Détection multiple de points de fuite horizontaux par deep learning

Abdelkarim Elassam, Gilles Simon, Marie-Odile Berger

Détection multiple de points de fuite horizontaux par deep learning

A. Elassam[1] G. Simon[1] M.-O. Berger[1]

[1] LORIA-INRIA-Université de Lorraine

prenom.nom@inria.fr

Résumé

Dans cet article, nous proposons des réseaux de neurones convolutionnels (CNN) pour l'estimation de plusieurs points de fuite horizontaux (PFh) à partir d'une seule image. Motivés par le succès des travaux récents qui utilisent des techniques basées sur l'apprentissage pour la détection des lignes d'horizon, nous proposons d'abord une méthode de détection des PFh qui intègre la puissance des CNN et la robustesse des méthodes a-contrario pour extraire des PFh significatifs. Pour détecter plusieurs PFh et tirer parti des informations structurelles de la carte des normales de surface, nous proposons un modèle multi-tâches qui estime conjointement la ligne d'horizon, les points de fuite horizontaux et la carte des normales de surface à partir d'une seule image. Enfin, nous introduisons un processus de fusion pour récupérer un ensemble plus étendu de points de fuite tout en évitant les doublons. Nous évaluons notre méthode de manière approfondie sur l'ensemble de données Holicity contenant des carte des normales et la vérité terrain des points de fuite. Nous la comparons également à une méthode de référence.

Mots Clef

Points de fuite, deep learning, localisation.

Abstract

We investigate in this paper the application of convolutional neural networks (CNN) for estimating multiple horizontal vanishing points (hVPs) from a single RGB image. Motivated by the success of recent works that use learning-based techniques for horizon line detection, we first propose a method for hVP detection that integrates CNN's power and the robustness of the a-contrario method for extracting meaningful hVPs. To detect multiple hVPs and take advantage of the surface normal map's structural information, we propose a multi-task model that estimates the horizon line jointly, hVPs and the surface normal map from a single RGB image. Finally, we introduce a fusion process to recover a more extensive set of hVPs while avoiding duplicate ones. We evaluate our method thoroughly on the Holicity dataset containing ground-truth surface normal maps and VPs. We also compare it to a state-of-the-art algorithm for hVP detection.

Keywords

Vanishing points, deep learning, visual localization.

1 Introduction

Dans une image acquise par une caméra sténopé, un point de fuite (PF) est le lieu de rencontre de lignes parallèles entre elles dans l'espace, et la ligne d'horizon (LH) est le lieu des points de fuite de direction horizontale. Détecter la ligne d'horizon ou des points de fuite est une étape essentielle de nombreux problèmes de vision par ordinateur et de robotique tels que la calibration de la caméra [4, 21], la reconstruction 3D [8], l'estimation de pose [9], l'estimation de cartes de normales [20], la localisation et la cartographie simultanées (SLAM) [12] etc...Des travaux récents ont montré que des méthodes d'apprentissage profond sont capables d'estimer la ligne d'horizon de manière robuste [3, 16, 23], mais la détection de points de fuite par réseaux de neurones convolutifs (CNN) s'avère plus problématique, principalement parce que le nombre de points de fuite diffère d'une image à l'autre. Par conséquent, la plupart des travaux se limitent à détecter un point de fuite dominant [16], ou des points de fuite contraints sur une grille à l'intérieur de l'image [3].

Dans les environnements fabriqués par l'homme, la plupart des points de fuite correspondent à une direction horizontale de l'espace. Une connaissance *a priori* de la ligne d'horizon, qu'elle soit obtenue à l'aide d'un CNN [23] ou d'une méthode plus classique [19], permet alors de restreindre à ce lieu la zone de recherche des points de fuite. C'est ce qu'exploitent les auteurs de [23] dont la méthode génère des hypothèses de ligne d'horizon en utilisant un CNN, puis retient la ligne sur laquelle le plus grand nombre de segments détectés dans l'image semble converger. La deuxième étape de cette méthode génère toutefois de nombreux faux positifs en raison d'une faible généralisation de ses paramètres. Les auteurs de [19] exploitent les segments de l'image pour retrouver à la fois la ligne d'horizon et les points de fuite. Une méthode dite *a contrario* permet dans les deux cas d'éviter des problèmes seuils, grâce à l'utilisation d'un critère universel, appelé *Nombre de Fausses*

Alarmes, qui doit tout simplement être inférieur à 1.

Notre méthode s'inscrit dans la continuité de ces travaux. Elle tire parti d'un apprentissage profond mais également de la robustesse des méthodes *a contrario* pour détecter de multiples points de fuite horizontaux (PFh) situés à l'intérieur ou à l'extérieur de l'image, et dont le nombre n'est pas connu d'avance.

2 État de l'art

2.1 Détection directe des points de fuite

Les méthodes existantes peuvent être séparées en deux catégories : d'un côté les méthodes basées sur un détecteur de segments de droites tel que LSD [7], de l'autre les méthodes – plus minoritaires – basées sur un apprentissage supervisé. Les premières regroupent les segments détectés par PF à l'aide d'algorithmes de type RANSAC [2], J-linkage [1] ou Transformée de Hough [11]. L'algorithme Expectation-Maximization (EM) [14] est également souvent utilisé en post-traitement pour affiner les positions des PF obtenus.

