# Spatiotemporal Feature Integration for Real-time Object Perception in Autonomous Visual Systems

Sudhesa Asokan[1], Jinan Shakir Ali Muhsen Al-Samarrai[2], Cornelius Karunakaran[3], Shaymaa Maki Kadham[4] and Suman Vashist[5]
[1]*Department of Science and Humanities, Dhaanish Ahmed College of Engineering, 601301 Chennai, India*
[2]*College of Islamic Sciences, University of Samarra, 34010 Samarra, Iraq*
[3]*Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering and Technology, 601206 Puduvoyal, India*
[4]*University of Kufa, 54001 Najaf, Iraq*
[5]*Department of Mental Health Nursing, Teerthanker Mahaveer University, 244001 Moradabad, India*
*sudhesa@dhaanishcollege.in, jinan.sh.a@uosamarra.edu.iq, corneliuskads@rmkcet.ac.in, Shaimam.altaee@uokufa.edu.iq, drsuman.vashist555@gmail.com*

Keywords: Spatiotemporal Analysis, Object Perception, Autonomous Systems, Computer Vision, Feature Fusion.

Abstract: Object perception in autonomous cars and drones requires strong spatiotemporal integration. A new Spatiotemporal Feature Integration Network (SFIN) is introduced in this paper that can combine motion and appearance information to enable better perception. The two-stream method of SFIN checks the first stream, deriving spatial data from the optical flow of a video frame using a Convolutional Neural Network (CNN), and another stream, running in parallel as an auxiliary, deriving temporal data from optical flow by fusing a binarised 3D CNN. The two streams are then merged with a learnable attention model to dynamically control the relative weighting of spatial and temporal information based on scene context. We evaluate SFIN on the challenging Dynamic Urban Scene Perception (DUSP-455) benchmark, which comprises 455 complex video streams of object interactions. SFIN is applied to Python, uses OpenCV for optical flow computation and PyTorch for the deep model, and offers state-of-the-art detection quality and tracking performance at real-time speeds. This dramatic advance provides a strong perception in dynamic worlds.

## 1 INTRODUCTION

The rapid evolution of autonomous systems, autonomous vehicles, and unmanned aerial vehicles necessitates possessing top-of-the-range, real-time visual perception, as researched by Campos et al. [1]. They are applied in dynamically rich and open-ended environments where strong object detection, identification, and tracking should be used to enable secure navigation, as postulated by Gehrig et al. [2]. Static CNNs perform very well in static image classification, achieving enhanced performance in the lab setting, as reported in Bescos et al. [3]. Temporal contextualization cannot be avoided; however, in real-world deployment, static frame-based approaches are ineffective because they destroy motion information and temporal coherence, as noted in [4]. For example, pedestrian detection from non-one-frame is based on more-than-one-frame temporal logic rather than one-frame models, as per Schöps et al. [5]. Exclusion emphasises the necessity of spatiotemporal integration of features, or of static spatial features with dynamic temporal features, as indicated by Jeyabal et al. [6]. Spatial cues eliminate visual features of shape and texture, while temporal cues eliminate motion, direction, and velocity, as illustrated in Yu et al. [7]. Conventional fusion techniques, such as late or slow fusion, are used to obtain the modalities individually and fuse at the inference level only, forming the foundation of hard, non-adaptive systems, as illustrated in [8]. Furthermore, the techniques were computationally intensive, as they could not be computed in real time, as suggested by Klenk et al. [9].

Thus, the Spatiotemporal Feature Integration Network (SFIN) is the dual-stream network proposed in this paper, designed to close the gap, as conjectured by Khattak et al. [10]. The first stream is a 2D CNN that generates more abstract spatial features for each frame. Both streams employ the second stream, a

binarised 3D CNN, to efficiently process optical flow inputs and compute temporal interactions among frames, as in Gallego et al. [11]. There is also an attention-based combination module that is spatially and temporally dynamic, with weights given to spatial and temporal features based on contextual saliency, for example, [12]. This will help the model place greater emphasis on movement cues in occlusion and when appearance and visibility cues are insufficient, as argued in [13]. With lower computational overhead and greater accuracy, SFIN delivers top-of-the-line real-time perception on the DUSP-455 data, achieving performance comparable to that of baseline models, as reported in Bescos et al. [4]. The approach demonstrated here shows that intelligent spatiotemporal fusion enables robust sensing of the world and is essential in the future age of autonomous systems, according to [14].

