Université Paris Cité

Master Vision et Machine Intelligente

État de l'art

Modélisation des relations spatiales à partir de données raster

Auteurs :

El Haj Samitt EBOU
Yassine FEKIH HASSEN

Encadrants :

Logan Servant, Camille Kurtz, Laurent Wendling

Année universitaire 2025-2026

Université Paris Cité - Master 2 VMI

Table des matières

1 Introduction

Les relations spatiales constituent un élément fondamental dans la perception humaine et dans la compréhension visuelle. Bien avant leur formalisation en vision par ordinateur, elles ont été étudiées dans des domaines tels que la psychologie cognitive, la linguistique et les sciences cognitives, où elles sont reconnues comme des primitives essentielles pour organiser l'espace et ensuite raisonner sur celui-ci. Les travaux fondateurs de Kuipers [1] proposent l'idée de cartes cognitives permettant aux individus de structurer leur environnement spatial de manière abstraite, au-delà de la simple perception sensorielle. Du point de vue linguistique, Landau et Jackendoff [2] ainsi que Vieu [3], montrent que les langues naturelles reposent sur un ensemble de relations spatiales symboliques (*à gauche, dans, entre, derrière, etc.)*, qui permettent de décrire des configurations spatiales mêmes complexes. Ces observations ont fortement influencé la manière dont l'informatique et l'intelligence artificielle ont tenté de formaliser et modéliser ces relations. Dans ce contexte, l'apport majeur de Freeman [4] a été de proposer la première liste formelle de relations spatiales élémentaires destinées à être manipulées par un système. Ces relations directionnelles (*left of, above...*), topologiques (*inside, touching*) ou métriques (*near, far*) constituent encore aujourd'hui une catégorisation importante. Freeman souligne aussi l'importance de modéliser l'imprécision et la gradualité inhérentes à ces relations, notamment via la théorie des ensembles flous de Zadeh [5]. Cet aspect est crucial, selon la forme des objets, leur distance ou leur concavité, une même configuration peut être décrite par plusieurs relations possibles. Historiquement, les travaux en modélisation spatiale en informatique se sont alors organisés selon deux grandes familles. D'un côté, les approches qualitatives, fondées sur des représentations symboliques (*RCC8, 9-intersections, méréotopologie*), capables d'exprimer la nature logique des relations mais limitées dans leur capacité à quantifier l'imprécision spatiale. De l'autre, les approches quantitatives, qui mesurent explicitement la configuration spatiale des objets à l'aide de modèles graduels tels que l'histogramme d'angles, l'histogramme de forces ou les descripteurs dérivés. Ces approches quantitatives ont progressivement permis d'intégrer la forme complète des objets, leur géométrie interne et la variabilité continue des relations spatiales, constituant la base des représentations dites *Relative Position Descriptors (RPD)*, encore largement utilisées aujourd'hui. Avec l'essor de l'apprentissage automatique et profond, un nouveau paradigme est apparu : reconnaître les relations spatiales directement à partir d'images ou à partir de représentations intermédiaires et hybrides, combinant géométrie, vision et modèles neuronaux. Cet état de l'art propose ainsi une analyse des principales approches, depuis les descripteurs spatiaux algorithmiques jusqu'aux modèles apprenants modernes, en montrant comment ces différentes méthodes tentent de capturer, explicitement ou implicitement, la richesse et l'ambiguïté propres aux relations spatiales. Afin d'organiser cet état de l'art, nous structurons la suite du document selon une certaine taxonomie des approches de modélisation des relations spatiales. Nous présentons d'abord les principaux domaines d'applications dans lesquels ces relations jouent un rôle fondamental (Section 2). La Section 3 décrit ensuite les trois grandes familles de méthodes (les représentations symboliques, géométriques explicites et les approches apprenantes). La Section 4 discute des problématiques et limites transversales rencontrées par ces approches, avant

de proposer une comparaison de ces différentes méthodes dans la Section 5.

2 Applications et contextes

La reconnaissance de relations spatiales entre objets occupe une place importante dans plusieurs tâches de vision par ordinateur. Ces relations fournissent un niveau d'interprétation intermédiaire entre la détection d'objets et la compréhension sémantique d'une scène. Elles constituent un élément fondamental pour structurer la perception visuelle, guider le raisonnement spatial et améliorer les capacités descriptives des modèles visuels.

2.1 Détection de relations visuelles et génération de graphes de scènes

La détection de relations visuelles (*Visual Relationship Detection, VRD*), constitue l'une des premières motivations directes pour la modélisation des relations spatiales. L'objectif est d'identifier, à partir d'une image, des triplets (sujet,prédicat,objet), décrivant les interactions spatiales ou sémantiques entre entités. Cette idée a conduit au développement de la génération de graphes de scènes [6] (*Scene Graph Generation, SGG*), popularisée par l'apparition du dataset *Visual Genome* [7]. Dans ce cadre, une image est représentée comme un graphe dont les nœuds correspondent aux objets et les arêtes aux relations, ce qui offre une représentation structurelle plus riche. Cette formalisation a permis l'introduction de modèles capables de raisonner sur l'ensemble des objets d'une scène plutôt que sur des paires isolées. Les réseaux de neurones de graphes (*GNN*) ont notamment été utilisés pour propager l'information contextuelle et mieux contraindre la prédiction de relations.

2.2 Interaction vision-langage et raisonnement spatial

Les relations spatiales occupent également une place essentielle dans les systèmes vision-langage, où elles assurent la cohérence entre le contenu visuel et les descriptions textuelles. Dans la génération automatique de légendes, la prise en compte des relations spatiales permet de produire des descriptions plus précises et plus fidèles à la structure de la scène [8]. Les graphes de scènes ont, dans ce contexte, souvent servi de représentation intermédiaire pour garantir une cohérence sémantique entre les objets, leurs attributs et leurs relations. Les tâches de *Visual Question Answering* (*VQA*), illustrent particulièrement la difficulté des modèles actuels à raisonner sur la structure spatiale. Le benchmark *SpatialSense* [9] montre que les systèmes fondés sur des CNN ou des Transformers échouent fréquemment sur des relations pourtant élémentaires, telles que *left of* ou *in front of*, ce qui indique que l'information spatiale n'est pas naturellement apprise par les architectures neuronales standards. Lorsque les relations nécessitent des informations 3D, cette difficulté est encore plus marquée. Le benchmark *Rel3D* [10] a ainsi démontré que les signaux monoculaires sont souvent insuffisants pour raisonner correctement sur la profondeur, ce qui rend indispensable l'utilisation d'indices géométriques supplémentaires. Les travaux antérieurs ayant recours à des descripteurs spatiaux tels que l'histogramme de force [11] ou le Φ-descripteur [12] ont également joué un rôle important dans la génération de

descriptions linguistiques. Ces méthodes établissent un effet un pont entre géométrie mesurable et langage naturel, en traduisant des configurations spatiales en prédicats linguistiques flous ou graduels.

2.3 Analyse structurelle, classification et indexation d'images

La modélisation spatiale ne se limite pas à la reconnaissance de relations et trouve de nombreuses applications dans l'analyse structurelle de formes et la classification d'images. Les descripteurs de position relative (*RPD*), permettent de capturer des signatures géométriques internes aux objets, indépendantes de leur apparence visuelle. Ces représentations ont été utilisées pour la classification de formes complexes [13], la reconnaissance structurelle et la construction de vocabulaires spatiaux.

Dans le domaine de la recherche d'images par le contenu (*CBIR*), les représentations spatiales jouent également un rôle déterminant. Le *Symmetric Force Banner* (*sFB*) [14] permet de comparer des images selon leur organisation spatiale plutôt que leur contenu sémantique, ce qui est essentiel pour retrouver des images partageant des structures similaires mais présentant des variations d'apparence. Les relations spatiales sont également centrales dans l'analyse des interactions humain-objet [15, 16, 17], où elles sont indispensables pour reconnaître des actions telles que *tenir*, *saisir*, ou *monter*. Les progrès récents de la segmentation d'objets ont par ailleurs mis en évidence les limites des boîtes englobantes pour le raisonnement spatial. Les masques de segmentation offrent une granularité plus fine qui se traduit par une meilleure précision dans l'évaluation des relations [18], un point particulièrement important pour les méthodes apprenantes modernes ou pour les approches multimodales, telles que *C-SIP* [14], où les *RPD* servant de contraintes spatiales explicites pour guider l'apprentissage auto-supervisé.

