Apprentissage contrastif de modèles de processus ponctuels pour la détection d'objets

Jules MABON [1] Mathias ORTNER [2] Josiane ZERUBIA [1]

[1] Inria, Université Côte d'Azur, Sophia Antipolis, France

[2] Airbus Defense and Space, Toulouse, France

Résumé – Cet article présente un modèle de détection d'objets fondé sur un processus ponctuel, qui permet de considérer des interactions entre les objets, tout en exploitant l'information extraite par un CNN. Nous proposons également une méthode d'apprentissage contrastif des paramètres du modèle d'énergie, puis appliquons cette méthode à la détection de véhicules dans des images satellitaires optiques.

Abstract – This paper presents a model for object detection based on point processes, which allows considering object interactions, while using the features extracted by a convolutional neural network. We also present a contrastive learning method to infer the parameters of the energy model, to then apply our method to the detection of vehicles in optical satellite images.

1 Introduction

La détection de petits objets – des véhicules pour le cas présent – dans des images satellitaires optiques est rendue complexe de par la nature de ces images : la résolution spatiale induit des tailles d'objets de quelques pixels. De plus, les ombres et occlusions limitent l'information visuelle, tandis que la densité spatiale élevée des objets rend la séparation des instances complexe.

Les réseaux de neurones à convolutions ont montré de bonnes performances pour les tâches de détection [12]. La plupart de ces méthodes extraient d'abord une représentation vectorielle de l'image (*features*) par convolution, puis proposent une série de boites (ou ancres, *anchors*) ensuite affinées par régression [13, 20]. Cependant, la plupart de ces approches ne considèrent pas les interactions entre les objets, outre leur non-superposition.

Les approches fondées sur les processus ponctuels marqués [14], pour leur part, modélisent la détection de l'ensemble des objets au lieu de considérer la détection de chaque objet indépendamment. De tels modèles combinent l'information de l'image (attache aux données) et des *a priori* propres aux objets étudiés et leurs interactions. Ces approches ont été appliquées, entre autres, à la détection dans des images de microscopie [2], ou à des données de télédétection [1, 9, 16]. Cependant, ces modèles se basent sur des mesures de contraste pour l'attache aux données. Ces mesures sont peu robustes aux scènes complexes (ombres, occlusions partielles), et aux variations d'aspect visuel des objets d'intérêt.

L'approche proposée dans cet article formule la détection sous forme d'une minimisation d'énergie, où l'attache aux données du processus ponctuel est issue d'un CNN, et les termes *a priori* régularisent les objets et leurs interactions. Nous proposons également une méthode pour apprendre les paramètres de ce modèle d'énergie. Enfin, nous présentons des résultats obtenus sur des images satellitaires optiques.

2 Processus ponctuels pour la détection

2.1 Processus ponctuels marqués

On définit une configuration de points Y comme un ensemble non ordonné d'éléments de $S \times M$, avec S l'espace de l'image, et M l'espace des marques. Une marque est une variable aléatoire comme le rayon d'un cercle ou une variable catégorielle discrète. Dans notre cas, un objet $y \in Y$ est défini par des coordonnées y_i, y_j dans S, et trois marques qui décrivent un rectangle : largeur y_a, longueur y_b et angle y_α. On dénote par $\mathcal{P} = \bigcup_{n=0}^{\infty}(S \times M)^n$ l'ensemble de toutes les configurations à nombre de points quelconque, et par $|Y|$ le cardinal de Y.

Une configuration de points Y pour une image X est modélisée comme la réalisation d'un Processus Ponctuel Marqué (MPP), de densité h, définie relativement au processus uniforme [14], par une densité de Gibbs à partir d'une énergie U :

$$h(Y|X) \propto \exp(-U(Y|X)) \tag{1}$$

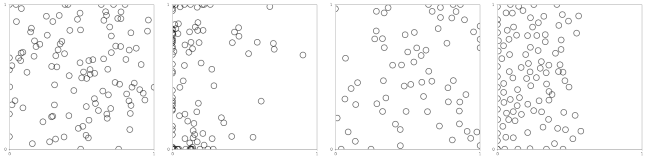
La configuration $\widehat{Y}$ maximisant la probabilité a posteriori pour une image X donnée, minimise l'énergie $U(Y|X)$.

