Estimation de profondeur à partir d'une seule image avec un réseau adversaire

Marcela CARVALHO[1,2], Bertrand LE SAUX[1], Pauline TROUVÉ-PELOUX[1], Andrés ALMANSA[2], Frédéric CHAMPAGNAT[1]

[1]ONERA, *The French Aerospace Lab*
F-91761 Palaiseau, France

[2]Télécom ParisTech
F-75013 Paris, France

Marcela.Carvalho@onera.fr, Bertrand.Le_Saux@onera.fr
Pauline.Trouve@onera.fr, Andres.Almansa@telecom-paristech.fr
Frederic.Champagnat@onera.fr

Résumé – Depuis peu, les réseaux génératifs adversaires (*Generative Adversarial Network*, GAN), originalement présentés par Goodfellow et al. [1], ont introduit un nouveau concept de génération d'images proches de la vérité terrain avec un apprentissage non-supervisé. Dans ce contexte, nous proposons l'exploration des GANs pour la prédiction de la carte de profondeur d'une scène à partir d'une seule image. La méthode présentée permet de nous approcher des performances de l'état de l'art avec une architecture plus simple que des méthodes existantes.

Abstract – Recently, the Generative Adversarial Networks (GANs), originally introduced by Goodfellow et al. [1], introduced us to a new concept of generation of very realistic images with unsupervised learning. In this context, we propose the explore the adversarial networks for the prediction of the depth map of a scene from a single image. The presented method allows us to get close to state-of-art results with a much simpler architecture than the existing ones.

1 Introduction

L'estimation de profondeur est aujourd'hui une problématique majeure de la vision par ordinateur. Les approches standard reposent sur la vision stéréoscopique, la lumière structurée ou bien la structure à partir du mouvement. Cependant ces techniques ont souvent des limitations selon l'environnement (soleil, texture) ou ont besoin de plusieurs vues de la scène.

Ces contraintes ont donc conduit plusieurs travaux [2, 3, 4] à exploiter les aspects géométriques d'une scène à partir d'un seul point de vue (une seule image) pour estimer la structure 3D à l'aide de réseaux de neurones convolutifs. Ceux-ci sont en général composés de couches de convolutions suivies ou pas des couches entièrement connectées, sont multi-échelle et optimisent une régression sur la carte de profondeur de référence (cf. Fig. 1). Le principal défi consiste alors à définir la fonction de perte adéquate pour la régression.

Les réseaux génératifs adversaires sont capables de générer des images particulièrement réalistes à partir d'un entraînement non-supervisé. Avec les *conditional GANs* (cGANs), des conditions supplémentaires pour cet apprentissage ont été proposées dans [5, 6] afin de contraindre encore plus la vraisemblance des sorties produites. Le principal avantage est la définition implicite de la fonction de perte par apprentissage d'une métrique dans l'espace des images. Dans cet article, nous proposons d'utiliser un cGAN pour l'estimation de profondeur.

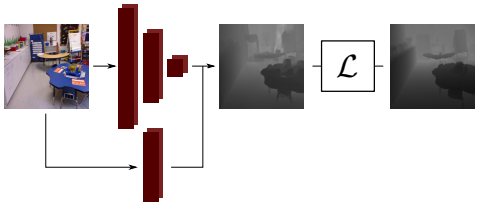

FIGURE 1 – Architecture de CNN multi-échelle pour l'estimation de profondeur [8].

Notre approche permet de générer une carte de profondeur à partir d'une seule image sans avoir besoin d'une fonction de perte spécifique pour cette tâche. Des comparaisons qualitatives et quantitatives de performance de notre méthode sur la base de données NYUv2 [7] montrent que notre méthode permet de nous approcher des performances de l'état de l'art [2, 3].

2 Contexte

Une des premières solutions au problème d'estimation de profondeur monoculaire a été présentée par Saxena *et al.* [9] qui formulent l'estimation de la profondeur comme un problème de champ aléatoire de Markov (*Markov Random Field*, MRF) avec des images alignées horizontalement.

Les travaux les plus récents se basent sur l'apprentissage des réseaux convolutifs profonds. Eigen *et al.* [8, 2] proposent une

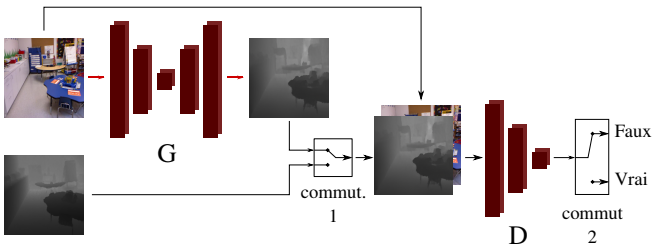

FIGURE 2 – Architecture proposée pour la génération de cartes de profondeur à partir d'une image.

architecture multi-échelle capable d'extraire des informations globales et locales de la scène pour l'estimation de la carte de profondeur. Liu et al. [3] explorent un réseau convolutif structuré associé à un champ aléatoire conditionnel continu (*Conditional Random Field*, CRF) avec des superpixels. Dans [10], Cao *et al.* utilisent le CRF à la sortie d'un réseau profond résiduel afin d'améliorer la fiabilité des prédictions par rapport à l'information des pixels voisins.

