Une nouvelle approche pour l'évaluation des méthodes monoculaires d'estimation de la profondeur basées sur l'apprentissage profond

Antoine Mauri, Redouane Khemmar, Benoit Decoux, Jean-Yves Ertaud, Madjid Haddad, Rémi Boutteau

Une nouvelle approche pour l'évaluation des méthodes monoculaires d'estimation de la profondeur basées sur l'apprentissage profond

Antoine Mauri [1,2]
Benoit Decoux [1]
Madjid Haddad[2]
Redouane Khemmar [1]
Jean-Yves Ertaud [1]
Remi Boutteau [3]

[1] Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France.
[2] SEGULA Engineering, 19 rue d'Arras, 92000 Nanterre, France.
[3] Normandie Univ, UNIROUEN, UNILEHAVRE, INSA Rouen, LITIS, 76000 Rouen, France.

antoine.mauri@esigelec.fr

Résumé

La détection d'objets, l'estimation de la profondeur et le suivi des objets sont des tâches très importantes pour la navigation autonome dans le contexte de la mobilité intelligente. Ces dernières années, l'apparition de nombreuses méthodes d'estimation de la profondeur basées sur l'apprentissage profond pour les caméras monoculaires a conduit à des progrès significatifs dans ce domaine. Dans cet article, nous proposons une évaluation des algorithmes de l'état de l'art pour l'estimation de la profondeur à partir d'images monoculaires sur les bases de données KITTI et NUScenes. Les modèles évalués dans cet article comprennent une méthode non supervisée (Monodepth2) et une méthode supervisée (BTS). Notre contribution réside dans l'élaboration de nouveaux protocoles d'évaluation de l'estimation de la profondeur : l'évaluation de la profondeur selon la classe de l'objet et l'évaluation sur des plages de distance. Nous avons validé nos nouveaux protocoles sur les bases de données KITTI et NuScenes, ce qui nous a permis d'obtenir une évaluation plus complète de l'estimation de la profondeur, en particulier pour les applications de compréhension de scènes dans des environnements routiers et ferroviaires.

Mots Clef

Estimation de la profondeur, évaluation des méthodes, apprentissage profond, vision par ordinateur, mobilité intelligente, approches monoculaires.

Abstract

In smart mobility based road navigation, object detection, depth estimation and tracking are very important tasks for improvement of the environment perception quality. In the recent years, a surge of deep-learning based depth estimation methods for monocular cameras has lead to significant progress in this field. In this paper, we propose an evaluation of state-of-the-art depth estimation algorithms based on single single input on both the KITTI dataset and the recently published NUScenes dataset. The models evaluated in this paper include an unsupervised method (Monodepth2) and a supervised method (BTS). Our work lies in the elaboration of novel depth evaluation protocols, object depth evaluation and distance ranges evaluation. We validated our new protocols on both KITTI and NuScenes datasets, allowing us to get a more comprehensive evaluation for depth estimation, especially for applications in scene understanding for both road and rail environment.

Keywords

Depth estimation, deep learning evaluation, computer vision, smart mobility, monocular approaches, KITTI dataset, NuScenes dataset.

1 Introduction

L'estimation précise de la profondeur est nécessaire pour la perception de l'environnement devant le véhicule et peut augmenter considérablement la sécurité en estimant la distance des piétons et des véhicules. La navigation autonome peut également améliorer la compétitivité des transports routiers, comme le montrent [1, 2], mais ce domaine présente encore de nombreux défis importants avant d'être complètement opérationnel. La mesure de la distance des objets peut être basée sur plusieurs types de capteurs : ultrasons, laser [3] ou caméra à temps de vol [4]. Ces solutions sont cependant encore très coûteuses. Dans ce travail, nous étudions l'utilisation de caméra(s) pour cette tâche. La méthode la plus courante pour estimer la distance avec des caméras est d'utiliser une paire stéréoscopique associée à des algorithmes d'appariement. Récemment, les algorithmes de vision basés sur les réseaux de neurones convolutifs (CNN) ont montré des performances de pointe dans l'estimation de la profondeur avec une seule caméra. Ces méthodes ont l'avantage de nécessiter un capteur relativement peu coûteux et facile à intégrer.

Ces méthodes manquent cependant encore d'une évaluation appropriée et précise. Dans ce contexte, nous avons déjà réalisé plusieurs travaux liés à la perception de l'environnement, comme le suivi d'une personne [5] ou la détection et le suivi d'objets, pour la mobilité intelligente sur route [6, 7].

Bien que les algorithmes d'estimation de la profondeur testés dans cet article offrent des résultats d'évaluation, ils ne fournissent qu'une évaluation de la performance globale d'une méthode. Il manque des informations sur la manière dont la distance des objets est prédite et sur la précision de la profondeur à des distances plus longues. Ces informations sont vitales pour les applications en environnement routier, notamment pour la conduite autonome. C'est ce qui a motivé notre contribution avec l'élaboration d'un protocole d'évaluation mieux adapté aux environnements routiers ainsi qu'une évaluation comparative, à l'aide de nos nouveaux protocoles, des méthodes les plus récentes sur deux grandes bases de données dans des environnements routiers : KITTI [8] et NuScenes [9].

