QualiNet: Acquiring Bird's Eye View Qualitative Spatial Representation from 2D Images in Automated Vehicle Perception

Nassim Belmecheri
Simula Research Laboratory, Oslo, Norway

Abstract

We present QualiNet, an end-to-end deep learning framework that acquires Bird's Eye View (BEV) qualitative spatial relations directly from 2D images, eliminating the need for depth sensors. The system combines 2D object detection, masking, and classification to infer Rectangle Algebra (RA) and Qualitative Distance Calculus (QDC) relations. Evaluated on NuScenes and PandaSet datasets, QualiNet achieves 91% accuracy for RA, 80% for QDC, and 99% top-2 accuracy, demonstrating robust performance for automated vehicle perception.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Spatial and physical reasoning; Computing methodologies → Scene understanding

Keywords and phrases Qualitative Spatial Representation, Deep Learning, Computer vision, Qualitative Scene Understanding, Spatio-temporal representation and reasoning models (including moving objects tracking)

Digital Object Identifier 10.4230/LIPIcs.TIME.2025.14

Category Short Paper

Supplementary Material *Software (Source code)*: `https://github.com/nassimbel/QualiNet.git` [3] archived at `swh:1:dir:a4900663aeb84632699b0217f7a7f98014466c00`

Funding This work is funded by the European Commission through the AI4CCAM project (Trustworthy AI for Connected, Cooperative Automated Mobility) under grant agreement No 101076911.

Acknowledgements I would like to thank my colleagues Arnaud Gotlieb, Nadjib Lazaar and Helge Spieker for the fruitful discussions and continuous support.

1 Introduction

Automated vehicle perception traditionally relies on quantitative methods that struggle with complex real-world scenarios [1]. Qualitative representations offer a promising alternative by capturing relative spatial relationships through calculi like Rectangle Algebra (RA) and Qualitative Distance Calculus (QDC) [15]. These methods simplify complex spatial information while aligning with human reasoning patterns [16, 2].

Current approaches for building qualitative representations [8, 4] typically require expensive sensors (LiDAR, depth cameras) and significant computational resources. We present QualiNet, an end-to-end deep learning framework that acquires Bird's Eye View (BEV) qualitative representations directly from 2D images using object detection and classification. Our method is validated on NuScenes [6] and PandaSet [18] datasets.

2 Related Work

Recent advances have demonstrated the effectiveness of qualitative representations across multiple domains. In action recognition, qualitative state transitions [19, 14] and relation chains [12] have proven valuable for capturing action semantics. For autonomous driving,

neuro-symbolic integration [17] and BEV qualitative representations [2] enhance both scene understanding and system explainability. Spatial knowledge acquisition methods show particular diversity, ranging from implicit templates [9] to force histograms [5] and hybrid symbolic-neural approaches [10]. While current methods typically depend on 3D sensors [13], our QualiNet system uniquely acquires BEV spatial relations directly from 2D images, eliminating the need for specialized depth sensing hardware.

Background

Perception in Automated Vehicles

Perception is crucial for automated vehicles to understand and interact with their environment. This involves processing data from various sensors like LiDAR, radar, and cameras [1].

2D perception, primarily using cameras, analyzes images to identify objects but lacks depth information, making distance and size estimation challenging [11].

3D perception overcomes this limitation by incorporating depth information from stereo cameras or LiDAR. This enables accurate spatial understanding and object localization, vital for safe and reliable autonomous navigation [7].

Qualitative Spatial Calculus

A qualitative calculus operates over domain $\mathcal{D}$ (e.g., $\mathbb{R}^2$) with binary relations $\Gamma = \{r_1, \ldots, r_m\}$ that are jointly exhaustive and pairwise disjoint. We employ:

- **QDC** [15]: Distance relations (very close, close, normal, far)
- **RA** [15]: Rectangle relations from Allen's Interval Algebra (before, meets, overlaps, etc.) on x/y axes

A scene $S = (V, O, R)$ consists of frames V, objects O, and their relations R.

Transforming Images into BEV Qualitative Constraint Networks

We represent detected objects and their spatial relations as a qualitative graph $\mathcal{G} = (\mathcal{O}, \mathcal{R})$, where $\mathcal{O}$ is the set of objects and $\mathcal{R}$ their qualitative relations from language Γ. Each object belongs to a single category (e.g., car, pedestrian).

Using the ego vehicle as reference object o_0, we construct a star graph $\mathcal{G}^* = (\mathcal{O}, \mathcal{R}^*)$ with relations between o_0 and other objects. The complete graph $\mathcal{G}$ is built through relation composition: $R_{ij} = R_{0i} \circ R_{0j} \quad \forall (o_i, o_j) \in \mathcal{O}^2, i \neq j$ using path consistency enforcement [15] until convergence.