Les techniques précédentes présentent des architectures à plusieurs échelles [2, 3], voire des étapes de post-traitement [10] pour améliorer la cohérence de la profondeur. Ils utilisent également des fonctions de perte spécifiques pour la tâche proposée, ce qui peut être très coûteux et difficile pour effectuer une régression ou une classification suffisamment robuste. La principale difficulté consiste à mesurer correctement la similitude entre la distribution des données réelles, $p_{donnees}$ et la distribution des données générées, p_{gen}, pour améliorer la prédiction.

Par ailleurs, des méthodes d'apprentissage récentes [1, 11] proposent d'utiliser un GAN dans lequel la distance entre la distribution des données réelle et celle des données générées est estimée pendant l'apprentissage. Ceci permet d'éviter la définition de nouvelles fonctions de perte spécifiques au problème.

Radford *et al.* [5] présentent la première utilisation efficace d'un GAN avec des réseaux convolutifs profonds. Mirza *et al.* [11] ont introduit les GANs conditionnels en ajoutant des labels à l'entrée du générateur (G) et du discriminateur (D). L'image générée par G, ainsi que la sortie de D sont ainsi conditionnées à ces informations discrètes. Isola *et al.* [6] proposent de conditionner l'image générée à une autre image.

La simplification apportée par les réseaux génératifs ainsi que la qualité des résultats générés par [1, 5, 11, 6] nous ont conduit à explorer cette approche pour une application d'estimation de profondeur monoculaire.

3 Méthode

Nous proposons d'utiliser les cGANs pour résoudre le problème de génération de carte de profondeur à partir d'une seule image. Contrairement à [2, 3, 10] l'estimation de profondeur est réalisée sans avoir à définir une fonction de coût.

L'idée de base est de former deux réseaux qui jouent des rôles opposés : le générateur (G) et le discriminateur (D). Le premier tente de créer des images réalistes pour tromper le second réseau, chargé de vérifier si l'échantillon provient de la base de données (échantillon réel) ou du générateur (faux échantillon). Le schéma de principe de notre approche est présenté dans la figure 2. Comme [6], nous utilisons aussi une image à l'entrée de G pour conditionner la génération de carte de profondeur. La première étape d'optimisation consiste à entraîner les poids de D. Pour chaque échantillon $y \in p_{donnees}(y)$ (commutateur 1 sur le terminal inférieur), D doit indiquer s'il s'agit d'une vraie carte de profondeur (commutateur 2 sur le terminal inférieur). En revanche, pour chaque image générée, $G(x, z)$, où $x \in p_{donnees}(x)$ (commutateur 1 sur le terminal supérieur), D doit indiquer que cet échantillon est faux (commutateur 2 sur le terminal supérieur). La deuxième étape consiste à entraîner G. Cette fois ci, la sortie de G passe par D comme un vrai échantillon. Par rétro-propagation, D actualise les paramètres de G pour que la distribution des images générées s'approche de $p_{donnees}(x)$. Formellement l'entraînement de D et G consiste à optimiser une fonction minmax :

$$\begin{aligned}\mathcal{L} \ =&\mathbb{E}_{x,y\sim p_{donnees}(x,y)}[log\, D(x,y)] \\ &+\mathbb{E}_{x\sim p_{donnees}(x)}[1-log\, D(x,G(x))].\end{aligned} \tag{1}$$

Nous évaluons l'adéquation de l'architecture d'un GAN aux données de type image et carte de profondeur. Le générateur utilisé est un réseau complètement convolutif de la forme codeur-décodeur, similaire à U-net [12]. Cette architecture permet de coder l'image d'entrée avec un haut degré d'abstraction, puis de la décoder dans l'espace associé à la sortie désirée. Ce générateur sera la partie conservée en test. Afin de lisser la prédiction, un terme d'erreur L1 est ajouté à la sortie de G. Le discriminateur est entièrement convolutif et optimise la perte définie en 1. Dans un premier temps, nous reprenons les paramètres indiqués dans [6]. Puis, pour bénéficier de l'apport d'information d'un ensemble d'entraînement plus important (Officielle 170k), nous utilisons un modèle plus complexe pour le générateur afin de mieux approcher la fonction de transfert image vers profondeur. L'optimisation est basée sur la descente de gradient stochastique (SGD) avec Adam [13].

4 Expérimentations

TABLE 1 – Métriques pour évaluer la performance de la méthode.