La principale contribution de notre travail est d'offrir un nouveau protocole d'évaluation pour les algorithmes d'estimation de la profondeur à partir d'images monoculaires, adaptés aux environnements routiers et ferroviaires pour les véhicules autonomes, ainsi qu'une évaluation des méthodes de l'état de l'art.

Le reste de cet article est organisé comme suit : dans la section 2, nous passons en revue les algorithmes de l'état de l'art qui sont évalués dans cet article et les bases de données qui sont utilisées. Dans la section 3, nous présentons les nouveaux protocoles que nous proposons, et décrivons plus en détail les nouvelles métriques spécifiques proposées, la méthodologie d'évaluation et certaines spécificités de l'apprentissage sur la base de données NuScene. Les résultats expérimentaux sont présentés dans la section 4. Enfin, la conclusion et les orientations futures sont présentées dans la section 5.

2 Etat de l'art

2.1 Méthodes monoculaires d'estimation de la profondeur

Quand les bases de données utilisées contiennent la distance de l'objet comme information de vérité terrain, cette dernière peut être utilisée pour superviser l'apprentissage d'un réseau neuronal avec une couche de sortie de régression. La plupart des modèles CNN pour l'estimation de la profondeur ont une structure encodeur-décodeur, similaire à celle utilisée pour l'application de la segmentation sémantique des images [10] [11], dans laquelle la couche de sortie du réseau possède la même taille que les images d'entrée. L'un des principaux problèmes rencontrés par ces modèles est de retrouver la pleine résolution à la sortie du réseau, en raison du goulot d'étranglement qui existe à la jonction entre les parties encodeur et décodeur.

Dans le modèle Multi-Scale Local Planar Guidance (appelé BTS pour Big-To-Small dans la suite, le nom donné par les auteurs dans leur article) [12], des couches situées à plusieurs étapes de la phase de décodage sont utilisées et leurs sorties sont combinées pour prédire la profondeur à pleine résolution. Un autre CNN de cette catégorie est DenseDepth [13], dans lequel la partie encodeur est basée sur un DenseNet pré-entraîné [14]. L'ensemble du réseau est ensuite entraîné avec les jeux de données NYU Depth v2 [15] et KITTI. Une fonction de perte spécifique pénalisant la distorsion à haute fréquence permet une reconstruction plus fidèle de la profondeur aux contours des objets.

Une autre façon d'obtenir des informations de supervision consiste à utiliser des paires d'images stéréoscopiques alignées pendant la phase d'apprentissage, puis à déduire des cartes de profondeur sur des images monoculaires. Cette approche est appelée auto-supervisée, car ces modèles n'ont pas besoin d'une base de données avec la vérité terrain (comme les mesures LiDAR ou les cartes de disparité) pour l'apprentissage. Dans Monodepth [16], le problème de l'estimation de la profondeur est transformé en un problème de reconstruction d'image. Plus précisément, le CNN reçoit en entrée l'une des images du couple stéréo, et fournit en sortie une estimation des disparités. La deuxième image du couple stéréo est ensuite synthétisée à partir de cette estimation. La différence entre l'image synthétisée et l'image réelle est alors utilisée dans la fonction de perte [17][18]. Dans le modèle MonoResMatch [18], les auteurs utilisent également des images stéréo avec un apprentissage de bout en bout pour obtenir les informations de profondeur. Ils n'utilisent pas la vérité terrain présente dans les jeux de données, mais un algorithme de mise en correspondance stéréo (Semi-Global Matching) pour générer cette information en interne.

Dans Monodepth2 [19], les auteurs présentent un ensemble d'améliorations par rapport à la première version de leur algorithme [16]. Le modèle est basé sur trois processus qui coopèrent afin d'améliorer l'estimation de la profondeur : le modèle peut être entraîné avec des données monoculaires, des données stéréoscopiques, ou les deux. Plus précisément, les 3 processus sont constitués de : (1) une perte de reprojection minimale calculée pour chaque pixel, (2) une perte d'auto-masquage pour éviter les confusions et (3) une méthode de multi-échantillonnage à pleine résolution pour réduire les artefacts visuels. L'efficacité du modèle est démontrée quantitativement et qualitativement sur le jeu de données KITTI.

D'autres méthodes utilisent des images monoculaires pour l'apprentissage et l'inférence. Plus précisément, des séquences d'images sont utilisées pour l'apprentissage, et au moment du test, la profondeur est estimée sur des images monoculaires. Dans [20], un CNN avec deux modules partageant les premières couches convolutives est utilisé pour donner conjointement des estimations pour la profondeur et la position de la caméra. Dans [21], la structure géométrique des objets et de la scène est introduite dans le processus d'apprentissage, en utilisant des estimateurs séparés du

mouvement des objets, de l'ego-motion et de la profondeur, ce qui rend le modèle bien adapté aux environnements hautement dynamiques.