2.2 Modèle d'énergie

L'énergie d'une configuration Y est calculée pour chaque point y comme la somme pondérée de K termes d'énergie $U_k(y)$. En définissant le voisinage de y dans Y comme $\mathcal{N}_y$ (l'ensemble des points $y' \neq y$ dans Y à une distance $d(y, y')$ inférieure à d_{max}) :

$$U(Y|X, \theta) = \sum_{y \in Y} \theta_{w,0} + \sum_{k=1}^{K} \theta_{w,k} U_k(y|X, \mathcal{N}_y, \theta) \tag{2}$$

On distingue les U_k en deux catégories ; d'une part des termes *a priori* ($U_k(y|\mathcal{N}_y, \theta)$) qui ne dépendent que de y et ses voisins $\mathcal{N}_y$, et mesurent la cohérence de la configuration indépendamment de l'image. D'autre part, les termes d'*attache aux données* ($U_k(y|X, \theta)$) sont fonction de y et l'image X et mesurent la correspondance des points à l'image.



(a) uniforme (b) gradient (c) répulsif (d) gradient + répulsif

FIGURE 1 : Processus ponctuels dérivés des énergies suivantes ; (a) $U_a(y) \propto 1$, (b) $U_b(y) \propto y_i$, (c) $U_c(y|\mathcal{N}_y) \propto \min_{y' \in \mathcal{N}_y} d(y, y')$, (d) $U_d(y|\mathcal{N}_y) \propto U_b + U_c$

Les paramètres du modèle d'énergie sont représentés par θ (poids des termes d'énergie $\theta_{w,0}, \ldots, \theta_{w,K}$ et le paramètre t_{pos}, voir Eq. 3).

La formulation sous forme de somme de l'équation 2 implique une bonne explicabilité du modèle, tout en permettant de composer les *a priori* et l'attache aux données qui décrivent des distributions de points (voir Fig. 1)

Dans cet article, nous présentons la construction des termes d'attache aux données à partir de la sortie d'un CNN, puis les *a priori* du modèle (illustré en Fig. 2). Une fois les U_k définis, nous détaillons la méthode de divergence contrastive pour apprendre les paramètres θ, ainsi que la méthode d'échantillonnage de $h(Y|X, \theta)$. Enfin, nous présentons des résultats obtenus sur des images satellitaires.

2.3 CNN pour l'attache aux données

Les modèles MPP classiques [1, 2, 9, 16] utilisent des mesures de contraste adaptées à chaque application, qui nécessitent un contraste fort entre les objets d'intérêt et le fond.

Dans cette partie, nous interprétons la sortie d'un modèle de CNN pour la détection d'objets, comme une énergie [4] mesurant la correspondance de la configuration par rapport à l'image.

Terme de position Pour un modèle de réseau de neurones pleinement convolutif (comme Unet [17]) qui produit une carte de probabilité des centres (*heatmap*) [8, 15] à partir d'une image de taille h, w, la carte est produite par l'application de la fonction sigmoïde à un tenseur A^{pos} de taille h, w. On construit alors l'énergie de position comme suit :

$$U_{pos}(y_i, y_j|X) = \ln(1 + \exp(-A^{pos}_X[y_i, y_j] + t_{pos})) \quad (3)$$

où $A^{pos}_X[y_i, y_j]$ est la valeur interpolée du tenseur A^{pos} aux coordonnées y_i, y_j, et t_{pos} un paramètre de seuil, appris avec θ (voir partie 3).

Termes sur les marques Un modèle de classification nous donne, pour une position dans l'image et chaque marque $m \in \{a, b, \alpha\}$, les probabilités sur les N_m marques discrétisées m_k, $k = 1, ..., N_m$, comme $p(m_k|y_i, y_j, X) = \text{Softmax}(A^m_{y_i,y_j,X})[k]$, avec $A^m_{y_i,y_j,X}$ un vecteur issu du CNN de taille N_m. L'énergie sur chaque marque $m = a, b, \alpha$ est obtenue par[1] :

$$\begin{aligned} U_m(y_m|y_i, y_j, X) = &- A^m_{y_i,y_j,X}[y_m] \\ &+ \ln \sum_{k=1}^{N_m} \exp(A^m_{y_i,y_j,X}[m_k]) \end{aligned} \quad (4)$$

[1] On note $V[k]$, pour $k = 1, \ldots, N_m$, la valeur d'un vecteur $V \in \mathbb{R}^{N_m}$ à l'indice k. On obtient $V[y_m]$, pour $y_m \in [m_{\min}, m_{\max}]$, en interpolant V à la position $N_m(y_m - m_{\min})/(m_{\max} - m_{\min})$.

