retour sur innovation

# Estimation de profondeur monoculaire avec un réseau adversaire

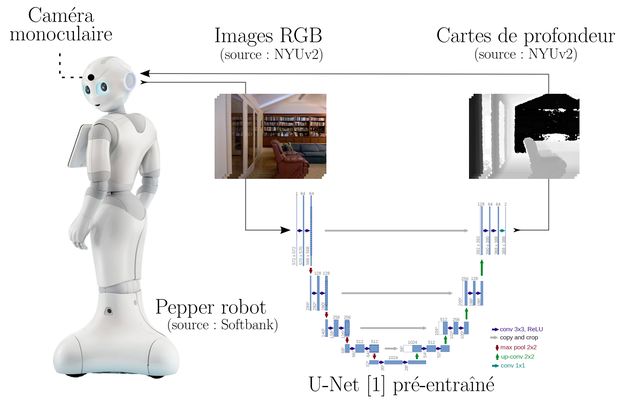
Marcela Carvalho[1,2], Bertrand Le Saux[1], Pauline Trouvé-Peloux[1],
Andrés Almansa[2], Frédéric Champagnat[1]

ONERA/DTIS[1], Université Paris Descartes[2]

## Contexte

➥ La **perception 3D** à partir d'**une seule image** est aujourd'hui un problème majeur de la vision par ordinateur. Les **approches actuelles** pour l'estimation de profondeur demandent plusieurs capteurs et ont souvent des **limitations** selon l'environnement (soleil, texture).

➥ Le ***machine learning*** obtient actuellement des résultats étonnants sur de nombreuses applications en vision par ordinateur. Nous proposons alors l'exploration de ces méthodes.



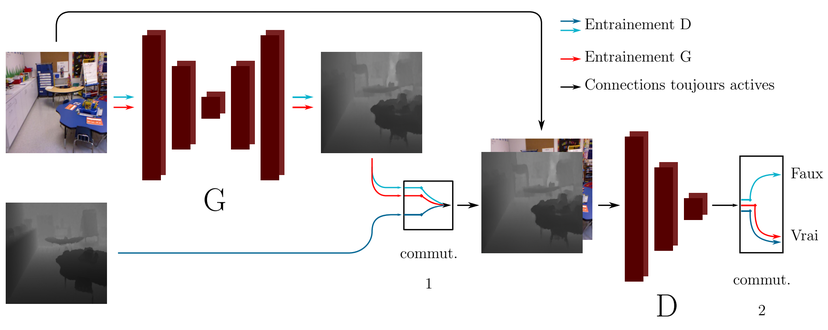

➥ Les **réseaux génératives adversaires** (GANs) [2] sont capables de générer des images réalistes à partir d'un bruit, sans la définition implicite d'une fonction de perte, en apprenant une métrique dans l'espace des images.

**État de l'art** :
➥ Eigen et al. [3] (**architecture multi-échelle**) ;
➥ Laina et al. [4] (***Residual Network*-ResNet** et perte **berHu**) ;
➥ Xu et al. [5] (**ResNet** et ***Conditional Random Fields*-CRF**).

**Motivation :**
Rendre le réseau capable d'apprendre tout seul à générer des cartes de profondeur **réalistes**.

**Notre contribution :**
Utiliser un réseau profond du type **codeur-décodeur** couplé avec un **entraînement adversaire conditionnel (cGAN)** pour l'**estimation de profondeur monoculaire**.

## L'apprentissage adversaire conditionnel




➥ Les **GANs** sont composés de deux réseaux avec des objectifs adversaires : le **générateur (G)**, entraîné à générer des images réalistes par rapport à la vérité terrain (VT) ; et le **discriminateur (D)**, un classificateur binaire entraîné à indiquer si l'image à son entrée est réelle ou artificielle. Après l'entraînement, seul le générateur est utilisé pour estimer la sortie.

➥ Les **GANs conditionnels** (cGANs) renforcent le réalisme des sorties à l'aide d'un élément d'entrée supplémentaire à G et D (*e.g.* label, image).

➥ **Objectif du discriminateur :**

$$\theta_D = max(\mathbb{E}_{x,y\sim p_{donnees}(x,y)}[log\, D(x,y)] + \mathbb{E}_{x\sim p_{donnees}(x)}[1 - log\, D(x, G(x))]).$$

➥ **Objectif du générateur :**

$$\theta_G = max(\mathbb{E}_{x\sim p_{donnees}(x)}[log\, D(x, G(x))]).$$

## Validation du cGAN pour la profondeur

Nous avons testé une architecture similaire à Isola *at al.* [6] sur la **base de données NYU-v2** [7] en faisant varier la taille de la base d'apprentissage. L'ajout d'une **perte L1** à la sortie du générateur guide l'entraînement pour générer des images plus lisses.

| Division | Perte | Erreur rel | Erreur rms | Précision $\delta<1.25$ | Précision $\delta<1.25^2$ | Précision $\delta<1.25^3$ |
|---|---|---|---|---|---|---|
| NYUv2-795 | cGAN+L1 | 0.306 | 1.040 | 51.8% | 80.5% | 92.4% |
| NYUv2-12k | cGAN+L1 | 0.246 | 0.887 | 60.8% | 86.0% | 95.2% |
| NYUv2-230k | cGAN+L1 | 0.240 | 0.864 | 62.8% | 87.2% | 95.5% |

➥ Validation du concept ;
➥ Performances augmentent avec le nombre d'exemples.

