Journal of Electronic Research and Application, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Failure Complementarity of Temporal and Single-Frame YOLO Detectors under Severe Motion Blur

Zhe Feng*

Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA

**Corresponding author: Zhe Feng, zf87@duke.edu*

Abstract: Object detection under high-speed motion remains challenging due to severe motion blur, which degrades spatial appearance and limits the effectiveness of single-frame detectors. While temporal modeling is widely explored to enhance performance, its specific behavior under extreme motion blur is not yet fully characterized. In this work, we conducted an experimental study comparing a single-frame YOLOv8n detector and a temporal-enhanced variant that incorporates multi-frame inputs and frame-difference cues within the backbone. Results on a highly blurred table tennis dataset show that although the temporal-enhanced model and single-frame baseline achieve similar average precision (AP~0.52), they exhibit markedly different failure modes. Quantitative analysis reveals a Jaccard index of only 0.43, demonstrating a pronounced complementarity between the detection outcomes of the two models. By exploiting this behavioral divergence through a simple ensemble strategy, we achieve a substantial aggregate performance gain, increasing AP50 from 0.52 to 0.73. These findings suggest that under extreme blur, temporal modeling can induce complementary detection behavior beyond improving individual detector accuracy, offering an alternative perspective for designing robust detection systems in highly degraded visual environments.

Keywords: Motion blur; Object detection; Temporal modeling; Detection complementarity; Ensemble; YOLO

Online publication: April 3, 2026

1. Introduction

With the increasing prevalence of vision systems operating in dynamic environments, object detection under high-speed motion has become increasingly important. In many practical scenarios, objects move rapidly relative to the camera, placing stringent demands on detection accuracy and robustness. However, fast motion often leads to motion blur in captured images due to limitations in exposure time, shutter speed, and sensor characteristics [1,2]. Motion blur degrades spatial details and distorts object appearance, significantly impairing detection accuracy [3]. While reducing exposure time can mitigate such blur, it may also introduce noise and underexposure [4]. As a consequence, motion blur remains a challenging factor for modern object detection systems.

Existing research addressing motion blur in object detection can be broadly grouped into several directions. One approach focuses on image restoration, aiming to recover sharp images prior to detection. Representative

methods employ deep neural networks or motion flow constraints to model and remove blur [5–9]. While these restoration-based techniques can improve visual quality, they often introduce additional computational cost and complexity to the detection pipeline due to the extra pre-processing stage. Another approach optimizes detectors to operate directly on blurred inputs without explicit restoration. Strategies such as knowledge distillation, dual-stream feature fusion, and attention mechanisms have been developed to enhance feature robustness against blur [10–12]. By integrating these mechanisms into the detection backbone, these recognition-oriented methods avoid the latency and complexity of multi-stage restoration-and-detection pipelines.

Within the recognition-oriented frameworks, an increasingly prominent direction involves the integration of temporal information across consecutive frames to mitigate motion-induced degradation. Representative work by Zhu *et al.* introduced flow-guided feature aggregation, which leverages optical flow to align and aggregate features from nearby frames, effectively enhancing feature quality against motion blur and video defocus [13]. Following this trajectory of utilizing temporal context, Alqaysi *et al.* demonstrate that a simplified temporal stacking approach within YOLO architectures can also enhance the detection of small moving targets [14]. Building on this, Fokkinga *et al.* further adapts temporal modeling for aerial platforms, introducing global motion compensation to handle camera movement and a history-only frame strategy to enable low-latency real-time inference [15].

Despite these architectural advances, the evaluation paradigm within this body of literature remains relatively narrow. The aforementioned studies predominantly focus on designing more effective fusion architectures, interpreting their value primarily through the quantitative enhancement of aggregate metrics. Under this prevailing perspective, temporal modeling is treated largely as a mechanism for incrementally increasing average detection success rates, obscuring the specific behavioral shifts in failure patterns.

