Spatial relationships knowledge integration in deep learning modeling for panoptic segmentation in urban driving scenarios

F E Benkirane, N Crombez, V Hilaire, Y Ruichek

Spatial relationships knowledge integration in deep learning modeling for panoptic segmentation in urban driving scenarios

FE. Benkirane[1], N. Crombez[1], V. Hilaire[1], Y. Ruichek[1]
[1]UTBM, CIAD UMR 7533, F-90010 Belfort, France
Fatima.benkirane@utbm.fr

Abstract - Panoptic segmentation is a computer vision task that aims to identify and analyze all objects present in an image. While semantic segmentation focuses on labeling each pixel in an image with a category label, panoptic segmentation goes further by not only assigning semantic labels but also identifying and distinguishing individual instance of objects. This task is valuable for various applications, such as robotics, surveillance systems or autonomous vehicle navigation. In this work, we propose a new informed deep learning approach that combines the strengths of deep neural networks for panoptic segmentation with additional knowledge about spatial relationships between objects. This is particularly important as spatial relationships can provide useful cues for resolving ambiguities, distinguishing between overlapping or similar object instances, and capturing the holistic structure of the scene. We propose a novel training methodology that integrates knowledge directly into the deep neural network optimization process. Our approach includes a process for extracting and representing spatial relationships knowledge, which is incorporated into the training using a specially designed loss function. The effectiveness of the proposed method is evaluated and validated on various challenging urban driving datasets.

Keywords: Hybrid AI, deep learning, panoptic segmentation, spatial relationships

I. INTRODUCTION

Panoptic segmentation, is a computer vision task designed to recognize and categorize all elements within an image by integrating information from both semantic and instance segmentation. Semantic segmentation divides an image into regions associated with non-quantifiable object classes, often referred to as “Stuff", which can include elements like the sky or the road. It is also able to categorize quantifiable objects, but it does not provide individual distinction. In contrast, instance segmentation, involves the precise identification of individual quantifiable objects in the image, referred to as “Things", such as cars or pedestrians. Panoptic segmentation ability to comprehensively describe and analyze images offers practical solutions across a range of applications. In the domain of mobile robotics, for example, it plays a pivotal role in the detection and tracking of moving objects [1]. Furthermore, this task significantly contributes to the field of autonomous driving, empowering vehicles to gain a deep understanding of their surroundings and make precise decisions [2] [3]. Since 2018, there has been a growing interest within the scientific community regarding the prediction of panoptic segmentation [4]. This approach is recognized as a collaborative one that combines the strengths of both the semantic and instance segmentation methods. Panoptic segmentation techniques find common use in image data, relying on various DL-based strategies. Some of these methods involve employing distinct neural sub-networks for semantic and instance predictions [4]. However, this dual-network approach can be complex and have limitations in terms of effectiveness, often necessitating complicated post-processing to merge the associated predictions [5]. To address these limitations, a novel category of panoptic segmentation techniques has emerged, based on the use of a shared backbone [5]. These approaches enhance the training PROCESS. Previous studies have demonstrated the significant impact of contextual information and object relationships in enhancing computer vision tasks, particularly in the domain of object detection [6] [7]. These investigations have primarily used post-processing techniques to reevaluate identified objects CONSIDERING object relationships, such as co-occurrence [8]. For example, certain objects, such as a sofa and a traffic sign, are not typically expected to co-exist within the same scene due to their associations with different environments, indoors and outdoors, respectively. It is worth noting that most of these studies were conducted before the widespread integration of DL techniques. Within the realm of DL research, there has been limited progress in employing object relations to enhance object detection tasks. Most current methods remain primarily focused on the recognition and identification of objects, regardless of their relationships [9]. One of the main challenges in this context lies in the complexity of modeling the spatial relations between objects, considering their potential disparities in position within an image, varying scales, and diverse shapes, to cite just a few. On the other hand, some research has demonstrated that Convolutional Neural Networks (CNNs) have certain abilities to learn contextual information automatically and implicitly during

training [10]. By using local receptive fields [11], CNNs capture contextual details within small local regions connected to each neuron. These outcomes highlight the importance of providing to deep neural networks explicit access to contextual information to further enhance their performance and accuracy. As the research on deep networks continues to explore their abilities to learn contextual information, it becomes evident that further enhancing their performance and accuracy can be achieved by incorporating explicit access to contextual knowledge. This is consistent with the concepts of hybrid intelligent systems which aim to combine the strengths of artificial intelligence with human expertise [12]. Within the field of hybrid Artificial Intelligence (hybrid AI), an outstanding approach is informed deep learning [13], which uses prior knowledge or domain expertise to enhance the learning performances of deep learning models. This knowledge can come in various forms, such as expert rules, ontologies, statistical information, to name but a few. By incorporating this prior knowledge, deep learning models can make more informed predictions, and improve the decision-making process. Based on the aforementioned analysis, incorporating contextual information into deep learning techniques should be advantageous for computer vision tasks. Contextual information can be globally defined as the surrounding cues in the environment that provide additional insights and understanding to aid in accurate estimations and predictions. In this context, we have observed that panoptic prediction in urban environments is particularly challenging due to the complex relationships between regions within an image. To address this issue, the key contributions of this paper are as follows:

1. the extraction and integration of knowledge about spatial relationships into deep neural network for panoptic segmentation,
2. the modeling of the spatial relationships as a loss function to optimize the network training,
3. the validation and evaluation of the proposed approach on urban scene dataset.