2.4 Robotique et navigation

La compréhension spatiale est également cruciale en robotique [19], où elle intervient dans la navigation, la manipulation d'objets et les interactions homme-robot. Les travaux en linguistique et en cognition visuelle, ont montré que la description spatiale repose sur des référentiels multiples qu'un robot doit être cabale d'interpréter pour exécuter correctement des instructions. Cette articulation entre perception, langage et action impose l'utilisation de modèles capables de représenter explicitement les relations entre objets, et de raisonner.

3 Approches de modélisation des relations spatiales

Les approches de reconnaissance et de modélisation des relations spatiales se sont historiquement structurées autour de deux grands paradigmes :

— Les approches qualitatives, qui décrivent les configurations spatiales au moyen de primitives symboliques
— Et les approches quantitatives, qui mesurent explicitement la configuration géométrique entre objets à l'aide de descripteurs numériques continus.

Ces deux perspectives, constituent le socle des représentations classiques des relations spatiales. Elles ont progressivement été étendues par une troisième famille de méthodes, fondée sur l'apprentissage automatique, qui cherche à extraire directement l'information spatiale à partir de données visuelles ou multimodales.

La figure 1 synthétise cette organisation en distinguant trois grandes familles d'approches : les représentations géométriques explicites, qui relèvent des méthodes quantitatives et mesurent directement la configuration spatiale entre objets, les représentations symboliques, qui relèvent des approches qualitatives du raisonnement spatial, et les représentations apprenantes, qui intègrent l'information spatiale au sein de modèles neuronaux.

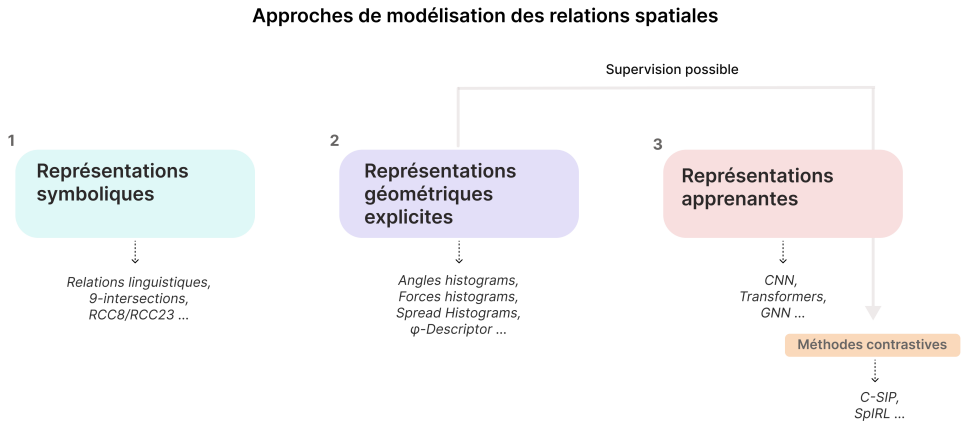

FIGURE 1 – Différentes approches de modélisation des relations spatiales.

Chacune de ces familles reflète une étape dans l'évolution des modèles, depuis les approches purement géométrique, jusqu'aux architectures apprenantes modernes.

3.1 Représentations symboliques

Les approches symboliques constituent l'une des plus anciennes familles de méthodes dédiées à la modélisation des relations spatiales. Elles s'inscrivent dans les approches dites qualitative, où la configuration entre objets est décrite à l'aide de primitives logiques, linguistiques ou topologiques plutôt que par des mesures continues. Ces représentations privilégient les catégories conceptuelles, telles que *left of*, *inside*, *touching* ou *near*, et se rattachent à des travaux fondateurs en linguistique, psychologie cognitive et raisonnement spatial.

Un jalon historique majeur est apporté par Freeman [4], qui propose une première systématisation des relations spatiales élémentaires. Sa formalisation regroupe trois grandes familles : les relations directionnelles, topologiques et de distance. Ces primitives qualitatives, directement inspirées des descriptions linguistiques humaines, servent de base à la plupart des modèles symboliques ultérieurs et peuvent être combinées pour décrire des configurations plus complexes.

Au sein de ces approches qualitatives, le formalisme des intersections introduit par Egenhofer et Franzosa [20] occupe une place centrale. Il encode la relation entre deux objets en examinant les intersections entre leurs régions intérieure, frontière et extérieure, ce qui permet de distinguer des configurations telles que *disjoint*, *overlap* ou

inside. À ce cadre s'ajoute le *Region Connection Calculus* (*RCC*), proposé par Randell *et al.* [21], qui formalise huit relations spatiales constituant une base axiomatique robuste pour le raisonnement spatial déclaratif. Cette axiomatisation a été étendue à des cas plus complexes, notamment pour gérer les concavités et les imbrications d'objets, dans *RCC23* [22].

Certaines méthodes symboliques dérivent également de modèles issus du raisonnement temporel. L'algèbre des intervalles d'Allen [23], initialement conçue pour la temporalité, a été transposée au domaine spatial en projetant les objets sur les axes horizontaux et verticaux. Cette analogie permet de décrire des relations directionnelles qualitatives à partir de comparaisons d'intervalles, mais se révèle moins adaptée pour des formes complexes ou non convexes.

Au-delà de ces fondements classiques, le raisonnement spatial qualitatif s'est élargi pour couvrir des dimensions plus complexes que les seules relations statiques entre objets. Un pan essentiel concerne les relations de mouvement et de trajectoires, notamment le *Qualitative Trajectory Calculus* (*QTC*) de Van de Weghe *et al.* [24], qui décrit le rapprochement, l'éloignement ou les changements directionnels entre objets mobiles. Plusieurs travaux ont introduit des formalismes capables de gérer l'incertitude et les frontières floues, tels que le modèle *egg-yolk* [25] ou les variantes floues du *RCC*. Une synthèse complète de ces développements couvrant topologie, direction, distance, mouvement, forme, incertitude et spatio-temporel, est fournie dans le papier de Chen *et al.* [26]. Dans le cadre du présent rapport, nous nous concentrons toutefois sur les relations spatiales statiques et bidimensionnelles pertinentes pour la vision par ordinateur.

3.2 Représentations géométriques explicites

Les premières approches de reconnaissance et de modélisation des relations spatiales reposent sur des représentations géométriques explicites, souvent appelées *Relative Position Descriptors (RPD)*. L'objectif de ces méthodes est de formaliser la configuration spatiale entre deux objets à partir de caractéristiques purement géométriques (positions, distances, orientations, recouvrements), indépendamment du contenu visuel ou sémantique de l'image. Ces représentations constituent les fondations de la modélisation des relations spatiales en vision par ordinateur, et sont à l'origine de nombreux travaux ultérieurs.

Principe général. Les représentations géométriques explicites assimilent les objets d'une scène à des entités géométriques (points, segments, régions), et décrivent leur arrangement relatif par des mesures quantitatives (ou qualitatives). Ces approches cherchent à définir une fonction de description $D(A, B)$, qui transforme une paire d'objets (A, B) en un vecteur ou histogramme caractérisant leur position relative. Les relations telles que *au-dessus*, *à gauche de*, ou *proche de* sont alors déduites de ces représentations à partir de seuils ou de règles simples.

3.2.1 Angles histogram

L'histogramme d'angles [27] (*Angles histogram)* introduit par Miyajima et Ralescu est l'un des premiers descripteurs utilisés pour représenter les relations direction-

nelles. Pour chaque paire de points (p, q) appartenant respectivement aux régions A et B, on calcule l'angle θ entre la droite reliant ces points et l'axe horizontal. L'ensemble des angles obtenus forme une distribution angulaire discrétisée, où chaque bin représente une direction moyenne (par exemple, *gauche*, *droite*, *haut*, *bas*). Cette approche simple permet de modéliser les relations directionnelles, mais elle reste sensible à la forme et à la taille des objets, et ne capture pas la notion d'intensité relationnelle.

3.2.2 Forces histogram

Pour pallier les limites de l'histogramme d'angles, Matsakis et Wendling (1999) ont introduit l'histogramme de force [11] (*Forces histogram*), un descripteur directionnel qui intègre non seulement l'orientation relative entre deux objets, mais aussi une notion d'intensité pondérée par la distance.

Le principe repose sur l'idée que chaque élément de l'objet A exerce une force sur les éléments de B, dont l'intensité décroît avec la distance et dépend de la direction considérée. Plutôt que de compter simplement les occurrences d'angles comme dans l'histogramme d'angles, l'histogramme de force agrège les contributions directionnelles pondérées issues des interactions entre segments des deux objets.

