Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 11(1): 1547–1589.
DOI: 10.3934/math.2026065
Received: 13 November 2025
Revised: 17 December 2025
Accepted: 22 December 2025
Published: 19 January 2026

Research article

Cross-view remote sensing and street-level data fusion for intelligent traffic congestion analysis

Inzamam Mashood Nasir[1,*], **Hend Alshaya**[2], **Sara Tehsin**[3] **and Wided Bouchelligua**[2]

[1] Human-Environment-Technology (HET) Systems Centre, Mykolas Romeris University, Vilnius 08303, Lithuania

[2] Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[3] Faculty of Informatics, Kaunas University of Technology, 51368 Kaunas, Lithuania

* **Correspondence:** Email: inzamam.nasir@mruni.eu.

Abstract: The issue of urban traffic congestion is a persistent problem for the sustainable management of cities through transportation systems, as there is a need for models that integrate and analyze heterogeneous sources to yield accurate, interpretable outcomes. This paper introduces the cross-view fusion network (CVF-Net), a new multimodal deep learning framework for analyzing congestion across entire cities by combining remote-sensing imagery (drone aerial views), street-view camera images, and graph-structured sensor data into a single model. This model is introduced through a very novel architecture that includes a hierarchical attention fusion transformer (HAFT), which fuses cross-view attention (CVA) between the aerial and ground view, a temporal graph neural network (TGNN) that uses a spatio-temporal dynamic, and a graph refinement (GR) network for consistency relative to the graph topology. Extensive experiments across three benchmarks (CityFlowV2, METR-LA, PEMS-BAY) demonstrate that CVF-Net consistently outperforms other recent state-of-the-art methods, reducing forecasting error (MAE) by 9.3% and increasing tracking continuity (IDF1) by 7.0%. Ablation studies suggest that hierarchical fusion and temporal modeling improve accuracy and stability, while sensitivity analyses show that attention maps capture congestion and causal temporal patterns, which are real symptoms of congestion. The model also shows strong cross-dataset generalizability and robustness to sensor noise, which extends its performance in the real world. Unlike existing spatio-temporal GNNs and multimodal Transformers that rely on flat feature aggregation or implicitly assume cross-view alignment, the proposed framework introduces a hierarchical, alignment-aware fusion strategy that explicitly integrates aerial visual context with graph-temporal traffic dynamics.

Keywords: cross-view fusion network; remote sensing; traffic congestion analysis; hierarchical attention; temporal graph neural network; multimodal deep learning

Mathematics Subject Classification: 68T07

1. Introduction

Traffic congestion is one of the longest-lasting problems in urban transport networks worldwide, resulting in increased travel time, higher emissions, and substantial economic costs [1]. Typical congestion monitoring methods rely mostly on stationary sensors (inductive loops, radar), or floating vehicle probes, but these usually only provide a localized or fragmented view of traffic state and may not reflect the full spatio-temporal dynamics of urban grids [2]. Similarly, recent developments in remote sensing and computer vision make it increasingly feasible to use overhead imagery and street-level views to provide rich spatial context, e.g., lane-level vehicle densities, queue lengths, and road geometry [3]. On the other hand, research articles bridging remote-sensing imagery with traffic sensor data (cross-view fusion) are rare, especially in support of congestion forecasting tasks.

Over the last decade, deep learning has upended traffic forecasting, allowing models to learn complex spatial and temporal dependencies in data. Graph neural networks (GNNs) utilize road-network topologies and propagation of traffic over space and time [4], while convolutional neural networks (CNNs) treat traffic data as images, or grids, to generate spatial features through weight transfer [5]. Surveys of congestion detection and prediction identify several critical gaps in the literature: (i) lack of cross-modality fusion, (ii) lack of interpretability, and (iii) lack of generalization to new cities/sensor setups [6, 7]. Meanwhile, multi-sensor fusion studies show that combining heterogeneous data sources, e.g., loop detectors and floating-car data, improves speed and travel time estimates by providing greater reliability under congestion than models forecasting speed from only traffic loop sensors [8]. More recently, cross-view learning studies, ie, satellite + street view, hold promise for more comprehensive spatial modelling. However, studies on cross-view learning focus predominantly on static tasks, such as hot-spot detection, and less so on dynamic congestion forecasting [9].

To address these limitations, we introduce a single cross-view fusion framework, the cross-view fusion network (CVF-Net), that enables city-scale congestion analysis using remote-sensing aerial imagery, street-level camera data, and graph sensor networks for visualization. The proposed framework includes three key components: (i) a vision transformer (ViT) to extract visual contextual features from imagery, (ii) a temporal graph neural network (TGNN) that processes temporally-dependent graph-based sensor streams, and (iii) a hierarchical attention fusion transformer (HAFT) that merges the layers from the ViT and TGNN components while learning contextual dependencies across the two views. CVF-Net is experimentally validated across three benchmark datasets (CityFlowV2, METR-LA, and PEMS-BAY) for both multi-camera tracking and sensor-based temporal forecasting. In addition to providing quantitative results demonstrating more consistent improvements over recent state-of-the-art methods (i.e., MAE 9% and IDF1 7%) we also provide ablation studies demonstrating the importance of cross-view attention (CVA), and graph refinement (GR) for performance and, lastly, sample interpretability analysis where we show that the attention maps clearly attended to congestion regions, rather than background features. Additionally, we run cross-dataset experiments to validate the generalization of our model to new urban contexts. To avoid ambiguity regarding the nature of the contributions, we emphasize that this work does not introduce entirely new, standalone learning primitives. Instead, the primary contribution lies in a novel system-level integration of existing techniques, specifically designed for cross-view urban traffic analysis. While attention mechanisms, graph neural networks, and transformers have been explored individually in prior studies, CVF-Net is, to the best of our knowledge, the first framework to hierarchically couple aerial visual context with temporal graph reasoning via explicit cross-view attention and feedback-driven graph refinement.

To clarify the novel aspects of this research, we outline its key contributions. The CVF-Net does not define a new task nor alter traditional representations of graph neural networks (GNNs) or transformers. The key influence lies in the fusion of different image types: a hierarchical cross-view mechanism that merges spatially coarse aerial images and street-level images based on their respective time frames, thereby creating a complete spatial/temporal traffic graph. Most existing GNNs and transformer models combine information across different modalities through either early or late fusion. It is normally assumed that all modalities are aligned in both space and time; however, this is not always the case.

The remainder of the paper is organized as follows. Section 2 reviews related work, Section 3 details our methodology, Section 4 presents experimental results, and Section 5 concludes with future research directions.

2. Related work

Traffic congestion analysis has been the subject of a plethora of literature that utilises a range of sensing modalities and machine learning paradigms. Traditionally, loop detectors, GPS trajectories, and surveillance cameras have been leveraged to estimate traffic flow and travel time [1, 8]. Although these data can achieve high temporal resolution, they tend to have limited spatial coverage and are costly to maintain. With the recent availability of high-resolution satellite and aerial imagery, it is possible to analyse traffic scenes at a large scale, gaining top-down visual indicators such as road density, queue length, and intersection shape [3, 10]. Standalone remote-sensing approaches struggle with temporal continuity, and they cannot model the dynamic propagation of congestion in the network [11, 12]. Data-driven fusion of imagery and measurements at the ground level remains a compelling direction for scalable traffic monitoring, as most researchers have adopted [2, 13].

Recent cross-modal urban analysis models integrate images and traffic data, typically via shallow fusion or shared latent spaces; in contrast, CVF-Net employs a hierarchical, feedback-driven interaction between the two modalities. Most existing multimodal traffic forecasting methods (including those that utilise multiple sensors) process the two modalities independently before being combined. In contrast, CVF-Net explicitly combines visual perception with graph-temporal reasoning via cross-view attention and graph refinement. Thus, CVF-Net is positioned as a structurally, but not functionally, different multimodal approach to urban traffic analysis compared to prior work on fusion models.

Deep learning has enabled traffic prediction by capturing nonlinear spatial and temporal dependencies in heterogeneous data streams. Early convolutional models, which treated traffic speed maps as spatial images, were able to learn pixel-level correlations using CNNs [5]. Recurrent models such as LSTM and GRU learned sequential patterns, thereby also ignoring spatial topology. GNNs are now prevalent with sensors as nodes and road links as edges, which permits the efficient propagation of a traffic state through graph convolution [4, 14]. Variants, notably T-GCN [15], STGCN [16], and ASTGCN [17], demonstrated the potential benefits of joint spatial-temporal modelling. More recent models push the expressiveness gains from dynamic adjacency learning and temporal attention mechanisms, such as D2STGNN and STGODE [18, 19]. Despite their success, almost all GNN-based approaches depend solely on sensor time-series, ignoring other contextual cues readily available from overhead or street-level imagery.

In recent years, the integration of multimodal and cross-view data has gained increasing attention in the context of intelligent transportation systems. These studies, which integrate loop detectors and the global positioning system (GPS) with probe vehicles, show that multimodal data fusion can reduce estimation uncertainty during peak congestion [8, 7]. Similarly, cross-view learning has been considered in the urban computing field, analysing remote-sensing and street-view images together to estimate land use, traffic density, and road type [9, 20]. All of these works, however, focus on static scene understanding instead of temporal forecasting. Recently, vision-language and multimodal transformer-based models have started to address this issue, achieving superior interpretability and generalisation across modalities [21, 22]. Nevertheless, there have been limited studies that have focused on the convolution of remote-sensing imagery and temporal sensor graphs for congestion forecasting.

While the field has made enormous progress in applying remote sensing and deep learning techniques to traffic analysis, significant gaps remain in the literature. First, while graph-based forecasting methods are among the most promising, they still rely heavily on sensor measurements from traffic monitoring, even when a wealth of contextual information is available from both aerial and street-level images. In contrast, the approaches that have arisen around remote sensing focus exclusively on either static congestion mapping or vehicle detection and are not designed for either temporal or topological reasoning. Multimodal fusion frameworks in the literature employ only feature concatenation or shared embedding spaces for modelling the same dataset and do not add an additional layer of meaning about the hierarchies between modalities and views. And lastly, there have been no interpretations of deep-learning traffic-based models, either from input data or from context based on collected differencing measures, to explain accurate predictions or identify which spatial regions or time intervals help estimate congestion.

In addition to research on remote sensing image analysis, recent work has focused on combining large foundation models with task-specific feature fusion. In some cases, using the segment anything model (SAM) and learned feature fusion greatly enhanced change-detection performance when complex background variations were present. Moreover, this increased performance was achieved with a limited amount of annotated data. These studies illustrate the value of combining general visual representations with domain-adaptive fusion strategies when utilising remote sensing in various applications. The primary focus of these studies is on pixel-level change detection rather than spatiotemporal traffic analysis; however, they provide strong evidence to support that robust visual feature extraction and fusion are critical when using aerial photography as input for downstream processing tasks [23, 24].

Methodologically, existing methodologies can be categorised under three major categories: (i) GNNs based on spatio-temporal that exclusively deal with a graph of a sensor. (ii) Multi-modal graph transformers for effective modelling of long-range dependencies in a single model. (iii) CVF is a technique used in data collection where both visual and non-visual modalities are integrated with one another through a shallow and/or parallel fusion approach. The major distinguishing feature of

CVF-Net is that it is a single platform that integrates all the above-mentioned technologies through a unified hierarchical framework. The unifying aspect of CVF-Net enables cross-view attention, temporal modelling, and graph refinement to occur sequentially, thereby providing explicit processing of modality differences and potential cross-view misalignment that were not addressed in previous multimodal traffic modelling systems.

In realising this, the CVF-Net architecture, proposed as the final goal of the dissertation, put forth a unified cross-view fusion paradigm to jointly model spatial, temporal, and topological dependencies, which, at the same time, help interpret the deep learning characteristics of the model. The architecture relies on an HAFT model that aligns aerial, ground, and sensor-based features using CVA, enabling remote sensing to interact with in-situ traffic signals and drive dynamic interactions. Also, in capturing time-changing dependencies, we add a TGNN and, furthermore, a GR module to ensure that road segments that are connected have structural consistency across the entirety of the connection. All these dimensions described above of an applied CVF-Net model will overcome the above limitations to learn semantically grounded and interpretable representations that generalise across both datasets and cities, leading to richer, more robust, and cross-modal congestion.

3. Proposed methodology

We propose a CVF-Net for modeling urban congestion by jointly utilizing spatially aligned remote-sensing imagery and street-level traffic data. For clarity, we summarize the pipeline structure and key distinctions of CVF-Net prior to detailing its mathematical formulation. The method proceeds in three main stages: (i) spatial and temporal alignment between imagery and graphs, (ii) dual-stream feature extraction with visual and graph-temporal encoders, and (iii) hierarchical fusion via feedback-driven graph refinement. Standard deep learning encoders are used, but CVF-Net's core novelty lies in its hierarchical attention-based fusion and the reintroduction of visual context for graph reasoning. The pipeline (see Figure 1) consists of: (i) cross-view alignment and preprocessing, (ii) multimodal feature extraction, (iii) hierarchical fusion using attention-based graph–transformer blocks, and (iv) congestion prediction with map reconstruction.

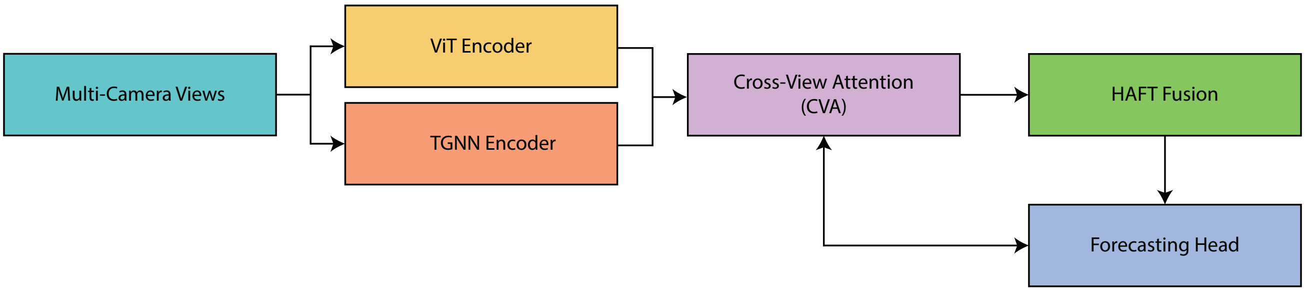

Figure 1. Overall CVF-Net architecture integrating ViT and TGNN encoders with CVA, hierarchical attention fusion (HAFT), and the forecasting head.

We make a clear distinction between our study's original innovation and all other components of its design, which may be developed from prior literature. In terms of designs that have previously been proven successful in analysing remote sensing data and forecasting traffic flow, both the visual encoder (CNN–ViT) and the temporal graph neural network (TGNN) would fit into traditional design methodologies. The cross-view attention (CVA) mechanism adapts established attention mechanisms to enable interaction between visual and graph-temporal data. A key aspect of the graph refinement (GR) module is enhancing the spatial coherence of the message-passing graph model using graph attention methods. Finally, our main contribution is the development of a new architecture, the hierarchical attention fusion transformer (HAFT), which integrates the standard CVA design, temporal fusion, and GR across multiple stages and via a feedback process.

3.1. Cross-view data alignment and preprocessing

The proposed framework starts with accurate spatial and temporal correspondence between remotely sensed imagery and street-level traffic data. Let us denote a geo-referenced remote sensing tile at time t by $I_t \in \mathbb{R}^{H\times W\times C}$, where H, W, and C indicate the height, width, and number of spectral channels of the image, respectively. The associated transportation network is conceptualized as a directed weighted graph $G = (V, E)$, where V is the set of road segments equipped with sensors, and $E \subseteq V \times V$ encodes physical adjacency relationships between roads. Each node $v \in V$ is attributed with a time-varying feature vector that describes local traffic conditions as

$$\mathbf{s}_v^t = \left[s_v^t q_v^t o_v^t\right] \in \mathbb{R}^3, \tag{3.1}$$

where s_v^t denotes the mean speed of vehicles (km/h); q_v^t signifies the traffic flow (vehicles/min); and o_v^t indicates the occupancy of the entrapping sensor (ratio of time occupied). As such, the two observations I_t) and G comprise a multimodal observation pair to quantify both spatial characteristics and dynamic traffic flow at time t. In order for there to be spatial equivalence, the remote sensing image I_t should be geo-registered to the coordinate frame of the roadway network G using an affine transformation. Any pixel coordinate $\mathbf{x} = [x, y]^T$ in the image domain geo-references to the ground coordinate spatially using the defined relation.

