Analyse d'images médicales (2023)

Listes des contenus disponibles sur ScienceDirect

Analyse d'images médicales

page d'accueil du journal www.elsevier.com/locate/media

Identification de la source : Une tâche d'auto-supervision pour la prédiction dense

Shuai Chen[a,*] , Subhradeep Kayal[a,*] , Marleen de Bruijne[a,b]

[a] *Groupe d'imagerie biomédicale de Rotterdam, Département de radiologie et de médecine nucléaire, Erasmus MC, Rotterdam, Pays-Bas.*
[b] *Machine Learning Section, Department of Computer Science, University of Copenhagen, DK-2110 Copenhagen, Denmark.*

ARTICLEENFO

Historique de l'article :
Reçu le 6 juillet 2023
Reçu en version finale le 6 juillet 2023
Accepté le 6 juillet 2023
Disponible en ligne le 6 juillet 2023

Communiquée par le nom

2000 MSC : 92C55, 68U10

Mots-clés : Apprentissage autosupervisé, prédiction dense, segmentation d'images, séparation aveugle des sources, imagerie médicale.

ABSTRACT

Le paradigme de l'auto-supervision se concentre sur l'apprentissage de la représentation à partir de données brutes sans nécessiter d'annotations fastidieuses, ce qui constitue le principal goulot d'étranglement des méthodes actuelles basées sur les données. Les tâches d'auto-supervision sont souvent utilisées pour pré-entraîner un réseau neuronal avec une grande quantité de données non étiquetées et extraire des caractéristiques génériques de l'ensemble de données. Le modèle appris est susceptible de contenir des informations utiles qui peuvent être transférées à la tâche principale en aval et améliorer les performances par rapport à l'initialisation aléatoire des paramètres. Dans cet article, nous proposons une nouvelle tâche d'auto-supervision appelée identification de la source (SI), qui s'inspire du problème classique de la séparation aveugle des sources. Les images syn- thétiques sont générées par la fusion de plusieurs images sources et la tâche du réseau consiste à reconstruire les images originales à partir des images fusionnées. Une bonne compréhension du contenu de l'image est nécessaire pour résoudre la tâche avec succès. Nous validons notre méthode sur deux tâches de segmentation d'images médicales : la segmentation des tumeurs cérébrales et la segmentation des hyperintensités de la matière blanche. Les résultats montrent que la tâche SI proposée surpasse les tâches traditionnelles d'auto-supervision pour les prédictions denses, y compris l'inpainting, le pixel shuffling, l'intensity shift et la super-résolution. Parmi les variations de la tâche SI fusionnant des images de différents types, la fusion d'images de différents patients est la plus performante.

1. Introduction

Le succès de l'apprentissage profond, et en particulier des réseaux de neurones convolutifs (CNN), peut être partiellement attribué à l'augmentation ex ponentielle de la quantité de données annotées disponibles. Cependant, dans des domaines hautement spécialisés tels que la segmentation d'images médicales, il est beaucoup plus difficile d'obtenir des annotations précises et denses. L'auto-supervision est un axe de recherche qui au réseau d'apprendre à partir des images elles-mêmes sans nécessiter d'annotations fastidieuses, les caractéristiques apprises pouvant être utiles pour les tâches en aval, telles que la classification et la segmentation.

En général, l'apprentissage auto-supervisé fait référence à un ensemble d'approches qui retiennent délibérément des informations dans les données originales et chargent un réseau neuronal de prédire les informations manquantes. à partir des informations incomplètes existantes. Ce faisant, le réseau est encouragé à apprendre des caractéristiques polyvalentes qui se sont avérées bien adaptées aux tâches en aval (Jing et Tian, 2019). Le pipeline d'autosupervision utilise souvent une stratégie de pré-entraînement et de réglage fin. La première étape consiste à pré-entraîner un CNN sur un grand volume d'échantillons non annotés à l'aide d'tâche de substitution conçue manuellement, dans laquelle le CNN explore et apprend des caractéristiques génériques des données elles-mêmes. Les caractéristiques apprises peuvent contenir des informations significatives sur les données d'image, par exemple la distribution de l'intensité, la cohérence spatiale et les connaissances anatomiques en imagerie médicale, etc. La deuxième étape consiste à affiner ce réseau pré-entraîné sur la tâche cible (principale) en aval qui nous intéresse le plus, et qui comporte généralement un petit ensemble de données annotées dans la pratique. Nous pensons qu'en exploitant les données non annotées et en recommençant l'apprentissage à partir d'un ensemble de caractéristiques pré-entraînées riches, un modèle plus robuste sur la tâche principale peut être formé.

* : S. Chen et S. Kayal ont contribué à parts égales.

Dans cet article, nous proposons une nouvelle tâche d'auto-supervision appelée *identification de la source* (SI), qui s'inspire du problème classique de la séparation aveugle de la source (BSS). La tâche proposée capable de former un réseau de prédiction dense de manière auto-supervisée en utilisant des données non étiquetées.

Contributions :

1. Nous proposons une nouvelle tâche d'autosupervision, SI, dans laquelle un réseau neuronal est (pré)entraîné à identifier une image (source) à partir de mélanges d'images. De cette manière, le codeur et le décodeur sont tous deux entraînés et le réseau est encouragé à apprendre non seulement les caractéristiques locales, mais aussi les caractéristiques sémantiques globales pour identifier et séparer le signal source cible. À notre connaissance, il s'agit de la première méthode auto-supervisée de type BSS pour les réseaux neuronaux profonds.
2. Nous étudions le problème mal posé de l'identification des sources et montrons dans quels contextes il peut être résolu par un réseau neuronal. La méthode SI proposée fournit un moyen simple d'éviter l'ambiguïté de la tâche.
3. Nous menons des expériences approfondies sur des ensembles de données publiques pour deux applications de segmentation d'images médicales : la segmentation des tumeurs cérébrales et la segmentation des hyperintensités de la matière blanche, toutes deux issues de l'IRM cérébrale. Nous les comparons à diverses tâches d'auto-supervision existantes. Les résultats montrent que la méthode SI proposée surpasse les lignes de base de l'auto-supervision, y compris l'in- painting, le mélange de pixels, le décalage d'intensité et la super-résolution en termes de précision de segmentation dans les deux applications.