Ainsi, pour chaque direction θ, l'histogramme de force $F^{AB}(\theta)$ mesure la quantité totale d'« attraction » de A vers B dans cette direction. Lorsque la pondération par la distance est supprimée, on retrouve naturellement l'histogramme d'angles, ce qui fait de l'histogramme de forces une généralisation de l'histogramme d'angles.

Ce descripteur de position relative (*RPD*) présente plusieurs avantages : il capture simultanément la direction dominante et la densité spatiale des interactions, il est robuste à la forme des objets et il fournit une mesure continue de la relation, reflétant mieux l'intuition humaine.

Il constitue la base de nombreux travaux ultérieurs, dont les Φ-Descripteurs ou les histogrammes d'enlacement directionnel.

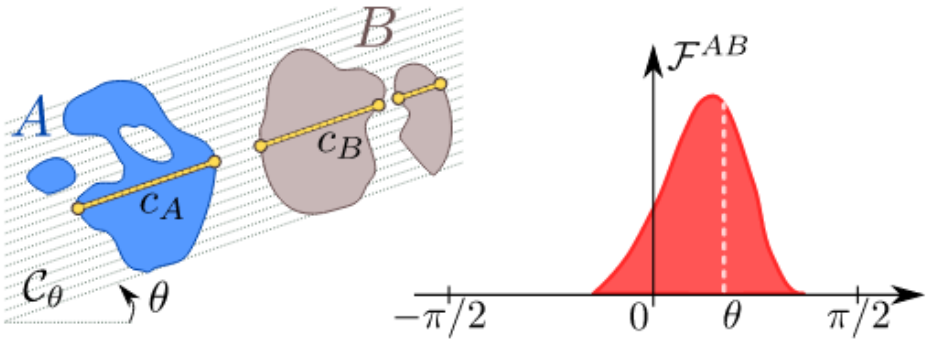

FIGURE 2 – Principe de l'histogramme de forces entre deux objets A et B (figure extraite de Clément *et al.* [28]).

3.2.3 Spread Histogram

Le *Spread Histogram* [29] a été proposé par Kwasnicka et Paradowski (2005) afin de compléter les descripteurs directionnels classiques, tels que l'histogramme d'angles

ou l'histogramme de forces, en introduisant une mesure permettant de distinguer des relations spatiales topologiques telles que *inside*, *outside* ou *encompass*. Contrairement aux méthodes strictement directionnelles, qui comparent uniquement les orientations formées par les paires de points des objets, le Spread Histogram quantifie l'« ouverture angulaire » disponible autour de chaque point de l'objet A par rapport à l'objet B.

Pour chaque pixel $p \in A$, l'ensemble des demi-droites reliant p aux pixels de B partitionne le plan en secteurs. Le descripteur associe alors à p l'angle du plus grand de ces secteurs, noté $\beta(p)$, qui reflète la manière dont l'objet B entoure ou au contraire laisse libre le point p. Intuitivement, cet angle est proche de zéro lorsque p est « entouré » par B (cas *inside* ou *encompassed*), tandis qu'il devient grand lorsque B occupe une zone restreinte autour de p, ce qui correspond à une configuration typique *outside*.

L'histogramme des valeurs β, constitue le *Spread Histogram*. Sa distribution permet alors de séparer différentes catégories spatiales que les histogrammes directionnels ne peuvent pas distinguer : les configurations *inside*, *outside*, *encompassed*, ainsi que les cas intermédiaires où un objet entoure partiellement l'autre. Le descripteur se combine naturellement à l'histogramme d'angles, le premier fournissant l'information topologique (ouverture angulaire), le second l'information directionnelle. Ensemble, ils permettent de reconnaître simultanément les relations directionnelles (*left*, *right*, *above*, *below*) et les relations d'inclusion ou d'entourance.

Le *Spread Histogram* présente également l'avantage d'être invariant par rotation et changement d'échelle, car l'angle maximal associé à chaque point ne dépend que de la géométrie relative des deux objets. Cette propriété en fait un outil robuste pour la reconnaissance de relations spatiales dans des scènes où les objets peuvent subir des transformations géométriques simples. Enfin, sa complexité de calcul reste comparable à celle de l'histogramme d'angles.

3.2.4 Φ-Descriptor

Introduit par Matsakis, Naeem et Rahbarnia, le Φ-Descripteur [12] constitue aujourd'hui l'un des descripteurs de position relative les plus complets. Il est conçu pour dépasser les limites des descripteurs histogrammes existants, dont chacun ne satisfait qu'une partie des propriétés fondamentales attendues d'un modèle spatial. À l'inverse, le Φ-Descripteur est le premier à satisfaire simultanément les treize propriétés identifiées dans [12], couvrant les relations directionnelles, topologiques, de distance, ainsi que la robustesse aux transformations affines.

Principe général. Le Φ-Descripteur repose sur le cadre des histogrammes de forces (F-Histogrammes), qui consistent à réduire l'analyse spatiale bidimensionnelle à un ensemble de coupes unidimensionnelles. Pour chaque direction θ, une droite L_θ intersecte les deux objets A et B, produisant des segments ($A \cap L_\theta$, $B \cap L_\theta$). Les points d'entrée et de sortie de ces segments sont classés selon douze catégories. Les couples consécutifs de points sur L_θ relèvent alors de *36 catégories*, regroupées en *10 familles sémantiques*, telles que *trails*, *overlaps*, *covers*, *uncovers*, *follows*, *leads* ou *starts*.

À chaque famille est associée une fonction $f_i(\theta, p, q)$, définie sur une paire de points successifs (p, q) et proportionnelle à la distance $|pq|$. Ces fonctions quantifient

la contribution locale d'une configuration de frontières à une relation spatiale dans la direction θ.

Les fonctions locales sont ensuite agrégées direction par direction pour produire *treize F-histogrammes* :

$$F_t^{AB},\ F_o^{AB},\ F_c^{AB},\ F_u^{AB},\ F_f^{AB},\ F_\lambda^{AB},\ F_s^{AB},\ F_v^{AB},\ F_a^{AB},\ F_r^{AB},\ F_e^{AB},\ F_d^{AB},\ F_w^{AB},$$

chacun représentant l'aire d'une région d'interaction particulière entre A et B pour la direction θ. Par exemple, F_t^{AB} mesure l'aire où un segment de A *traîne* (*trails*) derrière un segment de B.

Définition du descripteur. Le Φ-Descripteur associé au couple (A, B) est défini comme le vecteur :

$$\Phi_{AB} = \begin{pmatrix} F_t^{AB},\ F_o^{AB},\ F_c^{AB},\ F_u^{AB},\ F_f^{AB},\ F_\lambda^{AB},\ F_s^{AB}, \\ F_v^{AB},\ F_a^{AB},\ F_r^{AB},\ F_e^{AB},\ F_d^{AB},\ F_w^{AB}, \\ \text{measure}(A),\ \text{measure}(B) \end{pmatrix}$$

Il regroupe ainsi l'ensemble des treize F-histogrammes ainsi que les mesures des objets, constituant une représentation quantitative riche, continue et affine-invariante de la position de A par rapport à B.

Le Φ-Descripteur se distingue en fournissant une description unifiée et exhaustive des relations spatiales, à la fois directionnelles, topologiques et métriques, tout en restant calculable efficacement sur des objets raster ou vectoriels.

3.2.5 Directional Enlacement Histograms

Les *Directional Enlacement Histograms* [30] (*DEH*), introduits par Clément, Poulenard, Kurtz et Wendling (2017), ont été développés pour décrire des configurations spatiales plus complexes que celles capturées par les descripteurs directionnels classiques tels que les histogrammes d'angles ou les histogrammes de forces. Ces méthodes échouent en effet dès que les objets deviennent concaves, emboîtés, ou composés de plusieurs composantes, conduisant à des situations ambiguës d'encerclement, imbrication ou entrelacement (Fig. 1 du papier).

Principe général. L'idée centrale est que les relations telles que l'enlacement ne peuvent pas être décrites uniquement par des directions dominantes, elles nécessitent de quantifier dans quelle mesure un objet est « entre » deux parties de l'autre objet, selon chaque direction. Pour cela, les auteurs introduisent une définition rigoureuse de l'*enlacement unidimensionnel* basée sur les coupes longitudinales des objets. Pour une direction θ, toutes les droites parallèles $\Delta(\theta, \rho)$ découpent les objets en segments, l'enlacement local mesure alors le nombre (ou la quantité) de triplets ordonnés (b_i, a_k, b_j) où un point de A se situe entre deux points de B sur la même coupe (Sec. 4.1).

