Intégration de connaissances sur les relations spatiales dans la fonction de perte d'un réseau de neurones pour la segmentation panoptique

Fatima Ezzahra Benkirane, Nathan Crombez Nathan Crombez, Vincent Hilaire, Yassine Ruichek

Intégration de connaissances sur les relations spatiales dans la fonction de perte d'un réseau de neurones pour la segmentation panoptique

F.E. Benkirane[1] N. Crombez[1] V. Hilaire [1] Y.Ruichek[1]

[1] UTBM, CIAD, F-90010 Belfort, France

{fatima.benkirane, nathan.crombez, yassine.ruichek, vincent.hilaire}@utbm.fr

Résumé

La segmentation panoptique est une tâche de vision par ordinateur qui vise à fournir une compréhension globale de la scène en identifiant chaque élément quantifiable ou non dans une image. Une segmentation précise et complète est essentielle pour une variété d'applications telles que la navigation autonome des véhicules ou encore la surveillance de la circulation routière. Dans cet article, nous proposons une nouvelle méthode pour l'estimation de la panoptique. L'approche consiste en l'intégration des connaissances sur les relations spatiales entre les objets perçus d'une scène dans un réseau de neurones pendant son entraînement. Cette intégration est effectuée en proposant une nouvelle fonction de perte qui permet d'incorporer les connaissances dans le processus d'apprentissage. La validation et l'évaluation de notre approche sur le jeu de données CityScapes pour les scénarios urbains a montré des résultats prometteurs et encourageants.

Mots Clef

Segmentation panoptique, relations spatiales, réseaux de neurones.

Abstract

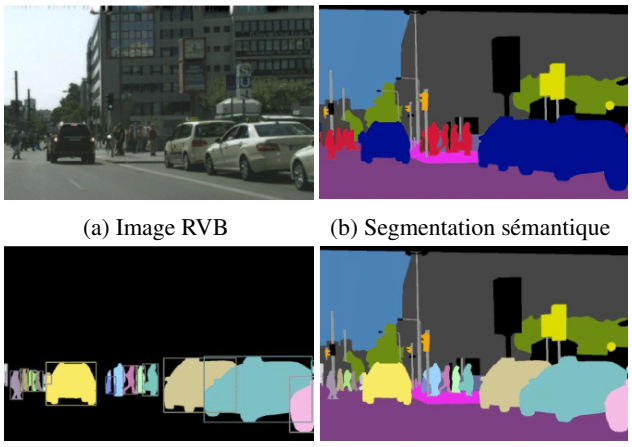
Panoptic segmentation is a computer vision task that consists in identifying all objects within an image to provide a holistic understanding of a given scene. Accurate and complete segmentation is crucial for a wide range of applications, including autonomous vehicle navigation and traffic tracking. In this paper, we propose a novel approach for panoptic segmentation that integrates knwoledge on spatial relationships between objects into a neural network during its training process. This is acheived through the introduction of a novel loss function that allows the knwoledge transfer into the learning process. Our approach is evaluated and validated on the CityScapes dataset for urban scenarios, demonstrating promising and encouraging results.

Keywords

Panoptic segmentation, spatial relationships, neural networks.

1 Introduction

La segmentation panoptique (Figure 1d) est une décomposition complète et cohérente de tous les éléments d'une scène perçus dans une image (Figure 1a). Elle combine les informations provenant de l'estimation de la sémantique (Figure 1b) et des instances (Figure 1c). La sémantique permet de décomposer une image en régions correspondantes à des classes d'objets non quantifiables (*Stuff*), tandis que les instances sont des identifications d'objets individuels et quantifiables dans l'image (*Things*).



(a) Image RVB (b) Segmentation sémantique

(c) Segmentation d'instances (d) Segmentation panoptique

FIGURE 1 – La segmentation panoptique d'une image peut être considérée comme une combinaison de la sémantique et des instances d'objets perçues.

La capacité des approches d'estimation de la panoptique à décrire et analyser les images permet de fournir des solutions pratiques à diverses applications. Nous citons à titre d'exemple la vidéo-surveillance, où la segmentation panoptique peut intervenir afin de détecter les objets en mou-

vement et suivre leur trajectoire [1]. Elle contribue également dans le domaine de la conduite autonome pour aider le véhicule à comprendre son environnement afin de prendre les décisions nécessaires en conséquence [2, 3]. La sémantique et les instances sont complémentaires et permettent de mieux comprendre et analyser les images d'une manière globale. L'estimation de la sémantique est caractérisée par sa capacité à mieux reconnaître les objets et à comprendre les significations des régions dans l'image. L'estimation des instances, quant à elle, est plutôt performante en termes d'identification des objets individuels et de suivi de leur évolution au fil du temps. Malgré leurs performances et leurs apports significatifs dans le traitement d'images et l'analyse de l'environnement, ces deux problématiques restent confrontées à des défis importants. Tout d'abord, l'apparence des objets peut compliquer la tâche de reconnaissance en raison des variations de l'environnement au niveau de l'éclairage, des occultations, de la perspective, etc. De plus, la complexité des scènes peut rendre difficile la segmentation précise de ses composantes, qui sont généralement de différentes tailles et de formes variées. En outre, il peut être difficile de distinguer les objets d'un arrière-plan complexe et peu distinct. Finalement, la distinction entre les éléments appartenant à la même classe peut s'avérer délicate, notamment lorsqu'ils se touchent ou se chevauchent.

