Utilisation d'un STN-Rotation pour la segmentation avec Florence-2

Jerome CARTIER[1] Alexis LECHERVY[1] Abderrahim ELMOATAZ[1]

[1]GREYC, 6 Boulevard Maréchal Juin, 14000 Caen, France

Résumé – La segmentation d'images est une tâche largement étudiée en vision par ordinateur, avec un vaste champ d'application. Récemment, des modèles de langage tels que Florence 2 ont été proposés pour générer des séquences de points permettant de délimiter des régions à segmenter. Bien que cette approche soit prometteuse, ces modèles sont très sensibles à l'ordre des points utilisé dans la base d'apprentissage. Dans cet article, nous proposons de résoudre ce problème en utilisant des *Spatial Transformer Networks* (*STN*) pour réaliser un alignement commun des images. Nous démontrons la pertinence de cette approche à travers des expérimentations sur deux problèmes de segmentation très différents.

Abstract – Image segmentation is a widely studied task in computer vision with a wide range of applications. Recently, language models such as Florence 2 have been proposed to generate sequences of points to delimit the regions to be segmented. Although this approach seems interesting, these models are very sensitive to the order of the points used in the learning base. In this paper, we propose to take this problem into account by using *Spatial Transform Networks* (STNs) to achieve a common alignment of the images. We demonstrate the relevance of this approach through experiments on two very different segmentation problem.

1 Introduction

Les modèles de fondation récents démontrent une capacité remarquable à s'adapter à diverses tâches, souvent surpassant les performances des modèles spécialisés. Parmi eux, Florence-2 [18] propose d'utiliser un modèle de texte unifié pour réaliser diverses tâches multimodales texte/image, telles que la description d'images, la détection de régions d'intérêt ou la reconnaissance optique de caractères (OCR). L'article propose notamment d'aborder la segmentation sémantique d'objet unique comme un problème textuel, en prédisant une succession de *tokens de coordonnées* décrivant les contours des régions à segmenter.

Lorsqu'un contour est décrit par une succession de points, le point de départ de la séquence n'a pas d'importance intrinsèque. Cependant, dans le cadre d'un modèle de langage, la séquence des points <A><B><C> n'est pas équivalente à la séquence <B><C><A> pour la métrique d'apprentissage, bien qu'elles décrivent la même région triangulaire. Le premier *token* de la séquence est donc crucial. Pour résoudre ce problème, nous proposons l'utilisation d'un *Spatial Transformer Network* (*STN*) [6] spécialisé dans les rotations. Ce module, placé avant l'utilisation de Florence-2, permet d'aligner les images selon une convention apprise, facilitant ainsi la génération de la séquence de points.

Après une présentation de l'état de l'art, nous détaillerons notre méthode, qui repose sur l'utilisation d'un *STN*-rotation suivi du modèle Florence-2 pour la segmentation. Nous démontrerons l'intérêt de cette approche sur deux jeux de données de segmentation très différents.

2 État de l'art

La segmentation d'images a été largement étudiée dans la littérature, notamment à travers des approches basées sur les réseaux de neurones. La plupart des méthodes prédisent directement des cartes de segmentation. Certaines architectures

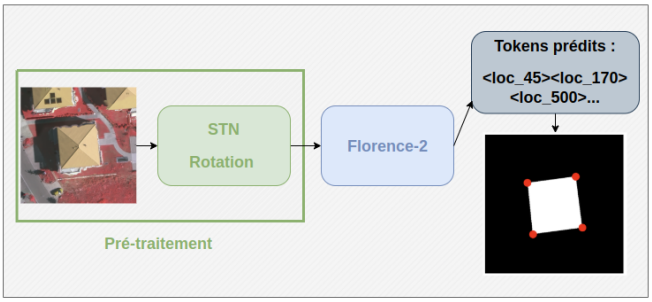

FIGURE 1 : La méthode proposée consiste à aligner les images à l'aide d'un *STN-rotation*, puis à prédire une séquence de points formant une segmentation avec Florence-2.

reposent sur des modèles encodeurs/décodeurs convolutifs, tels que U-Net [16] et ses variantes comme 2D nnU-Net [13] ou Super-Seg [7]. D'autres procèdent par affinement itératif à l'aide de modèles de diffusion, comme RNDiff [9]. Le modèle de fondation SAM [8] utilise quand à lui des prompts visuels pour guider la segmentation.

L'utilisation de la prédiction de séquences de points pour la segmentation a été explorée à travers des modèles génératifs de type *transformers*, tels que Florence-2 [18], ou encore par des techniques d'évolution de points de contours prédéfinis par minimisation d'énergie, comme dans DSAC [15].