Des travaux récents ont abordé le problème de la détection directe de PF par CNN. Borji [3] propose de diviser l'image en cellules et utilise un CNN pour prédire si un PF est présent dans chaque cellule. L'inconvénient de cette approche est qu'un seul PF est finalement détecté à l'intérieur de l'image. Kluger et al. [13] ont conçu un réseau qui opère sur la représentation en sphère gaussienne des lignes vues dans l'image plutôt que sur l'image brute. Afin d'exploiter les propriétés géométriques des PF en tant qu'intersection de lignes parallèles, Zhou et al. [25] ont développé un nouvel opérateur convolutif appelé *convolution conique*. Il s'agit d'une convolution régulière mise en œuvre dans un espace conique canonique, dans lequel les lignes structurelles locales sont toujours horizontales.

Les auteurs de [24] ont testé sur le jeu de données Holicity [24] à la fois la méthode NeurVPs [25] et une méthode classique basée sur le détecteur de lignes LSD [7] et la méthode de regroupement J-Linkage [1]. La méthode NeurVPS [25] est plus performante lorsqu'on ne considère que les PF zénithaux mais elle a des performances similaires à LSD [7] + J-Linkage [1] lorsqu'on considère tous les PF.

La plupart des travaux imposent une contrainte sur le nombre de PF possibles. Dans [3, 16], seul le PF dominant est détecté. D'autres travaux [1, 14, 21] utilisent l'hypothèse d'un monde dit *de Manhattan* dans lequel trois PF orthogonaux deux à deux sont supposés coexister : un PF vertical (appelé *zénith*) et deux PF horizontaux. Cette hypothèse permet de prendre en compte des scène urbaines bien choisies, mais ne convient pas, loin s'en faut, dans tous les cas. Une hypothèse moins stricte est celle d'un monde dit *d'Atlanta* dans lequel sont supposés représentées une direction verticale et un nombre illimité de directions horizontales. Plusieurs travaux [23, 25], dont ceux que l'on présente ici, reposent sur cette hypothèse.

2.2 Détection de points de fuite par horizon premier

La détection de PF avec pour première étape la détection de la ligne d'horizon [18, 23, 19] est basée sur le schéma suivant : (i) des hypothèses de LH sont proposées ; (ii) (dans [23] et [19] uniquement) des centaines d'hypothèses supplémentaires sont échantillonnées autour des hypothèses proposées ; (iii) des PF sont détectés le long des lignes candidates et (iv) la ligne obtenant les PF les plus consistants est sélectionnée comme LH, avec ses PF.

Dans [18] et [19], les LH candidates sont prédites perpendiculairement à la ligne zénithale, en exploitant la propriété géométrique que les lignes de l'espace inscrites dans le plan horizontal passant par le centre optique se projettent sur la LH quelle que soit leur direction horizontale. Cette propriété se traduit généralement dans l'image par une accumulation de segments horizontaux à hauteur de la LH. La principale différence entre [18] et [19] réside dans la manière dont les accumulations de segments horizontaux puis les PF sont détectés. Des résultats bien plus précis ont pu être obtenus avec la méthode [19] grâce à l'utilisation d'une méthode *a contrario* [6] pour résoudre les deux problèmes.

Dans [23], les prédictions de la LH sont obtenues sur la base d'un CNN puis, pour chaque prédiction, une mesure de consistance géométrique avec les segments de l'image est utilisée pour détecter les PF potentiels et attribuer un score à la ligne candidate. Le CNN permet d'obtenir des LH de manière robuste et précise, mais l'étape suivante, basée sur la détection de segments, introduit de nombreux PF parasites comme cela est démontré dans [19]. La procédure utilisée pour détecter des PF le long d'une ligne candidate sélectionne aléatoirement un sous-ensemble de segments de l'image et calcule leur intersection avec la ligne. Un sous-ensemble optimal de PF est ensuite extrait des points de rencontre obtenus de manière à ce que leur consistance soit maximale tout en garantissant qu'aucun PF de l'ensemble final ne soit trop proche. Par conséquent, tout ensemble de segments se rencontrant accidentellement sur la LH est suceptible de degénérer un faux positif. De plus, la pré-sélection aléatoire de segments, dont le but est de réduire la combinatoire, peut empêcher qu'un PF représenté par peu de segments soit détecté.

Les méthodes [19] et [23] obtiennent des résultats très proches en termes de précision de la LH, mais la méthode *a-contrario* [19] est généralement plus précise que la méthode utilisant un CNN [23] lorsque des structures architecturales sont bien représentées dans l'image, mais moins robuste lorsque ce n'est pas le cas, comme avec la plupart des images du jeu de données HLW [22]. Ces différences (précision vs. robustesse) sont finalement assez conformes à ce que l'on observe communément lorsqu'on compare une méthode « classique » de vision par ordinateur à une méthode basée sur un apprentissage profond.

2.3 Estimation directe de la ligne d'horizon

Dans la continuité des travaux de [23], Workman et al. ont conçu un CNN permettant d'estimer directement la LH [22]. Leur réseau est construit sur l'architecture GoogleNet et estime simultanément la pente et le décalage de la LH. J. Lee et al. [15] ont pour leur part proposé un détecteur de *lignes sémantiques* : il s'agit de lignes de haut niveau sémantique incluant la LH, mais n'étant pas limitées à ce type de droite. Les lignes sémantiques sont détectées par classification et régression de lignes candidates. En ce qui concerne l'estimation de la LH, la méthode atteint les performances les plus élevées sur le jeu de données HLW [22].

