Détection de petites cibles par apprentissage profond et critère *a contrario*

Alina CIOCARLAN[1], Sylvie LE HEGARAT-MASCLE[2], Sidonie LEFEBVRE[1], Clara BARBANSON[3]

[1]DOTA, ONERA, Université Paris-Saclay, F-91123 Palaiseau, France

[2]SATIE, Université Paris-Saclay, 91405 Orsay, France

[3]Safran Electronics & Defense, F-91344 Massy, France

alina.ciocarlan@onera.fr, sylvie.le-hegarat@universite-paris-saclay.fr
sidonie.lefebvre@onera.fr, clara.barbanson@safrangroup.com

Résumé – La détection de petites cibles est une problématique délicate mais essentielle dans le domaine de la défense, notamment lorsqu'il s'agit de différencier ces cibles d'un fond bruité ou texturé, ou lorsqu'elles sont de faible contraste. Pour mieux prendre en compte les informations contextuelles, nous proposons d'explorer différentes approches de segmentation par apprentissage profond, dont certaines basées sur les mécanismes d'attention. Nous proposons également d'inclure un module d'attention par canal au TransUnet, réseau à l'état de l'art, ce qui permet d'améliorer significativement les performances. Par ailleurs, le manque de données annotées induit une perte en précision lors des détections, conduisant à de nombreuses fausses alarmes non pertinentes. Nous explorons donc des méthodes *a contrario* afin de sélectionner les cibles les plus significatives détectées par un réseau entraîné avec peu de données.

Abstract – Small target detection is an essential yet challenging task in defense applications, since differentiating low-contrast targets from natural textured and noisy environment remains difficult. To better take into account the contextual information, we propose to explore deep learning approaches based on attention mechanisms. Specifically, we propose a customized version of TransUnet including channel attention, which has shown a significant improvement in performance. Moreover, the lack of annotated data induces weak detection precision, leading to many false alarms. We thus explore *a contrario* methods in order to select meaningful potential targets detected by a weak deep learning training.

1 Introduction

La détection de petites cibles est un grand défi en vision par ordinateur, principalement du fait de la petite taille des cibles et de leur environnement bruité qui peut conduire à de nombreuses fausses alarmes. Quelques méthodes d'apprentissage profond ont été étudiées dans des travaux antérieurs : elles sont basées sur des réseaux de neurones convolutifs (CNN) [1] et incluent parfois des mécanismes d'attention [2]. L'un des avantages de ces derniers est qu'ils modélisent mieux les dépendances à grande échelle comparés aux CNNs. Cette propriété est un atout pour la détection de cibles, celles-ci ne présentant pas de structure spécifique. Partant de cette observation, [3] utilise une version améliorée de U-Net qui inclut un encodeur Transformer (ViT) en plus de l'encodeur convolutif classique, ce qui conduit à des résultats très compétitifs.

Une autre difficulté de cette application est le manque de données annotées pour entrainer le détecteur, ce qui résulte en une détection comportant beaucoup de fausses alarmes. En effet, le réseau de neurones n'a pas suffisamment d'exemples pour apprendre à extraire les bonnes caractéristiques. Pour autant, en observant la carte des scores donnée en sortie du réseau, les cibles y apparaissent comme étant noyées dans du bruit. Ce bruit, bien que moins significatif perceptuellement, est détecté comme cible du fait de sa forte valeur pixellique. Cela est dû au seuil fixe appliqué en sortie du détecteur pour effectuer la détection, qui ne permet pas de prendre en compte certains critères de perception comme la forme ou la densité induite par les niveaux de gris (cf. Figure 1 colonne 2). Pour pallier cela, nous proposons d'explorer l'intérêt d'un critère *a contrario* appliqué sur la carte des scores obtenue en sortie du détecteur afin de sélectionner les détections les plus significatives au sens du Nombre de Fausses Alarmes (NFA, défini dans le paragraphe 2). La méthode proposée permet de considérer aussi bien des caractéristiques de niveaux de gris que des éléments de structuration spatiale tels que la densité ou la forme des nuages de points représentant des cibles potentielles.