La communauté scientifique s'est intéressée à partir de 2018 à l'estimation de la segmentation panoptique [4]. Les approches proposées, qui peuvent être vues comme collaboratives, permettent de bénéficier simultanément des avantages des méthodes de la sémantique et de de celles des instances. La plupart des approches de segmentation panoptique sont appliquée à des données de type image, avec des méthodologies variées, dont certaines utilisent deux sous-réseaux de neurones distincts pour l'estimation de la sémantique et des instances [4]. Cependant, ces approches sont complexes et nécessitent un post-traitement important pour fusionner les prédictions associées, ce qui limite leur efficacité [15]. Un autre type d'approches qui comprend un *Backbone* partagé a alors été proposée [15]. Ce type d'approches améliore le processus d'entraînement en permettant un échange de caractéristiques entre les modules d'estimation de la sémantique et des instances grâce à un *Backbone* commun, un encodeur partagé. De plus, elles sont moins complexes que les premières approches en termes de traitement nécessaire à la fusion et la génération de la prédiction panoptique [15]. L'utilisation de ce type d'approches a démontré que le partage réciproque de caractéristiques entre les différentes parties d'un réseau lors de son entraînement améliore considérablement les résultats de la segmentation panoptique.

Dans ce travail, nous proposons d'enrichir l'entraînement de réseaux de neurones en intégrant des connaissances significatives pour la panoptique. En effet, il a été démontré que l'apport d'informations et de connaissances supplémentaires à un réseau permet de guider efficacement son apprentissage et d'accroître ses performances [24]. Nous avons identifié que les problèmes et limites inhérents à la prédiction de la panoptique, notamment dans un contexte urbain, sont principalement liés aux relations spatiales complexes entre les régions au sein d'une image. À titre d'exemple, le chevauchement et les occultations entre les différents objets empêchent d'avoir une segmentation fine et précise. Par conséquent, nous supposons que l'intégration de connaissances sur les relations spatiales entre les différents objets peut apporter une valeur significative aux réseaux de neurones et de les aider à mieux orienter leurs entraînements. Les contributions dans le cadre de ce travail sont les suivantes :

- la proposition d'une approche améliorée d'entraînement des réseaux de neurones artificiels pour la segmentation panoptique, qui intègre des informations sur les relations spatiales entre les objets perçus,
- la proposition d'une nouvelle fonction de perte pour l'optimisation de l'entraînement du réseau,
- la valiation et l'évaluation de l'efficacité de l'approche proposée sur un jeu de données de l'état de l'art, le dataset CityScapes [9].

La suite de l'article est organisée comme suit. Les travaux de l'état de l'art traitant de l'estimation de la segmentation panoptique sont introduits dans la section 2. La section 3 décrit l'approche proposée. Nous présentons les résultats des expériences effectuées dans la section 4. Finalement, la dernière section conclut l'article et donne une ouverture sur les travaux futurs.

2 Panoptique via *Backbone* partagé

Dans cette section, nous nous focalisons sur les techniques les plus performantes proposées dans le domaine de la segmentation panoptique. En particulier, nous nous intéressons aux approches utilisant des réseaux de neurones artificiels ayant un *Backbone* partagé, qui sont considérés comme les plus avancés pour ce type de segmentation [5, 6]. Nous présentons les différentes méthodes proposées, ainsi que leurs avantages et inconvénients.

Les auteurs de [15] ont été les premiers à proposer une méthode dans le cadre de cette catégorie d'approches. Pour pouvoir améliorer l'utilisation des différentes supervisions et réduire la consommation de ressources informatiques, l'algorithme introduit dans ce papier explore le partage de caractéristiques entre les différentes branches et met en évidence l'intérêt de partager autant de caractéristiques que possible. De plus, pour faire face au problème des occultations entre les instances, les auteurs ont introduit un module de classement spatial simple mais efficace. L'approche proposée est innovatrice car elle met l'accent sur le partage des caractéristiques entre les deux branches pour améliorer les performances du réseau de neurones.

Les auteurs de [5] ont proposé une approche qui repose sur une pyramide spatiale de convolutions dilatées (ESPnet). Une structure composée de FPN (Feature Pyramid

Network) et de ResNet (Residual Network) est partagée entre les têtes de segmentation sémantique et d'instances. En vue d'améliorer les caractéristiques d'entrée, un module de fusion d'atténuation de couches croisées (CLA) est introduit. Le but de ce module est d'agréger les cartes de caractéristiques multi-couches dans le réseau FPN. Les expériences menées sur le jeu de données COCO [7] montrent des performances prometteuses avec une vitesse d'inférence significativement rapide. Malgré ces résultats encourageants, l'approche proposée présente certains inconvénients, à savoir l'utilisation d'un algorithme de fusion heuristique pour générer la segmentation panoptique. Cet algorithme ne permet pas d'avoir des résultats performants, il est donc important d'explorer d'autres algorithmes de fusion pour améliorer les résultats.

