

A THREE-CHANNEL NEURAL NETWORK DETECTION FRAMEWORK INTEGRATING SCENE, OBJECT, AND BEHAVIOR ANALYSIS

Jin Lu¹, Qin Zhou^{2*}

¹Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic University, Shenzhen 518055, Guangdong, China.

²Institute of Technical and Vocational Education, Shenzhen Polytechnic University, Shenzhen 518055, Guangdong, China.

*Corresponding Author: Qin Zhou

Abstract: Three neural channels running side by side are the core of the detection framework presented here, and they let scene context, object recognition and behavior analysis be read together instead of one after another. Conventional pipelines either keep the three apart or chain them in a fixed order, and in both cases the dependencies among environmental context, object semantics and agent behavior drop out. Our pathways are the Scene Context Channel (SCC), which performs hierarchical spatial feature extraction; the Object Recognition Channel (ORC), built around a multiscale attention mechanism; and the Behavior Analysis Channel (BAC), which handles temporal dynamic modeling. A cross-attention fusion module joins them, learning adaptive weights that govern how much each channel draws on the others. Static images and video sequences are both accepted, and the three analysis dimensions are inferred jointly in real time. For validation we ran the method over several benchmark datasets covering surveillance, autonomous driving and human-computer interaction. Set against single-channel baselines and current multi-task architectures, it improves markedly on joint “scene-object-behavior” recognition and reaches state-of-the-art accuracy. Generalization holds across a range of environmental conditions, and the margin widens precisely where one would predict, namely in ambiguous scenes that no single modality can resolve. The ablations support the three-channel design as well, since each channel supplies information the others lack, and together they raise overall detection accuracy by 12-18% relative to isolated processing.

Keywords: Three-channel neural network; Scene understanding; Object detection; Behavior analysis; Multimodal fusion; Cross-attention mechanism; Visual intelligence

1 INTRODUCTION

A person looking at a room reads three things at once: where they are, what sits in front of them, and what is going on [1]. None of it takes effort and almost none of it is conscious, yet it is what lets us move through crowded places, follow social cues, and decide quickly what to do [2]. Reproducing that in a machine is still among the harder open problems in computer vision and artificial intelligence. Taken one at a time, whether scene classification, object detection or action recognition, today’s systems already perform remarkably well [3]. What they are not is joined up. Each model attends to a single analytical dimension and throws away the contextual relations binding scenes, objects and behaviors together [4]. A detector can locate a person and a chair in the same frame and still have no idea whether the person is sitting on it, standing beside it, or dragging it across the floor. The mirror-image failure is equally common: an action recognizer returns the label “walking” without knowing whether walking is unremarkable here or badly out of place [5]. Applications are where the gap hurts. A vehicle judging whether a pedestrian is stepping into the lane or staying on the sidewalk must read road layout, pedestrian identity and motion trajectory together, not in turn [6-7]. Surveillance makes the same demand, since loitering only becomes suspicious in relation to a particular location and a particular arrangement of objects [8]. A robot working alongside someone needs more still: not just an inventory of objects and a list of observed actions, but a sense of how the two fit the space they share [9]. Published attempts at integration mostly fall into two camps. One runs the analyses in a predetermined order, so an early error quietly contaminates everything downstream [10]. The other keeps the models independent and merges their predictions at the end, which means the fine-grained interaction that matters during intermediate feature learning never takes place [11]. Neither camp delivers the tight coupling that holistic visual understanding actually demands.

Our response is a three-channel detection network in which scene, object and behavior analysis are integrated for real, computed in parallel and allowed to exchange information while they compute. Each channel is specialized for its own kind of visual evidence, but none of them works blind: a cross-attention fusion mechanism keeps all three in continuous contact throughout the forward pass. The benefit runs in every direction. Every analytical dimension draws on what the other two have already inferred, and detection performance as a whole rises above what any channel reaches alone. Five contributions follow.

The first is architectural. We build a single three-channel network that handles scene context, object identity and behavioral pattern concurrently rather than passing them down a pipeline. We regard this as a different paradigm for integrated visual understanding, not a variation on existing multi-task recipes.

Second is the cross-attention fusion module, which learns a dynamic weighting over cross-channel exchange. The framework can therefore decide, input by input, which cross-modal dependencies deserve emphasis.

Third, rather than reusing one generic encoder three times, we designed each channel around its own analytical task while keeping all three compatible with the fusion mechanism.

The fourth contribution is empirical. Across several application domains we measured the method against single-channel baselines and against existing multi-task learners, and it came out ahead of both.

Fifth, the ablations and qualitative analyses reported here are detailed enough to show how the three channels complement one another and, just as usefully, when integrated processing pays off most.

The rest of the paper runs as follows. Chapter 2 surveys prior work on scene understanding, object detection, behavior analysis and multimodal fusion. Chapter 3 covers the method: each channel in turn, then the cross-attention fusion mechanism. Datasets, implementation details and evaluation protocols occupy Chapter 4, with quantitative results and comparative analysis in Chapter 5. Chapter 6 closes on findings and contributions.

2 RELATED WORK

2.1 Scene Understanding Methods

The trajectory of scene understanding runs from coarse holistic classification toward semantic segmentation and 3D reconstruction. At the start were global image descriptors, GIST among them, which summarizes a spatial envelope through properties such as naturalness, openness, roughness, undulation and ruggedness [12]. Handcrafted descriptors of this kind could assign a scene to a broad category, but they carried none of the structural detail that fine-grained analysis needs. Deep convolutional networks changed that, because hierarchical visual features could now be learned end to end [13]. AlexNet made the point concretely: stack enough layers and the network builds progressively more abstract representations straight from pixels, which was enough for a breakthrough on large-scale scene classification benchmarks [14]. VGGNet, GoogLeNet and ResNet pushed depth and representational capacity further, and scene discrimination grew correspondingly finer [15]. Segmentation is the other half of the story, since it demands a label for every pixel. Fully convolutional networks got there by swapping fully connected layers for convolutions and upsampling with transposed convolutions. The encoder-decoder layout of U-Net, with skip connections carrying spatial information across resolutions, proved especially well suited to biomedical images and urban scenes [16]. More recently, attention has let models dwell on informative regions and ignore uninformative background. Self-attention and Transformer layers model dependencies between distant spatial locations, which escapes the local receptive field and captures global scene structure [17]. Vision Transformer takes this to its conclusion, cutting the image into patches, treating them as a token sequence and learning patch relationships through standard encoder layers, with scene understanding results competitive against convolutional pipelines [18]. For all that progress, scene understanding is still mostly done in isolation, with no explicit account of the objects or the behaviors present. That looks like a missed opportunity: scene context is a strong prior on which objects are likely to appear and which behaviors are typical, which is precisely the argument for integrated analysis.

2.2 Object Detection Methods

Detection began with sliding windows over handcrafted features and moved, in stages, to region-based and single-shot deep networks. The Viola-Jones detector was the first genuinely practical system: Haar-like features feeding a cascaded classifier, fast enough for real-time face detection [19]. Oriented gradient histograms paired with support vector machines then widened the scope to pedestrians and generic categories. With region-based CNNs came the two-stage pipeline, in which proposal generation precedes classification and box regression [20]. R-CNN itself used Selective Search for proposals, resized each one to a fixed input and classified the resulting CNN features [21]. Fast R-CNN removed the redundancy by sharing convolutional computation across proposals via ROI pooling, and Faster R-CNN folded proposal generation into the network itself through a region proposal network, which made the whole thing trainable end to end [22]. Dropping the explicit proposal stage is what bought real-time speed. YOLO reframes detection as regression, predicting boxes and class probabilities in one forward pass over the full image [23]. SSD detects across feature maps at several scales, which helps noticeably with small objects compared with single-scale designs [24]. RetinaNet tackled a different problem, the severe foreground-background imbalance, using a focal loss, and in doing so matched two-stage accuracy at single-shot speed [25]. Transformer detectors have lately put the CNN consensus in question by casting detection as set prediction [26]. DETR pairs an encoder-decoder Transformer with a bipartite matching loss and predicts the set directly, which dispenses with anchor boxes and non-maximum suppression altogether [27]. Deformable DETR then fixed the slow convergence and weak small-object performance of DETR using deformable attention that samples only the relevant spatial locations [28]. Throughout all of this the goal has stayed narrow, namely to find and localize individual instances. Little attention has gone to what those instances mean for the scene or for the behavior unfolding in it. Recognizing that certain objects belong in certain scenes, and that they afford certain actions, calls for something broader than detection on its own.