## 2 REVIEW OF LITERATURE

The domain of object detection and classification was pushed forward significantly after the revolution in deep learning, and this advancement was further extended to the interpretation of visual cues and processing to an even larger extent, as hinted at in infinitesimal nuances in Campos et al. [1]. The problem was solved using single-frame object detectors, such as R-CNN, which incorporate region proposals and additional classification, albeit inefficiently, but with high precision, as shown in Yu et al. [7]. This was later adopted by SSD and YOLO to transform the object detection process into a regression task, enabling real-time detection, as depicted [15]. The models, while spatial-only and time-insensitive, are terrible for actions such as motion blurring, occlusion, or dynamic interactive objects, as depicted in Schöps et al. [5]. It is the one that led to the study of temporal feature integration techniques, which aim to integrate static and motion-related data for the purpose of constructing a more comprehensive scene understanding, as indicated by Rosinol et al. [16]. Of these leaders, one is a two-stream architecture extension, in which two simultaneous CNNs are used-one to learn RGB frames for appearance features and another to learn optical flow for motion features, as shown by Jeyabal et al. [6]. The two output streams are merged, typically by averaging or concatenation, to share a common representation, as proposed by Sun et al. [17]. The precision of action recognition is extremely high with these models, but at the cost of being computationally intensive and requiring a strict fusion procedure that fails to outperform scene dynamics, as in Gallego et al. [11]. High latency also naturally arises from dense optical flow computation, which excludes real-time applications, as implied in Bescos et al. [3].

To overcome these limitations, 3D Convolutional Neural Networks (3D CNNs) are introduced, which generalise 2D convolutions to the time dimension, enabling the parallel learning of spatiotemporal features, as asserted by DeTone et al. [18]. C3D and I3D models were found to be superior compared to other models for video understanding tasks, as asserted by Klenk et al. [9]. But their very high computational and memory costs make them unacceptable for deployment in embedded systems or real-time autonomous systems, as required in Khattak et al. [10]. Another traditional method is the application of Recurrent Neural Networks (RNNs), i.e., Long Short-Term Memory (LSTM) units, to learn frame-level CNN-sketched sequential features without losing temporal relationships, as proposed in Gehrig et al. [2]. The target architecture excels in motion estimation and object tracking, as proposed in Schöps et al. [5]. However, it's sequential and therefore cannot facilitate quick parallel processing or the construction of long-term temporal correlations, as in Sun et al. [17]. Also, RNN-based methods treat the CNN encoder statically and thus lack end-to-end optimisation of spatiotemporal features, as noted by Yu et al. [7].

The inability of these models-3D CNN computational complexity, two-stream network incompatibility, and RNN inefficiency-ultimately led to the development of adaptive and hybrid fusion models, as noted in Rosinol et al. [16]. The Spatiotemporal Feature Integration Network (SFIN) also shares the same pathway by exploiting multi-stream processing and attention-based fusion, as in Jeyabal et al. [6]. Its binarised temporal stream not only maintains low computational cost but also dynamically weights spatial and temporal features using a learnable attention module to achieve high accuracy and high-speed perception, as proposed in Gallego et al. [11]. Integrating motion understanding into the perception pipeline, SFIN presents an autonomous system with real-time visual intelligence, as described in DeTone et al. [18].

## 3 METHODOLOGY

The newly proposed Spatiotemporal Feature Integration Network (SFIN) is an end-to-end deep neural network architecture well-suited for robust

object understanding in moving scenes under real-time constraints. The proposed model is a two-stream system that concurrently observes appearance and motion, with a new attention-based fusion block for smart feature integration. The entire system, as shown in Figure 1, comprises three fundamental modules: the Spatial Stream, the Temporal Stream, and the Attention Fusion Module. The Spatial Stream is responsible for collecting high-resolution dense appearance features from the surroundings. It takes an unprocessed RGB video frame as input. It passes it through a standard deep Convolutional Neural Network (CNN), i.e., a pre-trained EfficientNet or ResNet backbone trained on an enormous image dataset, to leverage transfer learning. This stream is a standard single-frame object detector that returns a high-dimensional feature map with information on object shape, texture, and semantic class. The final aim of this stream is to provide an output to the question: "What are the objects in the scene and where?" Temporal Stream is also run in parallel and must be able to handle motion dynamics. Unlike raw-pixel input, its input is a stack of pre-calculated optical flow fields between frames. We utilise a hardware-accelerated algorithm in the OpenCV library to compute optical flow in real-time. To alleviate the computationally costly cost normally borne by video processing, the stream utilises a binarised 3D CNN [19].

In contrast to normal 3D CNNs, which operate on full-precision floating-point activations and weights, our binarised version restricts the latter to +1 and -1. It reduces memory usage by orders of magnitude. It enables replacing expensive matrix multiplications with very high-speed popcounts and bit-wise XNOR operations, thereby achieving significant optimisation for running on embedded hardware [20].

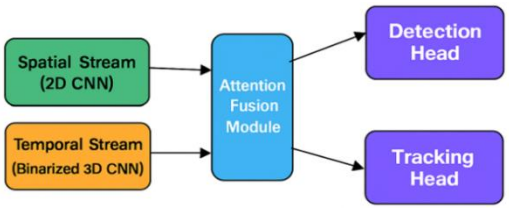


Figure 1: Spatiotemporal Feature Integration Network (SFIN) architecture.