Erreur absolue relative (abs)	$\frac{1}{N}\sum_{i=0}^{N}\frac{\lvert d_i-\hat{d}_i\rvert}{d_i}$
RMS linéaire (rmsl)	$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(d_i-\hat{d}_i)^2}$
RMS logarithmique (rmslog)	$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(log(d_i)-log(\hat{d}_i))^2}$
Précision avec seuil (thr)	$max(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i})=\delta<thr$

Pour tester notre approche sur des images réelles, nous utilisons la base de données NYU-depth v2 [7]. Elle comprend des séquences d'images RGB-D de la Kinect de Microsoft, en intérieur. Dans les essais, nous utilisons pour l'entraînement soit

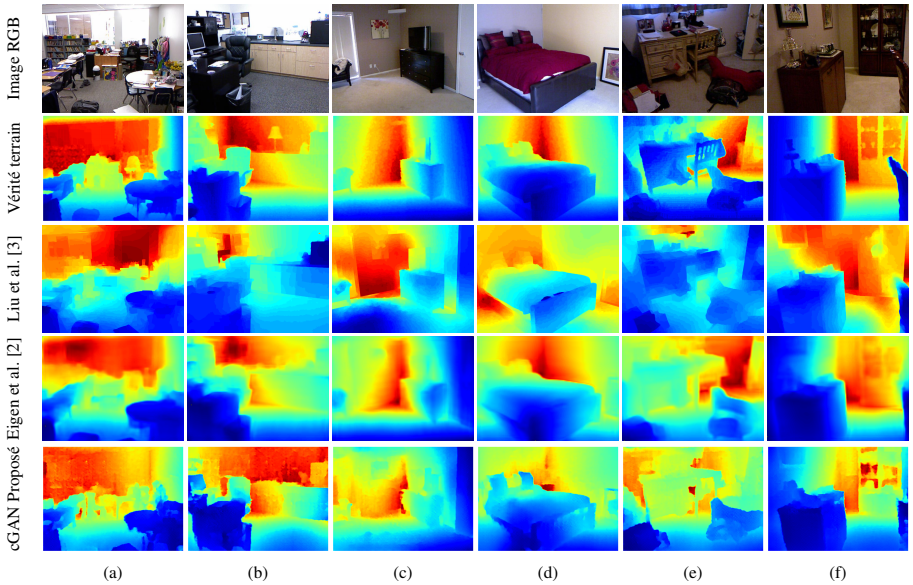

FIGURE 3 – Comparaison qualitative des différentes approches : **(a-d)** estimations réussies, **(e-f)** exemples difficiles.

la partition originale avec 795 images (Officielle 795), soit une partition augmentée avec plus d'images et le même jeu de test (Officielle 6.5k), soit une autre partition augmentée avec toutes les images de la base de données (Officielle 170k). Pour le test, afin de permettre des comparaisons avec des travaux existants, nous utilisons les 654 images officielles. Toutes les cartes de profondeur ont été recalées sur les images RGB.

Les performances d'estimation de profondeur à partir de l'image RGB sont évaluées avec les métriques de [8, 3] présentées dans la Table 1. d_i et $\hat{d}_i$ sont la vérité terrain et la prédiction, respectivement, et N est le nombre total de pixels par image.

5 Résultats

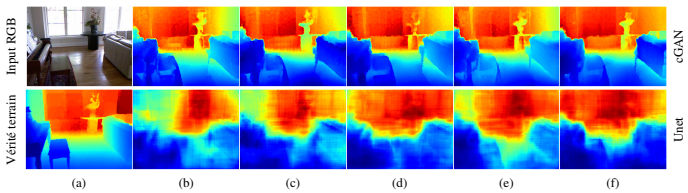

FIGURE 4 – Comparaison de la prédiction de profondeur à différentes époques avec cGAN couplé à une norme L1 (en haut) et un U-net avec une fonction de perte L1 (en bas). **(a)** Image couleur et profondeur associée. **(b-f)** Prédiction après entraînement de durées 20 à 100 époques (pas de 20).

La figure 3 compare des résultats d'estimation de notre cGAN avec deux approches de l'état de l'art [8, 3]. L'approche de Liu *et al.* présente des effets de bloc dus aux superpixels. L'approche d'Eigen *et al.* est plus lissée mais a une résolution moins importante que les images en entrée. Sur les exemples 3-a à 3-d, le cGAN produit des profondeurs réalistes qui correspondent à la vérité terrain. Les petits objets à des distances intermédiaires sont clairement séparés. Cependant, pour des scènes plus confuses, avec des dynamiques restreintes (que des objets proches 3-e) l'estimation du cGAN a tendance à repartir uniformément les prédictions sur la plage de valeurs possibles.