3 Méthode proposée

3.1 Approche générale

Pour améliorer l'alignement des images et la cohérence des prédictions de Florence-2, nous proposons d'utiliser un *Spatial Transformer Network* (*STN*) spécialisé dans les rotations. Ce module calcule automatiquement l'angle de rotation optimal θ, permettant d'orienter l'image selon une représentation favorable pour Florence-2. Le *STN* génère une image transformée en appliquant une rotation d'angle θ, déterminée par l'architecture. Cette image est ensuite utilisée comme entrée pour

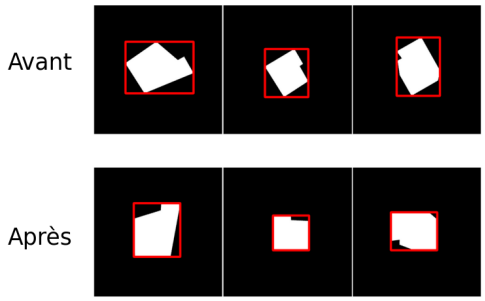

FIGURE 2 : Vérité terrain classique et modifiée pour l'apprentissage du *STN-rotation*.

Florence-2.

Notre méthode (cf. Fig. 1) se décompose en deux étapes distinctes et indépendantes :

1. **Phase d'apprentissage du *STN*** : Nous entraînons le *STN* de manière isolée sur une tâche d'alignement des images.

2. **Phase d'entraînement de Florence-2** : Le *STN* pré-entraîné est intégré dans le flux de traitement. Il pré-conditionne les images et leurs annotations avant leur passage dans Florence-2, qui peut alors être ré-entraîné.

Les sections suivantes détaillent les procédures d'optimisation spécifiques à chaque composant.

3.2 Spatial Transformer Networks (*STN*)

Les *Spatial Transformer Networks* (*STN*) sont des couches de réseaux de neurones placées au début d'une architecture pour réduire l'impact des transformations affines sur les performances. Proposés par [6], ils sont principalement utilisés dans des problèmes de classification, bien que quelques usages en segmentation aient été notés, comme dans [1].

Un *STN*, noté $\mathcal{T}$, prend en entrée une image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ (où W et H sont la largeur et la hauteur de l'image, et $C = 3$ le nombre de canaux couleurs) et prédit les paramètres d'une transformation spatiale à appliquer. Dans cet article, nous nous limitons aux rotations.

Ainsi, le *STN* prédit la matrice de rotation R_θ, définie par :

$$R_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{1}$$

En utilisant cette matrice, il est possible d'appliquer une rotation d'angle θ à l'image pour produire une nouvelle image $\mathcal{T}_\theta(I)$. Cela se fait en appliquant la rotation aux points de la grille des pixels et en interpolant les valeurs des nouveaux pixels pour les coordonnées non entières.

Pour entraîner notre modèle *STN-rotation*, nous générons des éléments cibles en effectuant la rotation qui permet d'obtenir la boîte englobante la plus compacte contenant notre objet (cf. Fig. 2), cette rotation nous donne ainsi l'angle cible $\hat{\theta}$ de l'apprentissage.

L'apprentissage du *STN* est réalisé en minimisant une fonction de coût composée de la perte *Mean Squared Error* entre les images prédites $\mathcal{T}_\theta(I)$ et les images cibles $\hat{I}$, ainsi que d'une distance L2 entre l'angle de rotation prédit θ et l'angle cible $\hat{\theta}$:

$$\mathcal{L} = \frac{1}{WH} \left\| \mathcal{T}_\theta(I) - \hat{I} \right\|^2 + \frac{1}{2} \left\| \theta - \hat{\theta} \right\|^2. \tag{2}$$

3.3 Florence-2

Florence-2 [18] est un modèle de fondation utilisant des *prompts*, capable d'effectuer diverses tâches de vision telles que la description d'images, la détection de régions d'intérêt, et la segmentation. Le modèle utilise une séquence de tokens visuels et textuels issus d'un *Dual Attention Vision Transformer* (DaViT) pour l'image et d'un tokenizer similaire à celui de BART [11] pour le texte. Cette séquence est ensuite traitée par un encodeur-décodeur similaire à BART, qui la considère comme une unique séquence de tokens multimodaux. La force de ce modèle réside dans sa représentation unifiée, permettant de traiter différentes tâches comme si une seule modalité textuelle était utilisée.

Grâce à ses capacités multi-tâches, Florence-2 a été utilisé de diverses manières, notamment pour la description de jeux de données [5] et la détection d'objets [17]. Cependant, il est peu exploité comme outil de segmentation, bien que des travaux comme [18] et [12] l'aient utilisé à cette fin.

