Apprentissage de modalités auxiliaires pour la localisation basée vision

Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, Cédric Demonceaux

Apprentissage de modalités auxiliaires pour la localisation basée vision

Nathan Piasco[1,2] Désiré Sidibé[1] Valérie Gouet-Brunet[2] Cédric Demonceaux[1]

[1] Le2i, ERL CNRS VIBOT 6000, Université Bourgogne Franche-Comté
[2] Univ. Paris-Est, LaSTIG MATIS, IGN, ENSG, F-94160 Saint-Mandé, France

nathan.piasco@u-bourgogne.fr

Résumé

Dans cet article nous présentons une nouvelle méthode d'apprentissage à partir de modalités auxiliaires pour améliorer un système de localisation basée vision. Afin de bénéficier des informations de modalités auxiliaires disponibles pendant l'apprentissage, nous entraînons un réseau convolutif à recréer l'apparence de ces modalités annexes. Nous validons notre approche en l'appliquant à un problème de description d'images pour la localisation. Les résultats obtenus montrent que notre système est capable d'améliorer un descripteur d'images en apprenant correctement l'apparence d'une modalité annexe. Comparé à l'état de l'art, le réseau présenté permet d'obtenir des résultats de localisation comparables, tout en étant plus compacte et plus simple à entraîner.

Mots Clef

Localisation basée image, Apprentissage via des données annexes, Fusion de modalités.

Abstract

In this paper we present a new training with side modality framework to enhance image-based localization. In order to learn side modality information, we train a fully convolutional decoder network that transfers meaningful information from one modality to another. We validate our approach on a challenging urban dataset. Experiments show that our system is able to enhance a purely image-based system by properly learning appearance of a side modality. Compared to state-of-the-art methods, the proposed network is lighter and faster to train, while producing comparable results.

Keywords

Image-based localization, Learning with Side Information, Modality Fusion.

1 Introduction

La localisation basée vision (LBV) est un sujet omniprésent dans les applications de vision par ordinateur et de robotique. Elle consiste à retrouver la localisation d'une requête visuelle par rapport à une référence (absolue). La LBV est utilisée dans de nombreuses applications comme la conduite autonome, la réalité augmentée, la navigation en milieu urbain ou encore la mise à jour automatique de référentiels. Les travaux de recherche récents se focalisent sur l'amélioration d'algorithmes existant de LBV, en manipulant des données qui appartiennent à la même modalité : les images. Les principales méthodes utilisées s'appuient sur la description globale d'images [1, 19, 11, 26] et l'appariement de descripteurs locaux [28, 7]. Les données de modalités différentes sont de plus en plus accessibles, ce qui a permis l'émergence de méthodes fusionnant plusieurs modalités [24, 25] ; des informations de profondeur [8] ou de sémantique [3] peuvent être utilisées en parallèle des images. Cependant, la plupart de ces travaux s'appuient sur des informations qui sont accessibles à la fois pendant l'étape de création des données de référence mais aussi pendant le traitement *en ligne* de la requête. Dans cet article, nous considérons une méthode de LBV utilisant des informations auxiliaires, par exemple lorsque les images et leurs cartes de profondeur associées sont disponibles pendant la création des données de références mais, au moment de la requête, seule la modalité image peut être utilisée pour la localisation.

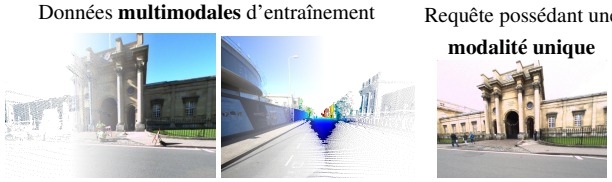

FIGURE 1 – **Entraînement d'un système de localisation basé vision en utilisant des informations auxiliaires :** Nous introduisons un nouvelle méthode qui bénéficie de modalités auxiliaires au moment de l'apprentissage (images de gauche) afin d'améliorer les résultats d'un système de localisation qui utilise seulement des images (image de droite).

Les véhicules de cartographie des milieux urbains permettent de collecter des données de plus en plus complètes [23], composées de plusieurs modalités [22, 31]. Néanmoins, les applications de LBV utilisent la plupart du temps des capteurs embarqués, comme l'appareil photo

d'un smart-phone. Nous proposons ici une technique permettant d'exploiter des données complètes (images et scan laser) collectées au préalable pour améliorer une application de LBV qui n'utilise que des données acquises par des capteurs légers. La figure 1 illustre le contexte de l'application considérée.

Afin de profiter des modalités auxiliaires dont on dispose, nous proposons d'apprendre à reproduire l'apparence de ces informations seulement à partir d'une modalité principale. Les réseaux de neurones profonds, spécifiquement les réseaux entièrement convolutifs, permettent de reproduire certaines modalités à partir d'une simple image [10]. Nous proposons une méthode, basée sur les réseaux convolutifs, qui permet de transférer l'information d'une modalité vers une autre. Cette information est ensuite utilisée pour améliorer notre système de LBV. Nous montrons que notre système permet d'augmenter la précision de la localisation, en utilisant lors du traitement de la requête une unique modalité, et ce d'une manière plus naturelle et efficace que les méthodes d'apprentissage avec modalités auxiliaires de la littérature [15].