Image to Relation Data Construction

We denote by $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{R}_i^*, o_0)\}_{i=1}^N$ a dataset of tuples consisting of 2D images $\mathcal{I}_i$, qualitative relations $\mathcal{R}_i^*$ between BEV detected objects and the reference object o_0 (ego vehicle). This dataset serves as the training data for QualiNet, enabling the model to learn the mapping between visual information and qualitative spatial relations.

Consider `I2BEVQR` as a function that maps a 2D image $\mathcal{I}$ to the set of qualitative relations $\mathcal{R}^*$ between the detected objects and the reference object o_0. `I2BEVQR` takes a 2D image $\mathcal{I}$ as input and outputs the set of qualitative relations $\mathcal{R}^*$ between the detected objects and the reference object o_0. $\mathcal{R}^*$ represents the labels for $\mathcal{I}$. For each image, the function extracts the objects with their categories and bounding boxes, including the ego-vehicle. It

then constructs a star graph with the ego-vehicle as the central node and other objects as peripheral nodes. Using the QXG-builder tool [4, 2] and the specified algebra, it determines the qualitative spatial relations between the ego-vehicle and each object. These relations, along with the image and the ego-vehicle information, are added to the dataset.

Learning to Acquire Qualitative Relations

We define $\mathcal{M}$ as a function that takes a 2D image $\mathcal{I}$ and returns a binary mask M indicating the presence of detected objects: $\mathcal{M} : \mathcal{I} \rightarrow M$ where M is a binary matrix of dimensions $H \times W$ such that:

$$M(i,j) = \begin{cases} 1 & \text{if pixel } (i,j) \text{ belongs to a detected object} \\ 0 & \text{otherwise} \end{cases}$$

This masking process helps to focus the attention of the deep learning model on the relevant regions of the image, improving the accuracy and efficiency of spatial relation extraction.

The QualiNet (Algorithm 1) takes as input the training dataset, learning rate, number of epochs, and predefined architectures for the CNN and the MLPclassifier. It outputs a trained QualiNet model.

Algorithm 1 QualiNet Training Algorithm.

Input: Dataset $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{R}_i^*, o_{0i})\}_{i=1}^N$, Epochs E, CNN architecture, MLPClassifier architecture
Output: Trained QualiNet model $\mathcal{F}_{\text{QualiNet}}$
Initialize CNN with a predefined architecture
Initialize Classifier with a predefined architecture
for *epoch* $\leftarrow 1$ **to** E **do**
  for $(\mathcal{I}, \mathcal{R}^*, o_0) \in \mathcal{D}$ **do**
    for $o_j \in \mathcal{O} \setminus \{o_0\}$ **do**
      $M \leftarrow \mathcal{M}(\mathcal{I}, o_j)$
      $F_j \leftarrow \text{CNN}(M)$
      $\hat{R}_{0j} \leftarrow \text{MLPClassifier}(F_j)$
      $L \leftarrow \text{Loss}(\hat{R}_{0j}, \mathcal{R}_j^*)$
      Update parameters of CNN and Classifier using gradient descent
return $\mathcal{F}_{\text{QualiNet}}$

3 Experiments

This section outlines the experimental setup and results for evaluating QualiNet's performance. Since our approach is novel, there are no direct baselines for comparison. Instead, we focus on showcasing QualiNet's capabilities and analyzing its behavior. Our evaluation aims to investigate how accurate is QualiNet's top predictions (Top 1 and Top 2 accuracy)?

We used the NuScenes [6] dataset for evaluation. NuScenes includes diverse urban scenes captured from multiple sensors. We utilized both the full dataset (NuScenes-Large) and a smaller subset (NuScenes-mini).

QualiNet is implemented in PyTorch using a ResNet-152 CNN for feature extraction and an MLP for classification. The model is trained with SGD and a learning rate scheduler. Data is split 70:30 for training and validation, and performance is evaluated over five runs using Top 1 and Top 2 accuracy.

The original dataset exhibited severe imbalance: QDC "far" (42% samples) vs "very close" (3%). After augmentation, all relations have 500±20 samples.

We assess QualiNet's performance using the following metrics:

Accuracy (Top 1): Percentage of correct top predictions. **Accuracy (Top 2):** Percentage of cases where the correct relation is among the top two predictions.

All datasets were transformed using the *I2BEV* function to generate the training and testing data. During data construction, impossible RA relations for each camera were removed to ensure the training data accurately reflects observable spatial relationships.

The code for QualiNet and the experiments presented in this paper is available at: `https://drive.google.com/drive/folders/1K9ViuyM4s_IwkcaCd3b1KYAhh0P3j1BH?usp=sharing`

3.1 Results

The results presented in this section are averaged over all the datasets test sets used in the evaluation.

Table 1 Top-1 and Top-2 Accuracy by camera: CF (Front), CFL (Front-Left), CFR (Front-Right), CB (Back), CBL (Back-Left), CBR (Back-Right).