## Travaux en cours : un réseau dédié

**Notre approche :** ➥ Réseau plus profond (U-net basée sur VGG-16) ;
➥ *Least Square* GAN (LSGAN).

| Architecture G | Réseau-base | # données | Erreur rel | Erreur log10 | Erreur rms | Erreur rmslog | Précision $\delta<1.25$ | Précision $\delta<1.25^2$ | Précision $\delta<1.25^3$ |
|---|---|---|---|---|---|---|---|---|---|
| Wang *et al.* [8] | | 2M | 0.220 | 0.094 | 0.745 | 0.262 | 60.5% | 89.0% | 97.0% |
| Laina *et al.* [4] | | 95k | 0.194 | 0.083 | 0.790 | - | 62.9% | 88.9% | 97.1% |
| Liu *et al.* [9] | | 795 | 0.213 | - | 0.087 | 0.759 | 65.0% | 90.6% | 97.6% |
| Notre approche | VGG-16 | 12k | 0.204 | - | 0.768 | - | 69.2% | 90.6% | 96.9% |
| Xu *et al.* [5] | | 4.7k | 0.169 | 0.071 | 0.673 | - | 69.8% | 92.2% | 98.1% |
| Notre approche | | 230k | 0.191 | - | 0.745 | - | 71.4% | 91.1% | 97.3% |
| Eigen *et al.* [3] | | 2M | 0.158 | - | 0.641 | 0.214 | 76.9% | 95.0% | 98.8% |
| Laina *et al.* [4] | ResNet | 95k | 0.127 | 0.055 | 0.573 | 0.195 | 81.1% | 95.3% | 98.8% |
| Xu *et al.* [5] | | 96k | 0.121 | 0.052 | 0.586 | - | 81.1% | 95.4% | 98.7% |

➥ Performances comparables à l'état de l'art ;
➥ Images plus réalistes (Fig. 1).

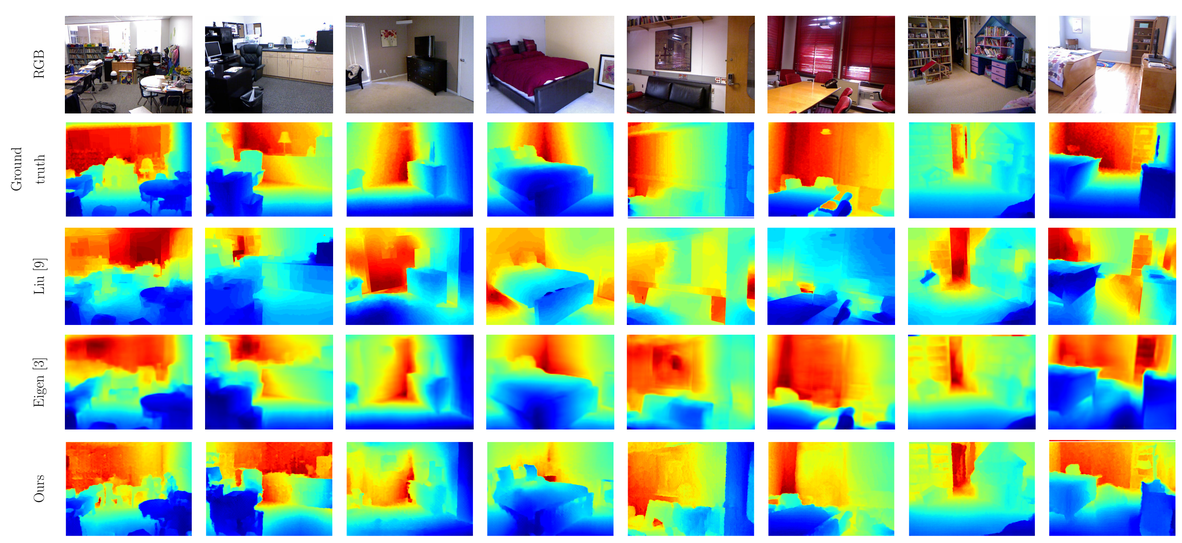



Fig. 1 Comparaison qualitative des différentes approches pour la prédiction de profondeur monoculaire

## Perspectives

➥ Utilisation de réseaux plus profonds (*e.g.* ResNet, DenseNet) ;
➥ Approche multiscale.

## Bibliographie

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

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

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

[4] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV)*. IEEE, 2016.

[5] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," *CVPR*, 2017.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

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

[8] 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.

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