TO present our approach, the remainder of this paper is organized as follows. Work related to panoptic segmentation are introduced and discussed in Section 2. The considered spatial relationships are described in Section 3. The proposed methodology including the loss function modeling of spatial relationships is described in Section 4. Section 5 presents the performed experiments, the results analysis and comparison with the state of the art. Finally, the last section concludes the paper and provides directions for future work.

II. STATE OF THE ART

We provide an overview of the existing panoptic segmentation approaches, specifically those that are based on a shared backbone architecture. These methods use a single neural network backbone for both ``Stuff'' and ``Things'' segmentation to achieve a unified panoptic segmentation of the image. Over the years, many frameworks have been developed following different techniques for panoptic segmentation. One effective approach is to use a shared backbone to encode features [14], as it has shown to yield high performance on benchmark datasets [15]. Within this category of techniques, there are two main approaches. The first one involves sharing a backbone between the two heads of semantic and instance segmentation and merges the outputs for the final panoptic generation. In addition to the shared backbone, the second category includes explicit connections between the two heads. Many methods have been proposed in the state of the art that can be classified into one of these two categories. In this section, we review some of the most important methods in each category and present their contributions to the field of panoptic segmentation.

The approach proposed in [16] performs instance and semantic segmentation separately and then applies the Non-Maximum Suppression (NMS) technique to obtain the Panoptic Quality (PQ) metric. The NMS procedure is used to produce non-overlapped instance regions, which are then combined with the semantic segmentation. The Efficient Spatial Pyramid of dilated convolutions (ESPnet) was introduced in [17]. This method involves several stages, including a shared backbone that consists of a Fetaure Pyramid Network (FPN) [18] and a Residual Network (ResNet) [19]. To enhance the input features, the method uses a Cross-Layer Attention (CLA) fusion module, which combines multi-layer feature maps in the FPN layer. The approach proposed in [14] introduces the Efficient Panoptic Segmentation (EfficientPS) architecture for scene understanding. The general architecture of the network consists of a shared backbone that encodes and fuses semantically rich multi-scale features. It includes a new semantic head that aggregates fine and contextual features consistently. For the instance segmentation head, a new variant of Mask R-CNN [14] augmented with depth-wise separable convolutions [20] is considered. A new system called Panoptic-DeepLab for panoptic segmentation is presented in [21]. The approach based on a dual-Atrous Spatial Pyramid Pooling (ASPP) and dual-decoder structure specific to semantic and instance segmentation respectively. The semantic branch follows the standard design of a semantic segmentation model, while the instance branch is class-agnostic and uses a simple instance center regression.

Some alternative cooperative techniques for panoptic segmentation have been proposed [22]. These techniques are also based on a shared-backbone architecture in addition to explicit connections between the instance and semantic segmentation heads. The approach outlined in [22], involves using a ShuffleNet [23] for feature extraction, as well as establishing explicit connections between the instance and semantic segmentation stages. These steps are followed by combining the results to produce the final panoptic output. A deep panoptic segmentation method that relies on a bidirectional learning technique is presented in [24]. To capture the intrinsic

interaction between semantic and instance segmentation, the authors introduce a Bidirectional Aggregation Network called BANet [24]. This network performs panoptic segmentation by leveraging two modules that extract rich contextual features from semantic and instance segmentation for recognition and localization. Finally, the bidirectional paths are used for feature aggregation, enhancing the overall segmentation performance.

On the other hand, the architecture proposed in [25] allows information exchange between the branches to leverage the benefits of both. Specifically, it involves leveraging semantic information to improve the instance segmentation. The output from the semantic segmentation branch is normalized and concatenated with the normalized features from the feature map. This concatenated information is passed through a convolutional layer and used as input to the instance segmentation branch. This allows relevant data from one branch to flow through the other, improving the performance of both semantic and instance segmentation branches.

Based on the state-of-the-art works, it is difficult to definitively conclude that one architecture always outperforms the other in all aspects considering panoptic segmentation. The choice depends on various factors such as the specific deep neural network architecture, the characteristics of the dataset, etc. Different datasets, tasks, and contexts may favor one architecture over the other. Ultimately the selection should be based on a careful consideration of the trade-offs between simplicity, computational efficiency, integration, and performance, as well as the available resources for training and inference.

