Amélioration de la segmentation de scènes par l'exploitation de relations spatiales : application aux sceaux byzantins*

Ege Şendoğan[1], Victoria Eyharabide[2], Isabelle Bloch[1]

[1] Sorbonne Université, CNRS, LIP6, Paris, France
[2] Sorbonne Université, STIH, Paris, France

ege.sendogan@lip6.fr

Résumé

L'analyse des scènes iconographiques sur les sceaux byzantins offre des informations précieuses sur les tendances religieuses, sociales et politiques de l'empire. La segmentation des éléments iconographiques permet d'extraire des informations détaillées et quantifiables sur chaque élément de design, facilitant ainsi l'analyse numérique des scènes iconographiques. Cependant, cette tâche est complexe en raison de la détérioration des sceaux, de la mauvaise qualité des images et du manque de données. Dans ce travail, nous proposons une fonction de coût intégrant des contraintes structurelles basées sur les relations spatiales attendues entre les éléments iconographiques lors de l'entraînement d'un réseau de neurones. Nous avons évalué la méthode proposée de manière quantitative et qualitative. Les résultats montrent l'intérêt d'intégrer des connaissances a priori sur les relations spatiales pour améliorer les performances dans une tâche de segmentation sémantique difficile.

Mots-clés

Intelligence artificielle, humanités numériques, segmentation, relations spatiales directionnelles, sceaux byzantins.

Abstract

The analysis of iconographic scenes on Byzantine seals provides invaluable insights into the empire's religious, social, and political trends. The segmentation of iconographic elements in Byzantine seals enables the extraction of detailed, quantifiable information about each design element, facilitating digital analysis of iconographic scenes. However, the task is non-trivial due to challenges such as seal deterioration, poor image acquisition, and the lack of data. In this work, we propose a loss function that incorporates structural constraints based on the expected spatial relationships between iconographic elements in the training of a neural network. The results of the proposed method are evaluated quantitatively and qualitatively. They show the benefit of integrating prior knowledge on spatial relationships into the network to improve performance in a challenging semantic segmentation task.

*Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (ANR), numéro de projet ANR-21-CE38-0001, https://anr.fr/Projet-ANR-21-CE38-0001.

Keywords

Artificial intelligence, digital humanities, segmentation, directional spatial relationships, Byzantine seals.

1 Introduction

Les sceaux en plomb byzantins sont de petits objets circulaires utilisés non seulement pour sécuriser et authentifier des documents, mais aussi pour exprimer l'identité de leur propriétaire. Leurs motifs personnalisés peuvent comporter des monogrammes et des inscriptions indiquant le statut, la fonction et la famille du détenteur. Les sceaux peuvent également porter des images religieuses. Leur analyse constitue une ressource précieuse pour les historiens souhaitant explorer les dynamiques religieuses, sociales et politiques de l'Empire. Toutefois, la compréhension des scènes iconographiques présentes sur les sceaux byzantins nécessite une analyse minutieuse de chaque élément, une tâche complexe qui requiert une expertise avancée en sigillographie ainsi qu'une connaissance approfondie du monde byzantin. L'analyse automatique fondée sur la segmentation sémantique représente ainsi un outil important pour assister les historiens.

L'état de conservation des sceaux byzantins ainsi que la qualité des images acquises limitent les performances des méthodes de segmentation traditionnelles. Les sceaux en plomb sont souvent détériorés en raison de la corrosion, de l'usure et d'autres formes de dégradation, tandis que les conditions d'éclairage médiocres sont fréquentes dans les images disponibles. Ces facteurs dégradent des éléments clés, entraînant une forte variance intra-classe des aspects visuels, ce qui complique l'identification et la différenciation fiables des motifs.

Les réseaux de neurones convolutionnels (CNN) ont révolutionné la segmentation d'images grâce à leur capacité à capturer des caractéristiques locales telles que les contours et les textures. Dans ce paradigme, la tâche de segmentation est formulée comme un problème d'optimisation d'une fonction de coût qui quantifie l'écart entre les prédictions du modèle et les résultats attendus. La conception de cette fonction de coût doit être en adéquation avec les objectifs spécifiques de l'application et tenir compte des caractéristiques du jeu de données, telles que le déséquilibre des