Dans la même optique, la méthode intitulée EfficientPS a été proposée dans [6]. Cette méthode s'appuie sur une structure EfficientNet [16] comme schéma d'extraction de caractéristiques multi-échelles partagé avec les modules de segmentation sémantique et de segmentation des instances. Les résultats de chaque module sont fusionnés et combinés grâce à un module de fusion panoptique à paramètres libres ou non contraints. Ce module est basé sur une approche adaptative qui vise à résoudre avec efficacité le problème de chevauchement entre les prédictions de la tête de segmentation sémantique et celle des instances. Il génère ensuite deux masques pour chaque instance à partir des deux têtes. Ces deux masques sont combinés de manière adaptative en utilisant le produit de Hadamard afin d'obtenir les masques fusionnés et générer la prédiction panoptique finale. Les résultats de cette approche ont montré des performances très intéressantes et compétitives à celles de l'état de l'art. Son efficacité est principalement due à l'architecture puissante proposée dans le module EfficientNet qui est caractérisé par une utilisation efficiente des ressources de calcul, une architecture flexible et une extraction de caractéristiques multi-échelles très adaptés aux tâches complexes comme la segmentation d'images. Cette approche a particulièrement démontré des performances sur de nombreux ensembles de données, notamment les ensembles de données d'images acquises en extérieur représentant les scénarios urbains complexes tels que Cityscapes [9] et Mapillary Vistas [8].

D'un autre côté, l'approche intitulée Axial DeepLab proposée dans [10] a également montré des résultats à la pointe dans l'état de l'art. En effet, la particularité de ce modèle réside dans l'intégration d'une nouvelle méthode d'attention 2D qui vise à réduire la complexité de calcul et permet une attention dans une région plus large ou même globale. Plus concrètement, la proposition est une nouvelle couche sensible à la position qui peut être utilisée pour former des modèles d'attention axiale pour la classification d'images et la prédiction dense, comme entre autres, la segmentation panoptique.

Finalement, le modèle Panoptic DeepLab [11] fait également partie des modèles les plus performants de l'état de l'art. Ce modèle comprend un mécanisme qui permet d'accorder plus d'importance aux zones et aux régions de l'image qui sont considérées plus pertinentes pour la tâche de segmentation panoptique. En effet, c'est l'un des premiers modèles de l'état de l'art proposant ce type de stratégie pour la segmentation panoptique avec un *Backbone* partagé. Enfin, le modèle a été conçu pour être efficace en termes de temps d'inférence afin que la tâche puisse être effectuée en temps réel.

3 Intégration des relations spatiales dans un réseau de panoptique

Dans cette section, nous décrivons la méthodologie mise en place pour intégrer des informations représentant les relations spatiales entre les objets perçus dans l'environnement afin d'améliorer l'apprentissage d'un réseau de neurones artificiel et donc de la segmentation panoptique estimée par ce dernier. Dans un premier temps, nous présentons les informations spatiales incorporées, puis nous procédons à une analyse détaillée de la méthode d'intégration de ces informations dans le réseau de neurones.

3.1 Relations spatiales qualitatives

Les objets 3D d'une scène urbaine sont projetés dans une image 2D en régions géométriques de formes et de tailles différentes. Ainsi, pour mettre en œuvre notre concept d'intégration des informations représentant la spatialité entre les objets, nous faisons référence aux relations spatiales qualitatives (QSRs)[14]. Plus précisément, nous nous intéressons au calcul des connexions entre régions (RCC), c'est-à-dire l'ensemble de relations spatiales standardisées pour décrire les relations possibles entre les régions géométriques dans un environnement. Elles permettent de décrire avec précision les relations topologiques entre les régions telles que la contiguïté, l'inclusion, l'occultation, etc. Il existe plusieurs variantes de cet ensemble de relations. Nous pouvons citer les variantes RCC-5 et RCC-8 qui comportent respectivement 5 et 8 relations fondamentales qui peuvent exister entre deux régions (Figure 2).