En utilisant des séquences d'images stéréoscopiques, il est également possible d'apprendre la profondeur et l'odométrie en même temps, puisque les erreurs photométriques spatiales et temporelles sont disponibles [22].

2.2 Bases de données pour la prédiction de la profondeur dans un environnement extérieur

Les méthodes d'apprentissage profond pour la détection d'objets et l'estimation de la profondeur nécessitent de plus en plus de jeux de données d'entraînement et d'évaluation contenant des données hétérogènes, comme des images, des vidéos, etc. C'est pourquoi il est très important d'identifier une base de données riche pour la détection d'objets en temps-réel et l'estimation de la profondeur visant les applications de navigation routière ou ferroviaire.

KITTI
Dans KITTI [8], les auteurs présentent l'un des jeux de données les plus utilisés en environnement routier pour la recherche en robotique mobile et en conduite autonome. KITTI est une base de données relative à la conduite autonome dont les données sont calibrées, synchronisées et horodatées et qui ont été capturées dans un large éventail de scénarios. La base de données KITTI a été collectée à l'aide d'un véhicule VW Station équipé de différents types de capteurs tels que des caméras stéréoscopiques (en couleur et en niveaux de gris), un scanner laser 3D Velodyne et un système de navigation GPS/IMU de haute précision. La plateforme contient des situations de trafic réelles avec des objets statiques et dynamiques, pour lesquels les étiquettes des objets sont présentées sous la forme de tracklets (éléments de trajectoire très courts) en 3D. Le jeu de données fournit des comparatifs en ligne pour différentes tâches telles que : la reconstruction 3D par stéréovision, l'estimation du flot optique et la détection d'objets.

NuScenes
NuTonomy scenes (NuScenes [9]) est un jeu de données multimodal pour la conduite autonome. NuScenes contient différents types de capteurs tels que : 6 caméras, 5 radars et 1 LiDAR. Il est entièrement annoté et comprend 1000 séquences (avec 20 acquisitions chacune) et des boîtes de délimitation 3D pour 23 classes. Il contient 100 fois plus d'images que la base KITTI. Il contient également une vérité terrain adaptée à la détection et au suivi basés sur le LiDAR et l'image [9].

3 Évaluation des méthodes monoculaires d'estimation de la profondeur basées sur l'apprentissage profond

3.1 Métriques des erreurs utilisées dans l'évaluation approfondie

Avant de présenter nos contributions pour l'évaluation de la profondeur, nous définissons ci-dessous les métriques d'erreur de profondeur qui sont utilisées dans la littérature et dans notre travail. Soit p la prédiction de profondeur d'un pixel de l'image, g sa vérité de terrain et N le nombre total de pixels de profondeur dans l'image.
Erreur relative : L'équation de l'erreur relative (appelé RE pour Relative Error) est détaillée dans l'équation (1).

$$RE = \frac{1}{N}\sum_i\sum_j \frac{|g_{i,j} - p_{i,j}|}{g_{i,j}} \tag{1}$$

Erreur relative au carré : L'équation de l'erreur relative au carré (appelé SRE pour Squared Relative Error) est détaillée dans l'équation (2).

$$SRE = \frac{1}{N}\sum_i\sum_j \frac{|g_{i,j} - p_{i,j}|^2}{g_{i,j}} \tag{2}$$

Root Mean Squared Error : L'erreur quadratique moyenne (appelé RMSE pour Root Mean Squared Error) est donnée par l'équation (3).

$$RMSE = \sqrt{\frac{1}{N}\sum_i\sum_j (g_{i,j} - p_{i,j})^2} \tag{3}$$

Erreur quadratique moyenne logarithmique : L'erreur quadratique moyenne logarithmique (logRMSE) est donnée par l'équation (4).

$$logRMSE = \sqrt{\frac{1}{N}\sum_i\sum_j (log(g_{i,j}) - log(p_{i,j}))^2} \tag{4}$$

Pourcentage de pixels non conformes : Le pourcentage de pixels non conformes (appelé BMP pour Bad Matching Pixels) est donné par l'équation (5), où C est un seuil utilisé pour définir une tolérance d'erreur.

$$[a]_{k=[1..3]} = \frac{1}{N}\sum_i\sum_j max(\frac{g_{i,j}}{p_{i,j}}, \frac{p_{i,j}}{g_{i,j}}) < C^k \tag{5}$$

Ces métriques donnent une évaluation statistique complète de la performance d'une méthode, mais nous pensons qu'elles peuvent être développées davantage. L'une de nos contributions réside dans de nouveaux protocoles d'évaluation de la profondeur que vous trouverez ci-après.

3.2 Évaluation de la profondeur selon l'objet

Alors que l'évaluation actuelle des estimations de la profondeur donne une évaluation complète de la performance globale d'une méthode donnée, elle est faite sur l'image globale et n'évalue pas la prédiction de la distance des objets. Or la distance des objets est un aspect fondamental pour les applications de conduite autonome et de perception des scènes. C'est pourquoi nous avons conçu un nouveau protocole d'évaluation de la profondeur qui nous permet de calculer l'erreur de prédiction de la profondeur pour les objets pertinents que l'on rencontre régulièrement dans les environnements routiers (personne, voiture, camion, etc.).