Après avoir présenté les concepts clés ainsi que l'état de l'art associé, nous décrirons une méthode de filtrage *a contrario* de la carte des scores obtenue en sortie de réseaux neuronaux, dont nous analyserons l'intérêt en dernière partie.

2 Définitions et travaux connexes

Méthodes *a contrario* Les méthodes de détection *a contrario* s'inspirent des théories de la perception, en particulier celle de Gestalt [4]. Elles reposent sur le principe d'Helmoltz qui stipule qu'une grande déviation d'un modèle aléatoire est probablement dûe à la présence d'une structure. Les méthodes *a*

contrario consistent à rejeter un modèle naïf caractérisant un fond destructuré en choisissant un seuil de détection. Ce seuil est choisi de sorte à contrôler le Nombre de Fausses Alarmes (NFA), souvent défini en chaque *objet* testé x_i par :

$$NFA(x_i) = N_{test} \times P(|X_i| \geq |x_i|), \tag{1}$$

où N_{test} représente le nombre de tests, $(X_i)_{i\in\mathbb{N}}$ une suite de variables aléatoires suivant la loi du modèle naïf et P la probabilité associée. Un évènement sera alors considéré comme étant ϵ-significatif si son NFA est inférieur au seuil de détection ϵ.

Parmi les travaux précédents utilisant un critère NFA pour de la détection, certains travaux ont défini le modèle naïf en termes de distribution des niveaux de gris et la détection est alors réalisée en chaque pixel [5], tandis que d'autres ont défini un modèle naïf en termes de distribution des pixels de valeur 1 dans une image binaire (c'est le cas de l'article [6]).

Mécanismes d'attention Malgré l'efficacité des CNNs pour extraire des informations significatives d'une image, l'invariance de translation induite par les convolutions semble nuire à la compréhension globale de la scène. Selon [7], cela induit un biais de texture élevé dans le processus de décision, tandis que les mécanismes d'attention semblent contourner cette limitation en imitant le cerveau et la perception humaine. Plusieurs types d'attention existent, conduisant à un large éventail de techniques discutées dans [8]. Les méthodes les plus récentes, telles que les ViT, reposent sur des attentions spatiales. Ces dernières améliorent significativement la modélisation des dépendances entre les différentes zones d'une image, conduisant à un biais de forme élevé [7]. Cependant, cela réduit la perception de petits objets, ce qui n'est pas souhaitable dans notre application. Dans cet article, nous avons décidé de nous appuyer sur des architectures hybrides CNN-Transformers à l'état de l'art, comme le TransUnet [9], afin de mieux prendre en compte l'information contextuelle tout en préservant la perception des petits objets. Parmi les autres mécanismes d'attention on peut citer l'attention par canal, qui permet de sélectionner les canaux pertinents en vue d'une réduction de la dimension. Il s'agit d'un des modules d'attention implémentés dans MA-Net [10] (bloc d'attention de fusion multi-échelle (MFA)), en plus d'un module d'attention spatiale.

3 Méthodologie explorée

3.1 Détecteurs de petites cibles

Sur la base de l'état de l'art actuel, nous évaluons plusieurs approches de segmentation pour la détection de petites cibles, inspirées de l'architecture U-Net. Ces détecteurs utilisent un encodeur ResNet-18 avec en entrée une image en niveaux de gris de taille 256×256 pixels. Les détecteurs sélectionnés sont :

— Un U-Net, qui représente notre algorithme de référence.
— Un MA-Net, dont l'architecture détaillée peut être consultée dans les Figures 1, 3 et 4 de [10].
— Un TransUnet, dont un schéma détaillé est présenté sur la Figure 1 de [9]. Il s'agit d'une méthode à l'état de l'art pour la détection des cibles [3].
— Un TransUnet modifié, appelé TransUnet-MFA, qui inclut le bloc MFA proposé dans le modèle MA-Net afin d'aider à la sélection des canaux pertinents lors de la concaténation au niveau du décodeur.