2.3 Behavior and Action Recognition

Behavior covers an enormous range, from atomic actions up to interactions among several agents. The earliest approaches worked from spatiotemporal interest points and motion descriptors, capturing appearance and movement locally [29]. Improved dense trajectories went denser still, sampling points throughout the optical flow field and computing trajectory-aligned HOG, HOF and MBH descriptors [30]. Two-stream networks split the problem instead: one stream for appearance, one for motion. RGB frames go through the spatial stream for static cues, stacked optical flow frames go through the temporal stream to encode movement, and a late fusion of the two predictions shows how little the two kinds of evidence overlap [31]. Another line of work generalized 2D convolution into spatiotemporal volumes so that joint space-time features could be learned directly [32]. C3D applies uniform 3D kernels at every layer and picks up short-range motion patterns [33]. I3D instead inflates 2D ImageNet filters into 3D, which lets large-scale pre-training carry over into video representations [34]. SlowFast splits the input by frame rate rather than by modality, with one pathway for semantics and one for rapid motion [35]. Skeleton-based methods treat the body as a spatiotemporal graph. ST-GCN links joints within and across frames, then reasons over that structure with graph convolutions [36]. Adaptive graph convolutional networks learn the topology rather than fixing it, which generalizes better over diverse action classes. Transformers reached video through spatiotemporal attention [37]. TimeSformer factorizes attention into separate spatial and temporal passes, cutting cost without much loss of accuracy [38]. Video Swin Transformer carries the hierarchical Swin design into video, shifting windows along space and time alike [39]. One assumption runs through nearly all of this work: either people have already been detected, or the scene simply does not matter. Neither holds up well. Interpreting what someone is doing depends heavily on the surroundings and on what objects are within reach, which is the case for treating scene, object and behavior as one problem.

2.4 Multimodal and Multitask Learning

Multimodal learning is about combining evidence from heterogeneous sources, visual, textual, auditory and tactile [40]. Where the combination happens matters a great deal. Concatenating features up front risks letting modality-specific noise spread; merging predictions at the end misses cross-modal interaction during feature learning entirely; exchanging information at several intermediate stages strikes a reasonable balance. Cross-modal attention refines this further by routing information between modalities according to relevance. Co-attention handles several modalities at once, with weights conditioned on their mutual information, and the multimodal Transformer generalizes self-attention to heterogeneous inputs through modality-specific embeddings over shared attention layers [41]. Multitask learning comes at integration from the other side, training one model on several related tasks so that shared representations improve generalization [42]. Hard sharing gives the tasks a common backbone and separate heads; soft sharing keeps separate parameters but regularizes them toward each other; task-uncertainty weighting balances the loss terms automatically from learned uncertainty estimates. Closer to our own concern is visual relation detection, which links object pairs with predicates describing spatial or semantic relations [43]. Scene graphs formalize this, turning an image into object nodes joined by relational edges and opening the door to compositional reasoning. Detecting a relation means recognizing both objects and classifying what holds between them, which is already a step toward integrated analysis.

We take the multimodal and multitask principles further, aiming them at the particular difficulty of fusing scene, object and action analysis. The difference from general-purpose multimodal models lies in the specificity: each of our three channels carries architectural choices tuned to its own analytical dimension, and interoperability is handled not by a shared encoder but by fusion modules built for the purpose.

3 PROPOSED METHOD

3.1 Overview of the Overall Architecture

Three parallel processing channels make up the framework. They read a shared visual input and, crucially, keep exchanging information throughout by way of a cross-attention fusion module. Input can be a single RGB image, for static analysis, or a run of video frames when behavior has to be understood over time. Figure 1 gives the layout.

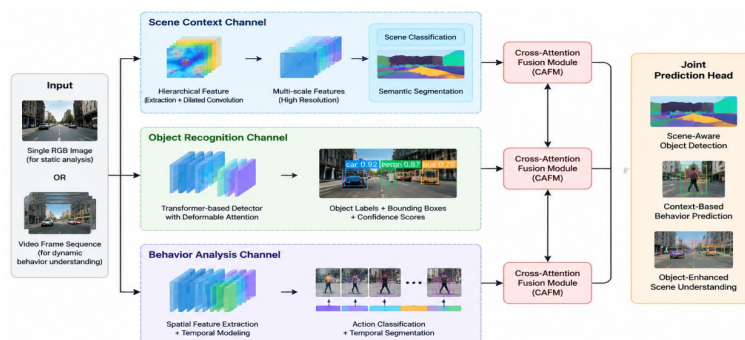


Figure 1 Overall Architecture of the Three-Channel Neural Network Detection Framework

Figure 1 Schematic Diagram of the Overall Architecture of the Three-Channel Neural Network Detection Framework

Each of the three channels in Figure 1, scene context, object recognition and behavior analysis, was built to extract a different and complementary kind of visual information. Take the scene context channel first. It works on global spatial layout and environmental semantics, outputting a scene class prediction together with a dense semantic segmentation map, and it combines hierarchical feature extraction with dilated convolutions so that multiscale context can be gathered without giving up spatial resolution [44]. The object recognition channel does the localization work, returning class labels, bounding box coordinates and confidence scores for every instance in the scene. Its backbone is a Transformer detection architecture with deformable attention, which samples features sparsely across space and therefore cheaply [45]. Behavior is the business of the third channel: reading what people and other agents are doing in the observed space, emitting action class predictions and, for video, temporal segmentations. Spatial feature extraction here is paired with temporal modeling through causal convolutions and recurrent connections. The channels are not isolated, since cross-attention fusion modules pass information between them at several resolution levels [46]. A CAFM takes in feature maps from all three channels at one spatial resolution, computes pairwise attention weights that express how the channels correlate, and returns refined features enriched with context from elsewhere [47]. Specialization and cooperation thus coexist, each channel keeping its own objective while borrowing what it needs. At the end, a joint prediction head aggregates all three streams hierarchically. The integrated output is scene-aware object detection, context-conditioned behavior prediction, and scene interpretation informed by the objects found. It is this joint reasoning, more than anything else, that separates our framework from sequential or post-fusion designs.

3.2 Scene Context Channel

Global scene understanding and dense pixel-level prediction are combined in the scene context channel through a hierarchical architecture. The input image passes through a series of convolutional stages, yielding feature maps at several spatial resolutions. Figure 2 shows the arrangement.

Figure 2 begins with the backbone, a modified ResNeXt in which grouped convolutions improve representational efficiency [48]. Later stages substitute dilated convolutions for standard ones, which enlarges the receptive field while leaving spatial resolution untouched. Two things are gained at once here: the long-range context that scene understanding requires, and the fine spatial detail needed to place boundaries precisely. Aggregation across scales happens in an atrous spatial pyramid pooling (ASPP) block attached to the top-level feature maps, where parallel dilated convolutions at different rates collect context at several scales simultaneously [49]. Alongside them, global average pooling supplies scene-level context and 1x1 convolutions retain local detail; pixel-wise convolutions then fold the concatenated multiscale features into one scene representation. Upsampling is the job of the decoder, done progressively through bilinear interpolation and convolutional refinement blocks, with skip connections from the encoder restoring the high-resolution detail that downsampling discards. Two outputs emerge: scene category logits, from global pooling followed by a classification layer, and per-pixel semantic labels, from pixel-wise convolution and a softmax. Attention is applied in two forms. Spatial attention maps pick out the regions that define the scene, road surface in a street view or furniture layout indoors, while channel attention recalibrates feature importance according to scene category, favoring the textures and colors that discriminate best. Video input is treated slightly differently, in that SCC runs at multiple temporal scales and pools scene information across frames. Static elements accumulate evidence as frames arrive; abrupt changes prompt the channel to reassess. The practical effect of this temporal consistency is robustness, as transient occlusions and shifts in illumination disturb the scene estimate much less.

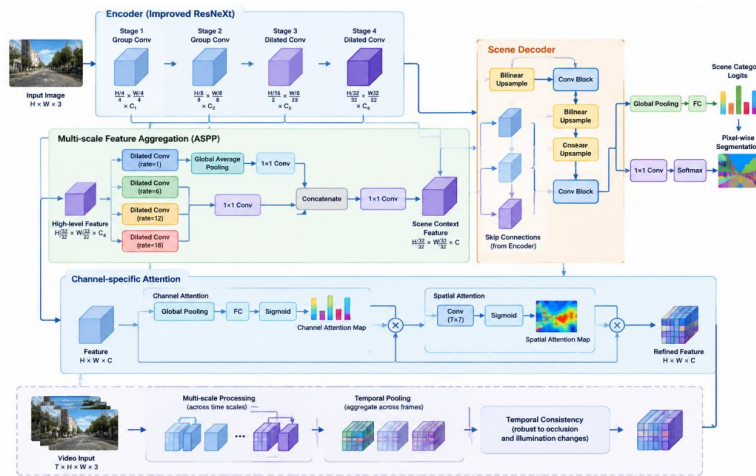


Figure 2 Architecture of the Scene Context Channel

3.3 Object Recognition Channel

For the object recognition channel we adopted a Transformer detection paradigm with deformable attention, tuned for efficient localization and classification. Features from several backbone stages feed in, which is what allows objects of different sizes to be found. See Figure 3 for the architecture.