Le plan de notre article est le suivant : dans la section 2 nous présentons une revue de la littérature liée à notre problématique. La section 3 est consacrée à la description détaillée de notre proposition reponsant sur l'apprentissage à partir d'informations auxiliaires. Nous démontrons ensuite l'efficacité de notre système au travers d'une application de description d'images pour la LBV dans la section 4. Enfin, la section 5 conclut ce travail.

2 Etat de l'art

2.1 Transfert de modalités par réseaux convolutifs

Recréer une modalité à partir d'une autre est un problème complexe. Récemment, les réseaux neuronaux profonds se sont montrés très efficaces dans la résolution de ce problème [21, 10, 32]. Les réseaux purement convolutifs (ConvNet) sont particulièrement adaptés à cette problématique [5]. Par exemple, les architectures convolutives décrites dans [21, 10] permettent d'estimer une carte de profondeur associée à une image RVB à partir de cette seule image. Des cartes thermiques jumelées à des images peuvent également être recréées par le biais de ConvNet [32]. Ces systèmes sont également capables de modéliser des changements propres à une même modalité, comme des variations d'illumination dans une scène [33] ou les changements visuels observés au cours des différentes saisons de l'année [35]. Ces travaux nous confortent dans l'idée d'utiliser une architecture purement convolutive pour apprendre l'apparence des modalités auxiliaires que nous souhaitons exploiter.

2.2 Localisation basée vision

Les techniques de localisation basée vision reposent très souvent sur des méthodes d'indexation d'images par le contenu. En retrouvant la donnée géo-référencée la plus proche de la requête, le système est capable de fournir une information globale de position de cette instance [34]. Les algorithmes communément utilisés reposent sur l'extraction de descripteurs locaux qui sont ensuite agrégés pour produire une signature globale associée à l'image [2, 30]. Les réseaux de neurones sont maintenant utilisés pour extraire des descripteurs globaux d'images spécialement destinés à la localisation [1, 19, 11, 26].

2.3 Apprentissage à partir d'une modalité auxiliaire

Gupta et al. [13] présentent une méthode pour classifier des objets à partir d'une carte de profondeur. Disposant de peu de données d'entraînement dans la modalité *profondeur*, les auteurs exploitent des images RVB jumelées aux cartes de profondeur afin d'augmenter leur base d'apprentissage. Dans les travaux de [16], les auteurs explorent le gain amené par l'ajout d'une modalité auxiliaire pour la classification d'images. Les résultats montrent que la modalité annexe peut améliorer les résultats de la classification même si aucun exemple pour certaines classes dans cette modalité n'est présent pendant la phase d'apprentissage. La contribution la plus proche de la nôtre a été exposée dans [15]. Dans ces travaux, les auteurs utilisent un réseau de neurones pour recréer l'information d'une modalité présente seulement pendant la phase d'apprentissage de la méthode. Leur système final permet *d'halluciner* la réponse d'un réseau de neurones qui serait produite par une carte de profondeur à partir d'une image. D'une manière similaire, Xu et al. [32] utilisent des images thermiques recréées pour améliorer la détection de piétons.

Dans cet article, nous proposons d'extraire l'information de la modalité annexe d'une manière différente de celles présentées dans [15, 32], en exploitant la capacité d'un réseau convolutif à récréer l'information d'une modalité en fonction d'une autre.

3 Utilisation de modalités auxiliaires par transfert d'apparence

Dans la section 3.1, nous décrivons deux systèmes qui sont entraînés à partir de données multimodales. Cependant, au moment du test, une unique modalité est utilisée pour évaluer ces systèmes. La modalité principale, c'est-à-dire celle utilisée au moment du test, est appelée *modalité* **centrale**. Les autres modalités utilisées pendant la phase d'entraînement sont appelées *modalités* **annexes**. Dans la section 3.2, nous présentons les détails et optimisation liées au système proposé.

3.1 Architecture des réseaux

Ci-dessous, nous revenons sur l'architecture du réseau d'hallucination proposée dans [15], avant de présenter notre contribution.

Hallucination de modalité. Hoffman et al. [15] présentent une architecture baptisée *réseau d'hallucination de*