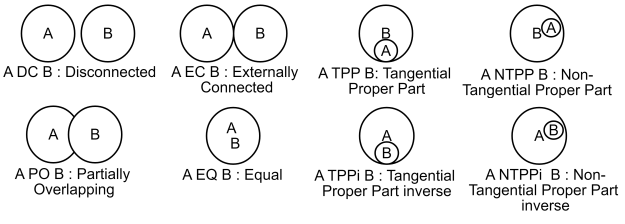
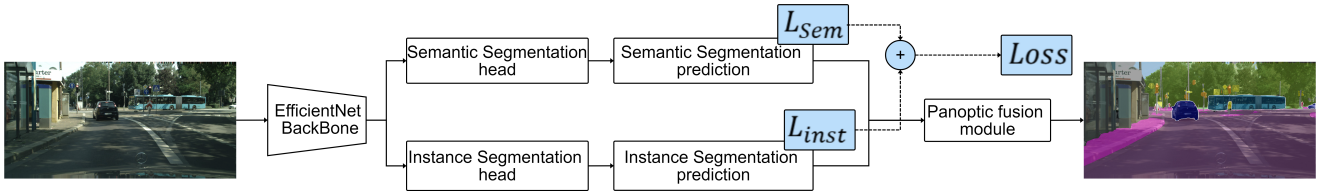
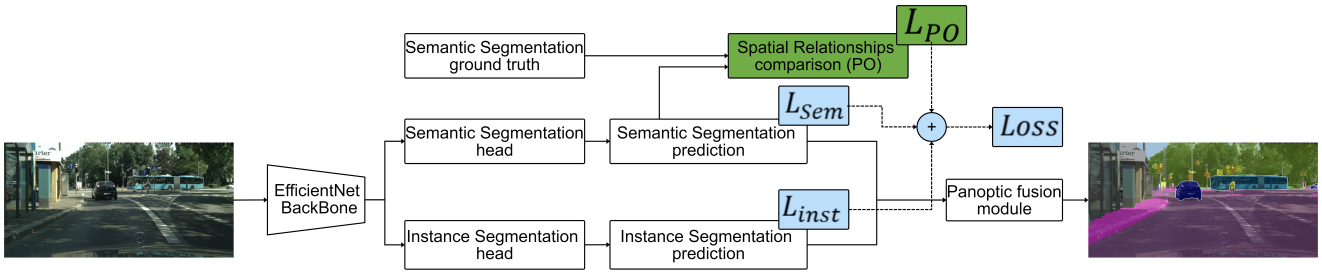


FIGURE 2 – Représentation et nomenclature des relations RCC-8

Pour modéliser les relations topologiques entre les régions d'un environnement, nous avons décidé d'intégrer les connaissances sur les RCC-8 dans notre réseau de neurones. Nous avons fait ce choix car les RCC-8 illustrent clairement les relations qui peuvent exister entre deux régions d'une image. À titre d'exemple, en considérant un



(a) Architecture générale du réseau EfficientPS [6]



(b) Architecture proposée pour l'intégration des informations spatiales dans le réseau de neurones artificiel

FIGURE 3 – Vue globale de l'architecture proposée : (a) représentation générale du réseau EfficientPS [6], (b) représentation de l'architecture proposée pour l'intégration des informations spatiales pendant l'entraînement du réseau.

contexte urbain, les régions représentant un bâtiment ont généralement une relation d'occultation partielle avec les régions de classe végétation. Dans une dimension plus fine, nous pouvons également dire que dans la plupart des cas, une plaque d'immatriculation se trouve à l'intérieur du châssis du véhicule, ce qui reflète une relation d'inclusion entre deux régions de ce type.

Parmi toutes les relations RCC-8 représentées dans la figure 2, on peut dire que la relation d'occultation partielle (*A PO B : Partially overlapping*) est l'une des relations les plus importantes qui nécessite une considération particulière. Ceci est principalement dû au fait que cette relation est fréquente en raison de la forme même des objets qui est généralement concave dans les images, favorisant ainsi la superposition partielle. Voici quelques exemples fréquents de scénarios représentant la relation d'occultation entre les objets dans une image de scène urbaine :

- un bâtiment peut occulter une partie du ciel, une partie d'un autre bâtiment ou une zone de végétation,
- une voiture peut occulter une partie d'un piéton ou d'une végétation,
- un lampadaire peut occulter une partie d'un bâtiment, etc.

En raison de tous ces éléments, nous avons choisi dans un premier temps d'intégrer uniquement la relation d'occultation partielle afin de valider le concept proposé.

3.2 Intégration des relations spatiales à un réseau de neurones (EfficientPS)

Nous présentons dans cette section l'architecture du réseau de neurones proposée qui intègre relation spatiale d'occultation partielle entre les objets perçus dans les images. Il est important de noter que le concept présenté est extensible et peut être appliqué sur diverses architectures de réseaux de neurones. Cependant, nous considérons dans notre cas un réseau de neurones performant de l'état de l'art pour évaluer concrètement l'apport de notre approche : EfficientPS [6].

De manière générale, l'objectif de notre approche est d'optimiser et d'améliorer les performances de la segmentation panoptique en incluant des informations supplémentaires sur la relation spatiale *PO* entre les régions, et donc entre les différents objets de l'image directement pendant l'entraînement du réseau de neurones. La figure 3a représente brièvement l'architecture globale du modèle EfficientPS [6]. Dans cet article, nous nous focalisons sur l'application de notre approche uniquement sur la tête de segmentation sémantique afin d'évaluer des premiers résultats et de construire notre preuve de concept. Par la suite, dans de futurs travaux, nous étendrons le procédé au second module, celui de la segmentation d'instances. La figure 3b illustre les niveaux spécifiques dans le réseau de neurones où notre approche est intégrée. Plus concrètement, nous proposons l'intégration d'un module intitulé "Comparaison des relations spatiales *PO*" dans l'entraînement du réseau qui comprend les différentes étapes décrites ci-dessous.