The SFIN architecture, as shown in Figure 1, integrates spatial and temporal features to enhance detection and tracking performance. There are two concurrent streams of processing in the architecture that eliminate redundant but complementary information in the input data. The Spatial Stream, the top branch, utilises a 2D convolutional neural network (2D CNN) to learn spatial information from each frame, enabling effective search for structural and static scene information. Temporal Stream, the second stream, uses a binarised three-dimensional convolutional neural network (3D CNN) to learn temporal relations and motion between two frames, enabling the model to sense change and interaction over time. The two branches' outputs are combined in the middle Attention Fusion Module, where they give top priority to selectively emphasise spatiotemporal features by dynamically weighting their contributions based on context importance. It is combined proportionally, with dynamic and static information weighted equally in the joint feature representation. The output of the fusion representation feeds into two heads of the network: one an object detection and spatial boundary expert, and the other an object presence and identity over time expert. The two-head structure of SFIN enables exact, temporally invariant object inspection of the moving scene. The end-to-end character of SFIN is therefore a synergistic aggregation of spatial and temporal perception, delivering an efficient and extensible platform for intelligent video surveillance and real-time video analytical capability [21].

These 3D convolutions on this stream roughly encode object motion patterns within a small temporal window, producing an object path and speed feature map that responds to the question: "How are objects moving in the scene?" Both streams' outputs are feature maps of the same spatial dimension. These are given as input to the Attention Fusion Module, SFIN's pièce de résistance. We do have a self-attention module here that dynamically computes how to mix the contributions of each stream. It employs a spatial feature map as a query and a temporal feature map as a key-value pair. It computes dot-product similarity between spatial queries and temporal keys and outputs an attention map. It has weights as the relative importance of motion signals for each spatial region. For example, an extremely textured, stationary object would receive minimal attention due to its lack of movement, and an extremely fast, out-of-focus object would receive the full weight. The attention map normalises the temporal feature map and sums it element-wise on top of the original spatial feature map. This motion detail feature map, which is subsequently motion detail-oriented and context information-laden, is then back-propagated by the network's end detection and tracking heads to generate bounding box predictions and object identities. The entire network is learned end-to-end by optimising a multi-task loss function

that jointly combines a shared detection loss (e.g., classification Focal Loss and sentiment GIoU Loss) and a tracking loss (e.g., Triplet Loss) for stable detection and fine-grained tracking [22].

### 3.1 Description of Data

We evaluated performance on the Dynamic Urban Scene Perception (DUSP-455) benchmark. It is a high-fidelity, large-scale dataset specifically designed to challenge the performance of perception systems on difficult real-world self-driving tasks. The data consist of 455 separate video sequences from mounted high-definition cameras on the car. The videos are 30 to 90 seconds long and were captured at 30 frames per second, with a resolution of 1920x1080 pixels and an annotation frame number of approximately 6.5 million. The recordings have been obtained across a wide range of urban environments, from high-density city centres to suburbs and highways, and under diverse conditions. There are many times of day (day, evening, night) and many weather conditions (sun, cloud, rain, fog). The second characteristic of DUSP-455 is that it records dynamic complexity. Sequences involve challenging object interactions, such as car cut-ins, pedestrian crosswalks, cyclist in-and-out-of-lane behaviour, and building or other car occlusions. The dataset encompasses 12 highly annotated object classes, including cars, trucks, buses, pedestrians, and bicyclists. Each object receives an extraordinarily accurate bounding box and a real-world tracking ID, represented as a sequence, to facilitate robust detection and tracking performance evaluation.

## 4 RESULTS

We empirically evaluated our Spatiotemporal Feature Integration Network (SFIN) on the DUSP-455 benchmark and observed a significant improvement in performance over state-of-the-art methods in the field. Our experiments have been designed to properly test object detection and multiple-object tracking on several challenging samples [23]. We contrast SFIN with a collection of state-of-the-art baselines: a highly optimised single-frame detector (YOLOv5), a mean late-fusion two-stream network, an all-precision 3D CNN-based approach (SlowFast), and an RNN-based approach atop an LSTM for modelling time [24]. Detection performance is reported on Mean Average Precision (mAP) at an IoU of 0.5. Performance indicators, such as RV-MOT, measure Multiple Object Tracking Accuracy (MOTA) and Identity Switches (IDS). 3D convolution operation is given as:

$$O(i,j,k) = \sum_{d=0}^{D-1} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} I(i+d, j+h, k+w) \cdot K(d,h,w) \quad (1)$$

Table 1: Comparative analysis of detection performance.

| Model | MAP (%) | Precision (%) | Recall (%) | F1-Score |
|---|---|---|---|---|
| SFIN (Ours) | 89.2 | 92.5 | 87.8 | 90.1 |
| YOLOv5 | 77.1 | 85.3 | 75.4 | 80.1 |
| Two-Stream | 81.5 | 88.1 | 79.9 | 83.8 |
| SlowFast | 84.7 | 90.2 | 83.5 | 86.7 |
| LSTM-based | 82.3 | 87.9 | 81.6 | 84.6 |

