particularly in low-quality or ambiguous images. In addition, prompt design is based on intuition and heuristics, limiting adaptability. Second, the integration of causal reasoning with object detection is still in the early stages. Although we use structural equation models for causal modeling, more research is needed to better link causal structures with image features, especially in complex environments.

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[2] Yecheng Huang, Jiaxin Chen, and Di Huang. Ufpmp-det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1026–1033, 2022.

[3] Nianxin Li, Mao Ye, Lihua Zhou, Song Tang, Yan Gan, Zizhuo Liang, and Xiatian Zhu. Self-prompting analogical reasoning for uav object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18412–18420, 2025.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.