Context-Aware Feature-Fusion for Co-occurring Object Detection in Autonomous Driving

Binay Kumar Singh Niels Da Vitoria Lobo
Department of Computer Science
University of Central Florida
Orlando, USA
{binay.singh, niels.davitorialobo}@ucf.edu

Abstract

Object detection in autonomous driving requires precise localization and an inherent understanding of the relational context between co-occurring objects. In extremely complex heterogeneous environments rare classes, small-scale objects, and frequently appearing objects are difficult for standard object detection frameworks to handle. In this paper, we propose a novel framework called Context-Centric Feature Fusion (CCFF), which utilizes two attention-based modules, Local Context Fusion Module (LCFM) uses the RoI-to-RoI self-attention mechanism to resolve spatial interactions, mainly considering small and partially obscured objects, while Global Context Attention Module (GCAM) converts the co-occurrence of objects priors by pooling top-K RoI features into a global context attention token, avoiding the computational overhead of pixel-level global pooling. This fusion of local and object-centric global features yields contextualized embeddings that enhance classification results and co-occurring objects detection. Our method is evaluated on two datasets, Cityscapes and BDD100K which demonstrate significant improvement on relational consistency, achieving a Category-level Consistency Strategy (CCS) of 0.973 and 0.969, respectively. Furthermore, our approach produces substantial gains in small object detection (AP_S: 14.1%) and successfully recovers rare classes such as "Train" that are typically lost in large distributions. Our efficiency report shows that the framework processes images in real time with a 0.2 FPS overhead.

1. Introduction

A key element of autonomous driving is object detection, which enables the model to perform subsequent operations [6] such as tracking, motion forecasting, planning, and decision-making. Modern detectors, such as [3] have achieved bottleneck performance in benchmark datasets, however, real driving scenes remain challenging due to frequent occlusions, cluttered intersections, diverse categories, and visually ambiguous instances (e.g., partially visible pedestrians or overlapping vehicles) [9]. In such scenarios, accurate detection depends not only on the local appearance of an object but also on the context [14] in which it appears. In modern object detectors, small and low-frequency objects [8], and objects that appear under heavy occlusion or in dense traffic scene, remain a challenging problem.

Inevitably, the structure of roads and contexts [11] in which autonomous driving operates is very much fixed in real-life environments. In many road scenes, intersections and occluded objects co-occur, pedestrians are commonly seen across sidewalks and crosswalks, and dense traffic patterns often indicate particular object distributions. These co-occurrence features presented here provide a strong prior guideline that can determine these complex environments. However, many traditional detection pipelines still primarily rely on region-level appearance features extracted around each proposal, which limits their ability to exploit object-object relationships and global scene cues.

Based on these rationale, we propose a model that understands how objects appear together in a scene. Our main contribution in this research work is as follows.

- We present a novel **Context-Centric Feature Fusion (CCFF)** framework that combines local context fusion module and object-centric global context module. This helps the model to learn about local and global contextual information required to enhance co-occurring object detection performance.
- First, we introduce **Local Context Feature Module (LCFM)** by using RoI-to-RoI attention to model nearby object interactions that commonly occurs in crowded traffic scenes.
- Second, our object-centric **Global Context Attention Module (GCAM)** that integrates global context through attention pooling over RoI features, explicitly encode

object co-occurrence priors in high-level environmental cues relevant to autonomous driving. Finally, by fusing on original RoI appearance features with locally and globally enhanced context features we produce **context-enriched embeddings** for classification and contextual priors.

- We evaluate our proposed work on Cityscapes and BDD100K datasets, demonstrating improved detection performance and co-occurring objects identification, specifically for partially occluded objects.

2. Related Work

2.1. Attention-Based Contextual Reasoning

For visual recognition tasks, transformer-based attention mechanisms have become an interesting tool for modeling long-range dependencies. By incorporating global contextual information, transformer-based architectures and attention-augmented convolutional networks [12] have proven to perform well. Attention has been used at the feature-map level in object detection to capture semantic and spatial context [2, 7]. Despite their effectiveness, these techniques lack explicit reasoning over instance-level representations generated by region-based detectors and usually work with dense backbone features.

