Apprentissage contrastif de représentations d'images guidé par les relations spatiales

Logan Servant, Michaël Clément, Laurent Wendling, Camille Kurtz

Apprentissage contrastif de représentations d'images guidé par les relations spatiales

Logan Servant[1] Michaël Clément[2] Laurent Wendling[1] Camille Kurtz[1]

[1] LIPADE, Université Paris Cité (France) prenom.nom@u-paris.fr
[2] LaBRI, CNRS, Bordeaux INP, Université de Bordeaux (France) prenom.nom@labri.fr

Résumé

Les informations spatiales contenues dans les images sont d'une importance cruciale pour la vision par ordinateur. Les approches actuelles pour les tâches spatialisées sont optimisées via la sémantique des relations spatiales. Cependant, les jeux de données de relations spatiales contiennent de nombreuses ambiguïtés au niveau des annotations (polysémie, référentiels) qui peuvent détériorer l'apprentissage. Les représentations d'images obtenues à partir de l'apprentissage supervisé contiennent des informations spatiales de moins bonnes qualités, car elles sont enchevêtrées avec d'autres modalités, comme la sémantique. Nous proposons C-SIP (Contrastive Spatial-Image Pre-training), une approche visant à obtenir des représentations d'images plus spatialisées, plus en accord avec la perception humaine, grâce à un apprentissage contrastif, visant à aligner les modalités visuel et spatial avec deux encodeurs. Le code source du projet est disponible : https://github.com/Logan-wilson/CSIP.

Mots Clef

Apprentissage contrastif, Relation spatiale.

Abstract

The spatial information contained in images is of critical importance for computer vision tasks. Current state-of-the-art approaches dealing with spatially-related tasks are typically trained in a supervised manner with the semantic information of spatial relations. However, datasets containing spatial relations contain many ambiguities(e.g. polysemy, reference frames) which might deteriorate the representation learning step. The image representations obtained from this training setup carry poor spatial information, as they are entangled with other modalities, such as semantic information. We introduce C-SIP (Contrastive Spatial-Image Pre-training), an approach aiming to learn better spatially-aware image representations, more in agreement with human perception, thanks to a contrastive training procedure aiming to align visual and spatial modalities of two encoders.

Keywords

Contrastive learning, spatial relations.

1 Introduction

Les représentations d'image traditionnelles contiennent une fusion d'informations visuelles et sémantiques, bien qu'elles ne suffisent pas toujours pour quelques tâches de vision par ordinateur. Tandis que les tâches de reconnaissance d'objets requièrent une vérité terrain d'objet dont la sémantique n'est pas ambigu, la reconnaissance de relations spatiales entre paires d'objets est composé de plus d'ambiguïtés qui doivent être traités pour des résultats optimaux. Les jeux de données existants sur les relations spatiales entre objets [1, 2, 3, 4] comportent des ensembles de relations qui ne sont pas toujours mutuellement exclusive (Visual Genome [2] propose *under*, *underneath*, *beneath*, *below* comme classes distinctes), pouvant provoquer des problème lors de l'apprentissage et lors de l'évaluation. La polysémie des relations peut également réduire les performances, lorsqu'une relation peut valoir pour deux configurations spatiales différentes : *Sur* peut être utilisé pour décrire *un objet sur une table* ou *une image sur un mur* [5].

Les annotations de ces jeux de données souffrent également à cause des points de vue linguistiques (*Reference frames*). En effet, une relation spatiale entre deux objets (comme *à gauche de*) peut être considéré selon deux points de vue : objet ou caméra, qui changent la manière dont un individu décrit une scène. Le point de vue caméra décrit les relations spatiales selon ce que la caméra voit, tandis que le point de vue objet prend l'objet comme référentiel principal de la description spatiale. Des études en psychologie cognitive et en linguistique ont démontré que l'utilisation de points de vue différents est souvent dépendant du contexte de l'image [6] ou de la langue natale de l'annotateur [7]. La grande majorité des jeux de données de relations spatiales ne définissent pas explicitement quel point de vue est utilisé mais ont tendance à contenir plus de relations au point de vue caméra que objet [8, 9].

Une façon de surmonter ces limitations est d'apprendre des représentations d'images sans utiliser les labels de relations comme vérité terrain, et d'entraîner un modèle en utilisant une stratégie auto-supervisée, qui peut permettre aux représentations d'images d'être plus en phase avec la perception humaine pour la compréhension de scène et des configurations complexes entre les objets. De grands modèles vision-langage [10, 11] entraînés de manière auto-

supervisée ont montré qu'ils étaient performants pour de nombreuses tâches, mais étonnamment décevants pour les tâches liées au raisonnement spatial. En effet, les descriptions des images ne contiennent pas de relations exhaustives entre les objets et souffrent de l'utilisation de différents référentiels des relations [12, 8]. Ces méthodes contrastives peuvent donc ne pas être optimales pour les représentations spatiales. Nous proposons dans ce travail de considérer l'information spatiale d'une image comme une modalité distincte que nous utilisons dans une nouvelle méthode contrastive de paires image-spatiale.