Pour utiliser Florence-2 pour la segmentation, il est nécessaire de produire une représentation textuelle des contours de la forme à segmenter. La procédure est la suivante :

1. **Acquisition des points de contour** : À partir des images cibles fournies dans les jeux de données, on extrait des points appartenant aux contours des objets.

2. **Encodage des coordonnées** : Chaque point de coordonnées (x, y) est transformé en deux tokens correspondant à la discrétisation de ces coordonnées. La transformation est la suivante :

$$x' = \left\lfloor \frac{x}{W} \times 1000 \right\rfloor, \quad y' = \left\lfloor \frac{y}{H} \times 1000 \right\rfloor. \tag{3}$$

3. **Formatage final** : Le point encodé prend la forme textuelle `<loc_x'><loc_y'>`, où `x'` et `y'` sont les valeurs calculées ci-dessus.

Nous utilisons le modèle Florence-2 pré-entraîné de ses auteurs, que nous adaptons à notre tâche et à nos données en utilisant *Low-Rank Adaptation* (LoRA) [4] et une fonction de perte de type *entropie croisée*.

Pour compenser l'utilisation du *STN* et être conforme à l'orientation de la vérité terrain attendue, nous appliquons en post-traitement la rotation inverse ${R_\theta}^{-1}$ à la prédiction de Florence-2 des points du contour de la zone segmentée.

4 Protocoles expérimentaux

4.1 Jeux de données

Pour évaluer l'efficacité de notre méthode, nous avons conduit des tests sur deux jeux de données de natures très différentes : *ISPRS Vaihingen* [2] et *CAMUS* [10].

ISPRS Vaihingen : Ce jeu de données comprend des images aériennes de scènes urbaines, présentant une grande variété de bâtiments en termes de taille et de forme. Il contient 168 images, chacune avec un objet cible à segmenter.

CAMUS : Il s'agit de l'un des plus grands jeux de données disponibles d'échocardiographies. Il comprend des images prises à différents moments du cycle cardiaque. Nous travaillons sur les 1000 images disponibles pour la fin de diastole (ED), qui correspond au moment de volume maximum du cœur. Nous utilisons ces données car elles représentent l'instant le plus couramment utilisé, et nous cherchons à segmenter la cavité supérieure appelée endocarde.

Les détails de la constitution des bases de données sont présentés dans le tableau 1.

TABLE 1 : Informations générales sur les jeux de données utilisés

Jeu de données	Nombre d'images			Type d'image
	Entraînement	Validation	Test	
Vaihingen	134	17	17	Vue aérienne
CAMUS	800	100	100	Échocardiographie

4.2 Détails d'implémentation

Pour évaluer notre méthode, nous avons comparé nos résultats à des modèles classiques de segmentation ainsi qu'à des modèles plus récents et spécialisés pour chaque jeu de données. Les méthodes basées sur Florence-2 ont été entraînées sur 10 époques pour le jeu de données ISPRS Vaihingen et sur 80 époques pour le jeu de données CAMUS, avec un taux d'apprentissage de 10^{-5} et un optimiseur AdamW. Nous utilisons l'implémentation de base du modèle Florence-2.

Nous avons appliqué l'augmentation de données sur le jeu de données Vaihingen en effectuant des rotations de $90°$, $180°$, et $270°$ pour les données d'apprentissage n'ayant pas observé de différence notable lors d'une plus large augmentation de données. Pour le jeu de données CAMUS, les données étant acquises selon un protocole médical bien défini, leurs orientations sont moins variables. Nous appliquons donc des rotations aléatoires restreintes à l'intervalle $[-20°, 20°]$.

5 Résultats

Nos résultats sont mesurés selon les métriques classiques pour ces deux bases de données, à savoir : *Intersection over Union* (IoU) et *Dice* pour ISPRS Vaihingen, et *Dice* et *Hausdorff Distance* (HD) pour CAMUS.

Les résultats obtenus par notre méthode (cf. Fig. 3) sur le jeu de données ISPRS Vaihingen (cf. Tab. 2) dépassent ceux de tous les autres modèles pour les deux métriques étudiées. Le gain de l'utilisation du *STN* par rapport à l'utilisation simple de Florence-2 est de 0,82 pour l'IoU et de 0,56 pour le Dice, ce qui montre l'intérêt de son ajout. On peut noter que Florence-2 aurait encore de meilleures performances si l'on utilisait la rotation optimale employée pour l'apprentissage du *STN*, comme le montre la dernière ligne du tableau. Cependant, en pratique, il n'est pas possible de réaliser cela car la rotation idéale est construite à partir de la vérité terrain (cf. Fig. 2).