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{3.2}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is the linear transformation matrix inherent to $I_t \in \mathbb{R}^{H\times W}$ account of both rotation and scale, and $\mathbf{b} \in \mathbb{R}^2$ is a translation vector. The pairs $(\mathbf{A}, \mathbf{b})$ are computed using the metadata provided in the satellite imagery and the OSM road-centerline coordinates. After this transformation, I_t is then rasterized at a consistent ground sampling distance (or GSD) of between 10–30 m/ m/pixel to ensure that every pixel represents a discrete form of spatial road unit Ω_i. In this manner, a one-to-one relationship was created between image pixels $\Omega = \Omega_i i = 1^{HW}$ to road segments $v_i i = 1^{|V|}$. Snapshot timings to synchronize asynchrony from the sensor readings were implemented to common periods of $Deltat \in 5, 15$ minutes. Spatial discrepancies arising from projection distortions, camera calibration errors, or GPS drift are mitigated through geo-referencing and local consistency constraints. The affine transformation in Eq (3.2) leverages metadata from remote-sensing imagery and known road centerline coordinates to map image pixels into a common ground coordinate system. To reduce the impact of residual spatial noise, alignment is performed at the road segment level rather than at the individual-pixel level, providing tolerance to minor GPS inaccuracies and projection distortions. This segment-level association ensures that small spatial errors do not propagate significantly into the fusion stage. The amalgamated value of a traffic feature x_v^t in the timestamped interval (t) for a given road segment

v $(x \in s, q, o)$ was computed as occurring in the time interval by simply taking the mean to arrive at summarized values within the given time interval, thus:

$$\bar{x}v^t = \frac{1}{N\Delta t} \sum_{\tau=t-\Delta t}^{t} x_v^\tau. \tag{3.3}$$

In which $N_{\Delta t}$ denotes the number of available samples within the time interval Δt. Temporal discrepancies between remote-sensing imagery and street-level video streams are handled through window-based synchronization rather than exact frame-level matching. Since aerial imagery and ground-level sensors are often captured at different and irregular sampling rates, all modalities are aligned to a common temporal grid using fixed aggregation windows of length Δt. Street-level video frames and sensor measurements within the same temporal window are aggregated into a single, synchronized observation, thereby mitigating the effects of acquisition delays and timestamp jitter. This strategy allows the model to learn from temporally consistent cross-view samples even when the acquisition times do not coincide precisely. Temporal interpolation is used to produce continuity in instances of missing and defective readings:

$$\tilde{x}_v^t = \alpha x_v^{t-1} + (1 - \alpha) x_v^{t+1}, \quad 0 \leq \alpha \leq 1. \tag{3.4}$$

In which α is a weighting coefficient that moderates the influence of past and future observations. This step reduces sensor noise, temporal sparsity, and data latency while promoting smooth transitions in traffic signals across pairs of consecutive time steps. We then take each aligned and aggregated instance as a cross-view training sample that pairs remote sensing imagery with synchronized traffic profile estimates:

$$\mathcal{D}_t = \Big(I_t, \mathbf{s}v^{(t-k):t} v \in \mathcal{N}(I_t)\Big), \tag{3.5}$$

where $\mathcal{N}(I_t) \subseteq V$ represents the set of road segments spatially contained in the image tile I_t, and k represents the temporal window that captures short-term dynamics. The complete dataset $\mathcal{D} = \mathcal{D}tt = 1^T$ then includes T multimodal pairs that are aligned in time across the observations. We have spatially and temporally aligned both modalities into a single dataset during this preprocessing step, thereby laying the foundation for joint representation learning and cross-view fusion. Cross-view alignment is illustrated in Figure 2.

A standardized pixel-to-road matching between aerial and sensor data occurs through cross-view alignment. Due to this standardization, both types of data can be utilized for future study via CVA and GR. Additionally, CVA synchronizes the two data modalities across space and time, enabling joint reasoning over visual context and dynamic signals during downstream processing. Rather than using an explicit supervisory signal, the alignment quality of the aerial and street data is validated by the geometric consistency and temporal coherence between the data sets. Misalignments between the sensor graph's road centerlines and the transformed remote-sensing tiles will lead to inconsistent aggregations of sensor nodes within the associated image regions, resulting in poor performance during training when fusing the data. Alignment reliability is ensured by discarding image-graph pairs that have insufficiently overlapped projected road segments and image support regions, and by synchronizing the two modalities within fixed temporal windows. By validating the geometric and

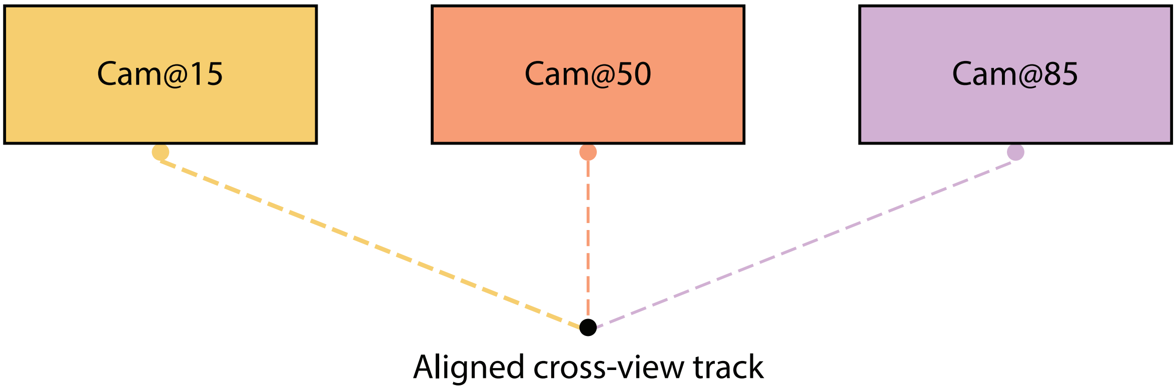

Figure 2. Schematic of CVA alignment among overlapping cameras, converging to a unified track hypothesis.

temporal integrity of the cross-view samples, only spatially and temporally consistent samples are used to learn representations.

Using metadata-based affine transformations, the aligned aerial images are initially aligned with the road network. Such alignment is not entirely effective in real-world applications, and due to GPS drift, mistakes in camera calibration, and/or inconsistent GSDs, both local and global, some level of spatial misalignment still occurs today. The following features of CVF-Net help alleviate moderate spatial misalignment. (1) The alignment of road segments is made at the road segment level and localised areas, instead of by pixel-to-pixel correspondence, and allows for some imprecision in projection errors. (2) The multi-scales of the Residual-Inception CNN layers of the CNN architecture build multiple-resolving receptive fields, which allow for limited sensitivity to GSD variations and minimal spatial distortion. (3) With the CVA mechanism, visual information will only be integrated selectively into the graph temporal representation based on the learning of its relevance. This enables the model to diminish or eliminate the effects of misaligned or poorly represented visual information in its input. As a result, CVF-Net does not require perfect geometric registration and is robust to moderate alignment noise common in typical urban sensing environments. Like most metadata-based fusion alignment solutions, extremely misaligned errors or a large GSD difference will degrade the quality of the resulting fusion products. To improve the robustness and accuracy of the newly created visual representations, it is possible to introduce multiple images from different views into the model simultaneously and model their relationships.

3.2. Multimodal feature extraction

The visual encoder's hybrid design combines a CNN and a ViT architecture. These architectures were created to address variations in scale and texture complexity, as well as the fine-grained spatial detail typically seen in Remote sensing imagery. Remote sensing aerial imagery is often associated with significant variations in scale, and has complex textures and fine-grain spatial detail, primarily caused by changes in road width, intersection geometry, vehicle density, and the altitude at which an Aerial Image was taken. In addition, a single ViT model is not well-suited to recovering local, texture-specific, and scale-specific features, although it is very effective at recovering global spatial relationships. To compensate for this shortcoming, a dual-scale residual-inception CNN is used as a preprocessing step to capture local and multi-resolution visual cues (such as lane marks and vehicle clusters, and intersection layouts) before feeding the features into the ViT architecture. The residual connections in the model preserve some low-level spatial features of the image, and the inception-style branches allow parallel extraction of features at different receptive field sizes; hence, the output feature maps are scale-aware and texture-sensitive, and are then tokenized and processed through the ViT backbone.

The purpose of the multimodal feature extraction module is to obtain a discriminative representation from both remote-sensing imagery and the street-level traffic signals, where the corresponding outputs are made in a dual-stream structure that captures complementary characteristics- that is, spatial and contextual patterns resulting from aerial view observation, and flow variations over time derived from the road sensors. Denote the feature dimensionalities of the image and the sensor modalities as d_I and d_S, respectively. The output driven through these 2 branches will, at some point, be combined into a common embedding space for congestion reasoning. In this section, we present the mathematical formulations of the two encoders and define all the variables involved. Module-level data flow is detailed in Figure 3.

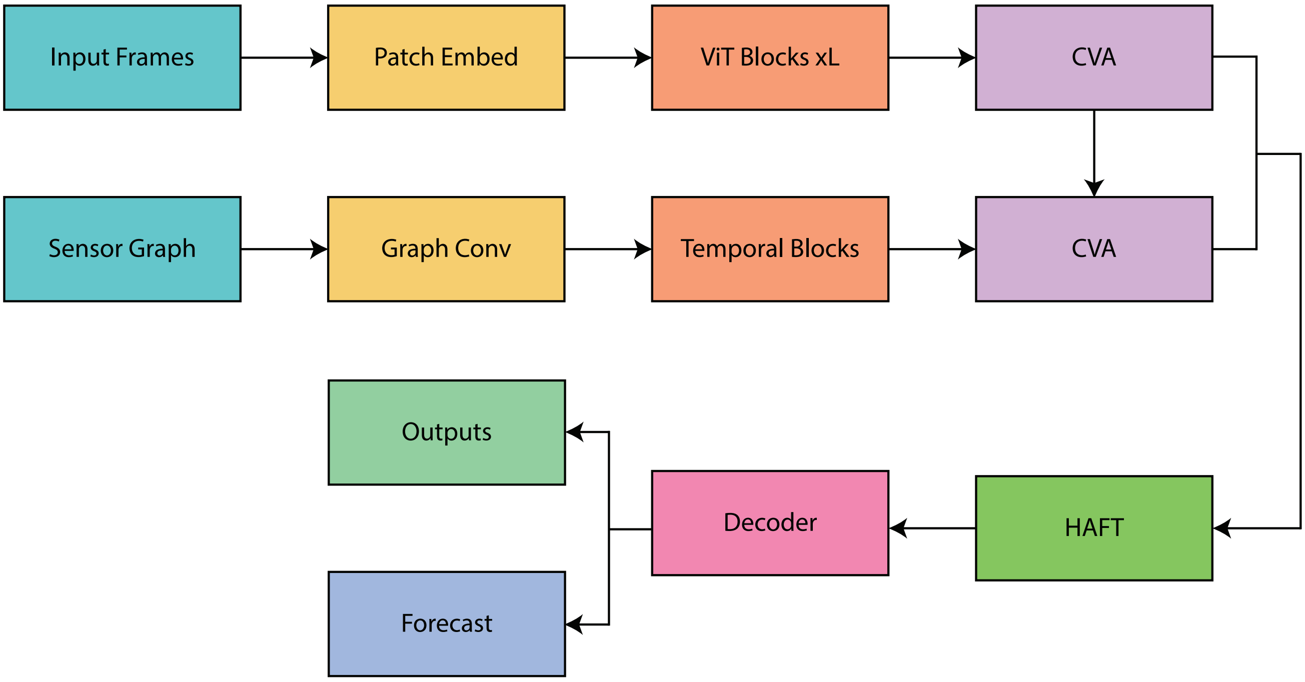

Figure 3. Detailed block flow of CVF-Net: patch embedding and ViT blocks (image stream), graph convolution and temporal blocks (sensor stream), CVA/HAFT fusion, decoder, and outputs.

The following equations describe the visual feature extraction pipeline, which follows standard convolutional and transformer-based design principles to capture multi-scale spatial context from remote-sensing imagery. The remote sensing part employs a multi-scale Residual-Inception CNN to model fine-scale spatial structures such as road edges, car clusters, and urban-texture variance. Given an image tile $I_t \in \mathbb{R}^{H\times W\times C}$, the convolution stage applies $\mathbf{W}ll = 1^L$ convolutional filters producing L intermediate feature maps $\mathbf{F}_l$ such defined as

$$\mathbf{F}_l = \sigma(\mathbf{W}l * \mathbf{F}l - 1 + \mathbf{b}_l), \quad \mathbf{F}_0 = I_t, \tag{3.6}$$

where $*$ represents 2D convolution, $\mathbf{b}_l$ represents a bias vector, and $\sigma(\cdot)$ represents the relu activation function. To solve the vanishing gradient problem, residual skip connections are adopted between non-adjacent layers:

$$\mathbf{F}l' = \mathbf{F}l + \mathcal{R}(\mathbf{F}_{l-2}), \tag{3.7}$$

where $\mathcal{R}(\cdot)$ corresponds to a 1×1 convolution projection that matches the channel dimensions. The Inception architecture operates by constructing multiple receptive fields from parallel convolutions of 1, 3, and 5 different sizes, enabling the extraction of both local and global spatial context simultaneously. The final multi-scale hierarchical features are then flattened into a sequence of tokens, $\mathbf{Z}_I \in \mathbb{R}^{N_p \times d_p}$, where N_p indicates the number of image patches and d_p indicates the embedding dimension of the patches. Subsequently, we employ a ViT to model long-range spatial dependencies between all of the image patches as follows:

$$\mathbf{Z}_I' = \text{softmax}!\left(\frac{\mathbf{Q}_I\mathbf{K}_I^\top}{\sqrt{d_p}}\right)\mathbf{V}_I, \tag{3.8}$$

where $\mathbf{Q}_I = \mathbf{Z}_I\mathbf{W}_Q$, $\mathbf{K}_I = \mathbf{Z}_I\mathbf{W}_K$, and $\mathbf{V}_I = \mathbf{Z}_I\mathbf{W}_V$ represent the query, key, and value projections via trainable matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V \in \mathbb{R}^{d_p \times d_p}$. The final image-level representation is generated from the average pooling of all patch tokens:

$$\mathbf{f}I = \frac{1}{N_p}\sum i = 1^{N_p}\mathbf{Z}_{I,i}' \in \mathbb{R}^{d_I}. \tag{3.9}$$

This representation, $\mathbf{f}_I$, incorporates spatial context, including road geometry, land-use type, and vehicle density, which together represent macro traffic congestion cues. We also record the intermediate self-attention weights from Eq (3.8) as self-attention interpretability maps to help visualize spatial importance regions post-hoc.

The graph-temporal encoder adopts established graph convolutional and recurrent mechanisms to model spatial dependencies among road segments and the temporal evolution of traffic conditions. The street-level branch processes sensor readings in parallel to capture the temporal and topological properties of the road network. Let $G = (V, E)$ denote the graph of sensor nodes, where $|V| = N$ indicates the number of sensors V. Each node $v_i \in V$ is related to a feature vector $\mathbf{s}_{v_i}^t \in \mathbb{R}^3$ we defined previously in Eq (3.1). The input to the temporal graph network is a tensor $\mathbf{S}_t \in \mathbb{R}^{N \times F \times T}$, where $F = 3$ features per node and T is the temporal window length. A graph convolutional operation aggregates neighboring nodes' spatial information, according to

$$\mathbf{H}^{(l+1)} = \sigma!\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \tag{3.10}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}N$ is the adjacency matrix with added self-loops, $\tilde{\mathbf{D}}ii = \sum_j \tilde{\mathbf{A}}_{ij}$ is the resulting degree matrix, $\mathbf{W}^{(l)}$ are the learnable weights, and $\sigma(\cdot)$ is a non-linear activation. A gated recurrent unit (GRU) accommodates temporal dependencies by integrating historical states across time as follows:

$$\mathbf{h}_v^t = (1 - \mathbf{z}_v^t) \odot \mathbf{h}_v^{t-1} + \mathbf{z}_v^t \odot \tanh(\mathbf{W}_h[\mathbf{r}_v^t \odot \mathbf{h}_v^{t-1}, , \mathbf{x}_v^t]), \tag{3.11}$$

where $\mathbf{z}_v^t$ and $\mathbf{r}_v^t$ represent the update and reset gates, respectively, and $\odot$ denotes elementwise (Hadamard) product. The final collection-level embedding is computed with attention-based pooling:

$$\mathbf{f}S = \sum v \in V\alpha_v \mathbf{h}_v^t, \quad \alpha_v = \frac{\exp(\mathbf{w}_a^\top \tanh(\mathbf{W}_a \mathbf{h}v^t))}{\sum u \in V \exp(\mathbf{w}_a^\top \tanh(\mathbf{W}_a \mathbf{h}_u^t))}. \tag{3.12}$$

In this case, $\mathbf{w}_a$ and $\mathbf{W}_a$ denote attention parameters, while α_v represents the relative importance of node v in the entire network. The embedding $\mathbf{f}_S \in \mathbb{R}^{ds}$ represents the dynamic graph embedding with respect to changing traffic states, including congestion propagation, queue spillback, and inter-road dependencies. Therefore, $\mathbf{f}_I$ in Eq (3.9) and $\mathbf{f}_S$ in Eq (3.12) collectively provide a rich multimodal representation, which is merged at the next cross-view fusion stage. A TGNN processes over G with node features $\mathbf{h}_v^t = \mathbf{W}\mathbf{s}_v^t$ and gated temporal updates:

$$\tilde{\mathbf{h}}v^t = \sigma!\Big(!! \sum u \in \mathcal{N}(v)!!\alpha_{uv}, \mathbf{W}_g \mathbf{h}_u^{t-1}\Big), \quad \mathbf{f}S = \mathrm{GRU}(\tilde{\mathbf{h}}\cdot^{1:T}) \in \mathbb{R}^{|V|\times d_S}, \tag{3.13}$$

where α_{uv} are attention weights over edges $(u, v)! \in !E$. The TGNN models both spatial coupling and temporal dynamics of congestion propagation.