classes ou le bruit. L'entropie croisée (CE), initialement introduite pour les tâches de classification, est l'une des fonctions de coût les plus utilisées en segmentation sémantique. Elle mesure à quel point la distribution de probabilité de la classe prédite s'aligne avec les étiquettes de la référence, pixel par pixel, en appliquant une pénalité logarithmique aux prédictions incorrectes. Cette fonction est sensible au déséquilibre des classes. Pour compenser ce problème, l'entropie croisée pondérée (WCE) est introduite comme une variante de la CE [12]. Elle accorde une importance plus élevée aux pixels appartenant aux classes plus petites et sous-représentées. Elle est sensible au choix des pondérations. La fonction de coût de Dice [9], également couramment utilisée en segmentation sémantique, évalue le degré de recouvrement entre les masques de segmentation prédits et ceux de la référence. Cet avantage est précieux pour gérer le déséquilibre des classes, car elle met l'accent sur les objets détectés. Cependant, la fonction de Dice a tendance à être peu sensible aux petits faux positifs : leur impact sur le score global est faible comparé à celui des régions plus grandes, ce qui peut entraîner la présence de petits faux positifs dans les prédictions du modèle. La fonction de coût focale (FL) [7] reformule fondamentalement le problème du déséquilibre des classes comme un défi de « dominance des exemples faciles », en introduisant un facteur de pondération dynamique qui réduit l'apport des exemples bien classés au coût global. La FL encourage ainsi le modèle à apprendre des exemples difficiles. Elle reste sensible au choix des paramètres. Ces fonctions de coût sont classiquement utilisées dans diverses architectures de CNN, telles que U-Net [12], ou encore DeepLabv3+ [4] que nous utilisons ici.

L'exploitation de connaissances externes est souvent essentielle pour améliorer le raisonnement contextuel des réseaux de neurones. Dans ce contexte, les fonctions de coût se sont révélées être un moyen efficace d'intégrer ces connaissances à leurs processus d'entraînement. Benkirane et al. [1] ont utilisé des connaissances a priori sur les relations spatiales topologiques (*region connection calculus*) intégrées dans la fonction de coût pour améliorer les performances des réseaux neuronaux profonds dans la segmentation panoptique. Riva [10] a également tiré parti des connaissances sur les relations spatiales dans deux nouvelles fonctions de coût intégrant la satisfaction des relations directionnelles.

La capacité des CNN à raisonner à partir des informations contextuelles, à condition que le champ réceptif soit suffisamment grand, reste limitée par les données dont ils disposent [11]. Le nombre limité de sceaux byzantins et les difficultés liées à leur annotation empêchent la création d'un grand jeu de données annotées, ce qui limite l'application des méthodes basées sur les CNN précédentes. Dans ce travail, nous introduisons une fonction de coût qui intègre des contraintes structurelles, basées sur les relations spatiales attendues entre les éléments iconographiques lors de l'entraînement du réseau. Les sceaux byzantins présentent des scènes religieuses structurées où les éléments de design des sceaux portant une iconographie similaire suivent des positionnements relatifs cohérents. En guidant l'entraînement du réseau à l'aide de connaissances a priori sur la structure des scènes iconographiques, nous visons à aider le réseau à mieux comprendre le design global, c'est-à-dire le contexte général du sceau, ce qui permet d'améliorer les résultats de segmentation. Les principales contributions de notre article sont les suivantes :

- une nouvelle fonction de coût pour guider l'entraînement des réseaux neuronaux en utilisant les connaissances a priori sur les relations spatiales,
- une méthode automatisée pour analyser les photos de sceaux byzantins,
- des résultats sur une série de sceaux, confirmant l'utilité des relations spatiales pour améliorer la segmentation sémantique.

2 Méthode

Afin d'améliorer la segmentation et la reconnaissance des éléments iconographiques des sceaux byzantins, nous proposons une méthode qui exploite les connaissances a priori sur l'agencement spatial de ces éléments. Au cœur de cette approche, nous introduisons une nouvelle fonction de coût basée sur les relations spatiales, permettant l'intégration de la structure inhérente (c'est-à-dire l'organisation spatiale) des sceaux byzantins dans un réseau neuronal pour la segmentation multi-objets. Cette approche répond aux défis posés par l'état actuel des sceaux byzantins, ainsi qu'aux problèmes liés à leur acquisition d'images, tels que des conditions d'éclairage médiocres.

