Trustworthy Temporal Feature Extraction for Object Detection and Tracking in Drone Video Streams

Executive Summary

The proliferation of drone technology has underscored the critical need for advanced AI-driven object detection and tracking capabilities in aerial video streams. However, the unique characteristics of drone footage—including significant ego-motion, the presence of tiny objects, pervasive motion blur, and dynamic, cluttered backgrounds—pose substantial challenges to traditional frame-by-frame analysis, often resulting in inconsistent or "flickery" tracking results.[1] Temporal features, which encapsulate motion and changes over time, are indispensable for overcoming these obstacles, enabling more robust and consistent object identification and trajectory estimation.[2]

Beyond mere accuracy, real-world drone applications, such as surveillance, autonomous navigation, and search and rescue, demand trustworthy AI systems. This necessitates models that are inherently robust to adverse conditions, capable of quantifying the uncertainty in their predictions, and able to provide explainable decisions.[6] This report details how modern AI, particularly deep learning architectures like Transformers and Graph Neural Networks, coupled with sophisticated fusion strategies and dedicated trustworthiness techniques, are addressing these complex requirements to achieve reliable and verifiable drone video analysis.

1. Introduction: The Imperative of Temporal Features in Drone Video AI

The ability to accurately detect and track objects in real-time from drone video streams is critical for diverse applications, including surveillance, infrastructure inspection, search and rescue, and autonomous navigation. Unlike static image analysis, video streams inherently contain temporal information—the dynamics of motion, change, and interaction over time. Leveraging these temporal cues is paramount for robust object detection and tracking, especially when spatial (appearance-based) features alone are insufficient or unreliable.

Temporal features in video analysis refer to information derived from changes or motion across successive frames, capturing dynamic variations and relationships over time.[8] This information is fundamental for tracking objects by continuously following their trajectories across multiple frames and helps maintain object identity, particularly when appearance features are unreliable due to occlusions or rapid changes.[2] For drone video, motion features and temporal information are critical cues to discriminate targets (e.g., other drones) from complex backgrounds, as appearance features alone are often insufficient.[3]

However, drone video presents unique and significant challenges that complicate temporal feature extraction:

- **Ego-motion and Camera Movement:** Drones are inherently mobile platforms, leading to complex camera movements that introduce global motion into the video, making it difficult to distinguish target motion from camera motion.[3] This can lead to significant motion blur and perspective distortions.[3]
- **Small and Tiny Objects:** Objects viewed from a drone's aerial perspective often appear extremely small, sometimes occupying only a few pixels.[2] This low pixel count leads to insufficient appearance and texture information, making traditional detection challenging, a problem often exacerbated by network downsizing and pooling operations.[3]
- **Dynamic and Cluttered Backgrounds:** Drone videos frequently capture complex and dynamic backgrounds, such as moving trees, water ripples, or urban clutter, which can easily obscure targets or be confused with object motion.[3]
- **Motion Blur and Rapid Movement:** High drone speeds and target velocities, coupled with camera shake, cause significant motion blur and rapid, irregular movements, degrading image clarity and detail.[3]
- **Occlusion and Disappearance:** Objects may be partially or fully occluded, or temporarily disappear from the frame, leading to tracking failures and identity switches.[2]
- **Varying Perspectives and Scale Changes:** The changing distance and angle between the drone and the target result in significant variations in object size, aspect ratio, and texture details, affecting detection and re-identification.[2]

For drone applications, the traditional computer vision paradigm, which often heavily relies on rich visual appearance, becomes less effective. The dynamic aspect—how pixels or objects move over time—becomes the primary signal for detection and tracking. This implies that AI models for drone video must prioritize and excel at motion modeling, even if it means sacrificing some spatial detail. This also suggests that datasets for drone vision need to emphasize temporal consistency and motion characteristics more than static image datasets.

The challenges of ego-motion, small objects, motion blur, and dynamic backgrounds are not isolated problems; they frequently co-occur and exacerbate one another. For example, a small, fast-moving object captured by a moving drone will likely suffer from motion blur and be easily lost in a cluttered background. This necessitates integrated, multi-faceted approaches that can simultaneously compensate for ego-motion, enhance tiny object features, suppress dynamic backgrounds, and maintain identity through occlusions. This points towards complex, end-to-end learning frameworks that can fuse various types of

information (spatial, temporal, multi-modal) and handle uncertainty.
This report delves into the methodologies for extracting temporal features and, critically, how to ensure their trustworthiness. Trustworthiness in AI systems for high-stakes drone applications demands not only high accuracy and efficiency but also robustness to real-world perturbations, quantifiable uncertainty in predictions, and explainable decision-making processes.

2. Core Temporal Feature Extraction Techniques

Temporal features capture the dynamic changes within a video sequence, providing crucial information about object movement, interactions, and state transitions. These features are fundamental for tasks like object tracking, action recognition, and anomaly detection.

2.1. Traditional Motion-Based Approaches

Historically, motion analysis in computer vision has relied on techniques that directly compute pixel or block displacements between frames.

- Optical Flow:
  Optical flow estimates the apparent motion of each pixel or feature point between two consecutive video frames.8 The underlying assumption is that the brightness of an object remains constant between frames and that neighboring pixels have similar motion.10 Algorithms like Lucas-Kanade address the ill-posed problem of solving for two unknowns (horizontal and vertical motion components) from a single brightness constancy equation by assuming local constancy of optical flow within a small window.10 Optical flow provides a dense, pixel-level motion field, offering fine-grained information about object trajectories, making it a useful method for object tracking, motion analysis, and video compression.10
  However, optical flow computation, especially dense optical flow, is computationally expensive.[8] This can hinder real-time applications, particularly for high-resolution drone video streams. Traditional optical flow methods are also sensitive to noise, lighting changes, and motion blur, which are prevalent in drone videos.[3] The projection of 3D motion onto a 2D plane makes it an ill-posed problem, requiring additional constraints that can introduce inaccuracies.[23] Furthermore, when the camera itself is moving, as is the case with a drone, a global motion is added to local object motion, complicating the distinction between background and moving objects.[10] Combining optical flow with image correlation or Kalman filters can help address this issue.[10] Despite these limitations, deep learning models such as PWC-Net, RAFT, and FlowNet 2.0 have significantly improved optical flow estimation by leveraging convolutional neural networks (CNNs) and recurrent units, with RAFT, for instance, extracting per-pixel features and iteratively updating a flow field using correlation volumes.[21]

- Motion Vectors (from Video Codecs):
  Motion vectors describe the transformation from one 2D image to another, typically from adjacent frames in a video sequence.23 They are a core component of modern video compression standards like H.264/MPEG-4 AVC and HEVC (High Efficiency Video Coding).23 Instead of encoding every pixel, these codecs represent how 8x8 or 16x16 blocks of pixels move from frame to frame, exploiting temporal redundancy to reduce file sizes and transmission bitrates.22 This encoding process involves motion estimation and compensation to generate these vectors.24
  Motion vectors are inherently efficient because they are already computed during video compression, requiring significantly less memory (e.g., 1/64th for 8x8 blocks) than fully decoded frames.[22] This makes them highly scalable for self-supervised training of large video models.[22] Being part of standard video formats, they are readily available without additional computation.[22] While optical flow provides dense, pixel-level motion, motion vectors are sparser and block-based.[22] However, for applications like motion-guided masking (MGM), motion vectors can serve as a proxy for determining regions of interest, leveraging their efficiency over the computational cost of optical flow.[22] New feature vectors based on motion structure can also be used alongside motion vectors to improve performance.[25]

The evolution of temporal feature extraction reflects a fundamental transition from explicit, hand-engineered motion features to implicitly learned temporal representations within deep neural networks. This allows models to capture more complex, non-linear motion patterns and temporal dependencies that are difficult to define manually, paving the way for more sophisticated video understanding. Furthermore, the efficiency of motion vectors, which are a byproduct of video compression, highlights the latent value of compressed domain information for real-time systems. For drone applications, where computational resources and bandwidth are often constrained, leveraging compressed-domain features like motion vectors offers a significant advantage. This suggests a future where AI models might increasingly operate directly on compressed video streams or utilize their metadata, rather than fully decoding and processing raw pixel data, to achieve higher efficiency and lower latency.