Undoubtedly, aggregate metrics such as average precision (AP) provide a comprehensive summary of overall detection capability and remain the standard for performance evaluation [16,17]. However, relying solely on these averages may obscure the fine-grained behavioral dynamics, as global statistics cannot distinguish between uniform improvements and structural shifts in detection patterns. Given that temporal modeling introduces decision cues that are distinctly different from those available in single-frame spatial evidence, such integration may not merely reinforce the spatial baseline [18,19]. Instead, it could potentially reshape the sample-level distribution of successes and failures, enabling the detector to succeed on previously missed samples while simultaneously introducing new failures on others. If this behavioral divergence fundamentally reshapes failure characteristics, changing which samples are detected, combining detectors with such complementary behaviors can yield substantial performance improvements relative to any individual model [20].

Nevertheless, the empirical characterization of such behavioral divergence remains limited. It is still unclear whether temporal integration under severe motion blur systematically reshapes failure patterns and whether such changes could yield meaningful, exploitable complementarity with single-frame baselines.

In this paper, we investigated object detection under severe motion blur by shifting the focus from aggregate metrics to detector failure patterns. We conducted a comparative analysis between single-frame baselines and detectors incorporating temporal information across consecutive frames to quantify the behavioral divergence in their detection outcomes. Our empirical results demonstrated that incorporating temporal cues reshapes the distribution of failures, creating a marked complementarity with spatial-only models. Furthermore, we validated that this orthogonality can be effectively leveraged through simple ensembles to achieve significant system-level performance gains without complex fusion strategies.

The main contributions of this paper are summarized as follows:

(1) We demonstrate that under severe motion blur, analyzing detectors from the perspective of failure modes serves as a valuable complement to standard aggregate metrics, revealing behavioral patterns and failure distributions that are obscured by global average statistics;
(2) We empirically confirm that incorporating temporal information induces failure patterns that are structurally different from single-frame baselines. This behavioral orthogonality validates that temporal cues provide independent evidence rather than merely refining spatial features;
(3) We demonstrate that this failure complementarity can be directly leveraged to achieve substantial system-level gains. A simple ensemble strategy yields a 39.9% relative improvement in AP50 (from 0.52 to 0.73), offering a practical and cost-effective solution for robust detection in motion-blurred environments.

2. Materials and methods

This study establishes a controlled experimental framework to evaluate the behavioral divergence between a spatial-only baseline and a temporal-enhanced detector. To isolate the specific contribution of temporal evidence, we employed a standard YOLOv8n model as the fixed reference and constructed a temporal variant through minimal backbone augmentation, while maintaining identical detection heads and training protocols. The comparative analysis utilizes a dual-metric evaluation system, combining standard aggregate performance indicators with a quantitative overlap analysis to explicitly characterize the orthogonality of failure patterns.

2.1. Baseline detector

To establish a consistent reference for analyzing detection behavior under severe motion blur, we adopted the standard YOLOv8n detector as the baseline model [21]. This architecture operates exclusively on individual RGB frames, deriving detection decisions solely from spatial appearance cues. The model retains its default configuration, including the CSP-based backbone, feature fusion neck, and anchor-free detection head. As a spatial-only detector, this baseline represents the standard recognition capability available from static observations, serving as the fixed reference point for characterizing the specific impact of temporal cues on failure patterns.

2.2. Temporal backbone augmentation

To investigate the impact of temporal evidence on detection outcomes, the baseline architecture is instantiated with a joint temporal injection mechanism, as illustrated in **Figure 1**. The augmentation is restricted to the backbone stage, while the remainder of the detection pipeline remains unchanged. This setup provides the network with access to temporal continuity cues that are physically absent in the single-frame baseline.