La modalité spatiale est une représentation qui modélise et décrit les relations spatiales entre les objets d'une image grâce aux descripteurs de position relative (RPD). Les RPD sont des représentations spatiales algorithmiques qui ont été utilisées historiquement pour la reconnaissance des relations spatiales et qui ont été largement étudiées au cours des dernières décennies [13]. Grâce à ces représentations spatiales, nous visons à entraîner un modèle multimodal utilisant des paires (image-spatial) avec une nouvelle méthode d'entraînement, que nous appelons C-SIP (*Contrastive Spatial-Image Pre-training*). En résumé, nos contributions sont les suivantes [1] :

- Nous proposons C-SIP, une stratégie d'entraînement qui permet d'obtenir de meilleures représentations spatiales des images, plus en phase avec la perception humaine ;
- Nous démontrons qu'utiliser des RPD comme modalité spatiale dans un processus contrastif est une stratégie puissante pour améliorer les capacités spatiales des encodeurs d'image ;
- Nous fournissons des résultats expérimentaux sur trois *downstream tasks* appliqué à des *datasets* publics démontrant que notre approche auto-supervisée est pertinente pour des tâches spatiales.

La Figure 1 illustre la composition de la méthode C-SIP. Dans la suite, nous passons d'abord en revue l'état de l'art lié aux informations spatiales dans les descripteurs algorithmiques et les réseaux de neurones (Sec. 2). Sec. 3 présente alors notre méthode se basant sur une nouvelle méthode d'apprentissage contrastive. Nous démontrons ensuite dans la Sec. 4 grâce à une étude expérimentale les forces de notre méthode avec les résultats associés (Sec. 5). Nous concluons ensuite au sein de la Sec. 6.

2 État de l'art

Dans cette section, nous commençons par décrire l'état de l'art sur les représentations spatiales algorithmiques qui sont liées à la modalité spatiale de notre étude. Nous examinons ensuite l'intégration des informations spatiales, explicitement et implicitement, dans les réseaux de neurones profonds.

Descripteurs de position relative. Les descripteurs de position relative visent à décrire la configuration spatiale entre les objets dans un référentiel centré sur la caméra. Le formalisme de Freeman [15] a ouvert la voie dans ce domaine de recherche en décrivant les relations spatiales dans les images avec des relations élémentaires (par exemple, topologiques, directionnelles). Parmi les RPD existants, l'histogramme de forces [16] calcule l'interaction de la force entre les objets dans chaque direction sur des images binaires (c'est-à-dire segmentées) grâce à un niveau de force, un paramètre fournissant différents poids dans l'image. Des travaux ultérieurs ont permis de traiter davantage de configurations spatiales en combinant plusieurs descripteurs, tels que le ϕ-descriptor [17], basé sur les intervalles de temps d'Allen, ou le modèle de ligne radiale étendue [18]. Le Bandeau de Forces (FB) [19] calcule l'histogramme de Forces avec une gamme de niveaux de force représentés dans une matrice, qui s'est avérée efficace avec les CNN pour classer les relations spatiales. Chaque combinaison de descripteurs présente des avantages et des inconvénients. Le FB permet une description spatiale fine des objets d'une image et peut facilement être intégré dans des réseaux de neurones. En raison de sa précision, nous considérons le FB comme notre modalité spatiale que nous utiliserons dans nos encodeurs spatiaux, mais d'autres RPD peuvent être employés en fonction du cas d'étude.

Apprentissage de représentation spatiale explicite. Les informations spatiales explicites dans les réseaux de neurones consistent généralement en de petits modules (tels que le perceptron multi-couches) qui reçoivent en entrée les informations des boîtes englobantes afin de créer des représentations spatiales aidant les modèles dans des tâches spatialisées telles que la reconnaissance de relations spatiales. Il existe deux types d'entrées : les coordonnées des boîtes englobantes et les images binaires. Les coordonnées des boîtes peuvent inclure les positions, les rapports de taille et l'IoU [3, 4], ou des métriques dérivées de la méthode de régression de boîte englobante [20, 21] utilisée dans la détection d'objets. Les images binaires spatiales représentent la configuration spatiale des boîtes englobantes des objets dans les images [22, 23], qui sont en général données en entrée de petits CNN. Quelle que soit la solution considérée, ces représentations spatiales ne tiennent pas compte de la forme des objets. Le barycentre de la forme d'un objet peut ne pas coïncider avec le centre d'une boîte englobante donnée, si l'objet n'est pas centré sur la boîte. Cette différence peut conduire à des représentations spatiales d'image incorrectes, en particulier lorsque les objets ont un chevauchement important entre le sujet et l'objet, comme c'est souvent le cas dans les jeux de données de relations spatiales. Notre représentation spatiale diffère : elle est calculée sur le masque de segmentation et non sur la base d'un masque de boîte englobante, ce qui permet au modèle de travailler avec une représentation spatiale plus proche du contexte de l'image.

Apprentissage implicite de la représentation spatiale. L'apprentissage implicite d'informations spatiales dans les modèles a été abordé de manière approfondie avec des stratégies d'apprentissage auto-supervisé. Des tâches prétextes