2.2. Deep Learning for Temporal Modeling

Deep learning has revolutionized temporal feature extraction by enabling models to learn complex spatio-temporal patterns directly from raw video data.

- Recurrent Neural Networks (RNNs) and LSTMs:
  Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTMs) networks, are specifically designed to process sequential data by maintaining a hidden state that captures temporal dependencies across frames.9 This internal "memory" allows them to learn patterns over time, such as how body positions change in action recognition or object trajectories evolve.26 RNNs excel at modeling temporal dependencies, making them suitable for video sequences where the order of elements

is important.9 LSTMs, in particular, overcome the vanishing gradient problem of standard RNNs through gating mechanisms (forget, input, and output gates), enabling them to selectively retain or discard information over long sequences.26 This is crucial for tasks requiring long-term memory, like predicting future frames or video captioning, and their ability to handle gaps or irregular timing between events makes them suitable for real-world video data where actions may unfold at varying speeds.26 LSTMs are used in visual tracking algorithms to model object trajectories over time 2 and can be part of hybrid models for video anomaly detection.27

However, standard RNNs struggle with long-term dependencies due to vanishing gradients, limiting their effectiveness for very long video sequences.[26] While LSTMs mitigate this, the problem can still arise. The sequential nature of RNNs also makes them less amenable to parallel processing compared to other architectures like Transformers [9], which can impact training and inference speed for large datasets. Training dynamical RNNs, especially continuous-time variants, can be computationally expensive due to the inflated number of time steps.[28]

- 3D Convolutional Neural Networks (CNNs):
  Unlike 2D CNNs that process images frame by frame, 3D CNNs apply convolutions across both spatial dimensions (height, width) and the temporal dimension (time). This allows them to inherently capture spatio-temporal features, integrating motion and appearance information simultaneously directly from video clips.29 3D CNNs learn filters that are sensitive to patterns evolving over time, such as specific movements or changes in object appearance across consecutive frames. They can learn robust feature representations against perturbations, especially when combined with data augmentation techniques like random sampling rates (to mimic target action speed) and random ego-motion (to simulate camera movement).29 The primary limitations of 3D CNNs include their high computational cost and memory requirements due to processing an additional dimension. Newer architectures like Transformers often outperform 3D CNNs with less computational complexity.30

The development of deep learning for temporal understanding reveals a complementary relationship between sequential and spatio-temporal deep learning for video. RNNs/LSTMs are strong for explicit sequence modeling and long-term dependencies, while 3D CNNs are strong for local spatio-temporal patterns. This suggests that combining these approaches could yield more powerful models. This foreshadows the rise of hybrid architectures. Furthermore, the field is moving towards more data-efficient and self-supervised methods for learning temporal representations. Techniques like Temporal Preference Optimization (TPO), which uses self-training and preference learning, are emerging to reduce reliance on manually annotated data.[31] This is critical for drone applications where annotating vast amounts of video data, especially for subtle temporal events or tiny objects, is extremely costly and time-consuming. Self-training and preference learning could unlock scalability for deploying AI in diverse drone scenarios.

Table 1: Comparison of Key Temporal Feature Extraction Methods

Method	Principle	Strengths	Weaknesses	Suitability for	Key Snippet

				Drone Video	References
Optical Flow	Estimates pixel-level motion between frames based on brightness constancy and local motion assumptions.	Provides dense, pixel-level motion; useful for precise event localization and trajectory tracking.	Computationally expensive; sensitive to noise, lighting changes, and motion blur; ill-posed problem; complicated by camera ego-motion.	High granularity but often too slow/resource-intensive for real-time on-drone processing, unless advanced deep learning methods are used.	[8]
Motion Vectors	Block-based offsets describing transformation between macroblocks in compressed video (e.g., H.264, HEVC).	Highly efficient and scalable (part of compressed stream); low memory footprint; readily available; good proxy for regions of interest.	Less granular (block-based) than optical flow; derived from compression, not direct scene motion.	Highly suitable for real-time, resource-constrained drone applications due to inherent efficiency and availability from video codecs.	[22]
Recurrent Neural Networks (RNNs) / LSTMs	Process sequential data by maintaining a hidden state (memory) that captures temporal dependencies; LSTMs use gates to manage long-term memory.	Capture time-dependent variations; LSTMs handle long-term dependencies and irregular timing; model object trajectories over time.	Standard RNNs suffer vanishing gradients; limited parallelization compared to Transformers; computationally expensive for continuous-time systems.	Useful for modeling object trajectories and temporal sequences, but parallelization limitations can be a bottleneck for long, real-time drone videos.	[2]
Transformer Networks	Utilize self-attention mechanisms to	Excellent at long-range interactions;	Quadratic complexity with input	Highly promising for drone video	[5]

	process entire sequences at once, modeling all-to-all relationships between tokens (patches, frames, clips).	highly versatile; parallelizable; can outperform 3D CNNs; robust and generalize well; effective with self-supervised learning.	length (requires efficient adaptations); lack strong inductive biases (needs large data or architectural modifications).	due to long-range dependency modeling and parallelization, especially with architectural adaptations for efficiency and self-supervised pre-training.	
Graph Neural Networks (GNNs)	Represent objects as nodes and their interactions as edges, modeling spatial-temporal relationships and dependencies across frames.	Model complex object interactions and relationships; dynamic graph structures enable adaptive tracking; good for occlusions and crowded scenes.	Can be computationally intensive for large, dense graphs; complexity in defining optimal graph structures for diverse scenarios.	Well-suited for multi-object tracking in dynamic drone environments, particularly for handling occlusions and complex interactions between targets.	[36]

3. Advanced Architectures for Spatio-Temporal Feature Learning

The limitations of traditional methods and early deep learning models have driven the development of more sophisticated architectures capable of capturing complex spatio-temporal dynamics in video.

3.1. Transformer Networks

Transformer models, initially dominant in Natural Language Processing, have shown remarkable success in computer vision, particularly for video, due to their ability to handle long-range interactions.

- Mechanism for Processing Sequential Video Data:
  The core of Transformers is the self-attention mechanism, which allows each token embedding to be augmented with information from all other embeddings in a

sequence.30 This enables modeling of all-to-all relationships, crucial for understanding motion cues and dynamic appearance changes over time.30 Videos are first divided into smaller units called "tokens".30 This tokenization can involve dividing frames into 2D or 3D patches (cubes) to capture local motion features, focusing on semantically meaningful foreground regions (instance-wise), processing entire frames (frame-wise), or condensing information from several frames into clip-wise tokens for longer temporal spans.30 Since self-attention is permutation invariant, positional embeddings (PEs) are added to signal the position of tokens in the sequence, exploiting the spatio-temporal structure of videos.30 These PEs can be fixed or learned, and absolute or relative.