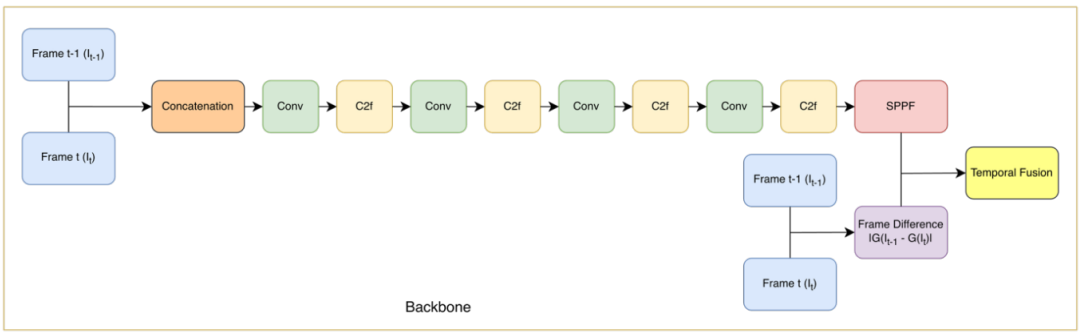

Figure 1. Pipeline of the proposed temporal-enhanced backbone architecture.

The temporal appearance context is established by expanding the input space to include historical visual data. We concatenate the current frame $I_t \in R^{H\times W\times 3}$ and its immediately preceding frame $I_{t-1} \in R^{H\times W\times 3}$ along the channel dimension to form a six-channel input:

$$X_t = Concat\,(I_{t\text{-}1}, I_t) \in R^{H\times W\times 6}$$

By processing this joint representation, the feature extractor gains the capacity to reference the immediate past context to resolve spatial ambiguities present in the current blurred observation.

To process this expanded input space, the first convolution layer of the pretrained backbone is adapted. Let the original first-layer kernel be $W^{(0)} \in R^{C\times 3\times k\times k}$, where C is the number of output channels. The expanded kernel $W_{6ch}^{(0)} \in R^{C\times 6\times k\times k}$ is initialized via scaled duplication:

$$W_{6\text{ch}}^{(0)} = \left[\frac{1}{2}W^{(0)}, \frac{1}{2}W^{(0)}\right]$$

In coordination with multi-frame stacking, we explicitly encode motion magnitude and location by computing a frame-difference representation $d_t \in R^{H\times W}$ as a lightweight motion-sensitive cue. We denote a standard grayscale conversion operator by G(·), mapping an RGB frame to a single-channel intensity image. The frame difference is defined as

$$d_t = |G(I_{t\text{-}1}) - G(I_t)|$$

The frame-difference map is injected at the Spatial Pyramid Pooling-Fast (SPPF) layer, the final stage of the YOLO backbone. This integration point is selected to align the motion evidence with high-level semantic features. Since severe motion blur often distorts the global shape and structure of the object, modulating the feature space at this semantic bottleneck is designed to allow the motion cue to influence the object-level representation.

To align spatial resolutions, the frame-difference map is resized via bilinear interpolation:

$$\widetilde{d}_t = \mathrm{R}(d_t; \mathrm{h}, w) \in R^{1\times \mathrm{h}\times w}$$

The resized motion map $\widetilde{d}_t$ is concatenated with the SPPF feature map:

$$Z_t = Concat\left(F_t^{SPPF}, \widetilde{d}_t\right) \in R^{(C+1)\times \mathrm{h}\times w}$$

where $F_t^{SPPF} \in R^{C\times h\times w}$ is the SPPF output feature map.

Feature fusion is performed using a 1×1 convolution without bias to generate the augmented SPPF features:

$$\widehat{F}_t^{\mathrm{SPPF}} = W_{fuse} * Z_t$$

where $W_{fuse} \in R^{C\times(C+1)\times 1\times 1}$ is the fusion kernel.

To ensure the optimization trajectory begins from a valid spatial feature representation, the fusion kernel is initialized in an identity-preserving manner:

$$W_{\mathrm{fuse}}[i,j,0,0] = \begin{cases} 1, i=j<C \\ 0, \text{otherwise} \end{cases}$$

Under this initialization, $\widehat{F}_t^{\mathrm{SPPF}}$ initially matches F_t^{SPPF} ensuring that the introduction of the motion map does not alter the backbone feature representation at initialization.

