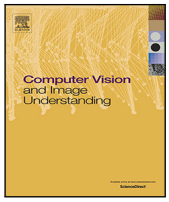
Computer Vision and Image Understanding 241 (2024) 103944

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



GMC: A general framework of multi-stage context learning and utilization for visual detection tasks

Xuan Wang [a,*], Hao Tang [a,b], Zhigang Zhu [a,c]

[a] The Graduate Center, The City University of New York, 365 5th Avenue, New York, NY 10016, USA
[b] The Borough of Manhattan Community College, The City University of New York, 199 Chambers Street, New York, NY 10007, USA
[c] The City College, The City University of New York, 160 Convent Avenue, New York, NY 10031, USA

ARTICLE INFO

Communicated by Shiliang Zhang

Keywords:
Context
Computer vision
Context integration
Object detection
Pedestrian detection

ABSTRACT

Various contextual information has been employed by many approaches for visual detection tasks. However, most of the existing approaches only focus on specific context for specific tasks. In this paper, GMC, a general framework is proposed for multistage context learning and utilization, with various deep network architectures for various visual detection tasks. The GMC framework encompasses three stages: preprocessing, training, and post-processing. In the preprocessing stage, the representation of local context is enhanced by utilizing commonly used labeling standards. During the training stage, semantic context information is fused with visual information, leveraging prior knowledge from the training dataset to capture semantic relationships. In the post-processing stage, general topological relations and semantic masks for stuff are incorporated to enable spatial context reasoning between objects. The proposed framework provides a comprehensive and adaptable solution for context learning and utilization in visual detection scenarios. The framework offers flexibility with user-defined configurations and provide adaptability to diverse network architectures and visual detection tasks, offering an automated and streamlined solution that minimizes user effort and inference time in context learning and reasoning. Experimental results on the visual detection tasks, for storefront object detection, pedestrian detection and COCO object detection, demonstrate that our framework outperforms previous state-of-the-art detectors and transformer architectures. The experiments also demonstrate that three contextual learning components can not only be applied individually and in combination, but can also be applied to various network architectures, and its flexibility and effectiveness in various detection scenarios.

1. Introduction

Contextual information plays a significant role in various computer vision tasks, encompassing both visual and non-visual data related to the appearance of a target, be it an object or an event. When objects are encountered without proper context, such as in object recognition, the task can become challenging. However, leveraging contextual cues can offer vital insights for accurate target recognition. In tasks involving videos, like action or event recognition, temporal context becomes crucial in predicting future occurrences. For instance, if a person walking is partially obscured by a car or a telegraph pole in the current frame, information from adjacent frames (previous or next) can aid in locating and detecting the occluded person.

In object detection tasks, the presence of other objects within the scene can influence the identification of a target object. These contextual cues can reveal co-occurrences and object locations. For instance, a painting should typically be found on a wall rather than on the ground. Knowing that there is a desktop on a table increases the likelihood of finding a keyboard and a mouse nearby. Furthermore, additional contextual information such as locations, dates, and environments can further enhance the likelihood of detecting objects or events.

A comprehensive survey on context understanding in computer vision can be found in our recent survey paper (Wang and Zhu, 2023). In this paper,we propose a *G*eneral framework of *M*ulti-stage *C*ontext learning utilization (the *GMC framework*) for visual detection tasks. The GMC framework incorporates different forms of contextual information, works for different visual detection tasks, and can use different network architectures (Fig. 1). The forms of context information include local context in the data labeling stage, semantic context in the model training stage, and spatial context among objects to be detected in the post-processing stage. This framework aims to offer the generality of using context in various tasks and with various architectures, in order to improve performance in various visual detection tasks.

In the domain of visual object detection, bounding boxes are widely used to represent the spatial location of objects. Crowdsourcing platforms like Amazon's Mechanical Turk (AMT) are commonly employed to annotate large datasets such as MSCOCO (Lin et al., 2014)

* Corresponding author.
E-mail address: xwang4@gradcenter.cuny.edu (X. Wang).

https://doi.org/10.1016/j.cviu.2024.103944
Received 11 September 2023; Received in revised form 14 December 2023; Accepted 21 January 2024
Available online 5 February 2024

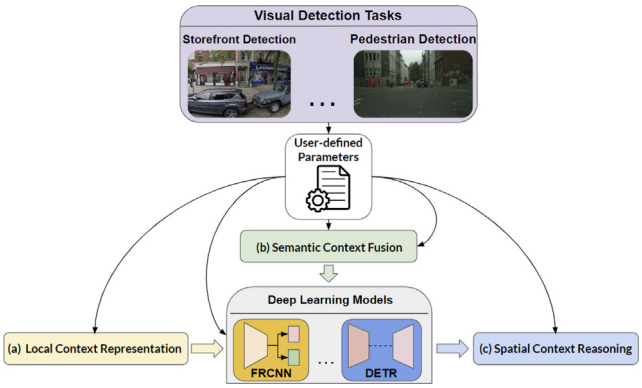


Fig. 1. The overview of GMC, our general framework of multi-stage context learning and utilization for visual detection tasks. We design a user configuration mechanism for automating the process for various detection tasks and with different network models. Each context component is guided by user-defined parameters with minimum modification of the system when applying to different deep learning models and visual tasks.

and ImageNet (Deng et al., 2009), heavily relying on human labelers. Typically, human labelers manually draw tight bounding boxes around objects to maintain label consistency. However, when dealing with small objects, using tight bounding boxes may not provide sufficient local contextual information for accurate recognition. In some cases, even human observers struggle to recognize small objects due to their small sizes. Moreover, viewing the entire scene allows for even easier recognition by incorporating a more global context, despite the size of a small object in the image.

Previous studies such as Lim et al. (2021), Leng et al. (2021) have demonstrated the importance of contextual information from the surrounding areas of small objects in achieving successful detection results. However, these studies typically utilize deep learning models to extract and refine features from these small objects, which can increase computational costs. In fact, one straightforward approach to leverage local context for small objects is to directly include their surrounding areas in the images during the labeling process, thereby providing explicit contextual information.

In this work, we propose an automatic *local context* representation that enhances the original bounding boxes for specific objects. This allows us to incorporate local context prior to the model training step, by simply using the two most commonly used definitions of small objects in computer vision tasks. By adopting this approach, we aim to exploit the benefits of local context while mitigating the potential increase in computational complexity associated with deep learning-based feature extraction methods.

Semantic context plays a crucial role in successful object detection by providing valuable information. Even without visual cues, knowing that a scene is set in an urban street environment allows us to make educated guesses about the presence of pedestrians, bicycles, vehicles, and other relevant objects. The labels assigned to objects within a scene in a training dataset can also provide prior knowledge regarding the co-occurrence relationships between different labels. Previous studies such as Li et al. (2014, 2016b), Lee et al. (2018) have demonstrated the effectiveness of using graphs to model label correlations. For instance, Chen et al. (2019) proposed a framework that leverages graph-based label dependencies for multi-label image recognition. Wang et al. (2022) model the highly correlated storefront objects using the co-occurrence of the related objects and leverage the context information for better detection performance.

Inspired by these approaches, we extend the idea we proposed in Wang et al. (2022) for storefront accessibility detection, and introduce a mechanism that allows easy user configuration to automate the generation of a contextual graph and the retrieval of word embeddings from pre-trained language models. This mechanism enables the adaptation of context learning models to various visual detection tasks. Within our framework, a Graph Convolutional Network (GCN) (Kipf and Welling, 2016) is utilized to learn from the contextual graph. By incorporating word embeddings, the GCN builds a semantic space and projects visual features extracted by the object detector into this space for the final classification stage. This integration of semantic context enhances the accuracy and performance of the object detection system.

Real-world scenes often exhibit spatial relationships between objects (i.e., *spatial context*), where certain objects tend to appear together or have specific spatial arrangements. For instance, a keyboard and a mouse are commonly found together, with the mouse typically positioned to the right of the keyboard. Yang et al. (2015) proposed a Faceness-Net that leverages spatial relationships between facial parts, such as the hair appearing above the eyes and the nose appearing below the eyes. Similarly, another work (Yang et al., 2019) introduced a spatial-aware network that models relative locations among different objects to improve object detection performance. Recent papers (Wang et al., 2022; Chacra and Zelek, 2022) have also utilized specific spatial relationships for tasks like storefront accessibility detection and scene graph generation. However, these methods often employ hard-coded spatial relationships tailored to their specific tasks, making it challenging to generalize them to other tasks without significant modifications.

To address this limitation and provide a more general approach to model spatial relationships, topological relationships can be beneficial for capturing object relations, as shown in Fig. 1. In this work, we extended the idea and propose a more generalized approach to model spatial relationships between objects for visual detection tasks. By utilizing a user configuration mechanism, we maximize flexibility in defining object relations without the need for code modifications.

While contextual information has been employed in specific computer vision tasks, such as data augmentation (Dvornik et al., 2018), semantic reasoning during training (Zhu et al., 2021; Chen et al., 2019; Wang et al., 2022, 2023), and post-processing (Fang et al., 2017; Wang et al., 2022, 2023), there is a lack of research on a comprehensive general framework that guides context learning across data labeling, model training, and post-processing stages in a generalized manner. In our previous work (Wang et al., 2022), we proposed a context learning framework for storefront accessibility detection that covered these stages. However, the framework was specifically designed with context learning mechanisms tailored to storefront accessibility detection. Therefore, significant code modifications were necessary to adapt it to different tasks. In a follow-up work (Wang et al., 2023), we proposed a framework for different visual detection tasks in urban scenes. However the framework only works with one single network architecture, and the experiment on the second example (the pedestrian detection) is very limited; There are no available contexts for spatial context reasoning and the result only achieves minor improvement over the baseline network.

In this work, we present a general context learning and reasoning framework with various deep learning models, applicable to various visual detection tasks, and therefore offering greater flexibility and adaptability without requiring extensive code changes. As an extended version of our previous work (Wang et al., 2022, 2023), this paper demonstrates the versatility and adaptability of our context components by successfully applying them to different deep learning models with minimal modification. The pedestrian detection task is greatly enhanced with more categories of contextual objects and includes all the three stages of context reasoning. We also tested the framework on a large detection benchmark—MSCOCO dataset, showing promising results.

The proposed context learning and reasoning framework for visual detection tasks offers several noteworthy aspects. Firstly, it introduces a comprehensive approach consisting of three key components: Local Contextual Representation (LCR), Semantic Context Fusion (SCF), and general Spatial Context Reasoning (SCR). The LCR component improves

recognition accuracy for specific objects especially small objects by incorporating their local context, while the SCF component models semantic relations using a contextual graph, capturing co-occurrence and contextual dependencies. Additionally, the SCR component leverages topological relationships and semantic masks to incorporate general spatial relations between objects. The framework's flexibility allows for easy adaptation to different tasks, without requiring extensive code modifications. Overall, this framework presents a valuable contribution to the field of visual detection by providing a comprehensive and adaptable solution that enhances context learning and reasoning capabilities.

As some highlights, the local context representation and semantic context fusion components are seamlessly integrated into diverse models, ensuring an automated adaptation process in using ground truth labels and prior knowledge. This integration is also designed to empower users with the flexibility to tailor the components according to their specific requirements through the utilization of user-defined parameters. Moreover, we have introduced a novel general spatial context reasoning component that combines topological relations between objects and semantic masks. This combination allows our framework to easily adapt to various visual detection tasks, providing a powerful tool for improving detection performance in diverse scenarios with more accurate results. We provide user flexibility to configure the spatial relations because the user configuration can offer meaningful definitions of important spatial relations as the first step, and then our Spatial Context Reasoning (SCR) component will autonomously generate relation parameters, such as overlapping thresholds based on the provided information in the user configuration, by modeling subject–object ground truth labels. Overall, our approach not only enhances the effectiveness of context learning and reasoning in visual detection but also simplifies the integration process, making it readily applicable to a wide range of deep learning models and tasks.