1. Contributions acceptées à WACV 2025 [14].

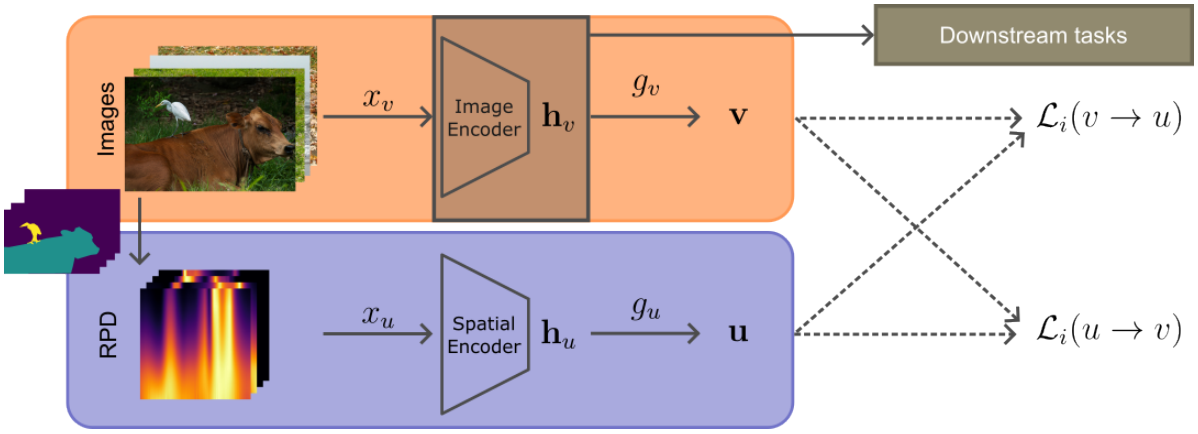

FIGURE 1 – Aperçu de la méthode C-SIP (Contrastive Spatial-Image Pre-training), visant à obtenir des représentations d'images plus spatialisées, en alignant les modalités visuelles (boîte orange) et spatiales (boîte violette). L'architecture de C-SIP permet aux deux encodeurs d'aligner leur représentation grâce aux fonctions de coût bidirectionnelles $\mathcal{L}_i(v \to u)$ et $\mathcal{L}_i(u \to v)$.

telles que la résolution de puzzle [24], la prédiction de rotation d'images [25] ou du contexte [26] ont permis d'obtenir des représentations puissantes contenant des informations spatiales. Plus récemment, l'apprentissage auto-supervisé a été utilisé pour entraîner de larges modèles de vision-langage [10, 11]. Cependant, des approches comme CLIP [10] ont donné des résultats médiocres sur des tâches liées au spatial, non pas en raison de l'architecture des modèles mais des données sur lesquelles ils ont été entraînés [12]. SpatialVLM [27] se concentre spécifiquement sur les informations spatiales et construit un nouveau jeu de données de questions-réponses visuelles, ce qui permet au modèle d'obtenir de meilleures capacités de raisonnement spatial. Notre contribution s'inscrit dans la lignée des méthodes d'apprentissage contrastives multimodales. C-SIP vise à aligner les représentations de deux encodeurs de modalité pour obtenir des représentations d'images plus spatialisées. Puisque les descriptions sémantiques des images ne fournissent pas d'informations spatiales suffisantes [12] et pour éviter le chevauchement des informations spatiales et sémantiques, nous alignons les modalités qui ne se concentrent pas sur les informations sémantiques.

3 CSIP : Contrastive Spatial-Image Pre-training

Pour apprendre de meilleures représentations spatiales, nous proposons le framework C-SIP dont la pierre angulaire est de guider la supervision d'un encodeur d'images à partir des informations spatiales contenues dans une scène. Pour ce faire, nous alignons les modalités visuelles et spatiales de façon auto-supervisée. Pour mettre en œuvre cette stratégie, nous calculons pour chaque image le descripteur de position relative des deux objets segmentés les plus saillants, capturant ainsi leur organisation spatiale. Pendant l'entraînement, les représentations des encodeurs d'image et spatiale sont ensuite alignées grâce à un apprentissage

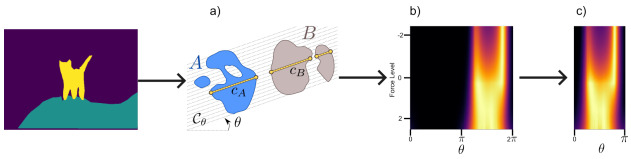

FIGURE 2 – Calcul du Bandeau de Forces symmétrique. (a) Calcul de l'Histogramme de Forces [16] entre les objets A et B. (b) Bandeau [19]. (c) Bandeau de Forces Symétrique. Les images avec des configurations spatiales similaires ont des bandeaux de forces similaires qui peuvent être utilisés pour déterminer les similarités spatiales entre les images. Les fortes valeurs de forces sont jaunes, les valeurs nulles sont noires.

contrastif. L'encodeur d'image entraîné à l'aide de ce framework peut ensuite être intégré dans toute *downstream task* nécessitant des informations spatiales. La Figure 1 illustre les différentes étapes du framework C-SIP. Nous décrivons d'abord le RPD que nous considérons, puis nous présentons la stratégie de supervision de manière plus détaillée.

3.1 Bandeau de Forces symétrique

Le Bandeau de Forces (FB) [19] est une extension de l'Histogramme de Forces [16] qui modèlise la relation spatiale entre deux objets segmentés selon un point de vue intrinsèque. Soient 2 points a et b à une distance d, la force entre a et b est calculée $\phi_r(a,b) = \frac{1}{d^r}$ avec r le niveau de force considéré. Celui-ci permet d'attribuer plus d'importance aux points qui sont soit plus proches ($r > 0$) soit plus éloignés ($r < 0$). Plutôt que d'étudier toutes les paires de points entre les deux objets, la force est considérée entre deux segments I et J et leur distance D^θ_{IJ} :

$$f_r(I,J) = \int_{D^\theta_{IJ}+|J|}^{|I|+D^\theta_{IJ}+|J|} \int_0^{|J|} \phi_r(u-v)dvdu \qquad (1)$$

Soient deux objets binaires A et B et une droite L_θ d'angle θ, la force d'attraction le long de L_θ entre les deux objets est défini par :

$$\mu_{AB}(L_\theta, r) = \sum_{I \in C_A} \sum_{J \in C_B} f_r(I, J) \tag{2}$$

avec C_A (respectivement C_B) l'ensemble de segments qui intersectent avec L_θ et l'objet A : $L_\theta \cap A$ (respectivement B : $L_\theta \cap B$). L'attraction globale $F_r^{AB}(\theta)$ est calculé sur l'ensemble des droites parallèles à un angle θ qui intersectent avec les deux objets (Fig. 2). L'Histogramme de Forces est alors calculé pour l'ensemble des angles $\Theta = \{0, 0 + \delta_\theta, \ldots, 2\pi - \delta_\theta\}$, où δ_θ représente le pas constant entre chaque angle.