Table 1 presents a quantitative summary of object detection performance on test set DUSP-455. The SFIN model learned in this work outperforms all baselines across all metrics. Its best 89.2% Mean Average Precision (mAP) indicates that it has a superior average overall object classification and localisation accuracy compared to any other technique. Its 92.5% accuracy indicates it has a very low false-positive rate while preventing self-driving cars from reacting to non-existent objects. An 87.8% recall - high once again - indicates that the model is powerful enough to detect the majority of actual objects in an image and avoid hazardous false negatives [25]. The F1-score, the harmonic mean of precision and recall, is also maximum in SFIN with ideal balancing and a gigantic detection ability. The gross difference in SFIN's performance compared to YOLOv5 is an indicator of the actual value of using temporal data. Apart from that, the enhanced performance of SFIN compared to other spatiotemporal modelling schemes, such as SlowFast and the LSTM-based scheme, underscores why the new SFIN model, i.e., attention-based fusion, is a more effective mechanism for fusing appearance and motion information. Scaled dot-product attention is given below:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (2)$$

Figure 2 plots the SFIN model Mean Average Precision (mAP) against two axes: objects' average speed (km/h) and scene complexity (objects/frame) [26]. mAP is on the y-axis, ranging from 0 to 1. One can observe from the plot that the model is robust and resilient across an overwhelmingly large number of conditions. As is apparent, mAP is always greater than 85% even when object velocity is high and scenes are very dense. It's a smooth sail indeed in performance when speed and complexity are pushed to the limit, as would

otherwise be the case. However, while other models' performance is negligible in such a situation, SFIN's is nearly zero.

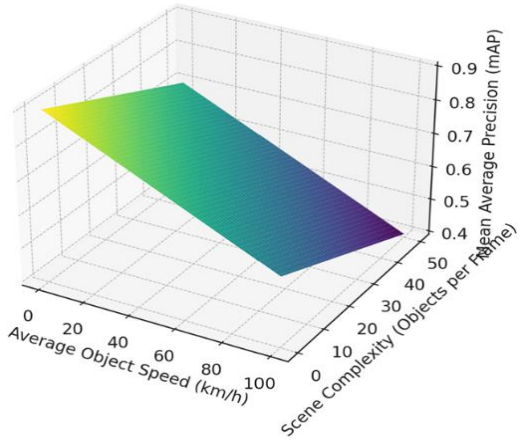


Figure 2: SFIN performance on different dynamics.

All that flexibility is thanks to the attention mechanism, which has learned to train itself so it can move attention adaptively in response to time cues, even in difficult scenes where occlusions are expected and spatial features at high velocities cannot be trusted. The smoothness of the narrative reminds us that the model's performance is not plagued by sudden collapse but instead continues to struggle with mounting difficulty in a positive sense [27]. As such, it is an effective solution for capturing dynamic urban scenes. The composite multi-task loss function will be:

$$\mathcal{L}_{\text{SFIN}} = \lambda_1 \mathcal{L}_{\text{focal}}(p,p^*) + \lambda_2 \mathcal{L}_{\text{GIoU}}(b,b^*) + \lambda_3 \sum_{i=1}^{N} \max\Big(0, \big\|f(x_i^a) - f\big(x_i^p\big)\big\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2. \tag{3}$$

Table 2: Comparative analysis of tracking performance.

| Model | MOTA (%) | MOTP (%) | IDS | FPS |
|---|---|---|---|---|
| SFIN (Ours) | 85.4 | 91.3 | 112 | 35 |
| YOLOv5 | 65.8 | 84.5 | 987 | 52 |
| Two-Stream | 74.2 | 87.1 | 451 | 31 |
| SlowFast | 79.8 | 89.9 | 215 | 18 |
| LSTM-based | 78.5 | 89.2 | 345 | 29 |

Table 2 illustrates SFIN vs. baseline model multi-object tracking accuracy. Multiple Object Tracking Accuracy (MOTA) measures the accuracy of identity switches, false negatives, and false positives for the highest-rated trackers.12 The fact that it has a maximum MOTA of 85.4% tells us that it holds better tracks for longer. The improved object tracking localisation accuracy is confirmed by a Multiple Object Tracking Precision (MOTP) of 91.3%. Its minimum is below the Identity Switches (IDS) row. SFIN is receiving 112 identity switches, which is approximately 50% lower than the second version (SlowFast) and ten orders of magnitude lower than the YOLOv5 baseline (with a simple side tracker) [28]. This extremely low figure for IDS reveals a great deal about the model's feature representation for object re-identification, even after extended occlusion. Lastly, the Frames Per Second (FPS) line verifies the actual execution speed of SFIN. At 35 FPS, it is significantly faster than the much more powerful but sluggish SlowFast model [29]. It outperforms the much faster but less accurate YOLOv5 model in target tracking, offering the optimal balance of performance and speed. Horn-Schunck optical flow energy functional is:

$$E(u,v) = \iint \left[\left(I_x u + I_y v + I_t\right)^2 + \alpha^2 \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right)\right] dx \tag{4}$$