3 Contributions

Notre première contribution est le réseau DIRECT qui infère directement à la fois la ligne d'horizon et un ensemble de PF, qui peuvent se trouver à l'intérieur ou à l'extérieur de l'image. Le réseau génère d'abord les distributions de probabilité des paramètres, puis une méthode *a-contrario* [6] est utilisée pour détecter les modes significatifs maximaux de la distribution de probabilité catégorielle correspondant aux coordonnées x des PFh. Pour la LH, nous extrayons la valeur ayant le meilleur score pour la pente et l'offset.

Dans l'idée de détecter un ensemble plus étendu de PF, nous avons étudié comment l'estimation de la carte des normales de la scène peut contribuer à la détection des PF. Les cartes de normales et les PF sont étroitement liés puisque les structures ayant des normales similaires contribuent à un même PF. Les PF ont été utilisés, par exemple, dans VPLNet [20] pour améliorer la qualité de la carte des normales estimée à partir d'images monoculaires. Nous proposons donc le réseau NM_VP (acronyme issu de l'anglais Vanishing Points from Normal Map), qui met en œuvre une stratégie d'apprentissage multi-tâches pour la détection des LH et des PF, la carte des normales étant une tâche auxiliaire. Bien que les ensembles de PF récupérés par les deux réseaux se recouvrent partiellement, le nombre de PF détectés augmente sensiblement grâce à la fusion de ces deux réseaux.

4 Architecture des réseaux

Dans nos deux réseaux, la LH est paramétrée par sa pente $\theta \in [-\pi/2, \pi/2]$ et son offset r $\in [-\infty, +\infty]$, où θ est l'angle entre la LH et l'axe x de l'image et r est l'ordonnée de la LH. Les coordonnées x des PF sont en effet suffisantes pour les localiser sur la LH. L'espace infini est ramené à l'intervalle $[-\pi/2, \pi/2]$, en utilisant l'inverse de la fonction tangente, afin de traiter les PF infinis.Nous traitons la détection de la LH et des PF comme un problème de classification : nous convertissons la pente, l'offset et les coordonnées des PF en classes catégorielles indépendantes en divisant leurs domaines respectifs uniformément en $N = 100$ baquets. L'utilisation d'une approche par classification nous permet d'estimer une distribution de probabilité sur les LH et PF possibles. Bien que l'entraînement soit effectué avec un seul PF pour chaque image, plusieurs pics peuvent être obtenus au stade du test dans les scores de sortie, correspondant à plusieurs PF.

4.1 L'architecture DIRECT

Nous utilisons l'architecture ResNet50 [10] comme base pour notre CNN. Les couches de convolutions sont laissées intactes tandis que trois classifieurs softmax disjoints remplacent le classifieur original avec 100 sorties. Les deux premières branches estiment la pente et l'offset de la LH, tandis que la troisième branche prédit les coordonnées x des PFh sur la LH (Figure 1). Nous introduisons des vérités terrain (VT) de PF un par un au cours de la phase d'apprentissage, ce qui donne lieu à une distribution de Dirac. Nous utilisons l'entropie croisée catégorielle pour l'apprentissage et la fonction Softmax comme activation pour la dernière couche entièrement connectée de chaque sortie du réseau.

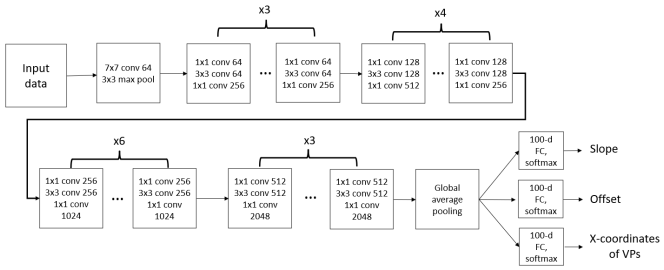

FIGURE 1 – Vue d'ensemble de l'architecture DIRECT pour la prédiction de la LH et des PFh.

4.2 L'architecture NM_VP

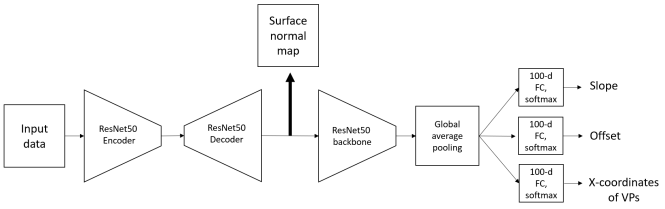

FIGURE 2 – Vue d'ensemble de l'architecture NM_VP pour la prédiction de la LH et des PFh.

Le modèle NM_VP est divisé en deux sous-modèles (Figure 2). Le premier sous-modèle estime la carte des normales à partir de l'image d'entrée. Nous utilisons un U-Net résiduel, qui tire profit à la fois des unités résiduelles profondes et de l'architecture U-Net [17]. Pour construire le modèle, nous utilisons des blocs résiduels au lieu des blocs neuronaux de base utilisés dans l'architecture U-Net [17]. Le deuxième sous-modèle est similaire à l'architecture DIRECT, il estime la LH et un ensemble de PF à partir de la carte des normales prédite.