2.3. Evaluation metrics

The fundamental correctness of a detection is established using the Intersection-over-Union (IoU) metric between a predicted bounding box B_p and the ground-truth bounding box B_{gt}, defined as:

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}$$

A prediction is classified as correct if $IoU(B_p, B_{gt}) \geq 0.5$. Based on this criterion, we compute precision and recall as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

where TP, FP, FN denotes the numbers of true positives, false positives, and false negatives respectively.

To quantify the overall aggregate performance, we utilize AP [17]. AP summarizes the precision-recall relationship by integrating precision over recall. Let P(r) denotes precision as a function of recall r, then AP can be expressed as:

$$\mathrm{AP} = \int_0^1 P(r)\ dr$$

In this work, we report AP50, which is the AP computed using the IoU threshold of 0.5. Since the dataset contains a single object category, AP50 is equivalent to mAP50 under this setting.

To explicitly quantify the behavioral orthogonality between models, we employ the Jaccard Index, which is a classical set-theoretic statistic utilized to quantify the similarity between finite sample sets, on the sets of successfully detected frames [22]. Let S_A and S_B denote the sets of unique frame indices correctly detected by Model A and Model B, respectively. The Jaccard similarity is defined as:

$$J(A,B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$$

A lower Jaccard value indicates that the intersection of successful detections is small relative to the union, implying that the models succeed on disjoint subsets of data. This metric serves as a direct quantitative proxy for behavioral complementarity.

2.4. Dataset

The dataset used in this study is constructed from video footage of a table tennis match recorded with a fixed camera viewpoint. We select table tennis as the experimental domain because it isolates high-velocity motion as the primary degradation factor. The ball's rapid acceleration creates severe directional smear while the background remains relatively static, providing a controlled setting to decouple motion blur from other environmental complexities. The camera remains stationary throughout the recording, capturing continuous rally segments in which the ball undergoes rapid motion. Frames within each rally are temporally consecutive, while brief discontinuities occur only between separate rallies. The dataset contains a single object category, namely the table tennis ball, which is manually annotated using axis-aligned bounding boxes.

In total, the dataset consists of 2010 annotated frames, of which 1710 frames are used for training and 300 frames for validation. All models evaluated in this work share the same data split to ensure consistency across experiments.

A defining characteristic of this dataset is the presence of severe motion blur caused by high-speed ball

motion. In many frames, the ball exhibits pronounced elongation along its direction of motion, resulting in substantial distortion of its spatial appearance. To characterize the severity of motion blur, we introduce the blur elongation ratio ρ as a quantitative proxy for degradation severity. By treating the ball's projection as elliptical, the blur elongation ratio is computed based on the ratio between its major and minor axes. This ratio serves as a compact proxy for the degree of motion-induced deformation, independent of object scale.

Table 1 summarizes the distribution of motion blur severity across the dataset, categorized from near-circular appearances to extreme blur cases. Notably, samples with extreme elongation (elongation ratio greater than 3) account for approximately 55% of the dataset, while the remaining samples are distributed across lower blur severity ranges. This distribution ensures that the dataset contains a high density of samples where spatial appearance is maximally compromised, effectively stressing the baseline detector's reliance on texture and shape integrity.

Table 1. Distribution of motion blur severity in the dataset

Blur level	Elongation ratio range	Number of samples	Percentage (%)
Near circle	1.0–1.2	70	3.5
Mild blur	1.2–1.5	219	10.9
Moderate blur	1.5–2.0	259	12.9
Severe blur	2.0–3.0	354	17.6
Extreme blur	> 3.0	1108	55.1

Representative examples of different motion blur levels are shown in **Figure 2**. For visualization consistency, all samples are presented using fixed-size cropped regions centered on the ball.

Figure 2. Representative samples from the motion-blurred table tennis dataset.

2.5. Experimental setup

All experiments were conducted on the Windows 10 system equipped with an AMD Ryzen 9 9955HX 16-Core Processor (2.50 GHz) and an NVIDIA GeForce RTX 5070 Ti Laptop GPU with 12GB of memory. The models were implemented using the Ultralytics YOLO framework, built upon PyTorch 2.11.0 and CUDA 12.8 for

hardware acceleration.