- Advantages:
  Transformers excel at modeling relationships between elements regardless of their distance in the sequence, which is vital for understanding motion cues and dynamic appearance changes over extended periods in video.30 This directly addresses the long-term dependency issues faced by RNNs.26 Their lack of strong inductive biases makes them highly adaptable to the complex spatio-temporal structure of video.30 Unlike RNNs, Transformers can process entire sequences at once, enabling greater parallelization during training and inference.26 Video Transformers can outperform 3D CNNs in action classification tasks, sometimes with less computational complexity.30 They have also shown robustness to various perturbations and may form more abstract semantic representations, leading to improved out-of-distribution (OOD) generalization.30 Furthermore, self-supervised learning strategies, particularly Masked Token Modeling (MTM), are highly effective for video Transformers, reducing reliance on large supervised datasets and yielding robust, general features.30
- Architectural Adaptations for Video:
  To mitigate the quadratic complexity of self-attention with input length, many approaches decompose full attention into smaller operations, such as restricted (local, axial, sparse) or aggregation (hierarchical, query-driven compression) methods.30 For long-term temporal modeling, approaches include memory-based methods (storing global frame features) or recurrence-based methods (propagating self-attention outputs forward in time as recurrent states).30 Multi-view approaches, defining multiple views of a video (e.g., varying resolution or sampling patterns), allow for cooperative task solving and interactions between views.30 Transformers are being explored for drone video anomaly detection, leveraging multi-scale feature maps and joint attention mechanisms to capture spatial and temporal information in dynamic backgrounds.35 They are also used in transformer-based tracking architectures for handling long-term motion dependencies in small object tracking.5

Transformers are consistently highlighted for their ability to capture long-range dependencies and outperform 3D CNNs, positioning them as a new foundation for video understanding. However, their quadratic scaling with input length is a major limitation, especially for high-dimensional video data.[30] This leads to ongoing research focusing on "efficient designs" and "architectural adaptations".[30] While Transformers are state-of-the-art, their direct application to raw, long video sequences is computationally prohibitive. The current trend is to develop clever architectural modifications (e.g., sparse attention, hierarchical designs) that retain the power of global context modeling while reducing computational complexity. This is crucial for real-time drone applications where resources are limited. The combination of Transformers' inherent ability to learn rich representations and self-supervised pre-training techniques, such as Masked Token Modeling [30], suggests a path toward more generalizable and data-efficient models for drone video. This is particularly valuable given the challenges of acquiring and annotating diverse drone datasets, allowing models to learn robust temporal features from unlabeled video.

3.2. Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) are powerful for modeling relationships and interdependencies between objects, making them well-suited for capturing spatio-temporal interactions in dynamic scenes.

- Modeling Spatial-Temporal Relationships:
  GNNs are designed for non-Euclidean data and excel at modeling complex relationships and interdependencies between objects.40 In video, they represent detected objects as nodes and their interactions (spatial proximity, temporal continuity) as edges, effectively capturing dependencies across consecutive frames.38 A key advantage is the ability to dynamically construct and update graph structures based on object motion and interactions.38 This enables adaptive and robust tracking in highly variable environments like urban traffic.38 GNNs are used for small object detection and tracking in traffic surveillance (DGNN-YOLO) 38, and for modeling spatio-temporal saliency cues in video.42
- Dynamic GNNs for Adaptive Tracking:
  The DGNN-YOLO framework integrates dynamic GNNs with YOLO for enhanced small-object detection and tracking.38 The DGNN module dynamically constructs graph structures where nodes are detected objects and edges capture their spatial-temporal relationships, allowing robust tracking across video frames.38 This helps address challenges like occlusion and complex object interactions.38 Message-passing algorithms, such as those in TrackMPNN, dynamically refine object associations.38 Recent advancements integrate attention mechanisms into GNNs to model precise relationships in crowded scenes, such as AST-GCN and DGNN, which improve association accuracy.38 DGNN-YOLO has shown superior performance in detecting and tracking small objects (pedestrians, cyclists, motorbikes) under diverse traffic conditions, demonstrating robustness and scalability.38
- Hybrid GNN-Transformer Approaches:
  The "Transformer-GraphFormer Blender Network" (TGBFormer) combines the strengths of Transformers (global representations, long-range dependencies) and Graph Neural Networks (local spatial and temporal relationships) for video object detection.33 It includes a spatial-temporal transformer module for global contextual information, a spatial-temporal GraphFormer module for local feature aggregation, and a global-local feature blender module to adaptively couple their outputs.33 Another hybrid model is

the Hybrid Spiking Vision Transformer (HsVT) for event-based object detection, which combines Spiking Neural Networks (SNNs) with Transformers, capturing spatio-temporal features efficiently and leveraging event cameras for low-latency and high-dynamic range benefits.45

GNNs, by explicitly modeling objects as nodes and their interactions as edges [38], offer a powerful solution for relational and dynamic object interactions. The emphasis on "dynamically capturing interactions" [38] and "adaptive graph construction" [38] is key. For complex drone scenarios with multiple interacting objects, occlusions, and unpredictable movements, GNNs allow AI to reason about the *relationships* between objects and their collective motion, leading to more robust tracking and better handling of identity switches, which are common failure modes in crowded scenes. The necessity of hybridization for comprehensive video understanding is evident, as no single architecture is a panacea. Both Transformers for global, long-range dependencies [30] and GNNs for local, relational interactions [38] have strengths. The TGBFormer explicitly states its goal is to "simultaneously leverage global and local information" by combining them.[33] This implies that the future of robust temporal feature extraction lies in intelligent hybridization that combines the best aspects of different deep learning paradigms, allowing models to handle both fine-grained local dynamics and broad contextual relationships, which is essential for the diverse and challenging conditions encountered in drone video.

3.3. Hybrid and Multi-Modal Fusion Strategies

Effective temporal feature extraction often relies on fusing information from various sources—different feature types, different levels of abstraction, or even different sensor modalities.

- Early, Intermediate, and Late Fusion Paradigms:
  These paradigms describe when different information streams are combined in a processing pipeline.46
  Early Fusion involves combining raw modalities *before* feature extraction, such as applying a depth mask to RGB images or projecting skeletal sequences onto the image.[46]
  Intermediate Fusion, also known as feature-level fusion, combines features extracted from each modality *before* classification.[46] This aims to produce a new, more expressive representation by merging the distinctive features of each data type.[46]
  Late Fusion, or decision-level fusion, combines modality-wise classification results, where each modality processes data independently, and their final decisions are merged.[46] In intermediate fusion, spatio-temporal patterns extracted from video sequences can be optimally combined.[46] In late fusion, temporal feature extraction occurs within each single-modality architecture, and then their temporally-informed decisions are merged.[46]
- Spatio-Temporal Feature Fusion:

Modern object tracking systems integrate both spatial (appearance-based) and temporal (motion-based) features for robust performance.2 The ViT Spatio-Temporal Feature Fusion (STFF) strategy enhances UAV tracking by applying object information from previous frames to improve real-time responsiveness and tracker performance.18 It leverages dynamic change information in space and time to enhance correlation and promote feature aggregation.18 YOLOMG, a motion-guided object detector for small drones, combines a motion difference map (pixel-level motion feature) with RGB images.3 This fusion helps discriminate small drones from complex backgrounds where appearance features are unreliable.3 Multi-scale feature fusion, as seen in algorithms for UAV target tracking, uses deep inter-correlation operations and global attention mechanisms to refine feature representation and reduce computational effort, addressing challenges like large target-scale variations.16