NM_VP produit quatre sorties, la carte des normales, les paramètres de la LH et un ensemble de PF. Pour entraîner

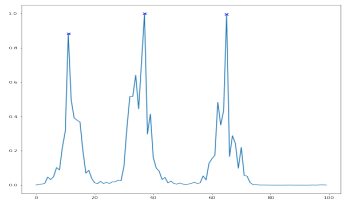

FIGURE 3 – Exemple de distribution de sortie obtenue avec le réseau NM_VP pour les coordonnées x des PF, avec trois pics détectés (croix) correspondant aux PF attendus.

notre modèle conjointement, nous utilisons l'entropie croisée catégorielle comme fonction de perte pour la détection des paramètres de la LH et des PF. La tâche d'estimation de la normale est traitée comme un problème de régression. Nous calculons la fonction de coût de Huber par pixel entre les normales prédites et les normales VT.
La fonction de coût globale est définie comme la somme pondérée des quatre pertes. Les expériences nous amènent à choisir (200,1,1,1) comme poids respectifs pour les fonctions de coût des cartes de normales, de la pente, du décalage et des PF.

4.3 Détection robuste des PF basée sur une approche *a-contrario*

Le cadre *a-contrario* [6] est utilisé dans une étape de post-traitement pour détecter les pics dans la distribution de probabilité catégorielle obtenue pour les coordonnées x des PF. Un intervalle de catégories (valeurs) est considéré comme significatif si la différence entre sa probabilité cumulée et celle d'une hypothèse de bruit (ici modélisée par une distribution uniforme) ne peut être due au hasard. Comme dans [6, 19], l'entropie relative H d'un intervalle est utilisée pour décider si celui-ci est significatif ($H > 0$). Cette estimation s'avère un peu moins précise que la queue de loi binomiale mais plus abordable en termes de calcul [6]. Enfin, un intervalle I est détecté comme un *mode significatif maximal* s'il s'agit d'un intervalle significatif (IS) et que pour tout IS $J \subset I$, $H(J) \leq H(I)$ et pour tout IS $J \supsetneq I$, $H(J) < H(I)$. Si un mode maximal significatif est constitué de plusieurs catégories, celle qui obtient le score le plus élevé est prise comme coordonnée x du PF (sinon l'unique catégorie déterminée est considérée comme coordonnée x). Les valeurs sont obtenues plus facilement dans le cas de l'offset et de la pente, puisqu'une seule valeur est attendue pour ces sorties : celle correspondant au score le plus élevé est donc retenue.
La figure 3 montre un exemple de distribution de sortie pour les coordonnées x des PF, avec trois pics détectés correspondant aux PF attendus. Les PF détectés sont affichés dans la figure 5, deuxième ligne. La méthode *a-contrario* est robuste aux pics locaux grâce à l'approche multi-échelle (tous les intervalles sont pris en compte) et son succès ne dépend pas d'un réglage délicat des paramètres. Un seul paramètre de seuil est utilisé, qui est universel ($H > 0$).

4.4 Fusion des deux ensembles de PF

Les réseaux DIRECT et NM_VP fournissent deux LH et deux ensembles de PF. Les différences entre les deux lignes obtenues étant très légères, nous utilisons systématiquement la LH récupérée avec DIRECT. Comme plusieurs PF sont communs aux deux ensembles de PF, nous devons les fusionner pour éviter autant que possible les PF dupliqués. Étant donné l'union des deux ensembles de PF, nous détectons d'abord ceux situés à une distance angulaire inférieure à un seuil donné (fixé à 5 degrés en pratique). Notre stratégie de fusion ne retient alors que celui qui présente l'entropie relative la plus élevée parmi un tel groupe.

5 Résultats

5.1 Données

Plusieurs jeux de données sont consacrés à l'évaluation de la ligne d'horizon (LH), mais seulement deux à notre connaissance contiennent la vérité terrain (notée PF VT) : York Urban [5], et Holicity [24] récemment publié. Malheureusement, York Urban ne contient que 102 images de scènes extérieures et intérieures et au plus deux PF par image, étiquetés à la main. Nous avons donc utilisé uniquement Holicity dans nos expériences. Cette base se compose de 54354 images couvrant le centre-ville de Londres et fournit des étiquettes pour différentes structures 3D telles que les carte des normales et les PF.
Afin d'avoir une évaluation réaliste, parfaitement indépendante des données d'apprentissage, nous avons isolé, parmi les 54 quartiers considérées, environ 10% des quartiers, soit 4205 images, qui ne sont jamais considérés dans le jeu de données d'entraînement. Cette ensemble est nommé UNSEEN. Le reste des données a été divisé en trois parties, 80% (37910 images) pour l'entraînement, 10% (4239 images) pour la validation, et 10% (4239 images, appelées SEEN) pour les tests. Les données de validation pour SEEN correspondent ainsi à des images non utilisées pour l'apprentissage mais plus ou moins proches de certaines images d'apprentissage. Les images d'entraînement ont été augmentées en appliquant des rotations dont les angles suivent une distribution normale avec un écart-type égal à $10°$, ce qui a conduit à un ensemble de 75820 images d'entraînement.
L'ensemble de données Holicity souffre en fait de certaines incohérences. Les cartes des normales VT ont été construites à partir de modèles CAO issus de SIG, ce qui conduit à des incohérences lorsque les bâtiments sont occultés (par exemple par des arbres ou des voitures) dans les images. Les images présentant une forte incohérence ont été écartées de l'apprentissage de la façon suivante : un U-net a été entraîné sur Holicity pour estimer la carte des normales de surface. Les images pour lesquelles la différence entre la normale VT et la normale estimée était supérieure à $30°$ en moyenne ont été retirées de l'étape d'entraînement.

Dans Holicity, les PF horizontaux de la vérité terrain ont également été générés à partir des modèles CAO 3D. Par conséquent, de nombreux PF horizontaux correspondent à des détails architecturaux de bâtiments (souvent sur les balcons), qui sont à peine perceptibles sur les images. Le nombre de PF VT est donc sensiblement plus important que celui des PF détectés dans les images.