In summary, the main contributions of this paper are:

- We introduce a general framework for multistage context learning and utilization, with three context components to leverage local context, semantic context and spatial context. This combination of components provides a holistic solution to address context learning and reasoning in visual detection tasks.
- Our framework proposed in this work is designed to be applicable to any deep learning models. This versatility makes the framework highly versatile and empowers users to leverage its benefits across different object detection tasks, regardless of the specific deep learning model employed.
- Our framework is not limited to a specific visual detection task but can be applied to various visual detection tasks, including storefront object detection and pedestrian detection as our examples. Its flexibility and adaptability enable users to utilize the framework across a wide range of visual detection tasks, benefiting from its context learning and reasoning capabilities.
- Our framework has the ability to incorporate different types of context information at various stages of the detection process. It provides a unified framework that can effectively integrate and utilize these contextual cues at the appropriate stages, such as during data preprocessing, model training, or post-processing. This capability enhances the overall performance and robustness of the detection system by harnessing diverse sources of contextual information to improve object understanding and localization.

The paper is organized as follows. Section 2 discusses related work. Section 3 proposes our general context learning and reasoning framework and describes each component in detail. Section 4 discuss how the general framework work with various deep learning network architectures with minimal modification of the code. Section 5 discuss the use of the general framework for three different tasks, including a description of the three datasets (Section 5.1), and the experimental results (Section 5.3). Finally, Section 6 provides a few concluding remarks.

2. Background and related work

In this section, we will start with a general survey of the literature in context learning and utilization for computer vision tasks, then move on the use of context information in object detection, and finally focus on pedestrian detection—a particularly important task that poses challenges and opportunities in using context information.

2.1. Context learning and utilization

Humans use visual context effortlessly to perceive the real world. An object hanging on the wall is probably a painting, not a car. A doorknob should be within the frame of a door, not on the ground. Contextual information provides critical information to help us visually find and recognize objects faster and more accurately. Not only in human perception, contextual information also plays an important role in many computer vision tasks, such as object detection (Du et al., 2012; Fang et al., 2017; Sun and Jacobs, 2017; Zhu et al., 2016, 2021), video event recognition (Wang and Ji, 2015, 2016), video action detection (Yang et al., 2019; Zhu et al., 2013), scene graph generation (Xu et al., 2017; Zellers et al., 2018), data augmentation (Dvornik et al., 2018), image classification (Mac Aodha et al., 2019), and image inpainting (Pathak et al., 2016). In these tasks, different forms of contextual information have been employed. The contextual information used in the literature includes: global context (Zellers et al., 2018), local neighborhood context (Pathak et al., 2016; Dvornik et al., 2018; Du et al., 2012), prior semantic knowledge (Wang and Ji, 2015, 2016), geographic information (Mac Aodha et al., 2019), spatial relation between objects (Sun and Jacobs, 2017; Xu et al., 2017; Zellers et al., 2018; Yang et al., 2019) and temporal information (Wang and Ji, 2015, 2016; Yang et al., 2019; Zhu et al., 2013).

Context information has been widely used in many computer vision tasks. Dvornik et al. (2018) show that the visual context surrounding objects is crucial to predict the presence of objects. A serial work (Wang and Ji, 2015, 2016) introduces a hierarchical context model to recognize events in videos. Wang et al. (2022) make use of various contextual information by applying a unified multi-stage framework in context learning and utilization from data labeling, model training, to object detection and result evaluation.

Context has been integrated in different ways in visual detection tasks. Many visual detection tasks (Yang et al., 2015; Chen et al., 2019; Pathak et al., 2016; Leng et al., 2021; Li et al., 2016a; Mac Aodha et al., 2019; Yang et al., 2018) implement context information into the backbone models and aggregate with the features extracted from context-free methods. Deep learning methods mainly have four stages: data pre-processing (including labeling), model training, post-processing, and result evaluation. Context information has either been aggregated during the training stage or used in the post-processing stage. No general pipelines have been proposed on how we can incorporate context through the whole process stages. Although different context integration can be used in a single stage or in multiple stages, a general pipeline is needed to guide the integration for context. Our proposed framework employs different forms of context information through the entire deep learning process, and each component is easy to add and remove from an object detector.

2.2. Object detection

Contextual information plays a crucial role in understanding natural scenes and images for object detection, as it provides rich

information about the relationships between objects and the overall scene. However, the evaluation of context models has primarily focused on improving object detection performance for particular tasks, overlooking more general applications of contextual information. In the domain of urban scene object detection, various methods have been proposed, addressing specific tasks such as text detection and recognition (Du et al., 2012; Zhu et al., 2016), zebra crossing detection (Ahmetovic et al., 2015), curb detection (Cheng et al., 2018; Sun and Jacobs, 2017), and storefront accessibility detection (Wang et al., 2022).

For example, Du et al. (2012) and Zhu et al. (2016) focused on text detection in street environments. Cheng et al. (2018) proposed a framework for road and sidewalk detection using stereo vision in urban regions. Sun and Jacobs (2017) aimed to identify missing curb ramps at street intersections by leveraging the pairwise existence of curb ramps. Our recent work (Wang et al., 2022) introduced a multi-stage context learning framework specifically designed for storefront accessibility detection, utilizing category-specific relations. These examples demonstrate that context modeling has been applied to various urban scene object detection tasks beyond traditional object recognition. It highlights the potential of exploiting different types of contextual information to improve the performance of detection systems in diverse real-world scenarios. In this paper we propose a general context learning and reasoning framework which could be adapted to various visual detection tasks.

Contextual information, particularly prior knowledge, has played a crucial role in advancing object detection tasks. Fang et al. (2017) introduced a knowledge-aware object detection framework that incorporates external knowledge, such as knowledge graphs, into object detection algorithms. By leveraging a knowledge graph, which represents real-world concepts and their interactions, this framework enables the modeling of semantic consistency. Even concept pairs that are not directly connected in the graph can benefit from this approach, leading to enhanced generalization capabilities.

Similarly, Zhu et al. (2021) explored the integration of semantic context and visual information for the task of few-shot object detection. Their work focused on explicit relation reasoning and utilized word embeddings to represent class labels. By establishing semantic relation consistency between base and novel classes, the aim was to bridge the domain gap between visual and language information. Incorporating semantic consistency principles, their framework improved object detection by optimizing for better alignment with prior knowledge.

Building upon these concepts, our general framework embraces the notion of semantic consistency to quantify and generalize knowledge, resulting in improved object detection performance through a re-optimization process. In addition, our framework adopts a context-aware approach to object detection, considering both visual context and prior knowledge context. By incorporating both types of context, our framework provides a more comprehensive and enriched understanding of the scene, leading to more accurate and robust object detection results.

Indeed, context can be leveraged not only for detecting objects but also for predicting their presence or absence in an image. Sun and Jacobs (2017) conducted a unique vision task focused on identifying the absence of objects in an image, specifically curb ramps. This work extensively utilized local and spatial context information to determine the locations where curb ramps should exist.

Similarly, in our proposed framework, we emphasize the importance of local context representation surrounding small objects. This local context provides valuable information that can indicate both the location and category of the object. By incorporating this local context into our general framework, we aim to enhance the detection and prediction capabilities, enabling more accurate understanding of the scene and object presence even in the absence of explicit object instances.

2.3. Pedestrian detection

Pedestrian detection in urban scenes presents unique challenges due to factors such as heavy occlusion and small-scale pedestrian images. Several papers have focused on addressing these challenges and improving the performance of pedestrian detection algorithms. For example, Cai et al. (2016) proposed a unified framework for pedestrian detection that incorporates contextual information to handle occlusion. Zhang et al. (2017) introduced the CityPersons dataset specifically for pedestrian detection in urban environments and proposed a scale-aware network to tackle the problem of detecting small-scale pedestrians.

Other works have explored different approaches to handle occlusion in pedestrian detection. Zhou and Yuan (2018) proposed an attention-based method that focuses on visible parts of partially occluded pedestrians, improving the detection accuracy in challenging scenarios. Wu et al. (2020) introduced a part-based detection framework that leverages feature transformation to handle occlusion and improve detection performance.

Despite the progress made by CNN-based pedestrian detectors, there are still limitations in detecting small-scale and heavily occluded pedestrians. These challenges require further exploration and innovation in the design of detection algorithms. For example, the integration of additional context information beyond a single image, such as global scene context and temporal context, could potentially improve the performance of pedestrian detection systems in real-world scenarios. This is beyond the scope of this paper; more details can be found in our recent survey paper (Wang and Zhu, 2023).

Pedestrian detection in urban scenes is a challenging task that has garnered significant attention in the computer vision community. Several papers have focused on addressing the unique challenges associated with detecting pedestrians in such environments. While approaches like Faster R-CNN have become popular for pedestrian detection, they often fall short in effectively handling heavily occluded pedestrians and small-scale pedestrians. Limited progress has been made in leveraging local context information specifically for these scenarios, resulting in sub-optimal detection performance.

To address this gap, our proposed novel framework integrates local context for small-scale and occluded pedestrian detection in urban scenes. Our approach incorporates general topological relations among objects to facilitate spatial reasoning. By considering the relationships (including occlusions) between objects, we can reason about the presence and location of pedestrians, even in challenging situations. Notably, our framework goes beyond improving pedestrian detection alone; it also enhances the detection results for other objects in the scene. By leveraging the synergistic effects of contextual components, our approach aims to achieve superior performance compared to existing methods.

By emphasizing the importance of local context and introducing general topological reasoning, our framework offers a comprehensive solution for pedestrian detection in urban scenes. Note that the general framework is not specially designed for pedestrian detection but the system can be configured to tackle these two challenges in pedestrian detection. Through the incorporation of contextual cues and the utilization of interplay between different components, we can overcome the limitations of traditional approaches and improve detection accuracy. Ultimately, our work contributes to advancing the understanding of urban scenes and objects, opening up new possibilities for real-world applications.