2.2. Instance-Level and Relational Context

To model object-object instance-level visual relationships, several studies have been investigated in the literature. Attention mechanisms are introduced to capture pairwise interactions between detected objects by methods such as Relation Networks [7] and object relation modules. Understandably, techniques mentioned here improve object localization and recognition by benefiting from instance-level context. The majority of relational reasoning frameworks, however, are only suitable as lightweight plug-and-play modules within conventional two-stage detectors because they are either made for relationship prediction tasks or require significant architectural changes.

2.3. Global Context in Two-Stage Detectors

To supplement local appearance features with scene-level cues, object detectors have also integrated global context. Previous research usually uses pooling over whole feature maps [2, 12] or other context branches [1] to aggregate global information. Recently introduced transformer-based detectors use self-attention across tokens to implicitly encode global context [3, 16]. However, the methods mentioned here frequently have higher computational costs or architectural complexity and do not explicitly model object co-occurrence priors at the ROI level.

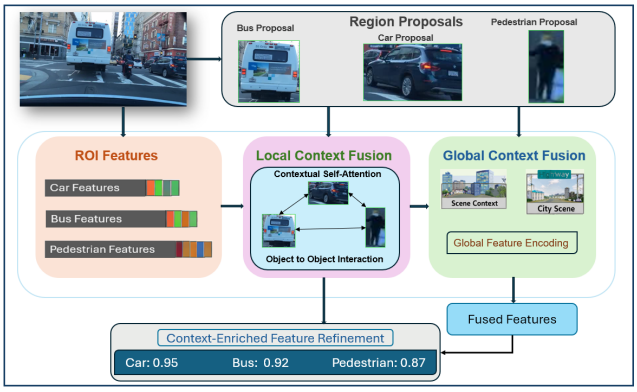

Figure 1. **Detailed schematic flow of the proposed Context-Centric Feature Fusion (CCFF) architecture.** Region proposals from the FPN are enhanced using parallel dual-stream contextual reasoning channels. Our (1) the Local Context Fusion Module (LCFM), handles localized spatial object interactions using regional self-attention, (2) the Global Context Attention Module (GCAM), maps global environmental priors by executing attention pooling over the top-K confident proposals biased by their normalized bounding coordinates. Then fused, context-enriched features pass directly to classification and regression heads to resolve ambiguities without relying on sequential tracking streams.

2.4. Query-Based Vision Transformers vs. Region Priors

Recent paradigms in perception have shifted toward query-based transformer detectors, such as Deformable DETR [3] and DINO [16]. While these networks implicitly capture small environment constraints layer-by-layer via broad self-attention maps, they suffer from prolonged training convergence, massive computational overhead, and a high deployment overload. For highly active autonomous systems, extracting explicit, instance-level spatial co-occurrences directly within structured region-of-interest (RoI) bounds offers a much more deployment-viable alternative. Our CCFF framework bridges this gap by introducing local-global relational reasoning strictly inside the local prediction head, resolving complex spatial occlusions without the underlying model complexity or convergence overhead typical for pure vision transformers.

3. The Proposed Method

3.1. Overview

Our proposed framework Context-Centric Feature Fusion (CCFF) is presented in Fig. 1. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we build our framework with a Detectron2 backbone and Feature Pyramid Network (FPN) that extracts multi-scale feature maps $\{F_\ell\}_{\ell=1}^{L}$, and a Region Proposal Network (RPN) produces a set of region proposals $\mathcal{P} = \{p_i\}_{i=1}^{N}$. Each proposal is mapped to a fixed-

dimensional RoI feature via RoIAlign:

$$\mathbf{r}_i = \text{RoIAlign}(F, p_i) \in \mathbb{R}^d, \tag{1}$$

where F denotes the respective pyramid feature level(s) selected for the proposal.