Sur le jeu de données CAMUS (cf. Tab. 3), nous obtenons la deuxième plus haute valeur pour les métriques Dice et Hausdorff Distance (HD) avec notre *STN* combiné à Florence-2. Notre approche est donc pleinement compétitive avec les autres méthodes de l'état de l'art (cf. Fig. 4). Le gain apporté par le *STN* par rapport à l'utilisation simple de Florence-2 est de 0,27 pour le Dice et de 0,61 pour le HD, démontrant à nouveau son intérêt.

TABLE 2 : Resultats sur le jeu de données ISPRS Vaihingen

Method	**IoU ± std**	**Dice ± std**
U-net [16]	64.35 ± 4.26	76.94 ± 4.83
DSAC [15]	84	-
RNDiff [9]	92.50	-
DCAC [3]	92.9	92.8
SAM [8]	91.75 ± 0.21	95.59 ± 0.26
Florence-2 [18]	92.73 ± 0.24	96.04 ± 0.30
STN+Florence-2	**93.55 ± 0.43**	**96.60 ± 0.45**
Rotation optimale+Florence-2	**93.78 ± 0.41**	**96.83 ± 0.39**

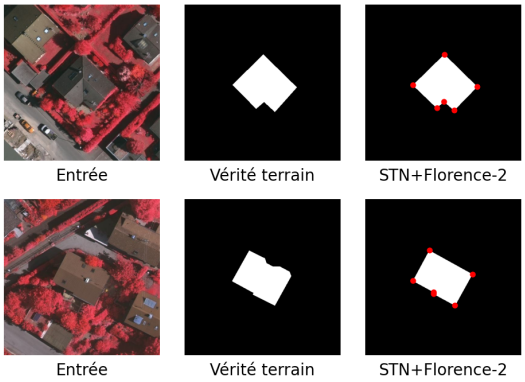

FIGURE 3 : Visualisation des résultats sur le jeu de données ISPRS Vaihingen

TABLE 3 : Resultats sur le jeu de données CAMUS

Method	**Dice ± std**	**HD ± std**
U-net [16]	92.19 ± 0.56	4.57 ± 0.74
EDMAE [14]	93.09 ± 0.22	**3.02 ± 0.81**
SAM [8]	93.52 ± 0.28	3.71 ± 0.52
2D nnU-Net [13]	93.92 ± 0.34	3.65 ± 0.44
Super-Seg [7]	**94.55 ± 0.26**	3.26 ± 0.46
Florence-2 [18]	94.15 ± 0.36	3.86 ± 0.49
STN+Florence-2	94.42 ± 0.38	3.25 ± 0.48
Rotation optimale+Florence-2	**94.66 ± 0.33**	**3.18 ± 0.45**

6 Conclusion

Nous avons présenté dans cet article l'utilisation d'un *Spatial Transformer Network* (*STN*) spécialisé dans les rotations afin de réduire l'impact du premier point dans la prédiction des points de contours pour la segmentation avec le modèle Florence-2. Sur l'ensemble des jeux de données utilisés, notre méthode produit des résultats compétitifs ou supérieurs à ceux des autres méthodes spécialisées. Grâce à l'ajout de notre *STN* de rotation, Florence-2 génère de meilleures segmentations en normalisant la convention des points, ce qui se traduit par un gain systématique de performance, démontrant ainsi l'intérêt de notre approche *STN-rotation* pour améliorer les performances de segmentation.

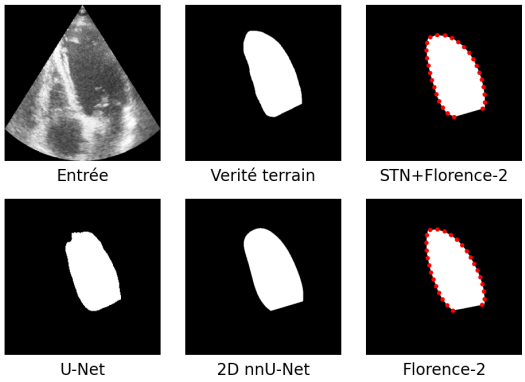

FIGURE 4 : Visualisation des résultats sur le jeu de données CAMUS

7 Remerciements

Ce travail est supporté par l'Agence Nationale de la Recherche (ANR) (ANR-23-CE19-0036) (project EXPERT). Ce travail a bénéficié des moyens de calcul du mésocentre CRIANN (Centre Régional Informatique et d'Applications Numériques de Normandie).