Étant donné que les relations spatiales sont sujettes à l'imprécision, nous nous appuyons sur la théorie des ensembles flous, qui offre un cadre robuste pour modéliser cette imprécision [3]. Pour toute paire d'objets, les relations spatiales sont calculées entre ces objets, à la fois dans la segmentation de référence et dans la prédiction, ce qui donne lieu à deux scores que nous définissons comme des degrés de satisfaction des relations. La fonction de coût utilisée dans le réseau neuronal compare ces deux scores, et sa minimisation favorise une segmentation sémantique conforme aux relations attendues. Étant donné un ensemble R de paires d'objets source-cible, où $|R| \geq 1$, nous définissons une fonction de coût qui compare les scores de relation de la manière suivante :

$$\mathcal{L}_{\text{Rel}} = \frac{1}{|R|} \sum_{(S,T)\in R} |\mu(S,T) - \mu(\hat{S},\hat{T})|, \qquad (1)$$

où $\mu(S,T)$ désigne le score d'une relation entre S et T dans la segmentation de référence, et $\mu(\hat{S},\hat{T})$ le score de cette même relation calculée pour $\hat{S}$ et $\hat{T}$, les prédictions correspondant à S et T.

2.1 Modélisation des relations spatiales

Les relations spatiales sont modélisées par une approche morphologique floue [2]. Cette approche repose sur une dilatation floue par un élément structurant flou qui caractérise la sémantique des relations. Dans cet article, nous utilisons principalement des relations directionnelles, pour

lesquelles l'élément structurant est modélisé sous la forme d'un cône flou : la direction d'intérêt entre deux objets S et T est définie par l'angle $\theta_{(S,T)}$ entre la direction horizontale et la ligne reliant les centres de S et T, puis pour chaque point (i, j) de l'élément structurant, noté $\kappa_{(S,T)}$, on calcule l'angle entre la ligne reliant l'origine à ce point et la direction souhaitée, et la valeur d'appartenance de ce point est une fonction décroissante de cet angle [2, 10], par exemple :

$$\kappa_{(S,T)}(i,j) = \max\left(1 - \frac{2\left|\arctan\left(\frac{j}{i+\epsilon}\right) - \theta_{(S,T)}\right|}{\eta}, 0\right). \quad (2)$$

Le paramètre η contrôle le taux de décroissance des degrés d'appartenance, définissant ainsi à quel point la relation spatiale entre S et T est interprétée de manière stricte ou souple.

Bien que cette définition de l'élément structurant gère efficacement les relations binaires, de nombreux cas réels impliquent des interactions spatiales plus complexes. Certains objets sont constitués de plusieurs composantes connexes (voir figure 1c-e), ce qui signifie qu'un seul objet source peut être en relation simultanément avec plusieurs parties d'un objet cible. Deux objets peuvent également satisfaire plusieurs relations, par exemple lorsqu'un objet en entoure un autre, c'est-à-dire qu'un objet cible peut se trouver dans toutes les directions par rapport à l'objet source (comme c'est le cas pour la relation entre l'Enfant et le médaillon dans la figure 1).

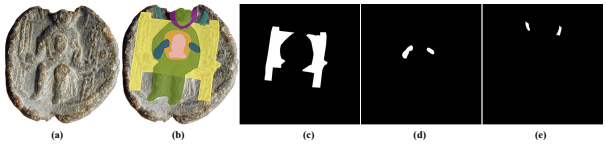

FIGURE 1 – Sceau byzantin (a) et annotations de la Théotokos avec un voile et un nimbe, assise sur un trône et tenant le médaillon de l'Enfant, accompagné de deux mains (b). Cartes de segmentation des classes trône, main et nimbe. Le trône et la Théotokos se superposent, ce qui entraîne une représentation du trône sous forme d'une structure à plusieurs composantes, composée de deux parties disjointes et visibles de part et d'autre de la Théotokos (c). Les mains sont par nature constituées de deux composantes (d), et le nimbe est divisé en deux parties en raison d'un fragment manquant du sceau (e).