While conventional RoI heads process each $\mathbf{r}_i$ largely independently, our goal is to enrich RoI representations with contextual cues that are critical in autonomous driving scenes. To this end, we introduce two complementary context modules: (i) *local context modeling* via RoI-to-RoI attention and (ii) *object-centric global context modeling* via attention pooling over RoIs to encode co-occurrence priors at the scene level [8]. The resulting features are fused and passed to standard classification and bounding box regression heads.

3.2. Local Context Feature Modeling via RoI-to-RoI Attention

In fixed-range contextual features there is a limitation, addressed recently in Dynamic Context Exploration (DCE) [14] that proposed sensing local information dynamically. In contrast, our Local Context Feature Module (LCFM) utilizes RoI-to-RoI self-attention to resolve spatial interactions without the need for manual surrounding searches. For that we need to model object interactions among the proposals generated earlier, by applying self-attention mechanism over the features generated by RoI. Let $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_N]^\top \in \mathbb{R}^{N \times d}$. Then we compute the following query, key, and value embeddings using learnable linear projections based on self-attention expessed as:

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{r}_i, \quad \mathbf{k}_j = \mathbf{W}_k \mathbf{r}_j, \quad \mathbf{v}_j = \mathbf{W}_v \mathbf{r}_j, \tag{2}$$

where, $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_a \times d}$ and d_a is the attention embedding dimension. Then attention weight from RoI i to RoI j is computed as:

$$\alpha_{ij} = \frac{\exp\left(\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_a}\right)}{\sum_{m=1}^{N} \exp\left(\mathbf{q}_i^\top \mathbf{k}_m / \sqrt{d_a}\right)}. \tag{3}$$

The local context feature for RoI i is obtained by aggregating value embeddings:

$$\mathbf{c}_i^{\text{loc}} = \sum_{j=1}^{N} \alpha_{ij} \mathbf{v}_j. \tag{4}$$

Finally, we form the locally enhanced RoI representation via residual fusion:

$$\tilde{\mathbf{r}}_i = \mathbf{r}_i + \mathbf{W}_{\text{loc}}\, \mathbf{c}_i^{\text{loc}}, \tag{5}$$

where, $\mathbf{W}_{\text{loc}} \in \mathbb{R}^{d \times d_a}$. This module enables each RoI to incorporate information from other objects in the scene, improving robustness in crowded and occluded driving scenarios.

3.3. Object-Centric Global Context Attention Modeling with Geometry Bias

To capture scene-level dependencies, we employ a global context fusion strategy inspired by the query-independent formulation presented in [2], which allows efficient long-range modeling without the computational overhead of standard non-local blocks. This global prior is then integrated with local features presented earlier to resolve semantic ambiguities in complex urban environments [14].

Local context captures short-range interactions but does not provide a compact representation of the overall object configuration of the scene. We therefore, construct an *object-centric global context vector* by aggregating RoI features rather than pixel-level feature maps in Global Context Attention Module (GCAM).

3.3.1. Top-K selection

From the locally enhanced RoIs $\{\tilde{\mathbf{r}}_i\}_{i=1}^N$ in LCFM presented earlier, we select the top-K proposals based on objectness or classification confidence, yielding:

$$\mathcal{S} = \{(\tilde{\mathbf{r}}_k, \mathbf{b}_k)\}_{k=1}^K, \quad K \ll N, \tag{6}$$

where $\mathbf{b}_k = (x_k, y_k, w_k, h_k)$ denotes the corresponding RoI box in image coordinates.