5.2 Métriques

L'erreur sur la ligne d'horizon est classiquement définie comme la distance maximale entre ligne d'horizon estimée et vérité terrain, calculée à l'intérieur des limites de l'image, divisée par la hauteur de l'image. Le tableau 1, première ligne, indique le pourcentage de la surface sous l'histogramme cumulé de cette erreur (AUC) dans le sous-ensemble [0, 0,25]x[0, 1], à la fois pour les parties SEEN et UNSEEN de Holicity.

La précision angulaire (**AA**) définie dans [25] est utilisée pour quantifier la qualité des PF estimés. Pour chaque PF, nous calculons l'angle entre le PF VT et le PF prédit. La précision angulaire (AA) définie dans [25] correspond à l'histogramme cumulatif des erreurs angulaires inférieures à un seuil prédéfini (5 ou 10° dans nos expériences). La valeur AA est définie comme l'aire sous cette courbe, divisée par le seuil.

Alors que la métrique AA permet de quantifier la précision des PF prédits, une seconde métrique appelée **AA-R** (en référence au terme *rappel*) est introduite afin de quantifier le ratio de PF VT détectés par les différentes méthodes. De manière symétrique, nous calculons pour chaque PF VT l'erreur par rapport au PF prédit le plus proche. L'AA-R est défini comme l'aire sous l'histogramme cumulé de ces erreurs, divisée par le seuil.

5.3 Résultats quantitatifs

La ligne d'horizon est significativement mieux estimée avec les méthodes DIRECT et NM_VP qu'avec la méthode *a-contrario* (LB) [19] (88% contre 67%). Les valeurs AA et AA-R sont fournies dans le tableau 1 pour deux seuils (5 et 10°) à la fois pour les ensembles de données SEEN et UNSEEN. Les valeurs globalement assez faibles de ces résultats (environ 50 %) sont dues au fait que les PF VT sont calculés à partir de modèles de ville CAO. Par conséquent, certains PF VT ne peuvent pas être détectés dans les images, soit parce qu'ils sont occultés par des objets réels des images non pris en compte dans les modèles CAO (arbres, voitures... et parfois de nouveaux bâtiments), soit parce qu'ils sont dus à des détails architecturaux 3D (balcons). Au contraire, certains PF sont manquants lorsque le modèle CAO est incomplet (voir ci-dessous).

Pour SEEN, 4604 PF ont été détectés par DIRECT et 4124 par NM_VP. Parmi eux, 3706 (contre 3664) ont un AA inférieur à 10° pour DIRECT (contre NM_VP). Bien que de nombreux PF soient détectés par les deux méthodes, 490 ont été détectés par DIRECT et non par NM_VP, tandis que 453 PF ont été détectés par NM_VP et non par DIRECT. Cela prouve l'intérêt de la stratégie de fusion qui permet de tirer profit des PF structurels détectés par la carte des normales. Par rapport à la méthode de référence [19], notre méthode démontre une meilleure précision des PF détectés (voir les valeurs AA qui sont plus grandes avec les trois réseaux). Ce tableau montre également que la fusion des deux réseaux permet de détecter un plus grand nombre de PF : la méthode de fusion obtient en effet les valeurs AA-R les plus importantes. Il faut noter que ce gain se fait au prix d'une précision (valeurs AA) légèrement inférieure à celle des méthodes DIRECT et NM_VP. En fait, obtenir un maximum de candidats PF peut être intéressant pour les applications où une méthode auxiliaire est disponible pour filtrer les faux positifs, par exemple pour détecter les PF de Manhattan.

SEEN	DIRECT	NM_VP	FUSION	LB [19]
AUC	**88.13%**	88.11%	88.11%	66.93%
AA (5°)	33.75%	**40.63%**	32.75%	22.74%
AA-R (5°)	23.23%	25.44%	**29.13%**	22.93%
AA(10°)	57.45%	**63.46%**	54.81%	40.12%
AA-R (10°)	37.94%	39.79%	**44.55%**	41.1%
UNSEEN	DIRECT	NM_VP	FUSION	LB [19]
AUC	88.56%	**88.62%**	88.56%	67.25%
AA (5°)	33.15%	**39.98%**	31.83%	22.29%
AA-R(5°)	24.01%	25.54%	**29.25%**	22.89%
AA (10°)	54.79%	**62.06%**	53.15%	40.36%
AA-R(10°)	38.99%	39.59%	**44.6%**	42.22%

TABLE 1 – Précision LH, précision PF et rappel avec les méthodes DIRECT, NM_VP, FUSION et [19] de Simon et al. pour les ensembles de données SEEN et UNSEEN.

1 PF VT	2 PF VT	3 PF VT	4 PF VT
51.71%	33.85%	30.46%	29.61%
74.21 %	51.06 %	46.26%	44.98%

TABLE 2 – Ratio d'images pour lesquelles plus de $N = 1..4$ PF VT sont détectés. Première ligne, avec un seuil de précision de 5°. Deuxième ligne, avec un seuil de 10°.

La table 1 montre aussi que les résultats sur SEEN et UNSEEN sont très proches, ce qui prouve la très bonne capacité de généralisation de la méthode.

Afin d'avoir une meilleure idée de la capacité de la méthode à détecter plusieurs PF, en particulier de $N = 1$ à $N = 4$, nous fournissons également quelques métriques supplémentaires. Étant donné N, nous considérons toutes les images avec plus de N PF VT. Une correspondance au plus proche voisin avec les PF prédits est réalisée pour chaque PF VT. Les erreurs angulaires correspondant aux N PF détectés les plus proches sont considérées pour le calcul de AA. Le ratio des images pour lesquelles plus de N PF sont en dessous du seuil est ensuite calculé. Les résultats sont fournis dans le tableau 2.