Le Bandeau de Forces est une généralisation de cet histogramme qui est calculé pour un ensemble de niveau de forces $R = \{r_{\min}, r_{\min} + \delta_r, \ldots, r_{\max}\}$, avec δ_r un pas constant entre les niveaux de forces. Le FB est donc une fonction à deux variables : l'angle θ et le niveau de force r, qui permet une représentation spatiale plus précise. Puisque nous comptons entraîner les modèles sans labels, l'ordre des objets dans la relation n'a pas d'importance mais peut conduire à des ambiguïtés. Pour résoudre ce problème, nous proposons le Bandeau de Forces symétrique (sFB), qui permet à la configuration spatiale *A-B* d'être identique à celle de *B-A* :

$$sFB^{AB} = \begin{pmatrix} \xi(\theta_0, r_s) & \cdots & \xi(\theta_{0+i\delta_\theta}, r_s) & \cdots & \xi(\theta_{\pi-\delta_\theta}, r_s) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi(\theta_0, r_{s+j\delta_r}) & \cdots & \xi(\theta_{0+i\delta_\theta}, r_{s+j\delta_r}) & \cdots & \xi(\theta_{\pi-\delta_\theta}, r_{s+j\delta_r}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi(\theta_0, r_e) & \cdots & \xi(\theta_{0+i\delta_\theta}, r_e) & \cdots & \xi(\theta_{\pi-\delta_\theta}, r_e) \end{pmatrix} \tag{3}$$

où $\xi(\theta, r)$ correspond à la moyenne des forces exercées par chaque objet sur l'autre :

$$\xi(\theta, r) = \frac{F_r^{AB}(\theta) + F_r^{BA}(\theta)}{2} = \frac{F_r^{AB}(\theta) + F_r^{AB}(\theta + \pi)}{2} \tag{4}$$

Le sFB conserve les mêmes propriétés d'échelle, de rotation et de translation que les descripteurs originaux [19, 16], qui sont fondamentales pour décrire avec précision les relations spatiales. Cette représentation 2D permet donc d'utiliser la représentation spatiale comme une image pour la stratégie contrastive multimodale décrite dans la section suivante.

3.2 Apprentissage de représentation spatiale contrastif

Considérons $x_v$une image dans un *batch*, x_u la représentation spatiale de x_v tel que $x_u = sFB(x_v)$ obtenue grâce aux masques de segmentations de l'image. E_v est l'encodeur visuel, E_u est l'encodeur spatiale. En accord avec les méthodes d'apprentissage auto-supervisé contrastif récentes [10, 11], nous cherchons à aligner les modalités dans un espace latent commun. Pour ce faire, nous entraînons conjointement les deux encodeurs et leurs têtes de projection respective, $g_v(\cdot)$ et $g_u(\cdot)$, de tel sorte à maximiser la similarité cosinus des représentations d'images et spatiales des N paires positives (image-spatial) du *batch*, tout en minimisant la similarité cosinus des $(N^2 - N)$ paires négatives, avec N la taille du *batch*.

Avec (x_{vi}, x_{ui}) les paires positives, nous définissions la fonction de coût contrastive *Image-to-Spatial* suivante qui correspond à une entropie croisée modifiée :

$$\mathcal{L}_i(v \to u) = -\log \frac{\exp(\cos(v_i, u_i)/\tau)}{\sum_{k=1}^{N} \exp(\cos(v_i, u_k)/\tau)} \tag{5}$$

de la même manière, la fonction *Spatial-to-Image* :

$$\mathcal{L}_i(u \to v) = -\log \frac{\exp(\cos(u_i, v_i)/\tau)}{\sum_{k=1}^{N} \exp(\cos(u_i, v_k)/\tau)} \tag{6}$$

où $\cos(\cdot, \cdot)$ correspond à la fonction de similarité cosinus, $\tau \in \mathbb{R}^+$ représentation un paramètre de température. Les fonctions de coûts sont similaires à la fonction de coût InfoNCE [28, 29].

Étant donné que certaines représentations spatiales peuvent manquer de variance au sein d'un *batch* (par exemple, des images peuvent avoir des objets à une distance et direction similaires), il est nécessaire d'apporter de la stabilité lors de l'entraînement. Nous le réalisons en introduisant une valeur de *label smoothing* [30] aux deux fonctions de coût décrites précédemment, qui permet également d'améliorer la généralisation.

Modèle	image	Boîte englobante
ResNet18+MLP [31]	✓	
2D [4]		✓
DRNet [22]	✓	✓
ResNet18+2D	✓	✓

TABLE 1 – Modèles considérés pour les expériences. Les modèles prennent en entrée soit le contenu de l'image, soit les informations des boîtes englobantes ou les deux. Chaque modèle est entraîné en supervisé et avec C-SIP.