3.3.2. Geometry-aware attention scoring

We incorporate a lightweight geometry bias into the attention logits. Specifically, we define a normalized geometry encoding:

$$\phi(\mathbf{b}_k) = \left[\frac{x_k}{W}, \frac{y_k}{H}, \log\frac{w_k}{W}, \log\frac{h_k}{H}\right] \in \mathbb{R}^4, \tag{7}$$

here, W, H are the image width and height. The geometry-aware attention logit for RoI k is:

$$s_k = \underbrace{\mathbf{u}^\top \sigma(\mathbf{W}_g \tilde{\mathbf{r}}_k)}_{\text{content}} + \underbrace{\mathbf{v}^\top \phi(\mathbf{b}_k)}_{\text{geometry bias}}, \tag{8}$$

where $\mathbf{W}_g \in \mathbb{R}^{d_g \times d}$, $\mathbf{u} \in \mathbb{R}^{d_g}$, $\mathbf{v} \in \mathbb{R}^4$ are learnable parameters, and $\sigma(\cdot)$ is a nonlinearity (e.g., ReLU). Then we normalize logits after applying Softmax.

The object-centric global context is computed as an attention-weighted sum of RoI features:

$$\mathbf{g}^{\text{obj}} = \sum_{k=1}^{K} \beta_k \tilde{\mathbf{r}}_k \in \mathbb{R}^d. \tag{9}$$

The above Eq. (9) encourages the model to emphasize RoIs whose spatial layout is informative for driving scenes (e.g., traffic lights frequently appear in upper regions), while remaining lightweight and fully compatible with end-to-end training.

3.3.3. Global context injection

We broadcast $\mathbf{g}^{\text{obj}}$ to all RoIs using a learnable projection as follows:

$$\hat{\mathbf{r}}_i = \tilde{\mathbf{r}}_i + \mathbf{W}_{\text{glob}}\, \mathbf{g}^{\text{obj}}, \tag{10}$$

where $\mathbf{W}_{\text{glob}} \in \mathbb{R}^{d\times d}$.

3.4. Context Feature Fusion and Detection Heads

We fuse appearance, local, and object-centric global context cues to form the final RoI representation. Using the original RoI feature $\mathbf{r}_i$, the local context feature $\mathbf{c}_i^{\text{loc}}$ from Eq. (4), and the global context vector $\mathbf{g}^{\text{obj}}$ from Eq. (9), we compute:

$$\mathbf{f}_i = \text{MLP}\Big([\mathbf{r}_i \,\|\, \mathbf{c}_i^{\text{loc}} \,\|\, \mathbf{g}^{\text{obj}}]\Big), \tag{11}$$

where $\|$ denotes concatenation and $\text{MLP}(\cdot)$ is a lightweight projection network. The fused feature $\mathbf{f}_i$ is fed into standard detection heads:

$$\hat{\mathbf{p}}_i = \text{ClsHead}(\mathbf{f}_i), \qquad \hat{\mathbf{b}}_i = \text{RegHead}(\mathbf{f}_i), \tag{12}$$

where $\hat{\mathbf{p}}_i$ denotes class probabilities and $\hat{\mathbf{b}}_i$ denotes predicted bounding boxes.

Finally, CCFF is trained end-to-end and the following loss function is calculated.

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{reg}}, \tag{13}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss, $\mathcal{L}_{\text{reg}}$ is the regression loss, and λ is a hyperparameter used to balance the two terms.

Because the local attention module (Eqs. (2)–(5)) and geometry-aware object-centric attention pooling (Eqs. (6)–(10)) are differentiable, gradients from Eq. (13) propagate through the fusion stage (Eq. (11)) into both context modules and the backbone. The top-K selection in Eq. (6) is discrete; however, the attention pooling weights and subsequent fusion remain fully differentiable and train stably in practice.

4. Experimental Results and Analysis

4.1. Evaluation Metrics

We evaluate our proposed model by considering two specialized metrics beyond standard object detection metrics. Our Co-occurring AP (CoAP) measures the precision of detections within contextual pairs, focusing on the model's ability to resolve individual objects through their spatial and semantic relationships [10], and Category-level Consistency Strategy (CCS) [15] quantify the alignment between the predicted and ground-truth object co-occurrence distributions [4]. This indicates that standard AP treats detection as independent events, while these metrics collectively demonstrate how effectively our Local Context Fusion Module (LCFM) and Global Context Attention Module (GCAM) capture the underlying structural logic of complex driving scenes, ensuring that high-confidence detections remain semantically and contextually enhanced as proposed in our research.