Ces résultats se retrouvent dans les tableaux 2 et 3 qui présente des résultats moyennés sur le jeu de test. Notre approche dépasse les performances de [9], mais fait moins bien que les autres approches : les mauvais résultats sur certaines images dégradent la performance moyenne.

TABLE 2 – Comparaison quantitative de notre méthode avec l'état de l'art sur le dataset NYUv2.

Méthode	Performance		
	abs	rmsl	rmslog
Saxena [9]	0.349	1.214	-
Liu [3]	0.230	0.824	-
Wang [4]	0.220	0.745	0.262
Eigen [2]	0.158	0.641	0.214
cGAN*+L1 170k	0.331	0.906	0.349

TABLE 3 – Comparaison quantitative de notre méthode avec l'état de l'art sur le dataset NYUv2.

Méthode	Précision		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Saxena[9]	44.7%	74.5%	89.7%
Liu [3]	61.4%	88.3%	97.1%
Wang [4]	60.5%	89.0%	97.0%
Eigen [2]	76.9%	95.0%	98.8%
cGAN*+L1 170k	53.0%	82.5%	93.4%

Dans le tableau 4, nous présentons les résultats avec les trois partitions présentées dans la Section 4. De plus, nous menons des études par ablation de différents modules du GAN : cGAN+L1

TABLE 4 – Comparaison de la performance de la méthode proposée avec différentes configurations sur le dataset NYUv2.

Partition	Méthode	Performance			Accuracy		
		abs	rmsl	rmslog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Officielle 795	cGAN	0.494	2.003	0.	42.4%	67.4%	82.4%
	GAN+L1	1.051	2.950	0.755	18.6%	37.8%	57.0%
	cGAN+L1	1.000	2.804	0.734	19.6%	39.4%	59.2%
Officielle 6.5k	cGAN	0.430	1.266	0.405	42.7%	72.0%	87.3%
	GAN+L1	0.414	1.057	0.439	41.8%	75.4%	90.3 %
	cGAN+L1	0.525	1.061	0.399	42.4%	75.9%	91.0%
Officielle 170k	cGAN+L1	0.371	0.999	0.378	46.5%	78.8%	92.1%
	cGAN*+L1	**0.331**	**0.906**	**0.349**	**53.0%**	**82.5%**	**93.4%**

est la méthode de [6], cGAN est la variante amputée du terme L1 dans la fonction de perte, GAN+L1 est la variante amputée de la conditionnelle (le discriminateur ne voit pas l'image RGB). CGAN*+L1 est la version modifiée de [6] avec un G plus profond (plus de couches).

Les GANs ont une tendance au surapprentissage reconnue [11] donc augmenter le nombre d'exemples permet de réduire cet effet et d'améliorer les performances. Avec peu d'exemples, la perte L1 se focalise trop sur des détails et dégrade les résultats. En revanche, avec suffisamment d'exemples, elle joue correctement son rôle de régularisation. Des expériences avec plus de données permettent d'améliorer encore les performances, ainsi que l'ajout des couches supplémentaires dans le générateur.

La figure 4 permet de comparer la vitesse d'entraînement d'un réseau encodeur-décodeur simple (U-net) avec une fonction de perte L1 et la variante avec une perte englobant le résultat d'un réseau discriminatif (cGAN+L1). D'une part, le cGAN converge beaucoup plus vite et obtient des bons résultats dès 20 époques alors que le U-net seul se stabilise après 60 ou 80 époques. D'autre part, la carte de profondeur obtenue par le cGAN est beaucoup plus précise et permet de distinguer des objets situés à une distance intermédiaire.

6 Conclusions

Nous avons présenté une nouvelle approche pour l'estimation de profondeur à partir d'une seule image RGB, basée sur des réseaux génératifs adversaires. Elle obtient des performances comparables à l'état de l'art sur la base NYUv2 et obtient dans de nombreux cas des images plus réalistes que les réseaux profonds multi-échelle concurrents. Ces résultats sont prometteurs mais pâtissent du surapprentissage quand le jeu d'entraînement est trop restreint. Nous chercherons dans nos travaux futurs à corriger ce phénomène, soit en travaillant avec une fonction de perte dédiée, soit en explorant des architectures multi-échelle, ou en essayant d'autres modèles de GAN.

Références

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NIPS*, 2014.

[2] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture," *ICCV*, 2015.

[3] F. Liu, C. Shen, G. Lin, and I. D. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *TPAMI*, 2015.

[4] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *CVPR*, 2015.

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv :1511.06434*, 2015.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv :1611.07004*, 2016.

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

[8] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NIPS*, 2014.

[9] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning Depth from Single Monocular Images," *NIPS*, 2006.

[10] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *arXiv preprint arXiv :1605.02305*, 2016.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv :1411.1784*, 2014.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net : Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

[13] D. Kingma and J. Ba, "Adam : A method for stochastic optimization," *arXiv preprint arXiv :1412.6980*, 2014.