2. Travaux connexes

2.1. Tâches d'autosupervision

L'auto-supervision est un axe de recherche actif en apprentissage automatique, qui s'étend de la vision par ordinateur au traitement du langage naturel (Mikolov et al., 2013 ; Kiros et al., 2015 ; Devlin et al., 2018). En imagerie, les premières tâches d'autosupervision peuvent être regroupées en deux catégories principales : celles basées sur la reconstruction et celles basées sur la prédiction du contexte. Par exemple, l'inpainting est une tâche d'autosupervision populaire basée sur la reconstruction (Pathak et al., 2016) où des zones d'une image sont cachées puis reconstruites à l'aide d'un CNN. De manière similaire, la recolorisation peut être effectuée en supprimant la couleur de l'image et en entraînant un CNN à la récupérer (Larsson et al., 2017), et la super-résolution en récupérant la résolution originale d'une image à partir d'une image sous-échantillonnée (Ledig et al., 2017). D'autre part, les tâches basées sur la prédiction du contexte font apprendre au réseau les relations entre les parties d'une image, comme le choix de tuiles arbitraires dans une image et la prédiction de leurs emplacements spatiaux relatifs (Doersch et al., 2015). Une version améliorée de cette méthode est présentée dans Noroozi et Favaro (2016), où les tuiles ont été choisies, mélangées et le réseau a appris à identifier le modèle de mélange, l'obligeant ainsi à apprendre comment les tuiles composent l'image originale. L'auto-supervision a également été appliquée à l'imagerie médicale (Zhou et al., 2019), notamment à l'inpainting (Chen et al., 2019 ; Kayal et al., 2020) et à la résolution de puzzles en traitant une image 3D comme un Rubik's cube mélangé (Zhuang et al., 2019).

Toutes les tâches d'autosupervision ci-dessus sont conçues pour apprendre des caractéristiques utiles à partir d'une seule image d'entrée en récupérant les informations retenues dans l'image elle-même. Cependant, les informations riches qui permettent de distinguer une image d'une autre ne sont pas explicites. considéré. La tâche d'identification de la source dans ce document vise apprendre non seulement les caractéristiques qui permettent d'identifier chaque image, mais aussi celles qui permettent de distinguer une image des autres images de l'ensemble de données.

La tâche que nous proposons présente certaines similitudes avec la méthode contemporaine d'apprentissage contrastif (Becker et Hinton, 1992), qui gagne également en popularité dans le domaine de l'imagerie médicale (Jiao et al., 2020 ; Li et al., 2021 ; Chaitanya et al., 2020 ; Feng et al., 2021 ; Li et al., 2020). Dans l'apprentissage contrastif, le réseau neuronal est chargé de reconnaître la similitude ou la dissemblance d'une paire d'images qui lui sont fournies, ce qui peut être considéré comme une méthode basée sur la prédiction du contexte plutôt que sur la reconstruction. Par exemple, la méthode de pointe connue sous le nom de *SimCLR* (Chen et al., 2020) consiste à tirer des échantillons aléatoires de l'ensemble de données original, à appliquer deux augmentations (toutes deux échantillonnées à partir de la même famille d'augmentations) sur les échantillons pour créer deux ensembles de vues. Ces vues passent ensuite par un CNN et une couche de réseau neuronal entièrement connectée pour générer des représentations latentes. Enfin, ces représentations sont utilisées pour former réseau, de sorte que les vues augmentées de la même classe sont rapprochées et que les vues augmentées de classes différentes sont repoussées à l'aide d'une perte contrastive. Cela peut encourager les caractéristiques latentes à être plus compactes et séparées, ce qui peut fournir une régularisation supplémentaire pour optimiser le réseau. Cependant, la plupart des approches d'apprentissage contrastif sont axées sur la tâche de classification en aval, et ne préapprennent que la partie en- coder du réseau. Ainsi, dans cet article, nous nous concentrons sur la comparaison entre les méthodes basées sur la reconstruction qui sont plus pertinentes pour notre tâche d'identification de source proposée, car elles pré-entraînent l'ensemble du réseau et se concentrent sur les tâches denses de prédiction en aval.

2.2. Séparation aveugle des sources

La séparation aveugle des sources (SSA), également connue sous le nom de séparation des signaux, est le problème classique de l'identification d'un ensemble de signaux sources à partir d'un signal mixte observé. Un exemple de SSA est le problème de la soirée cocktail, où un certain nombre de personnes parlent simultanément dans un environnement bruyant (une soirée cocktail) et où un auditeur essaie d'identifier et de séparer une certaine source individuelle de voix de la discussion. Le cerveau humain peut très bien gérer ce type de problème de séparation des sources auditives, mais il s'agit d'un problème non trivial dans le traitement des signaux numériques. Des méthodes traditionnelles telles que les variantes de l'analyse en composantes indépendantes (ICA) sont proposées pour résoudre le problème de la BSS (Bell et Se- jnowski, 1995 ; Amari et al., 1996 ; Hyvarinen, 2001 ; Choi et al., 2005 ; Isomura et Toyoizumi, 2016). À l'ère de l'apprentissage profond, les réseaux neuronaux convolutifs ont été utilisés pour résoudre les problèmes de BSS dans des applications de traitement du signal telles que la reconnaissance vocale (Drude et al., 2018 ; Drude et Haeb-Umbach, 2019) et la séparation des instruments cibles (Chandna et al., 2017). Ces travaux emploient généralement un réseau d'encodeurs pour apprendre les embeddings des signaux observés, puis utilisent des techniques traditionnelles comme les k-moyennes ou le clustering spectral pour regrouper les em- beddings en fonction du nombre de sources. Le regroupement peut également être effectué par un réseau neuronal profond (Hershey et al., 2016). Cet article présente une tâche auto-supervisée de type BSS sur des données d'images, dans laquelle un réseau neuronal est formé pour

identifier et restaurer le contenu de l'image source dans les mélanges d'images multiples.

2.3. Relation avec le débruitage

Une tâche connexe à l'identification de source proposée est le débruitage (Tian et al., 2020), qui est utilisé pour identifier et supprimer les artefacts d'imagerie non souhaités. Dans le débruitage, l'image et le bruit sont considérés comme deux sources différentes et un modèle est entraîné pour les séparer. Les propriétés statistiques du signal et du bruit sont très différentes, contrairement à notre cas, où une image mixte est construite à partir d'images appartenant au même ensemble de données. Un réseau de débruitage est susceptible d'apprendre davantage de caractéristiques locales pour distinguer le bruit des images propres plutôt que des caractéristiques sémantiques de haut niveau du contenu de l'image. Contrairement à la tâche de débruitage, l'approche proposée pour l'identification de la source tente de séparer une image d'une image fusionnée avec d'autres images plutôt que de l'identifier à l'aide d'un réseau de débruitage.

qu'avec du bruit. Il s'agit d'une tâche plus difficile qui est plus susceptible de capturer des caractéristiques sémantiques utiles à partir de l'ensemble de données.