4 Études expérimentales

Nous évaluons ici l'intérêt de la stratégie C-SIP pour apprendre de meilleures représentations spatiales sur trois *downstream tasks* nécessitant des informations spatiales : (1) Recherche d'Image par le Contenu (CBIR), (2) Reconnaissance de Relations Spatiales (SRR) et (3) Question-Réponse Visuelle (VQA) afin d'évaluer la capacité des modèles pré-entraînés à capturer et modéliser des informations spatiales complexes. Nous décrivons d'abord les modèles considérés avec C-SIP et les modèles de comparaison, puis nous présentons les *datasets* utilisés.

4.1 Modèles utilisés

C-SIP nécessite deux encodeurs : un encodeur spatial f_u et un encodeur d'image f_v (Fig. 1). Nous considérons pour f_u un CNN à 2 couches avec un *padding* circulaire [19] qui reçoit comme entrée les sFB dérivés du contenu de l'image, avec $r_{\min} = -2.24$ et $r_{\max} = 2.24$. Pour démontrer la capacité de la stratégie C-SIP à capturer l'information spa-

Entraînement	Modèle	SpatialSense [4]		UnRel [3]	
		$NDCG_{10}$	$NDCG_{25}$	$NDCG_{10}$	$NDCG_{25}$
Pre-trained	ResNet18 [31]	58,3	59,1	68,7	67,5
	ResNet50 [31]	58,4	58,9	69,2	68,1
	ViT-B/16 [32]	58,8	57,8	69,8	68,2
	$\text{CLIP}_{\text{Image}}$ [10]	60,6	61,1	64,4	64,9
Supervisé	ResNet18+MLP [31]	59,2	58,3	64,7	64,1
	2D [4]	73,3	72,1	77,5	76,0
	DRNet [22]	65,3	64,4	70,6	69,8
	ResNet18+2D	73,6	72,3	78,3	76,9
C-SIP (Nôtre)	ResNet18+MLP [31]	63,7	63,2	69,6	69,3
	2D [4]	76,0	74,5	80,9	79,8
	DRNet [22]	69,5	68,6	75,0	73,9
	ResNet18+2D	**76,3**	**75,6**	**81,5**	**80,5**

TABLE 2 – Résultats quantitatifs (gras = meilleur score) pour la tâche de CBIR.

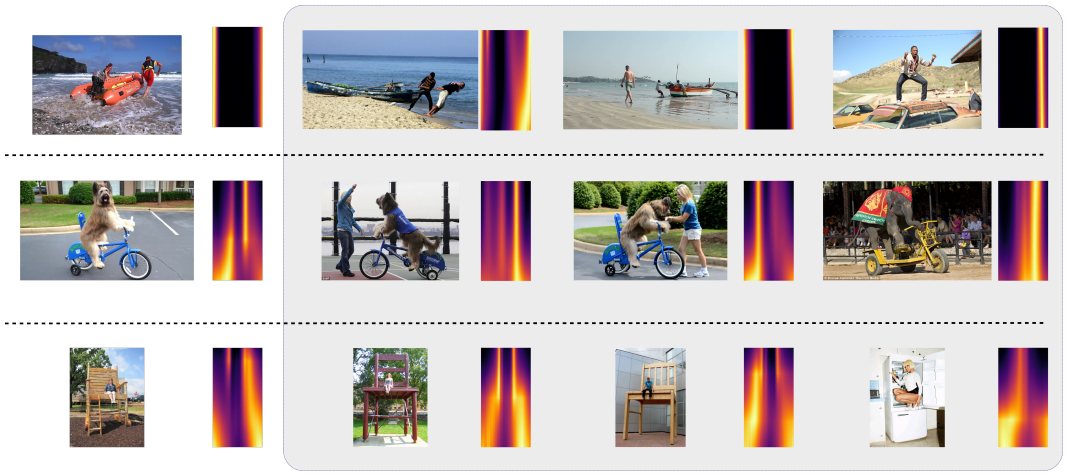

FIGURE 3 – Résultats qualitatifs du meilleur modèle entraîné avec C-SIP (ResNet18+2D, voir Tab. 2) sur la tâche de CBIR sur les *datasets* Unrel et SpatialSense. A chaque image requête (gauche) est associée le top 3 des images retournées (boîte grise). Le bandeau de forces symétrique de chaque image est présenté.

tiale et que la méthode est généralisable à de nombreuses architectures de modèles, nous expérimentons avec différents encodeurs d'images f_v, avec diverses entrées (voir la Table 1). Nous utilisons ResNet18 [31] comme extracteur de caractéristiques d'images, car les CNN ont tendance à être plus performants que les *Transformers* lorsque les *datasets* sont plus petits [33] (nous avons vérifié empiriquement que les modèles plus grands tels que ResNet50 ne compensaient pas leur plus grande profondeur par de meilleurs résultats). Dans un premier temps, nous considérons un ResNet18 [31] entraîné *from scratch* avec un MLP comme modèle *vision-only*, inspiré de [34]. Nous considérons aussi un modèle 2D [4] qui prend en entrée les coordonnées des boîtes englobantes des deux objets segmentés. Comme troisième et quatrième modèles, nous employons deux modèles qui combinent à la fois le contenu de l'image et les informations des boîtes englobantes : Un modèle similaire à DRNet [22], qui utilise une image binaire des boîtes englobantes, et ResNet18+2D.

En guise d'étude comparative, afin de fournir des comparaisons équitables avec notre *pipeline* C-SIP, nous considérons les mêmes modèles de la Table 1 et les entraînons de manière supervisée, avec une fonction de coût d'entropie croisée sur les classes des relations spatiales.