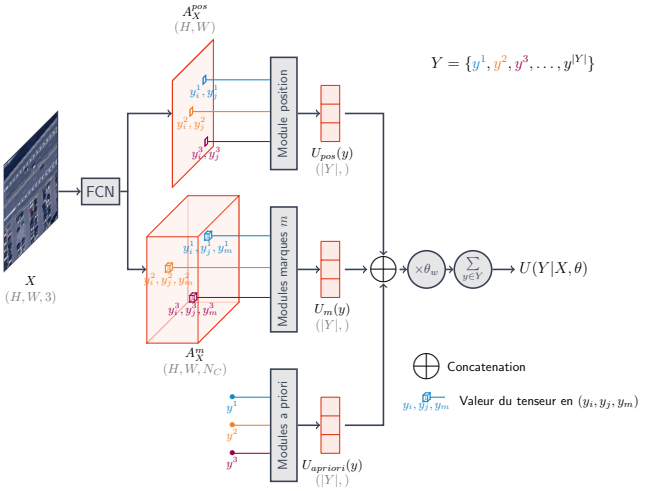


FIGURE 2 : Vue schématique du modèle

2.4 A priori sur les configurations

Aire et ratio On définit deux termes d'énergie U_{aire} et U_{ratio}, ces termes ne dépendent que du point courant y, avec $k = ratio, aire$:

$$U_k(y) = -\exp\left(-0.5(f_k(y) - \mu_k)^2 \sigma_k^{-2}\right) \quad (5)$$

avec $f_{ratio}(y) = y_a/y_b$ et $f_{aire}(y) = y_a y_b$, favorisant ainsi les points d'aire ou de ratio μ_k avec un écart type σ_k.

A priori d'interaction Le terme de l'Eq. 6 pénalise la superposition d'objets ; Eq. 7 favorise l'alignement entre les objets ($t_\alpha = 0$ favorise les objets parallèles, $\pi/2$ les objets perpendiculaires) ; Eq. 8 pénalise les objets proches ; enfin Eq. 9 permet d'ajuster l'énergie des points sans voisins.

$$U_{superp.}(y, \mathcal{N}_y) = \max_{y' \in \mathcal{N}_y} \left\{ \frac{\text{aire}(y' \cap y)}{\min\{\text{aire}(y'), \text{area}(y)\}} \right\} \quad (6)$$

$$U_{align}(y, \mathcal{N}_y) = \min_{y' \in \mathcal{N}_y} \{-\cos(|y_\alpha - y'_\alpha| - t_\alpha)\} \quad (7)$$

$$U_{repuls.}(y, \mathcal{N}_y) = \max_{y' \in \mathcal{N}_y} \left\{ 1 - \frac{d(y, y')}{d_{max}} \right\} \quad (8)$$

$$U_{const.}(y, \mathcal{N}_y) = \mathbb{1}_{|\mathcal{N}_y|=0} \quad (9)$$

3 Apprentissage des paramètres du modèle

Ici nous supposons le modèle de CNN pré-entrainé [15] ; il reste à inférer les paramètres du modèle d'énergie θ à partir des données $\mathcal{S}$ (ensemble de paires images/annotations X, Y^+). Les approches précédentes [1] construisent un ensemble de configurations perturbées Y^- à partir des vérités terrain Y^+, pour en construire des contraintes $U(Y^-) > U(Y^+)$; θ est alors obtenu par la résolution d'un problème linéaire sous contraintes. Cependant, cette approche est très sensible à la qualité des annotations, pouvant mener à un problème surcontraint.

Notre approche se fonde sur la maximisation de la vraisemblance [11] par rapport aux données $\mathcal{S}$:

$$P(Y_1^+, ..., Y_{|\mathcal{S}|}^+|X_1, ..., X_{|\mathcal{S}|}, \theta) = \prod_{(Y^+,X) \in \mathcal{S}} P(Y^+|X, \theta). \quad (10)$$

Maximiser la vraisemblance requiert de calculer une intégrale sur $\mathcal{P}$ [11], Hinton [7] propose plutôt de l'approximer par un unique échantillon contrastif et un gradient stochastique (SGD). Cet échantillon contrastif Y^- est issu d'une chaine de Markov simulant $h(Y|X,\theta_t)$. Du *et al.* [3] réduisent le nombre nécessaire d'itérations d'échantillonnage $N_{\text{ech.}}$, en initialisant l'échantillonnage de Y^- à l'étape t de l'estimation avec le Y^- obtenu à $t-1$ (via une mémoire β). Le coût à minimiser lors de la descente de gradient est formulé comme suit, avec $\gamma > 0$ et $\mathcal{R}$ la moyenne des énergies par objet au carré (régularisation limitant l'explosion de l'énergie) :

$$\mathcal{L}(\theta_t, Y^+, Y^-, X) = U(Y^+|X,\theta_t) - U(Y^-|X,\theta_t) + \gamma\mathcal{R} \tag{11}$$