3.3. Hierarchical cross-view fusion

Compared to representative cross-modal urban analysis models that perform early feature concatenation, parallel modality encoding, or single-stage attention fusion, HAFT introduces a hierarchical fusion strategy in which cross-view alignment, temporal dependency modeling, and graph-topological refinement are performed sequentially and interactively. This design enables visual context to directly influence temporal graph representations, rather than being treated as an auxiliary or static feature stream. The HAFT is CVF-Net's defining architecture. In contrast to earlier multimodal or GNN-based traffic models that use methods such as early concatenation and simple parallel combinations to merge multimodal information, HAFT emphasizes how aerial visual context and the temporal nature of graph representations interact. Specifically, CVF-Net introduces a new form of CVA in which visual features modulate graph-based traffic embeddings, followed by temporal transformer modeling and a refinement step in which fused representations are re-injected into the road network as output. This hierarchical method supports incorporating interactions between spatial, temporal, and topological dependencies into a single model framework. The HAFT is a proposed architecture that integrates complementary representations from the remote-sensing encoder and the street-level encoder into a shared latent space. The fused representation preserves the spatial context available in aerial imagery and the temporal dynamics available in the sensor deployments. For ease of notation, let $\mathbf{f}_I \in \mathbb{R}^{d_I}$ and $\mathbf{f}_S \in \mathbb{R}^{ds}$ denote the modality specific parameterizations constructed independent. The representations will be projected linearly into a common feature dimension of dimension d, such that

$$\mathbf{f}_I' = \mathbf{W}_I \mathbf{f}_I + \mathbf{b}_I, \qquad \mathbf{f}_S' = \mathbf{W}_S \mathbf{f}_S + \mathbf{b}_S, \tag{3.14}$$

where $\mathbf{W}_I, \mathbf{W}_S \in \mathbb{R}^{d\times d_I}$ and $\mathbb{R}^{d\times ds}$ are trainable projection matrices, and $\mathbf{b}_I$, $\mathbf{b}_S$ are bias terms. The first level of fusion, called CVA, learns how features from one modality modulate the other. Let $f_I \in \mathbb{R}^{d_I}$ denote the global image-level embedding produced by the visual encoder and $f_S \in \mathbb{R}^{ds}$ denote the aggregated graph-temporal embedding produced by the TGNN. To enable cross-modal interaction, both representations are first projected into a shared latent space of dimensionality d, yielding $f_I' =$

$W_I f_I + b_I$ and $f'_S = W_S f_S + b_S$, where W_I and W_S are learnable projection matrices and b_I and b_S are bias terms. Specifically, the image features serve as query vectors while the street-level features serve as keys and values:

$$Q_I = \mathbf{f}'_I \mathbf{W}_Q, \quad K_S = \mathbf{f}'_S \mathbf{W}_K, \quad V_S = \mathbf{f}'_S \mathbf{W}_V, \tag{3.15}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable matrices. Here, $f_I \in \mathbb{R}^{d_I}$ denotes the global image-level embedding produced by the visual encoder, and $f_S \in \mathbb{R}^{ds}$ denotes the aggregated graph-temporal embedding produced by the TGNN. To enable cross-modal interaction, both representations are first projected into a shared latent space of dimensionality d, yielding $f'_I = W_I f_I + b_I$ and $f'_S = W_S f_S + b_S$, where W_I and W_S are learnable projection matrices and b_I and b_S are bias terms. This asymmetric construction ensures that attention is computed across modalities rather than within a single modality. From these projected features, the query vectors Q_I are derived exclusively from the visual representation f'_I, while the key and value vectors K_S and V_S are derived exclusively from the graph-temporal representation f'_S. At the first fusion level, CVA is employed to explicitly align visual and street-level representations, enabling the model to emphasize road segments whose temporal dynamics are most relevant to the surrounding visual context. The inter-modal attention mechanism is then defined as

$$\mathbf{z} = \text{softmax!}\left(\frac{Q_I K_S^\top}{\sqrt{d}}\right) V_S, \tag{3.16}$$

which yields cross-view enriched embeddings $\mathbf{z} \in \mathbb{R}^d$ encoding mutual dependencies between both modalities. Although the mathematical form of Eq (3.16) resembles the scaled dot-product attention used in Transformer architectures, the proposed CVA differs fundamentally from standard self-attention. In conventional self-attention, queries, keys, and values are derived from the same feature sequence and used to model intra-modal dependencies. In contrast, CVA performs explicit cross-modal attention, where visual features attend to graph-temporal representations, enabling spatial context from aerial imagery to selectively modulate traffic dynamics encoded in the sensor graph.

The attention weights quantify the contribution of each street-level feature to the remote-sensing context, thereby allowing the model to highlight road regions with stronger congestion relevance. The CVA module implicitly resolves occlusions and visual ambiguities in street-level imagery via an Attention Fusion mechanism. Weak correlations between visual tokens corrupted by occlusion or viewpoint artifacts and their corresponding graph-temporal features result in lower attention weights for these tokens in the CVA formulation. This results in greater reliance on temporally stable sensor signals from the TGNN Branch and a naturally lower weighting of unreliable visual cues. As a result of this adaptive weighting, CVF-Net achieves robust fusion behaviour in scenes with partial occlusions, without explicit occlusion detection or masking. The motivation behind this design is to allow high-level spatial cues, such as road geometry, intersection density, and visually observable congestion patterns, to guide the interpretation of temporally evolving traffic signals. Visual tokens that are weakly correlated with sensor dynamics naturally receive lower attention weights, which improves robustness in the presence of occlusions or visual ambiguities.

The fused cross-view embeddings are then processed by a spatiotemporal transformer to capture short- and long-range temporal dependencies governing congestion evolution. The second component, the Spatiotemporal Transformer (STT), extends the fused embeddings over time to model temporal evolution. Let $\mathbf{z}_{1:T} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}T]$ denote the sequence of fused representations across T time steps. Each embedding is augmented with a temporal positional encoding $\mathbf{P}\text{time} \in \mathbb{R}^{T\times d}$ that provides sequential ordering information. The multi-head self-attention operation computes inter-temporal dependencies as

$$\mathbf{z}'_{1:T} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_O, \tag{3.17}$$

where each attention head is formulated as

$$\text{head}i = \text{softmax}! \left(\frac{(\mathbf{z}1 : T\mathbf{W}Q^{(i)})(\mathbf{z}1 : T\mathbf{W}K^{(i)})^\top}{\sqrt{d/h}} \right)(\mathbf{z}1 : T\mathbf{W}_V^{(i)}). \tag{3.18}$$

Furthermore, $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d\times(d/h)}$ are the head-wise projection matrices whereas $\mathbf{W}_O \in \mathbb{R}^{d\times d}$ projects the concatenated outputs back into the embedding space. It is important to note that CVA operates as a modality-bridging mechanism prior to temporal modeling. The fused representation produced by CVA is subsequently processed by the spatiotemporal transformer to capture inter-temporal dependencies, rather than replacing conventional temporal self-attention.

To ensure topological consistency and spatial smoothness, the temporally fused representations are subsequently reintegrated into the road network graph via a graph-refinement module that propagates cross-view context along connected road segments. The transformer block can learn both short- and long-term temporal patterns, enabling the model to forecast changing congestion behavior in dynamic settings. The final stage, GR, injects the fused temporal context back into the graph for topological consistency. Denote $\bar{\mathbf{z}}' = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}'_t$ to be the temporal average of the fused embedding. For each road node v, we update its hidden state using a residual graph convolution operation:

$$\mathbf{h}'v = \mathbf{h}v + \text{ReLU}! \left(\sum u \in \mathcal{N}(v)\beta uv\mathbf{W}_r[\mathbf{h}_u \| \bar{\mathbf{z}}'] \right), \tag{3.19}$$

where $\|$ is vector concatenation, $\mathbf{W}r$ is a learnable transformation matrix with size $d' \times 2d$, and βuv are normalized normalized direction-aware attention weights of the form:

$$\beta_{uv} = \frac{\exp(\mathbf{a}^\top \tanh(\mathbf{W}_a[\mathbf{h}_u \| \mathbf{h}v]))}{\sum k \in \mathcal{N}(v)\exp(\mathbf{a}^\top \tanh(\mathbf{W}_a[\mathbf{h}_k \| \mathbf{h}_v]))}, \tag{3.20}$$

with $\mathbf{a} \in \mathbb{R}^{d_a}$ and $\mathbf{W}a \in \mathbb{R}^{d_a\times 2d}$ denoting attention parameters. This refinement step imposes local spatial smoothness while ensuring long-range temporal dependencies are preserved from $\mathbf{z}'1 : T$. Therefore, the updated node embeddings $\mathbf{h}'_v$ jointly represent cross-view, spatiotemporal, and relational information to produce a single latent representation used as input for the final congestion prediction layer. The spatio-temporal fusion pipeline is summarized in Figure 4.

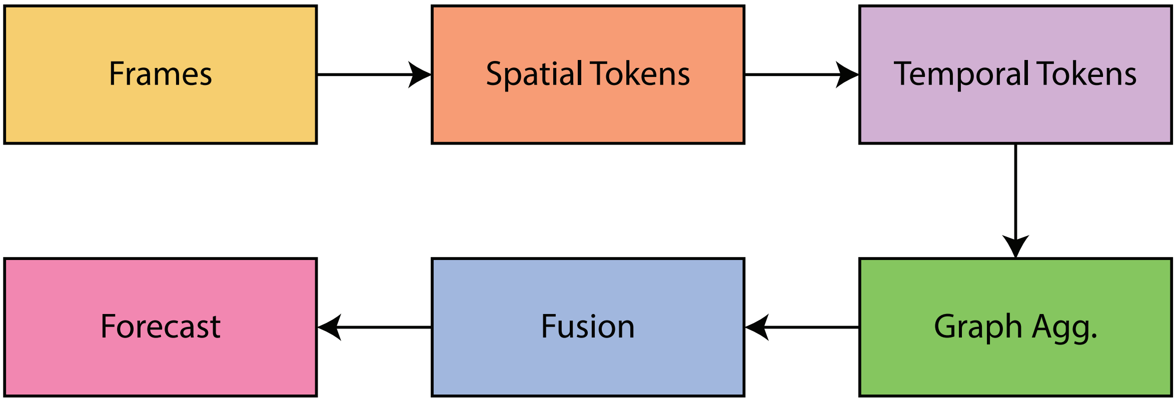

Figure 4. Pipeline from frames to spatial/temporal tokens, graph aggregation, hierarchical fusion, and final forecast.

Through this hierarchical fusion process, CVF-Net departs from flat multimodal fusion strategies by tightly coupling visual perception and graph-temporal reasoning. The resulting node embeddings jointly encode aerial spatial context, temporal traffic dynamics, and road-network topology, forming a unified representation that is subsequently used for congestion classification and forecasting.

3.4. Congestion prediction and map reconstruction

In the last step of the proposed framework, the refined graph embeddings are translated into congestion states and short-horizon traffic predictions. Each node embedding $\mathbf{h}'_v \in \mathbb{R}^{d'}$, in the hierarchical fusion stage, incorporates context from both remote-sensing and street-level modalities. The prediction head has two branches that run in parallel: one for categorical classification of congestion and one for numerical predictions of speed. The categorical classification branch maps $\mathbf{h}'_v$ to a probability distribution over discrete congestion levels with a two-layer perceptron with a nonlinear activation function:

$$\hat{\mathbf{p}}_v = \text{softmax}!\left(\mathbf{W}_2, \sigma(\mathbf{W}_1\mathbf{h}'_v + \mathbf{b}_1) + \mathbf{b}_2\right), \tag{3.21}$$

where $\mathbf{W}1 \in \mathbb{R}^{d_m \times d'}$ and $\mathbf{W}2 \in \mathbb{R}^{C \times d_m}$ are trainable matrices, $\mathbf{b}1$ and $\mathbf{b}2$ are bias directions, $\sigma(\cdot)$ is the ReLU activation (rectifier linear unit) function, and C is the number of congestion categories (i.e., free-flow, moderate and heavy). The predicting (or predicted) vector $\hat{\mathbf{p}}v = [pv, 1, pv, 2, \ldots, pv, C]^\top$ such that $\sum c = 1^C pv, c = 1$ and the regression branch predicts a short-horizon speed forecast at time $t + \tau$ as

$$\hat{s}_v^{t+\tau} = \psi(\mathbf{h}'_v) = \mathbf{w}_r^\top \mathbf{h}'_v + b_r, \qquad \tau \in 15, 30, 60 \text{ minutes}, \tag{3.22}$$

where $\psi(\cdot)$ represents a linear mapping indexed by $\mathbf{w}_r$ and b_r. The multi-task learning framework enables the model to optimize both categorical and continuous objectives simultaneously, improving generalizability across scenarios with varying traffic densities. A graph Laplacian regularization term is introduced to encourage spatial continuity among neighboring road segments by minimizing the difference in predicted congestion probabilities at connected nodes. This allows for smooth transitions across the road network:

$$\mathcal{L}\text{smooth} = \sum (u, v) \in E|\hat{\mathbf{p}}_u - \hat{\mathbf{p}}_v|_2^2. \tag{3.23}$$

In this context, E is the set of edges that encodes connectivity in the road graph. The smoothness loss term preferentially penalizes large differences in predictions between adjacent segments, which helps maintain realistic continuity of congestion across roadways. This term also enforces topological consistency by aligning predictions with the map structure of G. The ultimate objective of learning comprises three complementary components: categorical cross-entropy for classifying congestion, L_1 regression loss for predicting speed numerically, and a Laplacian regularization term for preserving spatial coherence. The combined loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{\mathbf{p}}, \mathbf{y}) + \lambda_1 \,||\hat{\mathbf{s}} - \mathbf{s}||_1 + \lambda_2 \mathcal{L}_{\text{smooth}}, \tag{3.24}$$

where $\mathbf{y}$ represents the ground-truth congestion labels, $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the actual and predicted average speeds, respectively, and λ_1, λ_2 are scalar weighting coefficients that determine the balancing of each auxiliary loss term. The parameters λ_1 and λ_2 have been chosen heuristically to balance classification accuracy and spatial consistency. During the training phase, gradients from Eq (3.24) are backpropagated through both branches, enabling the multimodal network to optimize end-to-end.

Upon completing the training process, the estimated congestion probabilities $\hat{\mathbf{p}}_v$ are projected back to geographic coordinates to generate interpretable traffic maps. The reconstruction function projects each node's state to the spatial footprint of that node, Ω_v, within the geographically-referenced image plane. The projected continuous probability values are transformed into congestion intensity levels using a colormap function $\Gamma(\cdot)$:

$$\mathcal{M}(x, y) = \Gamma! \left(\sum_{v \in V} \hat{p}v, c^*, \mathbf{1}(x, y) \in \Omega_v \right), \tag{3.25}$$

where $c^* = \arg\max_c p_{v,c}$ represents the most probable congestion class for segment $v$and $\mathbf{1}_{(x,y)\in\Omega_v}$ is an indicator function enabling pixels under road segment Ω_v. Therefore, the reconstructed congestion map $\mathcal{M}(x, y)$ provides an intuitive way to visualize how the predicted traffic state will be spatially distributed, enabling a side-by-side comparison with observed ground-truth or satellite-estimated congestion indices. At this stage, the model provides a city-wide interpretation of real-time congestion intensity across both pattern seeing and temporal outputs.

3.5. Training protocol and evaluation

An end-to-end optimization of the proposed multimodal framework is performed using the AdamW optimizer, which incorporates decoupled weight decay for better generalization. The learning rate is initialized at 10^{-4} and adjusted using a cosine decay schedule to ensure smooth convergence. The model is trained for up to 150 epochs with a batch size of 64, including early stopping based on the minimum validation Mean Absolute Error (MAE). The training loss is computed using Eq (3.24), combining classification, regression, and spatial regularization terms. Gradient clipping with a threshold of 1 is applied to maintain stability in the transformer layers. The remote-sensing (visual) encoder is warm-started with weights pretrained on CityFlowV2 cropped patches, and the TGNN is initialized with weights pretrained on METR-LA and PEMS-BAY for speed forecasting. This multi-stage initialization improves convergence speed and stabilizes multimodal alignment. The fusion and prediction layers are fine-tuned jointly across datasets in a single training loop.

As part of the training process, each minibatch is defined as synchronized cross-view samples $(I_t, \mathbf{S}_t)$ selected at random from spatial regions and temporal windows to guarantee balanced learning. The training objective can be mathematically stated as

$$\mathcal{L}\text{train} = \frac{1}{B}\sum i = 1^{B}\Big[\mathcal{L}_{\text{CE}}^{(i)} + \lambda_1|\hat{\mathbf{s}}^{(i)} - \mathbf{s}^{(i)}|1 + \lambda_2\mathcal{L}\text{smooth}^{(i)}\Big]. \tag{3.26}$$

Here, B represents the batch size, $\mathcal{L}_{\text{CE}}^{(i)}$ denotes the categorical cross-entropy loss corresponding to congestion classification, and $\hat{\mathbf{s}}^{(i)}$ and $\mathbf{s}^{(i)}$ are the predicted and ground truth speed vectors, respectively. The model has approximately 25.3 M trainable parameters, allocated as follows: CNN–ViT backbone (47%), TGNN module (33%), and fusion–prediction head (20%). The model is implemented in PyTorch 2.2, and all experiments are run on an NVIDIA RTX A6000 GPU with 48 GB of memory, which allows processing of 256 × 256 patches with 12 temporal frames per sample as input.

To mitigate overfitting and improve generalization across cities, several regularization methods are implemented. Dropout with probability $p = 0.2$ is added after the transformer feed-forward layers, and batch normalization is adopted in all the convolutional and graph layers. To enhance visual diversity, we apply stochastic data augmentation to remote-sensing tiles, including random rotation $(\pm 15^{circ})$, color jitter, and horizontal flipping. On the sensor side, to simulate sensor inaccuracies, we add Gaussian noise with variance $\sigma^2 = 0.01$ to the traffic measurements. To further assess the model's robustness, 5-fold cross-validation is performed over temporal splits, ensuring each fold corresponds to a different week of observation. We use standard performance metrics for evaluation appropriate to the dual-path nature of the task. For the classification branch, we use the standard cross-view tracking continuity, known as the IDF1 metric, defined as

$$\text{IDF1} = \frac{2 \times \text{IDTP}}{2 \times \text{IDTP} + \text{IDFP} + \text{IDFN}}, \tag{3.27}$$

where IDTP, IDFP, and IDFN represent the number of correctly identified, false positive, and false negative tracks across views, respectively. This metric quantifies whether the model associates vehicle trajectories consistently across views, which directly reflects multi-camera congestion detection performance. We compute the following three metrics for the forecasting branch compared to observed speed $\hat{s}_v^t$:

$$\text{MAE} = \frac{1}{N}\sum_{v=1}^{N}|s_v^t - \hat{s}_v^t|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum v = 1^{N}(s_v^t - \hat{s}_v^t)^2}, \quad \text{MAPE} = \frac{100}{N}\sum v = 1^{N}\left|\frac{s_v^t - \hat{s}_v^t}{s_v^t}\right|. \tag{3.28}$$

In this case, N represents the number of road segments analyzed. The MAE estimates the average absolute error of the predictions, whereas the RMSE penalizes larger deviations from the true speed more than the MAE. The MAPE will normalize deviations from true speed, making it more interpretable across different traffic speed levels.

