Recherche multimodale d'images aériennes multi-date à l'aide d'un réseau siamois

Margarita Khokhlova, Valérie Gouet-Brunet, Nathalie Abadie, Liming Chen

Recherche multimodale d'images aériennes multi-date à l'aide d'un réseau siamois

M. Khokhlova[1],[2] V. Gouet-Brunet[1]
N.Abadie[1] Liming Chen [2]

[1] LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mande, France
[2] Liris / Ecole Centrale de Lyon, Ecully, France

{margarita.khokhlova, valerie.gouet, nathalie-f.abadie}@ign.fr, liming.chen@ec-lyon.fr

Résumé

Cet article présente un réseau multimodal qui met en correspondance des images aériennes de territoires urbains et ruraux français prises à environ 15 ans d'intervalle. Il devrait être invariant à un large éventail de changements, tels que l'évolution du paysage au fil des années. Il exploite les images originales et les régions sémantiquement segmentées et étiquetées.

Le cœur de la méthode est un réseau siamois qui apprend à extraire des caractéristiques des paires d'images correspondantes dans le temps et des paires non correspondantes. Ces descripteurs sont suffisamment discriminants pour qu'un simple classifieur k-NN suffise comme critère de géo-correspondance final. Dans cet article, nous démontrons que notre descripteur siamois surpasse les autres descripteurs d'images en termes de recherche d'images par contenu à travers le temps.

Mots-clés

Réseaux siamois, recherche d'images multimodales par contenu, géolocalisation d'images multi-date.

Abstract

This paper introduces a multi-modal network that learns to match aerial images of french urban and rural territories taken about 15 years apart. This means it should be invariant against a big range of changes as the (natural) landscape evolves over time. It leverages the original images and semantically segmented and labeled regions.

The core of the method is a siamese network that learns to extract features from corresponding image pairs across time, and non matching pairs. These descriptors are discriminative enough, such that a simple kNN classifier on top, suffices as a final geo-matching criteria.

We demonstrate that our siamese descriptor outperforms other image descriptors for cross-time image retrieval.

Keywords

Siamese networks, multi-modal image retrieval, cross-time image geolocalization.

1 Introduction

Les images aériennes, telles que les images de satellites, avions ou d'autres dispositifs d'imagerie aérienne, sont nettement différentes des bases de données d'images telles que CIFAR [1], ImageNet [2], etc. Ces images sont sémantiquement plus similaires dans leur composition, les paysages naturels et urbains qu'elles présentent sont souvent constitués d'éléments visuellement identiques tels que la végétation, les surfaces d'eau et les structures artificielles.

Ces dernières années, un grand nombre d'images anciennes ont été numérisées, parmi lesquelles de nombreuses images aériennes [3]. Elles représentent une ressource unique qui permet d'étudier l'évolution du paysage, l'urbanisation, l'utilisation des terres, les événements historiques et autres. Le manque d'annotation de nombreux documents photographiques rend leur mise en relation avec les photographies modernes du même territoire extrêmement difficile. Le projet Alegoria [4], dans lequel ces travaux se situent, vise à créer un outil de recherche d'images anciennes par contenu (CBIR) pour aider les utilisateurs finaux.

Des approches de pointe pour la recherche d'images par contenu [5, 6] ont été conçues pour être robustes au changement des angles de vue et aux différentes modalités d'images (coloré et noir et blanc, moderne et historique). Généralement, elles sont mises au point et testées sur des référentiels composés d'objets architecturaux artificiels distincts [7, 8]. Ces images d'objets architecturaux sont très différents des images aériennes, que ce soit par la résolution ou par le contenu. Ceci représente une piste de recherche visant à savoir si ces méthodes peuvent être utilisées pour retrouver les images aériennes correspondantes prises à plusieurs années de décalage. Les images aériennes produites par les agences cartographiques nationales sont en outre souvent exploitées comme sources d'informations pour produire des référentiels de données géographiques vectorielles. Ces données représentent la forme et la localisation des entités géographiques figurant sur les images sous la forme de géométries, parfaitement superposées aux contenus des images. Cette représentation géométrique des entités géographiques s'accompagne d'attributs. Il est in-

téressant de savoir si les informations sémantiques provenant de ces référentiels de données vectorielles peuvent être utiles, et si c'est le cas, comment peuvent-elles être exploitées et encodées par un descripteur ?

Le travail présenté dans cet article vise à faire correspondre les images aériennes représentant des zones urbaines et rurales des territoires français, prises à 15 ans d'intervalle. Leur contenu reflète l'évolution des paysages au cours des années, ce qui fait de leur mise en correspondance une tâche compliquée. L'idée principale est de concevoir un descripteur qui conserve toutes les informations nécessaires pour retrouver l'apparence de la scène dans le temps et selon des conditions d'acquisition variables. L'ensemble des données dont nous disposons contient tout d'abord des paires d'images qui se correspondent à deux dates différentes, et par ailleurs des données géographiques vectorielles, sortes de cartes sémantiques. La Figure 1 montre un exemple de ces données. Ceci nécessite une méthode capable de faire la distinction entre des images sémantiquement proches, car toutes contiennent des éléments similaires, mais également être robuste aux changements, tel que l'apparition et la disparition des objets dans le temps ou encore les effets saisonniers.

Nos contributions sont les suivantes. Dans un premier temps, nous évaluons les performances des méthodes existantes pour cette nouvelle tâche de recherche par contenu d'images représentant une même zone géographique à des temporalités différentes. Nous utilisons un ensemble de données appelé FR0419 créé à cet effet. Nous déterminons la modalité la plus importante et évaluons différents scénarios de fusion de données multimodales. Ensuite, nous proposons un nouveau descripteur compact pour les données multimodales et nous l'affinons sur notre nouvel ensemble de données. Le cœur de notre méthode est formé par un réseau siamois qui prend en entrée des paires d'images. Ces paires, en plus des images naturelles, contiennent des étiquettes sémantiques correspondant à chaque image, ce qui rend notre approche multimodale. Le réseau génère un descripteur par paire d'images qui analyse la similitude tout en étant robuste aux changements survenus dans le temps. Ce descripteur est de faible dimensionnalité mais suffisamment puissant pour que la recherche finale d'images correspondantes puisse être effectuée par des méthodes non supervisées relativement simples telles que l'algorithme des k-plus proches voisins (k-NN).