- Multi-Modal Sensor Fusion:
  Traditional RGB cameras are vulnerable to environmental challenges like overexposure, low light, and motion blur.48 Therefore, fusion of multiple sensor modalities (e.g., camera, radar, LiDAR, event cameras) is crucial to overcome the limitations of individual sensors and achieve robust detection and tracking in complex, adversarial, and high-speed scenarios.2 Bio-inspired event cameras offer advantages like high temporal resolution, high dynamic range, and inherent privacy protection by asynchronously capturing pixel-level intensity changes.48 They excel under extreme conditions (overexposure, low light, fast motion) where RGB cameras struggle.48 SFDNet, a fully spiking neural network for object detection, integrates RGB frames and event streams for low-power and high-performance object detection.49 It uses a novel LIMF neuron model for enhanced spatio-temporal representation and a lightweight spiking aggregation module for efficient feature integration.49 Hybrid models like HsVT also combine SNNs with Transformers for event-based object detection.45 Kalman filters are used in motion estimation for object tracking, and hybrid tracking models combine them with CNN-based feature extraction for improved accuracy in real-time video streams.2 Kalman filters predict motion and correct errors, leveraging temporal information for robust tracking.2 SMART (Sensor Measurement Augmentation and Reacquisition Tracker) leverages high-frequency state estimates from Kalman filters to guide the search for new measurements, maintaining tracking continuity even with intermittent measurements.47 STTrack is a unified multimodal spatial-temporal tracking approach that explicitly leverages temporal context within multimodal video data to guide target localization and capture trajectories.51

Fusion is the cornerstone of robustness in drone video, as single-modality or single-feature approaches often struggle with the inherent challenges.[3] The solution consistently involves integrating both spatial (appearance-based) and temporal (motion-based) features [2] or fusing multiple sensor modalities.[47] This implies that future research and development should focus on sophisticated fusion architectures (e.g., adaptive, attention-based fusion) that can dynamically weigh and integrate information from various sources to maximize resilience to real-world variability. The emergence of event cameras as a game changer for low-latency and robust temporal sensing is also noteworthy. Event cameras are highlighted for their high temporal resolution, high dynamic range, and inherent privacy protection [48], and their ability to overcome motion blur and low light conditions where RGB cameras fail.[48] Their unique sensing principle offers a fundamental advantage for capturing rapid motion and operating in challenging illumination, potentially enabling new levels of low-latency, robust, and energy-efficient object detection and tracking for autonomous drones. This suggests a future where multi-modal systems, particularly those incorporating event-based vision, become standard for high-performance drone applications.

4. Ensuring Trustworthiness in Temporal Feature Extraction

Beyond raw performance, the trustworthiness of AI systems for drone applications is paramount, especially in safety-critical scenarios. Trustworthy AI encompasses principles of robustness, uncertainty quantification, and explainability.

4.1. Robustness to Adverse Conditions

Robustness ensures that the AI model performs consistently and reliably even when faced with unexpected or challenging real-world conditions.

- Addressing Motion Blur:
  Motion blur is a common artifact caused by relative movement between the camera and scene during exposure, reducing clarity and detail.17 In drone video, this is exacerbated by ego-motion.3 Practical solutions include adhering to the "180-degree rule" (shutter speed double frame rate) and using Neutral Density (ND) filters to manage exposure and add "natural" motion blur, thereby improving video quality at the source.12 Reducing flight speed or increasing shutter speed, or increasing Ground Sampling Distance (GSD) by increasing altitude, can also mitigate blur.13 Some missions can even stop the drone before taking a picture to eliminate motion blur entirely.13 Deep learning-based blind motion deblurring methods aim to restore clear images without prior knowledge of the blur kernel.17 These end-to-end CNN-based methods (multi-scale, multi-patch, multi-temporal structures) learn complex non-linear mappings from blurred to clear images, avoiding issues with inaccurate kernel estimation.17 Event cameras are inherently less susceptible to motion blur due to their asynchronous, pixel-level intensity change detection.48
- Handling Varying Perspectives and Rapid Movement:
  Drone video often involves significant scale variations and perspective distortions due to changing distances and camera angles.2 Rapid object movement or camera ego-motion can also lead to inconsistent detections and tracking "flicker".1 Adaptive algorithms are needed to update target scales.20 Data augmentation, simulating noise, rotation, or occlusion during training, helps models generalize to unseen conditions.11

Random sampling rates can mimic various target action speeds, and random ego-motion can simulate camera movement, enhancing temporal robustness.29 Transformer-based architectures, with their ability to handle long-range interactions and model dynamics, are suitable for rapid movement.5 Techniques like "UAV Motion Compensation" adjust bounding boxes to mitigate UAV ego-motion, preserving aspect ratios.15

- Mitigating Occlusion and Background Clutter:
  Occlusion and dynamic backgrounds are major challenges, leading to missed detections, identity switches, and objects being lost against moving backgrounds.2 Background subtraction methods like Mixture of Gaussians 2 (MOG2) and Visual Background Extractor (ViBe) isolate motion regions from complex, dynamic backgrounds, reducing search space and computational overhead.4 Integrating background subtraction with deep learning classification enhances robustness.4 Dynamic Vision Sensors (DVS) also offer inherent background suppression.50 Motion-guided detection, such as YOLOMG combining a motion difference map (pixel-level motion) with RGB images, can enhance detection of extremely small objects in complex environments.3 Re-identification (Re-ID) models are crucial for maintaining consistent tracking IDs despite occlusions and temporary disappearances by capturing distinctive visual characteristics.5 Deep SORT incorporates deep appearance features for re-identification.2 Using a low-pass filter to remove detections that do not reoccur often enough can help with flickery detections.1 Graph Neural Networks (GNNs), particularly dynamic ones, effectively capture dependencies across consecutive frames and handle occlusions and complex object interactions by dynamically updating graph structures.38
- Strategies for Noise Resilience:
  Preprocessing techniques like normalization, histogram equalization, temporal alignment, and frame interpolation adjust for lighting variations, inconsistent frame rates, or motion blur.11 Temporal smoothing, using moving averages or median filters, can reduce frame-to-frame feature jitter.11 Attention mechanisms can prioritize specific regions of interest, such as moving objects within a cluttered scene, making models more robust to noise.11 Data augmentation, including simulating noise during training, is also crucial.11 Furthermore, larger models with more parameters tend to exhibit greater robustness to adverse conditions.56

The pursuit of trustworthy AI for drone video moves beyond simply correcting errors after they occur. It demands a holistic design philosophy where robustness is built into every stage, from data acquisition and preprocessing to model architecture and training. This reduces reliance on brittle post-hoc fixes and leads to more inherently stable and reliable systems. For the prevalent challenge of tiny objects in drone video, motion is not just an additional feature; it becomes the *primary* distinguishing characteristic. This means that algorithms must be exquisitely sensitive to subtle pixel-level movements and temporal patterns, even when visual cues are minimal. This also highlights the importance of high frame rates and motion-centric data acquisition for drone video.