Algorithme 1 : Estim. θ par divergence contrastive

$\beta \leftarrow \emptyset$
pour $t = 0, \ldots, N_{estm.}$ **faire**
  pour $(X, Y^+) \in \mathcal{S}$ **faire**
    $\widetilde{Y}_0 \sim \beta$ avec proba. 0.99, sinon $\mathcal{U}(\mathcal{P})$
    $Y^- \leftarrow \texttt{SampleMPP}(Y_0, X, \theta_t)$ *(Voir Algo. 2)*
    $\beta \leftarrow \beta \cup Y^-$
    $\Delta\theta_t \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, Y^+, Y^-, X)$
    Mettre à jour θ_{t+1} avec $\Delta\theta_t$ via SGD
  fin
fin

4 Échantillonnage du MPP

4.1 RJMCMC

La méthode the Chaine de Markov Monte Carlo à Sauts Réversibles (RJMCMC) [5] permet d'échantillonner une configuration selon $h(Y|X,\theta)$. Le RJMCMC appliqué aux processus ponctuels [1, 9, 16] est une adaptation de Metroplis Hastings, permettant des *sauts* entre dimensions (nombre de points dans Y) par l'ajout –naissance– et le retrait –mort– de points dans la configuration courante. Ce noyau de naissance et mort uniforme est nécessaire à la convergence [5].

Pour accélérer la convergence, nous proposons des naissances de y dans $S \times M$ selon une densité $d(y)$ [9], issue des cartes d'énergies pré-calculées (Eq. 3, 4) :

$$d(y) \propto \exp(-\sum_{k\in\mathcal{K}} \theta_{w,k} U_k(y|X)),\ \mathcal{K} = \{pos, a, b, \alpha\} \tag{12}$$

La densité du noyau de naissance de Y à $Y' = Y \cup \{y\}$ (Eq. 13), est équilibrée avec celle du noyau de mort (Eq. 14) :

$$Q_B(Y \to Y') = \frac{d(y)}{\lambda} \tag{13} \quad Q_D(Y' \to Y) = \frac{1}{|Y'|} \tag{14}$$

4.2 Diffusion à sauts

Le RJMCMC appliqué aux MPPs utilise aussi des transformations aléatoires pour explorer $\mathcal{P}$ [5]. Les mouvements proposés sont indépendants de la topographie de l'énergie U atour de la configuration courante Y_t. Pour notre modèle, des librairies comme PyTorch[2] permettent la différentiation automatique de U par rapport à Y. Ainsi, propose-t-on d'exploiter ce gradient de l'énergie en Y_t pour explorer plus efficacement l'espace des configurations, via le processus de diffusion stochastique (ou dynamique de Langevin) [6, 10]. Pour une température T et un pas de gradient δ :

$$Y' = Y - \delta\nabla_Y U(Y|X,\theta) + dw\sqrt{2T},\ dw \sim \mathcal{N}(0,\delta). \tag{15}$$

La diffusion – qui ne peut atteindre qu'un sous-ensemble de $\mathcal{P}$ à nombre de points fixés – est alternée avec les noyaux de naissance et mort (sauts) pour explorer tout $\mathcal{P}$.

4.3 Parallélisation

Le RJMCMC canonique pour les MPPs ajoute et retire des points dans Y un par un. Cependant, il est possible d'exploiter la markovianité spatiale (les perturbations sur des points suffisamment distants induisent des variations d'énergie indépendantes) pour exécuter les noyaux en parallèle. Nous adaptons l'approche de [18] ; découpant l'espace S en cellules c de taille constante $2.d_{max}$, appartenant chacune à une couleur C, de façon à ce que deux cellules voisines ne soient pas de la même couleur. Cela garantit l'indépendance des perturbations dans les cellules d'une même couleur [18]. La procédure de choix des cellules à simuler $\widetilde{C}$, déroulée dans l'Algo. 2, permet de proposer (et retirer) en parallèle des points avec une densité $d(y)$ dans S.