2 Problème et contexte

2.1 Base de données

Nous proposons d'utiliser des données multimodales pour la tâche de recherche par contenu d'images représentant une même zone géographique à différentes temporalités. La modalité supplémentaire que nous utilisons est l'information sémantique sur la scène. Ces données sémantiques décrivent les entités géographiques de la scène, leur catégorie, leur forme et leur localisation [1]. Google maps ou OpenStreetMap [9] constituent de bons exemples de telles données sémantiques. L'institut national de l'information géographique et forestière (IGN) possède de vastes campagnes de données aériennes couvrant le territoire français. Toutes ces données géospatiales ont été annotées manuellement auparavant afin de créer des bases de données vectorielles qui sont maintenant disponibles sur le site Web de l'IGN sous la forme de fichiers au format shapefile [2]. Dans ce travail, nous utilisons la base de données vectorielle BD TOPO, dont la première version France entière remonte à 2004. Pour les images aériennes, nous utilisons la base BD ORTHO, disponible elle aussi sur la période 2004-2019. Des fonds d'images aériennes plus anciennes sont disponibles [3]. Les images aériennes sont de resolution 50 cm/pixel. Nous sélectionnons des zones d'une taille de 1 kilomètre carré, telles que les images représentent la même zone géographique. Ce scénario n'est pas réaliste pour un usage réel à large échelle, mais il nous permet de tester et de démontrer la robustesse des descripteurs face aux changements et évolutions du paysage.

Type d'entités géographiques	couleur RVB	# 2004	# 2019
route	(255,165,0)	380731	326882
église ou chapelle	(255,255,0)	1292	2195
tour ou donjon	(165,42,42)	56	17
fort ou blockhaus	(128,128,128)	481	734
chateau	(0,0,0)	36	245
bâtiment indifférencié	(255,0,0)	251294	3475104
surface d'eau	(0,0,255)	28043	12040
terrain de sport	(138,43,226)	1409	2859
cimetière	(75,0,130)	928	1299
zone de végétation	(0,255,0)	224101	164435
aéroport ou aérodrome	(95, 2, 31)	23	65
chemin de fer	(255,0,255)	3308	3972

TABLE 1 – Statistiques principales des catégories sémantiques. Moselle 2004-2019.

Trois départements français voisins situés au nord-est de la France ont été sélectionnés pour nos expériences : la Moselle, le Bas-Rhin et la Meurthe-et-Moselle. Les données résultantes couvrent différents types de paysages avec des distributions similaires. Dans les données vectorielles, nous avons sélectionné plusieurs catégories d'objets géographiques (voir Tableau 1). A noter que le nombre d'objets annotés peut différer considérablement, en partie en raison des différentes stratégies d'annotation entre 2004 et 2019 et en partie en raison de l'évolution du paysage. Nous avons enfin rastérisé ces données : cette opération fait perdre les possibilités d'analyse spatiale sur les géométries et les attributs des entités géographiques, mais permet de traiter ces données directement comme les images aériennes.

1. fournies à l'aide d'un ensemble de coordonnées exprimées dans un systeme de coordonnées de référence, aussi utilisé pour géoréférencer les images.

2. Les données IGN sont disponibles pour des travaux de recherche via la licence "Recherche et Enseignement" : https ://geoservices.ign.fr.

3. https ://remonterletemps.ign.fr.

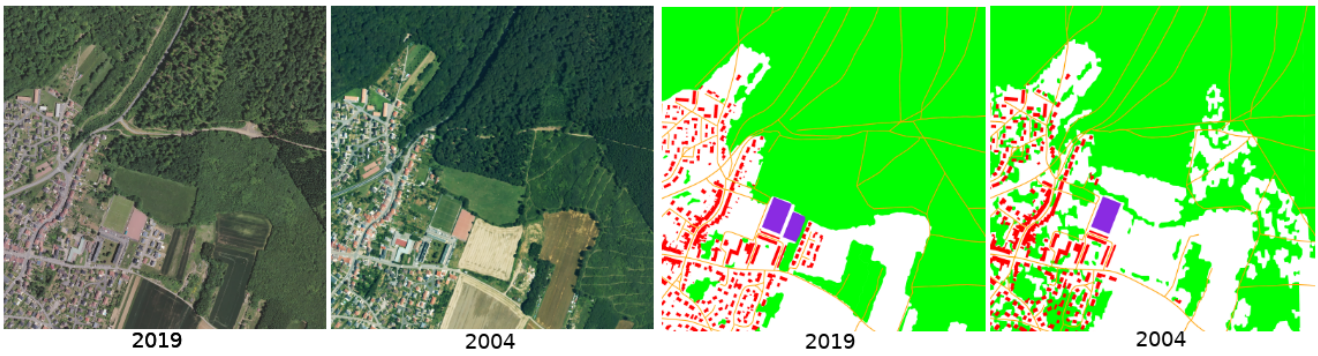

FIGURE 1 – Évolution des images et des données géographiques vectorielles 2004-2019. A noter également les changements saisonniers et les différences de conditions d'éclairage.

Usage	Département	# d'images
entraînement	Moselle	6000
validation	Bas-Rhin	4430
test	Meurthe-et-Moselle	5855

TABLE 2 – Configuration des jeux de données.

2.2 Problématique de recherche

Tout d'abord, nous cherchons à déterminer dans quelle mesure les descripteurs basés sur des réseaux convolutionnels existants peuvent être utilisés pour faire correspondre les images aériennes à différentes temporalités, et ensuite quelles informations (visuelles, sémantiques ou les deux) sont les plus pertinentes. Ensuite, nous testons différentes stratégies de fusion pour encapsuler toutes les modalités dans un seul descripteur D^R, où R est la dimension. Nous utilisons les images de 2019 comme images de requête par rapport à leurs homologues de 2004 et utilisons différentes zones géographiques pour séparer les données d'apprentissage, de validation et de test pour nos expériences (voir Tableau 2). Enfin, nous présentons notre nouveau descripteur basé sur le réseau convolutionnel siamois.