Table 2: Challenges in Drone Video and Corresponding AI Solutions for Robustness

Challenge	Impact on Detection/Tracking	AI Solution/Technique	Key Snippet References
Small and Tiny Objects	Insufficient appearance/texture info; low pixel count; easily confused with background.	Motion-guided detection (e.g., YOLOMG); leveraging temporal information/motion features; multi-scale feature fusion; DotD metric for evaluation.	[2]
Motion Blur	Reduced clarity/detail; blurred/stretched contours; caused by ego-motion or rapid target movement.	Deep learning deblurring (blind motion deblurring, CNNs); camera settings (180-degree rule, ND filters); reducing flight speed; event cameras.	[3]
Dynamic/Cluttered Backgrounds	Targets obscured; background confusion; moving elements mistaken for objects.	Background subtraction (MOG2, ViBe, Fuzzy Logic); motion-based pipelines; event cameras (inherent background suppression); motion-guided detection.	[3]
Occlusion and Disappearance	Tracking failures; identity switches; objects temporarily lost.	Re-identification (Re-ID) models; deep association metrics; long-term tracking strategies; Graph Neural Networks (dynamic graph updates).	[2]
Varying Perspective/Scale	Significant changes in object size, aspect ratio, texture details; missed detections.	Adaptive algorithms (multi-scale estimation, multi-scale feature fusion); data augmentation (simulating rotation);	[2]

		UAV motion compensation.	
Noise and Distortion	Degraded image quality; inconsistent feature extraction; sensor noise; compression artifacts.	Preprocessing (normalization, histogram equalization, temporal alignment); temporal smoothing; attention mechanisms; larger model size; adversarial training.	[11]

4.2. Uncertainty Quantification (UQ)

Uncertainty quantification is a critical aspect of trustworthy AI, especially in high-risk domains like autonomous systems, as it allows models to express the confidence in their predictions.[6] UQ is vital for reliable decision-making as it allows identifying when predictions are likely wrong, enabling adjustment of models or alerting human operators.[19] This is particularly important for autonomous driving, medical diagnosis, and disaster response.[61] Quantifying uncertainty improves the transparency of deep learning models and helps understand their decision-making process.[60] Trustworthy AI models should be able to quantify their uncertainty and recognize when they encounter novel or unreliable inputs, as making confident predictions on out-of-distribution (OOD) data can lead to critical failures.[6]

UQ techniques differentiate between **aleatoric uncertainty**, which is inherent randomness in the data itself (e.g., sensor noise, motion blur), and **epistemic uncertainty**, which is due to the model's lack of knowledge (e.g., insufficient training data, encountering OOD data).[7]

Methods for quantifying uncertainty in deep learning models include Bayesian methods, which are a predominant technique for UQ [7], and Bayesian Deep Learning approaches that can quantify uncertainty with theoretical guarantees, for example, using spatial-temporal neural processes.[63] Other methods include Monte Carlo Dropout, a widely used technique, though it may require multiple inference runs, making it impractical for real-time tasks [65], and Deep Ensembles, another common UQ method, also potentially computationally intensive.[65] Output-based methods (e.g., softmax confidence, energy-based models) and distance-based methods (e.g., Mahalanobis distance, k-NN) are also employed.[6] Contrastive learning can also improve generalization and robustness.[6]

In object tracking, UncTrack is a novel uncertainty-aware transformer tracker that predicts target localization uncertainty and incorporates this information for accurate target state inference.[19] It models localization uncertainty across continuous video frames as a prototype representation.[19] Spatio-Temporal Uncertainty Guided NMS (STU-NMS) incorporates spatial and temporal uncertainty to guide and improve the Non-Maximum Suppression procedure in video event detection, helping suppress detection probabilities when event instances exhibit significant spatial or temporal uncertainties.[66] Learned uncertainty can be meaningful and human-interpretable, especially when showing high uncertainty for occluded or small objects.[62] High uncertainty often correlates with inaccurate detections of occluded or small objects.[62] UQ can also guide "box relaxation" to increase intersection probability and associations.[62]

UQ is a prerequisite for autonomous drone operation, as it directly links to "high-risk domains" and "autonomous systems".[7] The ability to "identify when our predictions are likely wrong" [60] is not just a desirable feature but a safety-critical one for drones operating autonomously (e.g., collision avoidance, search and rescue). This implies that for drone AI, UQ is not an optional add-on but a fundamental requirement for deployment in real-world, high-stakes scenarios. It enables safe decision-making, allows for human intervention when confidence is low, and provides a crucial layer of accountability. Future drone AI systems will increasingly integrate UQ directly into their perception and decision-making pipelines. Furthermore, temporal features play a dual role in UQ for video. Temporal features are not just inputs for detection/tracking; they are also crucial for *quantifying uncertainty*. UncTrack uses "localization uncertainty across continuous video frames" [19], and STU-NMS integrates "spatial and temporal uncertainty".[66] This means temporal consistency and motion patterns can indicate the reliability of a prediction. For example, a sudden, inexplicable jump in an object's predicted position might correlate with high uncertainty. This suggests that UQ methods specifically designed for sequential data, leveraging temporal dynamics, will be more effective for drone video than static image UQ approaches.

4.3. Explainability (XAI) of Temporal Features

Explainability in AI is about providing human-intelligible insights into how models make decisions, fostering trust and enabling debugging. Transparency is crucial for trustworthy AI, as it allows insight into *how* a model reaches its decisions, ensuring they are made for the right reasons.[6] This helps justify the rationality behind a model's predictions.[67] Explaining temporal behavior is also crucial for verifying intelligent systems through formal methods.[68] Methods for explaining temporal models include the Temporal Motifs Explainer (TempME), a novel approach that uncovers the most pivotal temporal motifs (recurring substructures in dynamic systems) that guide the predictions of temporal Graph Neural Networks (GNNs).[67] TempME is theoretically grounded in the Information Bottleneck (IB) principle, which aims to extract the most interaction-related motifs while minimizing information redundancy, ensuring sparse and succinct explanations.[67] It samples candidate temporal motif instances (sequences of reversely time-ordered events) and learns motif-level representations.[67] TempME leverages a "null model" (randomized network) to define an empirical prior distribution, allowing it to distinguish motifs that are structurally significant from those that are random, thereby highlighting their importance to model predictions.[67] Automated learning of temporal properties from system executions, using formalisms like temporal logic and finite automata, can provide human-interpretable descriptions for Explainable AI.[68] This includes learning in the presence of noise and from positive data.[68] DGNN-YOLO also incorporates XAI techniques (Grad-CAM, Grad-CAM++, Eigen-CAM) to enhance the interpretability of its decisions for detecting and tracking small occluded objects.[44]

Explaining static image classification is already challenging; explaining dynamic systems like drone video tracking, where decisions depend on evolving temporal patterns, is even more complex.[67] The concept of "temporal motifs" [67] suggests that human-interpretable explanations for temporal AI will likely involve identifying recurring patterns of events or interactions, rather than just static features. This implies that developing effective XAI for temporal features is a significant frontier. It requires moving beyond saliency maps for individual frames to methods that can highlight *sequences of events* or *dynamic interactions* that lead to a particular decision. This is critical for debugging, auditing, and building user trust in autonomous drone systems, especially when failures occur. The pursuit of trustworthy AI for drones will also increasingly involve a convergence of machine learning and formal methods. The mention of "formal methods" and "temporal logic" [68] in the context of explaining and verifying system behavior suggests that machine learning can *learn* these formal properties, bridging the gap between black-box AI and verifiable systems. Learning temporal properties allows AI models to implicitly encode rules that can then be explicitly verified, offering a stronger guarantee of trustworthiness than statistical performance metrics alone. This is particularly relevant for safety-critical drone applications where formal verification is often a regulatory requirement.