Détection des régions visibles Le module proposé prend en entrée les régions de type *Stuff* de la sémantique estimée et celles de la vérité terrain. Dans la sémantique, les régions *Stuff* appartenant à une même classe sont étiquetées avec un label commun, bien qu'elles ne sont pas connectées entre elles. Par exemple, si une scène contient deux régions de bâtiments séparées, elles sont toutes les deux étiquetées avec un label identique. Cependant, il est important de considérer chaque région visible de manière indépendante afin d'exploiter au mieux les relations spatiales

entre toutes les régions d'une scène. Ainsi, nous avons mis en place un traitement qui permet de détecter toutes les régions visibles distinctes. Nous avons également ajouté des identificateurs afin de référencer ces régions à la fois dans la prédiction et dans la vérité terrain. Un processus est donc nécessaire pour vérifier l'adéquation des identificateurs des régions dans la sémantique estimée avec ceux de leurs homologues de la vérité terrain.

Extraction des propriétés des régions Afin d'identifier les relations entre les régions, il est d'abord primordial de déterminer les caractéristiques principales de celles-ci. Nous procédons en déterminant les coordonnées du centroïde de chaque région et les dimensions des axes principaux et secondaires qui sont utilisés pour délimiter les polygones convexes à l'intérieur d'ellipses. Chaque polygone fournit une approximation des régions qui peut être utilisée pour établir les relations spatiales entre elles. Encore une fois, cette approche est appliquée à la fois à la sémantique estimée et à sa vérité terrain correspondante.

Extraction des relations *PO* entre les régions Nous exploitons les propriétés des régions visibles de la sémantique pour extraire les relations *PO*. La qualité de la prédiction est évaluée en comparant les relations détectées avec celles de la vérité terrain, ce qui permet de calculer des métriques de performance. Pour intégrer ces éléments comparatifs et analytiques dans l'entraînement du réseau, nous proposons d'ajouter le terme de pénalité L_{PO} dans la fonction de perte globale. Ce terme vise à sanctionner les erreurs commises par le réseau au niveau des relations spatiales *PO* lors de la prédiction de la sémantique. Mathématiquement, il représente le rapport entre les erreurs commises par le modèle au niveau des relations *PO* et la somme entre les erreurs et les bonnes correspondances de ces relations avec la vérité terrain :

$$L_{PO} = \frac{Erreurs_{PO}}{Erreurs_{PO} + Exactitude_{PO}}. \quad (1)$$

Par conséquent, la mesure de performance L_{PO} varie entre 0 et 1 et représente la capacité du réseau de neurones à identifier les relations PO entre les objets de l'image traitée et ainsi de pénaliser son rendement.

La fonction de perte globale, telle qu'elle est décrite dans le modèle orignal de EfficientPS [6], est la somme de la fonction de perte qui optimise la tête de l'estimation de la sémantique L_{Sem} avec celle qui optimise la tête de la prédiction des instances L_{Inst} :

$$Loss = L_{Sem} + L_{Inst}. \quad (2)$$

En intégrant le terme de pénalité proposé, la nouvelle fonction de perte globale pour optimiser le réseau en considérant la relation spatiale *PO* est définie ainsi :

$$Loss = L_{Sem} + L_{Inst} + L_{PO}. \quad (3)$$

4 Validation expérimentale

Nous décrivons d'abord les détails techniques de l'implémentation dans la section 4.1. Nous présentons ensuite dans la section 4.2 les métriques standards que nous adoptons pour l'évaluation de la segmentation panoptique estimée, et dans la section 4.3, le jeu de données considéré pour l'entraînement et l'évaluation de l'apprentissage. Enfin, nous présentons une comparaison quantitative et qualitative des résultats obtenus dans la section 4.4.

4.1 Détails techniques de l'implémentation

Nous avons développé un algorithme permettant d'extraire les relations spatiales RCC-8 entre les objets. Pour ce faire, nous exploitons le module *Measure Region Properties* de la librairie *scikit-image* [18] afin d'obtenir des informations géométriques détaillées sur les régions. À partir de ces informations, nous utilisons la bibliothèque QSRLIB [17] pour déterminer les relations spatiales RCC-8 entre les objets.