2.3 Contexte et méthodes existantes

Les bases de données géographiques vectorielles peuvent être mises à profit comme sources d'annotations sémantiques pour les images aériennes couvrant les mêmes zones de l'espace. Récemment, avec les progrès de la segmentation réalisée par des architectures CNN, la tâche de segmentation entièrement automatisée est devenue possible [10]. Les annotations sémantiques sont une source supplémentaire d'informations qui peuvent potentiellement améliorer la géolocalisation, la classification, la recherche d'images (prises selon différents points de vue à différentes dates) selon le contenu. La combinaison de différentes sources d'informations et modalités pour améliorer les modèles (réseau de neurones) a été explorée dans [11, 12, 13, 14]. Cependant, à notre connaissance, l'utilisation de données multimodales pour mettre en correspondance des images aériennes prises au cours d'années différentes est nouvelle. Un problème de recherche connexe est la localisation basée image, où les acquisitions diffèrent en termes de conditions de visualisation et souffrent d'un large éventail de distorsions [5, 6, 15]. Traditionnellement, l'objectif est de représenter les caractéristiques de l'image comme des vecteurs de caractéristiques robustes. Les travaux récents se concentrent sur la famille des méthodes [5, 6, 14, 16] qui sont entrainées pour apprendre ces caractéristiques. Cette problématique comprend également les travaux de mise en correspondance des images au niveau du sol avec les images aériennes pour obtenir des estimations de localisation dans le cadre des changements de point de vue [17, 18, 19]. Cependant, toutes ces méthodes sont conçues pour gérer des fonctionnalités spécifiques à une scène géographique ou un objet et ne sont pas spécifiques à la mise en correspondance des images prises à des moments différents, où les scènes peuvent ne pas contenir un seul objet clé mais sont plutôt composées d'entités géographiques qui se répètent d'une image à l'autre. Dans ce cas seul l'agencement dans l'espace permet de distinguer les images représentant une même zone de celles représentant des paysages similaires, mais non identiques. Par conséquent, ces méthodes ne peuvent pas être appliquées directement dans le cas de grands changements de paysage au fil du temps avec des caractéristiques d'image non-distinctes qui se répètent d'une image à l'autre.

Descripteurs d'images basés sur les réseaux de neurones convolutionnels (CNN). Notre objectif est de concevoir des descripteurs de caractéristiques d'image qui sont à la fois discriminants en ce qui concerne le contenu sémantique, et robustes aux distorsions d'images. Alimentée par les progrès de l'apprentissage automatique, la recherche met l'accent sur le passage à des méthodes qui sont entraînées. En particulier, il a été démontré que les architectures CNN pré-entraînées peuvent être utilisées comme extracteurs de descripteurs pour des ensembles de données sur lesquels le réseau n'a jamais été entraîné. Généralement, les chercheurs recadrent le CNN à la dernière couche convolutive et le considèrent comme un extracteur de descripteur dense global. Les descripteurs résultants peuvent être comparés à l'aide de métriques relativement simples, telles que la distance euclidienne, ou introduits dans un classificateur. Il a été observé que cela fonctionnait bien pour la recherche d'instances [20, 21] et la reconnaissance de texture [22].

Des travaux récents ont démontré que le réglage fin des

descripteurs CNN existants sur de nouvelles données ainsi que de légères modifications d'architecture conduisent à de nouvelles améliorations des descripteurs. Arandjelovic et al. [23] a proposé une architecture NetVLAD, qui contient une nouvelle couche VLAD généralisée, inspirée d'une représentation d'image couramment utilisée dans la recherche d'images par contenu classique [24]. La couche VLAD regroupe les descripteurs extraits dans une représentation d'image fixe. Ses paramètres peuvent être appris par rétro-propagation et ils peuvent ensuite être utilisés par n'importe quel réseau. De même, GEM [5] est un descripteur basé sur ResNet [25] qui utilise un CNN pour extraire des éléments d'image visuels, ainsi qu'une couche d'agrégation de caractéristiques personnalisées et effectue un réglage fin pour réaliser la tâche de recherche d'images par contenu. Enfin, pour la recherche d'images multimodales avec des points de vues différents par contenu est le descripteur DELF (DEep Local Feature) [6] constitue la proposition la plus récente de l'état de l'art. L'algorithme est de nature locale et extrait des entités autour de points d'intérêts densément localisés à l'aide d'un bloc d'extraction d'entités ResNet50. Les cartes de caractéristiques obtenues sont considérées comme une grille dense de descripteurs locaux et sont utilisées pour la tâche de recherche d'images par contenu avec des points de vues différents. L'approche est très performante pour gérer les distorsions de perspective mais n'est cependant pas adaptée pour les images comportant des éléments répétitifs non caractéristiques telles que les images aériennes.

Multimodalité. La combinaison de différentes modalités ou types d'information pour améliorer les performances des algorithmes à modalité unique existants semble intuitivement simple [14, 13]. Cependant, il est difficile en pratique de combiner les différents niveaux de bruit et de gérer les conflits entre les modalités. De plus, les modalités peuvent avoir une influence quantitative différente sur la prédiction. La fusion des informations complémentaires peut être effectuée à différentes étapes de la chaîne de traitement. La stratégie la plus simple ainsi souvent utilisée est la fusion des listes de réponses retournées par chaque modalité unique, appelée fusion tardive. L'autre stratégie consiste à fusionner les différentes dimensions au début du processus (d'où le nom de fusion précoce) et à calculer ensuite un seul descripteur pour une combinaison de données multimodales. Bianco et al. [26] ont comparé ces deux stratégies de fusion pour l'appariement des caractéristiques visuelles sur plusieurs ensembles de données. Leurs résultats indiquent que les performances des différentes techniques de fusion dépendent fortement de l'ensemble de données. Cependant, ce travail utilise uniquement des descripteurs classiques tels que SIFT [27] et ne s'étend pas aux méthodes modernes basées sur les CNNs. Les informations multimodales peuvent être utilisées pour guider une transformation d'image aérienne en vue transversale exécutée par un GAN comme dans [11]. Schonberger et al. [14] utilisent la sémantique avec les informations visuelles et de profondeur pour obtenir un descripteur de scène 3D dense robuste à de forts changements de point de vue et des changements dans la géométrie de la scène. Les informations de profondeur jouent un rôle plus important car la correspondance finale est 3D-3D. Un autre exemple d'utilisation de données multimodales est ContextNet [28], où les auteurs exploitent des descripteurs d'image basées sur un CNN ainsi que des informations textuelles créant un graphe et utilisant une méthode de description de nœud [29] pour classer des objets.