3.2 Formulation du NFA

Pour le critère NFA, nous nous sommes inspirés de [11]. L'idée principale est de considérer simultanément des caractéristiques de niveaux de gris et de structuration spatiale (densité de points). Nous considérons qu'un ensemble de pixels susceptibles de représenter une cible est d'autant plus significatif qu'il contient beaucoup de points spatialement proches et de valeur élevée sur la carte des scores. Le modèle naïf est alors la distribution de Bernoulli de paramètre p représentant la présence d'un pixel à une position donnée dans un espace 3D, $\mathcal{E} \subset \mathbb{R}^3$, discret et borné, d'axes représentant les coordonnées spatiales et les valeurs des scores (troisième axe) transformées (cf. paragraphe suivant). La probabilité d'observer au moins κ pixels dans un pavé de volume ν est alors la loi Binomiale de paramètre p et le NFA s'écrit :

$$NFA_B\left(\kappa,\nu,p\right) = \eta \sum_{i=\kappa}^{\nu} \binom{\nu}{i} p^i \left(1-p\right)^{\nu-i}, \tag{2}$$

où η est le nombre de tests, ici pris égal au nombre de pavés de même taille que celui considéré dans l'espace 3D. En termes de mise en œuvre pratique, trois éléments sont à préciser.

Premièrement, les valeurs des scores sont transformées de façon à ce que les faibles valeurs soient étalées sur une forte dynamique et les scores élevés concentrés sur une faible dynamique. Dans les résultats présentés nous avons utilisé la fonction inverse : $\forall x > \tau, f(x) = \frac{1}{x-\tau}; \forall x \leq \tau, f(x) = +\infty$. En pratique le paramètre τ permet de ne pas considérer des scores trop faibles et de réduire la complexité algorithmique.

Deuxièmement, plutôt que de calculer le NFA de l'Eq. 2 qui est coûteux numériquement, nous utilisons la significativité définie par $S\left(\kappa,\nu,p\right) = -\ln(NFA_B(\kappa,\nu,p))$ et l'approximation de Hoeffding : si $\frac{\kappa}{\nu} > p$,

$$S\left(\kappa,\nu,p\right) \approx \nu\left[\frac{\kappa}{\nu}\ln\left(\frac{\frac{\kappa}{\nu}}{p}\right) + \left(1-\frac{\kappa}{\nu}\right)\ln\left(\frac{1-\frac{\kappa}{\nu}}{1-p}\right)\right] - \ln\eta. \tag{3}$$

Troisièmement, pour calculer les nombres de points de chaque pavé de $\mathcal{E}$, nous utilisons l'histogramme intégral comme dans [12]. Finalement, l'algorithme 1 résume les principales étapes du calcul du NFA avec une astuce consistant à ne calculer la significativité que pour les pavés a priori les plus significatifs (ceux de volume minimal à nombre de points inclus donné).

4 Résultats et discussions

En raison de la nécessité de disposer de capacités opérationnelles de jour comme de nuit, nos détecteurs de cibles ont été entraînés sur des images infrarouge annotées provenant du

Algorithm 1 Détection de cibles de taille maximale M pixels sur la carte des scores I_s ; en entrée : I_s, M, significativité minimale S_{min} ; en sortie : liste des cibles $\mathbf{C}$.