3. General framework and context components

Our proposed GMC framework, as detailed in Fig. 2, consists of three key context components: local context representation, semantic context fusion, and spatial context reasoning. These components can be applied

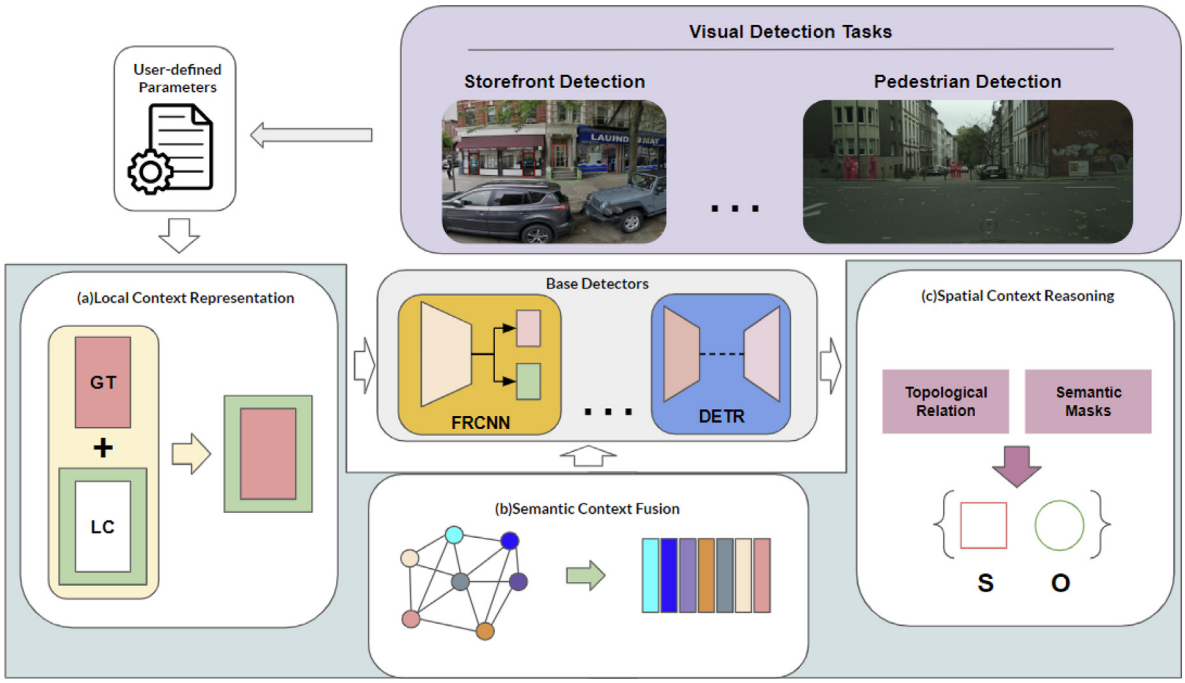


Fig. 2. Details of our GMC framework, the general framework of multi-stage context learning and utilization for visual detection tasks. We design a user configuration mechanism for automating the process for various detection tasks (e.g., storefront object detection, pedestrian detection), using different base detectors (e.g. a CNN model Faster R-CNN (FRCNN) and a transformer model DETR. Three context learning and utilization components—(a) Local Context Representation, (b) Semantic Context Fusion, and (c) Spatial Context Reasoning, guide the deep learning models during data labeling, model training and post-processing stages. Each component can be applied individually and in combination. *GT*: Ground Truth. *LC*: Local Context. *S*: Subject. *O*: Object.

individually or in combination with a given visual detection network architecture to enhance object detection performance.

The local context representation component (Section 3.1) focuses on capturing local contextual information specific to the objects of interest. By incorporating local context features in the data labeling stage, this component improves the accurate detection of objects, particularly small-scale or occluded ones, by leveraging relevant contextual cues. The semantic context fusion component (Section 3.2) integrates semantic information with visual context to capture object relationships. By combining prior knowledge and/or learning from the training dataset in the model training stage, this component enhances the detection network's understanding of the scene and improves its ability to discriminate and classify objects. The spatial context reasoning component (Section 3.3) introduces a general topological relation between object categories to optimize detection results. By considering the spatial relationships between objects in the post-processing stage, such as "above", "under", or "within", this component refines detection outputs based on their spatial arrangements. This spatial context reasoning enhances the detection network's localization accuracy and object classification performance by incorporating topological reasoning into the detection process.

An automated process is implemented for each component with simple user defined parameters. In local context representation component, we apply an automatic local contextual labeling approach to enhance the original bounding boxes for small objects in order to employ local context *before* the model training step, by using the two most used definitions of small object in computer vision tasks. In semantic context fusion component, we automate the process for generating a contextual graph by leveraging label occurrence knowledge from training data, and automatically searching the word embeddings from a pretrained language model. In spatial context reasoning component, we adopt user configuration for important spatial relations of objects as guidelines, to automatically generate the spatial relation thresholds, which maximize the flexibility for object relation definition, without code modifications.

In the following sections, we will provide detailed explanations of each component within our proposed general framework. Through some user-defined parameters related to a given visual detection task and the chosen base detector, the GMC framework can be easily configured to form an end-to-end model for the task.

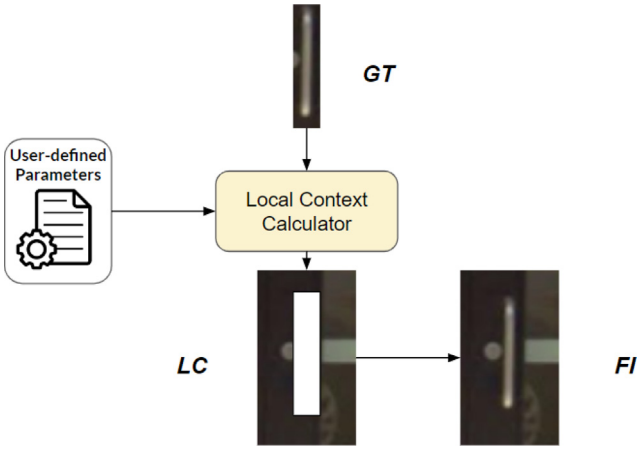


Fig. 3. An utilized local context representation. The local context calculator is guided by user-defined parameters and enhance the local context around the ground truth label of the object. *GT*: Ground Truth. *LC*: Local Context. *FI*: Final Input.

3.1. Local context representation

The concept of *local context* for objects, particularly small ones, takes center stage in the Local Contextual Representation (LCR) component. In the realm of computer vision, categorizing an object as "small" is not always clear-cut. Factors like shooting angles and environmental conditions can render an object that is deemed "small", such as a spoon, appearing quite "large" within an image. Hence, the notion of smallness hinges on an object's size relative to the context of the image, as explained further below. The procedural essence is graphically illustrated in Fig. 3. A local context calculator is at the heart of this process, guided

by user-defined parameters specific to LCR. This calculator works to enrich the local context surrounding the ground truth label of the targeted object. To initialize this local context calculator, we introduce two commonly embraced standards for characterizing small objects. The Local Context Representation (LCR) component operates during the data preprocessing stage, focusing solely on the labeling standard and the specified enlargement percentage for small objects (Table 1). This component automatically processes the labels before they are fed into the network, ensuring seamless integration without introducing additional inference complexity.

Within the COCO dataset (Lin et al., 2014), small objects are defined as those whose dimensions are 32 × 32 pixels or smaller, within the confines of an image with a fixed size of 640 × 480 pixels. Another definition, as detailed in Chen et al. (2017), relates to situations where the overlap area between the ground truth bounding box and the image remains below 0.58%. Given the robustness and widespread adoption of these definitions in the research community, we employ them as reference points for automating the labeling process for small objects. We include the surrounding local context of the bounding box B of an object O in image I if the object satisfies with the COCO standard for a small object as:

$$B'_O = \begin{cases} (1+\alpha)B_O, & \text{if } B_O < 32 \times 32 \\ B_O, & \text{otherwise} \end{cases} \tag{1}$$

If the small object satisfies with the second standard—the Small Object Dataset (SOD) Standard (Chen et al., 2017), we include the local context of the bounding box B of the object O in image I by:

$$B'_O = \begin{cases} (1+\beta)B_O, & \text{if } \frac{B_O}{R_I} < 0.58\% \\ B_O, & \text{otherwise} \end{cases} \tag{2}$$

The above equations introduce notations representing the original and updated bounding boxes of the ground truth label for a small object. These notations, B_O and B'_O respectively, are utilized in the context of the user-defined parameters for the Local Context Representation (LCR) component. Firstly, the parameters α and β hold significance as extending factors, expressed in terms of a percentage, from the original bounding boxes. These factors are related to two distinct standards: the COCO standard and the SOD standard. The resolution of the input image, denoted as R_I, is automatically determined. This automatically calculated resolution serves as a crucial component in the calculation of these factors. Secondly, the framework affords users the liberty to choose between the two contextual labeling standards. Should a given small object meet the criteria of both definitions, the user can opt for the standard that best aligns with their requirements. Importantly, both the original bounding boxes and the enlarged bounding boxes are retained for all small objects that conform to the user-selected standard for both training and testing sets. This dual retention strategy serves the dual purpose of integrating local contextual information and enhancing the detection's robustness. The forthcoming sections will delve into the specifics of the experimental settings in Section 5.2, elaborating further on these parameters and their implications.

3.2. Semantic context fusion

Semantic information indeed plays a crucial role in visual detection tasks, providing valuable insights to enhance the detection process. To ensure a seamless and automatic Semantic Context Fusion (SCF) into our framework, we have introduced the SCF user-defined parameters, namely, the categories of a given visual detection task and the text embeddings used in the task. For example, for a storefront object detection task, they are door, doorknob, stair. For pedestrian detection, they include pedestrian, vehicle, bicycle (bike), motorcycle, etc. These parameters act as guiding factors for the model to learn and incorporate semantic context using text embeddings. The text embeddings, obtained from pre-trained language models, are utilized to generate semantic spaces that can be effectively fused with the visual information obtained from the detection process. This integration of semantic context with text embeddings allows our framework to automatically leverage valuable semantic information to improve the overall detection performance, while minimizing the need for extensive component modification.

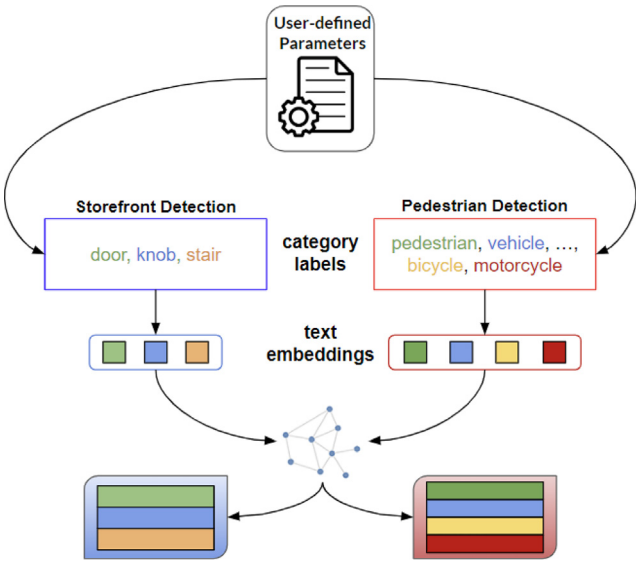


Fig. 4. The visualization of Semantic Context Fusion. We use category information as the semantic context cues to generate semantic spaces for visual detection tasks.

In our framework, the fusion of semantic context is depicted in Fig. 4. When the framework receives category information from the SCF user configuration, it proceeds to search for word embeddings $H_{labels} \in \mathbb{R}^{n \times d}$ from a pretrained language model (such as GloVe Pennington et al., 2014). Here, n represents the number of label categories, and d denotes the dimensionality of the word embeddings. Subsequently, an automatic generation of the contextual graph takes place. The Graph Convolutional Network (GCN) is then employed to learn semantic relations within the contextual graph, effectively constructing a semantic space. This semantic space is obtained by transforming the label feature representation, resulting in $H'_{labels} \in \mathbb{R}^{n \times D}$, where D represents the dimensionality of the region features extracted from the object detector. As illustrated in Fig. 2, the region features $f_{regions} \in \mathbb{R}^{D \times N}$ are projected into the semantic spaces H'_{labels}. Ultimately, the final output is derived from this process:

$$\mathbf{P}_{regions} = softmax(H'_{labels} f_{regions}) \tag{3}$$

where $\mathbf{P}_{regions}$ represents the classification probability distribution for each proposed region, and $\mathbf{P}_{regions} \in \mathbb{R}^{n \times N}$.

As the category information is provided by a given task, our system automatically generates a contextual graph between different categories, leveraging prior label occurrence knowledge extracted automatically from the training data. Additionally, we autonomously search for pretrained word embeddings from the dictionary (Pennington et al., 2014) without requiring extra information. The SCF (Semantic Context Fusion) component, armed with the prebuilt contextual graph and pretrained word embeddings, ensures minimal additional complexity. The user-defined parameters for the SCF module are detailed in Table 1.