Algorithme 2 : $\texttt{SampleMPP}(Y_0, X, \theta_t)$

pour $n = 0, \ldots, N_{ech.}$ **faire**
  Choisir *diffusion* avec proba. 0.8, sinon *saut*
  $Q \leftarrow Q_B$ avec proba. 0.5 sinon Q_D
  Choisir couleur C avec proba. $p(C) = d(C)$ [3]
  $\widetilde{C} \leftarrow \{c \in C$ avec proba. $p(c|C) = \frac{d(c)}{d(C)}\}$
  pour chaque $c \in \widetilde{C}$ **faire**
    si *diffusion* **alors**
      $dw \sim \mathcal{N}(0,\delta)$
      $Y^c_{n+1} \leftarrow Y^c_n - \nabla_{Y^c_n} U(Y^c_n|X,\theta_t)\delta + dw\sqrt{2T}$
    sinon
      $Y^{c\prime}_n \sim Q(Y^c_n \to \cdot)$ [4]
      $r \leftarrow \frac{Q(Y_n{}' \to Y_n)}{Q(Y_n \to Y_n{}')} \exp\left(-\frac{\Delta U(Y_n \to Y_n{}'|X,\theta_t)}{T}\right)$
      $Y^c_{n+1} \leftarrow Y^{c\prime}_n$ avec proba. $\min(1, r)$
    fin
  fin
fin
retourner Y_N

5 Application et résultats

Nous appliquons notre méthode à la détection de véhicules dans des images satellitaires à une résolution de 50 cm/pixel. Le modèle est entrainé sur une version sous-échantillonnée de DOTA [19], ainsi $\mathcal{S}$ correspond aux paires images/annotations de la subdivision d'entrainement de ce jeu de données.

La configuration $\widehat{Y}$ est obtenue avec l'Algo. 2, avec recuit simulé : $T_{n+1} = \alpha T_n$, $\alpha = 0.997$. Pour le score des détections (en vue de tracer des courbes précision-rappel), nous utilisons l'intensité de Papangelou [14] $\lambda(y|\widehat{Y} \setminus \{y\}) = \frac{h(\widehat{Y}|X,\theta)}{h(\widehat{Y}\setminus\{y\}|X,\theta)}$, $\forall y \in \widehat{Y}$, où $\lambda(y|Y)dy$ mesure la probabilité de trouver un point dans dy autour de y sachant Y.

La Table 1, illustre la précision moyenne (*Average Precision*, AP) pour différents seuils d'intersection sur union (pour la

[2] Librairie Python de calcul tensoriel, https://pytorch.org/

[3] $d(c)$ et $d(C)$; densité d resp. intégrée sur une cellule ou une couleur

[4] Notons que l'Eq. 13 devient alors $Q_B(Y^c \to Y^{c\prime}) = \frac{d(y)}{d(c)\lambda}$

FIGURE 3 : Comparaison des méthodes *CNN + local maxima* (rouge) et *CNN + MPP* (jaune)

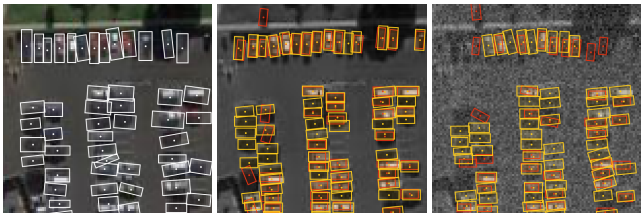


FIGURE 4 : Vérité terrain (blanc, gauche), *CNN + local max.* (rouge) et *CNN + MPP* (jaune) pour une image non bruitée (centre) puis bruitée (droite). L'image est en niveau de gris pour la lisibilité.

correspondance détection / vérité terrain). La Fig. 3 montre une comparaison entre notre méthode (*CNN + MPP*) et notre CNN seul (*CNN + local max.* : seuillage et extraction des maxima locaux) : on constate que le MPP permet une régularisation de la configuration de points et rend la détection plus robuste au bruit (Fig. 4).

Méthode	$AP_{0.10}$	$AP_{0.25}$	$AP_{0.50}$
BBA-Vec. [20]	0.84	0.82	0.41
CNN + local max.	0.86	0.86	0.64
CNN + MPP	**0.91**	**0.90**	**0.64**

TABLE 1 : Précision moyenne (AP) pour différents seuils d'intersection sur union.