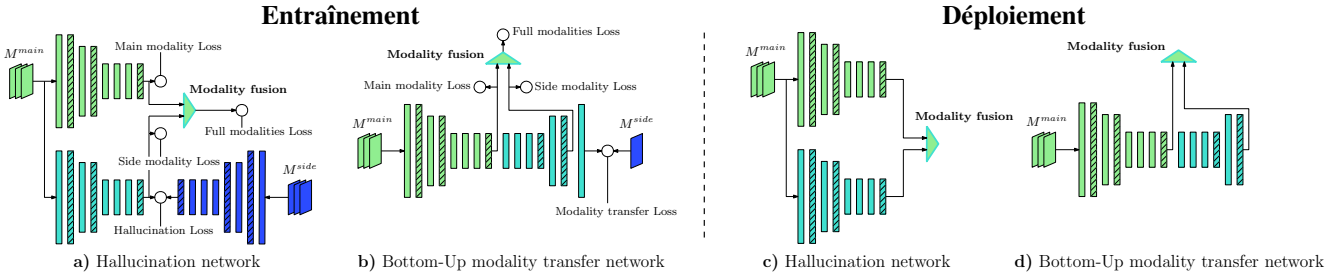

FIGURE 2 – **Aperçu de la méthode :** Nous présentons une nouvelle architecture de réseau permettant de profiter de modalités annexes au moment de l'apprentissage (*bottom-up modality transfer network* ou BU-TF). Les schémas de **gauche** (a-b) représentent les procédures d'entraînement du réseau d'hallucination [15] et de notre réseau BU-TF. Les réseaux en vert sont entraînés pour extraire de l'information dans la modalité centrale (dans notre cas des images), le bleu foncé est entraîné seulement sur des données de modalité annexe (dans notre cas des cartes de profondeur) tandis que les réseaux bleu clair servent à l'apprentissage des données annexes à partir des données centrales. Les figure (c-d) à **droite** présentent les réseaux au moment du déploiement, quand la modalité annexe n'est plus disponible. Les rectangles vides représentent des convolutions, les rectangles hachés des opérations de rectification linaire (`Relu`) et les changements de taille dans les blocs à des opérations de réductions ou d'extensions (polling, unpooling).

modalité afin de modéliser une information issue d'un réseau de neurones entraîné sur une modalité annexe à partir d'une modalité centrale. Le réseau est d'abord entraîné sur une modalité annexe M^{annexe} avant d'être ré-entraîné sur des données de modalité $M^{centrale}$. Au cours du ré-entraînement avec les données $M^{centrale}$, une contrainte supplémentaire est appliquée au réseau d'hallucination pour le forcer à reproduire la même réponse convolutive qu'au moment où les données M^{annexe} étaient présentées au réseau. La figure 2 résume l'entraînement (Figure 2-a) et le test (Figure 2-c) du réseau d'hallucination de modalités.

Transfert de modalité par encodeur/décodeur. Le réseau d'hallucination de modalités oblige un système ayant en entrée des données de modalité $M^{centrale}$ à reproduire une information normalement obtenue via des données de modalité M^{annexe}, sans essayer de comprendre le lien qui lie ces deux types de données. Au contraire, nous proposons un réseau qui fournit une information annexe en reproduisant l'apparence d'une modalité annexe d'une manière ascendante. La modalité $M^{centrale}$ est présentée à l'entrée du réseau tandis qu'un décodeur est ajouté pour permettre de récréer la modalité M^{annexe}. L'architecture encodeur/décodeur utilisée est inspirée des travaux de [5]. Finalement, l'information liée à la modalité M^{annexe} est extraite au niveau des réponses générées par le décodeur. La figure 2 illustre côte à côte les deux stratégies d'apprentissage avec modalité annexe étudiées dans cet article.

Notre méthode présente trois avantages principaux :

— Nous n'avons pas besoin d'entraîner un réseau sur la modalité M^{annexe}, à la différence du réseau d'hallucination. L'information annexe est automatiquement apprise durant la phase d'entraînement principale.

— Notre architecture est par nature plus compacte : $29k$ paramètres comparé à $41k$ pour le réseau d'hallucination, en considérant que les deux systèmes sont basés sur l'architecture de base d'Alexnet.

— Comme les réseaux neuronaux efficaces ont besoin en entrée d'images RVB sur 3 canaux, les méthodes classiques utilisant des modalités auxiliaires requièrent le pré-traitement des données pour en faire des données à 3 canaux. Par exemple, la représentation HHA [12] ou la colorisation [9] sont utilisées pour transformer une carte de profondeur possédant un seul canal en données à 3 canaux. Ces manipulations de données nécessitent des calculs supplémentaires, alors que notre approche n'a pas besoin de pré-traitement puisque nous fixons le nombre de canaux finaux de notre décodeur égal au nombre de canaux de la modalité annexe utilisée.

3.2 Entraînement

Dans les sous-sections suivantes, nous présentons les détails liés à l'entraînement de notre système. Nous désignons par la suite notre réseau *bottom-up modality transfer* (abréviation BU-TF) pour la description d'images.

Apprentissage d'un descripteur global. Une description discriminante d'une image peut être obtenue en extrayant la réponse (également appelée carte de descripteurs) aux couches de convolutions finales d'un CNN [4].

Agrégation des cartes de descripteurs. La sortie brute de la dernière couche convolutive d'un CNN est souvent trop grande pour être exploitée telle quelle pour la description d'images. Des opérations de regroupement sont alors appliquées sur cette carte de descripteurs pour en réduire sa taille et la rendre plus robuste aux changements de points de vue. Le *Maximum of Activation polling* (MAC) [27, 26], ou plus récemment le regroupement des descripteurs via NetVLAD [1] sont des méthodes de réduction largement utilisées. Nous utilisons dans ce travail le regroupement des descripteurs par MAC afin de valider notre architecture pour la description globale d'images.