Table 3: Trustworthy AI Principles and their Application to Temporal Feature Extraction

Principle	Definition	Relevance to Temporal Features in Drone Video	Key Techniques/Approaches	Key Snippet References
Scientific Validity	Model and predictions based on sound scientific principles; trained on correct data; fits underlying data well; recognizes limitations.	Ensures temporal features accurately capture motion and change; models are robust to data variations (e.g., motion blur, ego-motion).	Proper problem definition, training/evaluation; avoiding over/underfitting; preventing data leakage; robust feature extraction methods.	[6]
Fairness	AI systems should not discriminate or produce inequitable outcomes.	Ensuring temporal feature extraction does not introduce bias based on object	Careful dataset curation (diverse scenarios); fairness metrics; bias detection and	[6]

		type, speed, or environmental conditions (e.g., visibility of certain objects).	mitigation techniques.	
Transparency/Explainability	Insight into *how* a model makes decisions, ensuring right decisions for right reasons.	Understanding *which* motion patterns or temporal sequences led to a detection/tracking decision.	Temporal Motifs Explainer (TempME) for GNNs; learning temporal properties (temporal logic, finite automata); XAI techniques (Grad-CAM, Eigen-CAM).	[6]
Safety/Uncertainty Awareness	AI models quantify uncertainty; recognize novel/unreliable inputs; "know what they don't know."	Estimating confidence in object localization and trajectory prediction; identifying when tracking is unreliable due to occlusion, noise, or OOD data.	Uncertainty Quantification (UQ) methods (Bayesian, Monte Carlo Dropout); uncertainty-aware trackers (UncTrack); spatio-temporal uncertainty guided NMS.	[6]
Accountability	Mechanisms for addressing model mistakes or harm; multi-faceted.	Documenting model performance, limitations, and failure modes in dynamic, safety-critical drone operations.	Model documentation; performance tracking across diverse conditions; auditing mechanisms.	[6]

5. Performance Considerations and Evaluation Metrics

Evaluating the performance of temporal feature extraction for object detection and tracking in drone video requires a comprehensive set of metrics that go beyond simple accuracy, considering real-world operational constraints.

- Accuracy Metrics:
  Mean Average Precision (mAP) is a widely used metric in object detection, providing an evaluation of regression and classification accuracies.69 It represents the mean of Average Precision (AP) across categories.69 For video, mAP can be categorized by object speed (slow, medium, fast) using average IoU over time.69 Precision and Recall are fundamental metrics for classification accuracy.6 Precision (fraction of predicted positives that are true positives) and Recall (fraction of true positives classified as positive) are crucial.6 Intersection over Union (IoU) determines if a prediction is true based on overlap with ground truth.69 For small objects, traditional IoU can be overly sensitive to localization errors.5 Higher Order Tracking Accuracy (HOTA) is a more comprehensive metric for multi-object tracking that explicitly considers detection, localization, and association, making it particularly relevant for small object tracking challenges.5 Other detailed metrics for multi-object tracking include Identification Metrics (IDF1), Mostly Tracked (MT), Mostly Lost (ML), False Negative (FN), False Positive (FP), Identity Switches (IDs), Fragmentation (FM), Multi-object tracking accuracy (MOTA), and Multi-object tracking precision (MOTP).70
- Latency and Computational Efficiency for Real-time Drone Applications:
  Frames Per Second (FPS) measures the number of frames processed per second, directly indicating real-time capability.16 High FPS (e.g., >100 FPS for GAFFPFM, 84.7 FPS for MT-Track) is essential for applications like autonomous driving and surveillance.2 Many advanced models (e.g., Transformers, 3D CNNs) are computationally expensive.30 Lightweight network designs and efficient architectures (e.g., anchor-frame-free mechanisms, spiking neural networks) are crucial for low-latency performance.16 Spiking Neural Networks (SNNs) offer advantages in low energy consumption, making them promising for resource-constrained drone platforms.45
- Trade-offs between Accuracy, Speed, and Robustness:
  There is often an inherent trade-off between accuracy and computational speed; for example, high accuracy can come at the cost of higher latency.16 Models like GAFFPFM aim to balance speed and accuracy in UAV tracking.16 Adversarial robustness can sometimes trade off with clean accuracy.57 More complex models (larger number of parameters) tend to offer greater robustness but at higher computational cost.56 The choice of algorithms (e.g., Deep-SD Assignment vs. StrongSORT) involves balancing track continuity, false alarms, and processing time.70
- Adversarial Robustness Evaluation:
  Enhancing adversarial robustness is necessary for safety-critical tasks.57 Adversarial attacks involve subtle manipulations to input data that mislead detection models.71 Methods include Adversarial Training (AT), which involves training models on adversarial examples (intentionally perturbed inputs) to improve resistance.57 VFAT-WS (Video Fast Adversarial Training with Weak-to-Strong consistency) is a fast AT method for video, integrating temporal frequency augmentation.57 SPARK (Spatial-aware Online Incremental Attack) is an online adversarial attack that generates imperceptible

perturbations to mislead trackers along an incorrect or specified trajectory.72 It performs spatial-temporal sparse incremental perturbations online.72 Multi-agent Reinforcement Learning is used to identify sensitive spatial and temporal regions in videos for generating adversarial samples with minimal perturbations.73 Robustness toolkits like VOT-RT evaluate robustness to realistic image distortion scenarios (noise, compression artifacts).59

The performance metrics are not monolithic. There is a constant tension between accuracy, speed, and robustness.[16] For drone applications, "real-time" [16] and "low-latency" [48] are critical, often implying a willingness to accept slightly lower accuracy for higher speed. Furthermore, "robustness" [57] against adverse conditions and attacks becomes equally important as raw accuracy. This implies that developing AI for drones is fundamentally a multi-objective optimization problem. Engineers must carefully balance these competing demands based on the specific application's requirements. This often means choosing architectures that offer a good trade-off (e.g., lightweight models, efficient fusion strategies) rather than simply pursuing peak accuracy on benchmark datasets. Beyond traditional accuracy, adversarial robustness [57] is a critical area. The existence of methods like VFAT-WS [57] and SPARK [72] demonstrates that models are vulnerable to subtle attacks, and defending against them is a research priority. This directly ties into the "trustworthy" aspect of the query. As AI systems for drones become more autonomous and deployed in sensitive environments, their susceptibility to adversarial attacks poses a significant security and safety risk. Evaluating and enhancing adversarial robustness will become a standard part of the development and deployment lifecycle, moving from a niche research area to a core performance requirement for trustworthy drone AI.

6. Conclusion and Future Outlook

The trustworthy extraction of temporal features is foundational for robust object detection and tracking in challenging drone video streams. Significant advancements in AI, particularly in deep learning architectures, have enabled increasingly sophisticated temporal modeling.
Key Advancements:
The field has seen a fundamental shift from traditional motion methods, such as optical flow and motion vectors, towards advanced deep learning architectures that learn complex spatio-temporal dynamics directly from data. Architectures like Transformers excel at capturing long-range dependencies and offer parallelization benefits, while Graph Neural Networks effectively model object interactions and dynamic relationships. Hybrid architectures, which combine the strengths of these diverse paradigms, are emerging to provide comprehensive video understanding. Furthermore, sophisticated fusion strategies, including early, intermediate, and late fusion, as well as multi-modal approaches (e.g., RGB-Event fusion with Spiking Neural Networks), are effectively combining appearance, motion, and other sensor data to enhance robustness against drone-specific challenges. A growing emphasis on trustworthy AI principles, encompassing robust design, uncertainty