Notre protocole d'évaluation est composé de 4 étapes : (1) La carte de profondeur prédite est mise à l'échelle en utilisant une mise à l'échelle médiane. Soit p la carte de profondeur prédite et g la carte de profondeur de la vérité de terrrain, la mise à l'échelle médianne est décrite comme suit : $p = p * \frac{med(gt)}{med(p)}$. (2) Les masques d'objets sont générés en utilisant Mask-RCNN [23] (voir la Figure 1 pour un exemple de sortie du réseau) ; (3) Les masques des objets générés sont ensuite utilisés pour segmenter les cartes de profondeur et les erreurs de profondeur sont calculées pour chaque masque dans l'image ; (4) Enfin, la moyenne des erreurs est calculée pour chaque classe. Ce nouveau protocole d'évaluation permettra de mieux comprendre la façon dont une méthode donnée estime la distance des objets présents dans l'environnement routier. Il est particulièrement utile pour les applications de conduite autonome. Les différentes étapes sont illustrées dans la Figure 2.

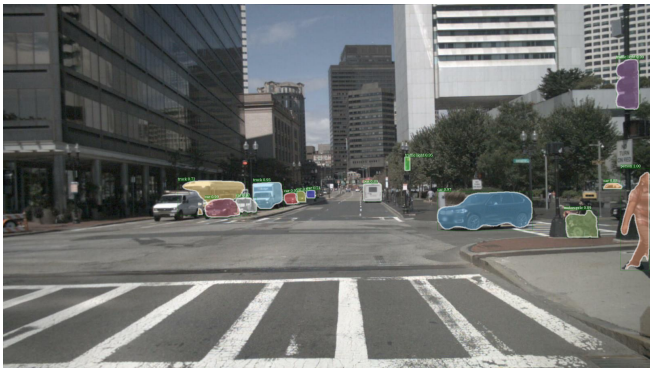

FIGURE 1 – Masques des objets obtenus avec Mask-RCNN sur une image du jeu de données NuScenes.

3.3 Évaluation de la profondeur sur des plages de distance

Un autre inconvénient du protocole classique d'évaluation de la profondeur est qu'il ne permet pas d'évaluer les performances d'une méthode selon la distance. La performance d'une méthode sur des distances plus longues est un paramètre important qui doit être pris en compte pour la compréhension de la scène. Nous proposons ici de suivre le travail de [24] qui a décrit un protocole d'évaluation sur des plages de distance, mais alors qu'ils ont utilisé ce protocole pour des scènes intérieures, nous l'avons utilisé dans un environnement routier où les plages de distance sont plus importantes. Le protocole que nous utilisons est donc composé des étapes suivantes : (1) La carte de profondeur prédite est mise à l'échelle en utilisant une échelle médiane ; (2) Des plages de distance de 10m à 80m ([0-10m], [10-20m], ..., [70-80m]) sont créées ; (3) Chaque pixel est affecté à une plage de distance en fonction de sa valeur dans la vérité terrain de la profondeur ; (4) Pour chaque plage de distance, les erreurs de profondeur sont calculées. Ce nouveau protocole permet d'évaluer la dégradation de l'estimation de la profondeur en fonction des distances.

4 Analyse des résultats

Dans cette section, nous présentons les résultats de notre évaluation des méthodes monoculaires d'estimation de la profondeur supervisées et auto-supervisées BTS et Monodepth2 (respectivement), sur les jeux de données KITTI et NuScenes.

4.1 Base de données KITTI

Pour l'évaluation sur la base KITTI, nous avons utilisé les modèles pré-entraînés fournis par les auteurs des modèles BTS et Monodepth2. Le modèle BTS a été entraîné avec des images provenant de l'entraînement d'Eigen [25] à une résolution de 704×352 et avec une vérité terrain dense. Les poids de Monodepth2 ont été entraînés en utilisant la vérité terrain monoculaire sur l'ensemble de Zhou[20], à une résolution de 1024×320. L'évaluation a été réalisée sur l'ensemble de test d'Eigen. Les résultats de notre évaluation de BTS et Monodepth2 sont présentés dans les tableaux 1 et 2. La valeur de *C* pour l'erreur de seuil définie dans l'équation (5) a été fixée à 1,25.