Fonction de coût par triplet. Il est commun d'utiliser une fonction de coût mettant en jeu des triplets d'images pour permettre au réseau de neurones de produire des descripteurs d'images performants [1, 11, 26]. Ces triplets sont composés d'une image requête q, d'un exemple positif q^+ (une image représentant la même scène que l'image requête mais avec un angle de vue ou un aspect visuel lé-

gèrement différent) et d'un exemple négatif q^-. Les poids du réseau sont ensuite modifiés pour minimiser la fonction définie dans l'équation 1 :

$$Loss_{triplet} = max\left(\left\|f(q) - f(q^+)\right\|^2 - \left\|f(q) - f(q^-)\right\|^2 + \lambda, 0\right), \quad (1)$$

$$avec \begin{cases} f(x) = \text{le descripteur de l'image } x \text{ extrait par le CNN} \\ \lambda = \text{une marge constante} \end{cases}$$

Fonctions de coût. Afin d'optimiser notre réseau, nous utilisons une procédure d'entraînement comportant plusieurs fonctions de coût. Au cours de l'entraînement, l'apprentissage de la modalité annexe est assuré par la fonction décrite dans l'équation 2 :

$$Loss_{transfer} = \frac{1}{n}\sum_{i=1}^{n}\left\|\widetilde{M}(M^{centrale})_i - M_i\right\|_1, \quad (2)$$

avec $\widetilde{M}(x)$ la sortie du décodeur en fonction de l'entrée x présentée au réseau. Cette contrainte permet d'apprendre au réseau les liens qui existent entre la modalité $M^{centrale}$ et la modalité M^{annexe}. La fonction décrite par l'équation 1 est appliquée d'une part pour optimiser le descripteur issu de la modalité centrale du réseau, mais également pour optimiser le descripteur calculé à partir de l'information annexe fournie par le réseau. La fonction global a minimiser devient :

$$Loss = Loss^{tot}_{triplet} + Loss^{annexe}_{triplet} * \sigma_{annexe} + Loss^{centrale}_{triplet} * \sigma_{centrale} + Loss_{transfer} * \sigma_{transfer}, \quad (3)$$

où $Loss^{tot}_{triplet}$ désigne l'équation 1 appliquée pour optimiser un descripteur composé à la fois du descripteur issu de la modalité centrale et du descripteur issu de la modalité annexe. Les paramètres $\{\sigma_{annexe}, \sigma_{centrale}, \sigma_{transfer}\}$ servent à pondérer les différentes fonctions.

Fonction de diversification. Afin d'éviter que l'information de modalité annexe soit redondante avec l'information de la modalité centrale, nous introduisons dans l'équation 4 une fonction de diversification :

$$Loss_{div} = max\left(Loss^{tot}_{triplet} - Loss^{centrale}_{triplet} + \lambda_{div}, 0\right), \quad (4)$$

avec λ_{div} une marge constante qui oblige la $Loss^{tot}_{triplet}$ à être toujours inférieur à $Loss^{centrale}_{triplet}$.
La fonction finale à minimiser devient :

$$Loss = Loss^{tot}_{triplet} + Loss^{annexe}_{triplet} * \sigma_{annexe} + Loss^{centrale}_{triplet} * \sigma_{centrale} + Loss_{transfer} * \sigma_{transfer} + Loss_{div} * \sigma_{div}. \quad (5)$$

FIGURE 3 – **Robotcar Dataset :** La carte de **gauche** représente la division des parcours utilisés : en vert la zone d'apprentissage, en bleu la zone de validation et en rouge la zone de test. Les images à **droite** sont extraites des données de test, avec en haut l'image requête et en bas l'image la plus proche dans la base de test. On constate des changements importants entre les images requête et les images de la base qui sont dus à des occultations ou des changements de points de vue et d'illumination.

Procédure d'entraînement. L'optimisation de notre réseau se fait en deux étapes : tout d'abord la partie encodeur est entraînée seule, puis le système complet, encodeur et décodeur, est optimisé par rapport à la fonction décrite dans l'équation 5. Les facteurs pondérant les différents critères de la fonction de coût finale sont déterminés empiriquement en faisant en sorte que toutes les normes du gradient d'erreur associées à chacune des expressions soient du même ordre de grandeur.
Si l'on compare avec la procédure d'entraînement du réseau d'hallucination, notre méthode est plus rapide à mettre en place, puisque nous n'avons pas besoin d'entraîner un réseau spécialement sur la modalité annexe (il s'agit du réseau bleu foncé dans la figure 2).

4 Expériences

Afin de valider la méthode proposée dans la section précédente, nous considérons la problématique de localisation basée vision en milieu urbain. Pour ce faire, nous disposons d'une base de données d'images géo-référencées couvrant une certaine zone et nous voulons retrouver l'image de cette base la plus proche d'une image requête donnée. Nous apprenons à décrire l'apparence de nos images au travers du réseau de neurones décrit dans la section 3.2 et nous évaluons la similarité entre deux descripteurs en calculant leur produit scalaire.
La section 4.1 décrit le cadre choisi pour l'évaluation de notre méthode, puis des détails sur l'implémentation des systèmes testés ainsi que des résultats préliminaires sont donnés dans la section 4.2. Enfin la section 4.3 est consacrée à la présentation et à la discussion des résultats obtenus.