III. QUALITATIVE SPATIAL RELATIONSHIPS (QSR)

The 3D objects of an urban scene are projected into acquired 2D images as geometric regions of different shapes, visual aspects, and sizes. To integrate information representing spatial relationships between these objects, we refer to Qualitative Spatial Relationships (QSRs) [26]. Our approach involves extracting all spatial relationships that exist between every pair of regions within an image and integrating this information into the training process of a deep neural network as extra knowledge. This integration of complementary relations is expected to enhance the model ability to better understand the spatial structure of the urban environment objects and improve the accuracy of panoptic segmentation prediction results. Specifically, we are interested in Region Connection Calculus (RCC) [27], which is a standardized set of spatial relations that are used to capture the possible connections and arrangements between regions, allowing for a comprehensive representation of their spatial interactions. There are many versions of these Region Connection Calculus such as RCC-5 and RCC-8. In our case, we considered the RCC-8 which describes 8 fundamental relations. It offers a fine level of detail that enables precise representation of relationships between two regions in an image. Consequently, it enables a more comprehensive spatial understanding of the environment.

RCC-8 specifically defines eight distinct relationships. Let U denotes the set of non-empty regular closed sets, also known as regions. Within the RCC-8 algebra, there are 8 topological relations that serve as its foundation. The Disconnected (DC) relationship (FIGURE 1) signifies that two regions have no shared points or boundaries. The Externally Connected (EC) relationship (FIGURE 4) denotes one region surrounding or enclosing another. The Tangential Proper Part (TPP) relationship (FIGURE 7) implies that one region is entirely contained within another, with at least one shared boundary point. On the other hand, the Non-Tangential Proper Part ($NTPP$) relationship (FIGURE 8) indicates complete containment without shared boundaries. The Partially Overlapping (PO) relationship (FIGURE 3) suggests that the regions have some common points or boundaries, without one region entirely encompassing the other. When both regions are identical in shape and size, they are considered Equal (EQ) (FIGURE 2). Finally, the Tangential Proper Part Inverse ($TPPi$) (FIGURE 5) and Non-Tangential Proper Part Inverse ($NTPPi$) (FIGURE 6) relationships mirror their respective counterparts but with the roles of the regions reversed.

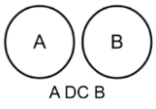

Figure 1 Disconnected

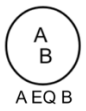

Figure 2 Equal

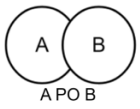

Figure 3 Partially Overlapping

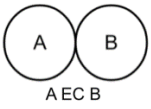

Figure 4 Externally connected

Figure 5 Tangential proper part inverse

Figure 6 Non Tangential proper part inverse

Figure 7 Tangential proper part

Figure 8 Non-Tangential proper part

The eight relations we have presented provide a comprehensive and detailed representation of spatial relationships between objects in the urban environment. These relations serve as a formal logic that captures essential spatial knowledge about the components within the environment. By combining this knowledge with the performances of a deep neural network, we can create an informed deep learning framework to enhance the network understanding and reasoning abilities.

In the next section, we describe the methodology to extract the RCC-8 relations and integrate them into a deep neural network.

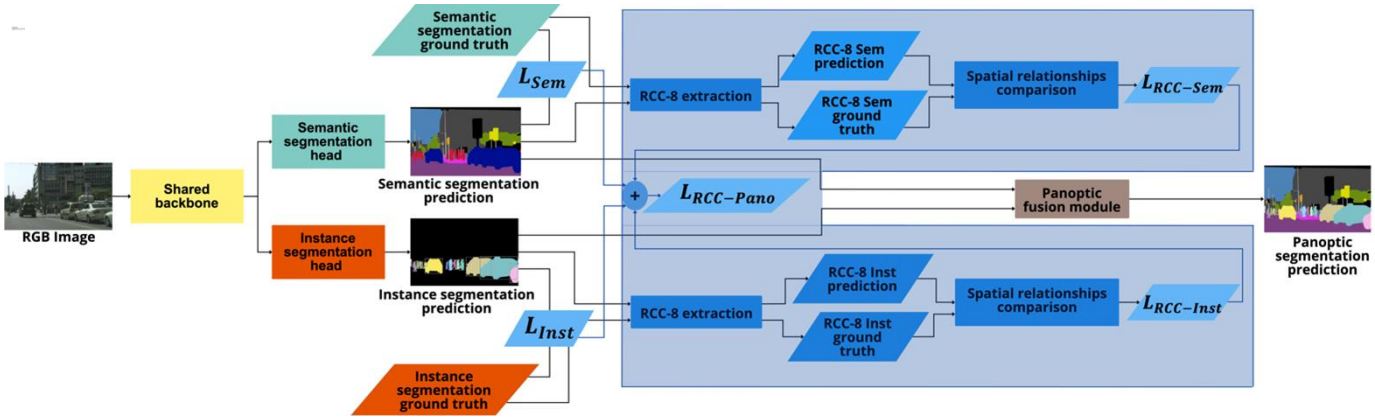

Figure 9 The proposed architecture for the integration of spatial relationships into a two-head panoptic segmentation deep neural network. The blue module is our contribution.