pour pour chaque pixel j de I_s **faire**
  $I_s(j) = f(I_s(j))$
fin pour
$\mathcal{P} \leftarrow$ nuage de points 3D issus des pixels $j / I_s(j) < +\infty$
$p \leftarrow \frac{|\mathcal{P}|}{\text{volume 3D de } \mathcal{P}}$
Initialiser un tableau Tab de dimension M à $+\infty$
Initialiser un tableau Idx de dimension M à -1
pour chaque pavé 3D $\mathcal{C}$, de projection 2D (x,y) d'aire inférieure à M **faire**
  $\kappa \leftarrow$ nombre de pixels dans $\mathcal{C}$; $\nu \leftarrow$ volume de $\mathcal{C}$
  si $Tab[\kappa] > \nu$ **alors**
    $Tab[\kappa] \leftarrow \nu$; $Idx[\kappa] \leftarrow$ indice de $\mathcal{C}$
  fin si
fin pour
pour $\kappa \in [\![1, T]\!]$ **faire**
  $\nu \leftarrow Tab[\kappa]$; $p_{\mathcal{C}} = \frac{\kappa}{\nu}$
  si $\nu < +\infty$ et $p_{\mathcal{C}} > p$ **alors**
    calculer $S(\mathcal{C})$ à partir de l'Eq. 3
  fin si
fin pour
$\mathcal{I} \leftarrow$ liste des indices des pavés triés selon $S(\mathcal{C})$
$S_{max} \leftarrow$ significativité du premier élement de $\mathcal{I}$
$\mathbf{C} \leftarrow \emptyset$
pour chaque indice i dans $\mathcal{I}$ **faire**
  $\mathcal{C}_i \leftarrow i^{th}$ pavé selon $\mathcal{I}$
  si $S(\mathcal{C}_i) > S_{min}$ et $S(\mathcal{C}_i) > 0.8 \times S_{max}$ **alors**
    $\mathbf{C} \leftarrow \mathbf{C} \cup \{\mathcal{C}_i\}$
  fin si
fin pour

jeu de données d'entraînement MFIRST [13]. Deux entrainements ont été effectués : l'un en utilisant 2500 images annotées, l'autre en se plaçant dans un contexte frugal et en n'utilisant que 100 images, choisies aléatoirement. Les détecteurs sont évalués sur l'ensemble de test MFIRST (100 images) en termes de précision, de rappel et de score F1 calculés au niveau objet. Les résultats sont donnés en pourcentage sous la forme $\mu \pm \sigma$, avec μ la moyenne et σ l'écart-type calculés sur cinq entrainements initialisés aléatoirement.

Performance des détecteurs Le Tableau 1 donne les métriques obtenues lors de l'évaluation des détecteurs de cibles entrainés sur un nombre suffisant d'images (2500). Il y a un écart notable en précision lorsque l'on compare le U-Net d'origine aux versions modifiées incluant des mécanismes d'attention. Ces dernières conduisent également à de meilleures performances globales, en particulier pour TransUnet-MFA qui obtient un score F1 moyen de 88.0%. On remarque également que les mécanismes d'attention, en particulier ceux spatiaux, réduisent fortement la variabilité des résultats. Le Tableau 2a donne les résultats lorsque l'on se place dans un contexte frugal pour U-Net et MA-Net (100 images d'entrainement). Dans ce cas, on remarque que la précision est très faible pour le

Méthodes	**Prec.** $\pm\sigma$**(%)**	**Rap.** $\pm\sigma$**(%)**	**F1**$\pm\sigma$ **(%)**
U-Net	50.1 ± 9.0	84.8 ± 2.9	62.5 ± 7.1
MA-Net	62.1 ± 9.4	$\mathbf{88.3 \pm 1.2}$	72.6 ± 6.2
TransUnet	$\mathbf{90.2 \pm 2.7}$	81.9 ± 3.8	85.8 ± 1.8
TransUnet-MFA	89.5 ± 3.1	86.7 ± 2.1	$\mathbf{88.0 \pm 0.6}$

TABLE 1 – Performance des détecteurs entrainés sur 2500 images, métriques calculées au niveau objet.

Méthodes		**Prec.** $\pm\sigma$ **(%)**	**Rap.** $\pm\sigma$**(%)**	**F1** $\pm\sigma$**(%)**
U-Net	S	20.7 ± 13.9	$\mathbf{53.6 \pm 33.5}$	29.7 ± 19.2
	S+F	35.6±28.3	43.4±39.2	35.4±33.2
	NFA	**76.9±14.3**	37.7±26.1	**44.9±23.5**
MA-Net	S	45.7 ± 11.6	**80.3 ±2.5**	57.4 ± 9.5
	S+F	47.0±12.0	79.6±1.4	58.3±9.7
	NFA	**74.6±4.8**	68.1±3.1	**71.2±2.8**

(a) Entrainement sur 100 images.