4.1 Robotcar Dataset

Nous avons choisi comme base de données pour nos tests le jeu de données *Oxford Robotcar* [22]. Il a la particularité de fournir des données de modalités différentes : images et nuages de points (géométrie + réflectance). Nous utilisons comme modalité centrale l'image et comme modalité auxiliaire la profondeur obtenue grâce au nuage de points 3D. D'autre part, ce jeu de données possède une redondance temporelle : le même trajet à été effectué plus d'une centaine de fois sur une période d'un an. Ainsi, nous pouvons

extraire des informations d'une même scène à différentes dates, ce qui est essentiel pour la création des triplets utilisés pour l'entraînement de nos CNN.

Division du jeu de données. La répartition des données d'entraînement, de validation et de test est visible sur la figure 3. La création des 400 triples utilisés pour l'entraînement a été faite à partir des trajets effectués aux dates : *02-10-2015, 05-19-2015, 08-28-2015* et *11-10-2015* (mois-jour-année). Le jeu de test est composé de 1688 images géo-référencées (trajet effectué le *06-26-2014*) prises tous les 5 mètres sur la zone décrite à la figure 3 et de 261 images requête de la même zone mais à une autre période (trajet effectué le *06-24-2014*). Les images requête et leurs images associées dans la base de test peuvent comporter des différences notables, comme l'illustre la figure 3.

Évaluation. Afin d'évaluer les performances des différents systèmes, nous considérons les métriques suivantes :

- **Rappel @N :** la requête est considérée comme localisée si l'un des N premiers candidats retournés par le système se situe dans un rayon de 25 mètres autour de la vraie position de la requête. Il s'agit d'une métrique standard utilisée pour évaluer les systèmes de localisation basée image [1]. Le rappel moyen @N (*mean recall @N*) est la moyenne des valeurs de rappel de 1 jusqu'à N.
- **Distance à la requête :** la distance entre le premier candidat retourné par la méthode et la position de la requête est reportée pour évaluer la précision absolue de la méthode. Les courbes présentées représentent le pourcentage de requêtes bien localisées pour des valeurs de distances allant de 1 à 100 mètres.

Pré-traitement. L'information de profondeur fournie par le dataset *Oxford Robotcar* est extraite d'un nuage de points. En projetant ce nuage de points 3D dans le repère d'une image, nous obtenons une carte de profondeur éparse de la scène (cf. figure 4). Or, les CNN ont besoin de données denses pour être utilisés conjointement avec des images. Afin de produire des cartes de profondeur denses à partir du nuage de point épars d'origine, nous utilisons l'algorithme *d'inpainting* de [6]. Notons tout de même que lorsque la densité du nuage de points est trop faible ou que certains points 3D sont en réalité occultés dans la scène, la méthode produit des cartes de profondeur incohérentes. La figure 4 présente un aperçu des données denses obtenues, qui seront utilisées lors des phases d'apprentissage.

4.2 Implémentation

Pour créer les descripteurs globaux d'images, la partie classification du réseau est supprimée (*i.e.* toutes les couches de neurones entièrement connectées) et la carte de descripteurs résultante de la dernière couche de convolution du réseau est agrégée par MAC. Enfin, le descripteur MAC obtenu est L_2-normalisé avant comparaison. Nous avons testé plusieurs architectures de réseaux pour décrire nos

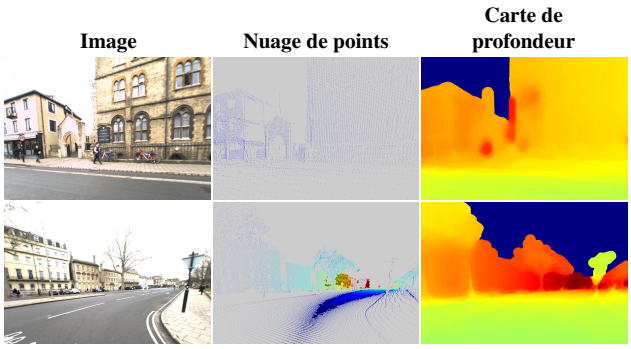

FIGURE 4 – **Deux exemples illustrant la densification des données de profondeur :** l'algorithme d'inpainting proposé par [6] permet d'obtenir une carte de profondeur dense qui peut être utilisée directement par nos CNN.

données : Alexnet [20], VGG [29], ResNet [14] et DenseNet [17]. Nous avons reporté dans le tableau 1 les résultats obtenus avec ces différentes architectures, en considérant seulement les réseaux pré-entraînés sur ImageNet, donc sans entraînement sur nos données.
Étonnamment, les réseaux possédant beaucoup de paramètres n'obtiennent pas de résultats significativement supérieurs à ceux du réseau le plus léger, Alexnet. Étant donné la taille deux à quart fois supérieure du descripteur produit par les concurrents d'Alexnet, nous avons décidé de choisir Alexnet comme réseau de base pour tester notre méthode.

Architecture détaillée du réseau BU-TF. Le descripteur d'image BU-TF est composé d'une partie encodeur immédiatement suivie d'un décodeur. Le descripteur associé à la modalité centrale est obtenu à la sortie de l'encodeur et le descripteur associé à la modalité annexe est extrait de la réponse aux couches convolutives du décodeur. Les deux descripteurs sont enfin concaténés pour former le descripteur final.