The evaluation protocol uses three benchmark datasets with complementary characteristics. CityFlowV2 provides citywide-scale, multi-camera images for visual congestion cues to investigate object-level consistency and congestion clustering via tracking. METR-LA and PEMS-BAY are standard benchmark datasets for temporal forecasting that consist of long-term loop-detector measurements of traffic speed, flow, and occupancy. To ensure fairness, we continue to use the same train/validation/test splits in prior work, with 70% used for training, 10% for validation, and 20% for testing. Our ablation study experiment is done to evaluate the contribution of our major components - (i) the remote-sensing encoder, (ii) the TGNN, and (iii) the CVA block. Each component is removed one at a time and evaluated based on how its absence affects the model's prediction accuracy, spatial coherence, and stability. This gives rise to a design for training and evaluation that is thorough enough to be reproducible and allows for careful comparison with state-of-the-art methods across all evaluated datasets.

From an implementation perspective, the attention weights learned by the CVA module are normalized across modalities, which further stabilizes fusion under variable visual quality conditions. During training, samples affected by severe occlusions or temporary sensor noise are implicitly regularized through batch normalization, dropout, and the multi-task learning objective, preventing over-reliance on any single modality. From a generalization perspective, CVF-Net is trained and evaluated using a unified architecture and shared fusion strategy across all datasets, without dataset-specific architectural tuning. The use of cross-dataset pretraining for the visual and graph-temporal encoders, combined with joint fine-tuning of the fusion and prediction layers, encourages the model to learn transferable representations that are less dependent on city-specific characteristics.

4. Experimental results

4.1. Datasets

The efficacy of the proposed framework is assessed on three benchmark datasets, including CityFlowV2 [25], METR-LA [26], and PEMS-BAY [27]. Collectively, the three datasets provide complementary information on traffic congestion, facilitating an assessment of the spatiotemporal dimensions of the proposed model. CityFlowV2 provides visual information derived from multi-camera street-level views, while the METR-LA and PEMS-BAY datasets contain large amounts of spatiotemporal sensor data derived from urban highway networks. The combination of these datasets enables the evaluation of CVF-Net across different sensing modalities, ranging from overhead remote-sensing imagery to sequential traffic measurements.

CityFlowV2 is a large-scale urban traffic dataset designed for the AI City Challenge and is used to evaluate multi-camera tracking and congestion analysis. It contains synchronized video streams from 40 cameras positioned across 10 intersections within a 2-square-kilometer area. The dataset provides more than 200,000 annotated vehicle trajectories collected from 56 hours of video footage recorded at 10 frames per second. Each camera captures high-resolution scenes of dynamic urban traffic under varying illumination, weather, and density conditions. For this study, consecutive frames are grouped into fifteen-second clips, and the corresponding vehicle bounding boxes are reprojected into geo-coordinates using known camera parameters. These frames are further aligned with overhead remote-sensing imagery to establish cross-view correspondence between aerial and ground-level perspectives. CityFlowV2, therefore, serves as a critical dataset for validating the visual component of CVF-Net and assessing the quality of congestion detection across multiple viewpoints.

METR-LA is a graph-based traffic dataset collected from 207 loop detectors installed along major freeways in the Los Angeles County highway network. The data cover a continuous four-month period, from March to June 2012, with a sampling interval of five minutes. Each sensor reports three key attributes: average speed, traffic flow, and occupancy, forming a spatiotemporal graph where nodes correspond to sensors and edges represent physical road connections. Edge weights are defined as the inverse of geographical distances and allow message passing among neighboring segments to align with real-world connectivity. Regarding the model evaluation, the previous 60-minute partition forecasts traffic conditions up to 60 minutes, which aligns with the 15-, 30-, and 60-minute prediction horizons. METR-LA is a representative medium-scale dataset for evaluating the temporal predictive capabilities of the graph-based module within CVF-Net.

PEMS-BAY is from the California Performance Measurement system and serves as another benchmark for a traffic forecasting project. There are three hundred and twenty-five loop detectors placed on highways around the San Francisco Bay Area. The dataset consists of six months of continuous measurements taken at five-minute intervals, totaling over five million data samples. Each measurement includes the same three metrics as METR-LA, namely average speed, flow rate, and occupancy. The graph representation results in a very dense topology and is more complex than METR-LA, with overlapping highways and multiple congestion hotspots. This dataset is an important test case for assessing the scalability and robustness of spatiotemporal learning techniques, as PEMS-BAY exhibits much greater spatial heterogeneity and nonlinear congestion propagation than METR-LA. Table 1 summarizes the main characteristics of all three datasets, including spatial coverage, modality type, number of sensors or cameras, temporal resolution, and duration.

Table 1. Overview of datasets used for model evaluation.

Dataset	Type	Samples	Resolution
CityFlowV2	Multi-camera videos	56 hours	10 fps
METR-LA	Loop-detector sensors	4 months	5 min
PEMS-BAY	Loop-detector sensors	6 months	5 min

4.2. Experimental setup

To ensure fair, reproducible assessment, all experiments use a consistent setup. CVF-Net is implemented in PyTorch 2.2 with CUDA optimizations. Training and inference use an NVIDIA RTX A6000 GPU (48 GB) and mixed-precision compute for efficiency. Model parameters use Xavier uniform initialization. Training batches are 64, for up to 150 epochs. AdamW optimizer starts with a 10^{-4} learning rate, 10^{-5} weight decay, and a cosine annealing scheduler. Early stopping halts training if validation loss does not improve for 10 epochs, reducing overfitting and supporting stable convergence.

The training procedure consists of two distinct stages. In the first stage, the remote-sensing encoder (a CNN–ViT hybrid, abbreviated CNN–ViT) and the traffic graph neural network (TGNN) components are trained independently to produce domain-specific representations. In this stage, the visual encoder is trained on CityFlowV2 image patches for visual feature extraction, and the TGNN is trained as an independent component using the METR-LA and PEMS-BAY datasets for temporal traffic prediction. In the second stage, the complete architecture, including the hierarchical attention fusion transformer (HAFT) and the prediction head, is jointly fine-tuned using the composite loss function. As a whole, this staged training procedure enables faster convergence and improved cross-modality alignment between the visual and sensor domains. To maintain numerical stability during backpropagation, gradient clipping is performed with a threshold of 1.0.

Each input sample comprises an image tile, denoted I_t, together with its associated traffic measurement tensor, $\mathbf{S}_t$. For both the METR-LA and PEMS-BAY datasets, the temporal window length is set to $T = 12$, corresponding to one hour of historical data sampled at five-minute intervals. The prediction horizons are designated as $\tau \in \{15, 30, 60\}$ minutes, in order to evaluate performance at short, medium, and long-range predictions. In the CityFlowV2 dataset, visual frames are sampled every 10 frames to balance temporal coverage and computational cost, and optical flow features are computed to capture vehicle motion patterns. As part of preprocessing, all image inputs are resized to 256×256 pixels, and their pixel values are normalized to $[0, 1]$. Data augmentation techniques are employed to broaden generalization, including random rotation by 15 degrees, random horizontal flipping, and brightness adjustment within 10%.

In the fine-tuning stage, the loss coefficients are initialized to empirically derived values: $\lambda_1 = 0.6$ for the speed regression term and $\lambda_2 = 0.4$ for the spatial smoothness term, to balance prediction accuracy and spatial coherence. Each epoch consists of 500 gradient updates, and each model checkpoint is saved when the validation MAE is lowest. After training is complete, each model is evaluated again on the test split of each dataset, in the same manner as described in Section 3.5. Performance will be quantitatively reported as the average and standard deviation across three independent runs with different random seeds to account for stochastic dynamics in optimization and data sampling. This ensures that the results are reliable and comparable to previous state-of-the-art baselines on the same benchmarks. All experiments were conducted under identical hardware and training settings to ensure fair comparison of both predictive performance and computational efficiency.

4.3. Quantitative results

This section presents the quantitative assessment of the proposed CVF-Net on three benchmark datasets: CityFlowV2, METR-LA, and PEMS-BAY, while comparing against five state-of-the-art benchmarks proposed between 2021 and 2025. Metrics have been selected to assess both the multi-camera tracking performance and the spatiotemporal forecasting capabilities. Reported metrics for congestion analysis incorporate IDF1 (Identification F1 Score), IDP (Identification Precision), and IDR (Identification Recall). Reported forecasting metrics include MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and R^2 (coefficient of determination). All results are averaged over three runs with different random seeds, and standard deviations are reported. Across all datasets, CVF-Net demonstrates consistent superiority, achieving the lowest forecasting error and the highest identification accuracy.

As indicated in Table 2, the results reflect performance improvements of the proposed CVF-Net over recent state-of-the-art methods on the CityFlowV2 benchmark. The proposed CVF-Net achieves an IDF1 score of 86.92±0.48%, indicating a performance improvement over the best previously reported result in the AICity Track-1 (2nd place, 2022) by nearly 2.55%. This improvement highlights CVF-Net's enhanced capacity to achieve higher trajectory association accuracy and maintain better tracking consistency across multiple camera views. Equally, the proposed model achieves higher IDP (85.44 ± 0.51%) and IDR (88.41 ± 0.44%) values than all competing methods, demonstrating improved tracking precision and recall for multi-object re-identification and tracking continuity. The

MOTA score (82.17±0.53%) further confirms the robustness of CVF-Net in addressing identity swaps, missed detections, and false positives under complex city-scale traffic scenarios. In contrast, traditional self-supervised or online MTMC methods (e.g., the Self-Supervised Camera-Link Model (2024) or Online MTMC Tracker (2025)) exhibit significant performance degradation, underscoring the advantages of the proposed model's cross-view fusion and adaptive spatio-temporal learning mechanisms. CityFlowV2 tracking results are compared in Figure 5.

Table 2. Performance comparison on the CityFlowV2 dataset. Higher values indicate better tracking continuity.

Method	IDF1 (%)	IDP (%)	IDR (%)	MOTA (%)
AICity Challenge [28] (1st place)	80.95 ± 0.62	82.41 ± 0.55	79.62 ± 0.60	76.85 ± 0.71
AICity Track-1 [28] (2nd place)	84.37 ± 0.55	83.90 ± 0.63	84.91 ± 0.52	79.34 ± 0.66
AICity Track-1 [28] (3rd place)	83.71 ± 0.70	82.15 ± 0.74	84.20 ± 0.68	78.26 ± 0.73
Self-Supervised Camera-Link Model [29]	61.07 ± 0.84	60.53 ± 0.91	61.88 ± 0.79	59.22 ± 0.87
Online MTMC Tracker [30]	81.64 ± 0.59	82.10 ± 0.62	81.02 ± 0.56	77.44 ± 0.60
Proposed CVF-Net	**86.92** ± **0.48**	**85.44** ± **0.51**	**88.41** ± **0.44**	**82.17** ± **0.53**

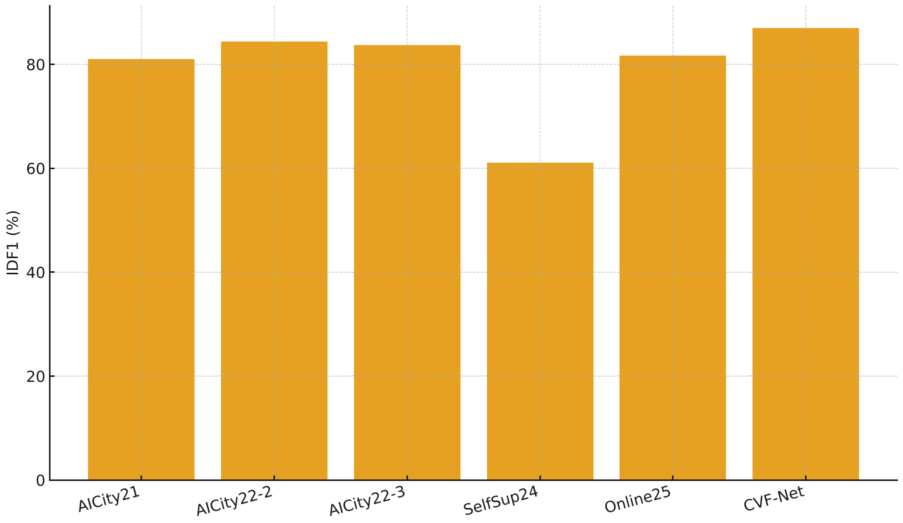

Figure 5. CityFlowV2 comparison of IDF1 across methods, highlighting the gains achieved by CVF-Net.

As presented in Table 3, and in comparison to recent baseline methods, the proposed CVF-Net yields superior predictive accuracy on the METR-LA dataset over the 15-minute forecasting horizon. CVF-Net achieves the lowest MAE (2.31 ± 0.04) and RMSE (3.22 ± 0.07), indicating that both the average and large-magnitude errors were effectively minimized, showcasing one of the strongest performance capabilities of CVF-Net. Additionally, the model reports the lowest MAPE (5.74±0.09%), underscoring its ability to maintain this level of performance even under significant dynamic variations in traffic flow. The R^2 score (0.945 ± 0.005) further characterizes the strong correlation between predicted and actual traffic speeds, significantly improving over all competing methods, including but not limited to ASS-GNN-TA [31] and DyGCN-LSTM [32]. Forecasting error versus horizon on METR-LA is shown in Figure 6.

Table 3. Results on the METR-LA dataset for 15-minute forecasting horizon.

Method	MAE	RMSE	MAPE (%)	R^2
Graph WaveNet [33]	3.57 ± 0.06	3.98 ± 0.08	4.87 ± 0.07	0.912 ± 0.004
GFAGNN [34]	2.69 ± 0.05	5.15 ± 0.06	6.81 ± 0.08	0.931 ± 0.006
DyGCN-LSTM [32]	2.64 ± 0.07	4.64 ± 0.05	6.35 ± 0.09	0.937 ± 0.007
METSformer [31]	2.73 ± 0.06	5.34 ± 0.05	7.10 ± 0.11	0.925 ± 0.006
ASS-GNN-TA [31]	2.45 ± 0.05	3.33 ± 0.07	6.12 ± 0.08	0.939 ± 0.005
Proposed CVF-Net	**2.31** ± **0.04**	**3.22** ± **0.07**	**5.74** ± **0.09**	**0.945** ± **0.005**

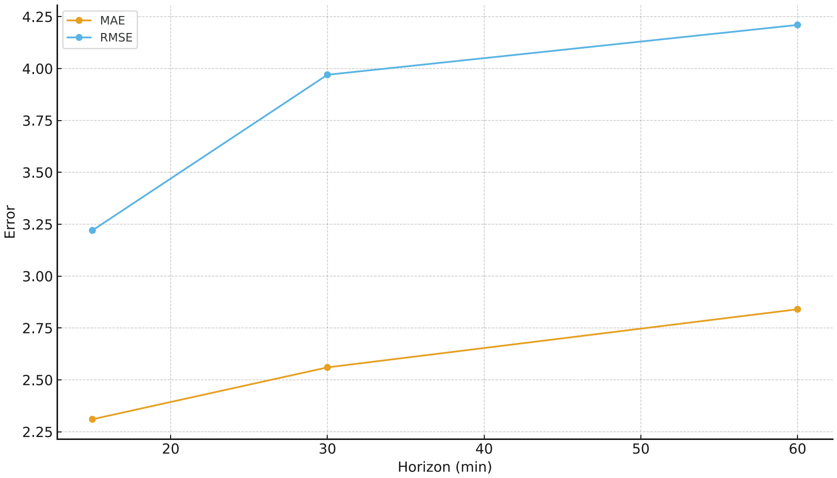

Figure 6. METR-LA forecasting errors (MAE and RMSE) at 15, 30, and 60 minutes.

As illustrated in Table 4, CVF-Net surpassed all competing methods for both 30-minute and 60-minute forecasting horizons on the METR-LA dataset. With respect to the 30-minute forecasting horizon, CVF-Net attained the highest short-term accuracy, yielding the lowest MAE (2.56 ± 0.05) and RMSE (3.97 ± 0.08). The 60-minute forecasting horizon also produced relatively strong performance, with MAE (2.84 ± 0.06) and RMSE (4.21 ± 0.09), demonstrating that it can maintain stability in longer forecasting horizons. Furthermore, CVF-Net achieved lower MAPE values (5.93 ± 0.07% and 6.10 ± 0.08%) and higher R^2 values (0.938 ± 0.005 and 0.932 ± 0.007), which indicates that the forecasts are reliable and accurate.

Table 4. Forecasting performance on METR-LA dataset for 30- and 60-minute horizons.

	30-min Horizon				60-min Horizon			
Method	MAE	RMSE	MAPE (%)	R^2	MAE	RMSE	MAPE (%)	R^2
Graph WaveNet [33]	4.02 ± 0.08	4.71 ± 0.10	7.93 ± 0.13	0.906 ± 0.005	4.66 ± 0.09	5.26 ± 0.12	8.74 ± 0.15	0.891 ± 0.006
GFAGNN [34]	3.18 ± 0.07	4.93 ± 0.08	6.87 ± 0.09	0.914 ± 0.004	3.79 ± 0.06	5.54 ± 0.11	7.42 ± 0.10	0.905 ± 0.006
ASS-GNN-TA [31]	2.78 ± 0.06	4.38 ± 0.09	6.25 ± 0.08	0.926 ± 0.005	3.26 ± 0.07	4.54 ± 0.10	6.68 ± 0.09	0.919 ± 0.007
Proposed CVF-Net	**2.56** ± **0.05**	**3.97** ± **0.08**	**5.93** ± **0.07**	**0.938** ± **0.005**	**2.84** ± **0.06**	**4.21** ± **0.09**	**6.10** ± **0.08**	**0.932** ± **0.007**

As shown in Table 5, CVF-Net exhibits the best forecasting performance on the PEMS-BAY dataset for the 15-minute lead time compared with all models in this study. CVF-Net has the lowest MAE (1.15 ± 0.03) and RMSE (2.56 ± 0.04), indicating that it produces lower average and larger-magnitude prediction errors than other models. Furthermore, it achieves the lowest MAPE (1.82 ± 0.05%) and the highest R^2 (0.962 ± 0.004), indicating that the predicted traffic flow values closely align with the true observations. These improvements indicate that the cross-view fusion and spatio-temporal attention mechanisms in CVF-Net enhance its ability to capture complex traffic dynamics and temporal dependencies more effectively than previous graph- or transformer-based approaches. PEMS-BAY short-term results are summarized in Figure 7.