L'entraînement du réseau de neurones a été effectué sur 2 *GPUs NVIDIA GeForce RTX 2080 Ti 11GB* avec un *Batch size* de 1 et un nombre d'époques de 160. Nous avons maintenu les autres hyper-paramètres inchangés pour respecter au mieux les spécifications originales décrites dans le papier d'EfficientPS [6]. Cependant, dans le papier d'EfficientPS [6], l'entraînement a été effectué sur 16 *GPUs NVIDIA Titan X 12GB*. Étant donné que nous disposons d'une machine moins puissante en termes de *GPU*, il nous était impossible d'entraîner le modèle dans les mêmes conditions. Pour remédier à cette difficulté technique, nous avons choisi d'utiliser comme *Backbone* partagé le EfficientNet-b4 à la place du EfficientNet-b5 considéré dans [6]. En effet, la version *b4* est plus légère que la version *b5* ce qui nous a permis d'entraîner le modèle selon les ressources computationnelles dont nous disposions.

4.2 Métriques d'évaluation

Nous utilisons la métrique standard de qualité panoptique PQ [4] pour quantifier les performances du modèle proposé. La métrique PQ est calculée comme suit :

$$PQ = \frac{\sum_{(p,g)\in TP} (IOU(p,g))}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (4)$$

où TP, FP et FN sont respectivement les vrais positifs, les faux positifs et les faux négatifs. IOU est le rapport d'intersection sur union donné par :

$$IOU = \frac{TP}{TP + FP + FN}. \quad (5)$$

Nous rapportons également les métriques de qualité de segmentation SQ et de reconnaissance RQ :

$$SQ = \frac{\sum_{(p,g)\in TP} IOU(p,g)}{|TP|}, \quad (6)$$

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}. \quad (7)$$

TABLE 1 – Comparaison des performances de l'approche proposée avec les méthodes de l'état de l'art sur l'ensemble de données de validation de CityScapes. EfficientPS-b5 et EfficientPS-b4 font respectivement référence aux performances du modèle original EfficientPS avec le *Backbone-b5* et le *Backbone-b4* . Les résultats surlignés en gris (EfficientPS-b4-PO) représentent les performances de l'approche proposée.

Méthode	**PQ**	**SQ**	**RQ**	$\mathbf{PQ_{th}}$	$\mathbf{SQ_{th}}$	$\mathbf{RQ_{th}}$	$\mathbf{PQ_{st}}$	$\mathbf{SQ_{st}}$	$\mathbf{RQ_{st}}$
AUNet [19]	$59,0$	–	–	$54,8$	–	–	$62,1$	–	–
UPSNet [20]	$59,3$	$79,7$	$73,0$	$54,6$	$79,3$	$68,7$	$62,7$	$80,1$	$76,2$
Seamless [21]	$60,3$	–	–	$56,1$	–	–	$63,3$	–	–
SSAP [22]	$61,1$	–	–	$55,0$	–	–	–	–	–
AdaptIS [23]	$62,0$	–	–	$58,7$	–	–	$64,4$	–	–
Panoptic-DeepLab [11]	$63,0$	–	–	$58,7$	–	–	$64,4$	–	–
EfficientPS-b5 [6]	$63,9$	$81,5$	$77,1$	$60,7$	$81,2$	$74,1$	$66,2$	$81,8$	$79,2$
EfficientPS-b4	$60,6$	$80,3$	$74,3$	$56,3$	$\mathbf{79,2}$	$70,9$	$63,8$	$81,1$	$76,7$
EfficientPS-b4-PO	$\mathbf{61,3}$	$\mathbf{80,3}$	$\mathbf{75,1}$	$\mathbf{57,2}$	$79,1$	$\mathbf{72,2}$	$\mathbf{64,2}$	$\mathbf{81,3}$	$\mathbf{77,2}$

En suivant les critères de *benchmarking* standard pour la segmentation panoptique, nous calculons PQ, SQ et RQ sur toutes les classes dans le jeu de données et nous les rapportons également pour les classes *Stuff* (PQ_{st},SQ_{st} et RQ_{st}) et les classes *Things* (PQ_{th},SQ_{th} et RQ_{th}).

4.3 Jeu de données Cityscapes

Le jeu de données Cityscapes [9] est l'un des ensembles de données les plus difficiles pour la segmentation panoptique en raison de sa grande diversité. En effet, il couvre des scènes de plus de 50 villes européennes capturées dans des conditions temporelles et météorologiques différentes. Le jeu de données contient des annotations à un niveau pixelique pour 19 classes d'objets, dont 11 sont des classes *Stuff* et 8 sont des classes *Things* spécifiques aux instances. Il se compose de 5000 images finement annotées capturées à une résolution de 2048×1024 pixels. Au total, 2975 images sont dédiées à l'entraînement, 500 images pour la validation et 1525 images pour l'ensemble de test. Les annotations pour l'ensemble de test ne sont pas publiquement disponibles. Nous rapportons donc les performances du modèle proposé sur l'ensemble de validation conformément au protocole expérimental considéré dans l'état de l'art [6].