Architecture détaillée du réseau d'hallucination. Nous comparons notre réseau BU-TF avec une adaptation du réseau d'hallucination proposé par [15]. Ce réseau est composé de deux encodeurs de type Alexnet produisant un descripteur pour la modalité centrale et un descripteur *halluciné* pour la modalité annexe. Ces deux descripteurs sont ensuite concaténés pour produire le descripteur final.
Le tableau 2 résume les deux implémentations considérées.

Entraînement. Nous suivons le schéma d'entraînement suivant :

- Entraînement de l'encodeur Alexnet + MAC sur notre jeu d'entraînement avec seulement la modalité centrale (image).
- Pour le réseau d'hallucination seulement : nous entraînons également un encodeur Alexnet + MAC sur les données de modalité annexe seulement (cartes de profondeur denses). L'encodeur *d'hallucination* du réseau est ensuite initialisé avec les poids de ce réseau, et nous suivons la procédure d'entraînement originale des auteurs.

TABLE 1 – **Comparaison des architectures :** Nous comparons différentes architectures de réseaux pour la description globale d'images. Les scores rapportés correspondent aux réseaux pré-entraînés sur ImageNet.

Nom	Nb de conv.	Taille du desc. MAC	Rappel @1	Rappel moyen @1 jusqu'à 50
Alexnet [20]	5	256	36.40%	79.50%
VGG [29]	16	512	31.03%	74.04%
ResNet [14]	51	2048	37.54%	73.27%
DenseNet [17]	201	1920	38.69%	78.60%

— Nous entraînons notre réseau BU-TF avec la partie encodeur initialisée avec les poids du premier réseau entraîné et la partie décodeur initialisée aléatoirement.

Étant donnée la taille réduite de notre jeu d'entraînement, pendant la dernière phase d'entraînement les poids de l'encodeur lié à la modalité centrale sont figés afin d'éviter un sur-apprentissage (pour le réseau d'hallucination et le réseau BU-TF). Chaque réseau est entraîné pendant approximativement 20 époques avec des mini-batch de 5 triplets par descente stochastique de gradient à un taux d'apprentissage de 0.001.

4.3 Résultats

Extraction de l'information annexe. De manière similaire à [15], nous comparons les résultats obtenus avec notre réseau BU-TF en fonction de l'endroit où est extraite l'information de modalité annexe. Nous pouvons extraire cette information à 4 endroits différents du décodeur après les couches convolutives suivantes, du plus proche au plus loin de l'encodeur : `deconv3.1`, `deconv3.2`, `deconv3.3` et `deconv2`. Le tableau 3 rapporte les différents scores obtenus en fonction de l'endroit où à été extraite l'information de modalité annexe. On constate que plus l'information de modalité est extraite *loin* dans le décodeur, plus elle participe efficacement à la description de l'image. Ces résultats nous confortent dans le fait que le transfert de modalité de façon ascendante est efficace. Dans la suite des expérimentations, nous utilisons la carte de descripteurs obtenue après la couche de convolution `deconv2` comme information de modalité annexe.

Hallucination contre BU-TF. Nous présentons dans la figure 5 les résultats des réseaux d'hallucination et BU-TF comparés au réseau utilisant seulement l'information centrale pendant la phase d'apprentissage. L'ajout d'informations auxiliaires est bénéfique dans les deux cas, les réseaux entraînés sur les données multimodales ont de meilleurs performances que le réseau entraîné sur une seule modalité. D'autre part, notre réseau BU-TF a des performances comparable au réseau d'hallucination, bien qu'il possède bien moins de paramètres pouvant être optimisés et qu'il produit un descripteur plus compact (320 dimensions contre 512 pour le réseau d'hallucination).

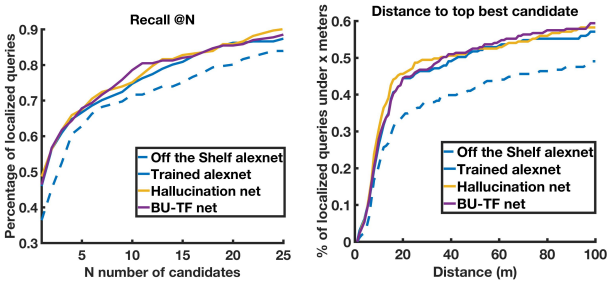

FIGURE 5 – **Résultats sur *Robotcar dataset* :** Tous les réseaux sont pré-entraînés sur ImageNet. Off-the-shelf alexnet désigne le réseau avant ré-entraînement sur le jeu de données *Robotcar dataset*.

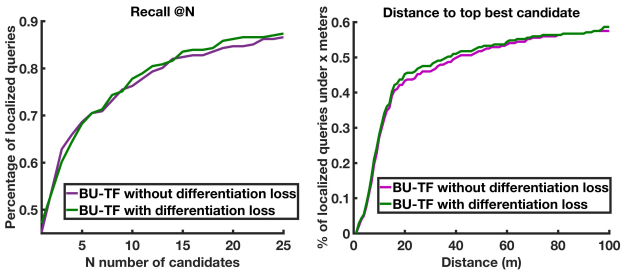

FIGURE 6 – **Validation de la fonction de diversification :** Nous comparons les résultats de notre réseau BU-TF avec et sans la fonction de diversification lors de l'entraînement.