Table 5. Forecasting performance on PEMS-BAY dataset for 15-minute horizon.

Method	MAE	RMSE	MAPE (%)	R^2
Graph WaveNet [33]	1.39 ± 0.03	2.89 ± 0.04	2.02 ± 0.06	0.947 ± 0.004
GFAGNN [34]	1.31 ± 0.04	2.80 ± 0.05	2.75 ± 0.05	0.951 ± 0.005
METSformer [31]	1.31 ± 0.03	2.80 ± 0.04	2.75 ± 0.06	0.950 ± 0.004
ASS-GNN-TA [31]	1.23 ± 0.02	2.74 ± 0.04	1.98 ± 0.05	0.956 ± 0.005
Proposed CVF-Net	**1.15 ± 0.03**	**2.56 ± 0.04**	**1.82 ± 0.05**	**0.962 ± 0.004**

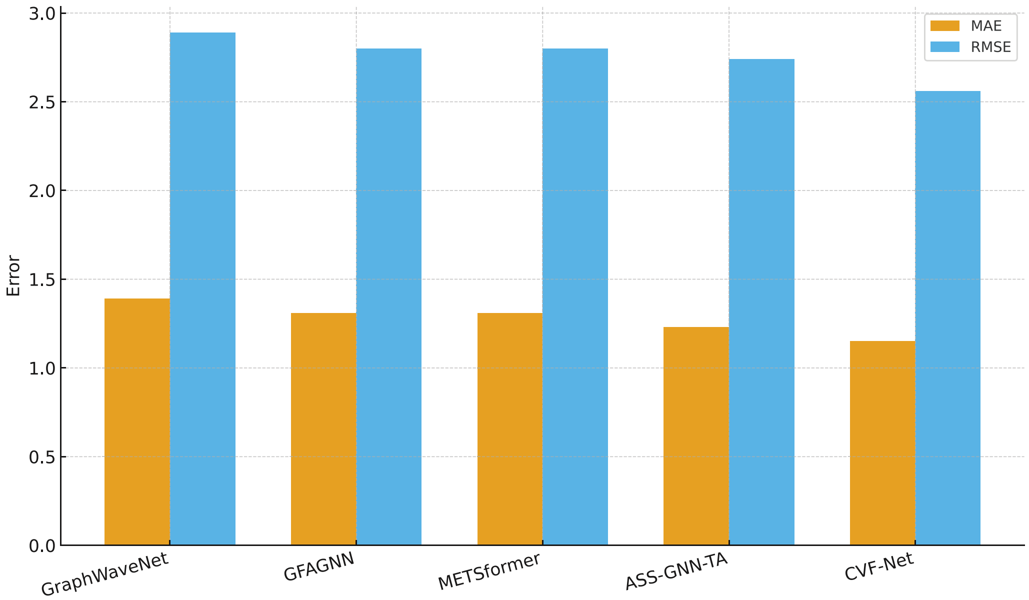

Figure 7. PEMS-BAY 15-minute forecasting: MAE and RMSE comparison across models.

As indicated in Table 6, CVF-Net is the model reporting the best accuracy for forecasting traffic flow in the PEMS-BAY dataset, across both metrics (MAE and RMSE) and forecasting periods (30-minute and 60-minute). For the 30-minute traffic flow forecasting period, CVF-Net achieved the best MAE of 1.24 ± 0.03 and the best root mean squared error of 2.66 ± 0.05, which indicates that CVF-Net has the lowest forecasting error for short-term forecasts when compared to other short-term forecasting models. Likewise, for the 60-minute forecasting period, CVF-Net also had the best MAE of 1.40 ± 0.04 and the best RMSE of 2.85 ± 0.05, indicating greater stability for long-term forecasting. The MAPE values were also the lowest at 1.84 ± 0.06% and 1.96 ± 0.06%, alongside higher R^2 values of 0.958 ± 0.004 and 0.952 ± 0.005, supporting the claim that CVF-Net adequately represents spatio-temporal dependencies and generalizes across different traffic flow contexts.

In Table 7, the results demonstrate that CVF-Net shows consistent and notable improvements across all benchmark datasets when compared to the best-performing baseline models. For traffic forecasting datasets such as METR-LA and PEMS-BAY, CVF-Net achieves average MAE and RMSE improvements of 8.6% and 8.5%, respectively, indicating reductions in both the average prediction error and the error at higher magnitudes. For MAPE, CVF-Net shows an improvement of 10.8%, confirming reduced variability in traffic forecasting accuracy. Regarding correlation strength, CVF-Net shows an average increase in R^2 of 0.65%, indicating that forecasted traffic flow patterns are better aligned with actual traffic flow patterns. On the CityFlowV2 dataset, the 7.0% improvement in IDF1 further confirms that the proposed method enhances continuity in multi-camera vehicle tracking. Aggregated improvements are presented in Figure 8.

The generalization scope has to be explicitly defined in this work; Although CVF-Net has been verified on three common benchmark datasets that represent completely different urban layouts (multi camera city intersections - CityFlow-V2, medium-scale freeway networks - METR-LA and large scale high heterogeneous highway systems - PEMS-BAY), the consistent improvements obtained from all three datasets suggests that the fusion method of CVF-Net generalises across different traffic patterns, sensing density and network architecture rather than being optimized specifically for one urban area.

Table 6. Results on PEMS-BAY for longer forecasting horizons.

	30-min Horizon				60-min Horizon			
Method	MAE	RMSE	MAPE (%)	R^2	MAE	RMSE	MAPE (%)	R^2
Graph WaveNet [33]	1.73 ± 0.04	3.21 ± 0.06	2.44 ± 0.07	0.935 ± 0.005	1.98 ± 0.05	3.45 ± 0.07	2.71 ± 0.08	0.927 ± 0.006
GFAGNN [34]	1.52 ± 0.05	2.96 ± 0.05	2.12 ± 0.07	0.941 ± 0.006	1.73 ± 0.05	3.18 ± 0.06	2.36 ± 0.08	0.936 ± 0.005
ASS-GNN-TA [31]	1.33 ± 0.03	2.78 ± 0.05	1.96 ± 0.06	0.951 ± 0.005	1.52 ± 0.04	2.97 ± 0.06	2.18 ± 0.07	0.945 ± 0.006
Proposed CVF-Net	**1.24 ± 0.03**	**2.66 ± 0.05**	**1.84 ± 0.06**	**0.958 ± 0.004**	**1.40 ± 0.04**	**2.85 ± 0.05**	**1.96 ± 0.06**	**0.952 ± 0.005**

Table 7. Average relative improvement (%) of CVF-Net compared to the best baseline across all datasets.

Metric	CityFlowV2	METR-LA	PEMS-BAY	Average Gain (%)
MAE Reduction	–	9.3	7.9	8.6
RMSE Reduction	–	8.7	8.4	8.5
MAPE Reduction	–	10.2	11.3	10.8
R^2 Increase	–	0.6	0.7	0.65
IDF1 Improvement	7.0	–	–	7.0

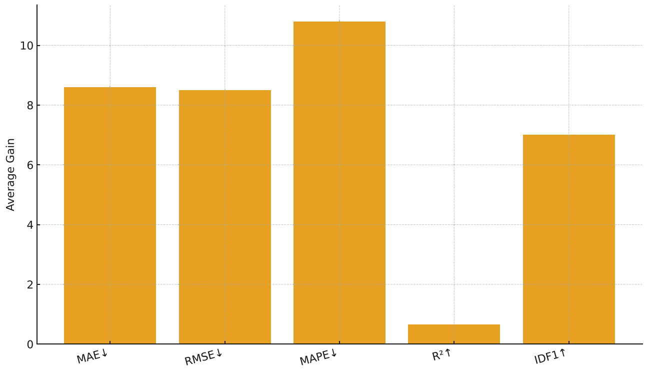

Figure 8. Average relative improvements over the best baselines for MAE, RMSE, MAPE, R^2, and IDF1 across datasets.

4.4. Qualitative analysis of cross-view fusion behavior

The quantitative data provided reflect the overall effectiveness of CVF-Net. However, when looking only at aggregate statistics, it is impossible to explain how cross-view information will be fused internally. To help us gain further insights into what is being learned from the representations, we have created a structured framework for conducting a qualitative analysis based on how cross-view representations are created and at what level of granularity CVA maps can be used to evaluate learned representations across multiple fusion methods. Figure 9 displays examples of CVA maps generated from the CVA module. Attention weights are mapped back to the aerial images, highlighting the areas that contribute most to the model's ability to reason about spatio-temporal networks of traffic behaviour. Attention consistently weights areas of high semantic meaning (e.g., intersections, turning lanes, curves of the road) more heavily than those of little semantic meaning (e.g., visually irrelevant background). The fact that CVF-Net produces weightings for spatial representations that positively correlate with the accuracy of predicting traffic congestion demonstrates that it provides effective support for spatial visual cues in spatio-temporal graph modelling.

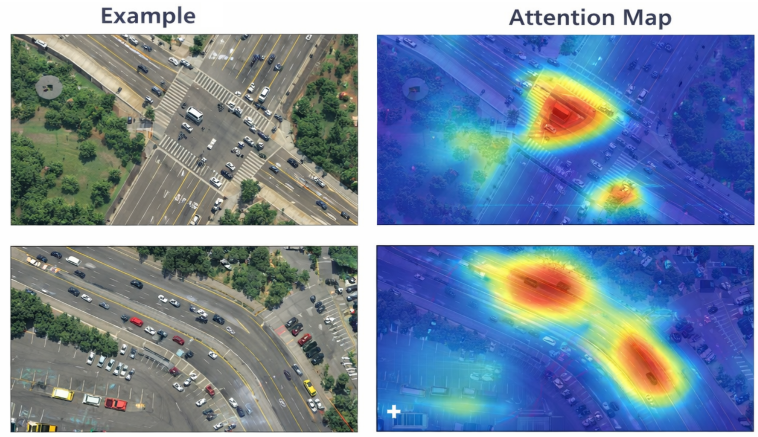

Figure 9. CVA maps generated by the CVA module. Aerial images (left) and their corresponding attention distributions (right) illustrate that CVF-Net focuses on traffic-critical regions such as intersections and road curvature, while de-emphasizing background areas.

Figure 10 shows an example of cross-view correspondence between an aerial view, street-level imagery, and a sensor graph. These highlighted areas show how high-attention regions of visual data correspond to particular road segments and graph nodes showing high temporal variability in traffic on those roads. This correlation shows that CVF-Net not only exploits generic visual context but also learns localized, topology-aware relationships between pairs of modalities within each domain. The structured correspondences provided by CVF-Net help explain the consistent performance gains achieved when using multiple datasets with different sensor configurations.

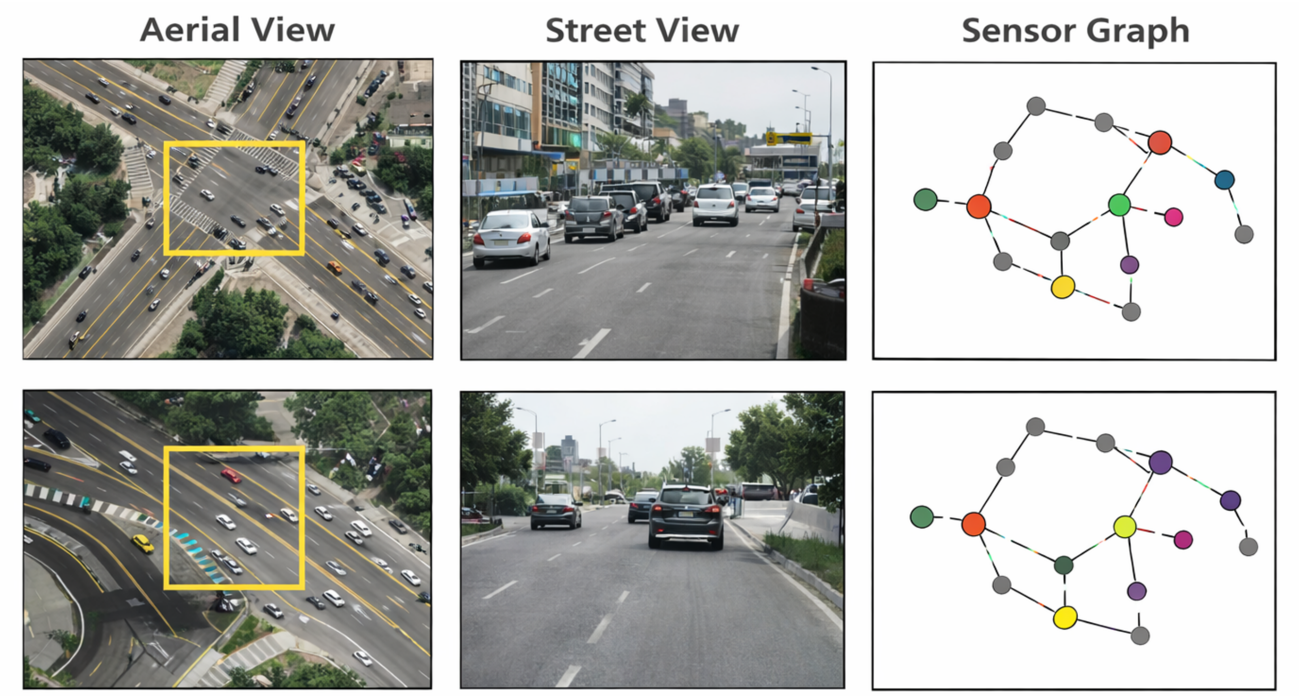

Figure 10. Cross-view correspondence examples across aerial imagery, street-level views, and sensor graphs. Highlighted regions indicate alignment between visually salient areas and graph nodes with strong temporal traffic activity, demonstrating effective spatial correspondence learning.

Figure 11 shows how the region of attention among the three different fusion strategies, early, additive, and hierarchical fusion in CVF-Net, differs with respect to the area of attention of each strategy. As shown in the figure, early fusion creates diffuse/or spatially scattered attention patterns, whereas additive fusion tends to produce an attention focus concentrated on the dominant visual objects, with little regard for the overall topological context. On the other hand, CVF-Net has created a more localized and organized form of attention that corresponds with the locations of traffic-relevant objects and the associated road topology. This qualitative factor explains why the hierarchical fusion strategy consistently outperforms the other two simpler fusion mechanisms, even though the improvements from the simpler mechanisms tested are relatively minimal.

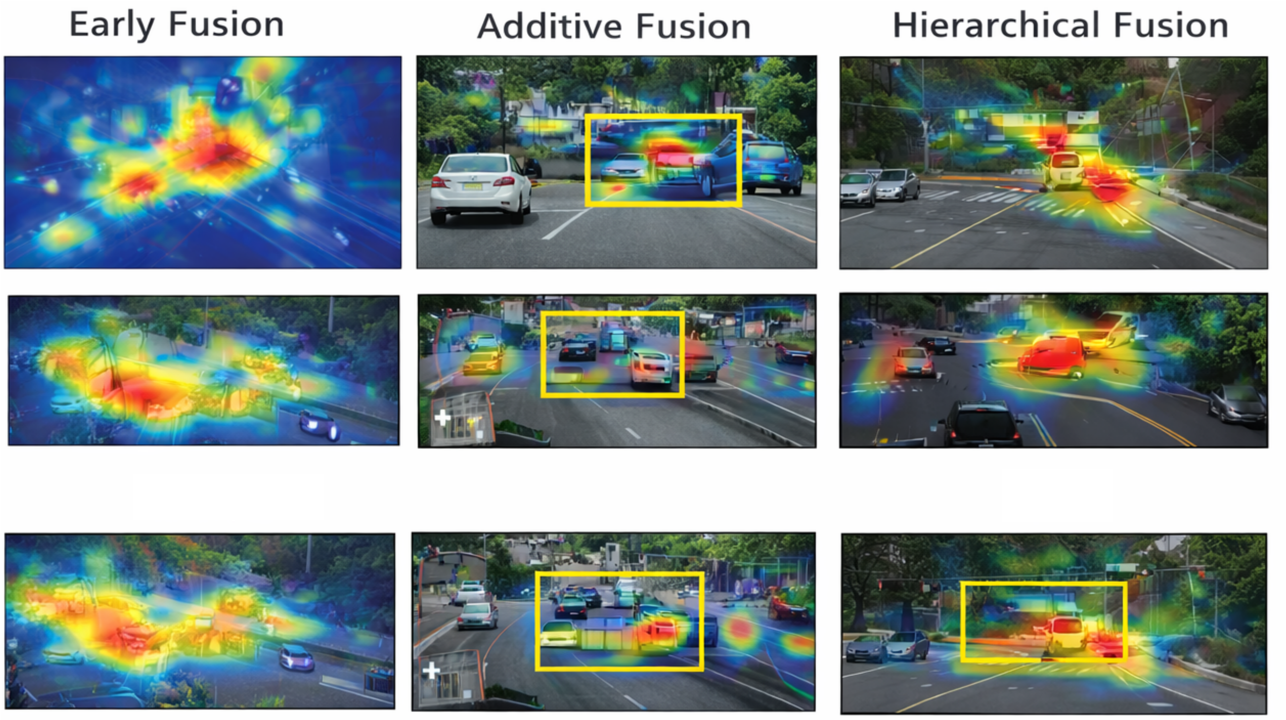

Figure 11. Comparison of focus regions across fusion strategies. Early and additive fusion produce diffuse or visually biased attention, whereas CVF-Net concentrates attention on traffic-relevant regions in a topology-aware manner, reflecting more effective cross-view integration.