Table 1. Comparative Analysis of the proposed approach on Cityscapes (all variants)

Experiment	AP ↑	AP_{50} ↑	AP_{75} ↑	AP_S	AP_M	AP_L	CoAP	CCS
Baseline	**36.44**	**60.59**	**36.11**	**12.03**	34.68	**60.89**	0.386	0.972
Local	35.34	59.12	34.12	11.08	34.88	59.72	0.382	0.972
Global	35.41	59.25	34.22	11.12	34.92	59.81	0.383	0.972
Ours	35.47	59.35	34.39	11.19	**34.99**	59.87	**0.389**	**0.973**
Ours_geom	35.51	59.42	34.45	11.22	35.05	59.92	0.385	0.972

4.2. Implementation Details

Our proposed CCFF framework is implemented using two-stage detector framework- Detectron2 that works on Faster R-CNN based FPN backbone. Then the proposed modules-LCFM and GCAM are inserted into the RoI head, as mentioned earlier, and trained end-to-end using the standard detection losses, reported in Eq. (13). Unless otherwise stated, we adopt the default Detectron2 model training schedule and augmentations commonly used for these datasets. We set the number of selected proposals for global context $K = \{32 \text{ or } 64\}$ and keep the attention embedding dimension d_a lightweight to limit overhead. We used Cityscapes [5] and BDD100K [13] datasets with resolutions (2048×1024), and (1280×720), respectively. All models are trained using SGD with a fixed learning rate schedule. To ensure fair comparison among the proposed model variants, we considered the same backbone, optimizer settings, batch size, and categories.

4.3. Quantitative Analysis on Cityscapes

We reported the quantitative performance analysis of our *CCFF* variants on the Cityscapes validation set in Table 1. This shows that the baseline achieves overall AP performance gain, our *Ours* variant demonstrates better capabilities in relational reasoning and scale-specific detection. We report scale-specific precision and our model demonstrates strong retrieval capabilities, particularly for medium-scale entities. Similarly, the proposed CCFF framework not only localizes objects accurately but also maintains a high capture rate for critical road factors.

Relational Consistency and Co-occurrence: As reported in Row 4 in Table 1, **Ours** variant achieves a **CoAP** of **0.389** and a **CCS** of **0.973**, outperforming all other configurations, including the baseline. This improvement indicates that our dual-stream context fusion successfully captures semantic dependencies between co-occurring objects. Additionally, the higher CCS score indicates that our model maintains higher spatial rank correlation, which is impor-

tant for scene understanding in autonomous driving.

Scale-Specific Performance: In **(AP_M)** column of Table 1, *Ours* variant reaches **34.99%** surpassing the baseline's 34.68%. This suggests that our object-centric fusion logic is particularly effective for medium-range interactions, where objects are close enough to provide mutual context. From this, we observe that feature enhancement from neighboring instances is required to resolve ambiguities.

Ablation of Fusion Variants: Comparing the *Local* and *Global* variants reveals that global scene context (AP 35.41%) provides a slight edge over purely local neighborhood features (AP 35.34%). However, fusion of both in *Ours* model provides the most balanced precision-context trade-off. Finally, the *Ours_geom* variant gives the highest raw AP (35.51%) among our proposed heads, ensuring that geometric priors such as distance and orientation further refine bounding box localization.

4.4. Quantitative Analysis on BDD100K

Our Table 2 shows the performance of the proposed variants on the BDD100K dataset. Here, our object-centric approach demonstrates superior efficacy in capturing complex urban dependencies and resolving small-object localization challenges inherent in the BDD100K benchmark.

Relational Integrity and Spatial Consistency: The proposed *Ours* variant (shown in bold in Row 4) achieves **CoAP** of **0.488** and **CCS** of **0.969**, significantly outperforming the baseline. This lead indicates that our dual-stream fusion head is uniquely robust at maintaining spatial ranking and contextual precision, even in the diverse lighting and weather conditions (main characteristic of BDD100K dataset).

