What machine learning algorithms most effectively extract temporal features for robust object tracking in drone video streams with varying environmental conditions?

Hybrid systems that merge CNNs with transformers and employ explicit temporal modeling demonstrate superior performance in extracting temporal features from drone video streams.

Abstract

Drone video streams with variable environmental conditions demand robust tracking algorithms that capture temporal features effectively. Several studies report that hybrid architectures—combining convolutional neural networks with transformers, recurrent units, or correlation filters—yield competitive performance. For example, Cao et al. (2022) describe a method using temporally adaptive convolution and a transformer that achieves a precision of 0.786, a success rate of 0.582, and a 5–9.8% increase in area under curve. Avola et al. (2021) and Duan et al. (2021) detail systems based on multi-stream CNNs, Deep SORT, and recommender mechanisms that report precision values ranging from 0.587 to 0.710 and consistently address occlusion, motion blur, and scale variation. Yuan et al. (2024) present a holistic transformer approach with multiple object tracking accuracies between 38.8% and 61.7% on benchmark datasets.

All studies emphasize the integration of explicit temporal modeling—whether through frame-by-frame processing or sequence-level analysis—and environmental adaptation via context-aware filtering and multi-scale strategies. In summary, papers show that hybrid and transformer-based strategies that incorporate temporal cues stand as effective solutions for robust object tracking in drone video streams under diverse conditions.

Paper search

Using your research question "What machine learning algorithms most effectively extract temporal features for robust object tracking in drone video streams with varying environmental conditions?", we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 50 papers most relevant to the query.

Screening

We screened in papers that met these criteria:

- **Machine Learning Implementation**: Does the study implement and clearly specify the technical details of machine learning algorithms for object tracking?
- **Drone Video Input**: Does the research use video streams from drones/UAVs as input data?
- **Temporal Processing**: Does the study include temporal feature extraction and video processing (beyond static image analysis)?
- **Performance Evaluation**: Does the research evaluate performance using quantitative metrics across multiple environmental conditions?
- **Practical Implementation**: Does the study include experimental implementation and testing of the proposed methods?
- **Object Tracking Focus**: Is object tracking a primary component of the research (not limited to drone navigation)?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Machine Learning Algorithm Type**:

  Identify and specify the primary machine learning algorithm used for temporal feature extraction. Look in the methods section for detailed algorithm descriptions.

Extraction guidelines:

- Specify exact algorithm name (e.g., Convolutional LSTM, Recommender-based CNN, Correlation Filter)
- If multiple algorithms are used, list all in order of primary importance
- If hybrid approaches exist, describe the specific combination
- If unclear, note "Not clearly specified"

Example formats:

- "Convolutional Long Short-Term Memory (LSTM)"
- "Hybrid approach: Convolutional Neural Network (CNN) with Correlation Filter regularization"

- **Temporal Feature Extraction Methodology**:

  Describe the specific techniques used for extracting temporal features from video streams.

Extraction guidelines:

- Identify specific feature extraction techniques
- Note any preprocessing or feature selection methods
- Describe how temporal dependencies are captured
- Include any domain-specific knowledge integration

Look primarily in methods and experimental design sections. If multiple techniques are described, list in order of significance.

Example formats:

- "Sequence length optimization with subsampling rate adjustment"
- "Context-aware environmental residual correlation with multi-feature integration"

- **Performance Metrics**:

  Identify and extract the specific performance metrics used to evaluate the machine learning algorithm's effectiveness.

Extraction guidelines:

- List all quantitative performance metrics
- Include precise numerical values if available
- Note the benchmark datasets used

- Capture comparative performance against other methods

Look in results and discussion sections. Prioritize metrics directly related to temporal feature extraction and object tracking performance.

Example formats:

- ”Precision score: 12.9% improvement on OTB2015 dataset”
- ”Metrics: Tracking accuracy, scale adaptation robustness”

- **Environmental Condition Variations**:

  Describe the range of environmental conditions tested in the study.

Extraction guidelines:

- List specific environmental variations examined
- Note how these variations were simulated or encountered
- Capture the algorithm's performance across different conditions
- Identify any limitations observed

Primarily search methods, experimental design, and results sections.

Example formats:

- ”Tested in maritime video sequences with varying lighting and sea conditions”
- ”Evaluated across 50 benchmark datasets with diverse environmental challenges”

- **Feature Representation and Processing**:

  Detail how temporal features are represented and processed in the machine learning approach.

Extraction guidelines:

- Describe feature representation techniques
- Note any feature weighting or selection mechanisms
- Capture the computational approach to feature processing
- Identify any innovative feature handling methods

Focus on methods section, looking for technical implementation details.