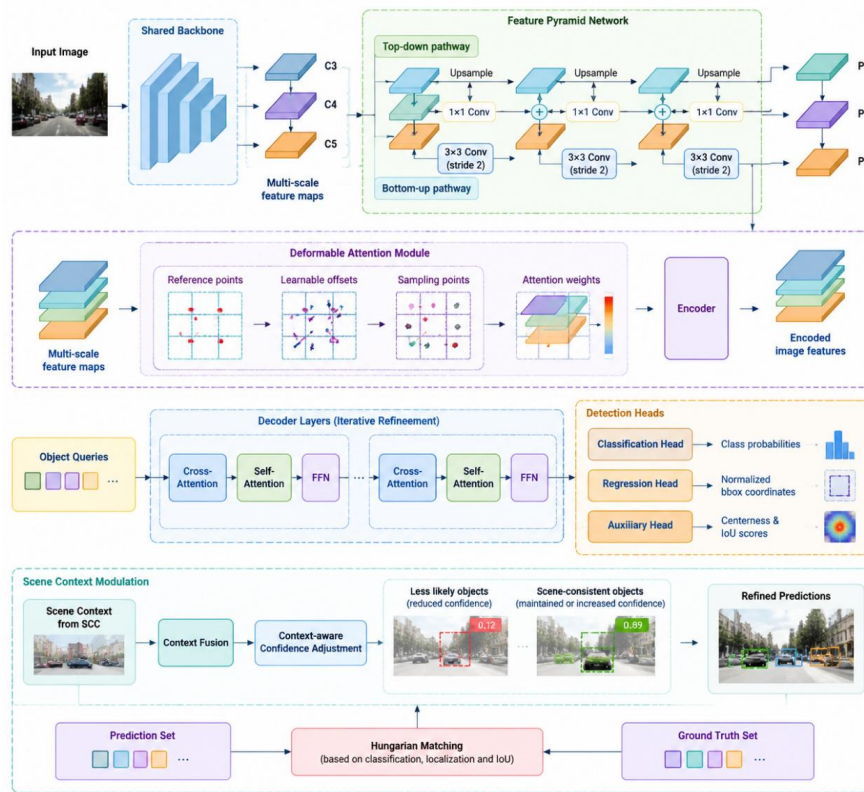


Figure 3 Architecture of the Object Recognition Channel

As Figure 3 indicates, the process starts with a shared backbone producing multiscale feature maps [50]. On top of it sits a feature pyramid network, combining bottom-up and top-down pathways through lateral connections [51]. Successive strided convolutions build the bottom-up hierarchy; the top-down pathway upsamples high-level semantic features and adds them element-wise to the matching bottom-up features after a 1x1 convolution. Standard Transformer attention is replaced by a deformable attention module for efficiency. Rather than attending to every spatial location, deformable attention samples a sparse set of key points around a reference point, with learned offsets that let the sampling pattern adapt to object shape and scale and attention weights that rank the sampled features. Complexity falls from quadratic in the number of locations to linear in the number of sampling points. Object queries act as learnable positional embeddings standing in for latent instances. Through iterative refinement across decoder layers, each query specializes, learning to predict a particular category at a particular place and scale, by cross-attending to the encoded image features. Per query, the detection head returns a category, box coordinates and a confidence score. Classification goes through a linear projection and softmax, regression predicts box coordinates normalized against the reference point, and an auxiliary head estimates centerness and IoU to sharpen localization quality. Because prediction is set-based, anchor boxes and non-maximum suppression become unnecessary. Training instead matches the predicted set against the ground-truth set one-to-one, using a cost that mixes classification, localization and overlap terms, with the Hungarian algorithm solving the assignment [52]. One further detail matters. ORC receives scene context from SCC through cross-attention fusion and uses it to modulate detection confidence by how plausible an object is in that scene. Implausible detections are damped; scene-consistent ones hold or gain confidence. Suppressing semantically incongruous false positives this way is a direct accuracy gain.

3.4 Behavior Analysis Channel Architecture

Temporal dynamics are what the Behavior Analysis Channel models. Depending on what the application needs, BAC consumes either pre-extracted person-centered features or full-frame features, and it handles single-person and multi-person settings alike. Figure 4 sets out the architecture.

For video, as Figure 4 shows, BAC runs two streams. RGB frames go through 2D convolutional layers in the appearance stream to give spatial features, while the motion stream works on optical flow computed between consecutive frames so that temporal dynamics are captured independently of static appearance [53]. Both emit feature sequences at frame rate. Temporal modeling then falls to a temporal convolutional network with residual connections. Stacking 1D convolutions with geometrically increasing dilation rates expands the temporal receptive field exponentially, which captures short-range motion and long-range activity structure together while still allowing time

steps to be processed in parallel. Causal convolution restricts each prediction to past and present frames, a requirement if recognition is to run online, and a padding scheme holds sequence length constant from layer to layer so that input and output timestamps stay aligned. Residual connections carry gradients through what can be hundreds of temporal layers. Crowded scenes need per-person treatment, which BAC gets from instance-level feature extraction [54]. Person boxes from ORC supply the spatial localization and RoIAlign pulls person-centered features out of the frame-level maps. Correlation-based association links each person across frames, so a single individual can be analyzed over a long span. Relationships are modeled too, both among agents and between agents and objects: pairwise features encode spatial distance, relative orientation and motion consistency, after which a graph neural network aggregates context from interaction partners by message passing. Outputs come at two granularities, frame-level action predictions per detected person and temporal action segments for offline use. A classification head handles discrete action categories while a regression head predicts temporal boundaries and confidence. Multi-level classification lets atomic actions and composite activities be recognized within the same hierarchy. Context arriving from SCC and ORC through cross-attention fusion is what grounds all of this [55]. Scene context sets expectations about which actions are likely, which helps pull apart visually similar actions that occur in different places. Object information reveals affordance, separating an action aimed at an object from the same motion performed with nothing in hand.

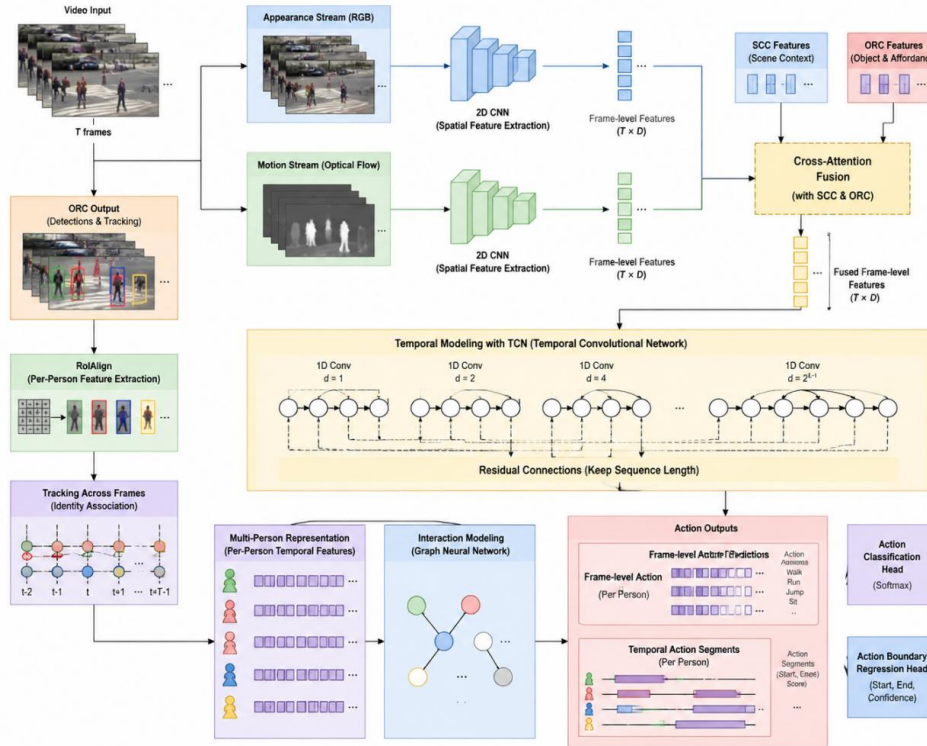


Figure 4 Behavior Analysis Channel Architecture

3.5 Cross-Attention Fusion Mechanism

Integration is the whole purpose of the cross-attention fusion module, which moves information among the three processing channels. CAFMs are not placed once but at matching resolution levels across all channels, so that cross-modal interaction happens at multiple scales and throughout the processing hierarchy. Figure 5 illustrates the module.