All detector configurations are trained under identical optimization settings, including optimizer type, learning rate schedule, batch size, and number of training epochs. No model-specific hyperparameter tuning is performed, ensuring that all comparisons are conducted under a controlled and consistent training setup.

For the temporal detector configuration, multiple independent training runs with different random initializations are performed to assess the variability of detection outcomes.

3. Results

3.1. Comparison between single-frame baseline and temporal-enhanced detector

We first evaluate the impact of temporal information using standard aggregate metrics. **Table 2** compares the detection performance of the single-frame baseline (YOLOv8n) and the temporal-enhanced detector. The results indicate that, in terms of global statistics, the two models exhibit nearly identical performance. The introduction of temporal cues does not produce a distinct improvement in AP50, precision, or recall under severe motion blur.

Table 2. Comparison between the single-frame and temporal-enhanced detectors

Model	AP50	Precision	Recall
YOLOv8n	0.5233	1.0000	0.5233
Temporal-enhanced detector	0.5212	0.9576	0.5267

This observation motivates a closer examination of detector behavior beyond aggregate metrics. While the overall success rates are similar, an analysis based on detection outcome partitioning reveals that the underlying sets of successful detections are substantially different. We explicitly decompose the detection outcomes into the common intersection and the model-specific exclusive sets.

As detailed in **Table 3**, the two models share an intersection of 95 frames. In contrast, the symmetric difference between the two success sets, comprising detections exclusive to either the baseline or the temporal model, totals 125 frames. Notably, this magnitude significantly exceeds the intersection count, indicating that instances of behavioral disagreement outweigh instances of consensus. This distribution confirms that the divergence in detection behavior is not a minor edge case but the dominant characteristic of the comparison. Such a pattern implies that under the severe motion blur, the baseline and temporal-enhanced detectors operate on fundamentally different feature dependencies, frequently succeeding in mutually exclusive scenarios.

Table 3. Overlap statistics between the baseline and temporal-enhanced detectors

Model A vs Model B	Intersection	A \ B	B \ A	A Δ B	Jaccard
YOLOv8n vs Temporal	95	62	63	125	0.4318

To formally summarize this degree of overlap as a standardized metric, we compute the Jaccard index. As reported in **Table 3**, the Jaccard index is 0.4318. Since this value falls below the theoretical majority threshold of 0.5, it mathematically confirms that the symmetric difference dominates the intersection, thereby validating the structural distinctness of the two detection mechanisms.

To confirm that this divergence is structurally induced by the temporal modality rather than by training

stochasticity, we analyze the overlap between two independently trained baseline detectors. As shown in **Table 4**, two spatial-only models exhibit a substantially higher degree of overlap (Jaccard index of 0.8659). This sharp contrast demonstrates that training randomness alone only induces minimal variation in failure patterns.

Table 4. Overlap statistics between two independently trained baseline detectors

Model A vs Model B	Intersection	A \ B	B \ A	A Δ B	Jaccard
YOLOv8n-1 vs YOLOv8n-2	142	15	7	22	0.8659

Furthermore, **Table 5** extends this analysis to multiple independent runs. The Jaccard indices between baseline and temporal models remain consistently low (~0.40), whereas comparisons between different temporal models show high consistency (~0.68). This systematic pattern verifies that the shift in failure distribution is an intrinsic property of the temporal augmentation. The incorporation of temporal information fundamentally reshapes the detection surface, creating a complementary failure profile that aggregate metrics alone fail to capture.