IV. SPATIAL RELATIONSHIPS INTEGRATION FOR PANOPTIC SEGMENTATION

This section presents the proposed deep neural network architecture that integrates RCC-8 relations between objects perceived in images. It is important to mention that the proposed approach is general and can be applied to any two-head (one for semantic segmentation and the other for instance segmentation) panoptic segmentation model. As mentioned previously, the main idea of the proposed technique is to optimize and enhance the performance of panoptic segmentation models by incorporating additional knowledge on the spatial relationships between different objects in an urban scene directly during the model training. We aim to integrate this knowledge by introducing a novel loss function that captures and represents the spatial relationships between objects. By incorporating this loss function into the training process, the model gains a comprehensive understanding of the urban environment, improving its ability to accurately segment objects by considering their contextual interactions. To extract the RCC-8 relations between the various object types of the image, including both ``Stuff'' and ``Things'', we integrated the proposed module in both heads during the training of the deep neural network (FIGURE 9). This module is designed to extract the RCC-8 relations between regions to define and compute the proposed $L_{RCC-Pano}$ loss function. To do so, distinct image regions should be separated, and then the different regions should be approximated before extracting the RCC-8 relations.

Separation of distinct regions: The proposed module takes as input the "Stuff" regions from the predicted semantic segmentation map and those from the ground truth (FIGURE 10). In the semantic map, "Stuff" regions belonging to the same class are labeled with a common label, even though they are not connected to each other. For example, in FIGURE 10, the two separate regions belonging to the class "Vegetation" were both labeled with the same label (V), despite being distinct and not connected. However, it is important in our case to consider each region independently of the others to accurately represent and integrate the spatial relationships between all the distinct regions in the scene. To solve this problem, we implemented an algorithm that separates all the distinct visible "Stuff" regions from the semantic maps. We also added some identifiers to reference the distinct regions belonging to the same label in both the prediction and the ground truth (FIGURE 10: Separation of distinct regions). Since the concept of instance segmentation itself involves identifying and separating individual objects within an image, we did not face the problem of identifying distinct regions regarding the "Things" regions related to the instance segmentation branch. Thus, each region belonging to an instance is basically segmented separately from the other instances of the same class. At the end of this step, we consider a set of distinct regions for each of the predicted maps (semantic and instance segmentation), along with their respective ground truths regions.

Region approximation: To identify the spatial relationships between regions, we initially extract the primary features and characteristics of each region. Specifically, the centroid coordinates and their principal and secondary axes are computed, which are used to generate a polygon approximation with a maximum of 50 vertices for each region (FIGURE 10: Region approximation). The polygons are used to establish the spatial relationships between each pair of regions.

RCC-8 extraction: The computed regions properties are used to extract the RCC-8 relations (FIGURE 10: RCC-8 extraction). The goal is to introduce a new penalty term to the global loss function of the panoptic segmentation deep neural network by comparing the 8 RCC spatial relations in the semantic and instance segmentation prediction maps with their corresponding ground truths. To incorporate these comparative elements into the network training, we propose the addition of two new penalty terms to the loss function, namely L_{RCC-S} and L_{RCC-I} which respectively correspond to the semantic and instance segmentation heads (FIGURE 9). These penalty terms aim to penalize the network errors made among the 8 RCC relations between the image regions during training.

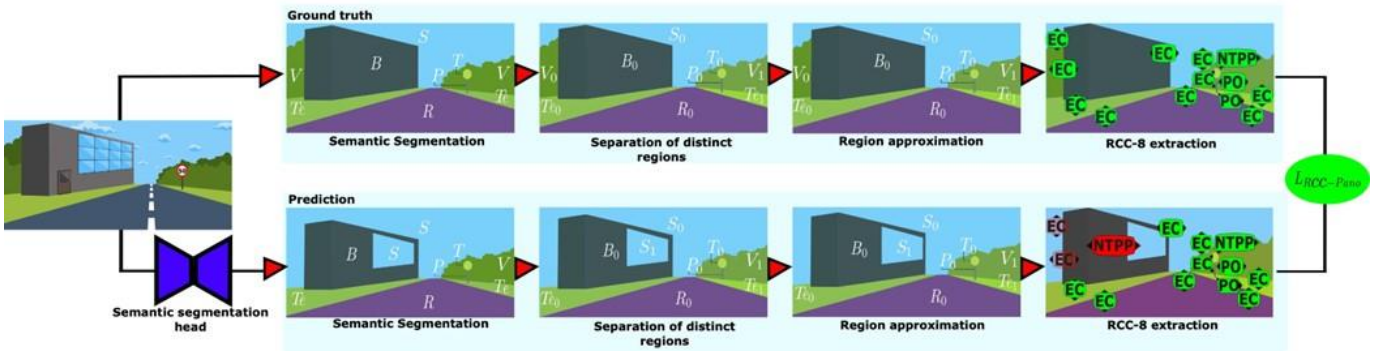

Figure 10 Methodology to extract the 8 RCC relationships between ``Stuff'' regions. The upper block presents the process considering the semantic segmentation ground truth, and the bottom block represents the process for prediction. In the final step, the green relationships indicate correct matches between the ground truth and prediction, the red ones represent false positives, and the red transparent ones represent false negatives.