D'un point de vue technique, nous utilisons AdamW pour l'entraînement de tous les modèles, un *batchsize* de 256, des taux d'apprentissage compris entre $1e^{-5}$ et $1e^{-3}$, et une valeur de *label smoothing* de 0,3 sur les deux fonctions d'entropie croisée de C-SIP.

4.2 Jeux de données

Différents jeux de données sont considérés pour les phases d'entraînement et d'évaluation, en fonction des tâches d'évaluation considérées.

Entraînement. Nous entraînons nos modèles de comparaisons sur un sous-ensemble de SpatialSense [4]. SpatialSense contient des relations positives et négatives pour chacune des 9 relations spatiales (*above, under, left of, right*

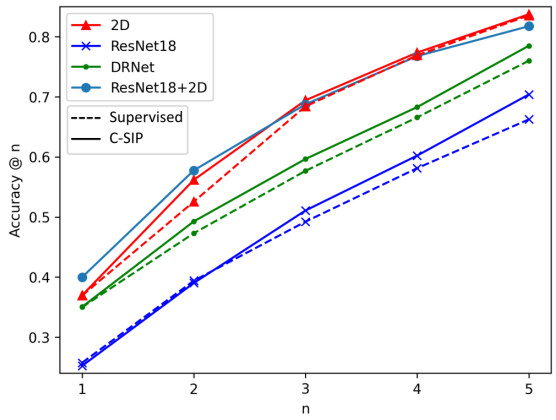

FIGURE 4 – Résultats quantitatifs de la tâche de SRR sur SpatialSense avec la métrique d'*Accuracy*@n pour $n \in \{1, 2, 3, 4, 5\}$. Les lignes en pointillés correspondent aux modèles comparatifs. Les lignes pleines sont les modèles C-SIP.

of, in, on, front of, behind, next to). Nous écartons les relations négatives, ce qui laisse 5433 images que nous divisons en 3606/681/1146 pour l'entraînement, la validation et l'évaluation. Puisque C-SIP ne nécessite pas de vastes jeux de données annotées de triplets de relations, nous tirons parti de la stratégie en entraînant nos modèles avec SpatialSense et un sous-ensemble de COCO [35]. Nous sélectionnons des images avec 3 annotations de segmentation d'objet ou moins et choisissons les deux masques d'objet les plus saillants pour obtenir des images où une relation entre les objets peut exister. Nous obtenons ainsi un sous-ensemble de 35284 images, dans lequel la seule étape supplémentaire consiste à calculer le sFB respectif entre les objets de chaque image. Des masques de segmentation sont nécessaires pour calculer les Bandeaux de Forces. Nous utilisons les masques de segmentation fournis par COCO et nous employons SAM [36] pour obtenir les masques des objets des autres jeux de données.

Évaluation. Selon les tâches d'évaluation, nous utilisons le jeu de test de SpatialSense [4], le jeu de données UnRel [3] et un sous-ensemble du jeu de données GQA [37]. UnRel est un jeu de données SRR de 1071 images naturelles avec des relations inhabituelles telles que (*Bike over woman*) qui vise à démontrer les biais sémantiques appris par les modèles. GQA est un jeu de données VQA à grande échelle avec des images réelles provenant du jeu de données Visual Genome et des paires de questions-réponses équilibrées. Nous utilisons un sous-ensemble de GQA pour notre tâche VQA afin de nous concentrer uniquement sur les questions relatives aux relations spatiales. Le sous-ensemble utilisé contient 56523 questions pour l'entraînement et nous utilisons 7933 questions pour l'évaluation. Les questions spatiales sont du type *verify rel*, c'est-à-dire des questions en oui/non qui visent principalement à vérifier la validité d'un triplet dans une image.

5 Résultats

Nous présentons maintenant les résultats obtenus pour trois *downstream tasks* considérées : CBIR, SRR et VQA.

5.1 Recherche d'image par le contenu

La tâche CBIR est utilisée pour évaluer la capacité des modèles à retrouver des images partageant des configurations spatiales similaires. Nous avons besoin d'une métrique qui évalue la similarité entre les configurations spatiales et utilisons donc la similarité entre les Bandeaux de Forces symétriques via le *Normalized Discounted Cumulative Gain* (NDCG), tel que :

$$\text{DCG}_k = \sum_{i=1}^{k} \frac{\cos(sFB(q), sFB(p_i))}{\log_2(i+1)} \tag{7}$$

avec q l'image requête, p_i l'image retournée au rang i et $\cos(\cdot, \cdot)$ la fonction de similarité cosinus. Outre les modèles de base et C-SIP, nous expérimentons également des modèles pré-entraînés sur ImageNet, tels que ResNet [31], ViT [32] et l'encodeur d'images de CLIP [10]. D'après les résultats de la Table 2, nous remarquons que les modèles pré-entraînés démontrent que la taille des modèles ne permet pas d'obtenir de meilleures capacités de modélisation spatiale. Ensuite, nous observons que les modèles entraînés avec C-SIP surpassent largement leurs homologues supervisés. Par exemple, le modèle *vision-only* (ResNet18+MLP) entraîné avec C-SIP gagne 4,5% sur SpatialSense et 5% sur UnRel, ce qui représente une amélioration plus importante que les autres modèles contenant des informations spatiales, et correspond presque au modèle DRNet supervisé.