Pour modéliser ces interactions spatiales complexes (qu'il s'agisse de la réunion de composantes ou de relations), nous définissons l'élément structurant sur un support carré comportant N distributions en forme de cône centré en $(0, 0)$. Par exemple, N représente le nombre de directions nécessaires pour capturer la relation "entoure" ou le nombre de composantes connexes de l'objet cible T, où $T = \cup_{n=1}^{N} t_n, N \geq 1$. L'élément structurant $\kappa_{(S,t_n)}$ représentant la relation entre l'objet source S et chaque composante t_n de T est calculé à l'aide de l'équation (2). Ces éléments structurants individuels sont ensuite agrégés en une représentation unifiée $\kappa'_{(S,T)}$, où la réunion floue utilisée dans cet article est le maximum :

$$\kappa'_{(S,T)}(i,j) = \max_{n\in\{1\ldots N\}} \kappa_{(S,t_n)}(i,j). \quad (3)$$

Une fois l'élément structurant $\kappa'(S, T)$ obtenu, il est utilisé pour construire le paysage flou $\varphi^{(S,T)}$ autour de S, où la valeur en chaque point représente le degré de satisfaction de la relation spatiale d'intérêt. Dans [2], $\varphi^{(S,T)}$ est modélisé comme une dilatation morphologique de S par $\kappa'(S, T)$. Cependant, les opérations morphologiques ne sont pas différentiables, ce qui rend difficile l'optimisation basée sur le gradient en apprentissage profond. Pour y remédier, une stratégie courante consiste à utiliser des approximations différentiables des opérations morphologiques, comme la moyenne contre-harmonique (CHM) [8], ou les fonctions α-softmax et LogSumExp [6, 13]. Une autre approche consiste à modifier le processus de rétropropagation, permettant l'utilisation directe des opérateurs morphologiques [5]. Ici nous adoptons une approximation par convolution [10], qui offre une solution à la fois efficace sur le plan computationnel, sans hyperparamètres, et bien adaptée à l'optimisation des réseaux par gradients :

$$\varphi^{*(S,T)}(i,j) = (S * \kappa'_{(S,T)})(i,j), \quad (4)$$

où $*$ désigne l'opération de convolution.

Enfin, nous calculons un score de relation pour comparer l'objet cible T avec le paysage flou généré autour de l'objet source S, défini comme la moyenne de l'intersection (définie ici comme un produit) de T et de $\varphi^{(S,T)}$:

$$\mu_{(S,T)} = \frac{\sum_{(i,j)\in\Omega} T(i,j)\varphi^{(S,T)}(i,j)}{\sum_{(i,j)\in\Omega} T(i,j)}, \quad (5)$$

où Ω représente l'espace des coordonnées, et $T(i, j)$ la valeur d'appartenance de l'objet cible T au pixel (i, j). Le terme $\varphi^{(S,T)}$ désigne le paysage flou $\varphi^{*(S,T)}$ normalisé dans $[0, 1]$. Ce score reflète dans quelle mesure T s'insère dans le paysage flou qui met en évidence les zones d'interaction entre S et son environnement dans une direction donnée, un score de 0 indiquant une absence totale d'alignement spatial, et un score de 1 traduisant une satisfaction parfaite de la relation spatiale.

2.2 Apprentissage intégrant des connaissances a priori sur les relations spatiales

Un jeu de données classique de segmentation sémantique, composé de paires d'images de sceaux et des images étiquetées correspondantes (x^k, y^k), peut être enrichi par l'ajout de chaque relation (S, T) appartenant à R, l'ensemble des relations présentes dans la $k^{\text{ème}}$ image. Dans ce jeu de données étendu, appelé jeu de données informé, chaque échantillon se compose désormais d'une image de sceau (x^k), de l'image étiquetée (y^k), d'un ensemble d'éléments structurants représentant les relations spatiales (κ'^k), et des scores de relations correspondants (μ^k).

Un réseau de neurones convolutionnel traditionnel pour la segmentation multi-objets, paramétré par ψ, peut être formulé comme suit :

$$f_\psi : \mathbb{R}^{H\times W\times C} \to [0,1]^{H\times W\times Z}, \tag{6}$$

où $\mathbb{R}^{H\times W\times C}$ représente le domaine spatial de l'image d'entrée, de taille $H \times W$ et C canaux, et Z désigne le nombre d'objets. Le réseau est entraîné à l'aide d'une fonction de coût au niveau du pixel, ici l'entropie croisée :

$$\mathcal{L}_{CE}(\hat{y}^k, y^k) = -\frac{1}{H\cdot W}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{z=0}^{Z-1} y^k_{(h,w,z)} \log(\hat{y}^k_{(h,w,z)}), \tag{7}$$