Enlacement directionnel. L'*enlacement de A par B dans la direction* θ, noté $E_{AB}(\theta)$, est obtenu en intégrant ces contributions élémentaires sur l'ensemble des coupes parallèles [30, Eq. (7)]. Le résultat est une fonction périodique, invariant par translation et mise à l'échelle, et quasi-invariante en rotation (Sec. 4.2).

Histogrammes d'enlacement et d'entrelacement. La fonction résultante, une fois normalisée par les aires des objets, définit l'*ε-histogramme*, qui décrit la façon dont A est enlacé par B selon chaque direction [30, Eq. (13)]. Comme l'enlacement n'est pas symétrique, les deux histogrammes E_{AB} et E_{BA} doivent être considérés conjointement. Les auteurs définissent alors un *I-histogramme* basé sur la moyenne harmonique des deux profils pour quantifier l'*interlacement*, c'est-à-dire l'enlacement mutuel [30, Eq. (15)].

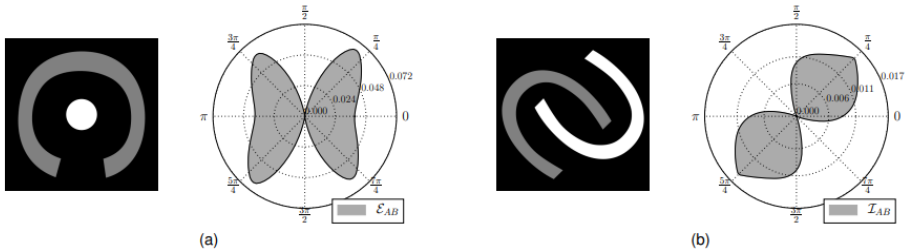

FIGURE 3 – Illustration des histogrammes d'enlacement (extrait de Clément *et al.*, 2017 [30]).

Intérêt du descripteur. Les *Directional Enlacement Histograms* permettent ainsi de capturer des configurations impossibles à distinguer avec les modèles directionnels classiques : *surrounding*, imbriqué, mutuellement enlacé, ou encore des structures fortement entremêlées. Ils fournissent une représentation continue, directionnelle, géométriquement robuste, adaptée aux objets concaves ou multi-composantes, et ont démontré leur efficacité dans plusieurs domaines applicatifs (médical, documents anciens, télédétection) [30].

Extensions floues de l'enlacement directionnel. Le modèle d'enlacement directionnel défini dans [30] a par la suite été étendu vers une formulation floue, notamment avec l'introduction des *Fuzzy Directional Enlacement Landscapes* (Fuzz-DELs) [31]. Alors que l'histogramme d'enlacement fournit une mesure globale de l'entrelacement entre deux objets pour chaque direction, les Fuzz-DELs proposent une évaluation *locale* de cette relation dans l'espace image, en attribuant à chaque point un degré flou d'enlacement pour une direction donnée. Cette extension ne redéfinit pas le descripteur initial, mais en propose une version fuzzifiée et spatialisée, mieux adaptée à l'analyse fine de la structure directionnelle d'objets complexes (concavités, régions enclavées, configurations multi-composants).

3.2.6 Force Banners

Le *Force Banner* [32] a été introduit par Deléarde, Kurtz, Dejean et Wendling comme une extension directe de l'histogramme de forces (*F-histogram*) proposé initialement par Matsakis et Wendling [11]. Alors que l'histogramme de forces classique repose sur une unique fonction de force $\varphi_r(d) = 1/d^r$, le Force Banner généralise ce principe en évaluant, pour chaque direction et niveau, un panel de forces allant des forces attractives aux forces répulsives.

Principe. Pour une direction θ et pour chaque paramètre $r \in [r_s, r_e]$, la force entre deux objets A et B est mesurée via l'histogramme directionnel $F_r^{AB}(\theta)$. Le Force Banner est

défini comme la fonction bidimensionnelle :

$$FB_{AB} : [0, 2\pi) \times [r_s, r_e] \to \mathbb{R}^+, \qquad (\theta, r) \mapsto F_r^{AB}(\theta). \tag{1}$$

Chaque ligne de la bannière correspond à un type de force (défini par r), et chaque colonne à une direction angulaire. L'ensemble forme une représentation matricielle riche qui encode simultanément direction et distance.

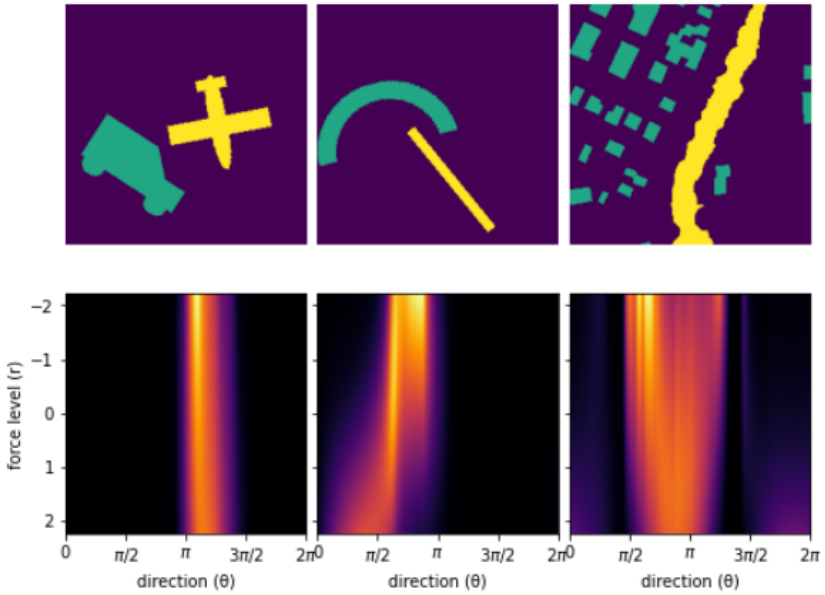

FIGURE 4 – Illustration du principe des Force Banners (extrait de Deléarde *et al.*, 2021 [32]).

Propriétés. Les auteurs démontrent que le Force Banner est invariant par translation et changement d'échelle (après normalisation) et est isotrope.

Intérêt. Le Force Banner constitue un descripteur bidimensionnel exploitable comme représentation intermédiaire pour l'apprentissage supervisé. Les auteurs montrent notamment qu'un réseau convolutionnel 2D (*SqueezeNet*) entraîné sur ces bannières reconnaît efficacement les relations directionnelles (*north, south, east, west*) et surpasse un CNN appliqué directement sur les images. La structure matricielle du descripteur permet de capturer des configurations spatiales complexes, y compris pour des objets non convexes ou disjoints.

Autres descripteurs géométriques. Outre les méthodes détaillées ci-dessus, plusieurs autres descripteurs de position relative ont été proposés dans la littérature, parmi lesquels les *R-histograms* [33], les *R-histograms* [34], ou encore les *Visual Area Histograms* [35]. Des modèles radiaux tels que le *Radial Line Model (RLM)* [36] et sa version étendue [37] ont également été proposés. Chacune de ces méthodes apporte une manière spécifique de décrire la position relative entre objets, mais aucune ne couvre l'ensemble des propriétés souhaitées pour un descripteur général, comme le souligne l'analyse de Matsakis *et al*. [12].

Dans cet état de l'art, nous choisissons de ne pas détailler ces méthodes, qui sont soit moins utilisées dans les travaux récents, soit davantage spécialisées. Nous concentrons notre analyse sur les descripteurs qui demeurent les plus influents aujourd'hui, car ils constituent le socle des approches modernes de modélisation des positions relatives.

3.3 Learning-based methods

3.3.1 Baselines simples : MLP sur boîtes englobantes

Avant l'essor des architectures à base de Transformers pour la reconnaissance de relations spatiales, une grande partie des travaux reposait sur une famille de modèles étonnamment simples mais particulièrement difficiles à surpasser : les *baselines géométriques* fondés uniquement sur les boîtes englobantes. Dans ces approches, chaque paire d'objets (A, B) est représentée par leurs coordonnées normalisées (x, y, w, h), éventuellement complétées par des caractéristiques dérivées telles que la différence des centres, les rapports de tailles ou l'aire relative, puis un perceptron multi-couches (*MLP*) est entraîné pour classifier la relation spatiale entre les deux objets.