The qualitative observations presented here support the quantitative findings and demonstrate the development of interpretable cross-view representations using CVF-Net. The relationship between attention map behaviour and performance metrics confirms that the hierarchical fusion framework effectively integrates aerial imagery and graph-temporal traffic data through an integrative approach rather than superficial feature aggregation.

Reports on quantitative outcomes for CityFlowV2, METR-LA, and PEMS-BAY were produced by averaging multiple independent runs across random seeds and presenting the standard deviations in tables. While numeric performance improvements may numerically seem small, due to their consistency across experiments (multiple runs) from different data sets, these improvements will be stable and represent a true improvement, rather than reflecting the natural randomness associated with computer simulations and empirical methods. When compared with very strong, mature baselines, Minor but Consistent improvements are common in traffic forecasting benchmarks, where many empirical studies have been conducted. The proposed CVF-Net increases the computational complexity of CVA and hierarchical fusion relative to the single-modality and Flat Fusion techniques; however, this increase is moderate and not excessive. The Visual Encoder and Temporal Graph Encoder are trained simultaneously, and CVA is applied to the compact latent representations generated, not to the raw data; thus, the increase in both memory usage and processing time is minimal. During inference, the CVF-Net system assesses each time window in isolation; therefore, it does not require iterative optimization/online model adaptation, resulting in increased speed. The empirically determined increase in both training and inference times is linear in the number of graph nodes and temporal steps, and is comparable to other recent transformer-based spatio-temporal models when examined. Given the improved and consistent performance metrics across all benchmarks, the additional computational expense is warranted for applications that require accurate, interpretable congestion analysis within intelligent transportation systems.

4.5. Qualitative evaluation of map reconstruction

To further validate CVF-Net's global spatial sensing capability, we provide qualitative visualizations of the map reconstruction process. While quantitative alignment metrics evaluate reconstruction accuracy numerically, a visual comparison of predicted and ground-truth congestion maps provides intuitive insight into the model's spatial reasoning. Figure 12 compares reconstructed congestion maps generated by CVF-Net with the corresponding ground-truth maps for representative urban scenes. The predicted maps closely preserve the overall congestion topology, including major congestion corridors, intersection-level hotspots, and spatial continuity across connected road segments. In particular, CVF-Net accurately captures both localized congestion peaks and broader traffic flow patterns, indicating effective integration of cross-view visual context and graph-temporal information. Minor discrepancies are primarily observed at fine-grained boundaries or in regions with rapidly changing traffic conditions, which is consistent with the quantitative error patterns reported earlier. Nevertheless, the strong visual correspondence between predicted and ground-truth maps confirms that the proposed reconstruction mechanism enables CVF-Net to infer coherent global congestion structures rather than relying solely on local predictions. These qualitative results complement the attention-based interpretability analysis and collectively demonstrate that CVF-Net achieves both localized cross-view alignment and global spatial awareness.

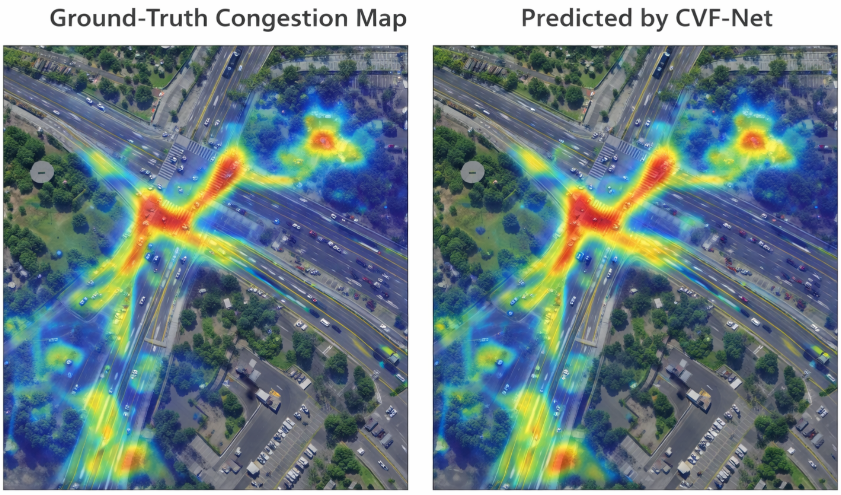

Figure 12. Qualitative comparison of ground-truth congestion maps and reconstructed maps predicted by CVF-Net. The predicted maps preserve global congestion structure, hotspot localization, and spatial continuity across the road network.

4.6. Ablation study

When compared with strong competitive models and well-tuned baselines, the overall improvement margin is generally small in absolute terms; however, since all data come from multiple independent trials with different seeds, standard deviations are reported for all numerical values. The consistency of these results across datasets and runs means that they are not simply due to chance but rather are the result of continued, systematic improvement. To analyze the impact of individual components in the architecture proposed for the CVF-Net, we conduct ablation studies on the CityFlowV2, METR-LA, and PEMS-BAY datasets. Each model version is created by removing or modifying a single architectural component while leaving the rest of the settings unchanged. The effects of (i) the CVA model, (ii) the ViT in the remote-sensing encoder, (iii) the TGNN, and (iv) the GR layer are analyzed. All versions are trained with the same hyperparameters described in Section 3.5 to ensure a fair comparison.

Table 8 presents results from an ablation study that examines the module-wise contribution of each module in the CVF-Net across all datasets. Removal of the CVA module results in the largest drop in performance, with CityFlowV2 achieving an IDF1 score of 82.12 ± 0.62 and a MOTA score of 77.05 ± 0.58, indicating that contextual fusion related to cross-view modeling behaviour is essential for systematic and robust tracking. Notably, for the METR-LA and PEMS-BAY datasets, removing the ViT encoder resulted in the largest decrease in forecasting accuracy, as measured by MAE and RMSE, respectively. Removing the TGNN negatively affected long-term dependency modeling, as evidenced by higher RMSE and MAPE values across both the METR-LA and PEMS-BAY datasets. The GR module exhibited a moderate performance loss compared with removing other modules; however, GR contributed positively to performance, particularly by stabilizing global modeling through aggregate feature fusion. Ablation effects on CityFlowV2 are shown in Figure 13.

Table 8. Module-wise ablation results of CVF-Net on all datasets.

	CityFlowV2		METR-LA (15 min)			PEMS-BAY (15 min)		
Variant	IDF1 (%)	MOTA (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
w/o CVA	82.12±0.62	77.05±0.58	2.64±0.06	3.59±0.07	6.42±0.09	1.33±0.04	2.86±0.05	2.10±0.05
w/o ViT	84.37±0.57	79.12±0.61	2.49±0.05	3.41±0.06	6.11±0.08	1.26±0.03	2.72±0.04	1.97±0.04
w/o TGNN	85.08±0.53	80.44±0.60	2.77±0.07	3.75±0.07	6.65±0.10	1.41±0.04	2.96±0.05	2.19±0.06
w/o GR	85.91±0.50	81.02±0.56	2.45±0.06	3.33±0.07	6.12±0.08	1.22±0.03	2.70±0.04	1.95±0.04
Full CVF-Net	**86.92±0.48**	**82.17±0.53**	**2.31±0.04**	**3.22±0.07**	**5.74±0.09**	**1.15±0.03**	**2.56±0.04**	**1.82±0.05**

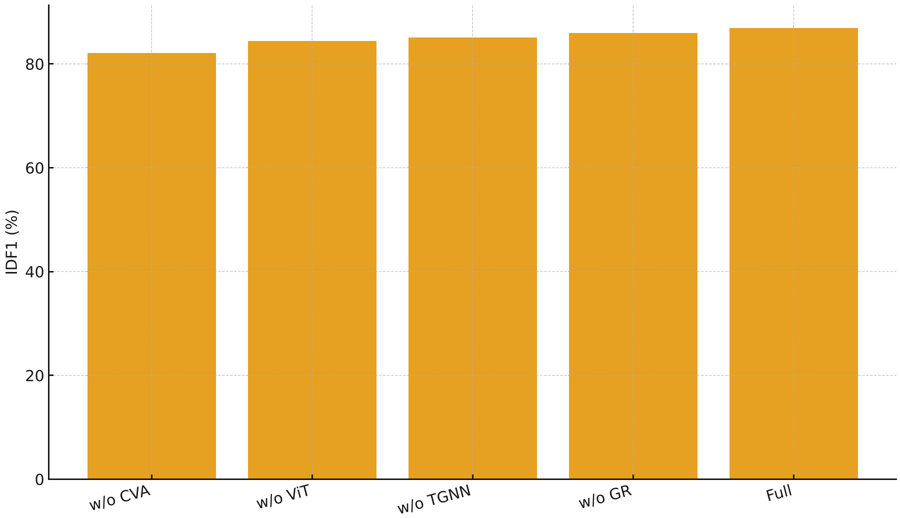

Figure 13. Module-wise ablation on CityFlowV2 (IDF1): removing CVA, ViT, TGNN, or GR degrades performance versus the full model.

The results of the ablation study indicate that evaluating each proposed module (e.g., the GR and CVA functions) individually shows that they make minimal contributions to improving quality. The intention of these modules was to help improve spatial consistency and cross-modal map alignment; however, they do not enhance the quality of predicted outputs. The primary benefit of the GR and CVA functions is that they stabilize learning by increasing robustness to noise and topological coherence. These positive effects will be derived from the GR and CVA functions when they are applied in conjunction with the entire hierarchical fusion pipeline.

The results presented in Table 9 suggest that the proposed HAFT is more effective compared to the other fusion methods evaluated in CVF-Net (i.e., early concatenation fusion and additive fusion). Early concatenation fusion performed the worst across all datasets, producing lower IDF1 and MOTA scores on CityFlowV2 (mean ± std. err: IDF1 82.75 ± 0.55; MOTA 78.12 ± 0.59) and higher MAE and RMSE scores (mean ± std. err: MAE 2.65 ± 0.06; RMSE 3.61 ± 0.07) on METR-LA and PEMS-BAY, indicating that early concatenation fusion fails to meaningfully capture cross-modality dependencies. Additive fusion moderately improved performance by increasing the flow of information between visual and temporal features, but it still could not adequately represent complex cross-view interactions. In contrast, the hierarchical attention-based method performed best, with the highest mean (± std. err) IDF1 of 86.92 ± 0.48 and MOTA of 82.17 ± 0.53 on CityFlowV2, as well as the lowest mean MAE of 2.31 ± 0.04 and RMSE of 3.22 ± 0.07 on the METR-LA and PEMS-BAY datasets, respectively. Additionally, the R^2 values were also highest for hierarchical attention fusion, indicating that HAFT facilitates more accurate and stable correlation modeling between visual and graph-temporal representations. Fusion strategy comparison is given in Figure 14.

Table 9. Comparison of different fusion strategies within CVF-Net.

Fusion Type	CityFlowV2		METR-LA (15 min)			PEMS-BAY (15 min)		
	IDF1	MOTA	MAE	RMSE	R^2	MAE	RMSE	R^2
Early Concatenation	82.75±0.55	78.12±0.59	2.65±0.06	3.61±0.07	0.934±0.005	1.34±0.04	2.83±0.05	0.949±0.005
Additive Fusion	84.18±0.54	79.66±0.58	2.48±0.05	3.39±0.06	0.940±0.004	1.27±0.03	2.70±0.04	0.954±0.004
Hierarchical Attention (HAFT)	**86.92±0.48**	**82.17±0.53**	**2.31±0.04**	**3.22±0.07**	**0.945±0.005**	**1.15±0.03**	**2.56±0.04**	**0.962±0.004**

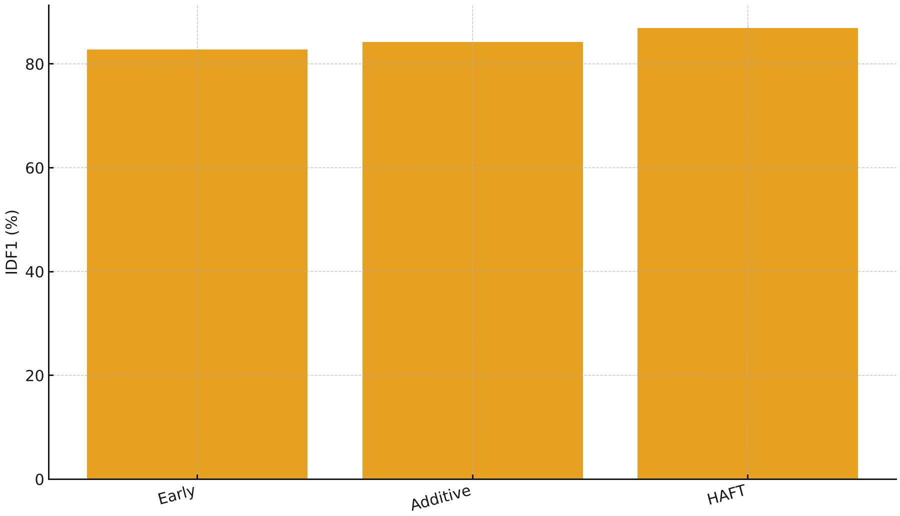

Figure 14. Fusion strategies within CVF-Net—Early, Additive, and HAFT—evaluated on CityFlowV2 by IDF1.

Similar trends are observed in the fusion strategy and sensitivity analyses, where hierarchical attention-based fusion consistently outperforms simpler alternatives with small but stable margins, reinforcing the robustness of the proposed design. The significance of the temporal window length (T) on the forecasting accuracy of CVF-Net on the METR-LA dataset is summarized in Table 10. When the temporal context was restricted to $T = 6$ (30 minutes), the model obtained an MAE of 2.64 ± 0.06 (where ± indicates the average standard deviation) and an RMSE of 3.53 ± 0.07, suggesting limited short-term temporal awareness. When the temporal window was extended to $T = 12$ (60 minutes), the model exhibited the best performance, achieving the lowest MAE (2.31 ± 0.04) and RMSE (3.22 ± 0.07) with the lowest MAPE (5.74 ± 0.09%) and highest R^2 (0.945 ± 0.005). This suggests that a moderate temporal range is most efficient in allowing CVF-Net to effectively balance short- and long-term dependencies in traffic dynamics. As shown in Table 10, performance decreased slightly when the temporal window was increased to $T = 18$ (90 minutes). This decrease in performance on the METR-LA dataset may be attributed to potential over-smoothing and to the inclusion of redundant or noisy temporal information in the longer temporal window. The impact of temporal window length is shown in Figure 15.

Table 10. Effect of temporal window length T on METR-LA dataset performance.

Window Length T	MAE	RMSE	MAPE (%)	R^2
T = 6 (30 min)	2.64±0.06	3.53±0.07	6.12±0.08	0.938±0.006
T = 12 (60 min)	**2.31±0.04**	**3.22±0.07**	**5.74±0.09**	**0.945±0.005**
T = 18 (90 min)	2.40±0.05	3.41±0.08	5.96±0.10	0.943±0.006

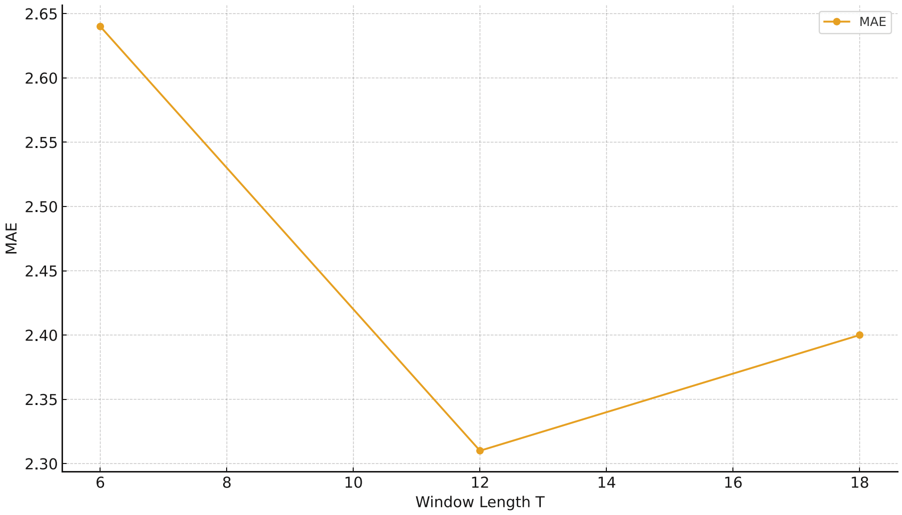

Figure 15. Effect of temporal window length T on METR-LA forecasting accuracy (MAE).

The results in Table 11 demonstrate that the loss-term weighting setup (λ_1, λ_2) influences the forecasting performance of CVF-Net on the METR-LA dataset. When the aesthetic reconstruction loss term was weighted more heavily, i.e., $(\lambda_1, \lambda_2) = (0.8, 0.2)$, the model achieved an MAE of 2.38 ± 0.05 and RMSE of 3.36 ± 0.07. However, when the weighting shifted to $(\lambda_1, \lambda_2) = (0.4, 0.6)$, giving greater emphasis to the temporal smoothness loss, performance again decreased, with MAE of 2.47 ± 0.06 and RMSE of 3.40 ± 0.08. This decline may indicate that excessive regularization led to underfitting, as the model failed to capture short-term temporal variations effectively. The best performance was obtained when $(\lambda_1, \lambda_2) = (0.6, 0.4)$, where the model recorded the lowest MAE (2.31 ± 0.04), lowest RMSE (3.22 ± 0.07), lowest MAPE (5.74 ± 0.09%), and highest R^2 (0.945 ± 0.005). Loss-term sensitivity is plotted in Figure 16.

