How can temporal features be extracted in a trustworthy manner for use in object detection and tracking in drone video streams using AI

Extracting temporal features in a trustworthy manner is crucial for accurate object detection and tracking in drone video streams using AI. The latest research emphasizes combining temporal and spatial information, leveraging attention mechanisms, and using adaptive aggregation strategies to ensure robust and reliable feature extraction.

Key Approaches for Trustworthy Temporal Feature Extraction

- Temporal-Spatial Feature Interaction: Integrating temporal feedback loops and spatial information from multiple drones enhances both localization and identification of objects, making tracking more robust against occlusion and visual decay. Utilizing prior knowledge from tracklets across frames further improves reliability in challenging scenarios (Wu et al., 2025).
- Temporal Embedding and Association: Separating detection and identification tasks while embedding temporal associations strengthens the model's ability to represent and track objects over time, reducing interference between tasks and improving overall tracking accuracy (Lin et al., 2022).
- Recurrent and Attention-Based Networks: Recurrent motion attention networks and attention mechanisms aggregate features from multiple frames, learning motion patterns and temporal context efficiently. These methods address issues like background clutter and appearance changes, leading to more trustworthy detection and tracking (Zhou et al., 2024; Fujitake & Sugimoto, 2022).
- Adaptive and Hierarchical Aggregation: Temporal-adaptive sparse feature aggregation and multi-level spatial-temporal frameworks selectively combine information from relevant frames and object proposals, filtering out irrelevant data and enhancing feature consistency across time (Xu et al., 2022; He et al., 2022).

Techniques and Mechanisms

Technique/Mechanism	Trustworthiness Contribution	Citations
Temporal-spatial feedback loops	Reduces sensitivity to visual decay, improves ID	(Wu et al., 2025)
Temporal embedding structures	Strengthens temporal representation, reduces errors	(Lin et al., 2022)
Recurrent motion attention	Learns long-term motion, handles background changes	(Zhou et al., 2024)
Attention-based temporal aggregation	Balances accuracy and speed, leverages past context	(Fujitake & Sugimoto, 2022)
Adaptive sparse sampling	Efficiently encodes informative frames, reduces noise	(He et al., 2022)
Hierarchical feature interaction Integrates multi-level info	filters irrelevant data	(Xu et al., 2022)

Ensuring Trustworthiness

- Cross-frame and cross-drone feature interaction increases robustness to occlusion and target similarity (Wu et al., 2025; Xu et al., 2022).
- Attention and memory mechanisms help focus on relevant temporal cues while maintaining real-time performance (Zhou et al., 2024; Fujitake & Sugimoto, 2022).
- Adaptive sampling and aggregation minimize the risk of propagating errors from noisy or irrelevant frames (Xu et al., 2022; He et al., 2022).

Conclusion

Trustworthy temporal feature extraction in drone video streams is achieved by combining temporal and spatial cues, using attention and memory mechanisms, and employing adaptive aggregation strategies. These approaches collectively enhance the reliability and accuracy of object detection and tracking in dynamic and challenging drone environments.

References

Wu, H., Sun, H., Ji, K., & Kuang, G. (2025). Temporal-Spatial Feature Interaction Network for Multi-Drone Multi-Object Tracking. IEEE Transactions on Circuits and Systems for Video Technology, 35, 1165-1179. https://doi.org/10.1109/TCSVT.2024.3478758

Lin, Y., Wang, M., Chen, W., Gao, W., Li, L., & Liu, Y. (2022). Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure. Remote. Sens., 14, 3862. https://doi.org/10.3390/rs14163862

Zhou, Z., Yu, X., & Wang, X. (2024). Object detection in drone video based on recurrent motion attention. Pattern Recognit. Lett., 183, 56-63. https://doi.org/10.1016/j.patrec.2024.04.029

Xu, C., Zhang, J., Wang, M., Tian, G., & Liu, Y. (2022). Multilevel Spatial-Temporal Feature Aggregation for Video Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 32, 7809-7820. https://doi.org/10.1109/TCSVT.2022.3183646

Fujitake, M., & Sugimoto, A. (2022). Temporal feature enhancement network with external memory for live-stream video object detection. Pattern Recognit., 131, 108847. https://doi.org/10.1016/j.patcog.2022.108847

He, F., Li, Q., Zhao, X., & Huang, K. (2022). Temporal-adaptive sparse feature aggregation for video object detection. Pattern Recognit., 127, 108587.

https://doi.org/10.1016/j.patcog.2022.108587