6 Conclusion

Nous avons proposé une nouvelle approche pour la détection d'objets qui considère des interactions entre les objets, tout en maintenant un nombre de paramètres peu élevés au sein d'un modèle probabiliste à l'échelle de l'image entière. Ce modèle permet de prendre en compte des *a priori* sur les configurations, et d'estimer l'importance de ces *a priori* par apprentissage contrastif. Cela permet une régularisation des configurations et une robustesse de l'estimation par rapport au bruit de l'image comme illustré dans les résultats ci-dessus. Même si cette méthode requiert un temps d'exécution plus long, du fait des chaines de Markov, son application à des contextes où les *a priori* sont forts (suivi d'objet avec des *a priori* sur la dynamique) ou avec un bruit élevé sur les données (imagerie radar *SAR*) pourra apporter améliorations dans ces cas.

Remerciements

Les auteurs sont reconnaissants envers l'infrastructure OPAL de l'Université Côte d'Azur (UCA) pour avoir fourni les ressources de calcul nécessaires à ce travail de recherche, ainsi qu'envers BPI France pour le soutien financier dans le cadre du contrat LiChiE.

Références

[1] P.CRACIUN, M.ORTNER et J.ZERUBIA : Joint detection and tracking of moving objects using spatio-temporal marked point processes. *In IEEE Winter Conf. on Applications of Computer Vision*, 2015.

[2] X.DESCOMBES : Multiple objects detection in biological images using a marked point process framework. *Methods*, 2017.

[3] Y.DU et I.MORDATCH : Implicit Generation and Modeling with Energy Based Models. *NeurIPS*, 2019.

[4] D.DUVENAUD, J.WANG, J.JACOBSEN, K.SWERSKY, M.NOROUZI et W.GRATHWOHL : Your classifier is secretly an energy based model and you should treat it like one. *In ICLR*, 2020.

[5] P. J.GREEN : Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 1995.

[6] U.GRENANDER et M. I.MILLER : Representations of knowledge in complex systems. *Journal of the Royal Statistical Society : Series B (Methodological)*, 1994.

[7] G. E.HINTON : Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

[8] Z.HUANG, W.LI, X.-G.XIA et R.TAO : A general Gaussian heatmap label assignment for arbitrary-oriented object detection. *IEEE TIP*, 2022.

[9] C.LACOSTE, X.DESCOMBES et J.ZERUBIA : Point processes for unsupervised line network extraction in remote sensing. *IEEE TPAMI*, 2005.

[10] F.LAFARGE, G.GIMEL'FARB et X.DESCOMBES : Geometric feature extraction by a multimarked point process. *IEEE TPAMI*, 2010.

[11] Y.LECUN, S.CHOPRA, R.HADSELL, M.RANZATO et F. J.HUANG : A tutorial on energy-based learning. *Predicting structured data*, 2006.

[12] K.LI, G.WAN, G.CHENG, L.MENG et J.HAN : Object detection in optical remote sensing images : A survey and a new benchmark. *ISPRS*, 2020.

[13] Y.LI, Y.XING, Z.WANG, T.XIAO, Q.SONG, W.LI et J.WANG : A framework of maximum feature exploration oriented remote sensing object detection. *IEEE GRSL*, 2023.

[14] M.-C. V.LIESHOUT : *Markov Point Processes and Their Applications*. Imperial College Press, 2000.

[15] J.MABON, M.ORTNER et J.ZERUBIA : Processus ponctuels marqués et réseaux de neurones convolutifs pour la détection d'objets dans des images de télédétection. *In GRETSI*, 2022.

[16] M.ORTNER, X.DESCOMBES et J.ZERUBIA : A marked point process of rectangles and segments for automatic analysis of digital elevation models. *IEEE TPAMI*, 2008.

[17] O.RONNEBERGER, P.FISCHER et T.BROX : U-Net : Convolutional networks for biomedical image segmentation. *In MICCAI*, 2015.

[18] Y.VERDIÉ et F.LAFARGE : Efficient Monte Carlo sampler for detecting parametric objects in large scenes. *In ECCV*, 2012.

[19] G.-S.XIA, X.BAI, J.DING, Z.ZHU, S.BELONGIE, J.LUO, M.DATCU, M.PELILLO et L.ZHANG : DOTA : A large-scale dataset for object detection in aerial images. *In IEEE CVPR*, 2018.

[20] J.YI, P.WU, B.LIU, Q.HUANG, H.QU et D.METAXAS : Oriented object detection in aerial images with box boundary-aware vectors. *In IEEE Winter Conf. on Applications of Computer Vision*, 2021.