2.4. Lien avec la confusion

Le mixage a d'abord été proposé comme stratégie d'augmentation des données lors de l'entraînement des CNN dans un cadre général (Zhang et al., 2018), et a été validé pour fonctionner correctement dans la segmentation d'images médicales.
tation (Eaton-Rosen et al., 2018). La mixité, dans un segmen-
à partir d'un ensemble d'images.
fonctionne en sélectionnant de manière aléatoire une paire d'images

les données d'apprentissage et de générer une combinaison pondérée des images d'entrée ainsi que des cartes de segmentation cibles. Ces images générées sont ensuite transmises à un CNN pendant l'apprentissage, plus de toute autre stratégie d'enrichissement des données qui pourrait être appropriée.

La similitude de notre travail avec Mixup réside dans la manière dont sont créées nos images mixtes, à partir desquelles, dans notre cas, le réseau apprend à identifier les sources. Cependant, notre approche est une stratégie d'autosupervision, dont l'objectif est d'enseigner au réseau l'utilisation de l'information.
de données lors du pré-entraînement, tandis que Mixup est un outil d'ajout de données qui permet d'améliorer la qualité des données.

de la méthode de mentation. Néanmoins, afin de comparer les deux, nous incluons également une série d'expériences avec Mixup en tant que stratégie supplémentaire d'augmentation des données.

3. Méthodes

Dans la section 3.1, nous donnons une définition générale de l'identification des sources. Dans la section 3.2, nous examinons si et quand la tâche d'identification de la source peut être résolue par un réseau neuronal. Dans la section 3.3, nous décrivons comment l'identification des sources peut être utilisée comme tâche de substitution pour un réseau autosupervisé. Enfin, nous décrivons quatre tâches d'autosupervision de base concurrentes populaires que nous comparons dans le présent document à la section 4.

3.1. Définition du problème de l'identification de la source

Considérons le domaine D, dans lequel chaque signal source peut être distingué des autres, par exemple, chaque signal est une image d'un patient différent dans un ensemble de données d'imagerie médicale. Plusieurs (N) signaux sources, $\mathbf{S_N} = (s_1...s_N)^T$, échantillonnés dans D sont linéairement "mélangés" pour produire M mélanges, $\mathbf{X_M} = (x_1...x_M)^T$, à l'aide d'une matrice $M\,N\times$ $\mathbf{W}$:

$$\mathbf{X_M} = \mathbf{W}\mathbf{S_N} \quad (1)$$

Le problème de la séparation aveugle des sources (BSS) consiste à reconstruire les signaux individuels qui constituent les mélanges sans connaître la transformation $\mathbf{W}$ et les signaux originaux $\mathbf{X}$.

Dans le contexte de l'utilisation de réseaux neuronaux pour cette tâche, à chaque lot d'itérations de formation, nous pouvons créer $\tilde{M}$ mélanges à partir de $\tilde{N}$ échantillons, comme le permet la taille de lot choisie. En règle générale, $\tilde{M} <= M$ et $\tilde{N} <= N$. Par exemple, deux signaux échantillonnés au hasard, s_1 et s_2, peuvent donner lieu à un mélange de signaux, x, créé par une méthode linéaire.
combinaison :

$$x = ws_1 + (1-w)s_2, \mathrm{w} \quad \in [0, 1] \quad (2)$$

où le poids, w, est un scalaire échantillonné uniformément entre 0 et 1. De nombreux signaux mixtes créés de la manière décrite ci-dessus constitueraient un lot sur lequel former le réseau neuronal.

Pour apprendre à séparer les signaux, nous pouvons entraîner un système multicanal.
modèle de réseau neuronal, $f^{\tilde{M}}(\cdot\,;\theta)$ paramétré par θ et avec $\tilde{M}$
et $\tilde{N}$ les canaux de sortie, pour apprendre les paramètres optimaux en
minimiser la perte, $\mathsf{L}(\theta)$:

$$\mathsf{L}(\theta) = \frac{1}{B}\times \ell\left(\mathbf{S}^b_{\tilde{N}}, f^{\tilde{M}}_{\tilde{N}}(\mathbf{X}^b_{\tilde{M}};\theta)\right)^b \quad (3)$$

où B est la taille du lot, $\mathbf{S}_N = (s_1...s_{\tilde{N}})$ est une collection de

$\tilde{N}$ sources choisies au hasard telles que $\mathbf{S}^b$ est la $b^{(ème)}$ collection dans le lot, et de même $\mathbf{X} = (x_{1...x_{\tilde{M}}})$ est une collection de

les "mélanges" créés à l'aide du processus décrit dans l'équation (2) appliqué aux sources de $\mathbf{S}^{\tilde{N}}$. Essentiellement, le réseau multicanal, $f_{\tilde{M},\tilde{N}}(\cdot\,;\theta)$, consomme un seul $\mathbf{X}^b$ comme entrée pour produire $\mathbf{S}^b$ comme sortie, sur laquelle la perte est calculée. La perte ci-dessus est représenté pour un lot d'une itération à travers toutes les données disponibles.

La fonction $\ell(\cdot,\cdot)$ est composée des normes L_1 et L_2 de la différence entre le signal source original et le signal corrélé.
de la sortie du modèle :

$$\ell(\mathbf{S}_{\tilde{N}}, f^{\tilde{M}}_{\tilde{N}}(\mathbf{X}_{\tilde{M}})) = \frac{1}{\tilde{N}}\times \left\| s_{(n)} - f(\mathbf{X}^{\tilde{M}}_{n}) \right\|_1 + \left\| s_n - f(\mathbf{X}^{\tilde{M}}_{n}) \right\|_2 \quad (4)$$

$n\ l\in....\tilde{N}$

où $f^{M\tilde{}}(\mathbf{X}_{M\tilde{}})$ est l'image de sortie du $n^{ième}$ canal réseau neuronal agissant sur l'entrée $\mathbf{X}_{M\tilde{}}$ décrite dans le paragraphe précédent.

En l'absence de contraintes ou d'hypothèses sur les propriétés des signaux sources et des mélanges, le problème BSS peut être mal posé. Par exemple, si nous n'observons que quelques mélanges pour le nombre de signaux sources à reconstruire (tels que $M\tilde{} <= N\tilde{}$), et/ou si les propriétés statistiques globales des signaux ne sont pas très différentes, il se peut que les signaux mélangés ne soient pas séparables par un réseau. Nous le montrons dans la section suivante à l'aide d'un exemple simple et nous expliquons comment cela motive la technique que nous proposons.