Mathematically, L_{RCC-S} and L_{RCC-I} represent the average of the 8 penalty terms of the 8 RCC relations:

$L_{RCC-S} = \frac{1}{8}(L_{PO-S} + L_{EO-S} + L_{TPP-S} + L_{NTPP-S} + L_{DC-S} + L_{EQ-S} + L_{TPPi-S} + L_{NTPPi-S})$.

For the instance segmentation, $L_{RCC-I} = \frac{1}{8}(L_{PO-I} + L_{EO-I} + L_{TPP-I} + L_{NTPP-I} + L_{DC-I} + L_{EQ-I} + L_{TPPi-I} + L_{NTPPi-I})$.

L_{RCC-S} and L_{RCC-I} range between 0 and 1 and represent the ability of the neural network to verify the 8 RCC relationships between objects in images. The penalty terms corresponding to the 8 RCC relations are defined as the ratio between the errors made by the model in the corresponding RCC relation and the sum of the wrong and the correct matches of the same relation with the ground truth. For example, if we consider the RCC relation ``PO'' (Partially Overlapping), the penalty term is defined as follow:

$$L_{PO} = \frac{Errors_{PO}}{Errors_{PO} + Correct_{PO}}$$

To provide a clear illustration, consider the example provided in FIGURE 10. We have an image with its corresponding semantic segmentation ground truth, which contains the following pairwise object ``EC'' relations: $(B_0, V_0)$ and $(V_0, S_0)$. On the other hand, the semantic segmentation map prediction of the same image does not include these relations and instead it contains the pairwise object ``NTTP'' relation: (S_1, B_0). From the comparison, we can identify two types of errors made by the model. The first error is the presence of the ``NTTP'' relation for the pairwise (S_1, B_0), which does not exist in the ground truth. This can be considered as a false positive since the model incorrectly identified a relationship between the region S_1 and the region B_0. The second error is the failure to detect the (B_0, V_0) and (V_0, S_0) relations, where the model did not recognize the ``EC'' connections between each pair of regions. These errors can be seen as a false negative since the model missed a true relation that should have been identified. Following the same methodology, all penalty terms for the 8 RCC relations are computed.

In general, deep learning models for panoptic segmentation that follow an architecture with two heads-one for semantic segmentation and the other for instance segmentation- typically employ a global loss function. The global loss function for these models is commonly defined as the sum of two individual loss functions: L_{Sem}, which optimizes the semantic segmentation head, and L_{Inst}, which optimizes the instance segmentation branch (FIGURE 9). Therefore, the general form of the loss function for such models can be expressed as: $L_{Pano} = L_{Sem} + L_{Inst}$. Using the proposed penalty terms, the new global loss function for optimizing the whole network while considering the integration of the spatial relationships knowledge between the objects is defined as follows : $L_{RCC-Pano} = L_{Sem} + L_{Inst} + L_{RCC-S} + L_{RCC-I}$.

V. EXPERIMENTS AND RESULTS

To validate, evaluate and demonstrate the performance of integrating spatial relationships knowledge into a deep neural network for panoptic segmentation, we consider a state-of-the-art panoptic segmentation network (EfficientPS [14]) as our base network. EfficientPS is a robust model that demonstrates exceptional performance in panoptic segmentation compared to other state-of-the-art approaches. It is also highly extensible, making it suitable for making modifications and adding modules to implement the proposed approach.

5.1. Architecture of the EfficientPS model

The EfficientPS architecture [14] includes a shared backbone with a 2-way FPN. The shared backbone is based on the EfficientNet architecture [28], which uses mobile inverted bottleneck units [29] and compound scaling to enhance its representational capacity with fewer parameters compared to other similar networks. EfficientPS incorporates a 2-way FPN that effectively fuses multi-scale features in both directions. This is achieved by spreading information flow in multiple

directions. After the 2-way FPN, two heads work in parallel: the semantic segmentation head and the instance segmentation head. To produce the panoptic segmentation output, EfficientPS employs a panoptic fusion module that combines the outputs from the semantic and instance heads. This module integrates the predictions from both heads to yield the final panoptic segmentation result.