Références

[1] Giammarco La BARBERA, Pietro GORI, Haithem BOUSSAID, Bruno BELUCCI, Alessandro DELMONTE, Jeanne GOULIN, Sabine SARNACKI, Laurence ROUET et Isabelle BLOCH : Automatic size and pose homogenization with spatial transformer network to improve and accelerate pediatric segmentation, 2021.

[2] Michael CRAMER : The dgpf-test on digital airborne camera evaluation overview and test design. *Photogrammetrie - Fernerkundung - Geoinformation*, 2010(2):73–82, 05 2010.

[3] Ali HATAMIZADEH, Debleena SENGUPTA et Demetri TERZOPOULOS : End-to-end deep convolutional active contours for image segmentation, 2019.

[4] Edward J. HU, Yelong SHEN, Phillip WALLIS, Zeyuan ALLEN-ZHU, Yuanzhi LI, Shean WANG, Lu WANG et Weizhu CHEN : Lora : Low-rank adaptation of large language models, 2021.

[5] Hochul HWANG, Krisha ADHIKARI, Satya SHODHAKA et Donghyun KIM : Synthetic data augmentation for robotic mobility aids to support blind and low vision people, 2024.

[6] Max JADERBERG, Karen SIMONYAN, Andrew ZISSERMAN et Koray KAVUKCUOGLU : Spatial transformer networks, 2016.

[7] Abbas KHAN, Muhammad ASAD, Martin BENNING, Caroline RONEY et Gregory SLABAUGH : Compositional segmentation of cardiac images leveraging metadata, 2024.

[8] Alexander KIRILLOV, Eric MINTUN, Nikhila RAVI, Hanzi MAO, Chloe ROLLAND, Laura GUSTAFSON, Tete XIAO, Spencer WHITEHEAD, Alexander C. BERG, Wan-Yen LO, Piotr DOLLÁR et Ross GIRSHICK : Segment anything, 2023.

[9] Benedikt KOLBEINSSON et Krystian MIKOLAJCZYK : Multi-class segmentation from aerial views using recursive noise diffusion, 2024.

[10] Sarah LECLERC, Erik SMISTAD, João PEDROSA, Andreas ØSTVIK, Frédéric CERVENANSKY, Florian ESPINOSA, Torvald ESPELAND, Erik Andreas Rye BERG, Pierre-Marc JODOIN, Thomas GRENIER, Carole LARTIZIEN, Jan D'HOOGE, Lasse LØVSTAKKEN et Olivier BERNARD : Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38:2198–2210, 2019.

[11] Mike LEWIS, Yinhan LIU, Naman GOYAL, Marjan GHAZVININEJAD, Abdelrahman MOHAMED, Omer LEVY, Ves STOYANOV et Luke ZETTLEMOYER : Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[12] Junliang LI, Kai YE, Haolan KANG, Mingxuan LIANG, Yuhang WU, Zhenhua LIU, Huiping ZHUANG, Rui HUANG et Yongquan CHEN : Grasp what you want : Embodied dexterous grasping system driven by your voice, 2024.

[13] Hang Jung LING, Nathan PAINCHAUD, Pierre-Yves COURAND, Pierre-Marc JODOIN, Damien GARCIA et Olivier BERNARD : *Extraction of Volumetric Indices from Echocardiography : Which Deep Learning Solution for Clinical Use ?*, page 245–254. Springer Nature Switzerland, 2023.

[14] Yiman LIU, Xiaoxiang HAN, Tongtong LIANG, Bin DONG, Jiajun YUAN, Menghan HU, Qiaohong LIU, Jiangang CHEN, Qingli LI et Yuqi ZHANG : Edmae : An efficient decoupled masked autoencoder for standard view identification in pediatric echocardiography. *Biomedical Signal Processing and Control*, 86:105280, septembre 2023.

[15] Diego MARCOS, Devis TUIA, Benjamin KELLENBERGER, Lisa ZHANG, Min BAI, Renjie LIAO et Raquel URTASUN : Learning deep structured active contours end-to-end, 2018.

[16] Olaf RONNEBERGER, Philipp FISCHER et Thomas BROX : U-net : Convolutional networks for biomedical image segmentation, 2015.

[17] Qi WU, Zipeng FU, Xuxin CHENG, Xiaolong WANG et Chelsea FINN : Helpful doggybot : Open-world object fetching using legged robots and vision-language models, 2024.

[18] Bin XIAO, Haiping WU, Weijian XU, Xiyang DAI, Houdong HU, Yumao LU, Michael ZENG, Ce LIU et Lu YUAN : Florence-2 : Advancing a unified representation for a variety of vision tasks, 2023.