Les descripteurs d'images aériennes peuvent grandement bénéficier des données des différentes modalités disponibles. Cela a été démontré avec succès par Audebert et al. [13], où de meilleures cartes sémantiques ont été obtenues en utilisant une architecture de type encodeur-décodeur, des images ainsi que des données sémantiques provenant du jeu de données Open Street Map. À notre connaissance, c'est le seul travail qui utilise directement des annotations sémantiques avec des images dans un contexte d'image aérienne. Li et al. [30] utilise avec succès une approche fondée sur la fusion tardive de modalités pour améliorer l'annotation des images aériennes. Cependant, les différentes modalités proviennent de la même image et sont extraites par différents algorithmes.

Réseaux siamois [31] Ce sont des réseaux de neurones contenant deux ou plusieurs composants de sous-réseau identiques. Les réseaux siamois sont un choix populaire pour les problèmes liés aux problèmes d'apprentissage ponctuels, lorsqu'un seul échantillon est disponible pour chaque classe [32, 33]. L'architecture de réseau siamois avec une métrique d'apprentissage appropriée vise à construire une intégration où deux entrées similaires vont produire des descripteurs proches et deux entrées différentes des descripteurs différents. [34, 35]. Les réseaux siamois ont été utilisés avec succès dans plusieurs applications basées sur des images de suivi des piétons [36, 37] et ré-identification [38] pour la segmentation [39] des objets similaires au premier plan simultanément et même dans certains scenarii basés sur du texte [40]. A notre connaissance, il n'y a que deux travaux qui utilisent les réseaux siamois dans le contexte des images aériennes. Daudt et al. [41] propose des architectures qui effectuent la détection de changements sur des paires multi-temporelles d'images d'observation de la Terre. Tout en étant similaire dans le choix de l'architecture, l'article de Daudt et al. cherche à détecter le changement (ou l'absence de changement comme la tâche de classification binaire) sur des images temporelles, là où nous essayons de trouver des caractéristiques robustes aux changements. La modalité des images est aussi différente, et les auteurs montrent que le réseau siamois produit de meilleurs résultats que le fine-tuning d'un réseau qui traite une paire des images concaténées. Récemment Hu et al. [19] ont proposé l'architecture CVM-Net pour la tâche de recherche des vues aériennes à partir des vues terrestres. Nous proposons d'utiliser une architecture de réseau personnalisée de type siamois pour l'en-

traîner sur des données multimodales géographiques afin d'obtenir une description qui capturera les caractéristiques visuelles de l'image et apprendra à ignorer les changements temporels survenus. L'approche basée sur un réseau siamois nous permet de former des descripteurs sur un ensemble de données à correspondance unique, et la conception de l'architecture de base est conçue pour bénéficier des informations multimodales.

3 Performances de référence

À notre connaissance, il n'y a pas de descripteurs d'images pré-entraînés sur des images aériennes disponibles au public, ni d'études quantitatives comparant des représentations d'images spécifiquement pour des images aériennes multimodales. Nous établissons donc notre propre référence de performances de base en utilisant les descripteurs d'images existants.

Les descripteurs d'images basés sur les CNN ont largement surpassé les méthodes traditionnelles pour les tâches de reconnaissance d'image. Les méthodes de référence sont ResNet [25] et GEM [5]. Nous testons leurs performances pour la mise en correspondance d'images représentant la même zone géographique et déterminons quel type de modalité d'information influence le plus la performance. Le descripteur basé sur ResNet50 classique [25] est pré-entraîné sur les données ImageNet [2]. Des variantes architecturales de la famille ResNet ont été déployées dans de nombreux descripteurs d'images récents tels que DELF [6] et GEM [5]. Ce dernier est pré-entraîné sur les jeux de données Oxford5k, Paris6k, Roxford5k et RParis6k [7]. Dans nos expériences, nous utilisons la sortie de la dernière couche convolutionnelle suivie d'une couche de *maxpooling* pour obtenir les caractéristiques de chaque image pour ResNet ; pour GEM, nous utilisons le modèle pré-entraîné de bout en bout, tel que fourni par les auteurs. Les deux descripteurs sont globaux, ce qui est intuitivement intéressant pour notre cas d'application où les caractéristiques individuelles des objets sont similaires et difficiles à capturer en raison de la résolution relativement faible.

La première étape de cet article consiste donc à évaluer les descripteurs GEM et ResNet à travers une étude comparative. Nous évaluons aussi les combinaisons de descripteurs et les paramètres utilisés pour la tâche de mise en correspondance d'images aériennes à différentes temporalités. La méthode mise en œuvre pour cette comparaison est schématisée dans la Figure 2.

3.1 Configuration de la chaîne de traitement

En utilisant la nouvelle base de données FR0419 comme référence, nous confirmons d'abord que les informations multimodales sont utiles pour améliorer la précision de la recherche d'images par contenu. L'expérience est également conçue pour établir un niveau de performance de référence pour évaluer les apports éventuels de nouvelles propositions. Nous évaluons la précision des réponses du système de recherche d'images par contenu fondé sur des descripteurs standards pour différentes modalités de données. Nous utilisons les descripteurs sans aucune modification et redimensionnons les images d'entrée à la taille de 512x512 pour ResNet50 et 1024x1024 pour GEM comme il est proposé par les auteurs [5]. De plus, pour le descripteur basé sur ResNet50, nous évaluons trois scénarios :

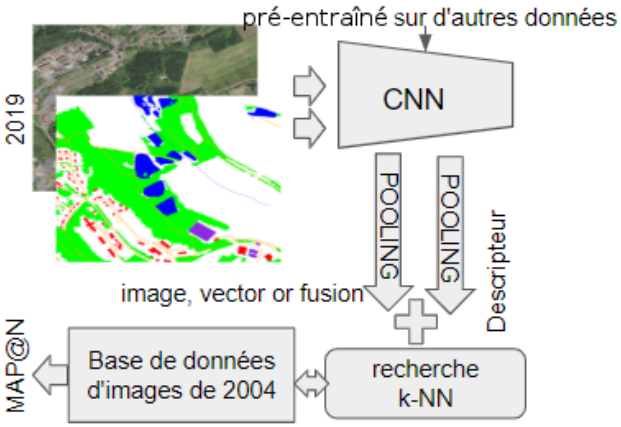

FIGURE 2 – Algorithme utilisé pour évaluer les performances des descripteurs GEM et ResNet sur la tâche de recherche d'images, selon les modalités utilisées en entrée.