5.2. Implementation details

Regarding the implementation of the algorithm for the extraction of the RCC-8 spatial relationships between objects (Section 3), we used the Measure Region Properties module of the Scikit-image library. Additionally, we considered the QSRLIB Library [30] to infer the RCC-8 spatial relationships. The official implementation code is available online. The EfficientPS model is implemented using Pytorch 1.7 Neural Network Libraries with CUDA GPU Toolkit 11.2. The hyper parameters set by the authors have remained unchanged. However, on the EfficientPS paper, the training was performed on 16 NVIDIA Titan X 12GB GPUs. The batch size was set to 1 and the number of epochs to 160. Due to our less powerful GPU resources available (2 NVIDIA GeForce RTX 2080 Ti 11GB GPUs), we were unable to train the model under the same conditions. To address this technical challenge, we chose to use the ``EfficientNet-b4" as the shared backbone instead of the ``EfficientNet-b5" used in [14]. Indeed, the b4 version is lighter than the b5 version, allowing us to train the model based on available computational resources.

5.3. Dataset and evaluation metrics

We use the standard Panoptic Quality metrics of the state of the art [16] to evaluate the performance of the proposed approach. These metrics are presented below. The Panoptic Quality (PQ) metric quantifies the accuracy of object instance segmentation as well as the correct prediction of ``Stuff" class. It is calculated as follows: $PQ = \frac{\sum_{(p,g)\in TP}(IOU(p,g))}{|TP|+\frac{1}{2}|FP|+\frac{1}{2}|FN|}$, where $\sum_{(p,g)\in TP}$ represents the sum over all pairs of prediction and ground truth objects that belong to the set TP, which represents the True Positives. FP, and FN, respectively, represent False Positives and False Negatives. IOU denotes the Intersection Over Union (IOU) ratio.

The Segmentation Quality (SQ) metric indicates the accuracy of the predicted segments in comparison to the ground-truth. It is calculated by averaging the IOU scores of all the TP segments. The SQ metric is defined as: $SQ = \frac{\sum_{(p,g)\in TP}(IOU(p,g))}{|TP|}$.

To consider the impact of incorrect predictions, the Recognition Quality (RQ) is introduced as a metric that combines precision and recall. The RQ metric is defined as: $RQ = \frac{|TP|}{|TP|+\frac{1}{2}|FP|+\frac{1}{2}|FN|}$.

Following the standard benchmarking criteria for panoptic segmentation, we calculate PQ, SQ and RQ for all the dataset classes, and report them separately for ``Stuff" classes ($PQ_{st}, SQ_{st}\ and\ RQ_{st}$) and ``Things" classes ($PQ_{th}, SQ_{th}\ and\ RQ_{th}$).

To evaluate the effectiveness of the proposed approach, we considered the CityScapes Dataset [15] that consists of diverse urban street scenes from more than 50 European cities, captured under different conditions. Recently, the CityScapes dataset introduced a benchmark for panoptic segmentation, with pixel-level annotations for 19 object classes, including 11 ``Stuff" classes and 8 ``Things" classes. The dataset contains 5000 finely annotated images captured using a stereo camera with a resolution of $2048x1024$ pixels. These images are divided into 2975 images for training, 500 images for validation, and 1525 images for testing.

5.4. Evaluation on CityScapes dataset

In this section, we present a comparative analysis of the proposed approach against current state-of-the-art panoptic segmentation methods. We evaluate our technique on Cityscapes dataset [15] and report the performance metrics in Table as mentioned in the corresponding papers of the state-of-the-art methods.

The baseline approach ``EfficientPS-b4" yields a PQ of 60.6, a SQ of 80.3 and a RQ of 74.3, with a PQ_{th} and a PQ_{st} of 56.3 and 63.8 respectively. However, the proposed approach, which incorporated additional knowledge about spatial relationships between objects in the loss function during model training, achieved higher scores. Specifically, it provided a PQ of 64.2, a SQ of 81.6 and a RQ of 77.5. The PQ_{th} and the PQ_{st} also respectively improved to 59.8 and 67.6. Furthermore, in comparison with prior state-of-the-art works, the proposed approach demonstrates superior performances regarding the panoptic evaluation metrics. These results highlight the effectiveness of integrating spatial relationships into the panoptic segmentation neural network. The improved PQ, SQ, and RQ scores signify that the proposed approach outperforms the baseline in terms of overall panoptic, segmentation, and recognition quality. More specifically, the improved RQ score indicates an enhanced recognition quality, suggesting that the proposed approach is better at accurately identifying and classifying objects in the scene. This means that the model developed a higher ability to recognize and assign correct labels to instances and semantic classes within the image thanks to the integrated RCC knowledge.