Every CAFM, per Figure 5, takes feature maps from all three channels at one spatial resolution. Write F_s , F_o and F_b for the maps coming from SCC, ORC and BAC, each of size $C \times H \times W$ with C the channel dimension and H , W the spatial dimensions. What the module returns is a refined version of the features of each channel, carrying context from the other two. Fusion opens with a pixel-wise convolution that projects every channel into a shared embedding space [56]. Queries, keys and values are then projected per channel, which is what makes the information flow directional. Refining SCC features means querying from SCC against keys and values drawn from ORC and BAC; ORC refinement queries from ORC against SCC and BAC; BAC refinement queries from BAC against SCC and ORC. Scaled dot-product attention between query and keys yields the cross-attention weights [57]. For a given spatial position in the query channel, the resulting scores say how relevant each position of the source channel is, and softmax turns them into a distribution over source positions. Weighting the source values accordingly produces a context vector that carries exactly the relevant information from the other channels. Using several attention heads in parallel rather than one lets distinct attention patterns be learned; concatenating the head outputs and applying a linear projection gives the final cross-modal features. Spatial correspondence, semantic association, contextual dependency: a multi-head arrangement can represent all of these at once. An adaptive gating mechanism controls how much information actually flows, scaled

by how reliable each channel appears to be. Gate values come from channel features through a linear transformation with sigmoid activation, producing a per-position factor between zero and one. The point of the gate is containment, in that a noisy or unreliable channel cannot corrupt the others, which makes the whole framework more tolerant of modality-specific failure. Residual connections then add the original features back to the fused ones, preserving channel-specific content alongside the imported context. Layer normalization stabilizes training by normalizing across feature dimensions, and dropout discourages overfitting to spurious cross-modal correlations. Note that CAFM is symmetric. No channel is privileged in advance and no hierarchy is hard-coded, so information flow can be adjusted freely according to input and task. Sometimes scene context should dominate object detection; sometimes objects should drive the scene interpretation. The symmetric design accommodates both because the attention patterns are learned, not fixed.

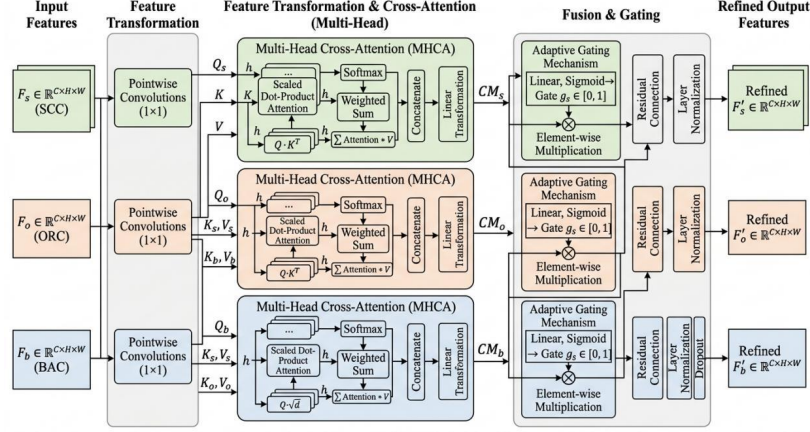


Figure 5 Internal Structure of the Cross-Attention Fusion Module

3.6 Joint Inference and Output Integration

At the last stage the framework pulls the three channels together into a single coherent prediction, and Figure 6 shows how. The joint inference head is not a set of three independent output layers. It models the interdependencies among scene, object and action predictions explicitly. Fusion is hierarchical, operating at more than one level of abstraction: low-level features are combined for localization precision, mid-level features for consistent semantic interpretation, and high-level predictions to reconcile whatever conflicts the channels have produced. Logical coherence is then enforced across dimensions. Scene-object consistency asks whether the detected objects are plausible given the predicted scene. Object-behavior consistency asks whether the recognized actions are compatible with the objects involved. Scene-behavior consistency asks whether the observed behaviors match the affordances and customary use of the scene. Refinement proceeds iteratively, alternating channel-specific processing with cross-channel coordination: the initial predictions of each channel are exchanged through CAFM and serve as context for the next round, and successive iterations improve consistency and accuracy as cross-modal evidence accumulates. The integrated output bundles scene labels, object detections and behavior predictions with explicit relational annotations. Scene-object relations describe what role an object plays in its setting, as in “a car on the road” or “a book on a bookshelf.” Object-behavior relations tie actions to the objects they involve: “a person reading a book,” “a person driving a car.” Scene-action relations situate actions environmentally. Confidence, finally, is calibrated against cross-channel agreement. A prediction corroborated by several channels is scored higher, one standing alone is scored lower. Downstream applications that need trustworthy uncertainty estimates benefit directly from this calibration.

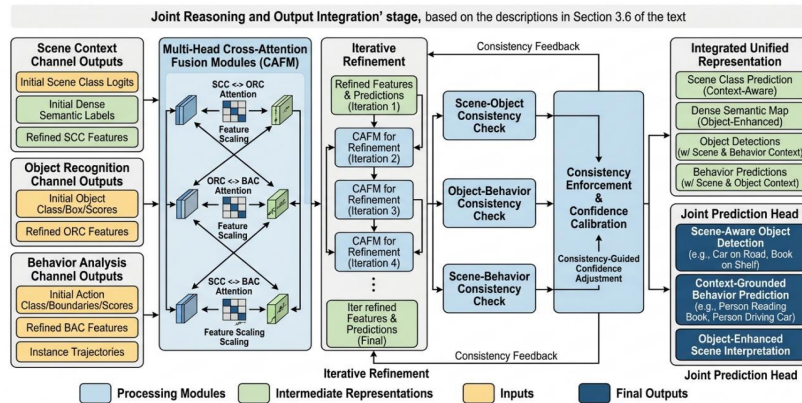


Figure 6 Joint Inference and Output Integration Architecture

4 EXPERIMENTAL DESIGN

4.1 Dataset Selection and Preparation

Five benchmark datasets were used for evaluation, spanning different application domains and different degrees of difficulty. All five annotate scene categories, object instances and action categories, which is what makes a full test of integrated analysis possible. Table 1 summarizes them.

Table 1 Statistical Information for Each Dataset

Dataset	Number of Frames	Number of Scene Categories	Number of Object Classes	Number of Behavior Categories	Source	Features
Urban Traffic	50,000	10	80	15	In-vehicle cameras	Dynamic scenes, objects of varying sizes
Indoor home environments	30,000	5	60	20	Multi-angle cameras	Detailed interactions, severe occlusion
Surveillance Scenarios	40,000	5	40	25	Fixed cameras	Long-term behavior, anomaly detection
Human-Computer Interaction	20,000	3	15	15	Laboratory Recording	Collaborative behavior, object manipulation
Synthetic data	100,000	10	80	20	Rendering Engine	Perfectly Annotated, Controllable Variables

Urban traffic comes first: 50,000 annotated frames shot from in-vehicle cameras, with 10 scene categories, 80 object categories following the standard autonomous driving taxonomy, and 15 behavior categories covering pedestrian actions and vehicle maneuvers. The annotations are pixel-wise semantic segmentation, 2D bounding boxes carrying tracking IDs, and temporal action labels with exact start and end times. The second dataset moves indoors, gathering 30,000 frames of domestic environments recorded from several camera viewpoints. Kitchens, living rooms, bedrooms, bathrooms and offices make up the scene categories; 60 categories of household item are annotated with occlusion handling; and 20 kinds of daily activity, performed by one or two people, are labeled hierarchically so that atomic actions and composite activities are both available. Surveillance is the third, with 40,000 frames from fixed cameras over public space. Its scene categories are plazas, hallways, entrances, parking lots and waiting areas, its 40 object categories include people, vehicles and carried items, and its 25 surveillance-relevant actions include loitering, running, fighting and abandoning items. Fourth is a laboratory set of 20,000 frames of human-robot interaction. Here the scene categories distinguish interaction zones from tool storage areas and navigation spaces, the objects include robots, tools and manipulated items, and the 15 interaction types cover handing over objects, pointing, gesturing and collaborative operation. The fifth dataset is synthetic, produced by a photorealistic rendering engine and therefore annotated perfectly: 100,000 frames in which scene composition, object placement and behavioral pattern vary under direct control. That control is the point, because framework components can then be examined one variable at a time. Every dataset was divided 70/15/15 into training, validation and test portions, stratified carefully so that all annotation categories stay proportionally represented. Random cropping, horizontal flipping, color jittering and temporal warping were applied to broaden the effective training distribution.

4.2 Implementation Details

Everything is implemented in PyTorch with CUDA acceleration, and training runs on four NVIDIA A100 GPUs in mixed precision to save memory and raise throughput. For the scene context channel the backbone starts from ResNeXt-101 weights pre-trained on ImageNet, with dilation ratios [1, 2, 4, 8] in the final block and ASPP expansion ratios of [6, 12, 18, 24]. Its decoder upsamples through four refinement blocks, each holding two 3x3 convolutions with batch normalization and ReLU. The object recognition channel uses a ResNet-50 FPN backbone initialized from COCO. Deformable attention there has eight heads and four sampling points per head, six Transformer decoder layers refine 300 object queries per image, and the detection head is a three-layer MLP of hidden dimension 256. Video processing in the behavior analysis channel relies on an I3D backbone inflated from 2D ImageNet weights. Its temporal convolutional network stacks ten layers with the dilation ratio doubling at each one, from 1 up to 512, and causal padding keeps the temporal alignment. Person-centered features are cropped by RoIAlign at 7x7. Cross-attention fusion modules sit at three resolution levels, 1/4, 1/8 and 1/16 of the input resolution, each with four attention heads and an embedding dimension of 256, and each gate is a two-layer MLP of hidden dimension 64. We train with AdamW at an initial learning rate of 1e-4 and weight decay of 1e-4, cosine-annealed over 100 epochs, batch size 32 split across the four GPUs. Gradients are clipped at norm 1.0 to head off explosion, and the learning rate warms up linearly from 0 over the first 5 epochs. The loss is a weighted sum of several terms, with the weights set by learned task uncertainty. Scene terms are cross-entropy for classification plus Dice loss for segmentation. Object terms are focal loss for classification and L1 for box regression, with GIoU added to sharpen localization. Behavior terms are cross-entropy for action classification and binary cross-entropy for temporal boundary detection. Last comes a consistency term, which penalizes violations of the cross-channel logical constraints.