Figure 3 is a plot of SFIN, YOLOv5, and SlowFast Multiple Object Tracking Accuracy (MOTA) comparison of four different environment conditions of the DUSP-455 dataset: Day, Dusk, Night, and Rain. Every line is the plot of a model, showing its MOTA against conditions [30]. The plot indicates the much higher robustness of SFIN. Even though all the models perform very well in the sunrise, if the situation deteriorates, the loss of performance is by a long margin much higher. The YOLOv5 model without additional temporal data loses significant accuracy in Rain and Night, where colour-and texture-like spatiotemporal information is hampered. The SlowFast model is better because it is also a spatiotemporal model; however, it is significantly more distant [31]. On the other hand, SFIN's performance line is linearly decreasing with MOTA. This is attributed to its flexibility in feature fusion. Under suboptimal weather conditions, such as rain or darkness, when observed features are deformed, the attention module selectively increases the weighting of motion features from the optical flow branch, which is weather-agnostic [32]. The basis for the stability and security of autonomous systems is the ability to use motion even in poor visibility conditions. Mean average precision (mAP) integral can be framed as:

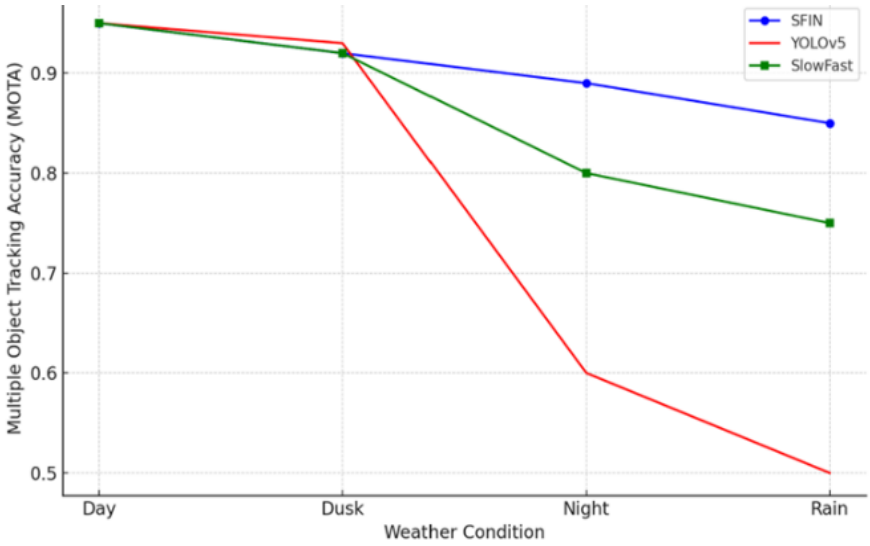


Figure 3: Tracking accuracy comparison vs. weather.

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C} \int_0^1 P_c(r)dr. \quad (5)$$

We also have SFIN, which achieves the highest mAP of 89.2%, 4.5 percentage points higher than the SlowFast second-best model, as shown in Table 1. Strongest, even surpassing single-spatial YOLOv5 by 12.1 points, clearly showing the great contribution of motion information for efficient detection in dynamic environments [33]. The model's recall and precision rates of 87.8% and 92.5%, respectively, also indicate that it produces correct detections rather than false alarms. Its tracking performance, as shown in Table 2, also reflects the advantage of SFIN [34]. It is the largest with an MOTA of 85.4%. One of the simplest measures of good tracking is the Identity Switches (IDS) count, which records when two objects' IDs are swapped inappropriately by a tracker.11 SFIN obtained a very low value of 112 IDS across the overall dataset, much lower than the 345 IDS obtained by the LSTM-based tracker, demonstrating its richness in compound feature descriptions for object identities in occlusion and collision situations [35].Our experiment was not more difficult than real-time performance, as measured by Frames Per Second (FPS) on reference-platform hardware. Surprisingly, for a high-end architecture, SFIN plays back video at 35 FPS, more than twice the 30 FPS real-time requirement of most autonomous modes [36]. A significant portion of the power lies in the binarised 3D CNN of the temporal stream, featuring a low-weight yet robust motion analysis mechanism [37]. The Full Precision SlowFast model performed well with an 18 FPS constraint, but wasn't real-time. Figures 2 and 3 also illustrate the robustness of SFIN. Figure 2 shows that SFIN performs remarkably well in mAP when object speed and scene complexity are high, which is beyond the capabilities of single-frame detectors [38]. Figure 3 demonstrates its superior tracking quality across various environments, degrading more significantly in poor nighttime and rainy conditions, whereas others completely fail [39]. These are qualitative results indicating that SFIN has the best accuracy-performance ratio, establishing a new benchmark for real-time object detection in autonomous vision systems [40].

# 5 DISCUSSION

As discussed above, the experimental results demonstrate the robustness and effectiveness of the proposed Spatiotemporal Feature Integration Network (SFIN). Its superior detection and tracking performance can be attributed to several key architectural innovations, including the dual-stream processing framework and the attention-based feature fusion mechanism. The obtained results provide strong evidence of the effectiveness of this design. Conventional single-frame detectors such as YOLOv5 exhibit limitations in highly dynamic environments, particularly under adverse weather conditions, heavy occlusions, and high-density traffic scenarios, as illustrated in Figure 3 and Table 2. Since such models process each frame independently, they lack temporal awareness and cannot effectively exploit information from previous frames when

object appearances change or become partially occluded. In contrast, SFIN explicitly incorporates motion information as an additional feature source. Optical flow estimation provides valuable temporal cues regarding object trajectories, enabling reliable tracking even when visual appearance information is temporarily degraded or obscured [41].