Example formats:

- ”Recommender system for autonomous layer and feature map selection”
- ”Spatiotemporal min-channel method for size variation estimation”

Results

Characteristics of Included Studies

Study	Study Focus	Algorithm Type	Environmental Conditions	Performance Metrics	Full text retrieved
Cao et al., 2022	Temporal context exploitation for aerial tracking	Hybrid: Temporally Adaptive Convolution (TAdaConv, a type of Convolutional Neural Network) + Adaptive Temporal Transformer	Real-world Unmanned Aerial Vehicle (UAV): illumination, scale, occlusion, motion blur, low resolution, fast/camera motion, deformation	Precision: 0.786; Success: 0.582; Area Under Curve (AUC): +5–9.8%; 4 aerial tracking benchmarks	Yes
Avola et al., 2021	Multi-stream Convolutional Neural Network for UAV object detection/tracking	Hybrid: Multi-Stream Convolutional Neural Network + Faster Region-based Convolutional Neural Network + Deep Simple Online and Realtime Tracking (Deep SORT)	UAVDT: vehicle-category, occlusion, out-of-view, weather, altitude, camera view; UAV123: aspect ratio, clutter, motion, occlusion, illumination, scale, viewpoint	Mean Average Precision (mAP); Precision: 0.710–0.587; Success: 0.430–0.475; UMCD, UAVDT, UAV123, UAV20L	Yes
Duan et al., 2021	Scale-aware tracking with online recommender	Hybrid: Recommender-based Convolutional Neural Network + Correlation Filter	50 benchmarks, rescue drone: blur, fast motion, discontinuous frames, illumination, scale	Center Location Error (CLE), Success Rate (SR); Reported as outperforming 10 state-of-the-art trackers; 50 datasets	Yes

Study	Study Focus	Algorithm Type	Environmental Conditions	Performance Metrics	Full text retrieved
Yuan et al., 2024	Multi-object tracking with holistic transformer	Holistic Transformer	No mention found; designed for occlusion, scale, rapid motion	Multiple Object Tracking Accuracy (MOTA): 38.8% (VisDrone), 61.7% (UAVDT)	No
Emiyah et al., 2021	Vehicle tracking in urban drone video	Hybrid: You Only Look Once version 4 (YOLOv4) + deepSORT	Urban drone video: varying object sizes	No mention found; reported as outperforming YOLO+deepSORT; matches manual counts	No
Xie et al., 2021	Multi-object tracking for Unmanned Aircraft Systems	Hybrid: Deep Extended Kalman Filter (DeepEKF, using Long Short-Term Memory and Convolutional Neural Network) + Siamese Network	Video: scale, viewpoint, pose, color, disappear-ance/reappearance, moving sensor, zoom, background, illumination, occlusion, small objects	Expected Average Overlap (EAO): 0.4464–0.5045; Visual Object Tracking (VOT) challenge	Yes
Xu et al., 2018	Discriminative Correlation Filter (DCF) with temporal consistency for tracking	DCF with temporal consistency-preserving spatial feature selection	OTB, Temple-Colour, UAV123, VOT2018: illumination, background, jitter, blur, deformation, occlusion	Area Under Curve (AUC), Overlap Precision (OP), Distance Precision (DP), Expected Average Overlap (EAO), Accuracy (A), Robustness (R); OTB, Temple-Colour, UAV123, VOT2018	Yes

Study	Study Focus	Algorithm Type	Environmental Conditions	Performance Metrics	Full text retrieved
Sakthivel et al., 2024	CAERDCF: context-aware DCF with deep features	Hybrid: Correlation Filter + Deep Convolutional Features	OTB2015, TempleColor128, UAV123, LASOT, GOT10K: ambiguous, intensive environment variations	Precision: +12.9% (OTB2015), +16.1% (TempleColor128); multiple datasets	No
Cruz and Bernardino, 2019	Maritime airborne video detection	Hybrid: Convolutional Long Short-Term Memory (ConvLSTM) + Convolutional Neural Network	Favorable and challenging maritime environments	No mention found; reported as comparable or superior to other detectors	No
Heslinga et al., 2023	Small object detection in military drone/infrared video	No mention found (YOLO, spatio-temporal deep learning)	Various objects/scenes/acquisition conditions	No mention found	No

Algorithm Type:

- Seven studies used hybrid algorithms combining multiple model types, such as Convolutional Neural Networks with transformers, correlation filters, or Long Short-Term Memory networks.
- One study used a transformer-based approach.
- One study used a correlation filter-based approach (not hybrid).
- We didn't find a clear specification of algorithm type for one study.