Sensor	Relation	Top-1 Accuracy	Top-2 Accuracy
CF	RA	0.93	0.98
CFL	RA	0.92	0.96
CFR	RA	0.92	0.96
CB	RA	0.94	0.99
CBL	RA	0.88	0.93
CBR	RA	0.89	0.96
CF	QDC	0.78	0.94
CFL	QDC	0.78	0.93
CFR	QDC	0.77	0.93
CB	QDC	0.77	0.94
CBL	QDC	0.78	0.94
CBR	QDC	0.78	0.93

Table 1 presents the Top-1 and Top-2 accuracy of QualiNet for different camera sensors and relation types. As shown, the Top-1 accuracy ranges from 76% to 94%. However, the Top-2 accuracy is consistently higher, ranging from 90% to 99%. This indicates that even when QualiNet's top prediction is not the exact ground truth relation, it often includes the true relation within its top two guesses. This observation answers **RQ1** by demonstrating that QualiNet exhibits high accuracy in predicting spatial relations, particularly when considering the Top-2 accuracy, which is important for building satisfiable qualitative graphs. The model has some limitation that will be addressed in future works. The limitations include: **Small/Distant Objects:** Performance degrades for objects <50px in size (15% accuracy drop) due to limited visual information. **Detection Sensitivity:** Sensitive to object detection errors (10% error propagation to relation classification).

4 Conclusion

We presented QualiNet, a novel framework for acquiring BEV qualitative spatial relations directly from 2D images, eliminating the need for expensive depth sensors. Experimental results demonstrated high accuracy (>90% Top-2) across multiple relation types and camera views, with 92% of predicted graphs being fully consistent. Future work will extend QualiNet to dynamic scenes and enhanced occlusion handling.

References

1 C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021. doi:10.1016/J.ESWA.2020.113816.

2 N. Belmecheri, A. Gotlieb, N. Lazaar, and H. Spieker. Toward trustworthy automated driving through qualitative scene understanding and explanations. *SAE Int. J. CAV*, 8(1), 2024. doi:10.4271/12-08-01-0003.

3 Nassim Belmecheri. QualiNet. Software, swhId: swh:1:dir:a4900663aeb84632699b0217f7a7f98014466c00 (visited on 2025-09-18). URL: https://github.com/nassimbel/QualiNet.git, doi:10.4230/artifacts.24755.

4 Nassim Belmecheri, Arnaud Gotlieb, Nadjib Lazaar, and Helge Spieker. Acquiring qualitative explainable graphs for automated driving scene interpretation. *arXiv*, August 2023. arXiv:2308.12755.

5 Rajkumar Bondugula, Pascal Matsakis, and James M Keller. Force histograms and neural networks for human-based spatial relationship generalization. In *Proceedings of the International Conference on Neural Networks and Computational Intelligence*, pages 185–190, 2004.

6 Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv*, 2019. arXiv:1903.11027.

7 Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

8 AG Cohn, C Burbridge, DC Hogg, M Alomari, N Hawes, P Duckworth, P Lightbody, Y Gatsoulis, Christian Dondrup, and Marc Hanheide. Qsrlib: a software library for online acquisition of qualitative spatial relations from video. In 29^{th} *International Workshop on Qualitative Reasoning (QR'16). New York City*, 2016.

9 Guillem Collell, Luc Van Gool, and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32(1), 2018.

10 Ivan Donadello, Luciano Serafini, and Artur S d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1596–1602, 2017. doi:10.24963/IJCAI.2017/221.

11 Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. doi:10.1109/CVPR.2012.6248074.

12 Hua Hua, Dongxu Li, Ruiqi Li, Peng Zhang, Jochen Renz, and Anthony Cohn. Towards explainable action recognition by salient qualitative spatial object relation chains. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-22)*, 2022. doi:10.1609/aaai.v36i5.20513.

13 Sang Uk Lee, Sungkweon Hong, Andreas Hofmann, and Brian Williams. Qsrnet: Estimating qualitative spatial representations from rgb-d images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8057–8064, 2020. doi:10.1109/IROS45743.2020.9341452.

14 Dongxu Li, Enrico Scala, Patrik Haslum, and Sergiy Bogomolov. Effect-abstraction based relaxation for linear numeric planning. In *IJCAI*, pages 4787–4793, 2018. doi:10.24963/IJCAI.2018/665.

15 Jochen Renz and Bernhard Nebel. Qualitative Spatial Reasoning Using Constraint Calculi. In *Handbook of Spatial Logics*, pages 161–215. Springer, 2007. doi:10.1007/978-1-4020-5587-4_4.

16 Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. Commonsense visual sensemaking for autonomous driving – on generalised neurosymbolic online abduction integrating vision and semantics. *Artificial Intelligence*, 299:103522, 2021. doi:10.1016/j.artint.2021.103522.

17 Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. Commonsense visual sensemaking for autonomous driving – on generalised neurosymbolic online abduction integrating vision and semantics. *Artificial Intelligence*, 295:103458, 2021.

18 Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021.

19 Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th acm international conference on multimedia*, pages 521–529, 2019. doi:10.1145/3343031.3351040.