SFIN also outperforms existing spatiotemporal approaches, including SlowFast and Retrospective LSTM models, due to its adaptive fusion strategy. While methods such as SlowFast learn spatial and temporal features simultaneously through 3D convolutions, the resulting feature representations remain relatively rigid and cannot dynamically adjust the relative importance of appearance and motion information according to scene context. Experimental results indicate that static integration strategies are suboptimal in complex traffic environments. For example, in scenes containing stationary pedestrians and moving vehicles, motion information becomes more informative than appearance cues. The attention-based fusion module of SFIN dynamically adjusts feature importance by selectively emphasizing relevant temporal and spatial information. Consequently, motion information associated with parked vehicles can be suppressed, while motion cues from approaching or maneuvering vehicles can be amplified. This adaptive behavior produces more discriminative feature representations, leading to lower identity-switching rates and higher overall tracking accuracy. As a result, the model remains effective even when either appearance or motion information becomes unreliable.

Another important factor contributing to the practicality of SFIN is the use of a binarized 3D convolutional neural network within the temporal stream. Conventional 3D CNNs are known to be computationally demanding, as reflected by the relatively low processing speed of SlowFast (18.14 FPS). Such computational requirements restrict their deployment in real-time embedded systems commonly used in autonomous vehicles. By employing network binarization, the proposed architecture significantly reduces computational complexity while maintaining sufficient temporal feature quality. As a result, SFIN achieves real-time performance of 35 FPS without substantial degradation in detection and tracking accuracy. The combination of a high-quality spatial stream, an efficient temporal stream, and an adaptive attention-based fusion mechanism enables SFIN to deliver superior overall scene understanding. This synergistic architecture allows the model to maintain reliable performance under complex and challenging conditions, as demonstrated by the experimental results obtained on the DUSP-455 benchmark.

# 6 CONCLUSIONS

This paper introduced the Spatiotemporal Feature Integration Network (SFIN), a novel deep learning architecture for real-time object detection and tracking in autonomous vision systems. The proposed framework addresses the limitations of conventional single-frame detectors by effectively integrating spatial appearance information with temporal motion cues across consecutive video frames. The architecture employs a dual-stream design, where spatial features are extracted using a 2D convolutional neural network and temporal motion features are obtained through a lightweight binarized 3D convolutional neural network operating on optical flow representations. The key contribution of this work is an attention-based fusion module that dynamically balances the influence of spatial and temporal information according to the contextual characteristics of the scene.

Comprehensive experiments conducted on the challenging DUSP-455 benchmark demonstrate the effectiveness of the proposed approach. SFIN achieved a mean Average Precision (mAP) of 89.2% and a Multiple Object Tracking Accuracy (MOTA) of 85.4%, outperforming both conventional spatial detectors and existing spatiotemporal methods. Furthermore, the model maintained real-time performance at 35 FPS, confirming its suitability for practical deployment in autonomous systems. The experimental findings validate the hypothesis that adaptive attention-based fusion provides more effective feature integration than conventional static or late-fusion strategies. By generating a richer and more context-aware representation of dynamic environments, SFIN enhances perception accuracy and contributes to safer and more reliable autonomous navigation in real-world conditions.

# 7 LIMITATIONS

Even with its impressive performance, the SFIN model also has some limitations. The quality of the temporal stream primarily depends on the baseline optical flow calculation. During low light at the start of incompressible shape change or in adverse weather, optical flow computation may be affected by noisy or ill-conditioned vectors. They will be

scattered along the temporal stream and have a debilitating effect on the network's overall operation. While an attention mechanism can learn to disregard an unhealthy temporal signal, it does catch an overwhelmingly good motion input. Second, it was separately trained and tested on the DUSP-455 driving dataset to mimic an urban environment. While large, its ability to be deployed in a range of environments, such as flying drones or off-road driving, has yet to be demonstrated. Motion appearance and style within such environments can necessitate significant modifications, and architectural modifications or tuning might be required. Lastly, while a binarised 3D CNN is extremely powerful, data is lost in a network of regular accuracy. This could limit the model's ability to distinguish extremely low-quality motion indications, which, under certain edge-case conditions, would be advantageous. Computational speed, if tolerable for real-time use, can perhaps be achieved at the cost of ultimate theoretical efficiency.