4.2 Base de données NuScenes

Pour l'évaluation sur NuScenes, nous avons dû entraîner les deux méthodes sur l'ensemble d'entraînement de ce jeu de données. Pour BTS, nous avons réalisé un entraînement de 50 epochs avec une taille de batch de 20 et une résolution de 192×192, avec les données éparses de la supervision LiDAR. Étant donné que Monodepth2 est une méthode non supervisée, elle repose sur la fonction de perte de reprojection monoculaire. Si les images ont une mauvaise visibilité (due aux conditions météorologiques, de l'obscurité, etc.), l'entraînement peut ne pas converger. C'est pourquoi nous avons sélectionné les scènes de l'entraînement où la visibilité est suffisamment bonne pour que l'entraînement converge. Nous avons également utilisé toutes les images de chaque scène et pas seulement celles qui étaient synchronisées avec le LiDAR afin d'obtenir une fréquence d'images suffisamment élevée pour que l'entraînement monoculaire fonctionne. Nous avons entraîné Monodepth2 pendant 20 epochs avec une taille de batch de 12 et une résolution de 446×224. Les résultats de notre évaluation de BTS et de Monodepth2 sont présentés dans les

TABLE 1 – Évaluation de la profondeur sur des plages de distance pour KITTI. Les algorithmes évalués sont des méthodes monoculaires d'estimation de la profondeur reconnues : Monodepth2 (MD2) et BTS. Les erreurs de profondeur globales sont indiquées ainsi que les erreurs de profondeur pour des distances comprises entre 10m et 80m. La RMSE est exprimée en mètres.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Plage de distance	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
$0 - 80m$	0.115	**0.060**	0.882	**0.249**	4.701	**2.798**	0.190	**0.096**	0.879	**0.955**	0.961	**0.993**	0.982	**0.998**
$0 - 10m$	0.102	**0.071**	0.503	**0.188**	1.489	**0.991**	0.141	**0.106**	0.929	**0.959**	0.979	**0.988**	0.990	**0.994**
$10 - 20m$	0.116	**0.088**	0.845	**0.395**	3.035	**2.198**	0.180	**0.149**	0.891	**0.924**	0.960	**0.971**	0.979	**0.985**
$20 - 30m$	0.168	**0.130**	1.866	**1.055**	6.208	**4.745**	0.261	**0.229**	0.773	**0.836**	0.916	**0.934**	0.957	**0.964**
$30 - 40m$	0.196	**0.160**	2.788	**1.945**	9.110	**7.476**	0.307	**0.279**	0.694	**0.764**	0.886	**0.906**	0.942	**0.947**
$40 - 50m$	0.209	**0.174**	3.504	**2.640**	11.682	**10.008**	0.318	**0.298**	0.641	**0.725**	0.865	**0.889**	**0.943**	0.941
$50 - 60m$	0.221	**0.190**	4.394	**3.739**	14.252	**12.852**	0.332	**0.326**	0.583	**0.675**	0.857	**0.868**	**0.927**	0.922
$60 - 70m$	0.212	**0.201**	4.657	**4.584**	15.855	**15.585**	**0.325**	0.334	0.609	**0.619**	0.854	**0.856**	**0.930**	0.923
$70 - 80m$	**0.181**	0.214	**4.340**	5.454	**15.800**	18.219	**0.284**	0.333	**0.652**	0.548	**0.873**	0.843	**0.945**	0.925

TABLE 2 – Évaluation de la distance selon l'objet sur KITTI. Les algorithmes évalués sont des méthodes monoculaires d'estimation de la profondeur reconnues : Monodepth2 (MD2) et BTS. Les erreurs de profondeur ont été calculées pour les classes d'objets ayant suffisamment d'instances dans l'ensemble de test. La RMSE est exprimée en mètres.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Classe de l'objet	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Personne	0.314	**0.166**	5.721	**1.786**	8.430	**5.892**	0.326	**0.253**	0.601	**0.772**	0.829	**0.894**	0.920	**0.947**
Deux roues	0.131	**0.116**	0.517	**0.467**	2.810	**2.669**	0.172	**0.163**	0.829	**0.839**	**0.964**	0.962	0.993	**0.994**
Voiture	0.206	**0.137**	3.132	**1.491**	7.924	**6.052**	0.271	**0.223**	0.773	**0.838**	0.883	**0.922**	0.938	**0.955**
Camion	0.215	**0.122**	2.769	**0.826**	6.978	**4.523**	0.259	**0.177**	0.694	**0.854**	0.903	**0.969**	0.964	**0.985**

tableaux 3 et 4. La valeur de *C* pour l'erreur de seuil définie dans l'équation (5) a été fixée à 1,25.

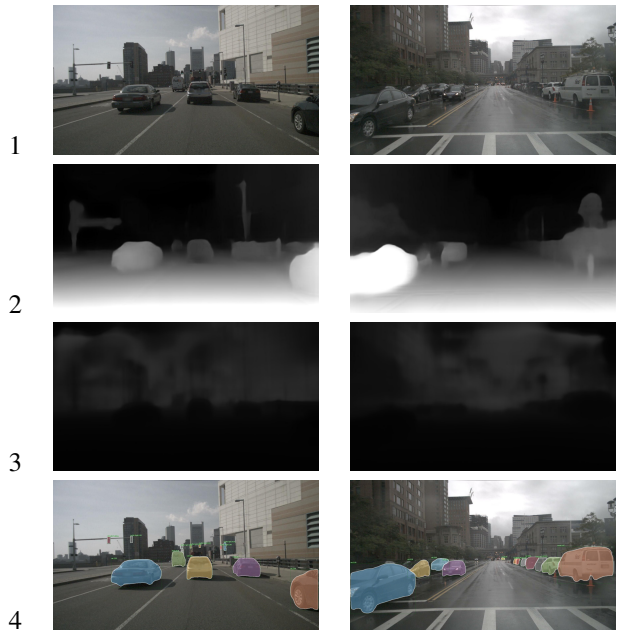

FIGURE 2 – Données d'entrée pour notre protocole d'évaluation de la distance selon la classe des objets. (1) l'image d'entrée alimentant l'algorithme de prédiction de la profondeur, (2) la carte de disparité, (3) la carte de profondeur normalisée après mise à l'échelle médiane et (4) les masques des objets de Mask-RCNN.