4.4 Validation et évaluation

Évaluation quantitative. Pour valider l'approche proposée, nous avons effectué des expériences sur l'ensemble de données Cityscapes. Les résultats de l'état d'art sur cet ensemble de données sont rapportés dans la première partie de la Table 1. Dans la seconde partie, les résultats de EfficientPS-b4 indiquent les performances du modèle EfficientPS sans intégrer de relation spatiale et en considérant le *Backbone* EfficientNetb4 afin de répondre à nos exigences en termes de ressources computationnelles (voir Section 4.1). Nous observons une différence de $3,3\%$ par rapport à la version originale EfficientPS-b5 [6]. Cette différence peut être principalement attribuée à l'utilisation d'un *Backbone* plus léger et moins performant ainsi qu'aux ressources matérielles différentes considérées lors de l'entraînement. Ainsi, les résultats de précision obtenus avec EfficientPS-b4 constituent notre base de référence pour évaluer l'effet de l'intégration de la relation spatiale *PO* dans la fonction de perte (Figure 3b).

Nous constatons une amélioration de $0,7\%$ de la métrique de base de la panoptique suite à l'intégration de la relation spatiale *PO* pendant le processus d'apprentissage. La métrique PQ est passée de $60,6\%$ avec EfficientPS-b4 à $61,3\%$, ce qui montre une amélioration de la qualité de la segmentation panoptique. En examinant de plus près les résultats, nous remarquons aussi une amélioration des métriques d'évaluation liées aux objets de classe *Stuff*. On note une évolution de $0,4\%$, $0,2\%$ et $0,5\%$ respectivement pour les métriques PQ_{st}, SQ_{st} et RQ_{st}, ce qui démontre l'impact positif sur la qualité de la segmentation des objets appartenant à ce type de classes. Cette amélioration est justifiée par le fait que, dans nos expériences, nous avons intégré les relations *PO* entre les objets de classe *Stuff*, ce qui permet d'améliorer la tête dédiée à l'estimation de la sémantique. Bien que nous ayons principalement mis l'accent sur les *Stuff*, il convient de souligner que le bloc d'entraînement dédié à l'estimation des instances a également été amélioré. Dans ce cadre, nous constatons une évolution de $0,9\%$ pour la métrique PQ_{th}. Autrement dit, la précision et la qualité de la panoptique pour les objets de classe *Things* est positivement impacté par les relations spatiales intégrées dans l'entraînement. En effet, la raison de cette amélioration au niveau des *Things* est l'architecture globale du réseau de neurones qui partage les caractéristiques des deux têtes (sémantique et instances) par le biais du *Backbone* commun. Ainsi, la fonction de perte proposée, qui intègre les connaissances sur la relation spatiale *PO* permet d'optimiser le bloc *Backbone* partagé et par consé-

(a) Images RVB (b) Panoptiques EfficientPS-b4 (c) Panoptiques EfficientPS-b4-PO

FIGURE 4 – Comparaison qualitative de l'approche proposée EfficientPS-b4-PO avec EfficientPS-b4

quent d'étendre l'apport de connaissances à l'ensemble du réseau.

Évaluation qualitative. Nous avons généré des visualisations de cartes de la segmentation panoptique pour évaluer davantage la qualité de l'approche proposée (Figure 4). Nous constatons sur le premier exemple que la voiture est partiellement masquée par une clôture. Le modèle EfficientPS-b4 détecte à peine les parties en contact externe avec la voiture sans considérer les pixels qui appartiennent à la classe clôture et se chevauchent avec la voiture. Notre approche (EfficientPS-b4-PO), quant à elle, a correctement détecté une partie importante qui occulte la voiture. Dans le deuxième exemple, EfficientPS-b4 n'a pas réussi à segmenter correctement le poteau qui occultait partiellement le bus, tandis que notre approche a nettement amélioré la segmentation au niveau des pixels concernés. Dans le dernier exemple, le modèle original EfficientPS-b4 a commis une erreur de segmentation au niveau de la partie de voiture masquée par le piéton. Le modèle a classé toutes les parties comme étant des pixels appartenant à la classe "piéton" et n'a pas pu identifier la voiture. En revanche, notre approche (EfficientPS-b4-PO) a correctement segmenté cette zone de la voiture, malgré sa position complexe par rapport à d'autres objets tels que le poteau et le piéton.

À partir des analyses quantitative et qualitative effectuées, nous pouvons conclure que notre approche améliore la qualité des résultats de la segmentation panoptique. Plus précisément, les performances du réseau de neurones en termes de compréhension des relations spatiales qui lient les objets d'une scène sont améliorées grâce à l'intégration des connaissances sur la relation spatiale PO pendant l'entraînement. Ces conclusions sont confirmées non seulement par les résultats quantitatifs prometteurs mais également par les résultats qualitatifs. Ces derniers montrent que notre approche résout avec une certaine efficacité les problèmes liés aux zones de l'image contenant des objets ayant des relations complexes entre eux telles que le chevauchement et les occultations.