où $\hat{y}$ représente la sortie de f_ψ, et $\sum_{z=0}^{Z-1} \hat{y}^k_{(h,w,z)} = 1, \quad \forall h, w$. Pour entraîner le réseau avec la nouvelle fonction de coût $\mathcal{L}_{Rel}$, nous introduisons un module original qui calcule le score de relation spatiale entre les objets prédits correspondant à ceux de la vérité terrain (S, T). Ce module prend en entrée la sortie de f_ψ. Pour chaque relation (S, T) présente dans l'image, il extrait les segmentations prédites $\hat{S}$ et $\hat{T}$. Le paysage flou $\varphi^{(\hat{S},\hat{T})}$ est ensuite calculé en utilisant $\hat{S}$ et l'élément structurant associé à la relation de référence (S, T), $\kappa'(S, T)$, tel que défini par l'équation (4). Enfin, le score de relation est calculé par l'équation (5).
Ce module peut être intégré dans des architectures CNN largement utilisées, avec seulement des modifications minimales à la structure existante du réseau. Pour nos expériences, nous avons utilisé l'architecture DeepLabV3+, qui est largement utilisée pour les tâches de segmentation sémantique. DeepLabV3+ a été choisi pour sa forte capacité à capturer le contexte local et global grâce à l'ASPP (*Atrous Spatial Pyramid Pooling*). Afin d'optimiser le modèle, nous avons combiné la fonction de coût proposée, pondérée par un paramètre α, avec $\mathcal{L}_{CE}$:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{Rel}. \tag{8}$$

3 Expériences

La Theotokos (Vierge Marie) occupe une place importante dans la sigillographie byzantine, et est représentée sur un nombre considérable d'images. Nous avons sélectionné 108 images de sceaux représentant la Theotokos comme figure centrale (voir figure 1). Ces images proviennent de collections prestigieuses, notamment celles d'Henri Seyrig, Yavuz Tatış et Georges Zacos, conservées à la Bibliothèque nationale de France (BnF). Ces images ont été annotées par des experts en sigillographie byzantine, identifiant huit classes cibles distinctes. Compte tenu de la complexité intrinsèque de l'annotation des sceaux byzantines, le processus a impliqué un consensus entre plusieurs annotateurs. La **Theotokos** est souvent représentée avec un **voile** et un **nimbe** au-dessus de sa tête. De plus, la Theotokos tient souvent l'**Enfant** portant un **nimbe crucifié**. Les **mains** de la Theotokos sont représentées dans diverses postures, telles que la position orante avec les mains levées, une posture de prière avec les mains élevées, un geste de guidage lorsqu'elle présente l'Enfant et d'autres variations vues dans différentes iconographies de sceaux. Deux représentations courantes sur les sceaux byzantins — la Theotokos trônant et la Theotokos tenant un Médaillon de l'Enfant — sont également présentes dans le jeu de données. Par conséquent, le **trône** et le **médaillon** figurent parmi les objets identifiés par les annotateurs.

Les relations spatiales entre ces éléments iconographiques présentent des propriétés cohérentes à travers différents sceaux. Par exemple, lorsque le trône est présent, il est toujours positionné à gauche et à droite de la Theotokos, tandis que le voile se trouve systématiquement au-dessus de sa tête. De même, le nimbe et le nimbe crucifié sont toujours situés au-dessus de la Theotokos et de l'Enfant, respectivement. Le médaillon entoure toujours l'Enfant. Lorsque ces éléments sont présents, nous ajoutons les relations suivantes au jeu de données informé afin d'enrichir l'apprentissage du réseau : (Theotokos, trône), (Theotokos, voile), (Theotokos, nimbus), (Enfant, nimbus crucifié), (Enfant, médaillon), en utilisant systématiquement la Theotokos et l'Enfant comme objets de référence (i.e., source) dans chaque relation, puisque tous les autres objets sont contextuellement liés à eux. De plus, la position relative des mains de la Theotokos et de l'Enfant varie selon le style iconographique, tout en restant constante au sein des sceaux appartenant à un même style. Certaines sceaux représentent la Théotokos tenant l'Enfant ou le médaillon de l'Enfant, avec ses mains placées près de lui. Dans ces cas, nous utilisons la relation (Enfant, mains). Pour les sceaux représentant une posture orante ou de prière, nous appliquons la relation (Théotokos, mains), et la relation (Théotokos, Enfant) lorsque la Théotokos est représentée tenant l'Enfant à ses côtés. Nous avons sélectionné ces relations spécifiques de manière à maximiser l'information structurelle, sans introduire une complexité computationnelle excessive lors de l'entraînement du réseau. Le jeu de données enrichi est obtenu comme expliqué dans la section 2.2. Des exemples sont illustrés dans la figure 2.