3.2. L'identification des sources peut-elle être résolue ?

Dans cette petite expérience, pour un ensemble de données d'imagerie médicale donné, D_I, nous échantillonnons aléatoirement deux signaux sources (images), s_1 et s_2, pour chaque itération d'apprentissage et nous les "mélangeons" linéairement. Le signal mélangé sert d'entrée à un réseau neuronal, tandis que s_1 est défini comme la source à reconstruire par le réseau. Comme on peut le ,

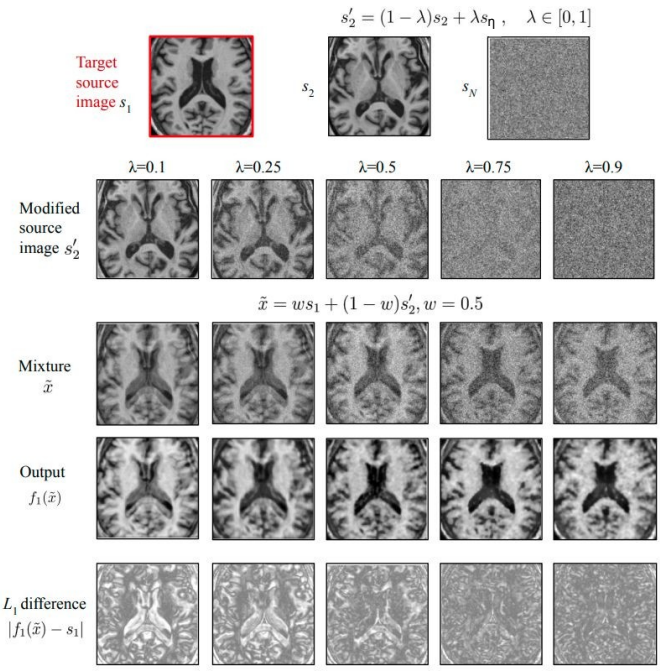


Fig. 1 : **Résultats qualitatifs de la récupération de s_1 avec différents λ.** Nous pouvons voir que le modèle est capable de séparer et de reconstruire s_1 à partir de $x\tilde{}$ gradu-
de plus en plus lorsque λ passe de 0,1 à 0,9. L'ensemble de données de cette expérience contient 30 IRM cérébrales de 30 patients différents. Meilleure visualisation avec le zoom.

nous avons $M = N\tilde{}$ (= 1) dans ce cas, et les sources sont échantillonnées

de la même distribution, ce qui rend la question mal posée.

Une façon de rendre la tâche de reconstruction moins ambiguë

serait d'échantillonner le signal s_2 dans un domaine différent de celui de s_1, par exemple en ajoutant du bruit à s_2, comme suit :

$$s_2' = (1-\lambda)s_2 + \lambda s_\eta, \qquad \lambda \in [0,1] \tag{5}$$

où $s_{\eta k} \sim \mathcal{N}(0,1)$ pour le kème voxel dans s_N, de sorte que lorsque λ= 1, s_2' est purement un bruit gaussien qui appartient à un domaine manifestement différent de celui de l'imagerie dans l'ensemble de données.

D_I. Le nouveau mélange, x, peut être créé en appliquant l'équation suivante

(2) sur s_1 et s', et la perte à minimiser par le réseau neuronal peut être calculée à l'aide des équations (3) et (4) en fixant $\tilde{M} = N\tilde{} = 1$.

Les résultats d'un réseau de neurones (nous utilisons ici 2D UNet (Ron- neberger et al., 2015)) optimisé pour minimiser la perte de recon- struction de s_1, s' avec différentes valeurs de λ sont visualisés dans le tableau suivant.

Figure 1. On peut observer que lorsque λ est petit (0,1), le taux d'émission de gaz à effet de serre est plus élevé que le taux d'émission de gaz à effet de serre.

put est une moyenne des deux images s_1 et s_2 et le modèle ne parvient pas à séparer s_1 du mélange x. Lorsque λ augmente progressivement (jusqu'à 0,9), s_1 devient plus clair et mieux séparé.

Comme l'illustre cette expérience, le réseau ne peut pas séparer sources lorsqu'elles sont échantillonnées à partir de la même distribution et que les mélanges sont faits arbitrairement. Pour imposer des contraintes supplémentaires et accroître la séparabilité, un moyen simple consiste à échantillonner des sources provenant de domaines différents, par exemple un scanner IRM et un bruit

Ces caractéristiques apprises peuvent contenir des motifs locaux triviaux et sont moins susceptibles de fournir des caractéristiques sémantiques utiles pour des tâches en aval telles que la segmentation. La technique que nous proposons ensuite repose sur la création d'un plus grand nombre de mélanges que d'échantillons à analyser.

tracé, tel que $\tilde{M} > N\tilde{}$, en ayant toujours une source fixe en
tous les mélanges créés, de sorte que le réseau puisse obtenir des informations ex-tra pour l'aider à identifier la source souhaitée (fixe). Nous appelons la première variante *Denoising SI* (DSI) et les dernières variantes *Cross-patients* SI (CSI) et *Within-patients SI* (WSI), en fonction des sources qui sont mélangées. Nous décrivons ces variantes en détail dans les sections suivantes.

3.3. *Tâche d'identification de la source proposée*

Dans cet article, nous proposons une variante simple de la tâche d'identification des sources qui résout le problème de la tâche mal posée. Dans cette tâche, nous échantillonnons les sources de manière à ce qu'une des sources soit présente dans chaque mélange d'entrée et produise la seule sortie cible. Cela suppose que le nombre de mélanges d'entrée, $\tilde{M}$, soit fixé à deux ou plus. Dans le cas de $\tilde{M}$ =2 et $N\tilde{}$=1, la tâche proposée consisterait à identifier et à séparer le signal cible, par exemple s_1, de deux mélanges x_1 et x_2 :

$$\begin{aligned} x_1 &= w_1 s_{(1)} + (1-w_1)s_2, \qquad w_1 \in [0,1] \\ x_2 &= w_2 s_1 + (1-w_2)s_3, \qquad w_{(2} \in [0,1] \end{aligned} \tag{6}$$

où w_1 et w_2 sont des scalaires, échantillonnés uniformément entre 0 et 1.

La fonction de perte pour cet arrangement s'écrit maintenant

(en utilisant l'équation (3)) comme suit :

$$L(\theta = \frac{1}{B} \sum_{b \in 1 \ldots B} \ell(s^b, f^2((x^b, x^b); \theta) \tag{7}$$

où l'exposant b désigne les échantillons d'un lot particulier. L'ordre d'entrée, (x^b, x^b) et (x^b, x^b) sont équivalents puisque l'ordre d'entrée $(x^{(b)}, x^{b)}$ est le même que l'ordre d'entrée $(x^{(b))}$.

gaussien. Toutefois, le cas λ= 1 est similaire à une tâche de débruitage autosupervisée où le modèle peut se concentrer sur l'apprentissage des différences entre le domaine de l'image et le domaine du bruit.