- la fusion précoce : une couche convolutionnelle fusionne l'image naturelle et l'image des annotations sémantiques avant le passage dans le réseau.
- la fusion par concaténation : les descripteurs obtenus à partir d'une image naturelle et d'une image contenant des annotations sémantiques sont obtenus séparément et sont concaténés.
- la fusion tardive : les informations multimodales sont utilisées séparément et les images correspondantes sont sélectionnées sur la base des k-nn voisins les plus proches pour les deux modalités.

Nous avons réalisé nos tests sur les images créées à partir des annotations sémantiques contenant 3 canaux originaux (RVB) et les avons converties en un masque en niveau de gris (NB). Ce dernier s'inspire de l'idée que les zones de ces images représentant des entités géographiques devraient avoir plus d'importance que les zones de fond.

La précision moyenne est une mesure couramment utilisée pour évaluer les algorithmes de recherche d'images par contenu. Puisque nous avons une seule correspondance par image dans notre jeu de données, l'équation map@N est simplement : $map@N = \sum_{n=1}^{N} \frac{r}{n}$ où R est égal à 1 si l'image retournée est correcte et 0 sinon. Une valeur élevée de map@N indique de meilleures performances. Dans toutes nos expériences, N est égal à 5. Nous avons réalisé nos tests avec différentes distances (euclidienne, cosinus) et différents types de techniques de standardisation (L1, L2, standardisation et pas de normalisation), et différents types de fusion.

Performance de référence : Les tableaux 3 et 4 résument les résultats de nos expériences utilisant respectivement les descripteurs GEM et ResNet50. Les descripteurs basés sur des annotations sémantiques donnent de meilleurs résul-

modalité utilisée	normalisation	fusion	distance	R	map@5 Moselle	Bas-Rhin	Meurthe & Moselle	moyenne
visuelle	n/a	none	cosine	2048	0,48	0,70	0,65	0,60
semantique	n/a	none	cosine	2048	0,67	0,57	0,72	0,65
vis + semantique	mean std	concat	cosine	4096	0,66	0,69	0,71	0,69
vis + semantique	none	conv precoce	cosine	2048	0,62	0,64	0,64	0,63
vis + semantique	none	tardive	cosine	2048	0,76	0,75	0,84	0,79

TABLE 3 – Précision map@5 obtenue par le descripteur ResNet50 [25]

modalité utilisée	normalisation	fusion	distance	R	map@5 Moselle	Bas-Rhin	Meurthe & Moselle	moyenne
visuelle	mean std	none	cosine	2048	0,54	0,63	0,59	0,60
semantique	mean std	none	euclidean	2048	0,63	0,64	0,61	0,63
vis + semantique	none	concaten	cosine	4096	0,66	0,69	0,71	0,68
vis + semantique	none	conv precoce	cosine	2048	0,51	0,52	0.56	0,53
vis + semantique	none	tardive	cosine	2048	0,75	0,73	0,84	0,77

TABLE 4 – Précision map@5 obtenue par le descripteur GEM pré-entrainé [5], architecture resnet101-gem-reg-whiten.

tats en termes de valeur de map@5 que les descripteurs basés sur des images naturelles. Les meilleurs résultats de map@5 ont été obtenus en utilisant les deux modalités, ce qui confirme que ces deux types d'informations sont complémentaires et pertinentes pour la mise en correspondance des images. De plus, la combinaison tardive de données visuelles et sémantiques permet d'améliorer encore les résultats pour les deux descripteurs testés. Nos expériences ont montré que l'utilisation d'une image réalisée à partir d'annotations sémantiques et convertie en niveaux de gris donne des performances similaires avec une image RVB. Dans l'ensemble, les résultats obtenus ne sont cependant pas très bons, montrant les limites des approches existantes basées sur CNN.

4 Architecture siamoise

Notre architecture de réseau siamois multimodal est illustrée schématiquement dans la Figure 3. L'architecture a deux copies de la fonction G_W, qui partagent le même ensemble de paramètres W. Elle se compose de deux branches et d'un module de coût. Un module de perte dont l'entrée est la sortie de cette architecture est placé dessus.

L'architecture du réseau est conçue pour gérer des entrées multimodales et affiner les descripteurs de la tâche de recherche d'images par contenu. L'entrée de l'ensemble du système est une paire d'images multimodales $(X_1, S_1; X_2, S_2)$ de la taille HxL et une étiquette Y. Une branche traite des paires d'images représentant la même zone géographique à deux dates distinctes (2019-2004), l'autre traite une paire d'images représentant des zones différentes, mais réalisées de la même année (2019). Les images sont traitées, donnant deux sorties $G_W(X_1, S_1)$ et $G_W(X_2, S_2)$. Le module de coût génère alors la distance $DW(G_W(X_1, S_1), GW(X_2, S_2))$. La fonction de perte combine DW avec l'étiquette Y pour produire la valeur de perte scalaire en fonction de l'étiquette Y :

$$\mathcal{L}(W) = -\frac{1}{N}\sum_{n=1}^{N} L(W, (Y, X_1, X_2, S_1, S_2)^n) \qquad (1)$$

$$DW(X_1, S_1, X_2, S_2) = |G_W(X_1, S_1) - G_W(X_2, S_2)| \qquad (2)$$

$$\mathcal{L}(W, (Y, X_1, X_2, S_1, S_2)^n) = Ylog(p) + (1 - Y)log(1 - p) \qquad (3)$$

où p, la probabilité prédite sur DW passe par une couche de classification connectée de dimensionalite R avec l'activation $sigmoid$, et n est le nombre de paires.