Small Object Detection Gains: A significant performance improvement is observed in the small object category of column (AP_S) in *Ours* method is **14.73%**. Here, by leveraging relational information from high-confidence neighboring object proposal embeddings, our model effectively improves the feature representation of distant or ambiguous objects, and as a result, our final overall **AP** is **32.95%**, establishing a competitive baseline for object-centric autonomous driving research.

Geometric Refinement: By adding relative spatial priors in the *Ours_geom* variant (reported in Row 5 in Table 2), further optimization of localization is achieved, yielding an overall **AP** of **32.21%**. These results show that explicit modeling of distance and orientation in highly competitive environments provides a crucial inductive bias for original scene understanding.

Contextual Constraints and Transformer Baselines: While global query-based transformers implicitly model multi-scale features, their lack of explicit localized geometric constraints often makes tiny, low-frequency, or heavily occluded categories difficult to capture under strict edge latency scenarios. In contrast, our dual-stream CCFF head explicitly adds geometric inductive biases into the regional level Eq. (8). This design allows our model to achieve a significant performance enhancement in small-scale category (AP_S: 14.73% on BDD100K) over standard two-stage baselines. Crucially, as detailed in our efficiency benchmarks (Sec. 4.5), this structural precision is achieved with an almost negligible latency penalty of $\approx$ 0.2 FPS. This indicates that dedicated context-centric feature fusion yields superior performance-to-latency trade-offs for real-world driving environments over generic, unconstrained pixel-level transformer baselines.

Table 2. Object Detection and Category-level Consistency Strategy Results on BDD100K (all variants)

Variant	AP ↑	AP_{50} ↑	AP_{75} ↑	AP_S	AP_M	AP_L	CoAP	CCS ↑
Baseline	31.02	54.21	30.50	12.12	32.08	52.12	0.408	0.952
Local	31.21	30.08	30.81	12.44	33.11	52.42	0.410	0.954
Global	31.41	54.81	31.09	12.06	33.42	52.81	0.415	0.955
Ours	**32.95**	**56.89**	**32.92**	**14.73**	**35.88**	**54.94**	**0.488**	**0.969**
Ours_geom	32.21	56.28	32.51	14.34	35.51	54.81	0.431	0.965

We evaluate the effectiveness of our proposed *CCFF* framework by visualizing the semantic relationships captured during inference on the Cityscapes dataset. As illustrated in Fig. 2, we present the original input image (left) alongside the corresponding detections and co-occurrence edges generated by our framework (right). Our model explicitly maps co-occurrences of objects through a network of colored semantic edges, such as *person* $\leftrightarrow$ *car* and *person* $\leftrightarrow$ *bicycle*. Using CCS, the framework is able to reason about the structural layout of the scene, achieving high-confidence detections even for challenging "tail" categories (e.g., *rider* 89%, *motorcycle* 99%) in cluttered urban environments. The dense network of edges demonstrates how our local-global feature fusion leverages co-occurrence data to resolve ambiguities and provide a more holistic understanding of the scene compared to standard localized detection pipelines.

Similarly, Fig. 3, visualizes the semantic relationship captured during inference on the Cityscapes dataset. Here, our proposed framework exhibits high robustness in complex street scenes containing diverse object scales. In this scenario, the *CCFF* accurately detects a large-scale tram (**train 99%**) while simultaneously identifying several smaller cars and pedestrians in the middle-to-background regions. By using CCS, our model generates a dense network of edges that link these heterogeneous objects. Specifically, the presence of the highly-confident train detection acts as a contextual anchor, where the resulting semantic edges—including the *person* $\leftrightarrow$ *car* (red) and other co-occurrences (white)—provide global relational cues that reinforce the detection of more distant or partially occluded instances. This explicit modeling of spatial and semantic