Ces résultats suggèrent que l'entraînement des modèles utilisant les informations spatiales permet de meilleures représentations spatiales, que la supervision classique ne peut pas obtenir en raison de l'enchevêtrement des informations dans la représentation. Le modèle 2D supervisé est le meilleur modèle de base pour capturer les configurations spatiales car il n'encode que les informations spatiales. En combinant les informations spatiales et visuelles, le modèle C-SIP semble s'améliorer sur les deux ensembles de données utilisés pour l'évaluation. Les modèles entraînés avec C-SIP ne sont pas biaisés par la sémantique de l'objet ni par la sémantique de la relation grâce à la stratégie auto-supervisée. En outre, la Figure 3 présente les résultats qualitatifs du meilleur modèle C-SIP (ResNet18+2D) sur les deux ensembles de données utilisés pour les évaluations quantitatives. La capacité du modèle à capturer des configurations spatiales similaires sans biais sémantiques est mise en évidence, dans la mesure où les images retournées ne contiennent pas toujours les mêmes objets que dans la requête (chien sur vélo - éléphant sur vélo). Les résultats des deuxième et troisième lignes montrent que les représentations optimisées avec C-SIP correspondent mieux à la perception humaine des configurations spatiales.

Question	e) Is the striped zebra to the left of a bird?	f) Is the helmet to the right of the white motorcycle that is to the right of the person?	g) Are the green palm trees behind the benches near the building?	h) Is the tennis player to the left of the person on the left?
Answer	No	Yes	Yes	No
ResNet18+BERT	Yes	No	No	Yes
ResNet18/C-SIP+BERT	No	No	Yes	No
ResNet18+ ResNet18/C-SIP+BERT	No	Yes	Yes	No

FIGURE 5 – Résultats qualitatifs de la tâche de VQA. Les bonnes réponses sont écrites en vert.

Modèle	Précision
Aléatoire	50.00
Text-only (BERT)	54.64
ResNet18 + BERT (Baseline)	70,23
ResNet18/C-SIP + BERT (Ours)	72,06
ResNet18 + ResNet18/C-SIP + BERT	75,91
ResNet50 + ResNet18/C-SIP + BERT	**77,10**

TABLE 3 – Résultats quantitatifs de VQA sur GQA. ResNet18 correspond au modèle pré-entraîné sur ImageNet. ResNet18/C-SIP est le modèle entraîné avec C-SIP.

5.2 Reconnaissance de relations spatiales

Pour la tâche de SRR, un *linear probing* est réalisé sur les représentations obtenues avec C-SIP pour prédire, à partir de l'image, une relation spatiale parmi les 9 classes de SpatialSense [4]. L'entraînement des modèles de base et le *fine-tuning* de C-SIP sont effectués à l'aide d'une fonction de coût d'entropie croisée. L'évaluation s'opère sur le jeu de test de SpatialSense, comme une tâche de classification dont l'objectif est de classer correctement la relation spatiale pour chaque image. Les résultats de cette expérience sont présentés dans la Figure 4. Nous comparons les résultats de SRR de la stratégie C-SIP (en trait plein) à la stratégie supervisée (en pointillés). Les lignes de couleurs différentes correspondent aux différents modèles de la Table 1. Alors que la précision des modèles tend à être la même pour *acc@1*, lorsque l'on compare les modes d'apprentissage, les modèles optimisés avec C-SIP surpassent systématiquement leurs modèles homologues supervisés lorsque la valeur de n augmente. En raison de leurs stratégies de supervision, les modèles supervisés sont entraînés à reconnaître pour chaque image une seule relation, empêchant les représentations d'obtenir une hiérarchie de validité des relations. Au contraire, grâce au changement de paradigme de C-SIP, les modèles sont plus à même de prédire la relation correcte dans leurs premières propositions.

5.3 Questions-Réponses Visuelles (VQA)

Nous ajoutons à chaque modèle utilisé dans cette étude expérimentale un encodeur de texte pour obtenir une représentation de la question. Un BERT [38] pré-entraîné est utilisé pour obtenir la représentation et nous gelons les encodeurs textuels et visuels pour *fine-tune* un MLP. Celui-ci est entraîné avec une fonction de coût d'entropie croisée binaire et évalué à l'aide d'une métrique de précision. Nous utilisons une technique de fusion par produit scalaire entre les modalités. La Table 3 présente les résultats obtenus. Le modèle textuel seul obtient une plus grande précision que l'aléatoire, ce qui montre un léger biais entre les questions et les réponses, comme on l'observe souvent dans les tâches de VQA. En combinant les modalités, les modèles donnent de meilleurs résultats, atteignant 70% pour le modèle de base. L'utilisation de ResNet18/C-SIP conduit à de meilleurs scores, même si le modèle n'est pas entraîné à l'aide des informations sémantiques et ne peut pas identifier correctement les objets mentionnés dans les questions. Bien que les encodeurs visuels et spatiaux aient la même entrée, la représentation obtenue est très différente en raison des fonctions objectifs. Nous avons poursuivi nos expériences avec la fusion des trois modalités pour montrer que les encodeurs visuel et spatial ne sont pas nécessairement interchangeables, mais plutôt complémentaires dans certaines tâches, comme le démontrent les résultats. La Figure 5 fournit des résultats qualitatifs sur l'ensemble de test de GQA. Tandis que le modèle ResNet18 éprouve des difficultés sur certaines questions, ResNet18/C-SIP réussit mieux sauf sur les questions complexes. La fusion des deux modèles avec l'encodeur textuel permet cependant de corriger les réponses qu'un seul modèle aurait mal déterminées. Ces résultats montrent que les modèles (ResNet18 pré-entraîné et ResNet18/C-SIP) se complètent et parviennent à atteindre ensemble de meilleurs scores sur la tâche de VQA.