Ces modèles sont utilisés comme *naive baselines* dans plusieurs benchmarks récents dédiés aux relations spatiales. Sur *SpatialSense* [9] et *Rel3D* [10], il a été montré que de nombreux modèles neuronaux fondés sur le contenu visuel ne parviennent pas à dépasser ces baselines purement géométriques. L'étude [38] de Wen *et al*. souligne ainsi qu'avant l'introduction de modèles exploitant l'attention globale, « aucune approche existante n'est capable de surpasser une approche de base naïve ne reposant uniquement sur les coordonnées des boîtes englobantes » [38].

Des observations similaires sont rapportées dans le benchmark Rel3D. Goyal *et al*. comparent plusieurs architectures de détection de relations spatiales (*DRNet, VTransE, VipCNN, PPR-FCN*) à différentes baselines géométriques [10]. Leur évaluation montre qu'un modèle « *BBox only* », fondé uniquement sur les coordonnées des boîtes englobantes, atteint une précision comparable, voire supérieure, à celle de ces architectures, basées sur des CNN, plus complexes. Ce résultat met en évidence la quantité importante de signal géométrique déjà encodée dans la configuration 2D des boîtes, notamment lorsque les relations sont principalement directionnelle, ou lorsque les scènes sont synthétiques.

Malgré leur simplicité, les baselines MLP sur boîtes englobantes demeurent des références difficiles à battre encore aujourd'hui. Ils capturent efficacement la géométrie relative des objets et fournissent un seuil minimal de performance que toute méthode apprenante moderne doit dépasser pour justifier l'exploitation d'informations spatiales, visuelles, ou sémantiques supplémentaires.

3.3.2 Méthodes apprenantes : CNN, Transformers, GNN et intégration de représentations spatiales

Au-delà des baselines géométriques fondées sur les boîtes englobantes, une grande partie de la littérature contemporaine s'appuie sur des modèles apprenants capables d'exploiter le contenu visuel, le contexte de la scène ou des représentations spatiales dérivées. Trois grandes familles d'approches émergent : les réseaux convolutionnels

(CNN), les architectures à base de Transformers, et les réseaux de neurones de graphes (GNN). Ces architectures diffèrent profondément dans la manière dont elles modélisent les interactions entre objets, leur capacité à intégrer le contexte global et les modalités spatiales disponibles.

Réseaux convolutionnels (CNN). Les premiers modèles de *Visual Relationship Detection* (VRD) et de *Scene Graph Generation* (SGG) reposent sur des CNN appliqués à des régions d'intérêt. Lu *et al*. [39] proposent un modèle combinant l'apparence visuelle du sujet, de l'objet et de leur région d'union avec des embeddings linguistiques pour prédire un prédicat. De leur côté, Peyre *et al*. [40] associent des caractéristiques d'apparence issus d'un CNN à des modèles spatiaux décrivant la configuration des boîtes englobantes.

Dans ces approches, la convolution opère sur des cartes locales. L'intégration du contexte dépend de mécanismes indirects tels que l'union des boîtes englobantes, des opérations de *pooling* ou des représentations globales. Ce biais structurel vers l'information locale limite la capacité des CNN à capturer des interactions spatiales à longue portée ou des configurations complexes, particulièrement lorsque les objets sont éloignés ou de formes non convexes.

Servant *et al.* [41] rappellent également que plusieurs extensions des architectures convolutionnelles cherchent à intégrer explicitement l'information spatiale. Une première ligne de travaux consiste à enrichir les modèles par des paramétrisations géométriques dérivées des boîtes englobantes, telles que l'IoU, les rapports d'aspect ou les positions relatives, utilisées notamment dans VTransE [42]. Une seconde approche repose sur la construction d'images binaires représentant les boîtes englobantes ou les masques d'objets, comme dans HO-RCNN [43] ou DRNet [44], où un CNN est entraîné non pas sur les pixels réels de l'image mais sur une carte spatiale synthétique destinée à isoler la structure géométrique. Bien que ces stratégies permettent de dissocier, au moins partiellement, les indices visuels et spatiaux, leur portée reste limitée. Les représentations binaires obtenues ne sont pas centrées sur la structure interne des objets et ne constituent pas de véritables caractéristiques visuelles, elles se réduisent à de simples indices spatiaux utilisés comme signal auxiliaire.

Intégration de représentations spatiales explicites dans les modèles apprenants. Une stratégie récente consiste à intégrer des *Relative Position Descriptors* (RPD), historiquement conçus pour la modélisation géométrique explicite comme expliqué plus haut, au sein de réseaux apprenants. Ces descripteurs (par exemple les histogrammes de forces, bandeaux de forces, Φ-Descripteur, modèles radiaux) capturent des informations directionnelles, topologiques ou métriques difficilement apprises par les modèles neuronaux supervisés classiques.

CNN appliqués aux RPD. Deléarde *et al* [32, 45] montrent que le *Force Banner*, une représentation bidimensionnelle structurée selon les directions et les niveaux de force, peut être traité comme une image pour un CNN, permettant au réseau d'exploiter directement la géométrie fine capturée par ce descripteur. Cette approche dépasse les CNN appliqués à des crops visuels pour la reconnaissance de relations directionnelles.

Apprentissage contrastif guidé par la géométrie. Plus récemment, l'approche C-SIP (*Contrastive Spatial-Image Pretraining*) proposée par Servant *et al*. [14] exploite explicitement un RPD (le *Symmetric Force Banner*) comme modalité spatiale. Un encodeur d'images est entraîné, de manière contrastive, à aligner ses représentations visuelles avec les signatures spatiales extraites des RPD. Ce paradigme contourne les ambiguïtés supervisées des benchmarks (polysémie, faibles liens entre sémantique et géométrie) en imposant directement une structure géométrique au modèle. Les auteurs montrent que l'encodeur visuel obtenu possède une sensibilité spatiale plus fine que les modèles supervisés traditionnels.

Transformers : modélisation globale et interactions à longue portée. Les Transformers, grâce à leur mécanisme d'auto-attention, modélisent explicitement les interactions entre toutes les régions de l'image. Contrairement aux CNN, ils ne sont pas limités par un champ réceptif local.

Dans RelTR, Cong *et al*. [46] utilisent un encodeur-décodeur Transformer pour générer directement des triplets relationnels, exploitant l'attention globale pour capturer le contexte. Wen *et al*. vont plus loin avec *RelatiViT* [38]. Selon les auteurs, le modèle RelatiViT serait la première architecture à surpasser les baselines géométriques sur *SpatialSense+*, leur dataset, ainsi que *Rel3D*. En injectant simultanément les patches du sujet, de l'objet et du contexte masqué dans le même ViT, le modèle laisse au mécanisme d'attention la tâche de modéliser les interactions spatiales correctes.

Bien que les Transformers aient profondément transformée la vision par ordinateur, plusieurs travaux montrent que leur capacité à modéliser la structure spatiale d'une scène, notamment en 3D et dans les configurations complexes, n'est pas intrinsèque [47, 38]. Dans les Vision Transformers (ViT), l'image est découpée en patchs, dont les relations spatiales doivent être entièrement apprises à partir des données via le mécanisme d'auto-attention, sans contrainte architecturale spatiale imposée par le réseau, contrairement aux CNN. Des analyses plus poussées [48, 38, 49] confirment que les Transformers, en particulier les grands modèles Vision-Langage (VLM), ont tendance à apprendre en priorité des corrélations globales (co-occurences d'objets, indices sémantiques ou linguistiques) plutôt que des contraintes géométriques fines. Ce phénomène, souvent appelé *shortcut learning*, explique pourquoi ces modèles sont poussés à privilégier les solutions basées sur des corrélations simples (comme les coordonnées des boîtes englobantes), ce qui les rend étonnamment décevants pour les tâches liées au raisonnement spatial précis. Ils ont du mal à surpasser les baselines géométriques simples sur des benchmarks ancrés physiquement.

Réseaux de neurones de graphes (GNN). Les GNN constituent une architecture particulièrement adaptée pour représenter la structure relationnelle d'une scène : les objets constituent les nœuds, les relations les arêtes. Xu *et al*. [50] introduisent un modèle de *Message Passing* pour raffiner conjointement les labels d'objets et de relations. Yang *et al*. [51] proposent *Graph R-CNN*, où un *Attentional GCN* propage le contexte entre objets et prédicats.