Fonction de diversification. Afin de valider la nouvelle fonction de coût que nous avons introduite dans l'équation 4, nous avons comparé les scores de localisation de notre réseau BU-TF avec et sans cette fonction lors de l'entraînement. Nous avons rapporté les résultats obtenus sur la figure 6. Nous observons une amélioration systématique des scores de localisation, tant dans la précision que dans le rappel. Le paramètre de marge de la fonction de diversification a été fixé à $\lambda_{div} = 0.1 * \lambda$, où λ est la marge de la fonction décrite par l'équation 1. Pour nos expérimentations nous avons utilisé $\lambda = 0.25$.

Fusion de modalités. Jusqu'à maintenant, le descripteur issu de la modalité centrale et celui issu de la modalité annexe ont simplement été concaténés. Nous proposons ici une nouvelle méthode de fusion inspirée des réseaux récurrents à porte. Le descripteur fusionné f^{tot} est alors défini par :

$$f^{tot} = g * f^{centrale} + (1 - g) * f^{annexe}, \qquad (6)$$

avec g un facteur pondérant obtenu par une couche de neurones entièrement connectés :

$$g = sig\left(W_g \cdot \left[f^{centrale}, f^{annexe}\right]\right). \qquad (7)$$

W_g représente les poids de la couche de neurones entièrement connectés et $\left[f^{centrale}, f^{annexe}\right]$ est la concaténation des descripteurs $f^{centrale}$ et f^{annexe}. La fonction sigmoïde permet d'assurer que la pondération obtenue appartient bien à l'intervalle $[0, 1]$.
Étant donné que les descripteurs $f^{centrale}$ et f^{annexe} doivent avoir la même taille, nous ne pouvons appliquer

TABLE 2 – **Détails des architectures :** le réseau d'hallucination (Hall) [15] est composé de deux encodeurs Alexnet suivis d'une agrégation par MAC pour produire le descripteur global de l'image. Notre réseau *bottom-up modality transfer* (BU-TF) concatène le descripteur obtenu d'un encodeur Alexnet avec un descripteur extrait de la partie décodeur du réseau.

Descripteur par Hallucination		
nom	**entrée**	**taille de sortie**
$alexnet\_feat_{centrale}$	`image rvb`	-
$alexnet\_feat_{annexe}$	`image rvb`	-
$MAC_{centrale}$	$alexnet\_feat_{centrale}$	256
MAC_{annexe}	$alexnet\_feat_{annexe}$	256
concaténation	$MAC_{centrale}$, MAC_{annexe}	512

Descripteur BU-TF		
nom	**entrée**	**taille en sortie**
$alexnet\_feat_{centrale}$	`image rvb`	-
decode_alexnet	$alexnet\_feat_{centrale}$	-
$MAC_{centrale}$	$alexnet\_feat_{centrale}$	256
MAC_{annexe}	decode_alexnet	64
concaténation	$MAC_{centrale}$, MAC_{annexe}	320

TABLE 3 – **Extraction de l'information annexe :** score de localisation rapporté en fonction de la couche convolutive où a été extraite l'information de modalité annexe dans la partie décodeur de notre architecture.

Couche de conv.	deconv3.1	deconv3.2	deconv3.3	deconv2
Taille du desc. MAC	256	384	192	64
Rappel @1	41.70%	44, 444%	42, 912%	45, 211%

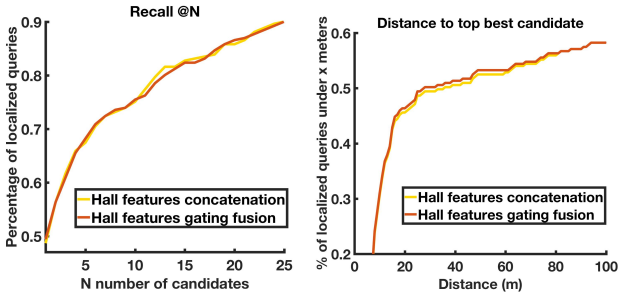

FIGURE 7 – **Évaluation de le routine de fusion :** Nous comparons ici la méthode de fusion décrite dans l'équation à la concaténation naïve des descripteurs.

cette fusion de descripteurs qu'au réseau d'hallucination. Nous pouvons cependant voir sur la figure 7 que cette nouvelle méthode de fusion permet d'améliorer légèrement les résultats de la localisation.

Discussion. Malgré la quantité limitée de données utilisées pour entraîner nos réseaux et la pauvre qualité des cartes de modalité utilisées, nous sommes capables d'extraire des informations complémentaires pour améliorer un système de localisation purement basé sur des images. Nous obtenons également des résultats comparables aux approches les plus modernes tout en ayant moins de paramètres et une procédure d'entraînement simplifiée. Ces résultats confirment notre intuition selon laquelle l'architecture entièrement convolutive est bien adaptée à la problématique d'apprentissage avec modalités auxiliaires.