6 Conclusion et perspectives

Les modèles de l'état de l'art entraînés avec une supervision classique souffrent de l'ambiguïté des relations sémantiques. Les différentes modalités utilisées (visuel, sémantique, spatial) aboutissent à une représentation enchevêtrée qui pose problème pour les tâches liées au spatial. Nous avons introduit C-SIP (Contrastive Spatial-Image Pre-training), qui se concentre sur les informations spatiales en tant qu'objectif d'entraînement auto-supervisé. C-SIP repose sur une stratégie contrastive multimodale qui aligne les représentations visuelle et spatiale. Empirique-

ment, nous démontrons que C-SIP appliqué à des tâches liées au spatial permet aux modèles d'obtenir de meilleurs scores, en plus d'être plus en phase avec la perception humaine d'une scène. En outre, l'absence d'annotations permet à C-SIP d'être entraîné avec n'importe quel jeu de données d'images. La seule exigence est celle d'un masque de segmentation, qui est être résolue par des modèles fondation de segmentation. Les représentations peuvent être utilisées en coopération avec d'autres modèles plus sémantisés pour obtenir de meilleurs résultats. Nous souhaitons étudier dans des travaux futurs l'entraînement parallèle de C-SIP avec la sémantique des objets pour obtenir un modèle spatio-sémantique pouvant être utilisé pour d'autres tâches telles que la détection de relations visuelles.

Références

[1] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *ECCV*, pp. 852–869, 2016.

[2] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome : Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, pp. 32–73, 2016.

[3] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *ICCV*, pp. 5179–5188, 2017.

[4] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense : An adversarially crowdsourced benchmark for spatial relation recognition," in *ICCV*, pp. 2051–2060, 2019.

[5] W. Liao, B. Rosenhahn, L. Shuai, and M. Y. Yang, "Natural language guided visual relationship detection," in *CVPRW*, pp. 444–453, 2019.

[6] B. Tversky and B. M. Hard, "Embodied and disembodied cognition : Spatial perspective-taking," *Cognition*, vol. 110, no. 1, pp. 124–129, 2009.

[7] S. C. Levinson, "Frames of reference and molyneux's question : Crosslinguistic evidence.," 1996.

[8] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *TACL*, vol. 11, pp. 635–651, 2023.

[9] A. Goyal, K. Yang, D. Yang, and J. Deng, "Rel3D : A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D," in *NeurIPS*, 2020.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[11] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, pp. 4904–4916, PMLR, 2021.

[12] A. Kamath, J. Hessel, and K. Chang, "What's "up" with vision-language models ? investigating their struggle with spatial reasoning," in *EMNLP*, pp. 9161–9175, ACL, 2023.

[13] Y. Wang, H. Peng, Y. Xiong, and H. Song, "Spatial relationship recognition via heterogeneous representation : A review," *Neurocomputing*, vol. 533, pp. 116–140, 2023.

[14] L. Servant, M. Clément, L. Wendling, and C. Kurtz, "Contrastive learning of image representations guided by spatial relations," in *WACV*, pp. XXX–XXX, 2025.

[15] J. Freeman, "The Modelling of Spatial Relations," *CGIP*, vol. 4, no. 2, pp. 156–171, 1975.

[16] P. Matsakis and L. Wendling, "A new way to represent the relative position between areal objects," *IEEE PAMI*, vol. 21, no. 7, pp. 634–643, 1999.

[17] P. Matsakis, M. Naeem, and F. Rahbarnia, "Introducing the ϕ-descriptor-a most versatile relative position descriptor," in *ICPRAM*, pp. 87–98, 2015.

[18] L. Servant, C. Kurtz, and L. Wendling, "An extension of the radial line model to predict spatial relations," in *VISAPP*, pp. 187–195, 2023.

[19] R. Deléarde, C. Kurtz, and L. Wendling, "Description and recognition of complex spatial configurations of object pairs with Force Banner 2D features," *PR*, vol. 123, p. 108410, 2022.

[20] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *CVPR*, pp. 5532–5540, 2017.

[21] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, "Relationship proposal networks," in *CVPR*, pp. 5678–5686, 2017.

[22] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *CVPR*, pp. 3076–3086, 2017.

[23] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *WACV*, pp. 381–389, 2018.

[24] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, vol. 9910, pp. 69–84, 2016.

[25] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[26] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, pp. 1422–1430, 2015.

[27] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. J. Guibas, and F. Xia, "Spatialvlm : Endowing vision-language models with spatial reasoning capabilities," in *CVPR*, pp. 14455–14465, IEEE, 2024.

[28] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv :1807.03748*, 2018.

[29] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *MLHC*, 2022.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, pp. 2818–2826, 2016.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words : Transformers for image recognition at scale," in *ICLR*, 2021.

[33] Z. Li, C. Xie, and E. D. Cubuk, "Scaling (down) clip : A comprehensive analysis of data, architecture, and training strategies," *arXiv preprint arXiv :2404.08197*, 2024.

[34] M. Haldekar, A. Ganesan, and T. Oates, "Identifying spatial relations in images using convolutional neural networks," in *IJCNN*, pp. 3593–3600, 2017.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO : Common objects in context," in *ECCV*, pp. 740–755, 2014.

[36] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *ICCV*, pp. 4015–4026, October 2023.

[37] D. A. Hudson and C. D. Manning, "GQA : A New Dataset for Real-World Visual Reasoning and Compositional Question Answering," in *CVPR*, pp. 6700–6709, 2019.

[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT : Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, pp. 4171–4186, 2019.