Méthodes		**Prec.** $\pm\sigma$**(%)**	**Rap.**$\pm\sigma$ **(%)**	**F1** $\pm\sigma$**(%)**
U-Net	S	50.1 ± 9.0	$\mathbf{84.8 \pm 2.9}$	62.5 ± 7.1
	S+F	79.2±7.8	83.6±2.5	81.1±3.7
	NFA	**84.9±6.6**	81.9±4.2	**83.1±2.0**
TransUnet-MFA	S	89.5 ± 3.1	86.7 ± 2.1	88.0 ± 0.6
	S+F	90.7±2.7	86.4±1.9	**88.5±0.7**
	NFA	**93.4±1.8**	78.3±3.0	85.1±2.1

(b) Entrainement sur 2500 images.

TABLE 2 – Performance des détecteurs entrainés sur 100 ou 2500 images. Les prédictions sont données soit par seuillage fixe (S), par filtrage du seuillage fixe (S+F) ou par calcul du NFA. Métriques calculées au niveau objet.

cas où l'on applique un seuil fixe (S) aux cartes de score en sortie des réseaux de neurones pour détecter les cibles, ce qui conduit à un taux de fausses alarmes élevé. La variabilité des résultats est également très élevée notamment du fait d'un cas de non convergence parmi les cinq expériences pour le U-Net. La Figure 1 permet de visualiser quelques résultats pour trois images. La première colonne montre l'image d'origine, la suivante la carte de sortie des réseaux, la troisième le résultat de prédiction en appliquant un seuil fixe et la dernière en y appliquant un critère NFA. On y voit les nombreux faux positifs induits par le bruit de fond présent dans la carte des scores, bien qu'ils soient moins pertinents perceptuellement que les cibles à détecter.

Contribution du NFA Pour pallier ce problème, nous appliquons un calcul de NFA sur les cartes de score pour effectuer les détections, et nous comparons les résultats avec ceux obtenus par un filtrage morphologique (ouverture) de la carte de détection donnée après seuillage par un seuil fixe (S+F). Nous avons testé les detecteurs U-Net et MA-Net dans un contexte frugal du fait de leur très mauvaise précision dans ce contexte. Le Tableau 2 donne les résultats obtenus pour les différentes méthodes testées. Nous avons aussi évalué deux détecteurs entraînés sur 2500 images et ayant une meilleure précision.

Nous pouvons observer que, dans un contexte frugal, le NFA

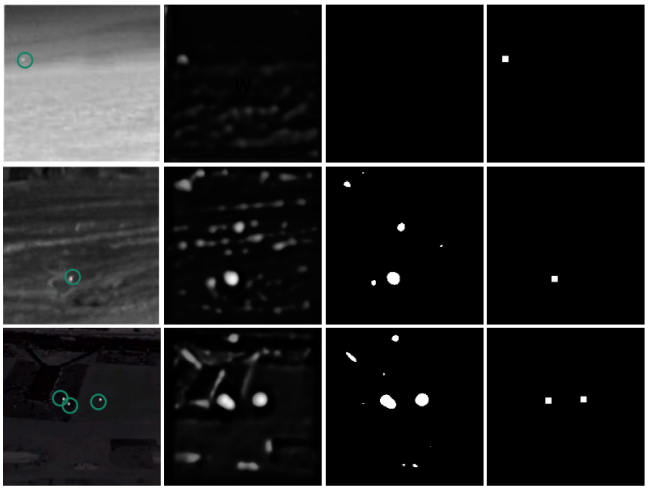

FIGURE 1 – Exemples de prédictions des détecteurs sur 3 images (lignes). De gauche à droite : image d'origine avec la vérité terrain encerclée, carte des scores en sortie du réseau, prédiction après seuillage avec seuil fixe, prédiction après NFA.

améliore fortement la précision, ce qui réduit le taux de fausses alarmes. Cela donne lieu à de meilleures performances globales comparé au seuillage fixe (S) ou à son filtrage (S+F). La Figure 1 rend compte de ces améliorations.