La première couche du réseau est une couche convolutionnelle avec des poids prédéfinis par l'équation 4.

$$BW = \begin{vmatrix} 0.5, 0.0, 0.0 \\ 0.0, 0.5, 0.0 \\ 0.0, 0.0, 0.5 \\ 0.5, 0.5, 0.5 \end{vmatrix}, RVB = \begin{vmatrix} 0.5, 0.0, 0.0 \\ 0.0, 0.5, 0.0 \\ 0.0, 0.0, 0.5 \\ 0.5, 0.0, 0.0 \\ 0.0, 0.5, 0.0 \\ 0.0, 0.0, 0.5 \end{vmatrix} \qquad (4)$$

Cette premiere couche avec des poids crée une image de la taille H,W,3, à partir d'une paire d'images : X_1, S_1 (S_1 l'image est en mode RVB ou en mode niveaux de gris, ce qui donne 3 ou 1 canaux en entrée). Les poids sont prédéfinis pour lancer et accélérer l'entraînement, ensuite ils sont entraînés avec l'ensemble du réseau. La partie principale du réseau est ResNet50 pré-entraîné sur ImageNet. La sortie de la dernière couche convolutionnelle de ResNet est ensuite passée à travers 3 autres couches convolutionnelles C_1 - C_3 suivies d'une couche entièrement connectée. Nous avons constaté que l'utilisation de l'activation $tanh$ et la normalisation par lots dans les couches convolutives ajoutées donne un meilleur résultat.

L'extraction de paires d'images difficiles est essentielle pour permettre un entraînement appropriée des réseaux siamois [42]. Nous avons adopté la stratégie d'apprentissage suivante. Toutes les 5 époques, le score map@5 est calculé pour l'ensemble des données d'entraînement. Les paires d'images difficiles sont celles pour lesquelles le système renvoie des réponses incorrectes (ne correspondant pas à la même zone géographique). Nous avons choisi la stratégie qui consiste à créer des lots composés simplement de paires aléatoires et d'images avec la map@5 pour cette image est inférieure à 0.5. Cette valeur signifie selon la map@N que la correspondance la plus proche estimée par k-NN n'était ni la première ni la deuxième zone retournée. Si à la fin de la formation de réseau il n'y a pas suffisamment d'échantillons difficiles pour remplir tous les lots, nous réutilisons simplement au hasard les échantillons difficiles pour parcourir l'ensemble des données.

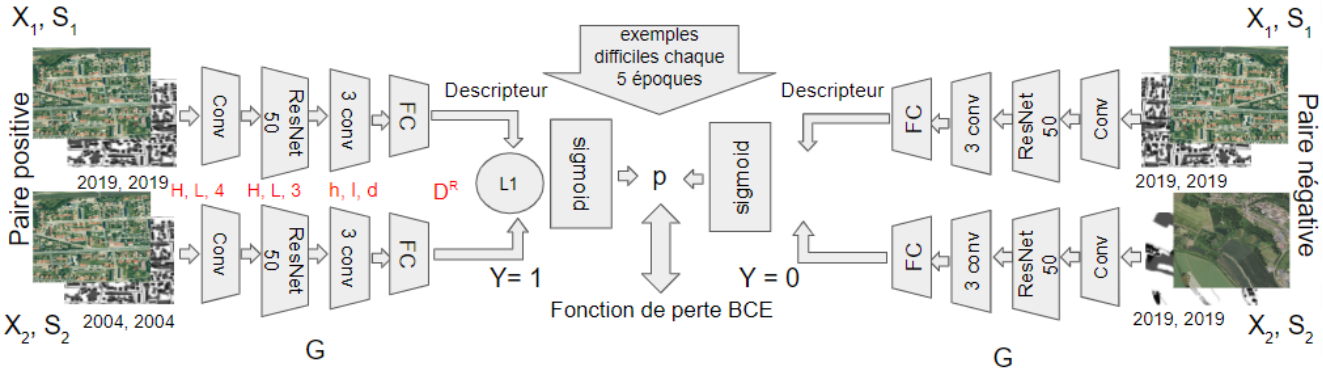

FIGURE 3 – Architecture siamoise pour améliorer les descripteurs d'images croisés exploitant des données multimodales.

5 Expériences et évaluations

Dans cette section, nous décrivons le cadre expérimental utilisé pour comparer notre méthode aux résultats de référence obtenus avec des algorithmes de descripteur d'image standards sur notre nouveau jeu de données FR0419. Nous entraînons une architecture siamoise complète personnalisée en utilisant les images de la Moselle. Les images de Bas-Rhin sont utilisées comme données de validation et celles de la Meurthe-et-Moselle forment l'ensemble de test. Tout au long de nos expériences, l'architecture de la Figure 3 est utilisée et la taille d'entrée des images aériennes naturelles et celles générées à partir de données vectorielles est de 256x256. La stratégie de fusion précoce est adoptée pour combiner les données multimodales et pouvoir effectuer un entraînement de bout en bout. La première couche convolutionnelle est pré-initialisée. Les 3 couches convolutives C_1-C_3 qui suivent le ResNet ont des noyaux 3x3 et le nombre de filtres est égal à 1024, 512 et 256, la dimension du descripteur final est D^{256}. On utilise la normalisation par lots des couches convolutives ajoutées. Nous avons essayé d'ajouter et de supprimer la normalisation par lots dans les couches ajoutées et nous avons observé que dans l'ensemble, ce n'est pas une opération nécessaire, mais le réseau s'entraîne plus rapidement avec la normalisation.

Choix des paramètres du réseau Nous avons testé l'influence de l'utilisation d'un canal de couleur unique ou d'une image sémantique RVB, celle de la *batchnorm* dans toutes les couches convolutives ajoutées et la taille optimale du descripteur final parmi 128, 256 et 512 (voir le tableau 5). Nous avons obtenu des résultats similaires pour les dimensions 128 et 256. En revanche, une augmentation supplémentaire de la taille du descripteur final entraîne une diminution de la valeur de map@5.

Le réseau apprend à prédire si deux descripteurs multimodaux correspondent à la même zone géographique à des temporalités différentes, en fonction de la distance $L1$ entre eux. L'idée est similaire à la perte de contraste [34] couramment utilisée dans les réseaux siamois, mais nous avons trouvé que cette approche fonctionnait mieux.

Nous utilisons la perte BCE, avec l'optimiseur Adam [43] avec un taux d'apprentissage variable en fonction des couches. Nous identifions à nouveau les échantillons difficiles toutes les 5 époques sur la base des scores map@5 sur l'ensemble d'entraînement. L'algorithme k-NN avec la distance cosinus est utilisé pour retrouver les images les plus similaires à partir d'une requête. Pendant l'entraînement, chaque lot est composé de 12 paires d'images positives et négatives, dont la moitié sont sélectionnées au hasard et la moitié sont des images difficiles. L'ensemble de données d'apprentissage contient 6000 paires. A chaque époque, nous parcourons toutes les images, sélectionnant chaque fois des négatifs aléatoires et ajoutant des échantillons difficiles dans un lot. Les ensembles de validation et de test ont 4430 et 5855 paires d'images. Le modèle final a été entraîne à 100 époques.