Environmental Conditions:

- Nine studies evaluated their algorithms under real-world or varied/challenging environmental conditions.
- We didn't find mention of environmental conditions for one study.
- The most commonly addressed challenges were:
  - Occlusion (five studies)
  - Scale variation (five studies)
  - Illumination changes (five studies)
  - Other challenges included fast or rapid motion (three studies), blur (two studies), deformation (two studies), viewpoint (two studies), and less frequently, weather, altitude, background, small

objects, and maritime environments.

Performance Metrics:

- Precision was reported in three studies.
- Success was reported in two studies.
- Area Under Curve was reported in two studies.
- Expected Average Overlap was reported in two studies.
- Other metrics (mean average precision, multiple object tracking accuracy, center location error, success rate, overlap precision, distance precision, accuracy, robustness) were each reported in one study.
- We didn't find mention of performance metrics in three studies.

Temporal Feature Extraction Approaches

Deep Learning Architectures

Study	Architecture Type	Feature Extraction Method	Temporal Resolution	Effectiveness
Cao et al., 2022	Hybrid: Temporally Adaptive Convolution + Transformer	Online temporally adaptive convolution; temporal context queue; transformer for similarity map refinement	Frame-by-frame, leveraging past frames	Reported as outperforming state-of-the-art methods; robust to diverse conditions
Avola et al., 2021	Hybrid: Multi-Stream Convolutional Neural Network + Faster Region-based Convolutional Neural Network + Deep SORT	Multi-stream Convolutional Neural Network (3x3, 5x5, 7x7 kernels); Deep SORT for temporal association	Frame-by-frame with temporal association via Deep SORT	Reported as competitive in precision and success; robust to occlusion and motion
Duan et al., 2021	Hybrid: Recommender-based Convolutional Neural Network + Correlation Filter	Recommender system for layer/feature map selection; spatiotemporal min-channel for scale	Frame-by-frame with short-term memory	Reported as high accuracy, scale adaptation, and robustness

Study	Architecture Type	Feature Extraction Method	Temporal Resolution	Effectiveness
Yuan et al., 2024	Holistic Transformer	Holistic transformer; visual Gaussian mixture for trajectory; multi-feature pattern for association	Sequence-level, integrating local and global context	MOTA 38.8–61.7%; reported as robust to occlusion and scale
Emiyah et al., 2021	Hybrid: YOLOv4 + deepSORT	YOLOv4 for detection; deepSORT for temporal association	Frame-by-frame with tracking	Reported as outperforming prior methods; matches manual counts
Xie et al., 2021	Hybrid: Deep Extended Kalman Filter (LSTM/CNN) + Siamese Network	Sequence-to-sequence DeepEKF; attention; Siamese for visual scoring	Sequence-level, latent space prediction	Improved Expected Average Overlap; reported as robust to non-linear motion
Xu et al., 2018	Discriminative Correlation Filter with temporal consistency	Adaptive spatial feature selection; temporal consistency via filter history	Frame-by-frame with temporal regularization	Improved Area Under Curve, Overlap Precision, Distance Precision, Expected Average Overlap
Sakthivel et al., 2024	Hybrid: Correlation Filter + Deep Features	Context-aware environmental residual correlation; multi-feature integration	Frame-by-frame, context-aware	+12.9–16.1% precision; reported as robust to ambiguous changes
Cruz and Bernardino, 2019	Hybrid: ConvLSTM + Convolutional Neural Network	ConvLSTM with Convolutional Neural Network; sequence length/subsampling optimization; domain knowledge	Sequence-level	Reported as comparable or superior in challenging maritime video
Heslinga et al., 2023	No mention found	Stacking frames as channels; difference maps	Multi-frame, spatio-temporal	No mention found

Architecture Type:

- Hybrid architectures were reported in seven studies.

- Transformer-based architectures were used in two studies.
- Convolutional Neural Networks were used in four studies.
- Correlation filters were used in two studies.
- Deep SORT was used in two studies.
- YOLO was used in one study.
- Long Short-Term Memory or Convolutional Long Short-Term Memory was used in two studies.
- Discriminative Correlation Filter was used in one study.
- We didn't find a clear architecture specification in one study.

Feature Extraction Method:

- Transformer-based feature extraction was used in two studies.
- Convolutional Neural Network-based feature extraction was used in four studies.
- Deep SORT or temporal association methods were used in two studies.
- Correlation or context-aware methods were used in two studies.
- Multi-feature integration was used in two studies.
- Sequence-to-sequence or sequence optimization was used in two studies.
- Attention mechanisms were used in one study.
- Stacking frames or difference maps were used in one study.