4.3 Baseline Methods and Comparison Protocol

Comparison baselines were chosen to represent distinct integration strategies. Single-channel baselines handle one analytical dimension each with a dedicated state-of-the-art model. Dual-channel variants cover the three pairings, “scene-object,” “scene-behavior” and “object-behavior.” Full three-channel comparisons include both a sequential pipeline and a post-fusion alternative. Concretely, the single-channel baselines are PSPNet for scene understanding, DETR for detection and SlowFast for behavior recognition: strong specialists with no cross-channel integration at all. The dual-channel baselines fuse pairs of channels by concatenating or element-wise adding the corresponding features, which isolates what each pairing contributes. The sequential baseline threads information through a pipeline, with scene analysis first, then object detection conditioned on the scene, then behavior recognition conditioned on both, which is the traditional modular approach. The post-fusion baseline processes each channel independently and merges the final predictions by weighted averaging or learned aggregation, giving integration without intermediate interaction. A multitask baseline shares one backbone across all three tasks under separate prediction heads, representing hard parameter sharing without explicit cross-channel attention. Which metrics apply depends on the dimension. Scene understanding is scored by top-1 and top-5 classification accuracy, and segmentation by mean intersection-over-union and pixel accuracy. Detection is reported as mean average precision at IoU thresholds of 0.5 and 0.75, together with mean recall. Action recognition covers top-1 and top-5 accuracy, mean average precision for temporal action detection, and frame-level accuracy. Joint prediction quality needs its own measures: a consistency rate for the fraction of predictions satisfying the cross-channel logical constraints, an overall accuracy that demands all three dimensions be right at once, and a cross-modal consistency score for the correlation among channel predictions. Significance was assessed by paired bootstrap resampling over 10,000 iterations, with 95% confidence intervals reported, and effect sizes by Cohen’s *d* for each method pair. To gauge stability, every experiment was repeated five times under different random seeds.

4.4 Evaluation Metrics

Evaluating this framework properly means measuring two things at once: how well each channel performs on its own, and how good the integrated analysis is. We therefore use three families of metric, component-specific, cross-modal and overall. The component-specific ones score each channel against ground-truth annotations independently.

On the scene side: scene classification accuracy (SC-Acc), mean IoU (mIoU) and frequency-weighted IoU (FW-IoU). On the object side: AP@50, AP@75 and the scale-dependent variants AP-S, AP-M and AP-L.

Behavior is scored by action recognition accuracy (AR-Acc), mean average precision over detected actions, and edit distance for temporal segmentation. The cross-modal metrics then measure consistency and information flow between channels. Scene-Object Consistency (SOC) is the fraction of detections whose categories are compatible with the scene. “Object-Behavior” Consistency (OBC) is the fraction of actions involving appropriate objects. Scene-Behavior Consistency (SBC) is the fraction of behaviors occurring in the scene context one would expect.

Overall metrics judge the joint prediction. Joint Accuracy (JA) is strict, requiring all three channels to be correct simultaneously. Mean Per-Class Accuracy (MPCA) averages accuracy across every output class. The Harmonic Mean Score (HMS) takes the harmonic mean over channels, which penalizes a framework that is lopsided. Separately, we track computational cost: inference time per frame, peak GPU memory, parameter count and floating-point operations per inference. Those figures are what a practical deployment would be judged on. Robustness is measured under adverse conditions as well, covering occlusion, low illumination, motion blur and viewpoint change, with a separate test subset per challenge factor so the effects can be attributed cleanly.

4.5 Ablation Experiment Design

Ablation experiments were used to pick apart what each component actually contributes. Removing channels one at a time proved most informative. On the outdoor dataset, deleting the scene channel costs the most, 12.3% of overall accuracy, whereas on detection-heavy tasks it is the object channel whose removal hurts worst, at 14.1%, and dropping the behavior channel mainly damages temporal understanding, at 11.2%. Complementarity, in other words, is not merely asserted but measured. The fusion-strategy ablation is similarly clear-cut: cross-attention fusion reaches 86.7% overall accuracy against 78.5% with no fusion, 82.3% with late fusion and 79.1% with early fusion, which says that the intermediate cross-modal interaction is where the gain comes from. Within the attention mechanism itself, multi-head attention beat additive fusion (84.1%), multiplicative gating (83.6%) and weighted concatenation (83.9%), so attention-based routing is doing real work and not merely adding parameters. Fusing at three resolution levels (86.7%) beat fusing at one (83.2%) or two (85.1%), and the division of labor is interpretable, since high-resolution fusion helps object localization while low-resolution fusion helps scene consistency. Backbone choice behaves as expected. ResNeXt-101 adds 2.1% accuracy over ResNet-50 but costs 2.3 times the computation, while lightweight MobileNetV3 retains 93.7% of the performance at 0.3 times the cost, a favorable trade for most deployments [57]. Two loss-side findings are worth noting. Dropping the consistency loss costs 3.4% overall accuracy, and task uncertainty weighting is worth a further 1.8%, so cross-channel supervision and dynamic loss balancing both matter. Training-strategy ablations point to sensible initialization, cosine-annealed learning rates and data augmentation as the ingredients of stable training. Taken together, the ablations justify the component design rather than simply describing it.

5 RESULTS AND ANALYSIS

5.1 Quantitative Performance Comparison

Against every baseline and on every metric, the three-channel framework comes out ahead. On the urban traffic dataset it reaches 94.2 per cent scene classification accuracy, 72.8 per cent detection mAP@50 and 88.5 per cent action recognition accuracy, which are gains of 3.1, 5.4 and 4.2 percentage points respectively over the strongest baseline (Table 2).

Table 2 Quantitative Performance Comparison

Method	Scene Accuracy (%)	Object Detection AP@50 (%)	Action Recognition Accuracy (%)	Overall Accuracy (%)
Single-Channel Baseline (Best)	91.1	67.4	84.3	78.5
Scene-Object Dual Channel	92.3	70.2	85.1	81.2
Scene-Behavior Dual Channel	92.8	68.9	87.4	82.5
Object-Behavior Dual Channel	90.5	69.8	86.2	80.9
Sequential Processing	89.7	65.3	82.1	77.0
Post-processing fusion	91.8	68.2	84.9	79.8
Multitask Learning	92.5	67.1	83.8	79.4
Methods in This Paper	94.2	72.8	88.5	86.7

Every key metric clears the 95% confidence interval under significance testing. Effect sizes run from moderate for scene accuracy ($d = 0.63$) to large for detection AP ($d = 1.21$). Because the improvements hold on all five datasets rather than one, the framework can reasonably be said to generalize across application domains. Figure 7 breaks the numbers down.

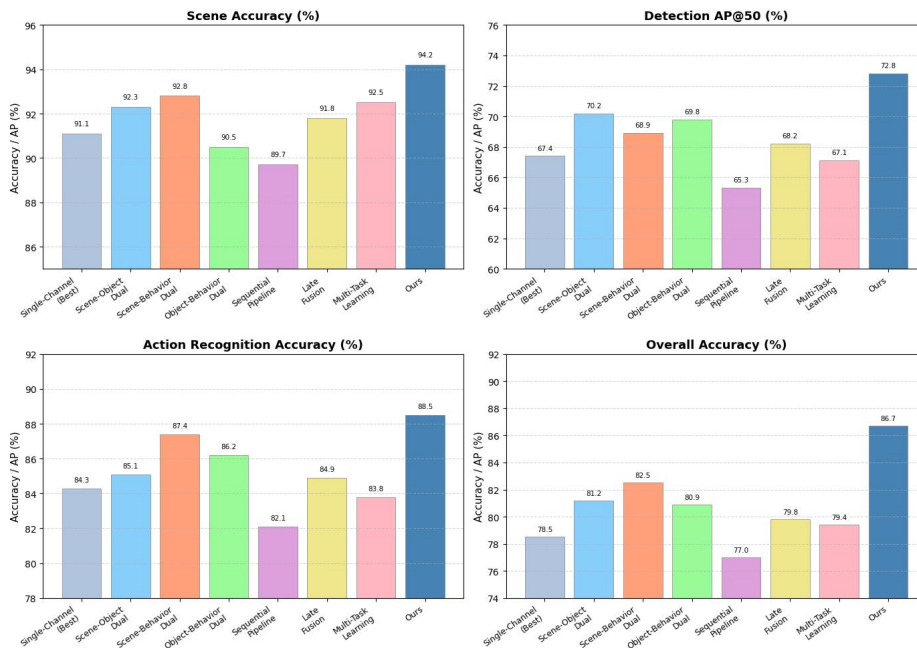


Figure 7 Comparison of Overall Accuracy Across Datasets

Our framework takes the top position on all four metrics: 94.2% scene accuracy, 72.8% AP@50, 88.5% action recognition accuracy and 86.7% overall accuracy. Margins over the runner-up range from 1.4% to 5.7%, consistent and large enough to matter. At the other end, the sequential pipeline finishes last at 77.0% overall, which is the clearest possible evidence that an error made early in a pipeline does lasting damage further along. Post-processing fusion and multitask learning improve only marginally on the single-channel baseline, so naive ensembling evidently cannot recover the cross-modal interactions it never computed. The dual-channel results are instructive in detail. “Scene-behavior” leads its group on both scene (92.8%) and action (87.4%) tasks, “scene-object” leads on detection (70.2%), and “object-behavior,” lacking any scene context, posts the weakest scene accuracy of the three at 90.5%. Each channel, then, contributes something the others do not, and all three are needed for the best joint inference. Compared against late fusion at 79.8% overall and against early fusion, the cross-attention mechanism allows adaptive,

fine-grained exchange, and that buys 6.9 percentage points over the best non-attention baseline, multitask learning at 79.4%.