En revanche, lorsque le détecteur est suffisamment performant pour donner en sortie des cartes de scores avec peu de bruit, l'apport du NFA est plus limité par rapport à un filtrage avec un seuil fixe : d'après le Tableau 2b la précision est augmentée de manière significative, au détriment du nombre de bonnes détections. En effet, par définition le NFA controlant le nombre de fausses alarmes va optimiser la précision des résultats (ce qui est bien observé sur le Tableau 2, tant au niveau de la précision moyenne que de sa variabilité) et non le F1 score. Par ailleurs, cette limitation provient également d'une extraction d'information réalisée par le détecteur de manière erronée. La cible ne ressortant pas sur la carte des scores, le calcul de NFA ne pourra pas réparer l'erreur commise en amont. Ainsi, le NFA aura plus d'intérêt pour les sorties de réseaux où la cible se trouve noyée dans un bruit, ce qui est le cas pour les réseaux faiblement entrainés. Dans des travaux futurs, nous envisageons d'intégrer un critère *a contrario* à un réseau de neurones dès l'apprentissage afin que les avantages du NFA puissent être bénéfiques à un plus grand nombre de scénarios.

5 Conclusion

Nous avons exploré plusieurs approches par apprentissage profond pour la détection de petites cibles, basées notamment sur les mécanismes d'attention. Celles-ci donnent des résultats globaux très compétitifs, mais manquent de précision dans un contexte frugal. Pour contrer ce problème, nous avons remplacé le seuillage fixe effectué sur les cartes de scores par un critère *a contrario*. Les conclusions de ces travaux préliminaires constituent une motivation pour explorer l'intégration d'un critère NFA dans la fonction de coût des réseaux.

Références

[1] A. d'Acremont, R. Fablet, A. Baussard, and G. Quin, "CNN-Based Target Recognition and Identification for Infrared Imaging in Defense Systems," Sensors, vol. 19, p. 2040, Apr. 2019.

[2] F. Chen, C. Gao, F. Liu, Y. Zhao, Y. Zhou, D. Meng, and W. Zuo, "Local patch network with global attention for infrared small target detection," IEEE Transactions on Aerospace and Electronic Systems, 2022.

[3] F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, "Infrared small-dim target detection with transformer under complex backgrounds," arXiv preprint arXiv :2109.14379, 2021.

[4] A. Desolneux, L. Moisan, and J.-M. Morel, From gestalt theory to image analysis : a probabilistic approach, vol. 34. Springer Science & Business Media, 2007.

[5] T. Ehret, A. Davy, M. Delbracio, and J.-M. Morel, "How to reduce anomaly detection in images to anomaly detection in noise," Image Processing On Line, vol. 9, pp. 391–412, 2019.

[6] A. Desolneux, L. Moisan, and J.-M. Morel, "A grouping principle and four applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 4, pp. 508–513, 2003.

[7] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are Convolutional Neural Networks or Transformers more like human vision ?," arXiv :2105.07197, July 2021.

[8] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention Mechanisms in Computer Vision : A Survey," arXiv :2111.07624, Nov. 2021.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet : Transformers Make Strong Encoders for Medical Image Segmentation," arXiv :2102.04306, Feb. 2021.

[10] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net : A Multi-Scale Attention Network for Liver and Tumor Segmentation," IEEE Access, vol. 8, pp. 179656–179665, 2020.

[11] A. Rezaei, S. L. Hégarat-Mascle, E. Aldea, P. Dondi, and M. Malagodi, "A-contrario framework for detection of alterations in varnished surfaces," J. Vis. Commun. Image Represent., vol. 83, p. 103357, 2022.

[12] S. L. Hégarat-Mascle, E. Aldea, and J. Vandoni, "Efficient evaluation of the number of false alarm criterion," EURASIP J. Image Video Process., vol. 2019, p. 35, 2019.

[13] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm : Adversarial learning for small object segmentation in infrared images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8509–8518, 2019.