3.3. Spatial context reasoning

In the proposed general Spatial Context Reasoning (SCR) component, we leverage topological relationships to model the spatial relations between different objects. Topological relationships provide a general and abstract representation of the relationships between objects, such as *overlap*, *within*, *touch*, and so on. These relationships

Table 1
Summary of the provided user-defined parameters for the contextual components.

Parameters	Context component	Definition
[***Subject***, ***Object***]	LCR\SCR	Subject and object pair
Labeling_standard	LCR	The standard for small object label enlargement
Enlarge_percentage	LCR	The enlarging percentage for small object labels
Categories	SCF	The object categories
Relation_descriptor	SCF	The contextual graph generation method
pred(optional)	SCR	Directional relationships between subject and object
t	SCR	Topological relationships between subject and object
Overlap_threshold(optional)	SCR	The threshold of overlap percentage between subject and object
$Search_{height}$(optional)	SCR	The height of search area for object
$Search_{width}$(optional)	SCR	The width of search area for object

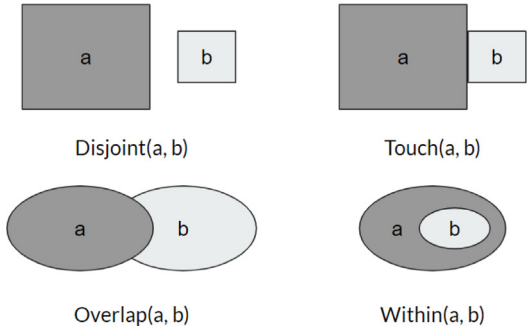


Fig. 5. The visualization of common used topological relationships from Clementini et al. (1993) and Egenhofer and Franzosa (1991).

Fig. 6. Bounding box vs. semantic masks for road and sidewalk.

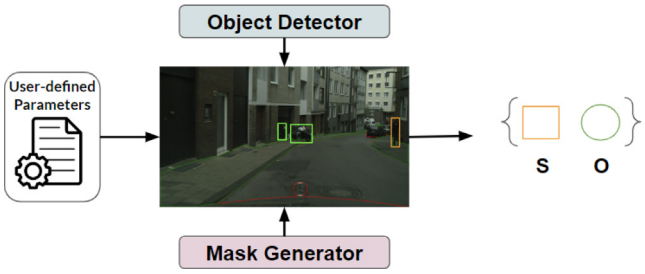


Fig. 7. The visualization of general Spatial Context Reasoning.

capture the overall spatial configuration and arrangement of objects in a scene, including next two each other, within, and occlusion. The visualization of topological relationships is depicted in Fig. 5, illustrating how different objects can be related in terms of their spatial positions and co-occurrence. By incorporating topological reasoning, our framework enables a more comprehensive understanding of the spatial context, enhancing the object detection performance and facilitating richer semantic interpretations of the scene.

We utilize a predicate *pred*, such as *above*, *under*, etc., to describe the directional relation between a subject and object pair $[S, O]$, along with the topological relationship t, such as *overlap* and *within*. This general relation R is defined as shown in Eq. (4):

$$R[S, O] = pred[t(S, O)] \tag{4}$$

For instance, in urban settings, a common spatial relationship is that a stair is usually located under a door, even if there might be overlaps or spatial misalignment between them. The general relationship between a pedestrian and sidewalk can be described as $R[pedestrian, sidewalk] = under[overlap(pedestrian, sidewalk)]$. It is important to note that the general spatial relation is inversible, meaning that a pedestrian is on the sidewalk, and sidewalk can be considered under a pedestrian. To effectively apply this spatial reasoning, we define a search area around the detected subject, and if an object is detected within this search area and satisfies the condition defined by Eq. (4). We propose it as a detection and send it for evaluation. In cases where multiple objects are detected within the search area, we propose the object with the highest score as the final prediction.

To enhance the applicability of our general framework to diverse visual detection tasks with more accurate detection, we have introduced *semantic masks* in our general spatial context reasoning component (see Fig. 7). As illustrated in Fig. 6, bounding boxes for entities like roads and sidewalks may not be suitable for effective spatial reasoning between objects. In contrast, semantic masks offer a more precise and appropriate means for modeling the relationships between subjects and objects. While segmentation poses its challenges, modern state-of-the-art segmentation models can yield accurate masks for larger entities such as roads and sidewalks, rendering them readily usable for spatial reasoning. This addition allows us to segment large stuff such as sidewalks and roads using a pretrained model, which could significantly improves spatial reasoning in larger scenes. To measure the overlap between subject–object pairs, we use the intersection over subject (IoS) metric to describe the general spatial relation, as defined as:

$$IoS = \frac{(A_s \cap A_o)}{(A_s)} \tag{5}$$

where A_s and A_o denote the area of the subject and area of the object. The area can be bounding box or semantic mask based on the specific scenarios. This formulation enables us to capture the relative spatial arrangement of objects in a scene, which is valuable for improving the accuracy of object detection and localization across various visual detection tasks. We also provide users with the flexibility to configure the general spatial relation for the categories in their own dataset, allowing them to adapt the framework according to their specific task requirements. Moreover, the user configuration can offer meaningful definitions of important spatial relations as guidelines, and then our Spatial Context Reasoning (SCR) component will autonomously generate relation parameters such as overlap thresholds based on the information obtained from the ground truth labels. Through this adaptation, users can furnish general spatial relations for specific subject–object pairs. For instance, according to common sense, a car should be on the road, or a keyboard typically appears under the monitor. Using the provided relations, we automatically analyze the training dataset and establish overlap thresholds accordingly. This approach enables the model to leverage contextual information based

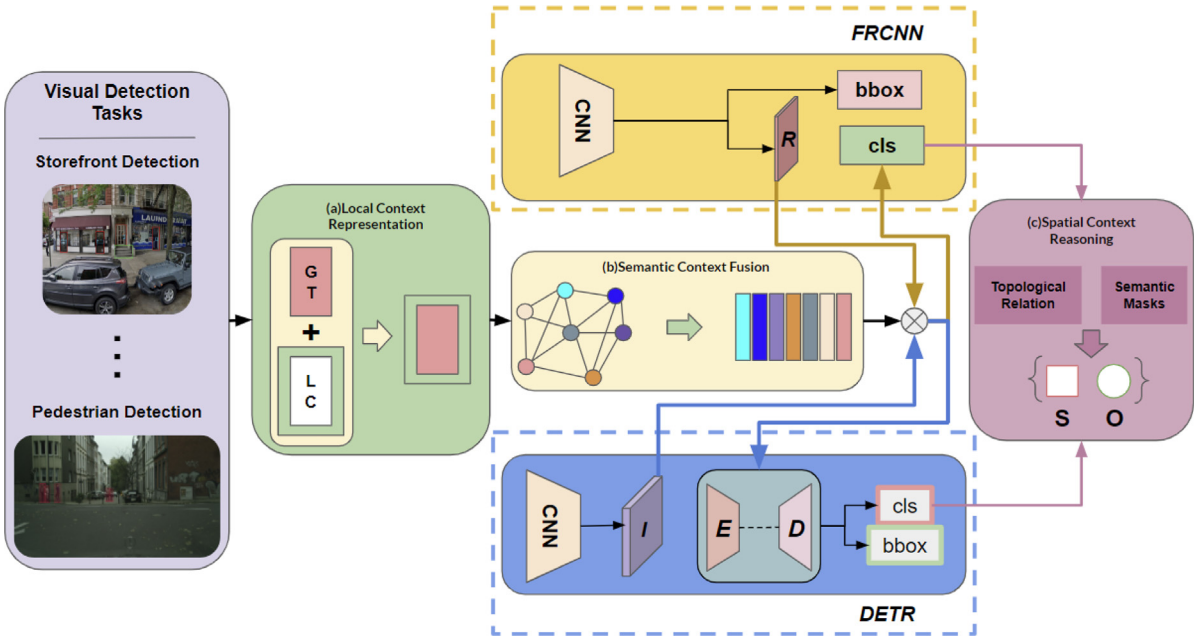


Fig. 8. Integration of contextual components with different deep learning network architectures: Faster R-CNN (FRCNN) and DETR. *GT*: Ground Truth; *LC*: Local Context; *S*: Subject; *O*: Object; *R*: Region features; *I*: Image features; *E*: Encoder; *D*: Decoder; *bbox*: bounding boxes; *cls*: classification.

on predefined spatial relationships, enhancing its understanding of the scene. The user-defined parameters for LCR, SCF and SCR components are summarized in Table 1.

4. Working with various network architectures

The GMC framework can work with various deep learning network architectures with minimal modification of the code. In this paper, we give two examples, both which will be used in the tasks of our experiments. We employ two popular object detection frameworks, Faster R-CNN (Ren et al., 2015) and DETR (Carion et al., 2020), as the underlying detectors for both storefront accessibility detection and pedestrian detection tasks. These frameworks have demonstrated strong performance in various object detection scenarios. The integration pipeline of the three context components with Faster R-CNN and DETR is shown in Fig. 8. We will detail how the three context components can be seamlessly integrated with different backbone models, with minimal code modification.

Prior to the input of the visual detection task dataset into the model, we incorporate the Local Context Representation (LCR) component to augment the local context of specific objects. While we begin with two widely adopted definitions of small objects, as detailed in Section 3.1, we also empower users to tailor the enhancement of local context according to their preferences by adjusting the enlarge percentage. This integration ensures that the LCR component can seamlessly adapt to diverse models without requiring any modifications to the underlying backbone models. This design approach not only increases the generality of our framework but also facilitates its ease of use and customization across different applications.

Within our Semantic Context Fusion (SCF) component, we harmonize semantic knowledge with visual features prior to the detection process. This integration is illustrated in Fig. 8. In the case of Faster R-CNN, we achieve this by mapping the extracted region features (R) from the feature extractor backbone into the semantic space, before subsequently feeding the resulting output into the classification (cls) head. In contrast, for a comparative scenario of DETR in Fig. 8, we first project image features (I) into the semantic space and subsequently input the resulting output into a transformer encoder–decoder (E&D) for generating predictions. This design allows users to exercise control over the nature of the pretrained word embeddings in the SCF component, with the default setting being GloVe (Pennington et al., 2014). The SCF component can be seamlessly integrated into each backbone architecture with minimal adjustments, signifying its adaptability and ease of incorporation into diverse models. This enables the enriched representation of contextual information in conjunction with visual cues, thereby enhancing the overall detection accuracy.

Moreover, the Spatial Context Reasoning (SCR) component can be seamlessly integrated to fine-tune the detected candidates by synergizing topological relationships and semantic masks among identified objects. The SCR component provides a valuable post-processing feature for both Faster R-CNN and DETR models, requiring minimal architectural adjustments. This adaptable SCR component can be easily integrated into the final stage of object classification (cls), offering a streamlined way to enhance object detection performance. Users retain the prerogative to exercise control over the component's parameters within the configuration file, ensuring adaptability and customization to distinct detection scenarios. This feature bolsters the accuracy of detection outcomes by leveraging not only the object-specific information but also the relationships and arrangements among objects within the scene.

5. Tasks and experiments

The general framework for context learning and utilization is designed not only for working with various visual detectors, but also for different tasks. In the following, we will showcase three examples: storefront accessibility detection, pedestrian detection, and COCO object detection. We will first introduce the three datasets, describe the experimental settings, and then detail the experimental results with the GMC framework.

5.1. Dataset description

Storefront Accessibility Image Dataset. For our experiments, we utilize the storefront accessibility image (SAI) dataset introduced in Wang et al. (2022). This dataset focuses on storefront accessibility in an urban environment and comprises three main categories: doors, knobs, and stairs. The SAI dataset is collected from Google Street

Fig. 9. An example of labeled objects. Red: Ground truth bounding box of Door. Cyan: Ground truth bounding box of Knob. Green: Ground truth bounding box of Stair. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Statistics of collected storefront accessibility data.