4.3 Analyse des résultats expérimentaux

Nos résultats sur les deux jeux de données montrent que, globalement, BTS donne de meilleurs résultats que Monodepth2. Notre évaluation sur les plages de distance montre également que les deux méthodes, comme prévu, ont tendance à avoir une précision plus faible lorsque la distance augmente. Notre évaluation de la profondeur des objets montre également que les erreurs de prédiction de la profondeur sont significativement plus élevées que les erreurs sur l'image globale (voir les figures 5 et 6). Cela peut s'expliquer par la grande variété de chaque classe d'objets qui rend l'apprentissage de la profondeur plus difficile pour les CNN, alors que l'environnement environnant est moins variable, ce qui facilite l'apprentissage de la profondeur par ces méthodes. Nous pouvons voir que les erreurs peuvent être 2 fois plus élevées que l'erreur globale pour les personnes et les voitures, et cela doit être pris en compte si ces méthodes sont utilisées dans le cadre du véhicule autonome. Enfin, nous pouvons voir que nos résultats sur la base de données NuScenes sont des erreurs plus élevées que sur KITTI, ce qui peut s'expliquer par les différences dans l'entraînement entre les deux jeux de données, notamment que NuScenes possède des scènes plus difficiles pour l'estimation de la profondeur. Par exemple, certaines scènes ont été acquises par temps de pluie et présentent des réflexions dues à la route mouillée. Des scènes ont également été capturées de nuit avec une mauvaise visibilité. Ces scènes ont une erreur beaucoup plus élevée que celles ayant une bonne visibilité et cela contribue à augmenter les erreurs moyennes utilisées pour calculer l'erreur glo-

TABLE 3 – Évaluation de la profondeur sur des plages de distance pour NuScenes. Les algorithmes évalués sont des méthodes monoculaires d'estimation de la profondeur reconnues : Monodepth2 (MD2) et BTS. Les erreurs de profondeur globales sont indiquées ainsi que les erreurs de profondeur pour des distances de 10m et jusqu'à 80m. La RMSE est exprimée en mètres.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Plage de distance	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
$0-80m$	0.176	**0.147**	2.521	**1.184**	7.746	**5.849**	0.271	**0.214**	0.787	**0.817**	0.911	**0.94**	0.955	**0.977**
$0-10m$	**0.115**	0.116	0.982	**0.430**	1.561	**1.347**	**0.139**	0.151	**0.919**	0.915	0.972	**0.974**	0.986	**0.987**
$10-20m$	0.187	**0.153**	2.690	**0.966**	4.854	**3.452**	0.243	**0.197**	0.794	**0.810**	0.916	**0.945**	0.958	**0.984**
$20-30m$	0.242	**0.162**	3.488	**1.386**	8.316	**5.427**	0.320	**0.217**	0.643	**0.768**	0.859	**0.933**	0.932	**0.978**
$30-40m$	0.256	**0.183**	4.296	**2.010**	11.327	**7.922**	0.368	**0.255**	0.569	**0.677**	0.807	**0.909**	0.904	**0.969**
$40-50m$	0.264	**0.206**	5.140	**3.047**	14.167	**10.952**	0.404	**0.300**	0.518	**0.598**	0.760	**0.845**	0.888	**0.946**
$50-60m$	0.270	**0.225**	6.298	**4.441**	17.267	**14.398**	0.440	**0.346**	0.483	**0.556**	0.746	**0.789**	0.854	**0.904**
$60-70m$	0.280	**0.248**	8.024	**6.275**	20.725	**18.243**	0.475	**0.385**	0.480	**0.495**	0.692	**0.736**	0.817	**0.868**
$70-80m$	0.289	**0.284**	10.299	**8.802**	24.621	**23.160**	0.505	**0.434**	**0.463**	0.394	0.645	**0.677**	0.776	**0.834**