Les GNN permettent de modéliser la structure de scène de manière explicite, mais sont souvent limités par l'usage des boîtes englobantes comme primitives spatiales. Pour atténuer cette limite, Khandelwal *et al*. proposent des approches *segmentation-*

grounded [52] qui remplacent les boîtes par des masques *pixel-level*, offrant une granularité spatiale plus riche. Ces modèles sont utilisés non seulement en SGG mais aussi pour l'interaction humain-objet ou certaines formes de raisonnement visuel.

Apport de la profondeur estimée et fusion multimodale. Une limitation fondamentale des approches basées uniquement sur les pixels RGB réside dans la perte d'information inhérente à la projection d'une scène 3D sur un plan 2D. Cette projection introduit une ambiguïté d'échelle-distance : un petit objet proche peut occuper la même surface en pixels qu'un grand objet lointain. Goyal *et al.* [10] démontrent dans le benchmark *Rel3D* que les modèles standards échouent fréquemment à distinguer des relations dépendantes de la profondeur, telles que *behind* ou *outside*, lorsqu'elles sont confondues avec des superpositions 2D triviales. L'intégration explicite d'une carte de profondeur (*depth map*) devient alors indispensable pour lever ces ambiguïtés et restaurer la structure topologique de la scène.

Historiquement, l'estimation de profondeur monoculaire (*Monocular Depth Estimation*) souffrait d'un manque de généralisation. Toutefois, l'état de l'art a récemment franchi un cap avec l'avènement des modèles fondations entraînés sur des données non étiquetées à très grande échelle. Après les travaux pionniers de *MiDaS* [53], le modèle *Depth Anything* [54] a établi un nouveau standard en traitant la profondeur comme un problème de modélisation du langage visuel. Contrairement aux méthodes antérieures limitées à des scènes d'intérieur, ces modèles fournissent des cartes de profondeur denses et robustes sur des images naturelles (*in-the-wild*). Dans le cadre de la modélisation des relations spatiales, cette carte de profondeur n'est pas une simple image supplémentaire, elle agit comme un *embedding* géométrique dense, permettant de réintroduire une métrique de distance relative continue là où les boîtes englobantes ne fournissaient que des coordonnées.

L'exploitation efficace de cette information nécessite des architectures de fusion multimodales avancées. Une fusion naïve par concaténation des canaux RGB et de profondeur est insuffisante, car elle mélange des modalités hétérogènes sans modéliser explicitement la relation entre apparence visuelle et structure géométrique. Les architectures récentes s'inspirent des mécanismes d'attention croisée (*Cross-Attention*) popularisés par les Transformers multimodaux [55, 56]. Dans ces schémas, une modalité peut jouer le rôle de *Queries* (souvent l'image), tandis que la profondeur ou le texte fournissent des *Keys* et *Values*. Cela permet au modèle d'aligner dynamiquement ses poids d'attention et le réseau peut apprendre à focaliser son attention sur la discontinuité de profondeur entre deux objets pour prédire une occlusion, tout en utilisant le contexte textuel pour désambiguïser la nature de la relation. Cette synergie entre vision, géométrie estimée et langage constitue une voie pour dépasser les limitations des modèles purement 2D.

Distinction entre approches symboliques et géométriques. Les approches symboliques se distinguent fondamentalement des représentations géométriques explicites, telles que les histogrammes d'angles, les histogrammes de forces, les bandeaux de force, les histogrammes d'enlacement ou le Φ-Descripteur, qui cherchent à *quantifier* la configuration spatiale à l'aide de mesures continues (distances, orientations, recou-

vrements, moments, etc.). Ces descripteurs géométriques, largement utilisés dans les modèles quantitatifs, fournissent des représentations numériques fines directement exploitables en classification ou en comparaison.

Les approches symboliques, au contraire, ne mesurent pas les relations : elles permettent de *raisonner* sur celles-ci en manipulant des règles logiques, des axiomes ou des structures qualitatives. Elles constituent un cadre adapté pour l'interprétation, la vérification de cohérence ou la déduction de relations, mais peinent à capturer la diversité morphologique et l'imprécision des scènes naturelles.

Dans l'ensemble, ces deux familles apparaissent comme profondément complémentaires : les approches symboliques fournissent une structure conceptuelle et inférentielle de haut niveau, tandis que les représentations géométriques capturent la richesse métrique et la variabilité des formes observées dans les images réelles. Les méthodes récentes combinent désormais raisonnement qualitatif, mesures géométriques et apprentissage automatique.

4 Problématiques

La reconnaissance et la modélisation des relations spatiales reposent sur la traduction de concepts spatiaux humains, naturels, ambigus et dépendants du point de vue, en représentations formelles. Cette relation entre la continuité des descriptions spatiales naturelles et la discrétisation imposée par les modèles explique pourquoi, malgré des années de travaux, les prédicats spatiaux restent difficiles à apprendre de manière fiable à partir d'images naturelles. Les difficultés naturelles rencontrées ne découlent pas uniquement de limites méthodologiques, mais de problèmes fondamentaux du domaine : ambiguïtés perceptives, incohérences sémantiques, biais des annotations humaines et appauvrissement géométrique causé par les primitives spatiales utilisées dans les datasets.

4.1 Ambiguïtés liées au point de vue et projection 2D d'un monde 3D

L'un des défis les plus importants est l'écart entre la nature tridimensionnelle des relations spatiales et la représentation bidimensionnelle de l'image. De nombreuses relations directionnelles changent radicalement d'interprétation selon le point de vue de la caméra ou selon le référentiel utilisé. Dans les images naturelles, une configuration peut paraître ambigüe, un objet peut être simultanément à gauche et légèrement en dessous d'un autre en 2D, alors qu'en 3D, la relation pertinente pourrait être simplement *devant*. Les jeux de données eux-mêmes reflètent ces ambiguïtés. *SpatialSense*, par exemple, annote certaines relations du point de vue de l'objet (*object-centric*), alors que l'image suggère un autre référentiel (*viewer-centric*). Cette discordance est aggravée par le fait que les annotations reposent principalement sur des boîtes englobantes 2D, qui ne permettent pas d'inférer des relations intrinsèquement tridimensionnelles.

4.2 Ambiguïtés sémantiques et variabilité des annotations humaines

La littérature en linguistique spatiale montre que les relations spatiales sont des catégories floues, dont les frontières dépendent fortement du contexte et des attentes des annotateurs. Cette subjectivité se retrouve dans des jeux de données modernes, par exemple les jeux de données comme *Visual Genome* ou *SpatialSense* présentent des désaccords entre annotateurs, même pour des relations directionnelles simples. Plusieurs facteurs expliquent cette variabilité :

— Les humains n'appliquent pas un référentiel unique,
— Certaines relations sont polysémiques,
— Certains prédicats sont si génériques qu'ils recouvrent des situations topologiques incompatibles (une tasse posée sur une table versus un tableau fixé sur un mur).

Ces ambiguïtés génèrent des incohérences structurelles dans les jeux de données, rendant difficile l'apprentissage de relations mutuellement exclusives dans les modèles apprenants.

4.3 Biais humains et asymétries cognitives

Les annotations sur Internet ne reflètent pas uniquement des erreurs individuelles, elles incorporent des biais cognitifs systématiques. Par exemple, la littérature montre que les annotateurs privilégient certaines constructions asymétriques. Dans *Visual Genome*, les humains apparaissent comme sujet dans la grande majorité des relations impliquant des personnes, ce qui biaise fortement la distribution des prédicats. De plus, certaines relations inverses ne sont presque jamais annotées, même lorsqu'elles existent physiquement dans la scène. Ce déséquilibre structurel induit des modèles qui apprennent des relations orientées par leur fréquence dans les données plutôt que par la géométrie réelle.

4.4 Dépendance aux primitives spatiales simplifiées

Une des difficultés fondamentales du domaine réside dans la manière dont les objets sont représentés. La plupart des jeux de données assimilent les objets à des primitives géométriques très grossières comme les boîtes englobantes rectangulaires, ou bien des centroïdes. Cette représentation appauvrit considérablement l'information spatiale. La forme réelle de l'objet, sa concavité, ou bien ses parties déconnectées disparaissent totalement et de nombreuses relations qualitatives deviennent impossibles à distinguer lorsque la représentation est réduite à deux rectangles. Il en résulte un espace descriptif trop pauvre pour capturer toute la diversité des configurations spatiales présentes dans des images naturelles.