5.4 Résultats qualitatifs

Des exemples de résultats sont donnés dans cette section et comparés à la méthode LB [19]. Dans tous les graphiques,

les résultats de la méthode LB sont représentés en jaune, ceux de la méthode DIRECT en rouge et ceux de la méthode NM_VP en bleu. Les données VT sont représentées en vert. Les PF estimés sont indiqués par une croix. Ceux qui sont proches d'un PF VT avec une précision inférieure à 5° (resp. entre 5 et 10°) sont entourés d'un cercle (resp. un carré). Les points de fuite sélectionnés par la méthode FUSION sont indiqués en noir. La figure 5 montre des exemples de détections réussies de PF dans des images difficiles. Les carte des normales VT et estimée sont d'abord présentées. La troisième image montre la ligne d'horizon et les PF estimés obtenus avec les méthodes LB, DIRECT et NM_VP. La quatrième image montre en noir le résultat de la méthode FUSION, toujours avec les résultats de la méthode LB. Sur la première ligne, des arbres occultent une des façades. Puisque la normale estimée est entraînée à partir d'images générées par des modèles CAO sans structures occlusives, il est intéressant de noter que la carte des normales n'est pas perturbée par les arbres. La LH obtenue avec DIRECT est très proche de la VT. En revanche, les troncs d'arbres imparfaitement verticaux corrompent l'estimation du zénith avec la méthode LB, et donc la pente de la HL, perpendiculaire à la ligne zénithale (Figure 4). Les deux PF VT (en vert) sont obtenus avec une précision de 5°, l'un avec la méthode DIRECT et l'autre avec la méthode NM_VP. Au contraire, un seul PF est détecté avec une précision de 10°, tandis que l'autre à droite est erroné.

La deuxième ligne est une autre image complexe avec trois PF VT. La carte des normales estimée est d'une bonne précision. Deux PF VT sont détectés à la fois par la méthode DIRECT et NM_VP, tandis que le troisième est détecté par la méthode NM_VP. À l'exception du point bleu le plus à gauche, tous les PF détectés ont une précision inférieure à 5°. Comme le processus de fusion est basé sur l'entropie relative des modes significatifs maximaux, les PF retenus correspondent aux PF les plus précis. Dans ce cas, la méthode LB ne détecte que 2 PF dont la précision est moins bonne que celle obtenue avec notre méthode.

L'image de la troisième ligne comporte également trois PF VT. La ligne d'horizon est bien détectée avec DIRECT mais n'est pas très précis avec LB. Les trois PF VT sont détectés soit par la méthode DIRECT soit par la méthode NM_VP, mais deux d'entre eux ont une précision angulaire de l'ordre de 5 à 10°. La méthode LB ne détecte dans ce cas que deux PF, dont l'un a une précision angulaire dans la plage [5, 10°].

Les exemples proposés dans la Figure 6 illustrent le fait que la précision de notre méthode est généralement meilleure que la méthode LB. La LH est correctement estimée dans la première image par la méthode LB et par nos méthodes. Un PF VT est détecté à la fois par LB et par nos méthodes, mais la précision de notre méthode est nettement meilleure. Cet exemple illustre également que certains PF VT peuvent être difficilement détectés à partir d'images. Le PF VT le plus à gauche (indiqué par une flèche) est lié à la façade sans fenêtre sur la partie droite de l'image (entourée d'un

FIGURE 4 – Segments contribuant à l'estimation des PF avec la méthode LB (une couleur par PF). Les troncs d'arbres sont pris en compte dans l'estimation du zenith. Mais, n'étant pas parfaitement verticaux, ils corrompent l'estimation du zénith et donc de la ligne d'horizon.

cercle), largement occultée par un arbre. La carte des normales estimée n'est pas précise, et aucune des méthodes ne parvient pas à détecter le PF correspondant.
La deuxième ligne de la figure illustre un cas où la LB et notre méthode détectent toutes deux deux des PF VT. Dans le cas le plus à gauche, le PF obtenu avec LB est plus précis, alors que dans le cas le plus à droite, celui détecté par notre méthode est le plus précis.
Dans la troisième ligne, la LH est extraite avec une meilleure précision avec notre méthode. Bien que l'environnement, avec de nombreux segments, soit théoriquement mieux adapté à la méthode LB, la détection LB du PF le plus à droite est moins précise qu'avec notre méthode. Le fait que les PF détectés soient contraints d'appartenir à la LH explique ce manque de précision. Il faut également noter qu'un autre PF, qui n'était pas considéré comme un PF VT, est détecté par la méthode LB. Il correspond aux bâtiments de l'arrière-plan qui manquent dans le modèle CAO de la vérité terrain (voir la carte des normales VT). Bien que ces bâtiments apparaissent dans la carte des normales estimée, le PF correspondant n'est pas détecté par NM_VP, probablement en raison de leur petite taille.