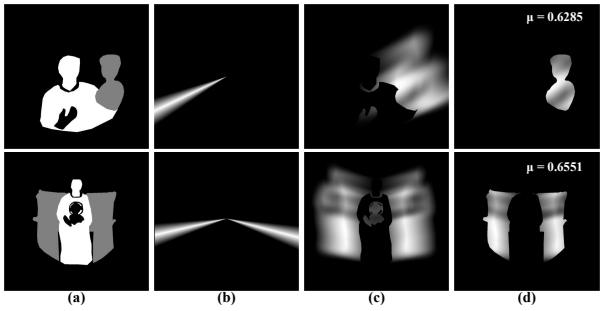

FIGURE 2 – Première ligne : (a) carte d'annotations contenant la Théotokos et l'Enfant, (b) élément structurant $\kappa'(Theotokos, enfant)$, (c) paysage flou, (d) son intersection avec l'Enfant et score de relation $\mu(Theotokos, enfant)$. Deuxième ligne : (a) carte d'annotations contenant la Théotokos et le trône, (b) élément structurant $\kappa'(Theotokos, trone)$, (c) paysage flou, (d) son intersection avec le trône et score de relation $\mu(Theotokos, trone)$.

Nous avons mené nos expériences en utilisant une validation croisée à 5 plis. Différentes valeurs de α ont été testées afin d'évaluer l'impact de $\mathcal{L}_{Rel}$. Nous avons utilisé l'optimisation par descente de gradient stochastique avec un momentum de 0,9, et un taux d'apprentissage initial fixé à 0,01, diminué progressivement jusqu'à 0,0001 au cours de l'entraînement. La taille de batch était de 2, et les modèles ont été entraînés pendant 25000 itérations. L'augmentation de données a été réalisée à la volée à l'aide d'une correction gamma aléatoire, d'une égalisation adaptative d'histogramme à contraste limité (CLAHE), ainsi que de retournements horizontaux aléatoires avec une probabilité de 0,5. Un redimensionnement et un recadrage aléatoires ont également été appliqués. Lorsqu'une image d'entrée est retournée ou redimensionnée, les mêmes transformations sont appliquées aux éléments structurants associés. Si une relation disparaît à la suite du recadrage, elle est exclue du calcul de $\mathcal{L}_{Rel}$. Les expériences ont duré entre 3 et 8 heures sur un GPU NVIDIA A100 Tensor Core avec CUDA 12.0.

4 Résultats et conclusion

Nous avons évalué les résultats à la fois quantitativement et qualitativement. L'évaluation quantitative a été réalisée par des mesures classiques de Dice, précision et rappel. Nous présentons les performances globales moyennes, à savoir la moyenne du coefficient de Dice (mD), de la précision (mP) et du rappel (mR), calculées sur l'ensemble des classes (voir table 1). Nous comparons notre méthode au modèle de référence, entraîné uniquement avec l'entropie croisée ($\alpha = 0$).

La table 1 présente les résultats des moyennes $\pm$ écarts-types de mD, mP et mR, calculés sur les 5 plis de validation croisée. Le meilleur score de mD (75,7 $\pm$ 1,9) ainsi que celui de mR (73,2 $\pm$ 2,5) ont été obtenus en utilisant la fonction de coût proposée avec $\alpha = 0{,}33$. La meilleure précision moyenne mP (80,4 $\pm$ 1,9) est atteinte avec $\alpha = 1{,}25$, également avec la fonction de coût proposée. Une amélioration de chacune des mesures est également observée pour $\alpha = 0{,}33$ par rapport au modèle de référence. La fonction de coût proposée produit des résultats cohérents sur les différentes divisions de données pour mD et mR, comme en témoigne la réduction de la variance pour plusieurs valeurs de α supérieures à 0.