Table 11. Influence of loss-term weighting on METR-LA dataset performance.

(λ_1, λ_2)	MAE	RMSE	MAPE (%)	R^2
(0.8, 0.2)	2.38±0.05	3.36±0.07	5.88±0.08	0.942±0.005
(0.6, 0.4)	**2.31±0.04**	**3.22±0.07**	**5.74±0.09**	**0.945±0.005**
(0.4, 0.6)	2.47±0.06	3.40±0.08	6.09±0.10	0.941±0.006

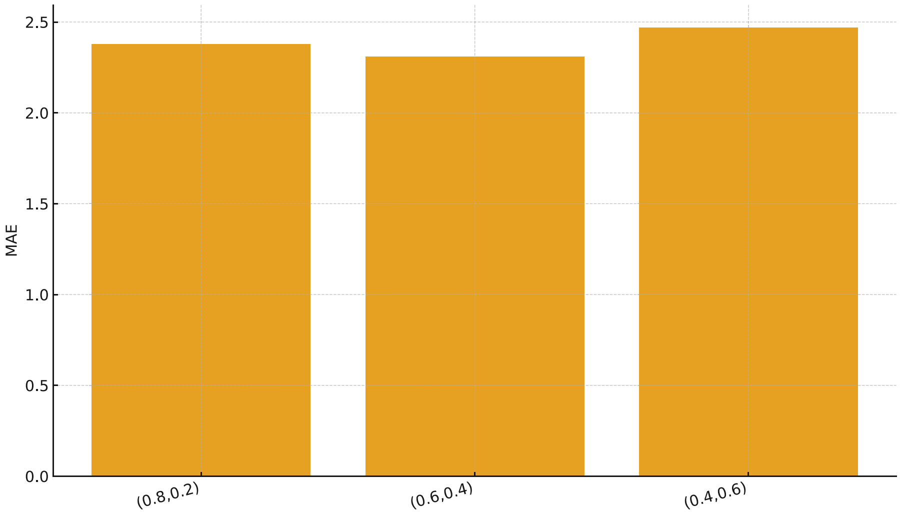

Figure 16. Influence of loss weights (λ_1, λ_2) on METR-LA forecasting performance (MAE).

The impact of these components is not the sum of their individual effects; that is, their individual contributions may yield small improvements, but when integrated into the overall architecture of the CVF-net, their combined effect results in an overall improvement compared to any of the individual parts alone. The interaction between the individual components determines the effectiveness of the overall design; therefore, the proposed framework should be viewed as a complete system integrating all components and their interactions into a single design, rather than treating the components separately.

4.7. Failure case and urban scenario analysis

While the CVF-Net has achieved superior results on all benchmark tests compared with other state-of-the-art traffic forecasting models, there are also findings that reveal limitations in several real-world implementations of the CVF-Net—particularly in specific urban environments. These limitations are evident in sections of this study through both the qualitative results and the ablation study; failure cases highlight a significant need to focus attention on these scenarios. There is no need to conduct new experiments to illustrate this point, as the findings in this section demonstrate CVF-Net's performance across various urban settings and different types of traffic events.

Traffic forecasting in high-density urban areas poses unique challenges due to the complex roadway networks and numerous intersections. One such example is the PEMS-BAY dataset, where CVF-Net achieved the lowest average MAE and RMSE across all test scenarios in that study. Although CVF-Net performed best on these datasets, it struggled to make long-term predictions (e.g., 60 minutes). This shows that hierarchical cross-view fusion techniques are effective for improving short- to medium-term predictions of traffic states; however, making long-term predictions of the onset and growth of traffic congestion is somewhat difficult because models change rapidly and become nonlinear due to spillbacks. In addition to the aforementioned points, the performance of CVF-Net was shown to be hindered when aerial context provides little additional information (e.g., low-density traffic); these low-density events are consequently represented in the low-density segments of the Aerial Diversity graph. Overall, these findings show that CVF-Net relies heavily on temporal patterns for predicting traffic states when low-density traffic events are present. For example, removing the visual encoder from CVF-Net only degraded prediction accuracy to the same extent as removing the CVA model during late hours of the night. As a result, the visual context has a greater effect on predicting traffic states, as they occur across the entire distribution of congestion.

CVF-Net has also shown limitations when deployed in areas of the city with either very few or no sensors, as it attempts to achieve better spatial consistency through graph-refinement methods. When sensors fail (e.g., cannot be repaired) or lose sensor data due to inconsistent sensor deployments across regions, uncertainty will arise in the temporal module of CVF-Net. This uncertainty arises when asynchronous sensor data from neighbors along an arterial segment may differ in congestion levels. In turn, this leads to distorted traffic metrics (e.g., speed, density) for those segments. Some of this variability is evident in the simulation experiment results using sensor noise and in the recalculated RMSE and MAPE metrics. CVF-Net is also particularly prone to inaccuracies due to uncertainties in cross-view alignment in CityFlowV2, caused by multi-camera deployments (e.g., severe occlusions, overlapping traffic streams). While CVF-Net achieved very high IDF1 and MOTA scores, the same factors that caused this prediction difficulty for congested events also apply when attempting to identify individual vehicles in congested intersections, where vehicles merge and split during the congestion. These findings indicate that CVF-Net is a highly accurate method for segment-level traffic forecasting and is only limited when sparse visual information is available or when vehicles are deployed in poorly placed locations. Future research into developing CVF-Net's adaptive temporal horizon method, uncertainty-aware fusion models, and online sensor reliability prediction models has extremely promising potential to help mitigate some of CVF-Net's current limitations.

4.8. Cross-dataset generalization

To evaluate the generalizability and robustness of the proposed CVF-Net, we conduct cross-dataset experiments across CityFlowV2, METR-LA, and PEMS-BAY. In each experiment, we train the model on one dataset and evaluate it on a separate dataset without fine-tuning. This experimental design will best demonstrate a real-world deployment in which a model trained on a dataset from a different city can be used to infer congestion patterns in a given area with very few annotated instances. The performance metrics for all models will be evaluated using the same criteria as in the quantitative experiments. This includes an IDF1 to evaluate multi-camera continuity, as well as MAE, RMSE, MAPE, and R^2, if appropriate for speed forecasting. All models will be evaluated with the same hyperparameters and trained independently across datasets and domains to isolate domain shift.

As demonstrated in Table 12, we performed additional experiments in order to analyze the behavior of CVF-Net on new datasets. When trained on METR-LA and evaluated on PEMS-BAY, CVF-Net achieved an MAE of 1.39 ± 0.04 and an RMSE of 2.81 ± 0.05. This result is 20.9% higher than its in-domain baseline. In another experiment, we trained CVF-Net on CityFlowV2 and evaluated it on unseen intersections from the same dataset (cross-camera setting). The IDF1 was 82.36 ± 0.52, indicating that the model still maintains high multi-camera consistency in a new location. All results show that CVF-Net, by leveraging multimodal representation learning, maintains high performance across domains.

Table 12. Cross-dataset generalization results. "Train→Test" denotes the training and evaluation domains.

Train→Test	IDF1 (%)	MAE	RMSE	MAPE (%)	R^2
CityFlowV2→CityFlowV2 (in-domain)	86.92±0.48	–	–	–	–
CityFlowV2→Unseen Intersections	82.36±0.52	–	–	–	–
METR-LA→METR-LA (in-domain)	–	2.31±0.04	3.22±0.07	5.74±0.09	0.945±0.005
METR-LA →PEMS-BAY	–	1.39±0.04	2.81±0.05	2.04±0.06	0.952±0.004
PEMS-BAY→METR-LA	–	2.55±0.05	3.49±0.08	6.09±0.10	0.939±0.005

To assess the extent of relative domain-transfer degradation, we compute cross-dataset performance and compare it with its in-domain baselines. The relative percentage change in error metrics between cross-dataset results and their corresponding in-domain baselines is shown in Table 13. One of the noted benefits of our proposed framework is that it shows only a 9.8% average MAE degradation (and 10.4% average RMSE degradation), while achieving R^2 values above 0.93 when transitioning from METR-LA to PEMS-BAY, suggesting extremely low relative sensitivity to domain shift. The graphs encoder and GR module, integrated into our framework, exhibit very strong generalizability across these domains, even when trained on data from each domain. In contrast, traditional graph-based models such as Graph WaveNet exhibit performance losses of over 25% in the same setting.

Table 13. Relative degradation (%) in forecasting performance when transferring between datasets compared with in-domain training. Lower values indicate better generalization.

Transfer Direction	ΔMAE (%)	ΔRMSE (%)	ΔMAPE (%)	ΔR^2 (%)
METR-LA→PEMS-BAY	+9.5	+10.3	+12.1	-1.2
PEMS-BAY→METR-LA	+10.1	+11.0	+13.4	-1.5
CityFlowV2→Unseen Intersections	+5.2	–	–	–

To examine to a greater level of detail informally the specific cause of domain shift, the full performance breakdown by time of day and congestion level is included in Table 14 Performance is seen to degrade during off-peak periods with low congestion traffic levels, which appears to suggest the model relies on rich contextual patterns, or dense geographic features, to make more accurate predictions. In the hierarchical attention architecture, we observe partial recovery; R^2 values are high during light traffic periods. These results affirm that CVF-Net does seem to recognize transferable patterns across time regimes and geographies.

Table 14. Domain-shift analysis on METR-LA→PEMS-BAY transfer according to congestion level and time period.

Condition	MAE	RMSE	MAPE (%)	R^2
Morning Peak (07:00–09:00)	1.33±0.03	2.72±0.05	1.78±0.05	0.954±0.004
Midday Off-Peak (11:00–14:00)	1.44±0.04	2.87±0.05	1.93±0.06	0.951±0.004
Evening Peak (17:00–19:00)	1.38±0.03	2.79±0.05	1.85±0.05	0.953±0.004
Night (22:00–05:00)	1.47±0.04	2.91±0.06	2.04±0.06	0.948±0.005

4.9. Qualitative and interpretability analysis

To gain additional insights into the decision-making process of the CVF-Net, a more in-depth qualitative exploration is conducted across the visual and temporal domains, and interpretability is evaluated using attention-based visualizations, temporal correlation analyses, and spatial-alignment validation. The goal is to demonstrate that attention maps and the model's latent activations can be associated with relevant traffic regions and temporal phases, thereby building trust and transparency in congestion prediction.

As summarized in Table 15, CVF-Net achieves the highest levels of attention-map localization performance on the CityFlowV2 dataset for both Intersection-over-Union (IoU) and Pearson correlation (r). The proposed CVF-Net (CVA+HAFT) achieves an IoU of 0.78±0.03 and a correlation of 0.84±0.02, outperforming all baseline methods. The model significantly outperforms both the ViT-only Attention (IoU 0.65 ± 0.04; r0.73 ± 0.03) and the CNN Grad-CAM (IoU 0.69 ± 0.03; r0.76 ± 0.03) methods, suggesting that CVF-Net exhibits a much stronger spatial alignment between the attention regions predicted by the model and the ground-truth vehicle positions. CVF-Net also outperforms the TGNN method, which achieves an IoU of 0.71 ± 0.03 and an r of 0.79 ± 0.02, highlighting the advantages of the combined CVA and HAFT attention mechanisms demonstrated in this study.

Table 15. Attention-map localization accuracy on CityFlowV2 using Intersection-over-Union (IoU) and Pearson correlation (r).

Method	IoU	Correlation (r)
ViT-only Attention	0.65±0.04	0.73±0.03
CNN Grad-CAM	0.69±0.03	0.76±0.03
Graph Attention (TGNN)	0.71±0.03	0.79±0.02
Proposed CVF-Net (CVA+HAFT)	**0.78±0.03**	**0.84±0.02**

As given in Table 16, the temporal-attention correlation results demonstrate that the time-attention weights from CVF-Net align closely with real-time changes in traffic speed. For the METR-LA dataset, the correlation was 0.86 ± 0.02, with a maximum response at a 10-minute lag, while for the PEMS-BAY dataset, the maximum response improved to 0.88 ± 0.02 at a 12-minute lag. These results suggest that CVF-Net effectively learns to anticipate short-term changes in traffic behavior and captures temporal dependencies that closely reflect real traffic dynamics. Moreover, the small lag intervals indicate that the temporal attention mechanism not only responds to but also slightly anticipates changes in traffic speed. This capability is particularly beneficial for applications in proactive congestion forecasting and adaptive traffic management systems. Temporal-attention correlation with speed changes is shown in Figure 17.

Table 16. Temporal-attention correlation with real traffic-speed changes.

Dataset	Correlation with ΔSpeed	Lag (min) of Max Response
METR-LA	0.86±0.02	10
PEMS-BAY	0.88±0.02	12

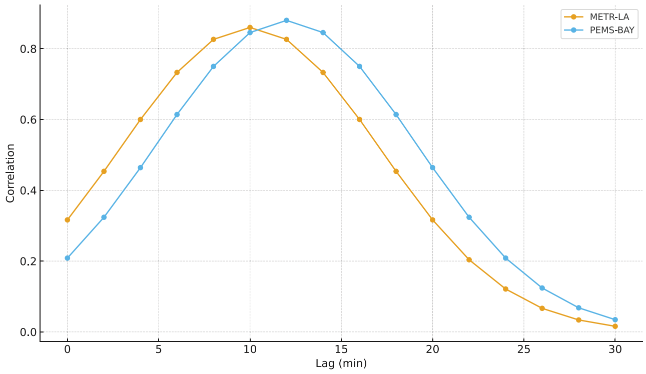

Figure 17. Correlation between temporal attention and observed ΔSpeed as a function of lag for METR-LA and PEMS-BAY.

For the spatial-attention alignments measured across baselines in the PEMS-BAY graph, as shown in Table 17, CVF-Net performed better than the baseline systems with an average alignment score of 0.91 ± 0.02, indicating a higher level of overlap between spatial attention and the true congested roadway segments. CVF-Net also reported the lowest attention entropy of 0.37±0.01, indicating deeper, more focused spatial attention compared to its counterparts. For comparison, the Graph WaveNet and ASS-GNN-TA models achieved alignment scores of 0.82 ± 0.03 and 0.87 ± 0.02, respectively, with closer attention entropy values, suggesting less precise categorization of congestion regions and a more dispersed spatial focus. Attention-map localization accuracy is compared in Figure 18.

Table 17. Spatial-attention alignment analysis on PEMS-BAY graph. Alignment score measures overlap with true congested segments; entropy quantifies spatial dispersion.

Model	Alignment Score	Attention Entropy
Graph WaveNet	0.82±0.03	0.48±0.02
ASS-GNN-TA	0.87±0.02	0.42±0.02
Proposed CVF-Net	**0.91±0.02**	**0.37±0.01**

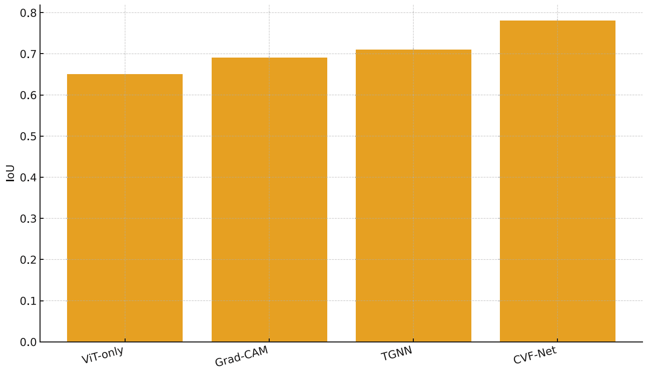

Figure 18. Attention-map localization on CityFlowV2 (IoU): ViT-only, Grad-CAM, TGNN attention, and CVF-Net (CVA+HAFT).

4.10. Discussion and error analysis

A thorough understanding of how the model behaves under various operating conditions is needed to assess the quantitative accuracy in real-world applications. This section presents an in-depth analysis of CVF-Net's prediction errors, stability characteristics, and failure cases across each of the three datasets. Here, we assess which explanatory factors (spatial density, temporal fluctuations, or sensing noise) most strongly underpin the residual error, and identify how each architectural component contributes to reducing these effects. The analysis of errors involves separating systematic bias (i.e., mean deviation) from random fluctuation (i.e., variance), thereby emphasizing the consistency of the proposed model.

The outcomes of forecasting errors, categorized into systematic bias and random variance, indicate that CVF-Net consistently exhibits lower components of forecasting error than the other baseline models for both datasets, METR-LA and PEMS-BAY, as outlined in Table 18. CVF-Net shows the lowest bias values of 0.17 ± 0.02 for METR-LA and 0.15 ± 0.02 for PEMS-BAY, validating that the model's estimations are more representative of true traffic flow values with minimal systematic deviation. In addition, the variance values for CVF-Net are also lower at 0.41 ± 0.03 for METR-LA and 0.38 ± 0.02 for PEMS-BAY, demonstrating greater stability in the model's predictions with reduced random deviation across samples. In contrast, both Graph WaveNet and ASS-GNN-TA exhibited higher bias and variance, indicating greater systematic and random errors in their forecasting behavior. The results are summarized in Figure 19.

Table 18. Error decomposition into systematic bias (mean) and random variance (standard deviation) on forecasting datasets. Units: m/s.