Les mélanges sont statistiquement interchangeables en raison de l'échantillonnage aléatoire, et ils partagent tous deux la même vérité de base, en l'occurrence s^b. Il convient de noter que même si tous les signaux sources sont échantillonnés dans le même domaine D_I, cette tâche peut être résolue par un réseau neuronal car le signal source cible est spécifique et invariant, et le nombre de mélanges est supérieur au nombre de signaux à séparer. Le déroulement de la tâche proposée est illustré à la figure 2.

Il convient de mentionner que, bien qu'il soit trivial de résoudre les équations linéaires de l'équation (6) et d'obtenir s_1, s_2 et $s_{(3)}$ de manière analytique, il n'est pas trivial pour le réseau de les résoudre lorsqu'elles sont formulées comme un problème d'apprentissage pour un réseau neuronal. exemple, avec l'utilisation de l'augmentation des données et l'échantillonnage uniforme des poids de mélange pendant l'apprentissage, la possibilité d'obtenir les mêmes entrées et sorties est très faible, et il est donc peu probable que le réseau apprenne à mémoriser des modèles. Cela fait de la variante SI proposée un moyen efficace d'apprendre des caractéristiques utiles à partir d'un ensemble de données, sans annotations laborieuses, en évitant l'ambiguïté. Par rapport à l'introduction d'un domaine différent pour résoudre le problème de l'ambiguïté dans la section 3.2, la méthode proposée se concentre sur le même domaine qui plus susceptible d'apprendre des caractéristiques utiles pour les tâches en aval.

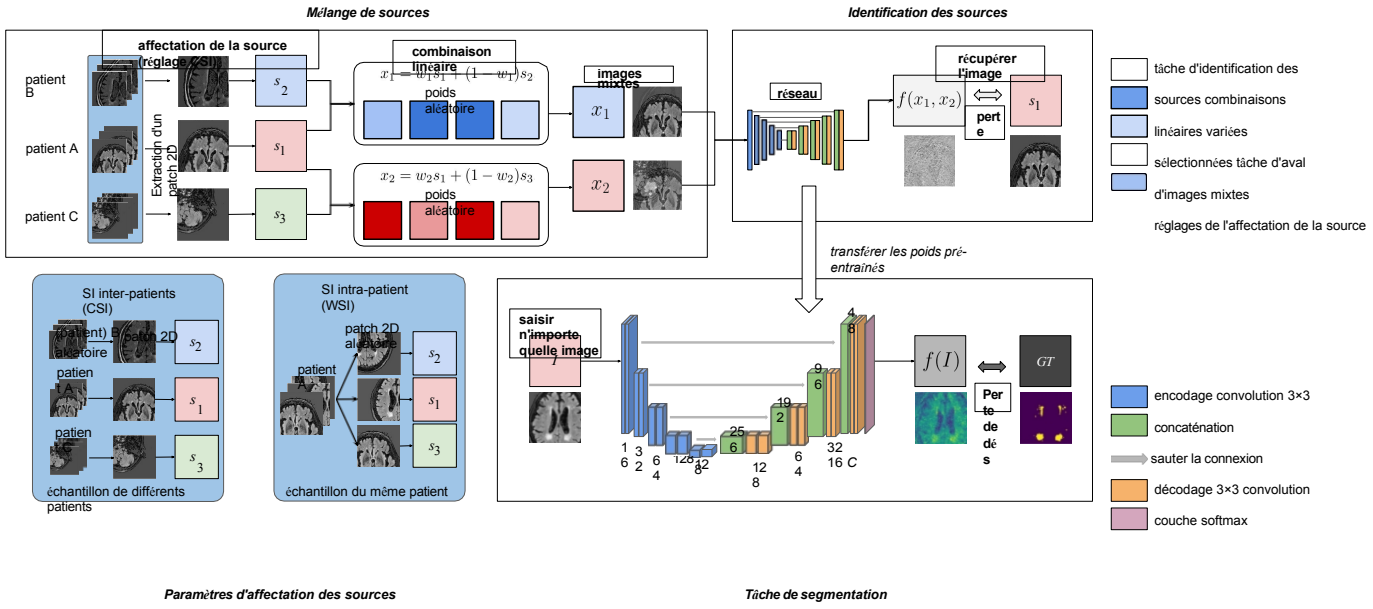


Fig. 2 : **La tâche d'identification de la source proposée.** Trois images de sources s_1, s_2 et s_3 sont utilisées pour cette illustration. L'identification transversale des patients (CSI) et l'identification intra-patient (WSI) sont deux stratégies différentes pour extraire les signaux de source, qui se concentrent sur l'apprentissage des caractéristiques entre les différents patients et au sein d'un patient individuel, respectivement. 2× 2 Le sous-échantillonnage et le suréchantillonnage sont appliqués entre les différentes résolutions de l'UNet. Meilleur affichage en couleur avec zoom.

4. Tâches d'autosupervision de référence

Nous comparons la méthode proposée à quatre tâches d'autosupervision largement utilisées pour la prédiction dense (Shurrab et Duwairi, 2021 ; Zhou et al., 2019). Les trois premières tâches se concentrent sur la re-construction et la prédiction basée sur le contexte dans une image, tandis que la dernière tâche se concentre sur la correction de l'intensité.

4.1. Peinture

L'inpainting d'image est le processus de reconstruction du contenu erroné ou endommagé d'une image.
la restauration de peintures et de photographies (Bertalmio et al., 2000). L'inpainting, en tant que tâche d'autosupervision, consiste à masquer intentionnellement des zones sélectionnées d'une image et un réseau doit apprendre à récupérer le contenu manquant.

Dans cet article, nous mettons en œuvre l'autosupervision de l'inpainting en superposant une image I à une grille régulière G de taille fixe et à une grille G *de* taille fixe.
le masquage aléatoire de cellules de grille sélectionnées. Formellement, une cellule de grille de pixels sélectionnée, indiquée comme $g(I)$, où $g \in G$, est transformée comme suit :

$$g'(I) = \begin{cases} g(I) \text{ si } B(\gamma) = 1, \\ 0 \quad \text{autrement.} \end{cases} \tag{8}$$

où $B(\gamma)$ suit une distribution de Bernoulli avec une probabilité γ
γ est un hyperparamètre compris entre 0 et 1. Cela signifie que dans chaque mini-lot, un réseau ne voit qu'environ γ contenus aléatoires des images d'entrée et tente de prédire le reste. En masquant les grilles d'une aussi non déterministe, nous évitons les cas où le réseau peut se concentrer sur des recon- structions faciles et sur l'apprentissage de caractéristiques triviales.