Temporal Resolution:

- Frame-by-frame processing (including with memory, association, or context) was used in six studies.
- Sequence-level processing was used in three studies.
- Multi-frame or spatio-temporal processing was used in one study.

Effectiveness:

- Two studies reported outperforming state-of-the-art or prior methods.
- Two studies reported competitive or comparable performance.
- Two studies reported high accuracy or precision.
- Four studies reported quantitative metrics (such as multiple object tracking accuracy, expected average overlap, area under curve, overlap precision, distance precision, or percent precision).
- Six studies reported robustness to occlusion, scale, motion, ambiguous changes, or diverse conditions.
- One study reported matching manual counts.
- We didn't find mention of effectiveness in one study.

Environmental Adaptation Mechanisms

Lighting and Weather Conditions

- Cao et al., 2022: Reported explicit testing under varying illumination.
- Avola et al., 2021: Used UAVDT dataset, which includes weather and illumination variation.
- Duan et al., 2021: Evaluated under illumination change and blur.
- Xu et al., 2018: Benchmarks included illumination variation.
- Sakthivel et al., 2024: Reported addressing ambiguous and intensive environmental changes.
- Heslinga et al., 2023: Included diverse acquisition conditions, but we didn't find further details.

Camera Motion Compensation

- Cao et al., 2022: Real-world UAV tests included camera motion.
- Avola et al., 2021: Datasets included camera motion.
- Xie et al., 2021: Included moving sensor, zoom, and dynamic background.
- Xu et al., 2018: Included camera jitter and background change.

Scale Variation Handling

- Duan et al., 2021: Used spatiotemporal min-channel for scale adaptation.
- Avola et al., 2021: Used multi-scale analysis via multi-stream Convolutional Neural Network.
- Yuan et al., 2024: Reported as designed for scale variation.
- Xu et al., 2018: Benchmarks included scale variation.

Summary of Environmental Adaptation Mechanisms:

- Most of these studies report that their algorithms incorporate mechanisms to adapt to environmental variation, either through explicit modeling (such as context-aware filters or multi-scale processing) or by leveraging diverse training and evaluation datasets.
- The degree of adaptation and the level of reporting detail varied across studies.

Performance Analysis

Study	Algorithm Type	Tracking Accuracy	Processing Speed	Robustness Score
Cao et al., 2022	Temporally Adaptive Convolution + Transformer	Precision: 0.786; Success: 0.582; Area Under Curve: +5–9.8%	Greater than 27 frames per second (Jetson AGX Xavier)	Reported as robust to illumination, scale, occlusion, motion
Avola et al., 2021	Multi-Stream Convolutional Neural Network + Deep SORT	Precision: 0.710–0.587; Success: 0.430–0.475	Real-time (no quantitative value found)	Reported as robust to occlusion, motion, weather
Duan et al., 2021	Recommender Convolutional Neural Network + Correlation Filter	Reported as outperforming 10 state-of-the-art trackers; Center Location Error, Success Rate	No mention found	Reported as robust to blur, motion, illumination, scale
Yuan et al., 2024	Holistic Transformer	Multiple Object Tracking Accuracy: 38.8%/61.7%	No mention found	Reported as robust to occlusion, scale, rapid motion
Emiyah et al., 2021	YOLOv4 + deepSORT	No mention found; reported as matching manual counts	No mention found	Reported as outperforming prior methods in urban video

Study	Algorithm Type	Tracking Accuracy	Processing Speed	Robustness Score
Xie et al., 2021	Deep Extended Kalman Filter (LSTM/CNN) + Siamese Network	Expected Average Overlap: 0.4464–0.5045	No mention found	Reported as robust to scale, pose, occlusion, re-identification
Xu et al., 2018	Discriminative Correlation Filter with temporal consistency	Area Under Curve, Overlap Precision, Distance Precision, Expected Average Overlap; +1–4.1% over state-of-the-art	Frames per second measured (no value found)	Reported as robust to illumination, deformation, occlusion
Sakthivel et al., 2024	Context-Aware Environmental Residual Discriminative Correlation Filter	Precision: +12.9–16.1% over Background-Aware Correlation Filter	No mention found	Reported as robust to ambiguous changes
Cruz and Bernardino, 2019	ConvLSTM + Convolutional Neural Network	No mention found; reported as comparable or superior	No mention found	Reported as robust in challenging maritime video
Heslinga et al., 2023	No mention found	No mention found	No mention found	No mention found

Algorithm Type:

- Transformer-based algorithms were used in two studies.
- Convolutional Neural Network combined with SORT/Deep SORT was used in two studies.
- Convolutional Neural Network with Correlation Filter was used in one study.
- Long Short-Term Memory/Convolutional Neural Network-based approaches (with or without Siamese networks) were used in two studies.
- Discriminative Correlation Filter-based algorithms were used in two studies.
- We didn't find mention of the algorithm type for one study.