Le trône, le médaillon et le nimbe cruciforme sont les classes les moins représentées dans le jeu de données. Le modèle de référence présente une forte variance du rappel pour ces classes. Le rappel montre une réduction notable de la variance lorsque $\alpha > 0$. La variance du rappel pour la classe trône passe de 19,1 à $\alpha = 0$ à des valeurs bien plus faibles (entre 8,3 et 11,4) lorsque α est compris entre 0,25 et 0,75. Les variances du rappel pour le médaillon et le nimbe cruciforme diminuent respectivement de 7,9 et 9,3 à $\alpha = 0$ à 1,8 et 5,2 lorsque $\alpha = 0,33$. Des améliorations mineures au niveau des pixels, qui n'affectent pas significativement les mesures basées sur les pixels, peuvent néanmoins avoir une importance sémantique considérable dans l'analyse des scènes iconographiques. La présence ou l'absence d'objets spécifiques, tels qu'un trône, un médaillon, un enfant ou des mains, ainsi que leur disposition spatiale et le nombre de composantes sont essentiels pour déterminer le type spécifique de Théotokos représenté, et donc pour l'interprétation de la scène. Par ailleurs, la détérioration des sceaux augmente l'incertitude des annotations. Cela introduit du bruit dans la vérité terrain, rendant les mesures basées sur les pixels moins fiables comme indicateurs de la performance réelle. Nous avons donc envisagé une autre évaluation par inspection visuelle. La figure 3 montre les résultats qualitatifs. Dans la première ligne, le modèle de référence a détecté de manière incorrecte des éléments qui ne devraient pas être présents dans la scène (comme montré sur le canevas, un médaillon et deux éléments représentant l'Enfant sont détectés). De telles prédictions erronées entraînent une mauvaise interprétation du type de Théotokos représenté sur le sceau, ce qui est plus problématique que de prédire à tort un élément qui fait partie de la scène. La méthode proposée améliore non seulement le résultat en omettant le médaillon et le premier élément de l'Enfant, mais elle réduit également de manière significative la présence du second élément de l'Enfant. De plus, notre méthode a amélioré la forme de la Théotokos et de ses mains. Dans la deuxième ligne de la figure 3, les mains de la Théotokos tenant le médaillon, un élément de grande importance dans cette représentation iconographique de la Théotokos, sont mieux localisées et détectées lorsque l'on utilise la fonction de coût proposée. De plus, la forme du médaillon et du nimbe a été améliorée.

En conclusion, nous avons proposé d'intégrer des contraintes structurelles reflétant les relations spatiales attendues entre les objets représentés sur les sceaux byzantins afin d'améliorer la segmentation des scènes iconographiques. Les résultats obtenus sont prometteurs et démontrent l'apport de l'intégration des connaissances a priori sur les relations spatiales dans l'entraînement du réseau. En guidant le réseau avec des contraintes structurelles via la fonction de coût proposée $\mathcal{L}_{Rel}$, nous avons amélioré l'identification des éléments iconographiques clés qui définissent des types spécifiques de scènes. De plus, cette approche a permis de supprimer les classes non pertinentes, conduisant à une interprétation sémantique plus cohérente et plus précise des scènes iconographiques. Ce travail souligne également le potentiel de la combinaison de l'apprentissage profond avec des connaissances externes sur les scènes iconographiques pour automatiser leur analyse, une tâche importante dans l'étude des sceaux byzantins.

Les travaux futurs porteront sur la modélisation d'autres relations et sur la conception de mesures d'évaluation qui reflètent mieux l'importance des résultats de segmentation dans le contexte des sceaux byzantins. Ils viseront également à comparer la fonction de coût proposée avec d'autres fonctions existantes. Une autre perspective sera l'utilisation de données synthétiques pour renforcer le jeu d'entraînement, en particulier avec des configurations iconographiques sous-représentées.

TABLE 1 – Dice moyen (mD), précision moyenne (mP) et rappel moyen (mR). Les valeurs sont indiquées sous la forme moyenne $\pm$ écart-type sur les 5 plis, pour différentes valeurs de α pondérant la fonction de coût proposée.

	0	**0,2**	**0,25**	**0,33**	**0,5**	**0,75**	**1**	**1,25**
mD	$75 \pm 2,4$	$74,4 \pm 0,6$	$74,1 \pm 1,7$	$\mathbf{75,7 \pm 1,9}$	$74,1 \pm 1,1$	$73,9 \pm 1,1$	$74,6 \pm 0,7$	$75,1 \pm 2,4$
mP	$79,8 \pm 0,9$	$78,7 \pm 1,3$	$79,5 \pm 2,5$	$80,1 \pm 1,3$	$79,5 \pm 1,6$	$79,6 \pm 1,1$	$78,5 \pm 1,3$	$\mathbf{80,4 \pm 1,9}$
mR	$72,6 \pm 3,6$	$72,2 \pm 1,1$	$71,4 \pm 3,6$	$\mathbf{73,2 \pm 2,5}$	$71,2 \pm 1,3$	$71 \pm 1,2$	$72,3 \pm 1,3$	$72,3 \pm 3,9$