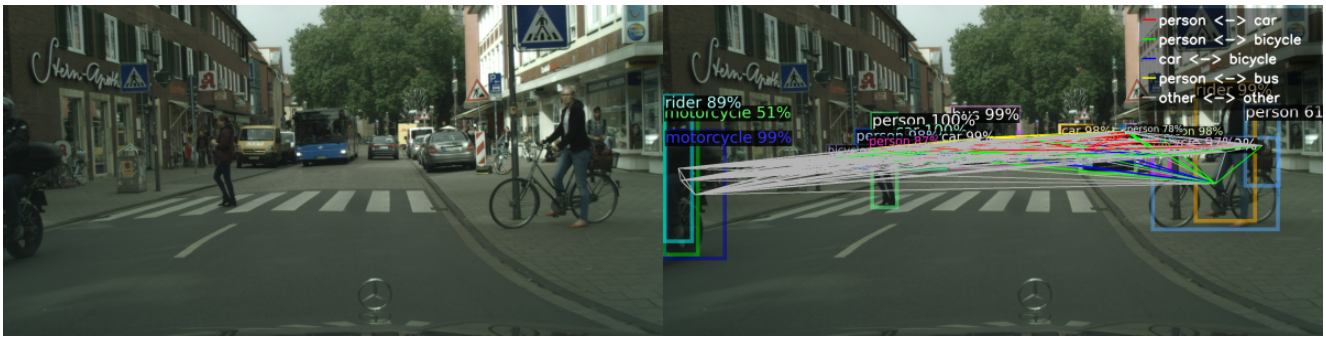

Figure 2. **Qualitative visualization of semantic co-occurrence links during inference on Cityscapes.** The left panel displays the raw input scene. The right panel illustrates finalized model predictions with our explicit relational logic links. Line colors distinguish discrete category configurations (e.g., red for *person* ↔ *car*, green for *person* ↔ *bicycle*). Line widths reflect attention confidence, demonstrating how the model utilizes highly visible anchors to stabilize the classification of ambiguous background elements.

proximity ensures that the final predictions remain consistent with the structural layout of the urban environment, achieving a more holistic scene interpretation than traditional, localized detection methods.

4.5. Efficiency Analysis

An efficiency analysis is reported in Table 3 reveals that our *CtxFusion* module adds 5.38M parameters, representing a manageable 13% increase in model size over the baseline. In terms of latency, the overhead remains minimal across varying resolutions: BDD100K (1280×720) incurs an additional 7.7 ms, while the high-resolution Cityscapes (2048×1024) sees a 16.5 ms increase. Consequently, our method maintains real-time viability with only a marginal FPS reduction ($\approx$ 0.2), demonstrating that our object-centric improvements do not sacrifice practical deployment speed.

Table 3. Model Complexity and Inference Efficiency. All models use ResNet-50 + FPN backbone and are evaluated on a single GPU.

Dataset	Variant	Params (M)	Lat. (ms)	FPS
Cityscapes	Baseline	41.33	308.04	3.25
	Ours	**46.71**	**324.57**	**3.08**
BDD100K	Baseline	41.33	160.53	6.23
	Ours	**46.71**	**168.24**	**5.94**
Overhead (Δ)	*Added*	*+5.38*	*+7.7–16.5*	*≈-0.2*

4.6. Ablation Study

We conduct ablation study to evaluate the performance of our *CtxFusion* components, as summarized in Table 4. The results indicate that while individual **Local** or **Global** modules establish stable relational baselines, their combination is required to effectively capture complex urban dependencies.

Effect of Dual-Stream Fusion: By analyzing the results reported in Table 4 we can understand that integrating local neighborhood interactions with global scene context (Row 3) provides the most significant performance leap. In the diverse BDD100K dataset, this dual-stream approach achieves a peak AP of 32.95 and a substantial boost in contextual precision (CoAP: 0.488). This suggests that modeling the relationship between an object and its immediate neighbors, while simultaneously accounting for the overall scene configuration, is essential for resolving ambiguities in complex driving environments.