6 Discussion et conclusion

Nous avons présenté dans ce travail la fusion de deux méthodes pour estimer les paramètres de la ligne d'horizon et détecter les points de fuite horizontaux à l'aide de réseaux de neurones convolutifs. La méthode DIRECT estime les paramètres directement à partir d'une image RVB, tandis que l'approche NM_VP estime conjointement la carte des normales de surface ainsi que la ligne d'horizon et les points de fuite. Nous avons montré dans ce travail l'importance de l'estimation de la carte des normales de surface comme tâche auxiliaire pour une meilleure détection des points de fuite.
La performance des méthodes basées sur les lignes dans

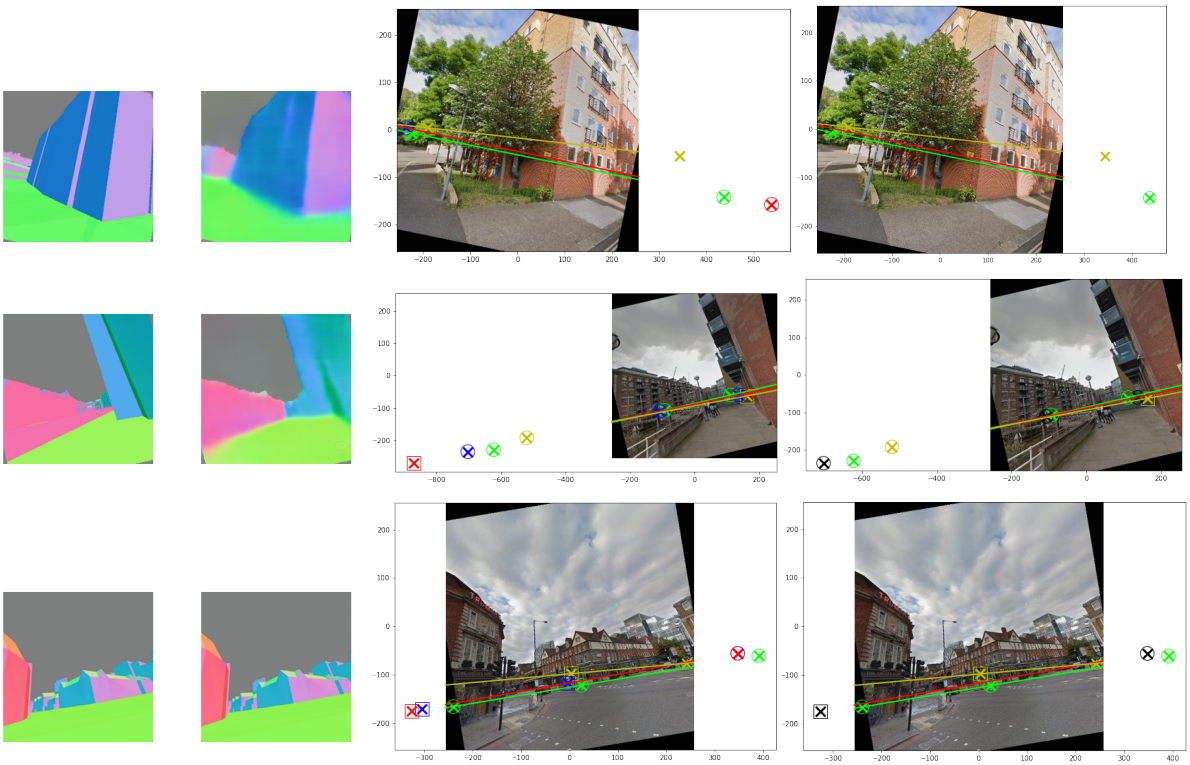

FIGURE 5 – Exemples d'images où notre méthode détecte un plus grand ensemble de points de fuite que la méthode LB, avec une meilleure précision en général. Code couleur : jaune pour LB, rouge pour DIRECT, bleu pour NM_VP, vert pour les données GT. Les PF après fusion sont représentés en noir.