TABLE 4 – Évaluation de la distance selon la classe des objets sur NuScenes. Les algorithmes évalués sont des méthodes monoculaires d'estimation de la profondeur reconnues : Monodepth2 (MD2) et BTS. Les erreurs de profondeur ont été calculées pour les classes d'objets ayant suffisamment d'instances dans l'emsemble de test. La RMSE est exprimée en mètres.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Classe de l'objet	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Voiture	0.346	**0.218**	6.853	**2.144**	10.420	**6.862**	0.448	**0.278**	0.546	**0.708**	0.736	**0.880**	0.920	**0.949**
Personne	0.501	**0.384**	8.312	**3.910**	9.291	**7.858**	0.531	**0.449**	0.438	**0.492**	0.679	**0.717**	0.803	**0.839**
Bus	0.448	**0.228**	11.837	**2.226**	13.929	**7.811**	0.448	**0.274**	0.465	**0.644**	0.729	**0.891**	0.848	**0.958**
Camion	0.324	**0.218**	6.803	**2.091**	11.425	**7.263**	0.378	**0.260**	0.574	**0.674**	0.793	**0.902**	0.887	**0.964**
Moto	0.284	**0.245**	1.671	**1.430**	4.509	**3.917**	0.320	**0.288**	0.512	**0.658**	0.868	**0.869**	0.935	**0.946**

bale. En combinant nos deux protocoles d'évaluation, nous avons également calculé l'évolution de l'erreur pour des objets tels que les voitures sur des plages de distance (voir figures 3 et 4). Ces résultats comparatifs peuvent être utilisés pour évaluer l'adéquation d'une méthode d'estimation de la profondeur à un scénario particulier dans les environnements routiers. Par exemple, pour la conduite sur une route dans des conditions idéales, où nous supposons que le véhicule roule à 90 km/h, la méthode doit être précise jusqu'à 60 m (distance de sécurité entre deux véhicules à cette vitesse).

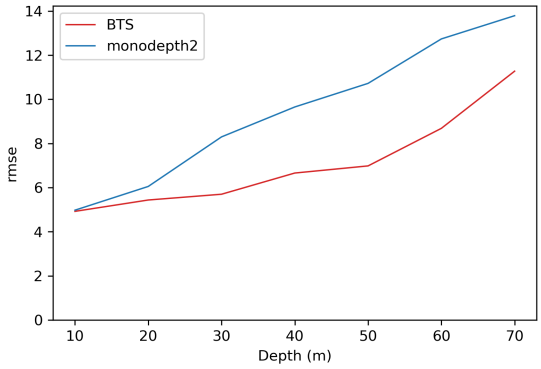

FIGURE 3 – RMSE pour la classe d'objets voiture selon la distance pour la base de données KITTI. La RMSE est exprimée en mètres.

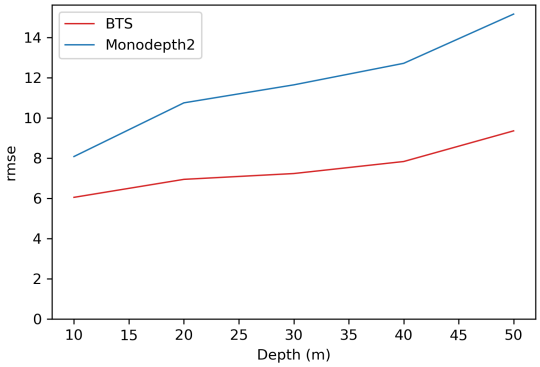

FIGURE 4 – RMSE pour la classe d'objets voiture selon la distance pour la base de données NuScenes. La RMSE est exprimée en mètres.

5 Conclusion

Nous avons présenté dans cet article un nouveau protocole d'évaluation de la profondeur mieux adapté aux applications de conduite autonome dans des scènes routières ainsi qu'une évaluation de BTS et Monodepth2, deux méthodes monoculaires d'estimation de la profondeur performantes, en utilisant nos nouveaux protocoles. En ce qui concerne les protocoles d'évaluation de la profondeur, nous avons proposé une méthode basée sur des plages de distance permettant d'évaluer l'évolution de la précision de la profondeur sur la distance et un protocole pour évaluer les prédic-

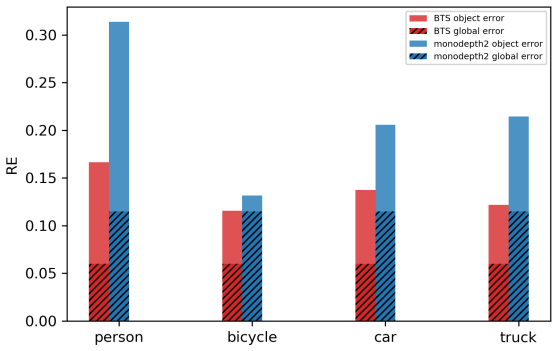

FIGURE 5 – Nos résultats d'erreur relative (RE) de BTS et Monodepth2 pour différentes classes d'objets comparés à l'erreur relative (RE) globale (hachuré) sur KITTI.