quantification (UQ) for reliable decision-making, and explainability (XAI) for transparency and debugging, is moving drone AI towards more reliable and interpretable systems.
Remaining Challenges and Open Research Questions:
Despite significant progress, several challenges persist. The scalability of models for very long-form video remains an issue, as the quadratic complexity of Transformers, for instance, still poses a computational hurdle. Achieving high accuracy and robustness simultaneously with low latency and computational efficiency for edge deployment on resource-constrained drones continues to be a trade-off that requires careful optimization. Existing methods often lack generalization to entirely new scenes or object types not seen during training, which is a critical limitation for real-world applications.3 Extreme conditions, such as severe motion blur, heavy occlusion, and highly dynamic backgrounds, still pose significant hurdles. Fully integrating UQ and XAI into real-time, end-to-end drone systems, and making them computationally feasible and truly human-interpretable, is an ongoing challenge. Finally, the continuous arms race between adversarial attacks and defenses requires constant innovation to ensure secure operation in safety-critical tasks.57
Potential Future Directions:
Future research will likely focus on several key areas. Further exploration and integration of event-based vision, leveraging event cameras for their inherent advantages in low-latency, high-dynamic range, and motion blur resilience, will be critical for next-generation drone AI. The development of adaptive and meta-learning architectures that can adjust on-the-fly to changing environmental conditions (e.g., lighting, weather) and learn from minimal new data will enhance generalization capabilities. A significant direction involves creating unified spatio-temporal-relational modeling frameworks that seamlessly integrate spatial appearance, temporal motion, and relational object interactions within a single, coherent architecture. For uncertainty quantification, deeper integration of probabilistic AI, such as Bayesian methods and active learning, will be crucial to provide more accurate and calibrated uncertainty estimates, especially for out-of-distribution detection. In explainability, moving beyond correlation to identify causal relationships in temporal data will provide more robust and actionable explanations for drone behavior. Finally, leveraging advanced generative models for synthetic data generation and simulation can augment real datasets, addressing data scarcity and improving robustness to varied conditions.

Works cited

1. Best approach for temporal consistent detection and tracking of small and dynamic objects : r/computervision - Reddit, accessed on June 13, 2025, https://www.reddit.com/r/computervision/comments/1jyehoz/best_approach_for_temporal_consistent_detection/
2. Deep Learning Object Tracking: Algorithms, Challenges, and Applications - FlyPix AI, accessed on June 13, 2025, https://flypix.ai/blog/deep-learning-object-tracking/
3. YOLOMG: Vision-based Drone-to-Drone Detection with Appearance and Pixel-Level Motion Fusion - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2503.07115v1

4. Evaluating a Motion-Based Region Proposal Approach with Background Subtraction Methods for Small Drone Detection - SCIEPublish, accessed on June 13, 2025, https://www.sciepublish.com/article/pii/491
5. Small Multi-Object Tracking for Spotting Birds (SMOT4SB) Challenge 2025 - MVA2025, accessed on June 13, 2025, https://mva-org.jp/mva2025/index.php?id=challenge
6. Trustworthy AI: Validity, Fairness, Explainability, and Uncertainty ..., accessed on June 13, 2025, https://carpentries-incubator.github.io/fair-explainable-ml/instructor/aio.html
7. From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence | Request PDF - ResearchGate, accessed on June 13, 2025, https://www.researchgate.net/publication/387797746_From_Aleatoric_to_Epistemic_Exploring_Uncertainty_Quantification_Techniques_in_Artificial_Intelligence
8. MOOSE: Pay Attention to Temporal Dynamics for Video Understanding via Optical Flows, accessed on June 13, 2025, https://arxiv.org/html/2506.01119v1
9. Recurrent neural network - Wikipedia, accessed on June 13, 2025, https://en.wikipedia.org/wiki/Recurrent_neural_network
10. Object Trajectory Estimation Using Optical Flow - DigitalCommons@USU, accessed on June 13, 2025, https://digitalcommons.usu.edu/context/etd/article/1469/viewcontent/ShouLiu_thesis.pdf
11. How do you ensure robustness in video feature extraction under ..., accessed on June 13, 2025, https://milvus.io/ai-quick-reference/how-do-you-ensure-robustness-in-video-feature-extraction-under-variable-conditions
12. How to fix motion blur problem when rotating the drone (yaw)? See example link - Reddit, accessed on June 13, 2025, https://www.reddit.com/r/dji/comments/1fkpfph/how_to_fix_motion_blur_problem_when_rotating_the/
13. Preventing Motion Blur in Drone Photogrammetry Flights - Hammer Missions, accessed on June 13, 2025, https://www.hammermissions.com/post/preventing-motion-blur-in-drone-photogrammetry-flights
14. Object Detection, Recognition, Tracking: Use Cases & Approaches - MobiDev, accessed on June 13, 2025, https://mobidev.biz/blog/object-detection-recognition-tracking-guide-use-cases-approaches
15. SFTrack: A Robust Scale and Motion Adaptive Algorithm for Tracking Small and Fast Moving Objects - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2410.20079v1
16. UAV target tracking method based on global feature interaction and ..., accessed on June 13, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0314485
17. Deep Learning in Motion Deblurring: Current Status, Benchmarks and Future Prospects, accessed on June 13, 2025, https://arxiv.org/html/2401.05055v2

18. ViT Spatio-Temporal Feature Fusion for Aerial Object Tracking, accessed on June 13, 2025, https://www.researchgate.net/publication/374931307_ViT_Spatio-Temporal_Feature_Fusion_for_Aerial_Object_Tracking
19. UncTrack: Reliable Visual Object Tracking with Uncertainty-Aware Prototype Memory Network - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2503.12888v1
20. A Multi-Scale Feature-Fusion Multi-Object Tracking Algorithm for Scale-Variant Vehicle Tracking in UAV Videos - MDPI, accessed on June 13, 2025, https://www.mdpi.com/2072-4292/17/6/1014
21. Optical Flow Estimation | Papers With Code, accessed on June 13, 2025, https://paperswithcode.com/task/optical-flow-estimation
22. Better foundation models for video representation - Amazon Science, accessed on June 13, 2025, https://www.amazon.science/blog/better-foundation-models-for-video-representation
23. Motion estimation - Wikipedia, accessed on June 13, 2025, https://en.wikipedia.org/wiki/Motion_estimation
24. HEVC/H.265 Codec (High Efficiency Video Coding) - VdoCipher, accessed on June 13, 2025, https://www.vdocipher.com/blog/hevc/
25. Temporal feature vector for video analysis and retrieval in high ..., accessed on June 13, 2025, https://digital-library.theiet.org/doi/full/10.1049/el.2017.3155
26. What role do recurrent neural networks (RNNs) and LSTMs play in ..., accessed on June 13, 2025, https://milvus.io/ai-quick-reference/what-role-do-recurrent-neural-networks-rnns-and-lstms-play-in-modeling-video-sequences
27. Awesome Video Anomaly Detection - GitHub, accessed on June 13, 2025, https://github.com/vt-le/Video-Anomaly-Detection
28. Recurrent neural network dynamical systems for biological vision ..., accessed on June 13, 2025, https://openreview.net/forum?id=ZZ94aLbMOK
29. Spatio-Temporal Filter Analysis Improves 3D-CNN for Action Classification - YouTube, accessed on June 13, 2025, https://www.youtube.com/watch?v=DdAgTEQl_I0
30. Video Transformers: A Survey - arXiv, accessed on June 13, 2025, http://arxiv.org/pdf/2201.05991
31. Temporal Preference Optimization for Long-Form Video Understanding - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2501.13919v1
32. NeurIPS Poster Recurrent neural network dynamical systems for biological vision, accessed on June 13, 2025, https://neurips.cc/virtual/2024/poster/94629
33. TGBFormer: Transformer-GraphFormer Blender Network for Video ..., accessed on June 13, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/32703
34. FETrack: Feature-Enhanced Transformer Network for Visual Object Tracking - MDPI, accessed on June 13, 2025, https://www.mdpi.com/2076-3417/14/22/10589
35. HSTforU: anomaly detection in aerial and ground-based videos with ..., accessed on June 13, 2025,