Method	PQ	SQ	RQ	PQ_{th}	SQ_{th}	RQ_{th}	PQ_{st}	SQ_{st}	RQ_{st}
PanopticFPN [16]	58.1	--	--	52.0	--	--	62.5	--	--
AUNet [31]	59.0	--	--	54.8	--	--	62.1	--	--
UPSNet [32]	59.3	79.7	73.0	54.6	79.3	68.7	62.7	80.1	76.2
Seamless [33]	60.3	--	--	56.1	--	--	63.1	--	--
SSAP [34]	61.1	--	--	55.0	--	--	--	--	--
Panoptic-DeepLab [21]	63.0	--	--	--	--	--	--	--	--
EvPSNet [35]	63.7	81.3	77.5	--	--	--	--	--	--
EfficientPS-b5 [14]	63.9	81.5	77.1	60.7	81.2	74.1	66.2	81.1	76.7
EfficientPS-b4	60.6	80.3	74.3	56.3	79.2	70.9	63.8	81.1	76.7
EfficientPS-b4-RCC	64.2	81.6	77.5	59.8	80.3	73.8	67.6	82.4	80.2

Table 1 Comparison of panoptic segmentation performance on the CityScapes validation set. (st) and (th), respectively, denote "Stuff" and "Things" classes. "–"indicates unreported metric for the corresponding method

Similarly, the higher SQ score indicates improved segmentation quality. This suggests that the proposed approach achieves more precise and accurate object boundaries, resulting in a better overall representation of the scene. To conclude, the integration of spatial relationships in the loss function during the training of EfficientPS model likely facilitated the model ability to capture contextual information, mainly the spatial layout of scene objects, which enhanced its panoptic segmentation accuracy. This additional knowledge enabled the model to better understand and use the spatial context of objects in the image, resulting in improved performance in terms of PQ, SQ and RQ metrics. In addition to the global PQ metric that has been increased thanks to our approach, the RQ and SQ metrics were also improved. This means that incorporating the 8 RCC relationships into the model loss function has also increased the model ability to accurately recognize and distinguish between instances of different objects, leading to higher RQ scores. Furthermore, the models ability to precisely segment objects have significantly enhanced as indicated by the SQ metric.

VI. Conclusion

In conclusion, we propose a new informed deep learning approach as part of hybrid AI, aiming to enhance the performance of deep neural networks for panoptic segmentation. By integrating prior knowledge into the deep learning networks, specifically focusing on spatial relationships between objects, our approach offers significant improvements. The integration of this additional knowledge allows the models to gain a deeper understanding of the scene beyond the visual cues present in the images. This integration enhances the models performance and accuracy by enabling them to capture complex object relationships, resolve ambiguities, and overcome panoptic segmentation challenges. Our approach offers several contributions, including the introduction of a new training methodology, the development of a new loss function, and the validation and evaluation of the proposed approach on urban scene dataset. The results of our experiments and evaluations consistently show that the proposed approach outperforms the state of the art and achieves better results regarding Panoptic Quality metrics. By incorporating meaningful knowledge during the training process, the proposed approach enables the model to better understand the context of the target environment. This leads to better performances and accurate decision-making. The significance of integrating additional knowledge is not limited to panoptic segmentation alone, it extends to other computer vision tasks where understanding context is important. As part of our future work, we aim to enhance the panoptic segmentation results by introducing a local loss function that specifically targets problematic regions. The goal is to provide the network with more precise and explicit knowledge transfer. Additionally, we aim to integrate other type of knowledge, beyond RCC-8, to further enhance the panoptic segmentation.

Bibliography

[1] C. Y. Z. Z. Z. L. Z. J. H. Zhu, «Fusing panoptic segmentation and geometry information for robust visual slam in dynamic environments,» *IEEE 18th International Conference on Automation Science and Engineering*, p. 1648–1653, 2022.

[2] J. B. C. M. C. S. A. Milioto, «Lidar panoptic segmentation for autonomous driving,» *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* , p. 8505–8512, 2020.

[3] M. S. B. R. M. M. C. B. O. Zendel, «Unifying panoptic segmentation for autonomous driving,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 21351–21360, 2022.

[4] K. e. al., «Panoptic feature pyramid networks,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 6399–6408, 2019.

[5] C. P. C. Y. J. W. X. L. G. Y. W. J. H. Liu, «An end-to-end network for panoptic segmentation,» *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 6172–6181, 2019.

[6] X. C. X. L. N.-G. C. S.-W. L. S. F. R. U. A. Y. R. Mottaghi, «The role of context for object detection and semantic segmentation in the wild,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[7] D. H. J. H. H. A. A. E. M. H. S. K. Divvala, «An empirical study of context in object detection,» *IEEE Conference on computer vision and Pattern Recognition*, p. 1271–1278, 2009.

[8] R. D. D. P.F.Felzenszwalb, «Object detectionwith discriminatively trained part-based models,» *IEEE Transactions on pattern analysis and machine intelligence*, p. 1627–1645, 2009.

[9] J. G. Z. Z. J. D. Y. W. H. Hu, «Relation networks for object detection,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3588–3597, 2018.