# 8 FUTURE WORK

SFIN's performance leaves ample scope for future research and development. In the short term, the system can be improved by adding information from other senses. By merging SFIN and LiDAR vision data with radar range and speed data, the system can be elevated to a higher, richer perception system under poor weather conditions with degraded camera performance. Multi-modal fusion could be enabled by improving the attention mechanism. Another promising direction is the self-supervised pre-training of the temporal stream. For large quantities of unlabelled video data, pre-learnt canonical motion patterns at the initial stage can be trained offline before fine-tuning on the labelled set, thereby enhancing generalisation and minimising reliance on high-quality labelled sets. Network optimisation and hardware acceleration also have immense potential. Techniques such as network pruning and quantisation are also employed to reduce the computational expense of the model, making it deployable even on more resource-constrained edge devices. Lastly, the building blocks in their raw form of SFIN would be leveraged to transition from 2D object detection to end-to-end 3D scene understanding, where even the 3D location, scale, and orientation of objects in the scene would be predicted -the holy grail for autonomous systems.

# REFERENCES


[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," IEEE Trans. Robot., vol. 37, no. 6, pp. 1874-1890, 2021.

[2] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A stereo event camera dataset for driving scenarios," IEEE Robot. Autom. Lett., vol. 6, no. 3, pp. 4947-4954, 2021.

[3] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," IEEE Robot. Autom. Lett., vol. 3, no. 4, pp. 4076-4083, 2018.

[4] V. S. A. Anala and S. Chintapalli, "Scalable data partitioning strategies for efficient query optimization in cloud data warehouses," FMDB Transactions on Sustainable Computer Letters, vol. 2, no. 4, pp. 195-206, 2024.

[5] T. Schöps, T. Sattler, and M. Pollefeys, "BAD SLAM: Bundle Adjusted Direct RGB-D SLAM," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, California, United States of America, 2019.

[6] S. Jeyabal, W. Sachinthana, S. Bhagya, P. Samarakoon, M. R. Elara, and B.-J. Sheu, "Hard-to-Detect Obstacle Mapping by Fusing LiDAR and Depth Camera," IEEE Sens. J., vol. 24, no. 15, pp. 24690-24698, 2024.

[7] K. Yu, T. Tao, H. Xie, Z. Lin, T. Liang, B. Wang, P. Chen, D. Hao, Y. Wang, and X. Liang, "Benchmarking the Robustness of LiDAR-Camera Fusion for 3D Object Detection," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023.

[8] K. Anitha, B. K. Nagaraj, P. Paramasivan, and T. Shynu, "Enhancing clustering performance with the rough set C-means algorithm," FMDB Transactions on Sustainable Computing Systems, vol. 1, no. 4, pp. 190-203, 2023.

[9] S. Klenk, J. Chui, N. Demmel, and D. Cremers, "TUM-VIE: The TUM Stereo Visual-Inertial Event Dataset," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021.

[10] S. Khattak, C. Papachristos, and K. Alexis, "Visual-Thermal Landmarks and Inertial Fusion for Navigation in Degraded Visual Environments," in 2019 IEEE Aerospace Conference, Big Sky, MT, United States of America, 2019.

[11] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al., "Event-based vision: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 154-180, 2020.

[12] M. Ataseveri and B. Kose, "Sustainable transformation in logistics: Environmental innovations and human-centered strategies," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 4, pp. 201-210, 2024.

[13] V. Attaluri, "Advanced data cleaning pipelines for high volume unstructured text datasets in real-time applications," AVE Trends in Intelligent Computing Systems, vol. 1, no. 4, pp. 209-218, 2024.

[14] S. Banala, "The future of site reliability: Integrating generative AI into SRE practices," FMDB Transactions on Sustainable Computer Letters, vol. 2, no. 1, pp. 14-25, 2024.

[15] D. Femi, C. Satheesh, M. Sakthivanitha, R. Maruthi, G. Gnanaguru, and M. Paslavskyi, "Sustainable environmental optimization of abrasive water jet machining parameters of AZ31D/$B_4C$ composite using TOPSIS technique," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 3, pp. 138-148, 2024.

[16] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping," in 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020.

[17] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, Tennessee, United States of America, 2021.

[18] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, United States of America, 2018.

[19] M. S. Islam, R. Islam, and M. S. H. Ridoy, "Investigating the design and performance optimization of spindle boxes in modular machine tools," FMDB Transactions on Sustainable Energy Sequence, vol. 2, no. 2, pp. 60-87, 2024.

[20] F. J. John Joseph, K. Chinnusamy, J. Jeganathan, A. J. Obaid, and S. S. Rajest, Eds., Machine Learning, Predictive Analytics, and Optimization in Complex Systems, IGI Global, USA, 2025, [Online]. Available: https://doi.org/10.4018/979-8-3373-5203-9.

[21] S. Kannan, S. Paneerselvam, M. N. Saroja, and M. A. Ahmad, "Real-time air quality prediction and role in environmental protection using machine learning," FMDB Transactions on Sustainable Environmental Sciences, vol. 2, no. 1, pp. 50-59, 2025.

[22] R. S. Madhuranthakam, "Scalable data engineering pipelines for real-time analytics in big data environments," FMDB Transactions on Sustainable Computing Systems, vol. 2, no. 3, pp. 154-166, 2024.