Dataset	# of Images	Doors	Knobs	Stairs
Train	992	1885	1614	420
Test	110	233	126	141

Fig. 10. The label example from CityPersons Dataset (Zhang et al., 2017). Red: Pedestrian. Blue: Rider. Yellow: Sitting person. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

View of New York City using the Google Street View API (Google, 2022). To create the dataset, we employ the methodology described in Cavallo (2015) to compose panorama images. Each panorama image captures building facades on both sides of a street in New York City. Subsequently, we divide each formed panorama image into two halves, with each half covering one side of the facade. To ensure clear and easily labelable storefronts, we crop the center of each image, where contains the necessary visual information for storefront accessibility labeling.

The SAI dataset consists of a total of 1,102 images, where each image has been labeled for three main categories of accessibility: Door, Knob, and Stair. The labeling process was carried out using the Labelbox platform (Sharma et al., 2019). To split the dataset for training and testing, a random sampling technique was employed, where 10% of the collected data was reserved for the testing set, while the remaining 90% was used for training. The data statistics are presented in Table 2, providing an overview of the dataset composition. Additionally, Fig. 9 showcases examples of labeled storefront objects within an image, providing a visual representation of the annotated data.

CityPersons and CityPersons+ Dataset. The CityPersons dataset is derived from the Cityscapes dataset (Cordts et al., 2016), focusing specifically on person annotations. It contains annotations for four categories: *pedestrian, rider, sitting person*, and *person (other)*. Table 3 provides an overview of the dataset, including information on the number of images and annotations for each category. Fig. 10 showcases

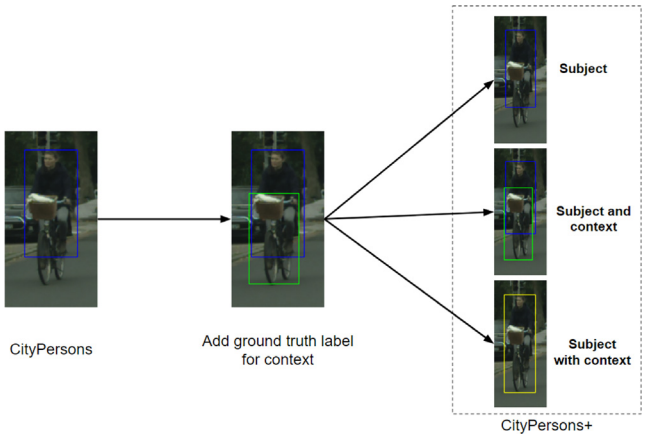


Fig. 11. The demonstration of riders in CityPersons+ dataset. We extend existing categories in CityPersons dataset, with context information, by adding the ground truth label for context things and combined with the existing subject class label.

Table 3
Statistics of CityPersons and CityPersons+ Datasets.

Dataset	# of Category	# of Training	# of Validation
CityPersons (Zhang et al., 2017)	4	2975	500
CityPersons+	6	2975	500

an example of labeled pedestrians from the dataset, providing a visual representation of the annotated data.

To incorporate various context information and leverage the general topological relations between different categories, we introduce the CityPersons+ dataset. This dataset expands upon the CityPersons dataset by incorporating additional object labels from the Cityscapes dataset, including more specific subcategories. Specifically, we categorize pedestrians and riders into four subcategories: *pedestrian on road, pedestrian on sidewalk, rider with motorcycle*, and *rider with bicycle*. Therefore CityPersons+ contains annotations for six categories. The purpose of adding subcategories is to better utilizing context information. Fig. 11 shows how we include more context information without changing existing labels. We also relate the six categories in CityPersons+ dataset to context information that are beyond these six categories. First, we add the bounding box ground truth labels for *context things*, including motorcycles, bicycles and vehicles, which are related to the existing subject class labels of rider with motorcycle, rider with bicycle, and pedestrian occluded by vehicle, respectively. Second, we include the semantic segmentation labels of *context stuff*, such as roads and sidewalks, which could provide precise spatial reasoning between different objects, namely, pedestrian on road, and pedestrian on sidewalk, in addition to pedestrian occluded by pedestrian. We also include word embeddings for both context things (motorcycles, bicycles and vehicles) and context stuff (roads and sidewalks) for Semantic Context Fusion (SCF) component. We use the pretrained model weights for Faster R-CNN and DETR to *detect* the context things, and Segformer (Xie et al., 2021) to *segment* the semantic masks for context stuff, to facilitate general topological reasoning within the Spatial Contextual Reasoning (SCR) component (see Table 4). Table 3 provides an overview of the statistics for the CityPersons+ dataset, comparing with CityPersons dataset: we double the class categories for pedestrian and riders (from 2 to 4), add 5 context objects (not shown in the Table), without changing the existing classes (2). For the 4 basic classes in CityPersons and 6 basic classes in CityPersons+, as shown in Table 3, the pretrained model weights for Faster R-CNN and DETR are finetuned using the two datasets, respectively, and the proposed GMC models will be evaluated.

MSCOCO-2017. MSCOCO is a standard benchmark in object detection and instance segmentation. It includes 80 object categories with 118k images for training and 5k for evaluation. The dataset is known

Table 4
Default user parameter settings for Spatial Context Reasoning in our experiments on the three datasets: SAI (Wang et al., 2022), CityPersons+, and COCO. O_T: Overlap_threshold.

Task	[Subject, Object]	Occlusion	Predicate	Topology	O_T	Search_area_height	Search_area_width
SAI	[door, knob]	–	–	within	–	–	–
	[door, stair]	–	under	overlap	0.2	$0.2height_{door} + height_{stair}$	$width_{door} + width_{stair}$
CityPersons+	[rider, bicycle]	Reasonable	under	overlap	0.48	$0.5height_{rider}$	$width_{bicycle}$
	[rider, motorcycle]	Reasonable	under	overlap	0.5	$0.5height_{rider}$	$width_{motocycle}$
	[pedestrian, vehicle]	Heavy	under	overlap	0.68	–	–
	[pedestrian, pedestrian]	Heavy	–	overlap	0.76	–	–
	[pedestrian, road]	Reasonable	under	overlap	0.2	–	–
	[pedestrian, sidewalk]	Reasonable	under	overlap	0.13	–	–
COCO	[person, person]	–	–	overlap	0.73	–	–
	[person, surfboard]	–	under	overlap	0.17	$0.2height_{person}$	$width_{surfboard}$
	[person, tie]	–	–	within	–	–	–
	[person, skateboard]	–	under	overlap	0.1	$0.2height_{person}$	$width_{skateboard}$
	[person, snowboard]	–	under	overlap	0.16	$0.2height_{person}$	$width_{snowboard}$
	[zebra, zebra]	–	–	overlap	0.83	–	–
	[baseball glove, person]	–	–	within	–	–	–
	[potted plant, vase]	–	under	overlap	0.45	–	–
	[frisbee, dog]	–	–	overlap	0.85	–	–

for its diversity, containing a wide range of objects and scenes. It features a maximum of 93 object instances per image, with an average of 7 objects.

5.2. Experimental settings

Faster R-CNN. In our implementation, we utilize ResNet-50 (He et al., 2016) as the backbone feature extractor along with the Feature Pyramid Network (FPN) (Lin et al., 2017), which are both pretrained on the COCO dataset. For the semantic context fusion, we employ a 2-layer graph convolutional network (GCN) with LeakyReLU (Maas et al., 2013) as the activation function. The GCN takes 300-dimensional word embeddings from GloVe (Pennington et al., 2014) as the input label feature vector. During training, we employ Stochastic Gradient Descent (SGD) as the optimizer, with a momentum of 0.95 and a weight decay of 1e-4. The initial learning rate is set to 0.005 and is reduced by a factor of 0.25 every 8 epochs. We train the model for a total of 40 epochs for storefront accessibility detection, 60 epochs for pedestrian detection, and 50 epochs for COCO object detection.

DETR. Following the methodology described in Carion et al. (2020), we utilize ResNet-50 as the feature extractor and a transformer encoder–decoder for our visual detector. The learning rate for both ResNet-50 and the transformer encoder–decoder is set to 0.005, and a weight decay of 1e-4 is applied. To train the model effectively, we set the maximum number of training epochs to 120 for storefront accessibility detection and 200 for pedestrian detection. During the training process, we log the results every 5 epochs, allowing for detailed monitoring of the model's performance and progress. These settings ensure a comprehensive and robust training process for achieving accurate detection results.

To ensure a fair comparison, we fine-tuned the pretrained parameters on COCO of the two baseline models on both SAI and CityPersons+ datasets. The configurations of the SCR component for the three tasks are shown in Table 4.

5.3. Experimental results

In this section, we present the comparison results for object detection on the SAI dataset (Section 5.3.1) and pedestrian detection on the CityPersons dataset (Section 5.3.2). We conduct comparisons with baseline detectors, including Faster R-CNN and DETR, as well as our previous context learning approaches (Wang et al., 2022, 2023), considering various combinations of our context learning and utilization components. The evaluation focuses on performance metrics such as precision, recall, and mean average precision (mAP), providing insights into the effectiveness of our proposed framework in enhancing object and pedestrian detection tasks.

To ensure a fair comparison between our proposed framework and the previously designed MultiCLU particularly for storefront accessibility detection (Wang et al., 2022), we initially adopt the same settings as described in Wang et al. (2022). Specifically, we utilize the Small Object Dataset (SOD) standard to represent the local context for small objects in the SAI dataset. For this, we set the enlarge percentage to 15 percent, denoted as $\beta = 0.15$. Similarly, we employ the same small object standard for the CityPersons dataset, with the enlarge percentage set to 10 percent, denoted as $\beta = 0.10$. By using these consistent settings, we aim to facilitate a direct performance comparison between our proposed framework and MultiCLU.

5.3.1. Storefront object detection

In order to assess the effectiveness of our proposed general framework, we conducted a thorough comparison with two baseline detectors—Faster R-CNN (Ren et al., 2015) and DETR (Carion et al., 2020), and two of our previous context learning approaches (Wang et al., 2022, 2023), using the SAI dataset. Here we use MultiCLU to represent the specially designed multi-stage context framework with the CNN-based model Faster R-CNN, as reported in Wang et al. (2022), GMC-C to represent the GMC framework with the CNN-based model in this paper and also as reported in Wang et al. (2022), and GMC-T to represent the GMC framework on the DETR-based model. To gauge the effectiveness of our approach on small objects within the SAI dataset, we adopted the evaluation methodology outlined in Wang et al. (2022). Here, for the scenarios where the local context representation is employed, we leveraged both the original and expanded labels for small objects adhering to the defined criteria. In cases where both labels were detected for the same small object, we considered just one to eliminate any possibility of duplicate detections. The evaluation primarily focused on two key performance metrics: mean average precision (mAP) and recall. These metrics were measured at a standard Intersection over Union (IoU) threshold of 0.5, which is commonly used in object detection tasks.

Performance comparison on Faster R-CNN (Ren et al., 2015)**.** Our comparative analysis revealed significant performance improvements when applying our framework to the CNN-based models (represented in rows 1 to 3 of Table 5). Note for the SAI dataset, the GMC-C results have been reported in Wang et al. (2023), and the configuration is the same in this paper. Specifically, our GMC-C model outperformed Faster R-CNN, achieving substantial increases in both mAP (+13.6%) and recall (+15.3%). This highlights the effectiveness of our general context framework in enhancing object detection performance, surpassing the baseline detector. Furthermore, our GMC-C model exhibited a slightly higher mAP (+0.3%) compared to the special MultiCLU model, which employed specialized context mechanisms. However, there was a slight decrease in recall (−0.5%).