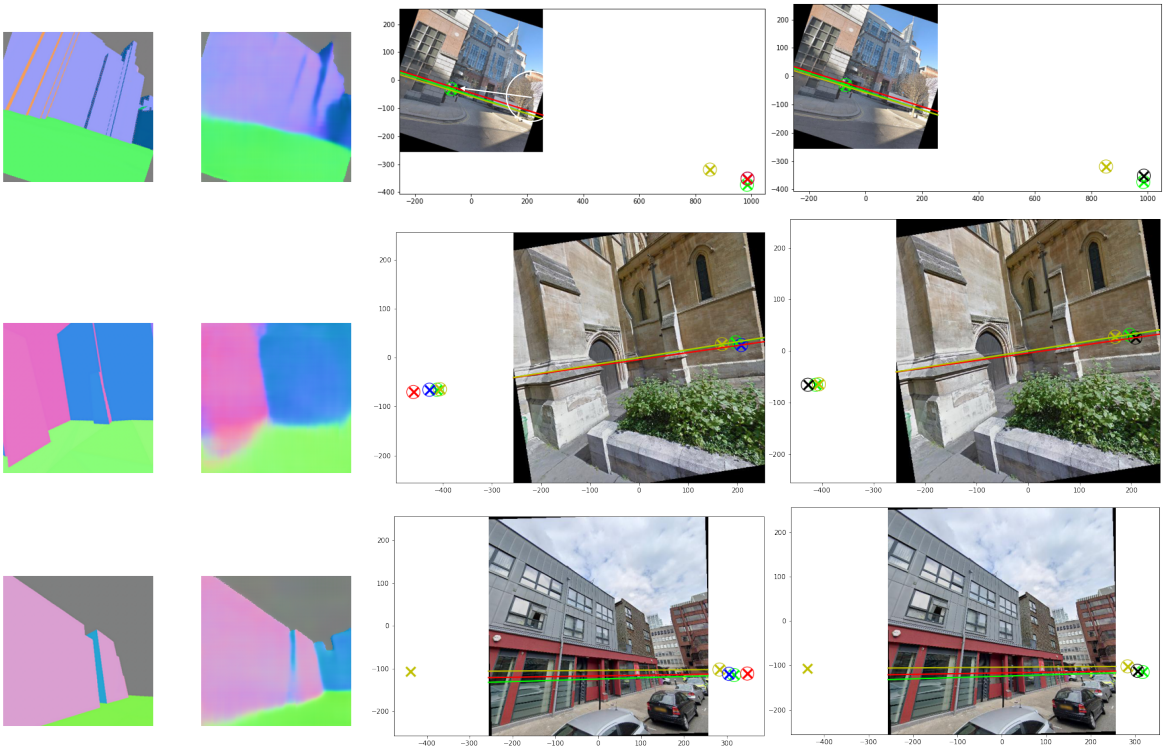

FIGURE 6 – Comparaison de la précision de détection des points de fuite. De gauche à droite : carte des normales (GT), carte normal estimée, points de fuite détectés et VT, points de fuite détectés après l'étape de fusion.

la détection des PF dépend de la pertinence des segments détectés. Dans les environnements où les segments sont nombreux, les méthodes conventionnelles telles que LB peuvent extraire les PF avec précision et souvent avec une précision légèrement meilleure que les méthodes basées sur l'apprentissage profond. Dans les autres situations, notre méthode détecte généralement un ensemble plus étendu de PF VT avec une meilleure précision.
Le système actuel fusionne deux réseaux par un processus robuste de raisonnement *a-contrario*. On pourrait se demander si un réseau unique englobant à la fois l'image RVB et la détection basée sur la carte des normales pourrait être envisagé pour éviter le test de fusion. Nous avons envisagé un tel réseau mais n'avons pas réussi à obtenir de meilleurs résultats que ceux obtenus avec deux réseaux distincts et une étape de fusion supplémentaire.

Références

[1] J.-C Bazin, Yongduek Seo, Cédric Demonceaux, Pascal Vasseur, Katsushi Ikeuchi, In-So Kweon, and Marc Pollefeys. Globally optimal line clustering and vanishing point estimation in manhattan world. In *CVPR*, pages 638–645, 2012.

[2] R. Bolles and M. Fischler. A ransac-based approach to model fitting and its application to finding cylinders in range data. In *IJCAI*, 1981.

[3] Ali Borji. Vanishing point detection with convolutional neural networks. *CoRR*, abs/1609.00967, 2016.

[4] William Chen and Bernard C. Jiang. 3-d camera calibration using vanishing point concept. *Pattern Recognition*, 24(1) :57–67, 1991.

[5] Patrick Denis, James Elder, and Francisco Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *ECCV*, pages 197–210, 2008.

[6] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis : A Probabilistic Approach*. Springer Publishing Company, Incorporated, 1st edition, 2007.

[7] Rafael Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd : A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32 :722–32, 04 2010.

[8] E. Guillou, Daniel Méneveaux, E. Maisel, and K. Bouatouch. Using vanishing points for camera calibration and coarse 3d reconstruction from a single image. *The Visual Computer*, 16 :396–410, 2000.

[9] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] P. Hough. Machine analysis of bubble chamber pictures. 1959.

[12] Yonghoon Ji, A. Yamashita, and H. Asama. Rgb-d slam using vanishing point and door plate information in corridor environment. *Intelligent Service Robotics*, 8 :105–114, 2015.

[13] Florian Kluger, Hanno Ackermann, Michael Ying Yang, and Bodo Rosenhahn. Deep learning for vanishing point detection using an inverse gnomonic projection. *CoRR*, abs/1707.02427, 2017.

[14] Jana Košecká and Wei Zhang. Video compass. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *ECCV*, pages 476–490, 2002.

[15] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Semantic line detection and its applications. In *ICCV*, pages 3249–3257, 2017.

[16] Yin-Bo Liu, Ming Zeng, and Qing-Hao Meng. Dvpnet : A network for real-time dominant vanishing point detection in natural scenes. *CoRR*, abs/2006.05407, 2020.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[18] Gilles Simon, Antoine Fond, and Marie-Odile Berger. A Simple and Effective Method to Detect Orthogonal Vanishing Points in Uncalibrated Images of Man-Made Environments. In *Eurographics 2016*, 2016.

[19] Gilles Simon, Antoine Fond, and Marie-Odile Berger. A-Contrario Horizon-First Vanishing Point Detection Using Second-Order Grouping Laws. In *ECCV*, pages 323–338, September 2018.

[20] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet : Deep single view normal estimation with vanishing points and lines. In *CVPR*, pages 686–695, 2020.

[21] Horst Wildenauer and Allan Hanbury. Robust camera self-calibration from monocular images of manhattan worlds. In *CVPR*, pages 2831–2838, 2012.

[22] Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild. *CoRR*, abs/1604.02129, 2016.

[23] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manhattan world. *CoRR*, abs/1608.05684, 2016.

[24] Yichao Zhou, Jingwei Huang, Xili Dai, Linjie Luo, Zhili Chen, and Yi Ma. HoliCity : A city-scale data platform for learning holistic 3D structures. *CoRR*, 2020. arXiv :2008.03286 [cs.CV].

[25] Yichao Zhou, Haozhi Qi, Jingwei Huang, and Yi Ma. NeurVPS : Neural Vanishing Point Scanning via Conic Convolution. *CoRR*, abs/1910.06316, 2019.