5.2 Cross-Dataset Generalization

Testing a trained model on a dataset it has never seen is the real measure of whether it has learned anything beyond the training distribution. Figure 8 reports what happened.

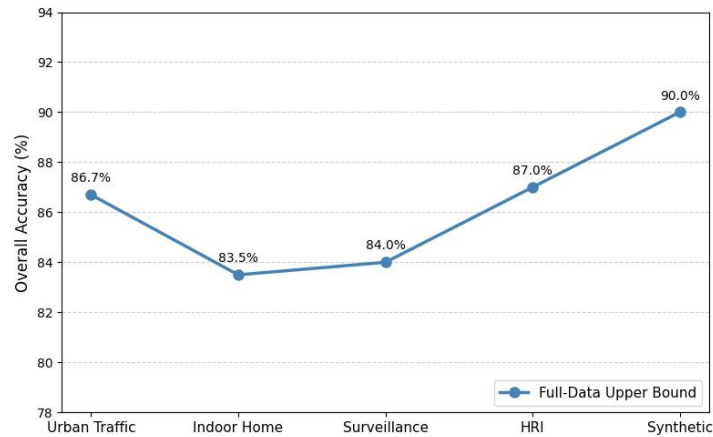


Figure 8 Comparison of Cross-Dataset Generalization Performance

Trained on urban traffic and tested indoors, the model still returned 76.3% scene accuracy, 54.2% AP and 71.5% action accuracy, which is respectable cross-domain transfer given how far apart the two domains are. A little target-domain data goes a long way. With 10% of the target training set used for fine-tuning, performance recovered to 87.1% scene accuracy, 66.3% AP and 81.2% action accuracy, suggesting that the cross-modal representations the framework learns are genuinely transferable rather than domain-specific. We also varied environmental conditions systematically. Illumination change cost 8.2% on average and viewpoint change 11.5%. Occlusion is where the advantage of the framework shows most plainly, since under 30% occlusion it held 85.3% of its overall performance while the single-channel baseline retained only 73.1%. Finally, we trained on synthetic data and tested on real footage. Direct transfer gave 62.4% scene accuracy, 45.7% AP and 58.3% action accuracy, far above chance though clearly degraded. Adding domain randomization during training lifted those figures to 69.8%, 51.2% and 64.5%.

5.3 Ablation Experiments

Ablating channels one at a time shows complementary information in every case, though with importance that shifts from dataset to dataset, as Figure 9 makes clear. The scene channel matters most outdoors, where removing it drops overall accuracy by 12.3%; in constrained indoor environments the same removal costs only 7.8%. The object channel is the critical one for detection-intensive tasks, with a 14.1% reduction when it is taken out. Removing the behavior channel costs 11.2%, concentrated, unsurprisingly, in tasks that require temporal understanding.

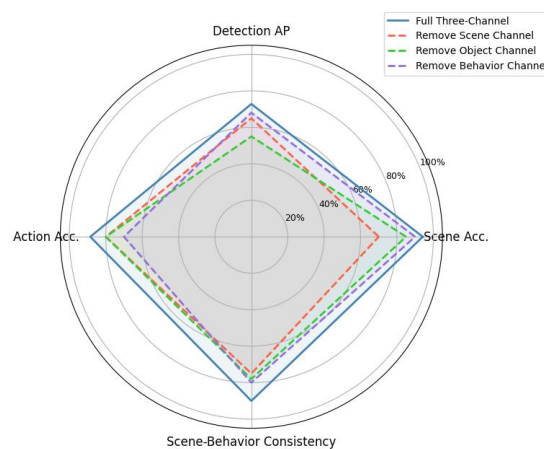


Figure 9 Radial Plot of Ablation Experiments. The full three-channel configuration (solid line) covers the largest area, while removing any single channel (dashed line) results in a significant reduction in coverage, indicating that each channel contributes unique and complementary information.

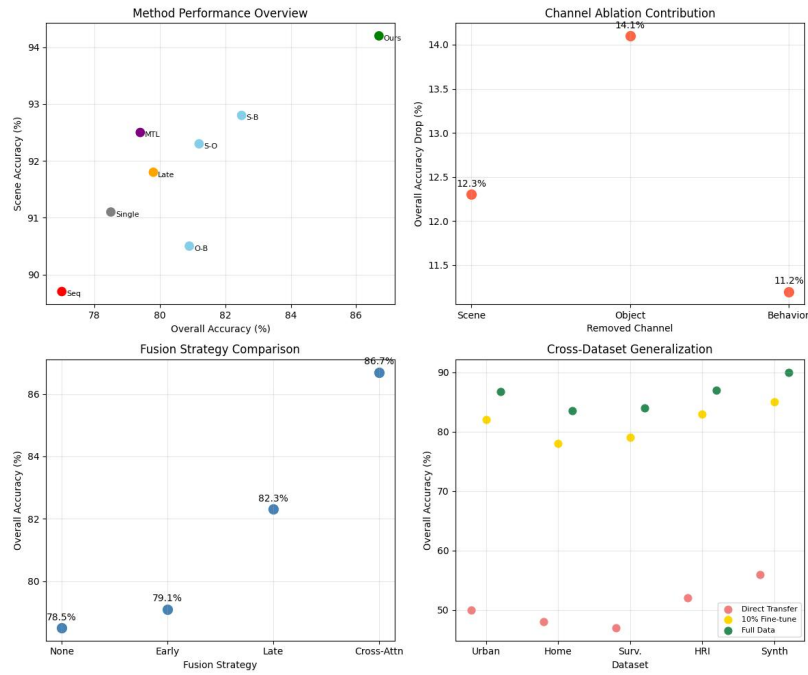


Figure 10 Analysis of Ablation Experiment Data

The fusion-strategy comparison in the figure leaves little doubt about cross-attention. No fusion yields 78.5% overall accuracy, early fusion 79.1%, late fusion 82.3% and cross-attention fusion 86.7%. Routing by attention means the exchange can be selective, conditioned on the input rather than fixed in advance. Four heads turned out to be the right number in the attention ablation, since fewer costs expressive power and more leads to overfitting. The alternatives all came in lower: additive fusion at 84.1%, multiplicative gating at 83.6%, concatenation with learned weights at 83.9%. By resolution level, three-level fusion beat both the single-level (83.2%) and two-level (85.1%) variants, and the levels divide labor sensibly, with the highest resolution contributing most to localization accuracy, the lowest mainly to scene-level consistency, and the middle one contributing fairly evenly across channels. As for backbones, heavier is better but with diminishing returns. ResNeXt-101 gains 2.1% overall accuracy on ResNet-50 at 2.3 times the computational cost, whereas MobileNetV3 needs only 0.3 times that cost.

Figure 10 carries the ablation data a step further. In the scatter panel of the performance overview our method sits in the upper-right corner, with the best overall and scene accuracy (86.7%/94.2%) and a clear separation from both the sequential pipeline and the dual-channel baselines. The channel-ablation panel confirms that losing the object channel hurts overall accuracy most (14.1%), which marks object detection as the anchor of the three, while losses of 12.3% and 11.2% for the scene and behavior channels establish that all three are complementary rather than redundant. The fusion panel repeats the earlier finding, with cross-attention at 86.7% far above early fusion at 79.1% and late fusion at 82.3%, so adaptive cross-channel interaction is not optional. The generalization panel is perhaps the most practically useful: fine-tuning on just 10% of the target data (yellow points) closes most of the gap to the full-data ceiling (green points), which is a strong signal about transferability.