Table 5
Comparison results on SAI dataset (Wang et al., 2022) with baseline detectors and previous context learning approaches. IT: Inference Time (s).

Model	IT	Precision ↑			Recall ↑			mAP ↑	Recall ↑
		Door	Knob	Stair	Door	Knob	Stair		
Faster R-CNN (Ren et al., 2015)	0.029	75.6	17.7	66.0	87.5	47.6	73.1	53.1	69.4
MultiCLU (Wang et al., 2022)	0.036	78.0	51.2	70.0	92.3	80.4	83.0	66.4	85.2
+LCR	0.029	78.1	41.3	66.8	88.9	77.7	74.5	62.1	80.4
+SCF	0.036	78.0	19.0	68.5	90.1	53.0	79.4	55.2	74.2
+SCR	0.029	77.8	18.6	67.2	88.8	52.4	74.5	54.5	71.9
+LCR+SCF	0.036	78.4	50.0	69.2	90.8	75.0	79.4	65.9	81.7
+SCF+SCR	0.036	78.2	21.2	69.6	90.3	55.8	80.8	56.3	75.6
+LCR+SCR	0.029	79.2	41.2	67.8	89.2	77.8	74.5	62.7	80.5
GMC-C (Wang et al., 2023) & (this paper)	0.036	78.2	52.3	69.6	92.0	79.9	82.3	66.7	84.7
DETR (Carion et al., 2020)	0.040	75.9	23.8	69.2	91.8	58.4	77.8	56.3	76.0
+LCR	0.040	77.0	45.6	68.5	90.5	75.4	79.4	63.7	81.7
+SCF	0.045	77.8	27.6	70.0	91.4	61.5	81.2	58.5	78.0
+SCR	0.040	77.4	25.2	69.6	90.8	60.8	79.0	57.4	76.9
+LCR+SCF	0.045	80.2	55.1	71.2	92.7	81.2	82.3	68.8	85.4
+SCF+SCR	0.045	78.2	29.8	69.2	91.4	62.3	81.5	59.1	78.4
+LCR+SCR	0.040	78.8	50.8	69.2	92.0	77.8	80.4	66.3	83.4
GMC-T (this paper)	0.045	80.6	55.8	71.2	92.7	82.0	82.6	**69.2**	**85.8**

The comprehensive comparison outcomes demonstrate the compelling performance of our framework when integrated into CNN-based models. By incorporating various context learning and utilization components, our framework successfully enhances both mAP and recall, surpassing the performance of baseline detectors and previous context learning approaches. This reaffirms the potential and value of our general context framework in advancing the field of computer vision and object detection tasks.

Performance comparison on DETR (Carion et al., 2020)**.** To evaluate the flexibility and general applicability of our proposed framework, we extended its integration to the detection transformer architecture, represented by the DETR model (Carion et al., 2020). By incorporating the context learning components into the detection transformer, we conducted a comprehensive analysis of its impact on the detection performance. The evaluation results (rows 4 to 5 in Table 5) demonstrated significant improvements of our GMC-T model in both mean average precision (mAP) and recall compared to the baseline transformer model (DETR). Specifically, we observed a noteworthy increase of 12.9% in mAP and 9.8% in recall, highlighting the effectiveness of our context learning components in enhancing detection performance within the transformer framework. These findings further emphasize the adaptability and efficacy of our proposed framework, as it consistently improves detection performance across different model architectures. Note here that the transformer-based model already has context information learnt within the model, this is probably why the improvement (from DETR to GMC-T) is not as high as that on the CNN-based models (from Faster R-CNN to GMC-C). Nevertheless, the GMC-T model, which incorporates our context learning components into the detection transformer, emerged as the top-performing model among the evaluated configurations. This outcome underscores the versatility and effectiveness of our framework in enhancing detection capabilities across diverse model architectures, showcasing its potential for various object detection tasks.

Our proposed framework demonstrates superior performance on the SAI dataset, exhibiting significant improvements over the baseline detectors and delivering competitive results compared to our previous specially-designed context learning model MultiCLU (Wang et al., 2022). These findings support the efficacy of our general context framework in improving object detection accuracy and recall rates, meanwhile adapting to different visual detector architectures. By efficiently leveraging contextual information, our framework enhances object detection accuracy and recall rates, demonstrating its flexibility and effectiveness in various detection scenarios.

Performance comparison with different context components. We embarked on a comprehensive performance comparison across various combinations of our three contextual components. The outcomes, presented in Table 5, illuminate compelling insights.

First we analyze the performance improvements when using various combinations of contextual components on Faster RCNN. When each contextual component was applied in isolation, notable enhancements in recall (from 2.8% to 11%) and mAP (from 1.4% to 9%) over the baseline were discernible. Furthermore, it is intriguing to observe that when deploying individual contextual components, the impact of local contextual labeling was more pronounced than that of the other two components.

Upon considering combinations of two contextual components, a noteworthy trend emerged, with each combination outperforming the baseline detector. The improvements ranged from +3.2% to 12.8% for mAP and from 6.2% to 12.3% for recall. Strikingly, when the combinations encompassed the Local Context Representation (LCR) component, they exhibited substantial superiority over other combinations, showcasing considerable gains in both mAP (+6.4% to 9.6%) and recall (+4.9% to 6.1%). This outcome underscores the value of incorporating contextual information around small objects, notably accentuating the detection efficacy of vital elements like doorknobs. Moreover, in relation to the single LCR component, both Semantic Context Fusion (SCF) and Spatial Context Reasoning (SCR) exhibited positive impacts. These components further improved results over a single LCR component, influencing both mAP and recall positively. Intriguingly, when contrasting the application of both SCF and SCR against their individual application, the combined utilization marginally enhanced both mAP and recall compared to using them in isolation.

The apex of our proposed framework's performance emerged with the integration of all three components (GMC-C), attaining a notable 13.6% improvement in mAP and an impressive 15.3% enhancement in recall over the baseline model Faster R-CNN. An interesting observation lies in the fact that our general framework enhances mAP across all categories in contrast to MultiCLU (Wang et al., 2022), albeit with only minimal reductions in recall. This suggests that the specifically designed MultiCLU might introduce more false positives than accurate predictions, positioning our framework to offer heightened precision at the cost of slightly reduced recall.

One notable distinction between the two base models lies in the impact of the Local Context Representation (LCR) component. Specifically, the improvements achieved by using LCR with DETR are not as substantial as those observed with Faster R-CNN. When solely applying the LCR component to Faster R-CNN, there is a remarkable enhancement in Precision and Recall for the "knob" category, with improvements of 23.6% and 30.1%, respectively. In contrast, when the LCR component is applied to DETR alone, the precision and recall see improvements of 21.8% and 17.0%, respectively, which are comparatively less effective than with Faster R-CNN. Moreover, the mAP and recall for Faster R-CNN see enhancements of 9.0% and 11.0%,

Table 6
Comparison results on Citypersons dataset (Zhang et al., 2017) with baseline detectors and previous context learning approaches. IT: Inference Time (s).

Model	IT	Reasonable ↓	Heavy ↓
Faster R-CNN (Ren et al., 2015)	0.062	13.4	36.9
+LCR	0.062	12.3	35.6
+SCF	0.068	13.3	37.1
+SCR	0.063	13.0	36.5
+LCR+SCF	0.068	12.2	35.2
+SCF+SCR	0.069	13.2	36.5
+LCR+SCR	0.063	12.0	36.0
GMC-C (Wang et al., 2023)& (this paper)	0.069	12.0	**35.2**
DETR (Carion et al., 2020)	0.059	11.8	40.8
GMC-T (this paper)	0.063	**10.5**	39.5

whereas DETR experiences improvements of 7.4% and 5.7%, respectively, when the LCR component is added. This discrepancy could be attributed to the inherent self-attention mechanism of the transformer architecture, which inherently incorporates context information of local context especially for small objects, a feature that Faster R-CNN lacks. Nevertheless, the performance improvements achieved through various combinations of contextual components on DETR exhibit similar trends, indicating the consistent and robust functionality of the GMC framework across different backbone models.

5.3.2. Pedestrian detection

We conducted further evaluation of our general context learning and reasoning framework on pedestrian detection task using CityPersons dataset, comparing it with the baseline detectors, Faster R-CNN (Ren et al., 2015) and DETR (Carion et al., 2020), without any code modifications. Here again, we use GMC-C to represent the general framework of context learning with the CNN-based model, and GMC-T to represent the general framework on DETR-based model, on the original CityPersons dataset (without considering the subcategories or additional context for spatial context reasoning). In summary, in the labeling stage, we employ the small object standard for the CityPersons dataset to enhance the labeling of small objects with local context labeling. We further leverage the fine-grained category rider in CityPersons dataset to enable the semantic context fusion in the training stage, and the spatial context reasoning in the postprocessing stage. Note that the GMC-C model in this paper is the same as that in Wang et al. (2023).

Further, we use GMC-C+ and GMC-T+ to represent the general framework with more spatial context reasoning, using the CityPersons+ dataset with subcategories of pedestrians and riders, as well as information of vehicle, road and sidewalk. We compared the evaluation results on the *reasonable* and *heavy* subsets of the data using the standard evaluation metric in pedestrian detection, MR^{-2} (where lower values indicate better performance). Here, the subsets were defined based on the height (h) and visible ratio (v) of pedestrians: Reasonable subset: $h \in [50, \infty]$, $v \in [0.65, 1]$; Heavy subset: $h \in [50, \infty]$, $v \in [0, 0.65]$.

Overall comparison with baseline detectors. The comparison results presented in Table 6 provide insights into the performance of the GMC framework on different architectures on both the reasonable and heavy subsets. It is observed that DETR and transformer-based GMC model (GMC-T) generally exhibits superior performance on the reasonable subset (+1.6% and +2.9%, respectively, compared to the Faster-RCNN base model), indicating its effectiveness in capturing contextual information and enhancing detection accuracy. However, DETR and GMC-T demonstrates lower performance on the heavy subset (−2.6% and −3.9% respectively, compared to the Faster-RCNN base model), which could be attributed to the absence of design elements such as the feature pyramid network (FPN) (Lin et al., 2017) employed in the Faster R-CNN framework. In contrast, the CNN-based model GMC-C may not achieve the same level of performance on the reasonable subset as transformer-based model GMC-T, but it often demonstrates better performance on the heavy subset (+1.7% compared to the Faster-RCNN base model). This suggests that the CNN-based model are able to effectively handle challenging scenarios with heavily occluded pedestrians, where precise localization and robust feature extraction are crucial. This evidence supports our rationale of the general context framework in working with various backbone models depending on the task requirements.

Performance comparison with different context components on Faster-RCNN. Upon applying the Local Context Representation (LCR) component alone on Faster R-CNN, there was a noticeable enhancement of 1.1% on the reasonable subset and 1.3% on the heavy subset (as illustrated in Table 6). To further amplify our framework's capabilities, we introduced a fine-grained category (rider) into the CityPersons dataset during training to facilitate the Semantic Context Fusion (SCF) and Spatial Context Reasoning (SCR) components. As observed in the results analogous to those from the SAI dataset, configurations with the LCR component consistently yielded superior performance compared to other settings. However, it is worth noting that both SCF and SCR modules had a minor impact on pedestrian detection, possibly attributed to the relatively weak correlation between pedestrians and other urban objects. In summation, our comprehensive framework, encompassing all three components, achieved the most impressive performance across both the reasonable subset (1.4% lower) and the heavy subset (1.7% lower), outperforming the baseline detector and alternative combinations.