Table 5. Pairwise overlap statistics among baseline and temporal-enhanced detectors

Model A vs Model B	Intersection	A \ B	B \ A	A Δ B	Jaccard
YOLOv8n-1 vs M1	81	76	51	127	0.3894
YOLOv8n-1 vs M2	85	72	53	125	0.4048
YOLOv8n-1 vs M3	93	64	56	120	0.4366
YOLOv8n-1 vs M4	95	62	63	125	0.4318
M1 vs M2	107	25	31	56	0.6564
M1 vs M3	115	17	34	51	0.6928
M1 vs M4	119	13	39	52	0.6959
M2 vs M3	113	25	36	61	0.6494
M2 vs M4	119	19	39	58	0.6723
M3 vs M4	132	17	26	43	0.7543

3.2. Ensemble results

The profound behavioral divergence identified above implies that spatial and temporal models capture non-redundant information. We validate the practical utility of this complementarity through an ensemble analysis, where detections from multiple models are merged via standard non-maximum suppression (NMS) [23]. This analysis serves to confirm that the observed orthogonality translates into system-level robustness.

Table 6 reports the detection performance of individual baseline detectors alongside ensemble configurations. Among the single-model baselines, YOLOv8n and YOLOv11n exhibit comparable AP50 and recall, with both outperforming the older YOLOv5n architecture. However, the performance saturation observed between different versions of YOLO model suggests that under severe motion blur, improvements in spatial feature extraction yield diminishing returns.

Table 6. Comparison of individual models and ensemble configurations.

Model	AP50	Precision	Recall
YOLOv5n	0.4856	0.9865	0.4867
YOLOv8n	0.5233	1.0000	0.5233
YOLOv11n	0.5246	0.9814	0.5267
B+B	0.5398	0.9878	0.5400
T+T	0.5736	0.9508	0.5800
B+T	0.7323	0.9821	0.7333

Building on these individual baselines, we examine homogeneous ensembles formed by combining models of the same architecture. We denote the ensemble of two independently trained baseline detectors (YOLOv8n) as "B+B", and the ensemble of two independently trained temporal-enhanced detectors as "T+T". The B+B ensemble yields only marginal gains over the single baseline, indicating that independent spatial models suffer from correlated failures. Similarly, the T+T ensemble shows limited improvement, consistent with the high behavioral overlap observed among temporal models in **Table 5**.

In sharp contrast, the heterogeneous ensemble combining the spatial baseline and the temporal-enhanced detector (denoted as "B+T") delivers a substantial performance breakthrough. As shown in **Table 6**, the B+T configuration achieves an AP50 of 0.7323 and a recall of 0.7333. Compared to the single-frame YOLOv8n baseline (AP50 = 0.5233), this represents an absolute increase of 0.209 and a relative improvement of 39.9%. Importantly, this improvement is realized without a drop in precision, indicating that the gain stems from the valid recovery of previously missed objects rather than an increase in false positives.

These results establish a direct link between the statistical orthogonality observed in the Jaccard analysis and tangible performance gains. The limited efficacy of B+B and T+T ensembles highlights that simply increasing model count or training diversity is insufficient when failure patterns remain structurally correlated. Conversely, the success of the B+T ensemble demonstrates that temporal modeling functions as a powerful source of orthogonal diversity. Under severe motion blur, the value of temporal cues extends beyond merely refining the spatial baseline; it offers a complementary detection mechanism that covers the blind spots of spatial-only processing. This finding suggests that for degradation-heavy scenarios, system design should prioritize the integration of structurally diverse experts, such as the spatial-temporal pair demonstrated here, as a critical strategy alongside the optimization of single-model metrics.

4. Conclusion

Conventional approaches typically interpret temporal modeling as a mechanism for boosting average precision. However, this study reveals that under severe motion blur, temporal information provides a distinct and critical value by inducing complementary failure patterns. Our analysis demonstrates that incorporating temporal cues does not merely reinforce spatial features but systematically reshapes detection behavior, creating a failure profile that is structurally orthogonal to spatial-only baselines. Empirically, while the temporal-enhanced detector performed similarly to the single-frame baseline in aggregate metrics, quantitative analysis confirmed a high degree of behavioral divergence (Jaccard index~0.43). By leveraging this orthogonality, a simple heterogeneous ensemble achieved a 39.9% relative improvement in AP50 (from 0.52 to 0.73) without sacrificing precision.