4.2. Mélange local de pixels

Le mélange local de pixels est connu pour aider un réseau à apprendre les informations locales au sein d'une image, sans compromettre les structures globales (Zhou et al., 2019). Cette tâche est similaire à l'inpainting, mais avec des informations supplémentaires sur la distribution des intensités à inpaint. Dans cette tâche, les images synthétiques sont générées en mélangeant aléatoirement les pixels dans la cellule de grille sélectionnée, comme le montre l'équation suivante :

$$g'(I) = \begin{cases} PgQ(I) \text{ si } B(\gamma) = 1, \\ g(I) \quad \text{sinon.} \end{cases} \tag{9}$$

où γ est un hyperparamètre compris entre 0 et 1, similaire à celui de l'inpainting ; P et Q sont des matrices de permutation. Une matrice de permutation est une matrice carrée binaire qui peut permuter les lignes d'une matrice arbitraire lorsqu'elle lui est prémultipliée, et permuter les colonnes lorsqu'elle lui est postmultipliée. Ainsi, dans le premier cas de l'équation (9), une nouvelle cellule de grille de pixels est générée en mélangeant à la fois les lignes et les colonnes de la grille originale.

4.3. Super-résolution

La super-résolution peut être mise en œuvre comme une tâche d'auto-supervision (Zhao et al., 2020), dans laquelle un réseau est entraîné à dé-flouter l'image à faible résolution. Pour créer l'image à faible résolution, le réseau est entraîné à dé-flouter l'image à faible résolution.

L'image synthétique résultante I' est constituée de toutes les grilles sélectionnées, $g'(I)$, conservant ainsi $1-\gamma$ fraction de l'im- age originale.

à partir des images à haute résolution pour l'apprentissage, nous brouillons les images à haute résolution en transformant chaque cellule de la grille en remplaçant toutes ses valeurs par celles du centre de la grille :

$$g'(I) = g(I)_{(\lfloor w/2 \rfloor, \lfloor h/2 \rfloor)} \tag{10}$$

où w et h sont largeur et la hauteur de la cellule de la grille $g(I)$. Au cours du processus d'apprentissage, le réseau apprend, à partir dune image transformée, à prédire la version haute résolution, c'est-à-dire l'image originale avant transformation.

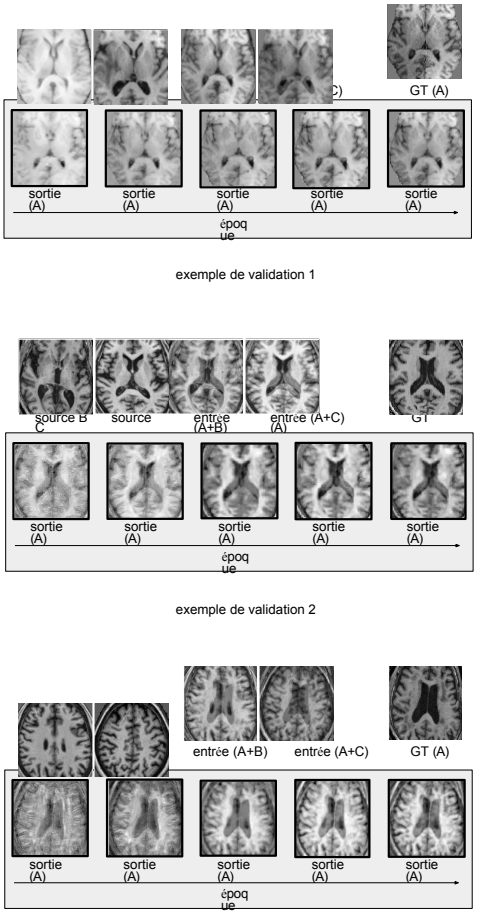


Fig. 3 : **Visualisation du réseau résolvant la tâche SI lors de lentraînement sur l'ensemble de données BraTS (exemple 1) et WMH (exemples 2 et 3).** L'image T1 est utilisée pour la visualisation. Les sous-figures montrent trois échantillons de validation différents composés d'images d'entrée mixtes, la deuxième ligne de chaque sous-figure montrant la sortie du réseau pour cinq époques intermédiaires. La configuration de l'IS inter-patients (CSI) (avec trois sources) est visualisée. Les détails de la configuration des sources sont présentés à la section 5.2. Nous pouvons voir que le réseau est capable d'identifier et de reconstruire le signal source cible A à partir des mélanges d'entrée $A+$ B et $A+$ C progressivement au cours de l'apprentissage.

4.4. Déplacement non linéaire de l'intensité

Le mécanisme de décalage d'intensité est proposé par Zhou et al. (2019), où chaque valeur de pixel dans l'image est traduite de manière monotone à l'aide d'une courbe de Bézier (désignée par la fonction B) (Morten- son, 1999). En imagerie médicale, comme les valeurs d'intensité dans une image correspondent généralement aux détails anatomiques sous-jacents, cette tâche peut être utilisée pour encourager un réseau à apprendre des caractéristiques anatomiques utiles.

Étant donné une valeur de voxel v qui est normalisée entre [0, 1], les points d'extrémité p_0, p_3, et deux points de contrôle p_1, p_2, la valeur trans- formée du pixel est donnée par :

$$v' = B(v) = (1-v^3)p_0 + 3x(1-v^2)p_1$$

$$+3v^2(1-v)p_2 + v^3 p_3 \quad (11)$$

5. Expériences

5.1. Ensembles de données

Nous appliquons notre méthode à deux problèmes de segmentation en imagerie médicale : la segmentation des tumeurs cérébrales et la segmentation des hyperintensités de la matière blanche. Les deux ensembles de données contiennent des images IRM du cerveau.