5 Conclusion et perspectives

Nous avons proposé une nouvelle approche d'apprentissage à partir de modalités auxiliaires pour la localisation basée vision. Notre méthode a permis d'améliorer un système de localisation se basant sur des CNN, en proposant une architecture plus légère et une procédure d'entraînement plus rapide comparée à l'état de l'art dans le domaine. Nous avons également introduit une nouvelle fonction de coût permettant une meilleure optimisation de notre système. Enfin, une méthode de fusion supervisée de descripteurs a été présentée afin de combiner efficacement des informations issues de modalités différentes.

Travaux futures. D'autres méthodes d'agrégation de descripteurs peuvent être considérées afin d'améliorer les résultats de la localisation, comme NetVLAD [1]. Il serait également intéressant de considérer une autre application reliée au domaine de la localisation, comme la régression de pose [18], afin de vérifier que la méthode proposée est généralisable à d'autres problématiques.

Remerciements

Les auteurs souhaitent remercier Marco Bevilacqua pour avoir fourni le code source de son algorithme d'inpainting [6]. Nous remercions également le projet français ANR pLaTINUM (ANR-15-CE23-0010) pour son support financier. Enfin nous remercions NVIDIA Corporation pour le don du GPU Tesla K40c utilisé dans cette recherche.

Références

[1] Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2017). NetVLAD : CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 5297–5307.

[2] Arandjelović, R. and Zisserman, A. (2014). DisLocation : Scalable descriptor. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

[3] Ardeshir, S., Zamir, A. R., Torroella, A., and Shah, M. (2014). GIS-assisted object detection and geospatial localization. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 8694 LNCS, pages 602–617.

[4] Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014). Neural Codes for Image Retrieval. In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, pages 584–599.

[5] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12) :2481–2495.

[6] Bevilacqua, M., Aujol, J. F., Biasutti, P., Brédif, M., and Bugeau, A. (2017). Joint inpainting of depth and reflectance with visibility estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 125 :16–32.

[7] Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., and Rother, C. (2017). DSAC - Differentiable RANSAC for Camera Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Cavallari, T., Golodetz, S., Lord, N. A., Valentin, J., Di Stefano, L., and Torr, P. H. S. (2017). On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2015-Decem :681–687.

[10] Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2017). End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision (IJCV)*, 124(2) :237–254.

[12] Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 8695 LNCS(PART 7) :345–360.

[13] Gupta, S., Hoffman, J., and Malik, J. (2016). Cross Modal Distillation for Supervision Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[15] Hoffman, J., Gupta, S., and Darrell, T. (2016a). Learning with Side Information through Modality Hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834.

[16] Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., and Darrell, T. (2016b). Cross-modal adaptation for RGB-D detection. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 2016-June, pages 5032–5039.

[17] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3.

[18] Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Kim, H. J., Dunn, E., and Frahm, J.-M. (2017). Learned Contextual Feature Reweighting for Image Geo-Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. In *Communications of the ACM*, pages 1097–1105.

[21] Kuznietsov, Y., Stückler, J., and Leibe, B. (2017). Semi-Supervised Deep Learning for Monocular Depth Map Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016). 1 year, 1000 km : The Oxford RobotCar dataset. *The International Journal of Robotics Research (IJRR)*, page 0278364916679498.

[23] Paparoditis, N., Papelard, J.-P., Cannelle, B., Devaux, A., Soheilian, B., David, N., and Houzay, E. (2012). Stereopolis II : A multi-purpose and multi-sensor 3D mobile mapping system for street visualisation and 3D metrology. *Revue française de photogrammétrie et de télédétection*, 200(1) :69–79.

[24] Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2018). A survey on Visual-Based Localization : On the benefit of heterogeneous data. *Pattern Recognition*, 74 :90–109.

[25] Piasco, N., Sidibé, D., Gouet-Brunet, V., and Demonceaux, C. (2017). Localisation basée vision : de l'hétérogénéité des approches et des données. In *ORASIS, Journées francophones des jeunes chercheurs en vision par ordinateur*.

[26] Radenović, F., Tolias, G., and Chum, O. (2017). Fine-tuning CNN Image Retrieval with No Human Annotation. *arXiv*, pages 1–13.

[27] Razavian, A. S., Sullivan, J., Carlsson, S., and Maki, A. (2014). Visual Instance Retrieval with Deep Convolutional Networks. *arXiv preprint*, 4(3) :251–258.

[28] Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., and Pajdla, T. (2017). Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization ? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*.

[30] Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., and Pajdla, T. (2015). 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., and Urtasun, R. (2017). TorontoCity : Seeing the World with a Million Eyes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[32] Xu, D., Ouyang, W., Ricci, E., Wang, X., and Sebe, N. (2017). Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[33] Yin, P., He, Y., Liu, N., Han, J., Xu, W., and Zealand, N. (2017). Condition directed Multi-domain Adversarial Learning for Loop Closure Detection.

[34] Zamir, A. R. and Shah, M. (2014). Image geo-localization based on multiplenearest neighbor feature matching using generalized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8) :1546–1558.

[35] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. ICCV.