We conclude that for robust perception in high-degradation environments, integrating complementary experts represents a feasible and effective strategy. This approach expands the system design space, suggesting that optimizing behavioral diversity is a valuable pathway alongside the pursuit of individual model accuracy.

Disclosure statement

The author declares no conflict of interest.

References

[1] Potmesil M, Chakravarty I, 1983, Modeling Motion Blur in Computer-Generated Images. SIGGRAPH Comput. Graph. 1983(17): 389–399.

[2] Zhang X, Zhang T, Yang Y, et al., 2020, Real-Time Golf Ball Detection and Tracking Based on Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2808–2813.

[3] Wu Q, Yang Y, Han R, et al., 2026, LDA-YOLO: A Lightweight Deblurring-Aware Network for Real-Time Object Detection in Blurred Aerial Images. Journal of Real-Time Image Processing 2026(23): 23.

[4] Moreira C, Ferreira L, Coelho P, 2025, A Comprehensive Review of Ball Detection Techniques in Sports. PeerJ Computer Science, 2025(11): e3079.

[5] Su S, Delbracio M, Wang J, et al., 2017, Deep Video Deblurring for Hand-Held Cameras, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Truong N, Lee Y, Owais M, et al., 2020, SlimDeblurGAN-Based Motion Deblurring and Marker Detection for Autonomous Drone Landing. Sensors, 2020(20): 3918.

[7] Sun R, Li X, Liu G, et al., 2025, An Improved MPRNet for the Improvement of Blurred Concrete Crack Images. Structures, 2025(73): 108416.

[8] Dai S, Wu Y, 2008, Motion from Blur. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 1–8.

[9] Gong D, Yang J, Liu L, et al., 2017, From Motion Blur to Motion Flow: A Deep Learning Solution for Removing Heterogeneous Motion Blur, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Cho S, Kim S, Jung S, et al., 2022, Blur-Robust Object Detection Using Feature-Level Deblurring via Self-Guided Knowledge Distillation. IEEE Access, 2022(10): 79491–79501.

[11] Li Q, Zhang Y, Fang L, et al., 2025, DREB-Net: Dual-Stream Restoration Embedding Blur-Feature Fusion Network for High-Mobility UAV Object Detection. IEEE Transactions on Geoscience and Remote Sensing, 2025(63): 1–18.

[12] Zhu X, Lyu S, Wang X, et al., 2021, TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2778–2788.

[13] Zhu X, Wang Y, Dai J, 2017, Flow-Guided Feature Aggregation for Video Object Detection, In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[14] Alqaysi H, Fedorov I, Qureshi F, et al., 2021, A Temporal Boosted YOLO-Based Model for Birds Detection around Wind Farms. Journal of Imaging 2021(7): 227.

[15] Fokkinga E, van Leeuwen M, Kuijf H, 2025, Air-to-Ground Real-Time Temporal Small Object Detection from a

Flying Platform, In Proceedings of the Artificial Intelligence for Security and Defence Applications, 2025(13679): 1367907.

[16] Everingham M, Van Gool L, Williams C, et al., 2010, The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 2010(88): 303–338.

[17] Lin T, Maire M, Belongie S, et al., 2014, Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision – ECCV 2014, 740–755.

[18] Simonyan K, Zisserman A, 2014, Two-Stream Convolutional Networks for Action Recognition in Videos, In Proceedings of the Advances in Neural Information Processing Systems, 27.

[19] Feichtenhofer C, Fan H, Malik J, et al., 2019, SlowFast Networks for Video Recognition, In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[20] Razinkov E, Saveleva I, Matas J, 2018, ALFA: Agglomerative Late Fusion Algorithm for Object Detection, In Proceedings of the International Conference on Pattern Recognition (ICPR), 2594–2599.

[21] Jocher G, 2023, YOLOv8: Ultralytics Next-Generation YOLO, https://github.com/ultralytics/ultralytics.

[22] Jaccard P, 1912, The Distribution of the Flora in the Alpine Zone. New Phytologist, 1912(11): 37–50.

[23] Girshick R, Donahue J, Darrell T, et al., 2014, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).