Comparison with DETR. Upon comparing our newly introduced GMC-T model with the baseline Detection Transformer (DETR) model, our GMC-T model consistently demonstrated superior performance across both the "reasonable" and "heavy" subsets. This was marked by a substantial enhancement in detection performance, exhibiting an impressive 1.3% improvement on both subsets. These results provide compelling evidence for the effectiveness of our context learning and reasoning components in bolstering the detection capabilities of diverse architectural frameworks. Moreover, our framework's adaptability is evident as it showcases its prowess not only in CNN-based models but also in transformer-based models. The ease with which our framework can be integrated and customized underscores its potential to cater to a range of visual detection tasks beyond just pedestrian detection.

Overall, the comparison results highlight the potential and versatility of our proposed context learning and reasoning components in improving object detection performance across different datasets and tasks. The framework offers a flexible and effective solution for incorporating context information and enhancing the detection capabilities of various deep learning models, contributing to advancements in the field of computer vision and object detection.

The effectiveness of the general Spatial Context Reasoning (SCR). We also conducted an extensive study to evaluate the effectiveness of the general spatial context reasoning (SCR) component within our framework. In order to achieve a more comprehensive and robust topological reasoning, we leveraged both bounding boxes for objects (such as bicycles, motorcycles, cars, pedestrians) and semantic masks for stuff (such as sidewalks and roads) in CityPersons+ dataset. This allowed us to capture and utilize the spatial relationships between various entities in the scene. To assess the impact of the enhanced general SCR component, we evaluated its performance in two enhanced models—GMC-C+ and GMC-T+, as well as its use on the two baseline object detection models—Faster R-CNN and DETR. Table 7 presents the comparative results of these models with and without the SCR component.

(1). SCR performance on Faster R-CNN. When we solely applied the SCR component to the Faster R-CNN model, we observed notable improvements in performance for both the reasonable and heavy subsets, achieving an increase of 0.6% and 0.8%, respectively. However, it is important to note that the Faster R-CNN model, without the inclusion of the local context and semantic context components, did not achieve the same level of performance as the GMC-C model. By replacing the initial spatial context reasoning component with our enhanced SCR component in the GMC-C model, leading to the GMC-C+ model, we

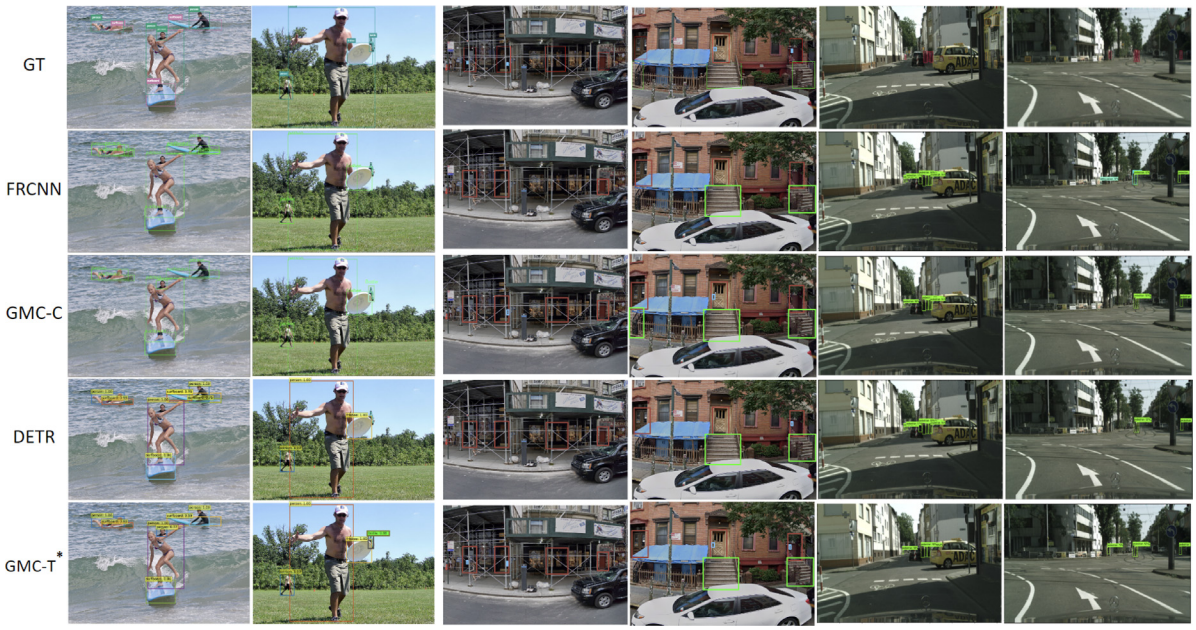


Fig. 12. Qualitative results on the three datasets: COCO (columns 1 & 2), SAI (columns 3 & 4) and CityPersons+ (columns 5 & 6). GMC-T*: We only evaluate the SCF and SCR components on COCO dataset, and the GMC-T was evaluated on the other two datasets.

observed a slight performance improvement of 0.2% on the reasonable subset and 0.4% on the heavy subset, over the GMC-C model. These results indicate that the integration of the enhanced SCR component can enhance the performance of the GMC-C model to some extent. However, when comparing these results with the performance of the enhanced SCR component alone (i.e., Faster R-CNN + SCR), it is evident that the GMC-C+ model with the combined local context, semantic context, and enhanced SCR component outperformed both subsets, achieving a significant improvement of 1.0% on the reasonable subset and 1.3% on the heavy subset. This demonstrates the synergistic effect of incorporating multiple context sources within the framework. our evaluation confirms that the integration of the enhanced general SCR component can effectively improve the performance of object detection models, particularly when combined with the local context and semantic context components. Overall, GMC-C+ achieves performance improvements of 1.6% on the reasonable and 2.1% on the heavy, compared to the Faster-RCNN base model.

(2). SCR performance on DETR. We also study whether our enhanced general SCR component can improve over the DETR model, which already incorporates a self-attention mechanism to leverage context information. Not surprisingly, even with the existing self-attention mechanism, the application of the enhanced SCR component to the DETR model led to performance improvements. Specifically, we observed an increase of 0.6% on the reasonable subset and 1.0% on the heavy subset, indicating that the SCR component can effectively enhance the context utilization capabilities of the DETR model. Furthermore, when we combined the general SCR component with the other two contextual components (local context and semantic context), our GMC-T+ model achieved additional performance improvements over the DETR model and the GMC-T model on both evaluation subsets. The results showed a significant improvement of 1.6% on the reasonable subset and 2.2% on the heavy subset, compared to the DETR base model, and a visible improvement of 0.3% on the reasonable subset and 0.9% on the heavy subset, compared to the GMC-T model. This highlights the complementary nature of the contextual components and their ability to further enhance the detection performance of the DETR model.

Our evaluation on pedestrian detection task confirms that the integration of the more general SCR component can effectively improve the performance of the detection models, particularly when combined with the local context and semantic context components. Our three contextual components, when integrated with the DETR model, demonstrated the best performance on the reasonable subset. On the other hand, the three contextual components combined with the CNN-based model Faster R-CNN exhibited better performance on the heavy subset. These findings indicate that the choice of model architectures, in combination with the specific context components, can have an impact on the overall detection performance, with different configurations achieving better results on different evaluation subsets. This also highlights the importance of leveraging multiple context sources and considering the spatial relationships between objects for achieving more accurate and robust detection.

Table 7
Comparison results on general spatial context reasoning (SCR) component with baseline detectors and previous designed component.

Model	Reasonable ↓	Heavy ↓
Faster R-CNN (Ren et al., 2015)	13.4	36.9
Faster R-CNN + SCR	12.8	36.1
GMC-C (Wang et al., 2023) &(this paper)	12.0	35.2
GMC-C+ (this paper)	11.8	**34.8**
DETR (Carion et al., 2020)	11.8	40.8
DETR + SCR	11.2	39.8
GMC-T (this paper)	10.5	39.5
GMC-T+ (this paper)	**10.2**	38.6

5.4. COCO object detection

In order to check the scalability of our proposed general framework, we evaluate our framework on a large detection benchmark COCO dataset. We conducted comparison with two baseline detectors—Faster (Ren et al., 2015) and DETR (Carion et al., 2020). We focus on two performance metrics: average precision (AP) and average precision for small objects (AP_S). The comparison results are shown in Table 8.

Performance comparison on Faster R-CNN (Ren et al., 2015)**.** Our comprehensive comparison results underscore the efficacy of our proposed GMC-C model, revealing significant improvements in key metrics. The average precision (AP) metric, a crucial indicator of overall detection performance, exhibited a notable enhancement of +0.7% when employing our framework compared to the baseline Faster

Table 8
Comparison results on COCO dataset (Lin et al., 2014) with baseline detectors. IT: Inference Time (s).

Model	IT	AP ↑	AP_S ↑
Faster R-CNN (Ren et al., 2015)	0.028	37.4	21.2
+LCR	0.028	37.6	21.5
+SCF	0.040	37.6	21.3
+SCR	0.030	37.5	21.2
+LCR+SCF	0.030	37.9	21.6
+SCF+SCR	0.040	37.8	21.4
+LCR+SCR	0.028	37.7	21.6
GMC-C	0.040	38.1	21.7
DETR (Carion et al., 2020)	0.036	42.0	21.0
+SCF	0.042	42.3	21.4
+SCR	0.037	42.2	21.2
+SCF+SCR	0.042	42.7	21.5

R-CNN. Moreover, our model demonstrated a noteworthy advancement in AP for small objects, registering an improvement of +0.5%. This targeted improvement underscore the effectiveness of our proposed framework, particularly in addressing the detection challenges associated with smaller objects within the visual scene. The results substantiate the adaptability and enhanced performance of our GMC-C model, positioning it as a valuable asset in scenarios demanding precise and comprehensive object detection.

The application of the Local Context Representation (LCR) component in isolation on the Faster R-CNN model resulted in a modest improvement, with a 0.2% increase in average precision (AP) and a 0.3% enhancement in AP_S (as detailed in Table 8). Remarkably, when the LCR component was synergistically combined with the Semantic Context Fusion (SCF) component, this pairing exhibited the most substantial improvement compared to other combinations. The joint application yielded a 0.5% boost in AP and a 0.4% increase in AP_S. It is noteworthy that the individual application of the SCF and Spatial Context Reasoning (SCR) modules had a comparatively minor impact on the COCO dataset. In summary, our holistic framework, encompassing all three components, demonstrated the most remarkable performance improvement across both AP (+0.7%) and AP_S (+0.5%), surpassing the baseline detector and alternative component combinations.

Performance comparison on DETR (Carion et al., 2020)**.** In our evaluation using DETR, the impact of our context components becomes apparent when applied individually. Since we have to fine-tune the large DETR model for LCR, we only tested performance improvements for the other two components (SCF and SCR) as the DETR can be frozen when training SCF and no re-training is needed for SCR. The Semantic Context Fusion (SCF) component, when introduced on its own, yields notable enhancements with a relative increase of +0.3% on AP and +0.4% on AP_S. This signifies that incorporating semantic relationships between objects contributes positively to the overall detection performance.

Conversely, the Spatial Context Reasoning (SCR) component, when applied independently, demonstrates a more modest impact, with only a +0.2% improvement on both AP and AP_S. This result is suggestive of the challenges associated with defining meaningful relations between objects in the COCO dataset, where the provided relations are limited.

Interestingly, the synergy between SCF and SCR components becomes evident when they are combined. Their complementary nature enhances each other's contributions, resulting in a more substantial improvement. The joint application of SCF and SCR leads to a further increase in performance, with a +0.7% improvement on AP and +0.5% on AP_S. This collaborative effect underscores the value of integrating both semantic and spatial context reasoning for more effective object detection within the DETR framework.