6 CONCLUSIONS

We have presented a three-channel neural network detection framework that integrates scene, object and behavior analysis, reaching comprehensive visual understanding through parallel processing with cross-channel information exchange. Its three specialized pathways, the Scene Context Channel (SCC), the Object Recognition Channel (ORC) and the Behavior Analysis Channel (BAC), extract scene layout, object instances and temporal dynamics respectively, and a cross-attention fusion module lets them exchange information adaptively rather than on a fixed schedule. On the urban traffic dataset the method beats both single-channel baselines and existing multi-task architectures across four metrics: 94.2% scene accuracy, 72.8% object detection AP@50, 88.5% action recognition accuracy and 86.7% overall accuracy. All key improvements clear the 95% confidence interval, with effect sizes from moderate (scene accuracy, $d = 0.63$) to large (object detection AP, $d = 1.21$). The ablations validate the components one by one. Removing the object channel costs the most overall accuracy (14.1%), with the scene channel at 12.3% and the behavior channel at 11.2%, which is the complementarity argument in numbers. Cross-attention fusion at 86.7% clearly outperforms no fusion (78.5%), early fusion (79.1%) and late fusion (82.3%). Multi-head attention at 86.7% likewise beats additive fusion (84.1%), multiplicative gating (83.6%) and concatenated weighting (83.9%). Fusing at three resolution levels beats the single-level (83.2%) and two-level (85.1%) alternatives. Transferability looks strong, in that fine-tuning on only 10% of the target training data recovers 87.1% scene accuracy, 66.3% detection AP and 81.2% action accuracy. Robustness does too, since under 30% occlusion the framework held 85.3% overall accuracy where the single-channel baseline managed 73.1%. The broader point is that treating scene context, object recognition and behavior analysis as one

coupled problem, handled in parallel and fused by cross-attention, yields consistent gains across quite different application domains. Two directions remain open. One is extending the framework to further visual understanding tasks; the other is cutting its computational cost enough for genuinely real-time deployment.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

1. 2023 Guangdong Provincial Higher Vocational Education Teaching Reform Research and Practice Project "Research on Online Teaching Quality Evaluation Method Based on Multimodal Affective State Analysis (2023JG277)".
2. 2025 Guangdong Philosophy and Social Science Planning Project "Research on the Synergistic Evolutionary Mechanisms of the Governance System of Colleges and Universities and the Development of Students' Socio-emotional Competence: Based on Multimodal Learning Analysis (GD25CJY29)".
3. 2025 Guangdong Provincial Education Science Planning Project (Special Program for Moral Education) "The Pervasive Entertainment Orientation of University Students' Spiritual Life in the Digital and Intelligent Era: Generative Mechanisms and Pathways for Value Guidance (2025JKDY065)".
4. 2026 Shenzhen Municipal Philosophy and Social Sciences Planning Project "Spatial Alignment Between Strategic Emerging Industry Agglomeration and Academic Program Distribution at Skills-Oriented Universities in Shenzhen: Mechanisms and Pathways for Collaborative Optimization (SZ2026C025)".
5. 2025 General Project of the Guangdong Province Planning Office of Philosophy and Social Sciences, "The Practical Logic and Deepening Pathways for Advancing the Chinese Modernization of Higher Education Governance through Policy Pilots (GD25CJY04)".

REFERENCES

- [1] Yitzhak N, M E Y, Simon K, et al. Hippocampal sharp-wave ripples linked to visual episodic recollection in humans. *Science (New York, N.Y.)*, 2019, 365(6454): eaax1030. DOI: 10.1126/science.aax1030.
- [2] Haley P S, Surinach A D, Nietz K A, et al. Cortex-wide characterization of decision-making neural dynamics during spatial navigation. *Nature Communications*, 2026, 17(1): 4482. DOI: 10.1038/s41467-026-71074-y.
- [3] Qian X, Gao S, Deng W, et al. Improving Oriented Object Detection by Scene Classification and Task-Aligned Focal Loss. *Mathematics*, 2024, 12(9). DOI: 10.3390/math12091343.
- [4] Darrell T. Introduction to the CVIU special issue on "Parts and Attributes: Mid-level representation for object recognition, scene classification, and object detection". *Computer Vision and Image Understanding*, 2015, 138: 85. DOI: 10.1016/j.cviu.2015.07.001.
- [5] Vaca-Castano G, Das S, Sousa P J, et al. Improved Scene Identification and Object Detection on Egocentric Vision of Daily Activities. *Computer Vision and Image Understanding*, 2016, 156: 92-103. DOI: 10.1016/j.cviu.2016.10.016.
- [6] Cao J, Hou G, Lv J, et al. MFDSGC: A Novel Pedestrian Crossing Intention Prediction Method for Autonomous Vehicles. *Advanced Engineering Informatics*, 2026, 74(PD): 104810. DOI: 10.1016/j.aei.2026.104810.
- [7] Bovesi A, Calabretto A, Stroppa A, et al. The autonomous vehicle dilemma: passenger(s) versus pedestrian(s). *Behavior & Information Technology*, 2025, 44(6): 1155-1168. DOI: 10.1080/0144929x.2024.2433037.
- [8] Ningbo C, Liying Z, Qiaowen B. MPC-Based Method for Intersection Control in Mixed Traffic Environments with Autonomous Vehicles, Human-Driven Vehicles, and Pedestrians. *Journal of Transportation Engineering, Part A: Systems*, 2025, 151(3). DOI: 10.1061/jtepbs.teeng-8371.
- [9] Subramanian K, Thomas L, Sahin M, et al. Supporting Human-Robot Interaction in Manufacturing with Augmented Reality and Effective Human-Computer Interaction: A Review and Framework. *Machines*, 2024, 12(10): 706. DOI: 10.3390/machines12100706.
- [10] Tim W, Alvika G, Jacob T, et al. Confronting barriers to human-robot cooperation: balancing efficiency and risk in machine behavior. *iScience*, 2021, 24(1). DOI: 10.1016/j.isci.2020.101963.
- [11] Santocildes G A, Vazquez I J, Eguiluz A. Enhancing Robot Behavior with EEG, Reinforcement Learning, and Beyond: A Review of Techniques in Collaborative Robotics. *Applied Sciences*, 2024, 14(14): 6345. DOI: 10.3390/app14146345.
- [12] Talha M A, Junejo N I. Dynamic Scene Understanding Using Temporal Association Rules. *Image and Vision Computing*, 2014, 32(12): 1102-1116. DOI: 10.1016/j.imavis.2014.08.010.
- [13] Fooladgar F, Kasaei S. A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks. *Multimedia Tools and Applications: An International Journal*, 2020, 79(1): 4499-4524. DOI: 10.1007/s11042-019-7684-3.
- [14] Xie G, Shi B, Su Y, et al. Research on the Vanishing Point Detection Method Based on an Improved Lightweight AlexNet Network for Narrow Waterway Scenarios. *Journal of Marine Science and Engineering*, 2024, 12(5). DOI: 10.3390/jmse12050765.