Le score map@5 sur l'ensemble de validation est utilisé pour sélectionner les meilleurs paramètres de descripteur. Ceux-ci sont utilisés pour obtenir le score final map@5 sur les données du test. Nous avons également procédé à une validation croisée en échangeant les départements pour l'entraînement et la validation et pour tester et ré-entraîner le réseau à partir de zéro une fois que les paramètres d'entraînement optimaux ont été sélectionnés. Une fois qu'un réseau est entraîné, nous utilisons une seule branche pour calculer le descripteur multimodal.

Comparaison avec l'apprentissage contrastif. Pour montrer l'intérêt de l'architecture siamoise proposée, nous avons fait des tests en utilisant la perte contrastive NT-XENT proposée par [44] sur des paires d'images 2004-2019. Dans ce cas, nous avons utilisé l'architecture spécifiée dans l'article [12] avec le ResNet50 classique, en ajoutant une seule couche convolutionnelle pour fusionner les modalités comme cela été defini précédemment par l'équation 4. Le réseau a été entraîné sur 100 époques avec une taille de lots de 26 paires, la perte NT-XENT avec τ égal à 100, sans normalisation. Nous avons utilisé l'optimiseur Adam avec le taux d'apprentissage 1e-3 et decay égal à 1e-6. La méthode nécessite une augmentation des données. Nous avons utilisé la rotation et le changement de couleur aléatoires des paires d'images. Après l'entraînement du réseau, comme cela a été fait dans la méthode originale, la sortie de ResNet50 suivie par regroupement moyen (*average pooling*) a été utilisée comme le descripteur final d'une paire d'images. Le tableau 5 contient l'évaluation quantitative de résultats obtenus et montre que la méthode est capable d'apprendre les caractéristiques de l'image sur l'ensemble d'apprentissage mais se généralise moins bien que l'architecture proposée sur les nouvelles données pour la tâche de mise en correspondance d'images. A noter également que contrairement à l'apprentissage contrastif utilisant la fonction de perte NT-Xent, notre méthode ne bénéficie pas de la suppression de toutes les couches ajoutées dans l'architecture finale. Nous attribuons ces résultats au fait que la configuration de données bénéficie de l'absence de la couche *pooling*, qui supprime les informations spatiales.

6 Résultats

Le tableau 5 résume les scores map@5 obtenus par la méthode présentée dans la section 4. Ils montrent que l'architecture de réseau siamoise proposée améliore les résultats de référence et est capable de traiter des images temporellement décalées. Nous at-

département	perfomance de référence	tests			
	map@5	lot	map@5	cross-validation	map@5
Moselle	0,76	entraînement	0,82	validation	**0,92**
Bas-Rhin	0,75	validation	0,89	test	**0,91**
Meurthe-and-Moselle	0,84	test	**0,93**	entraînement	0,91

TABLE 5 – Meilleure performance de référence comparée aux performances de notre modèle affiné.

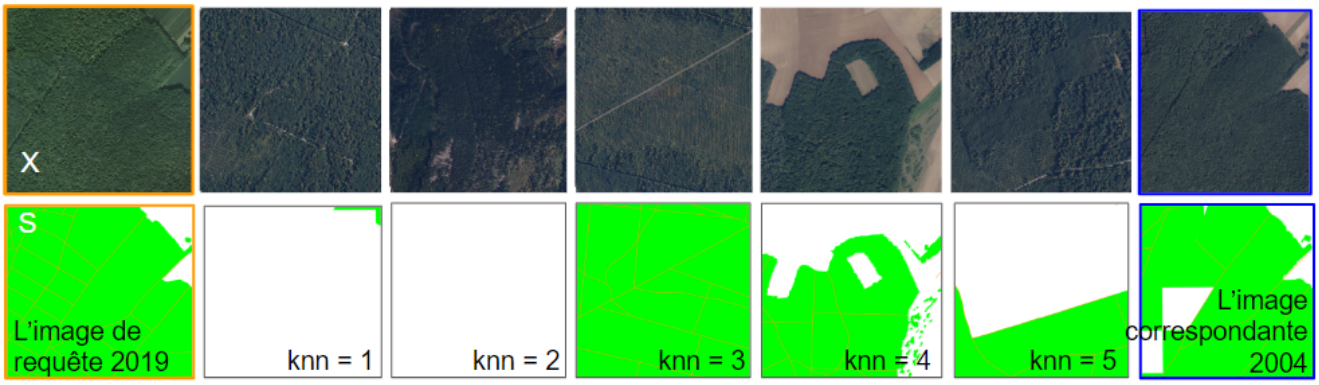

FIGURE 4 – Les résultats non-corrects de recherche d'images par k-NN. La bonne correspondance pour l'image de 2019 dans la base de données 2004 est surlignée en bleu.

R	map@5 entraînement Moselle	map@5 validation Bas-Rhin
128	0,92	0,88
256	0,93	0,92
512	0,86	0,70

TABLE 6 – Précision d'entraînement maximale obtenue selon la taille du descripteur final.

méthode	R	map@5		
		Moselle entraînement	Bas-Rhin validation	M et M tests
NT-Xent	128	0,77	0,62	0,52
NT-Xent	2048*	0,83	0,73	0,61
La nôtre	128	0,92	0,88	0,92
La nôtre	2048	0,61	0,71	0,58

TABLE 7 – Comparaison avec [44].

teignons une précision moyenne de 0,91 pour nos ensembles de données de validation et de test, ce qui représente une amélioration de 10% par rapport aux meilleurs résultats de référence. De plus, le descripteur résultant est presque 10 fois plus compact que ses homologues de l'état de l'art avec seulement 256 dimensions au lieu de 2048 (ou même 4096 s'ils sont concaténés), ce qui lui permet de mieux évoluer pour les grandes bases de données et de réduire les temps de réponse du système de recherche d'images par contenu. Le descripteur peut être encore réduit à 128 dimensions et donner des résultats légèrement moins précis (nous avons observé des résultats très similaires pour les dimensions 128 et 256, comme indiqué dans le tableau 5). De plus, la résolution des images d'entrée est également inférieure à celles utilisées par ResNet (512) et GEM (1024) plus tôt. Malgré les performances considérablement améliorées après l'entraînement, il y a encore 10% de correspondances erronées renvoyées par le k-NN. La Figure 4 illustre un résultat typique d'images renvoyées par erreur par notre système. Nous avons observé que l'algorithme n'est pas très bon pour les zones rurales avec des zones de boisement, où le descripteur ne peut pas capturer la configuration exacte du paysage et renvoie à la place d'autres zones forestières visuellement similaires même après l'entraînement.