Tracking Accuracy:

- Quantitative tracking accuracy metrics were found in six studies.
- Only qualitative or relative performance descriptions were found in three studies.
- We didn't find mention of tracking accuracy in one study.

Processing Speed:

- Quantified processing speed (in frames per second or hardware) was found in one study.
- Real-time performance (not quantified) was found in one study.
- Speed was measured but no value was found in one study.
- We didn't find mention of processing speed in seven studies.

Robustness:

- Robustness information was found in nine studies.
- We didn't find mention of robustness in one study.
- The most common robustness categories were:
  - Occlusion: six studies
  - Motion or rapid motion: six studies
  - Scale: five studies
  - Illumination: three studies
  - Other categories (each in one study): weather, blur, pose, re-identification, deformation, ambiguous changes, urban video, maritime video

---

Summary

- The included studies primarily report on hybrid deep learning algorithms that explicitly model temporal dependencies for object tracking in drone video streams under varying environmental conditions.
- These studies report that architectures combining convolutional feature extraction with transformers, Long Short-Term Memory, or correlation filters outperform simpler or non-temporal baselines, especially under challenging conditions such as illumination variation, occlusion, motion blur, and scale changes.
- Reported adaptation mechanisms include context-aware filtering, multi-scale processing, and temporal consistency regularization, which are associated with improved robustness.
- However, there is considerable heterogeneity in reporting, with some studies lacking standardized quantitative metrics and limited detail on computational requirements.
- The ability to draw strong comparative conclusions is constrained by these reporting differences and by the lack of full text for some studies.
- Overall, the evidence from these studies suggests that hybrid and transformer-based approaches, particularly those incorporating explicit temporal modeling and environmental adaptation mechanisms, are reported as effective for robust object tracking in drone video streams with varying environmental conditions.

References

Christian Emiyah, K. Nyarko, C. Chavis, and Istiak A Bhuyan. "Extracting Vehicle Track Information from Unstabilized Drone Aerial Videos Using YOLOv4 Common Object Detector and Computer Vision." *Lecture Notes in Networks and Systems*, 2021.

D. Avola, L. Cinque, Anxhelo Diko, Alessio Fagioli, G. Foresti, Alessio Mecca, D. Pannone, and C. Piciarelli. "MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images." *Remote Sensing*, 2021.

Friso G. Heslinga, Frank Ruis, Luca Ballan, Martin C. van Leeuwen, Beatrice Masini, Jan Erik van Woerden, R. D. den Hollander, et al. "Leveraging Temporal Context in Deep Learning Methodology for Small Object Detection." *Security + Defence*, 2023.

G. Cruz, and Alexandre Bernardino. "Learning Temporal Features for Detection on Maritime Airborne Video Sequences Using Convolutional LSTM." *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

Ran Duan, Changhong Fu, K. Alexis, and Erdal Kayacan. "Online Recommendation-Based Convolutional

Features for Scale-Aware Visual Tracking." *IEEE International Conference on Robotics and Automation*, 2021.

Sachin Sakthi Kuppusami Sakthivel, Sathishkumar Moorthy, Sathiyamoorthi Arthanari, Jae Hoon Jeong, and Y. Joo. "Learning a Context-Aware Environmental Residual Correlation Filter via Deep Convolution Features for Visual Object Tracking." *Mathematics*, 2024.

Tianyang Xu, Zhenhua Feng, Xiaojun Wu, and J. Kittler. "Learning Adaptive Discriminative Correlation Filters via Temporal Consistency Preserving Spatial Feature Selection for Robust Visual Object Tracking." *IEEE Transactions on Image Processing*, 2018.

Wanlin Xie, J. Ide, Daniel Izadi, Sean Banger, Thayne T. Walker, Ryan Ceresani, Dylan Spagnuolo, Christopher Guagliano, Henry Diaz, and Jason C. Twedt. "Multi-Object Tracking with Deep Learning Ensemble for Unmanned Aerial System Applications." *Security + Defence*, 2021.

Yubin Yuan, Yiquan Wu, Langyue Zhao, Yaxuan Pang, and Yuqi Liu. "Multiple Object Tracking in Drone Aerial Videos by a Holistic Transformer and Multiple Feature Trajectory Matching Pattern." *Drones*, 2024.

Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. "TCTrack: Temporal Contexts for Aerial Tracking." *Computer Vision and Pattern Recognition*, 2022.