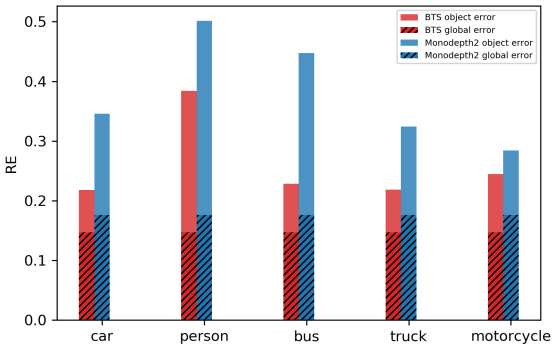

FIGURE 6 – Nos résultats d'erreur relative (RE) de BTS et Monodepth2 pour différentes classes d'objets comparés à l'erreur relative (RE) globale (hachuré) sur NuScenes.

tions de la profondeur des objets en utilisant les masques des objets générés par Mask-RCNN, l'un des meilleurs détecteurs d'objets. Nous avons ensuite réalisé une évaluation d'une méthode non supervisée et d'une méthode supervisée avec Monodepth2 et BTS sur deux jeux de données à grande échelle pour les scènes routières, KITTI et NuScenes. Les résultats comparatifs de l'évaluation montrent que BTS a de meilleures performances que Monodepth2 sur tous les aspects. Nous avons également montré que les erreurs d'estimation de la profondeur des objets étaient significativement plus élevées que les erreurs sur l'image entière pour les deux méthodes. Cependant, bien que nous offrions une évaluation complète de l'estimation de la profondeur dans des environnements routiers, d'autres méthodes, y compris les algorithmes basés sur la stéréovision, pourraient également être évaluées pour une étude comparative complète. Nous avons également l'intention d'évaluer les algorithmes d'estimation de la profondeur dans des environnements ferroviaires, mais en raison de l'absence de bases de données publiques avec des images de caméra et des vérités terrain de profondeur provenant d'un LiDAR, nous devrons acquérir notre propre base de données pour cette tâche. Ainsi, nous proposons de développer un système d'acquisition comprenant une caméra stéréoscopique et un LiDAR afin de pouvoir collecter notre propre jeu de données dans l'environnement ferroviaire.

Remerciements

Cette recherche est soutenue par SEGULA Technologies et le projet M2SINUM (Ce projet est cofinancé par l'Union européenne avec le Fonds européen de développement régional (FEDER, 18P03390/18E01750/18P02733) et par le Conseil régional de Haute-Normandie via le projet M2SINUM). Nous tenons à remercier SEGULA Technologies pour leur collaboration et les ingénieurs du Laboratoire de Navigation Autonome de l'IRSEEM pour leur soutien. Ce travail a été réalisé en partie sur des ressources informatiques fournies par le CRIANN (Centre Régional Informatique et d'Applications Numériques de Normandie, Normandie, France).

Références

[1] H. Mukojima, D. Deguchi, Y. Kawanishi, I. Ide, H. Murase, M. Ukai, N. Nagamine, and R. Nakasone, "Moving camera background-subtraction for obstacle detection on railway tracks," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3967–3971, IEEE, 2016.

[2] S. Yanan, Z. Hui, L. Li, and Z. Hang, "Rail surface defect detection method based on yolov3 deep learning networks," in *2018 Chinese Automation Congress (CAC)*, pp. 1563–1568, IEEE, 2018.

[3] J. Palacín, T. Pallejà, M. Tresanchez, R. Sanz, J. Llorens, M. Ribes-Dasi, J. Masip, J. Arno, A. Escola, and J. R. Rosell, "Real-time tree-foliage surface estimation using a ground laser scanner," *IEEE transactions on instrumentation and measurement*, vol. 56, no. 4, pp. 1377–1383, 2007.

[4] B. Kang, S.-J. Kim, S. Lee, K. Lee, J. D. Kim, and C.-Y. Kim, "Harmonic distortion free distance estimation in tof camera," in *Three-Dimensional Imaging, Interaction, and Measurement*, vol. 7864, p. 786403, International Society for Optics and Photonics, 2011.

[5] R. Khemmar, M. Gouveia, B. Decoux, and J.-Y. Ertaud, "Real time pedestrian and object detection and tracking-based deep learning. application to drone visual tracking," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 01 2019.

[6] Z. Chen, R. Khemmar, B. Decoux, A. Atahouet, and J.-Y. Ertaud, "Real time object detection, tracking, and distance and motion estimation based on deep learning : Application to smart mobility," in *2019 Eighth International Conference on Emerging Security Technologies (EST)*, pp. 1–6, IEEE, 2019.

[7] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Deep learning for real-time 3d multi-object detection, localisation, and tracking : Application to smart mobility," *Sensors*, vol. 20, no. 2, p. 532, 2020.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics : The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes : A multimodal dataset for autonomous driving," *arXiv preprint arXiv :1903.11027*, 2019.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet : A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, p. 2481–2495, Dec 2017.

[12] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small : Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv :1907.10326*, 2019.

[13] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv :1812.11941*, 2018.

[14] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[15] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[16] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279, 2017.

[17] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[18] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9809, 2019.

[19] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv :1806.01260*, 2018.

[20] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

[21] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors : Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8001–8008, 2019.

[22] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017.

[24] T. Koch, L. Liebel, F. Fraundorfer, and M. Körner, "Evaluation of cnn-based single-image depth estimation methods," *CoRR*, vol. abs/1805.01328, 2018.

[25] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2658, 2015.