[10] K. V. A. Z. A. Simonyan, «Deep inside convolutional networks: visualising image classification models and saliency maps,» *Proceedings of the International Conference on Learning Representations (ICLR), ICLR.*, 2014.

[11] G. B. Z. K. L. V. C. Huang, «Local receptive fields based extreme learning machine.,» *IEEE Computational intelligence magazine 10*, p. 18–29, 2015.

[12] L. Medsker, «Hybrid intelligent systems,» *Springer Science & Business Media*, 2012.

[13] G. K. I. L. L. P. P. W. S. Y. L. Karniadakis, «Physics-informed machine learning,» *Nature Reviews Physics 3*, p. 422– 440, 2021.

[14] R. V. A. Mohan, «Efficientps: Efficient panoptic segmentation.,» *International Journal of Computer Vision*, p. 1551–1579, 2021.

[15] M. O. M. R. S. R. T. E. M. B. R. F. U. R. S. S. B. Cordts, «The cityscapes dataset for semantic urban scene understanding,» *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[16] A. H. K. G. R. R. C. D.́. P. Kirillov, «Panoptic segmentation,» *Proceedings of the IEEE/CVF Conference on Computer Vi- sion and Pattern Recognition* , p. 9404–9413., 2019.

[17] C. C. S. H. P. F. L. Chang, «Epsnet: efficient panoptic segmentation network with cross-layer attention fusion,» *Proceedings of the Asian Conference on Computer Vision*, 2020.

[18] T. D.́. P. G. R. H. K. H. B. B. S. Lin, «Feature pyramid networks for object detection,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2117–2125, 2017.

[19] C. I. S. V. V. A. A. Szegedy, «Inception-v4, inception- resnet and the impact of residual connections on learning,» *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[20] S. P. L. K. P. Bulo, «In-place activated batchnorm for memory-optimized training of dnns,» *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 5639–5647, 2018.

[21] B. C. M. Z. Y. L. T. H. T. A. H. C. L. Cheng, «Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation,» *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[22] Y. Z. G. X. H. L. X. L. L. Wu, «Auto-panoptic: Cooperative multi-component architecture search for panoptic segmentation,» *Advances in Neural Information Processing Systems 33*, p. 20508–20519, 2020.

[23] X. Z. X. L. M. S. J. Zhang, «Shufflenet: An extremely efficient convolutional neural network for mobile devices,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 6848–6856, 2018.

[24] Y. L. G. L. S. B. O. W. Y. W. F. F. J. X. M. L. X. Chen, «Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation,» *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[25] D. M. P. D. G. De Geus, «Single network panoptic segmentation for street scene understanding,» *IEEE Intelligent Vehicles Symposium (IV)*, p. 709–715, 2019.

[26] A. B. B. G. J. G. N. Cohn, «Qualitative spatial representation and reasoning with the region connection calculus.,» *geoinfor- matica 1,*, p. 275–316, 1997.

[27] D. C. A. Randell, «Modelling topological and metrical properties in physical processes,» p. 357–368, 1989.

[28] M. L. Q. Tan, «Efficientnet: Rethinking model scaling for convolutional neural networks,» *International conference on machine learning*, p. 6105–6114, 2019.

[29] S. G. R. D.́. P. T. Z. H. K. Xie, «Aggregated residual transformations for deep neural networks,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 1492–1500, 2017.

[30] Y. A. M. B. C. D. C. D. P. L.-. b. P. H. M. H. N. H. D. C. A. e. a. Gatsoulis, «Qsrlib: a software library for online acquisition of qualitative spatial relations from video,» 2016.

[31] Y. C. X. Z. Z. X. L. H. G. D. D. W. X. Li, «Attention-guided unified network for panoptic segmentation,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , p. 7026–7035, 2019.

[32] Y. L. R. Z. H. H. R. B. M. Y. E. U. R. Xiong, «Upsnet: A unified panoptic segmentation network,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8818–882, 2019.

[33] L. B. S. C. A. K. P. Porzi, «Seamless scene segmentation,» *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8277–8286, 2019.

[34] N. S. Y. W. Y. Z. X. Y. Y. Y. M. H. K. Gao, «Ssap: Single-shot instance segmentation with affinity pyramid,» *Proceedings of the IEEE/CVF international conference on computer vision*, p. 642–651, 2019.

[35] K. M. S. B.̈. D. B. W. Sirohi, «Uncertainty-aware panoptic segmentation,» *IEEE Robotics and Automation Letters* , 2023.

[36] J. L. K. S. K. T. Z. Lazarow, «Learning instance occlusion for panoptic segmentation,» *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 10720–10729., 2020.

[37] H. Z. Y. G. B. A. H. Y. A. C. L. Wang, «Axial-deeplab: Stand-alone axial-attention for panoptic segmentation,» *Computer Vision–ECCV: 16th European Conference*, 2020.