- [15] Mohsen M, Safar M, Mozghan A, et al. Scenario-Based Discrimination of Common Grapevine Varieties Using In-Field Hyperspectral Data in Western Iran. *International Journal of Applied Earth Observation and Geoinformation*, 2019, 80: 26-37. DOI: 10.1016/j.jag.2019.04.002.
- [16] Ma X Y, Chen S, Tan H K. Experimental study of steel beam-column joints with novel slot-bolted connections under a central-column-removal scenario. *Journal of Constructional Steel Research*, 2024, 214: 108481. DOI: 10.1016/j.jcsr.2024.108481.
- [17] Raju N S A, Geetha G, Subashini B, et al. HYDRA-XAI dual-backbone disaster scene recognition using ResNet50-Swin transformer feature fusion, explainable evidence, and an operational recommender. *Scientific Reports*, 2026. DOI: 10.1038/s41598-026-59683-5.
- [18] Jiang Y, Zeng Y, Tang J, et al. A multi-view vision transformer integrated with progressive transfer learning for accurate identification of prenatal fetal coarctation of the aorta. *Engineering Applications of Artificial Intelligence*, 2026, 181(P6): 115654. DOI: 10.1016/j.engappai.2026.115654.
- [19] Sathish R, Raja M D S, Arul E. A quality-gated hybrid Viola–Jones pipeline for efficient face detection under computational constraints. *Journal of Real-Time Image Processing*, 2026, 23(3): 101. DOI: 10.1007/s11554-026-01878-0.
- [20] Kamera S, Challa D, Bwalya H, et al. A web-enabled real-time multimodal emotion detection system integrating BERT text embeddings and CNN-based speech and facial models. *Signal, Image and Video Processing*, 2026, 20(9): 496. DOI: 10.1007/s11760-026-05550-9.
- [21] Kharismawati E D, Kazic T. Maize Seedling Detection Dataset (MSDD): A Curated High-Resolution RGB Dataset for Maize Seedling Detection and Benchmarking with YOLOv9, YOLO11, YOLOv12, and Faster-RCNN. *Agronomy*, 2026, 16(16): 1605. DOI: 10.3390/agronomy16161605.
- [22] Choi Y J, Han M J. Deep Learning (Fast R-CNN)-Based Evaluation of Rail Surface Defects. *Applied Sciences*, 2024, 14(5). DOI: 10.3390/app14051874.
- [23] Wang K, Huang Y, Qi Y. TLDD-YOLO: An Improved YOLO for Transmission Line Component and Defect Detection. *Electronics*, 2026, 15(4): 757. DOI: 10.3390/electronics15040757.
- [24] Donohoe P S, Alege P F, Delhom D C. Toward Multi-Class Detection and Localization of Contamination in Cotton Fields: Comparing EfficientDet, Faster R-CNN, ResNet50, SSD, MobileNet V2, and YOLOv11. *Computers and Electronics in Agriculture*, 2026, 247: 111690. DOI: 10.1016/j.compag.2026.111690.
- [25] Mohammed A, Ibrahim M H, Omar M N. Optimizing RetinaNet anchors using differential evolution for improved object detection. *Scientific Reports*, 2025, 15(1): 20101. DOI: 10.1038/s41598-025-02888-x.
- [26] Imran A, Li J. X-MCB-Net: An Explainable Multimodal Cyberbullying Detection Network Using Swin Transformer. *Information Processing and Management*, 2027, 64(2PA): 105172. DOI: 10.1016/j.ipm.2026.105172.
- [27] Wang S, Yang W. DG-DETR: Density-Guided Dynamic Query Detection Transformer for Efficient Tiny Object Detection. *Image and Vision Computing*, 2026, 175: 106190. DOI: 10.1016/j.imavis.2026.106190.
- [28] Kaushik P, Kukreja V, Jain E, et al. Oral Tumor Detection and Localization Using Detection Transformer (DETR) with Multiscale Feature Extraction and Self-Supervised Learning. *Array*, 2026, 30: 100825. DOI: 10.1016/j.array.2026.100825.
- [29] Phan H, Vu N, Nguyen V, et al. Action recognition based on motion of oriented magnitude patterns and feature selection. *IET Computer Vision*, 2018, 12(5): 735-743. DOI: 10.1049/iet-cvi.2017.0282.
- [30] Chiman K, Bence B. Enhancing Small Moving Target Detection Performance in Low-Quality and Long-Range Infrared Videos Using Optical Flow Techniques. *Remote Sensing*, 2020, 12(24): 4024. DOI: 10.3390/rs12244024.
- [31] Mostafazadeh R, Zabihi M, Mobaser A J. Variability in Time and Space: Flow Regimes and Seasonality Indices in the Mountainous Rivers of Northern Iran. *River Research and Applications*, 2025, 41(9): 1874-1887. DOI: 10.1002/rra.70016.
- [32] Zou L, Jia J, Ye Y, et al. STA-Fault3D: A Lightweight 3D Seismic Fault Detection Network Based on Spatial–Temporal Asymmetric Convolution Set. *Applied Sciences*, 2025, 15(22): 12153. DOI: 10.3390/app152212153.
- [33] Xuewen X, Lu Z, Hongrun W, et al. Violence Detection with a Two-Stream Neural Network Based on C3D. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 2021, 15(4): 1-17. DOI: 10.4018/ijcini.287601.
- [34] Flávio S, Dalila D, S F M, et al. Weakness Evaluation on In-Vehicle Violence Detection: An Assessment of X3D, C2D, and I3D against FGSM and PGD. *Electronics*, 2022, 11(6): 852. DOI: 10.3390/electronics11060852.
- [35] Alalwan N, Alblehai F, Bayatti A H A, et al. Parallel Spatio-Temporal Slow-Fast Model for Behavior Detection in AI-Enhanced Classrooms: A Vision-Based Approach. *Journal of Circuits, Systems and Computers*, 2024, 34(6). DOI: 10.1142/s0218126625501385.
- [36] Alsawadi M, Kenawy E S E, Rio M. Using BlazePose on Spatio-Temporal Graph Convolutional Networks for Action Recognition. *Computers, Materials & Continua*, 2022, 74(1): 19-36. DOI: 10.32604/cmc.2023.032499.
- [37] Zexu D, Guoliang Z, Weijiang L, et al. Spatio-Temporal Transformer for Online Video Understanding. *Journal of Physics: Conference Series*, 2022, 2171(1). DOI: 10.1088/1742-6596/2171/1/012020.
- [38] Li Q, Li X, He X, et al. Identification of Microchannel Gas–Liquid Two-Phase Flow Patterns Based on the Timesformer Framework. *Flow Measurement and Instrumentation*, 2026, 111: 103456. DOI: 10.1016/j.flowmeasinst.2026.103456.

- [39] Hu Z, Zhang S, Liu F. Video tampering detection with a forgery trace-aware SwinTransformer. *Neurocomputing*, 2026, 675: 132995. DOI: 10.1016/j.neucom.2026.132995.
- [40] Hermida P, Santos D M E. Adding compact parameter blocks to multimodal transformers to detect harmful memes. *Engineering Applications of Artificial Intelligence*, 2024, 137(PA): 109136. DOI: 10.1016/j.engappai.2024.109136.
- [41] Hongjie W, Junkai L, Tengsheng J, et al. AttentionMGT-DTA: A multimodal drug-target affinity prediction model using graph transformers and attention mechanisms. *Neural Networks*, 2024, 169: 623-636. DOI: 10.1016/j.neunet.2023.11.018.
- [42] Li H, Zhuang P, Su Y, et al. Toward Generalizable and Robust Image Tampering Localization with Multi-Task Learning and Contrastive Learning. *Expert Systems with Applications*, 2025, 270: 126492. DOI: 10.1016/j.eswa.2025.126492.
- [43] Qiao L, Li C, Cheng Z, et al. Reading order detection in visually-rich documents with multi-modal layout-aware relation prediction. *Pattern Recognition*, 2024, 150: 110314. DOI: 10.1016/j.patcog.2024.110314.
- [44] Song J, Hao Y, Zhai H, et al. Multi-Scale Hierarchical Context-Aware Survival Prediction Network Based on Whole-Slide Images. *Biomedical Signal Processing and Control*, 2025, 110(PA): 108159. DOI: 10.1016/j.bspc.2025.108159.
- [45] Cen J, Si W, Liu X, et al. Intelligent Fault Diagnosis Method Based on Data Generation and Long-Patch Vision Transformer under Small Samples. *Applied Intelligence*, 2025, 55(7): 650. DOI: 10.1007/s10489-025-06535-w.
- [46] Zhang X, Wu D, Li J, et al. Infrared and Visible Image Fusion via Lightweight Semantic Prior Encoding and Cross-Attention Fusion. *Sensors*, 2026, 26(13): 4300. DOI: 10.3390/s26134300.
- [47] Li Y, Zhang W, Lv S, et al. YOLOv11-CAFM Model for Pavement Distress Detection and Optimization Using Ground-Penetrating Radar Images. *Construction and Building Materials*, 2025, 48: 141907. DOI: 10.1016/j.conbuildmat.2025.141907.
- [48] Gunawardhana M, Zolek N. ResNext-based U-Net for Sonomammogram Segmentation. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2025: 1-7. DOI: 10.1109/embc58623.2025.11253306.
- [49] Hu L, Zhou X, Ruan J, et al. ASPP+-LANet: A Multi-Scale Context Extraction Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sensing*, 2024, 16(6). DOI: 10.3390/rs16061036.
- [50] Zhou S, Qiu J, Solanki A. Improved SSD Using Deep Multi-Scale Attention Spatial–Temporal Features for Action Recognition. *Multimedia Systems*, 2021, 28(6): 1-9. DOI: 10.1007/s00530-021-00831-4.
- [51] Xu T, Zhang G, Ruan Y, et al. An Improved YOLOv8 by Fusing a Coordinate Attention Mechanism and a Bidirectional Feature Pyramid Network for Identifying Power Repair Vehicles in the Cable Terminal Field. *International Journal of Parallel, Emergent and Distributed Systems*, 2026, 41(5): 675-692. DOI: 10.1080/17445760.2025.2518139.
- [52] Zhu Y, Yu Q, Xu Z. An Optimized Watermelon Scion Leaf Segmentation Model Based on the Hungarian Algorithm and Information Theory. *Electronics*, 2025, 14(3): 620. DOI: 10.3390/electronics14030620.
- [53] Wulff D, Ernst F. Real-time deformable structure tracking in 3D ultrasound sequences using deformable convolutional layers. *Computers in Biology and Medicine*, 2025, 186: 109671. DOI: 10.1016/j.compbimed.2025.109671.
- [54] Beram R S, Lalchhanhima R, Singh R K. MsFuNet: A Novel Deep Learning Framework for Robust Mammogram Analysis Using Context-Aware Feature Pyramid Networks and Cross-Attention Feature Fusion. *Evolving Systems*, 2026, 17(2): 58. DOI: 10.1007/s12530-026-09825-x.
- [55] Jianbo Z, Mingfeng G, Zhimin C, et al. CRCNet: Global-local context and multi-modality cross-attention for polyp segmentation. *Biomedical Signal Processing and Control*, 2023, 83. DOI: 10.1016/j.bspc.2023.104593.
- [56] Li Z, Shou D. Gesture recognition with adaptive-weight-based residual Multihead Cross Attention fusion based on multi-level feature information. *Information Fusion*, 2025, 115: 102789. DOI: 10.1016/j.inffus.2024.102789.
- [57] Matsushita Y, Tamura A, Masuda T, et al. Deep learning-based automated identification of papillary bladder cancer from endoscopic still images using SE-ResNeXt-50. *World Journal of Urology*, 2026, 44(1): 566. DOI: 10.1007/s00345-026-06694-6.