4.5 Confusion entre indices spatiaux et sémantiques

Il a été démontré à plusieurs reprises que les modèles apprenants exploitent massivement les co-occurences sémantiques et non les configurations spatiales elles-mêmes (phénomène de *shortcut learning*). L'existence d'un cheval et d'une personne suffit souvent à prédire *person riding horse*, même si la spatialité contredit cette relation. De

même, des prédicats comme *on* ou *under* peuvent être prédits uniquement à partir du nom des objets, sans utiliser l'image, ce qui démontre à quel point les modèles apprenants, s'appuient davantage sur la sémantique que sur la structure spatiale. Ces biais proviennent en grande partie de la construction du dataset, dans lequel certaines relations sont fortement corrélées à certaines paires d'objets, mais également à la nature de certains modèles neuronaux (comme les Transformers), qui n'intègrent pas forcément de module de raisonnement spatial. Tout cela limite la capacité des modèles à apprendre des relations réellement géométriques et non pas juste statistiques.

5 Comparaisons des méthodes

5.1 Représentations symboliques (qualitatives, logiques)

Forces. Les approches symboliques offrent une interprétabilité maximale, les relations étant définies explicitement (par exemple : RCC, modèle des 9-intersections, logiques de description). Elles constituent un cadre formel robuste, fondé sur des règles logiques bien définies permettant un raisonnement explicite et vérifiable. Elles ne nécessitent aucune donnée d'entraînement et demeurent indépendantes du contenu visuel, seule la structure spatiale qualitative est prise en compte.

Limites. Ces approches reposent sur des relations strictes (*inside*, *meet*, *overlap*, etc.) manquant de flexibilité pour modéliser les variations continues. Elles sont peu sensibles à la géométrie fine des objets (concavités, angles, forme détaillée), puisqu'elles opèrent sur des régions abstraites. Elles sont rigides : une relation est « vraie » ou « fausse », ce qui les rend peu adaptées aux scènes naturelles où les formes sont irrégulières. Elles ne capturent pas la variabilité du réel, en l'absence de continuité, de mesure ou de gradients.

5.2 Représentations géométriques explicites (quantitatives)

Forces. Ces approches permettent une modélisation quantitative et continue des relations spatiales. Elles capturent la forme, la répartition, la direction et parfois la concavité (comme dans les *DEH*). Elles sont souvent conçues pour être invariantes à des transformations usuelles (rotation, translation, échelle), comme c'est le cas pour les histogrammes d'angles, les histogrammes de forces, les histogrammes d'enlacement, le Φ-Descripteur ou encore les bandeaux de force. Elles ne nécessitent aucune donnée annotée et constituent des méthodes déterministes. Elles fournissent des représentations directement exploitables par des classifieurs classiques (SVM, MLP, etc.).

Limites. L'expressivité de ces descripteurs dépend des paramètres choisis. Ils restent essentiellement dyadiques, ce qui rend difficile la modélisation de relations complexes impliquant plusieurs objets ou un contexte global. Ils ignorent les informations visuelles (texture, couleur, sémantique) et se concentrent uniquement sur la géométrie. Ils nécessitent des primitives fiables (contours, masques), ce qui peut limiter leur applicabilité dans des scènes complexes. Enfin, les méthodes géométriques explicites peuvent également devenir coûteuses en calcul lorsqu'elles sont appliquées à grande échelle.

5.3 Méthodes apprenantes (CNN, Transformers, GNN, multimodalité)

Forces. Les méthodes apprenantes peuvent extraire automatiquement des caractéristiques visuelles pertinentes (forme, texture, contexte, indications 3D implicites). Elles permettent d'apprendre des représentations latentes complexes, difficiles à formaliser manuellement. Elles peuvent combiner vision, géométrie, profondeur ou texte dans des architectures multimodales. Elles sont adaptables : un même modèle peut apprendre différentes relations selon les données utilisées. Elles capturent des interactions globales via l'attention ou structurées via le *message passing* dans les GNN.

Limites. Elles nécessitent des quantités importantes de données annotées, ce qui constitue une contrainte structurelle. Elles sont sensibles à la qualité des annotations, à la définition du vocabulaire relationnel (taxonomie) et au choix du référentiel (2D/3D, *viewer-centric* ou *object-centric*). Elles fonctionnent souvent comme des boîtes noires, avec une interprétabilité limitée. Elles souffrent de *shortcut learning*, notamment via l'utilisation excessive des boîtes englobantes, la dépendance aux co-occurrences sémantiques ou la difficulté à apprendre la géométrie fine. Enfin, elles manquent d'invariances naturelles (rotation, échelle), qui doivent être acquises par augmentation des données ou par architecture dédiée.

6 Conclusion

La reconnaissance des relations spatiales occupe une place importante au sein de la vision par ordinateur. A la différence des tâches plus classiques comme la détection ou la segmentation d'objets, elle nécessite de raisonner non pas sur des entités isolées, mais sur la structure d'une scène, ses interactions, sa géométrie, son organisation et parfois son interprétation sémantique. Cette complexité découle directement de la manière dont les humains perçoivent et conceptualisent l'espace. Les relations telles que *left of*, *inside*, *behind* ou *next to* ne sont pas de simples opérations géométriques, elles sont ancrées dans des capacités cognitives, mêlant perception, langage, catégorisation et conventions culturelles. L'espace n'est pas seulement mesuré, il est interprété.

Les travaux étudiés au sein de cet état de l'art montrent que cette dualité entre perception métrique continue et catégorisation qualitative discrète, s'est reflétée dans les différentes familles d'approches proposées pour modéliser les relations spatiales. Les modèles symboliques hérités du raisonnement qualitatif offrent un cadre conceptuel robuste, explorant les primitives relationnelles qui organisent l'espace à un niveau abstrait. Les descripteurs géométriques explicites, quant à eux, permettent de capturer des signatures fines, invariantes et quantitatives, rendant possible une analyse métrique précise des configurations spatiales. Les méthodes apprenantes, enfin, capturent des représentations latentes complexes susceptibles de s'aligner sur les structures géométriques comme sur les régularités sémantiques.

Pourtant, aucun paradigme ne résout pleinement le problème. Les approches symboliques manquent de granularité, les descripteurs géométriques peinent à intégrer le contexte global et les modèles apprenants souffrent de biais structurels, de dé-

pendance aux données et d'une difficulté à distinguer les indices spatiaux des co-occurences sémantiques. Cette tension entre ce que les modèles voient et ce qu'ils comprennent demeure un enjeu central. Elle rappelle que la reconnaissance des relations spatiales, loin d'être un simple problème de classification, constitue une véritable question de représentation. Comment encoder l'espace, la forme, les interactions et les règles implicites qui les organisent ?

Enfin, les défis liés aux annotations humaines, aux ambiguïtés inhérentes au passage 3D/2D, aux référentiels multiples ou aux variations de point de vue soulignent que la scène visuelle n'est jamais donnée telle quelle. Elle est un construit, interprété différemment selon les observateurs, les langues et les tâches. Toute tentative de formalisation des relations spatiales doit donc s'inscrire dans un dialogue entre géométrie, cognition et apprentissage.

Les progrès récents montrent qu'une fusion entre représentations explicites et approches apprenantes constitue une voie particulièrement prometteuse pour dépasser les limites actuelles. Comprendre et modéliser l'organisation spatiale des scènes demeure un défi fondamental, non seulement pour la vision par ordinateur, mais pour toute intelligence artificielle cherchant à raisonner sur le monde physique de manière cohérente, interprétable et robuste.

L'ensemble des travaux étudiés met en évidence une tendance récurrente. Les modèles purement neuronaux apprennent surtout des corrélations statistiques, tandis que les représentations géométriques explicites capturent des propriétés spatiales essentielles. Ces constats convergent vers l'idée qu'un système fiable de reconnaissance de relations spatiales doit combiner apprentissage statistique et modélisation géométrique explicite, plutôt que de s'appuyer exclusivement sur l'un ou l'autre paradigme.

Références

[1] B. Kuipers, "Modeling spatial knowledge," *Cognitive Science*, vol. 2, no. 2, pp. 129–153, 1978.

[2] B. Landau and R. Jackendoff, "Whence and whither in spatial language and spatial cognition ?," *Behavioral and Brain Sciences*, vol. 16, no. 2, p. 255–265, 1993.

[3] L. Vieu, *Semantique des relations spatiales et inferences spatio-temporelles : une contribution a l'etude des structures formelles de l'espace en langage naturel*. PhD thesis, 1991. Thèse de doctorat dirigée par Borillo, Mario Sciences appliquées Toulouse 3 1991.

[4] J. Freeman, "The modelling of spatial relations," *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 156–171, 1975.