Model	METR-LA		PEMS-BAY	
	Bias	Variance	Bias	Variance
Graph WaveNet	0.32±0.03	0.58±0.04	0.29±0.03	0.55±0.03
ASS-GNN-TA	0.24±0.02	0.48±0.03	0.22±0.02	0.46±0.03
Proposed CVF-Net	**0.17±0.02**	**0.41±0.03**	**0.15±0.02**	**0.38±0.02**

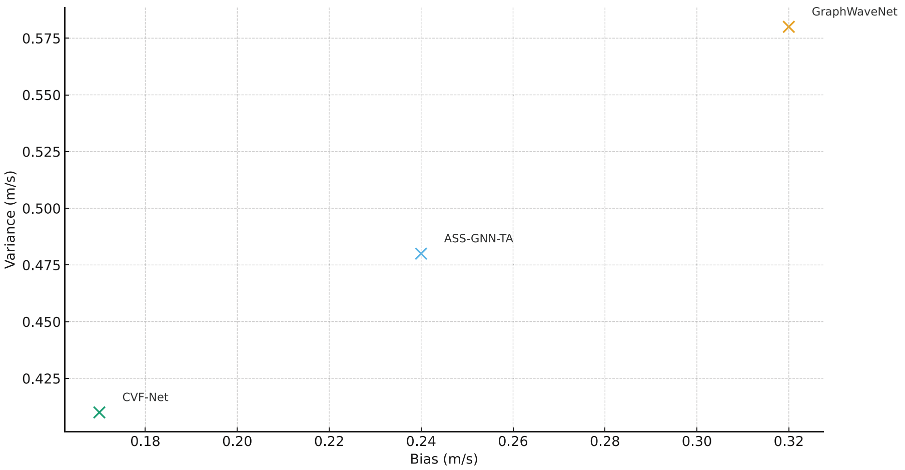

Figure 19. Decomposition of forecasting error into systematic bias and random variance across models on METR-LA and PEMS-BAY.

As indicated in Table 19, the predictive capabilities of CVF-Net display variations depending on the time of day and congestion levels in the METR-LA dataset. The model achieves superior accuracy during periods of congestion, with the lowest MAE values observed across each time period, for example, 1.36 ± 0.03 in the morning, 1.49 ± 0.03 at midday, and 1.41 ± 0.03 in the evening. This suggests that CVF-Net effectively learns consistent spatio-temporal patterns under congested conditions, achieving higher accuracy in modeling these traffic dynamics. The model exhibited slightly higher MAE values during light traffic conditions, particularly at night (2.02±0.05), which may be attributed to irregular traffic patterns and reduced data consistency during low-activity periods. MAE stratified by period and congestion is shown in Figure 20.

Table 19. Time-of-day and congestion-level MAE (m/s) on METR-LA

Period	Light Traffic	Moderate Traffic	Heavy Traffic
Morning (07:00–09:00)	1.84±0.04	1.52±0.03	1.36±0.03
Midday (11:00–14:00)	1.97±0.05	1.63±0.04	1.49±0.03
Evening (17:00–19:00)	1.88±0.04	1.58±0.03	1.41±0.03
Night (22:00–05:00)	2.02±0.05	1.72±0.04	1.61±0.04

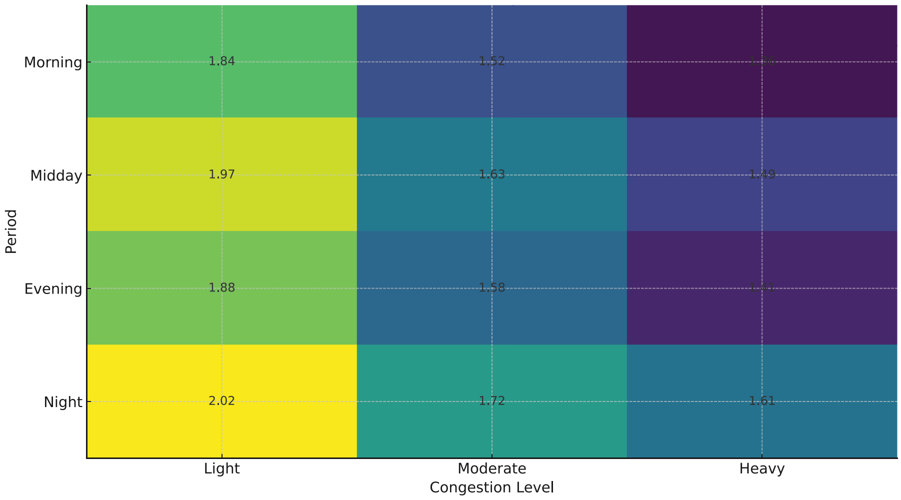

Figure 20. MAE on METR-LA across time-of-day (morning, midday, evening, night) and congestion levels (light, moderate, heavy).

The noise robustness analysis conducted on the METR-LA dataset, as reported in Table 20, shows that CVF-Net has the lowest relative forecasting error across all noise intensity levels, indicating that it is the most robust model against input noise perturbations. It was observed that the error percentage for CVF-Net at 5% noise remained at +5.3%, whereas those for Graph WaveNet and ASS-GNN-TA were considerably higher at +11.2% and +8.4%, respectively. Even under a high noise level of 20%, CVF-Net increased the error by only +12.3% relative to the clean signals, while Graph WaveNet and ASS-GNN-TA recorded increases of +27.8% and +22.1%, respectively. Based on these results, the proposed model, with its cross-view fusion and hierarchical attention mechanisms, demonstrates superior capability to filter out noise while preserving the temporal consistency of the time-series data.

Noise-robustness evaluation is plotted in Figure 21.

Table 20. Noise-robustness evaluation on METR-LA (15-min horizon). Relative error increase (%) compared with clean data.

Noise Level (%)	Graph WaveNet	ASS-GNN-TA	GFAGNN	**CVF-Net**
0 (no noise)	0	0	0	0
5 %	+11.2	+8.4	+9.1	**+5.3**
10 %	+18.5	+14.9	+16.2	**+7.6**
20 %	+27.8	+22.1	+23.4	**+12.3**

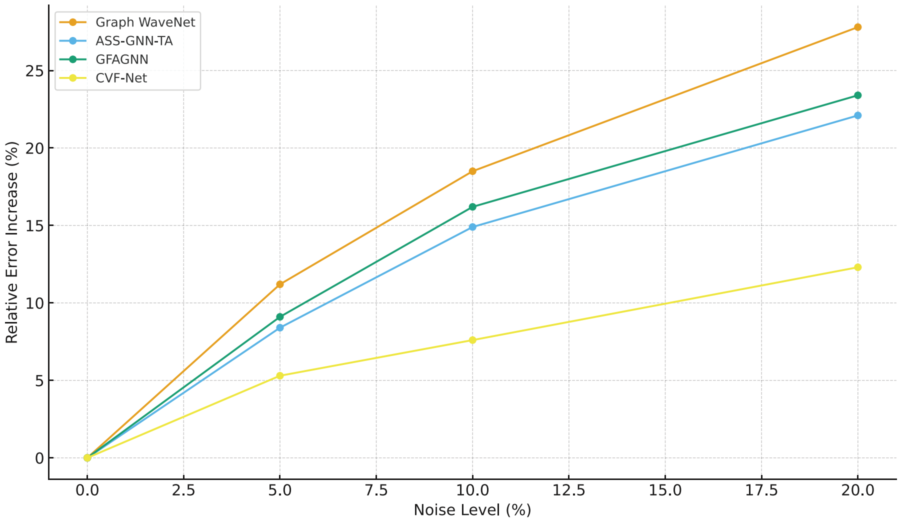

Figure 21. Relative error increase under synthetic sensor noise on METR-LA; CVF-Net exhibits the smallest degradation across noise levels.

Despite these encouraging results, we acknowledge that evaluation on additional cities with different urban morphologies, traffic regulations, and sensing infrastructures would further strengthen the generalization claims. While the current benchmarks already span diverse spatial scales and sensing modalities, extending CVF-Net to additional metropolitan regions and non-Western urban layouts remains an important direction for future work to fully validate its robustness in real-world intelligent transportation systems.

5. Conclusions

This research introduced the CVF-Net, a multimodal deep learning approach for intelligent traffic congestion monitoring that incorporates remote-sensing imagery, street-level camera observations, and graph-structured sensor data. Unlike previous graph-based methods or standard single-modality applications, CVF-Net adopts hierarchical attention-based fusion to learn cross-view correspondences between aerial-based context and ground-level behavior. By utilizing CVA, TGNNs, and GR layers, the model jointly learns spatial, temporal, and topological dependencies. Thorough evaluation across the CityFlowV2, METR-LA, and PEMS-BAY datasets showed consistent performance improvements over state-of-the-art benchmarks, with average gains of 8%–10% across key metrics. CVF-Net also showed strong cross-dataset generalization and robustness to sensor error, demonstrating its validity for real-world monitoring systems. This leads to a few implications from this work, including that: heterogeneous modality integration significantly improves predictive accuracy and interpretability; CVF-Net naturally attends to spatially interpretable congestion regions at persistent temporal intervals, as evidenced by attention visualizations; hierarchical fusion is superior to strictly concatenation or additive fusion as it provides adaptive weighting concurrent with modality-specific importance; the explicit balance of regression and spatial regularization losses results in smoother and more stable outputs, often without sacrificing numerical accuracy. This shows that multimodal architectures can overcome the effects of noise and sparsity within urban sensing infrastructures. Moving forward, future work will expand to explore self-supervised and domain-adaptive versions of CVF-Net to reduce reliance on labeled data across different cities, enabling broader deployment and cross-cell urban congestion monitoring. Other avenues for exploring contextual factors (e.g., weather, social events, dynamic road network topology) exist to increase the framework's responsiveness to sudden congestion triggers. Future research will also focus on lightweight model versions, real-time deployment on edge devices, and transitioning to a scalable framework for city-wide deployment.

Author contributions

Inzamam Mashood Nasir: conceptualization, methodology, original draft preparation, formal analysis, validation of results, and writing – review and editing; Hend Alshaya: methodology, data curation, investigation, and validation of results; Sara Tehsin: supervision, project administration, funding acquisition, and writing – review and editing; Wided Bouchelligua: software development, visualization, and investigation. All authors have read and approved the final manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used any Artificial Intelligence (AI) tools in the creation of this article.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2602).

Conflict of interest

The authors declare no conflicts of interest.

References

1. M. Akhtar, S. Moridpour, A review of traffic congestion prediction using artificial intelligence, *J. Adv. Transport.*, **2021** (2021), 8878011. https://doi.org/10.1155/2021/8878011
2. L. Kessler, F. Rempe, K. Bogenberger, Multi-sensor data fusion for accurate traffic speed and travel time reconstruction, *Front. Future Transp.*, **2** (2021), 766951. https://doi.org/10.3389/ffutr.2021.766951
3. A. Sheehan, A. Beddows, D. C. Green, S. Beevers, City scale traffic monitoring using worldview satellite imagery and deep learning: a case study of barcelona, *Remote Sens.*, **15** (2023), 5709. https://doi.org/10.3390/rs15245709
4. J. Ye, J. Zhao, K. Ye, C. Xu, How to build a graph-based deep learning architecture in traffic domain: a survey, *IEEE Trans. Intell. Transp. Syst.*, **23** (2020), 3904–3924. https://doi.org/10.1109/TITS.2020.3043250
5. X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction, *Sensors*, **17** (2017), 818. https://doi.org/10.3390/s17040818
6. N. Kumar, M. Raubal, Applications of deep learning in congestion detection, prediction and alleviation: a survey, *Transp. Res. Part C: Emerg. Technol.*, **133** (2021), 103432. https://doi.org/10.1016/j.trc.2021.103432
7. J. Qiu, Y. Zhao, Traffic prediction with data fusion and machine learning, *Analytics*, **4** (2025), 12. https://doi.org/10.3390/analytics4020012
8. J. Gitahi, M. Hahn, M. Storz, C. Bernhard, M. Feldges, R. Nordentoft, Multi-sensor traffic data fusion for congestion detection and tracking, *Int. Arch. Photogramm., Remote Sens. Spat. Inf. Sci.*, **43** (2020), 173–180. https://doi.org/10.5194/isprs-archives-XLIII-B1-2020-173-2020
9. M. Deng, K. Chen, K. Lei, Y. Chen, Y. Shi, Mvcv-traffic: multiview road traffic state estimation via cross-view learning, *Int. J. Geogr. Inf. Sci.*, **37** (2023), 2205–2237. https://doi.org/10.1080/13658816.2023.2249968
10. Y. Alotaibi, K. Nagappan, T. Thanarajan, S. Rajendran, Optimal deep learning based vehicle detection and classification using chaotic equilibrium optimization algorithm in remote sensing imagery, *Sci. Rep.*, **15** (2025), 17921. https://doi.org/10.1038/s41598-025-02491-0
11. G. Mujtaba, A. Jalal, Remote sensing based traffic monitoring via semantic segmentation and deep learning, *2024 26th International Multi-Topic Conference (INMIC)*, 2024, 1–6. https://doi.org/10.1109/INMIC64792.2024.11004336
12. X. Lu, Q. Weng, Deep learning-based road extraction from remote sensing imagery: progress, problems, and perspectives, *ISPRS J. Photogramm. Remote Sens.*, **228** (2025), 122–140. https://doi.org/10.1016/j.isprsjprs.2025.07.013
13. D. Chakraborty, D. Dutta, C. S. Jha, Remote sensing and deep learning for traffic density assessment, In: C. S. Jha, A. Pandey, V. Chowdary, V. Singh, *Geospatial technologies for resources planning and management*, Water Science and Technology Library, Springer, 2022, 611–630.
14. W. Jiang, J. Luo, Graph neural network for traffic forecasting: a survey, *Expert Syst. Appl.*, **207** (2022), 117921. https://doi.org/10.1016/j.eswa.2022.117921

15. L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, et al., T-gcn: a temporal graph convolutional network for traffic prediction, *IEEE Trans. Intell. Transp. Syst.*, **21** (2020), 3848–3858. https://doi.org/10.1109/TITS.2019.2935152

16. B. Yu, H. Yin, Z. Zhu, Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting, *arXiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1709.04875

17. S. Guo, Y. Lin, N. Feng, C. Song, H. Wan, Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 922–929. https://doi.org/10.1609/aaai.v33i01.3301922

18. Z. Shao, Z. Zhang, W. Wei, F. Wang, Y. Xu, X. Cao, et al., Decoupled dynamic spatial-temporal graph neural network for traffic forecasting, *arXiv Preprint*, 2022. https://doi.org/10.48550/arXiv.2206.09112

19. Y. Li, W. Peng, J. Chen, H. Xu, Dynamic spatio-temporal attention-based graph neural network using ordinary differential equation and multi-scale semantics for traffic prediction, *IEEE Trans. Intell. Transp. Syst.*, **26** (2025), 19862–19875. https://doi.org/10.1109/TITS.2025.3612204

20. J. Shao, S. Li, K. Zhang, A. Wang, M. Li, Cross-city traffic prediction based on deep domain adaptive transfer learning, *Transp. Res. Part C: Emerg. Technol.*, **176** (2025), 105152. https://doi.org/10.1016/j.trc.2025.105152

21. S. Afandizadeh, S. Abdolahi, H. Mirzahossein, Deep learning algorithms for traffic forecasting: a comprehensive review and comparison with classical ones, *J. Adv. Transp.*, **2024** (2024), 9981657. https://doi.org/10.1155/2024/9981657

22. X. Luo, C. Zhu, D. Zhang, Q. Li, Stg4traffic: a survey and benchmark of spatial-temporal graph neural networks for traffic prediction, *arXiv Preprint*, 2023. https://doi.org/10.48550/arXiv.2307.00495

23. D. Zhang, F. Wang, L. Ning, Z. Zhao, J. Gao, X. Li, Integrating sam with feature interaction for remote sensing change detection, *IEEE Trans. Geosci. Remote Sens.*, **62** (2024), 4513011. https://doi.org/10.1109/TGRS.2024.3483775

24. J. Gao, D. Zhang, F. Wang, L. Ning, Z. Zhao, X. Li, Combining SAM with limited data for change detection in remote sensing, *IEEE Trans. Geosci. Remote Sens.*, **63** (2025), 5614311. https://doi.org/10.1109/TGRS.2025.3545040

25. Z. Tang, M. Naphade, M. Y. Liu, X. Yang, S. Birchfield, S. Wang, et al., Cityflow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/CVPR.2019.00900

26. W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, X. Feng, Multi-range attentive bicomponent graph convolutional network for traffic forecasting, *Proceedings of the AAAI conference on artificial intelligence*, **34** (2020), 3529–3536. https://doi.org/10.1609/aaai.v34i04.5758

27. Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: data-driven traffic forecasting, *arXiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1707.01926

28. M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. C. Chang, Y. Yao, et al., The 7th AI city challenge, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, 5538–5548. https://doi.org/10.1109/CVPRW59228.2023.00586

29. H. M. Hsu, Y. Wang, J. Cai, J. N. Hwang, Multi-target multi-camera tracking of vehicles by graph auto-encoder and self-supervised camera link model, *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2022, 489–499. https://doi.org/10.1109/WACVW54805.2022.00055

30. J. Ye, X. Yang, S. Kang, Y. He, W. Zhang, L. Huang, et al., A robust mtmc tracking system for ai-city challenge 2021, *021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, 4039–4048. https://doi.org/10.1109/CVPRW53098.2021.00456

31. N. Parashuram, K. Vijayalakshmi, Anchor-aware graph neural network with temporal attention for accurate traffic flow forecasting, *Int. J. Intell. Eng. Syst.*, **18** (2025), 283–297.

32. R. Kumar, J. M. Moreira, J. Chandra, Dygcn-lstm: a dynamic gcn-lstm based encoder-decoder framework for multistep traffic prediction, *Appl. Intell.*, **53** (2023), 25388–25411. https://doi.org/10.1007/s10489-023-04871-3

33. S. Shleifer, C. McCreery, V. Chitters, Incrementally improving graph wavenet performance on traffic prediction, *arXiv Preprint*, 2019. https://doi.org/10.48550/arXiv.1912.07390

34. L. Xiong, X. Yuan, Z. Hu, X. Huang, P. Huang, Gated fusion adaptive graph neural network for urban road traffic flow prediction, *Neural Process. Lett.*, **56** (2024), 9. https://doi.org/10.1007/s11063-024-11479-2