https://paperswithcode.com/paper/hstforu-anomaly-detection-in-aerial-and

36. Graph Transformer Networks - NIPS, accessed on June 13, 2025, https://proceedings.neurips.cc/paper/9367-graph-transformer-networks.pdf
37. Batch3DMOT: 3D Multi-Object Tracking Using Graph Neural ..., accessed on June 13, 2025, https://www.youtube.com/watch?v=hPhC2MJubqE
38. arxiv.org, accessed on June 13, 2025, https://arxiv.org/html/2411.17251v2
39. Graph Neural Networks in Point Clouds: A Survey - MDPI, accessed on June 13, 2025, https://www.mdpi.com/2072-4292/16/14/2518
40. [1901.00596] A Comprehensive Survey on Graph Neural Networks - arXiv, accessed on June 13, 2025, https://arxiv.org/abs/1901.00596
41. The basics of spatio-temporal graph neural networks - YouTube, accessed on June 13, 2025, https://m.youtube.com/watch?v=RRMU8kJH60Q&pp=ygUJI3NwYXRpb2Fs
42. A Motion-aware Spatio-temporal Graph for Video Salient Object ..., accessed on June 13, 2025, https://openreview.net/forum?id=VUBtAcQN44&referrer=%5Bthe%20profile%20of%20Yongjian%20Deng%5D(%2Fprofile%3Fid%3D~Yongjian_Deng1)
43. MCBLT: Multi-Camera Multi-Object 3D Tracking in Long Videos - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2412.00692v3
44. (PDF) Interpretable Dynamic Graph Neural Networks for Detecting ..., accessed on June 13, 2025, https://www.researchgate.net/publication/386143622_Interpretable_Dynamic_Graph_Neural_Networks_for_Detecting_and_Tracking_Small_Occluded_Objects_in_Urban_Traffic
45. Hybrid Spiking Vision Transformer for Object Detection with Event Cameras (ICML 2025), accessed on June 13, 2025, https://arxiv.org/html/2505.07715v1
46. (PDF) Early, intermediate and late fusion strategies for robust deep ..., accessed on June 13, 2025, https://www.researchgate.net/publication/354984828_Early_intermediate_and_late_fusion_strategies_for_robust_deep_learning-based_multimodal_action_recognition
47. Multi-Modal Sensor Fusion and Object Tracking for Autonomous Racing - ResearchGate, accessed on June 13, 2025, https://www.researchgate.net/publication/370450915_Multi-Modal_Sensor_Fusion_and_Object_Tracking_for_Autonomous_Racing
48. Towards Low-Latency Event Stream-based Visual Object Tracking: A Slow-Fast Approach - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2505.12903v1
49. Efficient Spiking Neural Network for RGB–Event Fusion-Based ..., accessed on June 13, 2025, https://www.mdpi.com/2079-9292/14/6/1105
50. Towards Real-Time Fast Unmanned Aerial Vehicle Detection Using Dynamic Vision Sensors - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2403.11875v1
51. Exploiting Multimodal Spatial-temporal Patterns for Video Object Tracking, accessed on June 13, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/32372/34527
52. 1 Introduction - arXiv, accessed on June 13, 2025,

https://arxiv.org/html/2401.05055v1

53. Top Video Object Detection Algorithm in Computer Vision | Encord, accessed on June 13, 2025, https://encord.com/blog/video-object-tracking-algorithms/
54. 80 Norwegian Journal of development of the International Science No 153/2025 ADAPTIVE MULTI-KERNEL BACKGROUND SUBTRACTION USING, accessed on June 13, 2025, https://nor-ijournal.com/wp-content/uploads/2025/03/NJD_153-80-83.pdf
55. Evaluating a Motion-Based Region Proposal ... - SCIEPublish, accessed on June 13, 2025, https://www.sciepublish.com/index/article/download_article/id/491.html
56. Robustness Benchmark Evaluation and Optimization for Real-Time ..., accessed on June 13, 2025, https://www.mdpi.com/2076-3417/15/9/4950
57. Fast Adversarial Training with Weak-to-Strong Spatial-Temporal Consistency in the Frequency Domain on Videos - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2504.14921v1
58. Robust Deep Object Tracking against Adversarial Attacks | Request ..., accessed on June 13, 2025, https://www.researchgate.net/publication/384363936_Robust_Deep_Object_Tracking_against_Adversarial_Attacks
59. Explaining and verifying the robustness of Visual Object Trackers to noise, accessed on June 13, 2025, https://aiia.csd.auth.gr/wp-content/uploads/2022/06/IVMSP_2022__Explaining_and_verifying_the_robustness.pdf
60. Robust adversarial uncertainty quantification for deep learning fine-tuning - PMC, accessed on June 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9957691/
61. A Survey on Uncertainty Quantification Methods for Deep Learning - arXiv, accessed on June 13, 2025, https://arxiv.org/html/2302.13425v5
62. Exploiting Detection and Localization Uncertainty in Multi-Object Tracking - YouTube, accessed on June 13, 2025, https://www.youtube.com/watch?v=7qzHMZF1C5s
63. KDD 2023 - Deep Bayesian Active Learning for Accelerating Stochastic Simulation, accessed on June 13, 2025, https://www.youtube.com/watch?v=Z65xOXHmuao
64. Bayesian Deep Learning and Probabilistic Model Construction - ICML 2020 Tutorial, accessed on June 13, 2025, https://www.youtube.com/watch?v=E1qhGw8QxqY
65. Uncertainty Quantification for Collaborative Object Detection Under ..., accessed on June 13, 2025, https://arxiv.org/pdf/2502.02537?
66. Spatiotemporal uncertainty guided non maximum suppression for ..., accessed on June 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11978777/
67. TempME: Towards the Explainability of Temporal ... - OpenReview, accessed on June 13, 2025, https://openreview.net/pdf?id=6OOgw4boZI
68. Learning Temporal Properties for Explainability and Verification, accessed on June 13, 2025, https://kluedo.ub.rptu.de/frontdoor/index/index/docId/8425
69. (PDF) A Review of Video Object Detection: Datasets, Metrics and ..., accessed on June 13, 2025,

https://www.researchgate.net/publication/345311452_A_Review_of_Video_Object_Detection_Datasets_Metrics_and_Methods

70. A comparative study of joint video tracking and classification ... - V-City, accessed on June 13, 2025, https://vcity.diginext.fr/images/news/publications/CS_GROUP_publication_-_video_tracking_and_classification_for_countering_UAS.pdf
71. Improving DeepFake Detection: A Comprehensive Review of ..., accessed on June 13, 2025, https://jcbi.org/index.php/Main/article/download/572/532/1757
72. tsingqguo/AttackTracker - GitHub, accessed on June 13, 2025, https://github.com/tsingqguo/AttackTracker
73. (PDF) Robustness Evaluation for Video Models with Reinforcement ..., accessed on June 13, 2025, https://www.researchgate.net/publication/392423233_Robustness_Evaluation_for_Video_Models_with_Reinforcement_Learning