[5] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[6] P. xu, X. Chang, L. Guo, P.-Y. Huang, and X. Chen, "A survey of scene graph : Generation and application," 04 2020.

[7] 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," *International Journal of Computer Vision*, vol. 123, pp. 32 – 73, 2016.

[8] D. Runyan, W. Zhang, G. Zhi, and S. Xian, "A survey on learning objects' relationship for image captioning," *Computational Intelligence and Neuroscience*, vol. 2023, 05 2023.

[9] K. Yang, O. Russakovsky, and J. Deng, " SpatialSense : An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition ," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 2051–2060, IEEE Computer Society, Nov. 2019.

[10] A. Goyal, K. Yang, D. Yang, and J. Deng, "Rel3d : A minimally contrastive benchmark for grounding spatial relations in 3d," *ArXiv*, vol. abs/2012.01634, 2020.

[11] P. Matsakis and L. Wendling, "A new way to represent the relative position between areal objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 634–643, 1999.

[12] P. Matsakis, M. Naeem, and F. Rahbarnia, "Introducing the ϕ-descriptor - a most versatile relative position descriptor," in *International Conference on Pattern Recognition Applications and Methods*, 2015.

[13] M. Clément, C. Kurtz, and L. Wendling, "Bags of spatial relations and shapes features for structural object description," 12 2016.

[14] L. Servant, M. Clément, L. Wendling, and C. Kurtz, "Contrastive learning of image representations guided by spatial relations," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2124–2133, 2025.

[15] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Affordance transfer learning for human-object interaction detection," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 495–504, 2021.

[16] L. Bai, K. Li, J. Pei, and S. Jiang, "Main objects interaction activity recognition in real images," *Neural Computing and Applications*, vol. 27, pp. 335–348, Feb. 2016. Publisher Copyright : © 2015, The Natural Computing Applications Forum.

[17] W. Xu, Z. Miao, X.-P. Zhang, and Y. Tian, "A hierarchical spatio-temporal model for human activity recognition," *Trans. Multi.*, vol. 19, p. 1494–1509, July 2017.

[18] D. A. Chacra and J. Zelek, "The topology and language of relationships in the visual genome dataset," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4859–4867, 2022.

[19] Z. Zeng, A. Röfer, and O. Jenkins, "Semantic linking maps for active visual object search," 06 2020.

[20] M. J. EGENHOFER and R. D. FRANZOSA, "Point-set topological spatial relations," *International journal of geographical information systems*, vol. 5, no. 2, pp. 161–174, 1991.

[21] D. A. Randell, Z. Cui, and A. G. Cohn, "A spatial logic based on regions and connection," in *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, KR'92, (San Francisco, CA, USA), p. 165–176, Morgan Kaufmann Publishers Inc., 1992.

[22] A. Cohn, B. Bennett, J. Gooday, and M. Gotts, "Qualitative spatial representation and reasoning with the region connection calculus," *GeoInformatica*, vol. 1, 10 1997.

[23] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, p. 832–843, Nov. 1983.

[24] N. Van de Weghe, B. Kuijpers, P. Bogaert, and P. De Maeyer, "A qualitative trajectory calculus and the composition of its relations," vol. 3799, pp. 60–76, 11 2005.

[25] A. G. Cohn and J. Renz, "Chapter 13 qualitative spatial representation and reasoning," in *Handbook of Knowledge Representation* (F. van Harmelen, V. Lifschitz, and B. Porter, eds.), vol. 3 of *Foundations of Artificial Intelligence*, pp. 551–596, Elsevier, 2008.

[26] J. Chen, A. Cohn, L. Dayou, S. Wang, J. Ouyang, and Q. yu, "A survey of qualitative spatial representations," *The Knowledge Engineering Review*, vol. 30, pp. 106–136, 05 2013.

[27] K. Miyajima and A. Ralescu, "Spatial organization in 2d segmented images : representation and recognition of primitive spatial relations," *Fuzzy Sets Syst.*, vol. 65, p. 225–236, Aug. 1994.

[28] M. Clément, C. Kurtz, and L. Wendling, "Learning spatial relations and shapes for structural object description and scene recognition," *Pattern Recognition*, vol. 84, pp. 197–210, 2018.

[29] H. Kwasnicka and M. Paradowski, "Spread histogram — a method for calculating spatial relations between objects," in *Computer Recognition Systems* (M. Kurzyński, E. Puchała, M. Woźniak, and A. żołnierek, eds.), (Berlin, Heidelberg), pp. 249–256, Springer Berlin Heidelberg, 2005.

[30] M. Clément, A. Poulenard, C. Kurtz, and L. Wendling, "Directional enlacement histograms for the description of complex spatial configurations between objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2366–2380, 2017.

[31] M. Clément, C. Kurtz, and L. Wendling, "Fuzzy directional enlacement landscapes for the evaluation of complex spatial relations," *Pattern Recognition*, vol. 101, p. 107185, 2020.

[32] R. Deléarde, C. Kurtz, P. Dejean, and L. Wendling, "Force banner for the recognition of spatial relations," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6065–6072, 2021.

[33] Y. Wang and F. Makedon, "R-histogram : quantitative representation of spatial relations for similarity-based image retrieval," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, (New York, NY, USA), p. 323–326, Association for Computing Machinery, 2003.

[34] Y. Wang, F. Makedon, and A. Chakrabarti, "R*-histograms : efficient representation of spatial relations between objects of arbitrary topology," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, (New York, NY, USA), p. 356–359, Association for Computing Machinery, 2004.

[35] K. Zhang, K.-p. Wang, X.-j. Wang, and Y.-x. Zhong, "Spatial relations modeling based on visual area histogram," in *2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 97–101, 2010.

[36] K. C. Santosh, L. Wendling, and B. Lamiroy, "Unified pairwise spatial relations : An application to graphical symbol retrieval," in *Graphics Recognition. Achievements, Challenges, and Evolution* (J.-M. Ogier, W. Liu, and J. Lladós, eds.), (Berlin, Heidelberg), pp. 163–174, Springer Berlin Heidelberg, 2010.

[37] L. Servant, C. Kurtz, and L. Wendling, "An extension of the radial line model to predict spatial relations," in *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2023, Volume 4 : VISAPP, Lisbon, Portugal, February 19-21, 2023* (P. Radeva, G. M. Farinella, and K. Bouatouch, eds.), pp. 187–195, SCITEPRESS, 2023.

[38] C. Wen, D. Jayaraman, and Y. Gao, "Can transformers capture spatial relations between objects ?," in *The Twelfth International Conference on Learning Representations*, 2024.

[39] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," vol. 9905, pp. 852–869, 10 2016.

[40] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Weakly-supervised learning of visual relations," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5189–5198, 2017.

[41] L. Servant, M. Clément, L. Wendling, and C. Kurtz, "Spirl : Spatially-aware image representation learning under the supervision of relative position descriptors," *Pattern Recognition*, vol. 170, p. 112013, 2026.

[42] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3107–3115, 2017.

[43] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 381–389, 2018.

[44] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3298–3308, 2017.

[45] R. Deléarde, C. Kurtz, and L. Wendling, "Description and recognition of complex spatial configurations of object pairs with force banner 2d features," *Pattern Recognition*, vol. 123, p. 108410, 2022.

[46] Y. Cong, M. Y. Yang, and B. Rosenhahn, "Reltr : Relation transformer for scene graph generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 04 2023.

[47] C. Islam, O. Mamo, S. Chacko, X. Liu, and W. Yu, "Spatial-vilt : Enhancing visual spatial reasoning through multi-task learning," 10 2025.

[48] X. Ding, Y. Li, Y. Pan, D. Zeng, and T. Yao, "Exploring depth information for spatial relation recognition," pp. 279–284, 08 2020.

[49] A. Kamath, J. Hessel, and K.-W. Chang, "What's "up" with vision-language models? investigating their struggle with spatial reasoning," *ArXiv*, vol. abs/2310.19785, 2023.

[50] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3097–3106, IEEE Computer Society, 2017.

[51] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, *Graph R-CNN for Scene Graph Generation : 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pp. 690–706. 09 2018.

[52] S. Khandelwal, M. Suhail, and L. Sigal, "Segmentation-grounded scene graph generation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15859–15869, 2021.

[53] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation : Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022. Published online 2020, print 2022.

[54] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything : Unleashing the power of large-scale unlabeled data," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, IEEE Computer Society, 2024.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 5998–6008, 2017.

[56] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip : Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, PMLR, 2022.