5 Conclusion

Dans le cadre de cet article, nous proposons une nouvelle fonction de perte qui permet d'incorporer des connaissances sur les relations spatiales entre les objets dans le processus d'apprentissage d'un réseau de neurones pour la segmentation panoptique. L'approche proposée a montré des résultats prometteurs et encourageants sur le jeu de données CityScapes pour les scénarios urbains. Il est important de souligner que, dans le cadre de ce travail, nous intégrons uniquement la relation spatiale *Partially Overlapping PO* entre les objets de type *Stuff* afin de construire une preuve de concept et valider notre approche. Toutefois, notre idée dans une dimension plus globale est d'intégrer d'autres relations spatiales qualitatives pertinentes pour transférer davantage de connaissances aux réseaux de neurones. Nous visons également à étendre notre approche à l'ensemble des objets de type *Stuff* et *Things*. Une direction envisageable dans les travaux futurs inclut également l'exploitation des connaissances sur les relations spatiales à travers une modélisation et un raisonnement ontologique. Cette approche permettrait de pousser les réseaux de neurones à imiter au mieux le raisonnement humain afin de bien interpréter les relations spatiales entre les objets d'une scène et améliorer ainsi la précision de la segmentation panoptique.

Références

[1] J. Miao, X. Wang, Y. Wu, W. Li, X. Zhang, Y. Wei, Y. Yang. Large-scale video panoptic segmentation in the wild : A benchmark. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21033-21043, 2022.

[2] O. Zendel, M. Schörghuber, B. Rainer, M. Murschitz, C. Beleznai. Unifying panoptic segmentation for autonomous driving. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21351-21360, 2022.

[3] A. Milioto, J. Behley, C. McCool, C. Stachniss. Lidar panoptic segmentation for autonomous driving. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8505-8512, 2020.

[4] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár. Panoptic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.

[5] C.Y. Chang, S.E. Chang, P.Y. Hsiao, L.C. Fu. Epsnet : Efficient panoptic segmentation network with cross-layer attention fusion. *Proceedings of the Asian Conference on Computer Vision*, 2020.

[6] R. Mohan, A. Valada. EfficientPS : Efficient panoptic segmentation, *International Journal of Computer Vision*, 129(5), pp. 1551–1579, 2021.

[7] H. Caesar, J. Uijlings, , V. Ferrari. Coco-stuff : Thing and stuff classes in context. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209-1218, 2018.

[8] G. Neuhold, T. Ollmann, S.R. Bulo, P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. *IEEE International Conference on Computer Vision*, pp. 4990-4999, 2017.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele. The cityscapes dataset for semantic urban scene understanding. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213-3223, 2016.

[10] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L.C. Chen. Axial-deeplab : Stand-alone axial-attention for panoptic segmentation. *Springer Conference on European Conference on Computer Vision*, pp. 108-126, 2020.

[11] B. Cheng, M.D. Collins, Y. Zhu, T. Liu, T.S. Huang, H. Adam, L.C. Chen. Panoptic-deeplab : A simple, strong, and fast baseline for bottom-up panoptic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12475-12485, 2020.

[12] J. Behley, A. Milioto, C. Stachniss. A Benchmark for LiDAR-based Panoptic Segmentation based on KITTI. *IEEE International Conference on Robotics and Automation*, pp. 13596-13603, 2021.

[13] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom. Pointpillars : Fast encoders for object detection from point clouds. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697-12705, 2019.

[14] A.G. Cohn, B. Bennett, J. Gooday, N.M. Gotts. Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1, pp. 275-316, 1997.

[15] H. Liu, C. Peng, C. Yu, J. Wang, X. Liu, G. Yu, W. Jiang. An end-to-end network for panoptic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6172-6181, 2019.

[16] B. Koonce. EfficientNet. Convolutional Neural Networks with Swift for Tensorflow : Image Recognition and Dataset Categorization, *Springer Apress*, 2021.

[17] Y. Gatsoulis, M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, et al. Qsrlib : a software library for online acquisition of qualitative spatial relations from video, *International Workshop on Qualitative Reasoning*, 2016.

[18] S. Van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, et al. scikit-image : image processing in Python. *PeerJ*, 2, e453, 2014.

[19] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, X. Wang. Attentionguided unified network for panoptic segmentation. *IEEE Conference on Conference on Computer Vision and Pattern Recognition*, pp. 7026–7035, 2019.

[20] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, R. Urtasun. Upsnet : A unified panoptic segmentation network. *IEEE Conference on Conference on Computer Vision and Pattern Recognition*, pp 8818–8826, 2019.

[21] L. Porzi, S.R. Bulo, A. Colovic, P. Kontschieder. Seamless scene segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8277–8286, 2019.

[22] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, K. Huang. Ssap : Single-shot instance segmentation with affinity pyramid. *International Conference on Computer Vision*, pp. 642–651, 2019.

[23] K. Sofiiuk, O. Barinova, A. Konushin. Adaptis : Adaptive instance selection network. *IEEE/CVF International Conference on Computer Vision*, pp 7355–7363, 2019.

[24] W. Cao, J. Yuan, Z.He, Z.Zhang. Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection. *IEEE Access*, pp. 8990-8999, 2018.