5.1.1. Ensemble de données BraTS

Multimodal Brain Tumor Segmentation Challenge 2018 (Menze et al., 2014 ; Bakas et al., 2017, 2018) se concentre sur l'évaluation des méthodes de segmentation des tumeurs cérébrales dans les scans d'imagerie par résonance magnétique (IRM) multi-modaux. Il y a au total 210 images IRM acquises auprès de différents patients. Chaque

L'image RM contient quatre modalités : pré-contraste pondéré en T1, post-contraste en pondération T1, pondération T2 et FLAIR. Trois classes de tumeurs cérébrales sont fournies en tant qu'annotations manuelles : 1) le noyau tumoral nécrotique et le noyau tumoral sans rehaussement (NCR&NET) ;
2) l'œdème péritumoral (ED) ; et 3) la tumeur qui se renforce.
(ET). Étant donné que les classes d'évaluation du défi sont les classes combinées : tumeur entière (NCR&NET+ ED+ ET), cœur de la tumeur (NCR&NET+ ET) et tumeur améliorante (ET), nous utilisons ces classes combinées pour l'entraînement proprement dit. Nous avons divisé aléatoirement l'ensemble de données en 1) 100 sujets pour l'entraînement aux tâches d'autosupervision et à la tâche principale de segmentation ; 2) 10 sujets pour la validation ; et 3) 100 sujets pour le test. Pour chaque sujet, nous
images MR recadrées/padrées dans une taille constante de 200× 200× Z (Z est le nombre de coupes axiales de l'image) où les principaux éléments sont les suivants
Les tissus cérébraux sont préservés. Après le prétraitement des données de
nnUNet (Isensee et al., 2018), la normalisation gaussienne (soustraction de la moyenne et division par l'écart-type) est appliquée sur l'avant-plan cérébral pour chaque modalité et pour chaque image individuellement.

5.1.2. Ensemble de données WMH

Le White Matter Hyperintensities (WMH) Segmentation Challenge (Kuijf et al., 2019) évalue les méthodes de segmentation automatique des WMH dans les images RM du cerveau. Les images MR fournies contiennent des séquences MR pondérées en T1 et FLAIR et sont acquises auprès de 60 patients, où chaque groupe de 20 patients provient d'un hôpital différent. La ségrégation manuelle des lésions WMH est également fournie pour chaque image. Nous avons divisé aléatoirement l'ensemble de données en 1) 30 sujets pour l'entraînement tâches d'auto-supervision et à la tâche principale de segmentation ; 2) 10 sujets pour la validation ; et 3) 20 sujets pour le test. Pour chaque

nous avons recadré les images MR dans une taille constante de 200× 200× Z, où Z est le nombre de coupes axiales dans limage 3D. Le recadrage/adaptation de a été nécessaire car les images
Les images provenant des différents hôpitaux ont des tailles légèrement différentes et il était pratique d'avoir des images de taille constante pour qu' soient toutes traitées de la même manière par le réseau. En outre, il était pratique d'avoir des images de taille constante pour que le réseau les traite toutes de la même manière,
la taille de 200× 200 couvre le tissu cérébral principal, ce que le réseau doit consommer pour apprendre. Nous utilisons la méthode Gaus-
La normalisation sienne pour normaliser les intensités à l'intérieur de l'avant-plan du cerveau similaire à l'ensemble de données BraTS.

5.2. Paramètres de la tâche SI proposée

où les points de p_0 à p_3 sont échantillonnés indépendamment à chaque époque à partir d'une distribution uniforme continue entre 0 et 1.

Il y a deux hyperparamètres à ajuster dans la tâche proposée. Le premier concerne le processus de création des images "mixtes". Dans la tâche proposée, il y a deux hyperparamètres à ajuster.

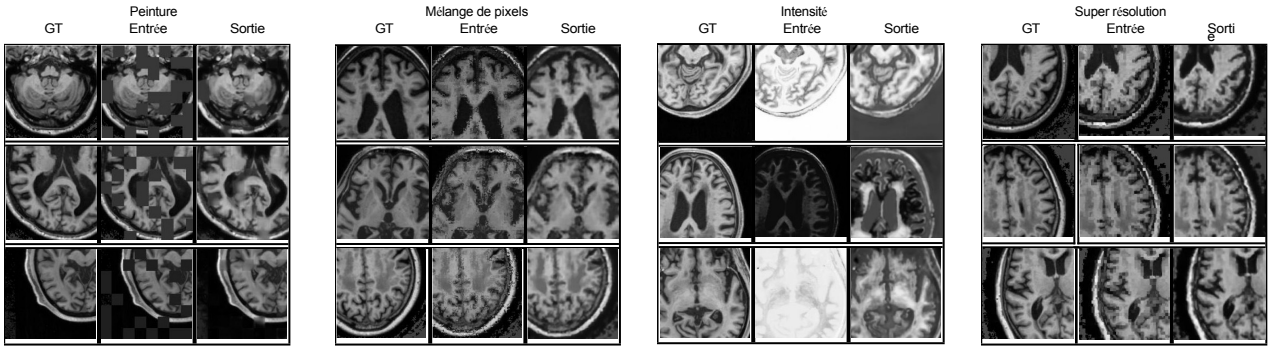


Fig. 4 : **Exemples visuels de tâches auto-supervisées de base.** Les résultats présentant la meilleure performance de validation sont utilisés pour la visualisation. Meilleure visualisation avec le zoom.

Dans la section 3.3, nous avons examiné l'exemple de la combinaison linéaire de trois signaux s_1, s_2 et s_3 pour créer deux images mélangées. On peut généraliser en choisissant $\tilde{N}$ images pour créer $\tilde{M}$ mélanges en utilisant des poids uniformément échantillonnés, de sorte que pour chaque mélange, la somme des poids échantillonnés soit égale à 1. Par exemple, lorsque N= 5, $\tilde{N}$=3, $\tilde{M}$ =2 :

$$\begin{aligned} x_1 &= w_1 s_1 + w_2 s(+ w_3 s(_3), & w_1 + w_2 + w_3 &= 1 \\ x_2 &= w_4 s_1 + w_5 s_4 + w_6 s_5, & w_4 + w_5 + w_6 &= 1 \end{aligned} \tag{12}$$

où tous les poids sont des scalaires échantillonnés aléatoirement entre 0 et 1 dans les conditions, et l'image commune s_1 doit être reconstruite. Rappelons que N est l'ensemble des images sources choisies au hasard pour la création d'images mixtes en un seul lot d'une itération d'apprentissage, tandis que $\tilde{N}$ est le nombre d'images composantes mixées, qui est un sous-ensemble *de N*.

Le deuxième hyperparamètre est la stratégie d'affectation des sources. Dans le présent document, nous considérons trois types de stratégies d'affectation des sources :

5.2.1. Patients croisés SI

Pour que le réseau apprenne à identifier l'image source cible et à la distinguer des autres images sources, N patients aléatoires sont utilisés pour extraire des signaux (coupe 2D par patient) respectivement dans chaque échantillon d'apprentissage. Nous appelons cette variante de l'IS l'*IS inter-patients* (*ISIC*).