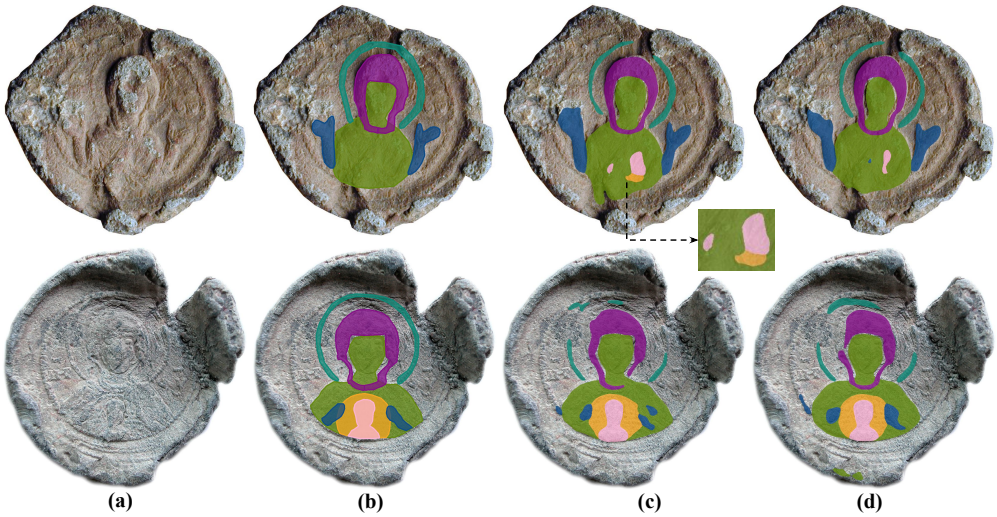

FIGURE 3 – Images de sceaux (a) et images étiquetées (b) contenant les objets Théotokos, voile, nimbe, et mains, et sur la deuxième ligne également les objets Enfant et médaillon. (c) Prédictions du modèle de référence. Dans la première ligne, une petite zone agrandie illustre la prédiction. (d) Résultats de l'approche proposée.

Références

[1] F. E. Benkirane, N. Crombez, V. Hilaire, and Y. Ruichek. Hybrid AI for panoptic segmentation : An informed deep learning approach with integration of prior spatial relationships knowledge. *Comput. Vis. Image Underst.*, 240 :103909, 2024.

[2] I. Bloch. Fuzzy relative position between objects in image processing : a morphological approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(7) :657–664, 1999.

[3] I. Bloch. Fuzzy sets for image processing and understanding. *Fuzzy Sets Syst.*, 281 :280–291, 2015.

[4] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. conf. Computer Vision (ECCV)*, pages 801–818, 2018.

[5] Gianni F., Amin F., and Angela Y. Deep morphological networks. *Pattern Recognition*, 102 :107246, 2020.

[6] A. Kirszenberg, G. Tochon, É. Puybareau, and J. Angulo. Going beyond p-convolutions to learn grayscale morphological operators. In *Int. conf. Discrete Geometry and Mathematical Morphology*, pages 470–482. Springer, 2021.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *IEEE Int. conf. Comput. Vis. (ICCV)*, page 2980–2988, 2017.

[8] J. Masci, J. Angulo, and J. Schmidhuber. A learning framework for morphological operators using counter–harmonic mean. In *Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 329–340, 2013.

[9] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net : Fully convolutional neural networks for volumetric medical image segmentation. In *4th Int. conf. 3D Vision (3DV)*, page 565–571, 2016.

[10] M. Riva. *Spatial Relational Reasoning in Machine Learning : Deep Learning and Graph Clustering*. PhD thesis, Institut Polytechnique de Paris, Télécom Paris, 2022.

[11] M. Riva, P. Gori, F. Yger, and I. Bloch. Is the U-Net directional-relationship aware ? In *IEEE Int. conf. Image Processing (ICIP)*, pages 3391–3395, 2022.

[12] O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional networks for biomedical image segmentation. In *Int. Conf. Med. Image Comput. Comput.-Assist (MICCAI)*, pages 234–241, 2015.

[13] Y Shen, F. Y. Shih, X. Zhong, and I.-C. Chang. Deep morphological neural networks. *Int. J. Pattern Recognit. Artif. Intell.*, 36(12) :2252023, 2022.