5.5. Performance discussions for different tasks/datasets

With more in-depth examinations, we sought to delineate the specific object categories that exhibit significant influence from the Spatial Context Reasoning (SCR) component across the diverse datasets we scrutinized. As shown in Table 9, within the SAI dataset, the SCR component dynamically integrates contextual relationships for all three categories—door, knob, and stair. Transitioning to the CityPersons+ dataset, the SCR component extends its reach across the entire spectrum of object categories. Notably, contextual elements like road and sidewalk draw upon insights from a state-of-the-art segmentation model, leading to a pronounced impact on 75% of the dataset's categories. In the case of the COCO dataset, the SCR component centers its focus on the person category, given its preeminence as the most abundant class in the dataset. While other categories also experience influence, the overall impact encompasses approximately 13.75% of all object categories within the COCO dataset.

Table 9
Impacted categories for all datasets in SCR component.

Datasets	Impacted categories	Percentage
SAI	3/3	100
Citypersons+	6/8	75
COCO	11/80	13.75

We further conducted evaluations to assess how our components perform on the most impacted categories across all datasets, and the summarized results are presented in Table 10. In the SAI dataset, the substantial improvement of +21.8% in AP for the "knob" category, achieved by applying the Local Context Representation (LCR) component with DETR, underscores the pivotal role of contextual information in detecting and delineating small objects. This result suggests that leveraging local context in tandem with transformer-based models significantly benefits the identification of intricate details in specific categories. Moving to the CityPersons+ dataset, where the "pedestrian" category exhibited the most notable enhancement of +1.1% on the reasonable set and +1.3% on the heavy set with the LCR component on Faster R-CNN, we observe the importance of local context in urban scenes. The improved detection performance for pedestrians, a crucial element in urban scenarios, emphasizes the significance of considering context for specific object classes. This insight becomes especially valuable in the domain of object detection, where capturing fine-grained details is essential.

In the COCO dataset, the "person" category's substantial improvement of 3.6% in AP with the Spatial Context Reasoning (SCR) component applied to the DETR model suggests that accounting for spatial relationships is particularly beneficial in datasets characterized by a larger scale and diverse object categories. Spatial reasoning plays a crucial role in refining the predictions, especially in scenarios where objects interact in complex spatial configurations. Although Semantic Context Fusion (SCF) did not exhibit standout improvements compared to the other two components, its role in contributing to enhanced performance, especially when combined with LCR and SCR components, underscores its potential in capturing contextual semantics. This holistic approach, leveraging different forms of context throughout the entire deep learning process, demonstrates promising results and sets the stage for further exploration in context-aware computer vision tasks.

Furthermore, we conducted a thorough comparison of the inference times (expressed in seconds) across our results(Tables 5, 6 and 8). The findings revealed that our framework incurs only a marginal increase in time complexity. Furthermore, the qualitative results visualized in Fig. 12 provide a compelling illustration of how the proposed method enhances performance across all three datasets (COCO, SAI, and CityPersons+), offering a comprehensive validation of its efficacy.

6. Conclusions and discussion

In summary, we have proposed a general framework of multistage context learning and utilization for visual detection tasks. Our proposed framework consists three context components to utilize local

Table 10
Component performance on most impacted categories on all dataset. D:DETR. F:Faster R-CNN.

Dataset	Category	Model	AP ↑	Reasonable↓	Heavy↓
SAI	knob	D+LCR	23.8 → 45.6	–	–
CityPersons+	pedestrian	F+LCR	–	13.4 → 12.3	36.9 → 35.6
COCO	person	D+SCR	47.3 → 50.9	–	–

context, semantic context and spatial context information. The three context components have the flexibility and adaptability to utilize the framework across various visual detection tasks, with different visual detectors. The proposed framework are evaluated and verified on complex street scenes for a storefront object detection task and a pedestrian detection task. Compared to the state of the art methods, the evaluation demonstrates that our framework can efficiently leveraging contextual information at various stages such as data preprocessing, model training and post-processing. Our comparison results also show that the proposed contextual components can effectively improve the performance over different baseline models, with the support of different context information.

However, there is still space for improvements over the proposed framework. In this work, we only explore local, global and semantic context, mostly in the spatial domain. Other context types need more attention, and new architectures particularly designed for context learning and utilization as summarized in Wang and Zhu (2023) have not been considered.

Despite our attempt in conducting experiments on the extensive MSCOCO dataset to show promising results, defining general spatial relations of all object categories becomes a challenge, especially when dealing with a dataset that encompasses numerous categories. The task of establishing meaningful and universally applicable spatial relations becomes intricate due to the diversity of object categories present in the dataset. Addressing this challenge requires a thoughtful approach to derive spatial relations that can effectively generalize across a wide range of object types. Further exploration and research may be needed to develop a robust and adaptable method for defining spatial relations that accommodates the inherent diversity of categories within the dataset.

Furthermore, there are many works focus on the real world detection scenarios, where the standard evaluation metrics may not work well. A contextual evaluation based on the requirements of real-world applications is needed not only for object detection task, but may also benefit other computer vision tasks.

CRediT authorship contribution statement

Xuan Wang: Conceptualization, Formal analysis, Investigation, Visualization, Writing – original draft. **Hao Tang:** Conceptualization, Funding acquisition, Supervision. **Zhigang Zhu:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The work is supported by the National Science Foundation (NSF), USA through Awards #2131186 (CISE-MSI), #1827505 (PFI), and #1737533 (S&CC). The work is also supported by the US Air Force Office of Scientific Research (AFOSR), USA via Award #FA9550-21-1-0082, a College-wide Research Vision (CRV) Fund from the CCNY Provost's Office, and the ODNI Intelligence Community Center for Academic Excellence (IC CAE) at Rutgers University, USA (#HHM402-19-1-0003 and #HHM402-18-1-0007).

References

Ahmetovic, D., Manduchi, R., Coughlan, J.M., Mascetti, S., 2015. Zebra crossing spotter: Automatic population of spatial databases for increased safety of blind travelers. In: Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility. pp. 251–258.

Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N., 2016. A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 354–370.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, pp. 213–229.

Cavallo, M., 2015. 3D city reconstruction from google street view.

Chacra, D.A., Zelek, J., 2022. The topology and language of relationships in the visual genome dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4860–4868.

Chen, C., Liu, M.Y., Tuzel, O., Xiao, J., 2017. R-CNN for small object detection. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (Eds.), Computer Vision – ACCV 2016. Springer International Publishing, Cham, pp. 214–230.

Chen, Z.M., Wei, X.S., Wang, P., Guo, Y., 2019. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186.

Cheng, M., Zhang, Y., Su, Y., Álvarez, J.M., Kong, H., 2018. Curb detection for road and sidewalk detection. IEEE Trans. Veh. Technol. 67, 10330–10342.

Clementini, E., Felice, P.D., Oosterom, P.v., 1993. A small set of formal topological relationships suitable for end-user interaction. In: International Symposium on Spatial Databases. Springer, pp. 277–295.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, http://dx.doi.org/10.1109/CVPR.2016.350.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Du, Y., Duan, G., Ai, H., 2012. Context-based text detection in natural scenes. In: 2012 19th IEEE International Conference on Image Processing. IEEE, pp. 1857–1860.

Dvornik, N., Mairal, J., Schmid, C., 2018. Modeling visual context is key to augmenting object detection datasets. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 364–380.

Egenhofer, M.J., Franzosa, R.D., 1991. Point-set topological spatial relations. Int. J. Geogr. Inf. Syst. 5 (2), 161–174.

Fang, Y., Kuan, K., Lin, J., Tan, C., Chandrasekhar, V., 2017. Object detection meets knowledge graphs. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI-17, pp. 1661–1667. http://dx.doi.org/10.24963/ijcai.2017/230.

Google, 2022. Google Street View API. Google, URL https://developers.google.com/maps/documentation/streetview/overview.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lee, C.W., Fang, W., Yeh, C.K., Wang, Y.C.F., 2018. Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1576–1585.

Leng, J., Ren, Y., Jiang, W., Sun, X., Wang, Y., 2021. Realize your surroundings: Exploiting context information for small object detection. Neurocomputing 433, 287–299.

Li, Y., Huang, C., Loy, C.C., Tang, X., 2016a. Human attribute recognition by deep hierarchical contexts. In: European Conference on Computer Vision. Springer, pp. 684–700.

Li, Q., Qiao, M., Bian, W., Tao, D., 2016b. Conditional graphical lasso for multi-label image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2977–2986.

Li, X., Zhao, F., Guo, Y., 2014. Multi-label image classification with a probabilistic label enhancement model. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, vol. 1, (2), pp. 1–10.

Lim, J.S., Astrid, M., Yoon, H.J., Lee, S.I., 2021. Small object detection using context and attention. In: 2021 International Conference on Artificial Intelligence in Information and Communication. ICAIIC, IEEE, pp. 181–186.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing.

Mac Aodha, O., Cole, E., Perona, P., 2019. Presence-only geographical priors for fine-grained image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9596–9606.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 1532–1543.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, Advances in neural information processing systems, vol. 28.91–99,

Sharma, M., Rasmuson, D., Rieger, B., Kjelkerud, D., et al., 2019. Labelbox: The best way to create and manage training data. https://www.labelbox.com.

Sun, J., Jacobs, D.W., 2017. Seeing what is not there: Learning context to determine where objects are missing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5716–5724.

Wang, X., Chen, J., Tang, H., Zhu, Z., 2022. Multiclu: Multi-stage context learning and utilization for storefront accessibility detection and evaluation. In: Proceedings of the International Conference on Multimedia Retrieval. ICMR '22, Association for Computing Machinery, New York, NY, USA, pp. 304–312. http://dx.doi.org/10.1145/3512527.3531361.

Wang, X., Ji, Q., 2015. Video event recognition with deep hierarchical context model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4418–4427.

Wang, X., Ji, Q., 2016. Hierarchical context modeling for video event recognition. IEEE Trans. Pattern Anal. Mach. Intell. 39 (9), 1770–1782.

Wang, X., Tang, H., Zhu, Z., 2023. A general context learning and reasoning framework for object detection in urban scenes. In: Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP. SciTePress, INSTICC, pp. 91–102. http://dx.doi.org/10.5220/0011637600003417.

Wang, X., Zhu, Z., 2023. Context understanding in computer vision: A survey. Comput. Vis. Image Underst. 229, 103646.

Wu, J., Zhou, C., Zhang, Q., Yang, M., Yuan, J., 2020. Self-mimic learning for small-scale pedestrian detection. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 2012–2020. http://dx.doi.org/10.1145/3394171.3413634.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L., 2017. Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5419.

Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D., 2018. Graph r-cnn for scene graph generation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 670–685.

Yang, S., Luo, P., Loy, C.C., Tang, X., 2015. From facial parts responses to face detection: A deep learning approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3676–3684.

Yang, X., Yang, X., Liu, M.Y., Xiao, F., Davis, L.S., Kautz, J., 2019. Step: Spatio-temporal progressive learning for video action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 264–272.

Zellers, R., Yatskar, M., Thomson, S., Choi, Y., 2018. Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5831–5840.

Zhang, S., Benenson, R., Schiele, B., 2017. Citypersons: A diverse dataset for pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3221.

Zhou, C., Yuan, J., 2018. Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of the European Conference on Computer Vision.

Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M., 2021. Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8782–8791.

Zhu, A., Gao, R., Uchida, S., 2016. Could scene context be beneficial for scene text detection? Pattern Recognit. 58, 204–215.

Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K., 2013. Context-aware modeling and recognition of activities in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2491–2498.