7 Conclusion

Dans cette étude, nous avons abordé un nouveau problème de comparaison d'images aériennes prises à des moments éloignés en temps. Nous avons proposé une nouvelle approche d'apprentissage à partir de données multimodales pour affiner n'importe quel descripteur d'images basé sur un CNN concu pour cet effet.

Nous avons effectué une comparaison complète de différentes stratégies permettant d'utiliser des informations multimodales dans notre algorithme et proposé une architecture de réseau siamois personnalisée permettant d'affiner un descripteur d'image basé sur un CNN pour la tâche à accomplir. Le descripteur résultant est suffisamment puissant pour faire la distinction entre des images sémantiquement proches et il est robuste vis-à-vis de l'évolution temporelle du paysage. Les résultats expérimentaux montrent que la méthode proposée peut gérer avec succès les différences entre les images dues à l'évolution du paysage et aux changements saisonniers. Il améliore considérablement les résultats de référence et les descripteurs d'images les plus récents disponibles sur dans la littérature. Nous avons proposé une manière originale d'utiliser deux modalités dans un seul descripteur. La méthode est applicable à tout extracteur de fonctionnalités basé sur CNN, facile à utiliser et simple. Nos résultats fournissent une référence pour les comparaisons futures et montrent que la correspondance des images acquises à plusieurs années de décalage à l'aide de descripteurs basés sur CNN reste un problème ouvert. Bien que ce travail se limite intentionnellement à la mise en correspondance d'images représentant exactement les mêmes emplacements géographiques pour établir comment les changements temporels affectent les descripteurs d'images, il représente une introduction pour les recherches futures visant à explorer la mise en correspondance d'images multimodales à différentes dates et selon différents points de vue pour un scénario réel. Une direction de recherche intéressante est notamment l'étude de correspondance entre les données visuelles X et sémantiques S.

Références

[1] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," tech. rep., Citeseer, 2009.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[3] "Catalogue collectif de france. fonds lapie de photographies aériennes." https ://ccfr.bnf.fr/portailccfr/ark :/06871/0033535.

[4] "Alegoria : Advanced linking and exploitation of digitized ge0graphic iconographic heritage." http ://www.alegoria-project.fr.

[5] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[6] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3456–3465, 2017.

[7] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris : Large-scale image retrieval benchmarking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5706–5715, 2018.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization : Improving particular object retrieval in large scale image databases," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, 2008.

[9] OpenStreetMap contributors, "Planet dump retrieved from https ://planet.osm.org." https ://www.openstreetmap.org, 2017.

[10] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6054–6068, 2017.

[11] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3501–3510, 2018.

[12] K. Chen, K. Fu, X. Gao, M. Yan, W. Zhang, Y. Zhang, and X. Sun, "Effective fusion of multi-modal data with group convolutions for semantic segmentation of aerial imagery," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3911–3914, IEEE, 2019.

[13] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint learning from earth observation and openstreetmap data to get faster better semantic maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 67–75, 2017.

[14] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6896–6906, 2018.

[15] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," *arXiv preprint arXiv :1907.05021*, 2019.

[16] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2614–2620, IEEE, 2017.

[17] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," *arXiv preprint arXiv :1903.12351*, 2019.

[18] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 627–637, 2017.

[19] S. Hu, M. Feng, R. M. Nguyen, and G. Hee Lee, "Cvm-net : Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267, 2018.

[20] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marques, and X. Giro-i Nieto, "Bags of local convolutional features for scalable instance search," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327–331, ACM, 2016.

[21] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5007–5015, 2015.

[22] T.-Y. Lin and S. Maji, "Visualizing and understanding deep texture representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2791–2799, 2016.

[23] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad : Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

[24] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3304–3311, IEEE Computer Society, 2010.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[26] S. Bianco, D. Mazzini, D. P. Pau, and R. Schettini, "Local detectors and compact descriptors for visual search : a quantitative comparison," *Digital Signal Processing*, vol. 44, pp. 1–13, 2015.

[27] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features.," in *iccv*, vol. 99, pp. 1150–1157, 1999.

[28] N. Garcia, B. Renoust, and Y. Nakashima, "Understanding art through multi-modal retrieval in paintings," *arXiv preprint arXiv :1904.10615*, 2019.

[29] A. Grover and J. Leskovec, "node2vec : Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, ACM, 2016.

[30] K. Li, C. Zou, S. Bu, Y. Liang, J. Zhang, and M. Gong, "Multi-modal feature fusion for geographic image annotation," *Pattern Recognition*, vol. 73, pp. 1–14, 2018.

[31] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, pp. 737–744, 1994.

[32] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, 2015.

[33] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1983–1991, 2017.

[34] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[35] T. Trzcinski, J. Komorowski, L. Dabala, K. Czarnota, G. Kurzejamski, and S. Lynen, "Scone : Siamese constellation embedding descriptor for image matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

[36] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*, pp. 850–865, Springer, 2016.

[37] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking : Siamese cnn for robust target association," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–40, 2016.

[38] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European conference on computer vision*, pp. 791–808, Springer, 2016.

[39] P. Mukherjee, B. Lall, and S. Lattupally, "Object cosegmentation using deep siamese network," *arXiv preprint arXiv :1803.02555*, 2018.

[40] S. Maheshwary and H. Misra, "Matching resumes to jobs via deep siamese network," in *Companion Proceedings of the The Web Conference 2018*, pp. 87–88, International WWW Conferences Steering Committee, 2018.

[41] R. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Détection dense de changements par réseaux de neurones siamois," 2018.

[42] B. Harwood, B. Kumar, G. Carneiro, I. Reid, T. Drummond, *et al.*, "Smart mining for deep metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2821–2829, 2017.

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

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv :2002.05709*, 2020.