Role of Geometric Priors: The addition of geometric embeddings $\phi(\mathbf{b}_k)$ in the **Ours_geom** variant (Row 4) acts as a structural stabilizer. While the content-only fusion in Row 3 provides higher raw AP on BDD100K, the geometric variant maintains high reliability across datasets. By anchoring features to physical coordinates and scales, the model moves toward a structural reasoning approach that effectively filters background noise and resolves position ambiguities for distant objects.

Co-occurrence Supervision: We observe that the model's ability to maintain high *CoAP* and *CCS* scores depend on the joint optimization of the local and global modules. Even without a separate supervision term, the dual-stream architecture implicitly learns semantic co-occurrence priors.

Table 4. Ablation study of context components on Cityscapes and BDD100K validation sets.

Components			Cityscapes			BDD100K		
Loc.	Glob.	Geom.	AP	CoAP	CCS	AP	CoAP	CCS
Baseline			**36.44**	0.386	0.972	31.02	0.408	0.952
✓	×	×	35.34	0.382	0.972	31.21	0.410	0.954
×	✓	×	35.41	0.383	0.972	31.41	0.415	0.955
✓	✓	×	35.47	**0.389**	**0.973**	**32.95**	**0.488**	**0.969**
✓	✓	✓	35.51	0.385	0.972	32.21	0.431	0.965

Figure 3. **Qualitative results illustrating scale robustness in heterogeneous urban driving environments.** The left panel represents the raw street environment. The right panel displays the corresponding CCFF inference result. Our model simultaneously maps the co-occurrences by utilizing a highly confident macro landmark (e.g., the 99% confidence *train* extraction) as a structural anchor. The co-occurrence links (e.g., yellow line represents *person* $\leftrightarrow$ *bus*) act as a contextual representation to confidently identify and recover distant or heavily occluded background instances.

5. Conclusion

In this paper, we introduce Context-Centric Feature Fusion (CCFF), a framework that improves autonomous driving perception by leveraging explicit relational reasoning. Rather than treating objects in isolation, CCFF uses two key components: the Local Context Fusion Module (LCFM) to capture spatial relationships between regions, and the Global Context Attention Module (GCAM) to model how objects typically co-occur. Together, these modules shift the focus from independent detection to holistic scene understanding. Our evaluations on Cityscapes and BDD100K show significant gains, especially for small or rare objects that are often misidentified. Furthermore, a higher CCS confirms that our model successfully learns the underlying structural layout of urban environments.

References

[1] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2874–2883, 2016. 2

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1971–1980, 2019. 2, 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 1, 2

[4] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4087–4096, 2017. 4

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 4

[6] Xuyao Guo, Feng Jiang, Quanzhen Chen, Yuxuan Wang, Kaiyue Sha, and Jing Chen. Deep learning-enhanced environment perception for autonomous driving: MDNet with CSP-DarkNet53. *Pattern Recognition*, 160:111174, 2025. 1

[7] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018. 2

[8] Guoguang Hua, Fangfang Wu, Guangzhao Hao, Chenbo Xia, and Li Li. ELFT: Efficient local-global fusion transformer for small object detection. *PLoS ONE*, 20(9):e0332714, 2025. 1, 3

[9] Jiayao Li, Chak Fong Cheang, Xiaoyuan Yu, Suigu Tang, Zhaolong Du, and Qianxiang Cheng. A segmentation network for enhancing autonomous driving scene understanding using skip connection and adaptive weighting. *Scientific Reports*, 15(1):36692, 2025. 1

[10] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 852–860, 2016. 4

[11] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014. 1

[12] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 2

[13] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-

rell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. 4

[14] Ziji Zhang, Ping Gong, Haotian Sun, Pingping Wu, and Xuanyuan Yang. Dynamic local and global context exploration for small object detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 1, 3

[15] Siyu Zhu, Yingjie Tian, Fenfen Zhou, Kunlong Bai, and Xiaoyu Song. COCM: Co-occurrence-based consistency matching in domain-adaptive segmentation. *Mathematics*, 10(23):4468, 2022. 4

[16] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 2