[23] I. O. Mathew, I. Otaraku, A. Oji, and P. N. Ikenyiri, "Addressing methane venting: Strategies and implications on the environment," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 1, pp. 41-56, 2024.

[24] M. S. Minu, S. S. Subashka Ramesh, R. Canessane, M. Al-Amin, and R. B. Sulaiman, "Experimental analysis of UAV networks using oppositional glowworm swarm optimization and deep learning clustering and classification," FMDB Transactions on Sustainable Computing Systems, vol. 1, no. 3, pp. 124-134, 2023.

[25] M. Mokdad, "Environmental strategic dynamics of global energy management and security," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 2, pp. 91-106, 2024.

[26] S. Panyaram, "Integrating artificial intelligence with big data for real-time insights and decision-making in complex systems," FMDB Transactions on Sustainable Intelligent Networks, vol. 1, no. 2, pp. 85-95, 2024.

[27] S. Panyaram, "Optimization strategies for efficient charging station deployment in urban and rural networks," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 2, pp. 69-80, 2024.

[28] P. Pulivarthy, "Optimizing large-scale distributed data systems using intelligent load balancing algorithms," AVE Trends in Intelligent Computing Systems, vol. 1, no. 4, pp. 219-230, 2024.

[29] A. Raj, A. K. Gupta, R. K. Pachauri, and V. Sharma, "Performance evaluation of Walrus optimization-based MPPT with other algorithms under partial shading condition," FMDB Transactions on Sustainable Energy Sequence, vol. 2, no. 2, pp. 88-101, 2024.

[30] S. S. Rajest, S. Moccia, B. Singh, R. Regin, and J. Jeganathan, Eds., Advancing Intelligent Networks Through Distributed Optimization, Advances in Computer and Electrical Engineering, IGI Global, USA, Aug. 2024, [Online]. Available: https://doi.org/10.4018/979-8-3693-3739-4.

[31] J. I. Ramos, R. Lacerona, and J. M. Nunag, "A study on operational excellence, work environment factors and the impact to employee performance," FMDB Transactions on Sustainable Social Sciences Letters, vol. 1, no. 1, pp. 12-25, 2023.

[32] P. Reena, G. Gnanaguru, S. B. V. J. Sara, D. Suresh, and H. K. Ibrahim, "Rheological and environmental performance of drilling muds enhanced with organic ash," FMDB Transactions on Sustainable Applied Sciences, vol. 1, no. 2, pp. 99-112, 2024.

[33] Y. M. Sabti, R. I. N. Alqatrani, M. I. Zaid, B. Taengkliang, and J. M. Kareem, "Impact of business environment on the performance of employees in the public-listed companies," FMDB Transactions on Sustainable Management Letters, vol. 1, no. 2, pp. 56-65, 2023.

[34] O. J. Singh, S. Mahapatra, S. R. Bose, A. Szeberényi, and C. C. Angelin, "A new distribution system power quality mitigation method using recursive least squares controlled dynamic voltage restorer," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 4, pp. 191-200, 2024.

[35] A. Thirunagalingam, "Transforming real-time data processing: The impact of AutoML on dynamic data pipelines," FMDB Transactions on Sustainable Intelligent Networks, vol. 1, no. 2, pp. 110-119, 2024.

[36] R. Vani, K. Lalitha, S. B. V. J. Sara, V. Brindha, G. Gnanaguru, and D. Lale, "Real-time air traffic monitoring system with Raspberry Pi-based environmental monitoring," FMDB Transactions on Sustainable Environmental Sciences, vol. 1, no. 4, pp. 211-220, 2024.

[37] E. Zanardo, "Chronostamp: A general-purpose run-time for data-flow computing in a distributed environment," AVE Trends in Intelligent Computing Systems, vol. 1, no. 2, pp. 106-115, 2024.

[38] A. S. Abdulbaqi, A. J. Obaid, and M. H. Abdulameer, "Smartphone-based ECG signals encryption for transmission and analyzing via IoMTs," Journal of Discrete Mathematical Sciences and Cryptography, vol. 24, no. 7, pp. 1979-1988, 2021.

[39] S. R. Bose, J. A. Jeba, O. J. Singh, R. Regin, S. S. Rajest, and G. M. A. Sagayee, "Machine learning-based real-time stampede and crowd risk prediction," AVE Trends in Intelligent Computer Letters, vol. 2, no. 1, pp. 53-66, 2026.

[40] T. Kanimozhi and S. B. V. J. Sara, "Oppositional pufferfish optimization algorithm-based cluster head selection and congestion trust-aware energy-efficient clustering routing protocol in IoT-WSN," FMDB Transactions on Sustainable Intelligent Networks, vol. 3, no. 1, pp. 26-44, 2026.

[41] L. Y. Reddy, M. N. S. Sumanth, R. Regin, S. R. Bose, S. S. Jeev, and S. B. Sherine, "SECUREPATCH: Specialized multi-agent architecture with automated validation for security vulnerability repair," AVE Trends in Intelligent Computing Systems, vol. 3, no. 1, pp. 23-47, 2026.