5.2.2. SI intra-patient

Pour que le réseau se concentre sur chaque source particulière de l'ensemble de données, nous utilisons la même image du patient pour extraire tous les N signaux (toutes les coupes 2D du même patient). Étant donné que les informations d'un seul patient sont utilisées dans chaque mélange, il est peu probable que le réseau apprenne les informations transversales entre les différents patients. Nous appelons cette variante de l'IS l'*IS à l'intérieur d'un patient* (WSI).

5.2.3. Débruitage SI

Pour étudier la différence entre la tâche SI proposée et la tâche de débruitage traditionnelle, nous remplaçons les sources s_2 et s_3 dans CSI par un bruit gaussien aléatoire de moyenne nulle et de variance unitaire. Cette tâche est similaire à une tâche de débruitage traditionnelle et encouragerait le réseau à apprendre des caractéristiques représentatives qui distinguent des sources distribuées différemment, telles que l'image et le bruit.

explicitement le bruit. Nous appelons cette variante SI "*Denoising SI*". (DSI).

Toutes les expériences de la section 5 appliquent la combinaison linéaire de trois signaux dans deux mélanges, comme indiqué dans l'équa- tion (12), où au total N= 5 signaux sont utilisés pour générer un échantillon d'apprentissage. Ce paramètre est réglé sur l'ensemble de validation pour les deux ensembles de données. Pour éviter que le réseau n'apprenne des
toutes les combinaisons sans chevauchement entre régions cérébrales sont exclues en tant qu'échantillons d'entraînement.

5.3. Paramètres des tâches de base

Pour l'inpainting, la taille de la grille est réglée entre [2, 2] et [64, 64] et le pourcentage de masquage entre 0% et 100% ; pour le mélange local de pixels et la super-résolution, la taille de la grille est réglée de la même manière que pour l'inpainting. Il n'y a pas d'hyperparamètre à régler pour le décalage d'intensité non linéaire. Tous les hyperparamètres sont réglés sur l'ensemble de validation de la performance de la tâche principale.

5.4. Architecture du réseau

Nous utilisons le même réseau pour les tâches proxy d'autosupervision et la tâche principale de segmentation. est basé sur 2D UNet (Ronneberger et al., 2015) et les détails du réseau sont présentés dans la figure 2. Le réseau comporte deux couches 'entrée et de sortie : 1) pour l'entraînement de la tâche SI proposée, la couche d'entrée-sortie est définie comme suit
La couche d'entrée comporte T× 2 canaux, T étant le nombre de modalités d'imagerie pour les deux mélanges d'entrée. La couche de sortie comporte
T canaux pour reconstruire toutes les modalités de s_1 ; 2) pour la segmentation, la couche d'entrée est remplacée par une nouvelle couche avec T canaux pour l'image d'entrée x et la couche de sortie est remplacée par une couche avec C canaux pour les prédictions de segmentation où C est le nombre de classes. Toutes les couches intermédiaires sont partagées entre la tâche proxy pré-entraînée et la tâche principale. Lorsqu'aucun réseau pré-entraîné n'est utilisé, les poids de toutes les couches convolu- tionnelles sont initialisés par l'initialisation de Kaiming (He et al., 2015).

Le choix des paramètres du réseau est influencé par le modèle de pointe nnUNet (Isensee et al., 2018), décrit dans la section 5.5.3.

5.5. Stratégie de formation et augmentation des données

Nous menons les principales expériences dans un cadre entièrement supervisé et dans un cadre semi-supervisé pour les deux ensembles de données.

Tableau 1 : **Résultats du paramétrage entièrement supervisé.** Chaque expérience est répétée 3 fois avec différentes répartitions aléatoires des données. Pour BraTS, les mêmes 100 images sont utilisées pour l'entraînement de la tâche principale de segmentation (avec étiquettes) et les tâches proxy d'auto-supervision (sans étiquettes) ; pour WMH, les mêmes 30 images sont utilisées à la fois pour les données étiquetées et non étiquetées. Le score moyen de Dice (écart-type) sur toutes les données de test de l'expérience est indiqué pour chaque classe individuellement, où WT**=** tumeur entière, TC**=** cœur de la tumeur, ET**=** tumeur améliorante, *All WT***=+** TC**+** ET (uniquement pour BraTS), WMH**=** matière blanche.
hyperintensités. * : significativement meilleur que le CNN de référence ($p<$ 0.05).◦ : significativement moins bon que le CNN de référence ($p<$ 0.05). Les valeurs P sont calculées par un test t bilatéral pour chaque classe. En gras : meilleurs résultats et non significativement différents des meilleurs résultats.

	BraTS				
Méthodes/Classe	WT	TC	ET	*Tous*	WMH
CNN	0.866(0.11)	**0.835**(0.17)	0.785(0.16)	0.846(0.11)	0.775(0.11)
CNN-restart	0.868(0.11)	0.825(0.19)◦	0.786(0.16)	0.848(0.11)	0.781(0.11)
Peinture	0.867(0.11)	**0.838**(0.17)	0.788(0.16)	0.850(0.11)	0.782(0.13)
Mélange de pixels	0.859(0.13)	0.829(0.20)◦	0.777(0.18)◦	0.844(0.13)	0.777(0.12)
Intensité	0.865(0.12)	**0.838**(0.18)	0.787(0.16)	0.846(0.12)	0.775(0.13)
Super résolution	0.852(0.13)◦	**0.838**(0.18)	0.786(0.17)	0.842(0.13)	0.776(0.12)
Débruitage SI	0.868(0.09)	0.821(0.17)◦	0.783(0.16)	0.850(0.10)	0.771(0.12)
SI intra-patient	0.869(0.09)	0.817(0.19)◦	0.781(0.16)	0.851(0.10)	0.769(0.12)
Patients croisés SI (le nôtre)	**0.878**(0.09)*	**0.837**(0.17)	**0.796**(0.15)*	**0.861**(0.09)*	**0.793**(0.11)*

5.5.1. Un cadre entièrement supervisé

L'entraînement du réseau de manière autosupervisée se fait en deux étapes. Tout d'abord, nous devons pré-entraîner le réseau avec la tâche de substitution correspondante, comme décrit dans les sections 3.3 et 4. La tâche de substitution utilise le même ensemble de données que la tâche principale ; par exemple, pour l'ensemble de données BraTS, nous pré-entraînons et affinons le réseau sur les mêmes 100 images (étiquetées) de l'ensemble d'entraînement. Une taille de lot de 1 est utilisée pour la tâche de substitution dans toutes les expériences de ce document, réalisée par un réglage de 1 à 4, sur la base de l'ensemble de validation. Ensuite, pour la tâche principale, nous utilisons une taille de lot de 8 et 4 pour la formation de la tâche principale dans BraTS et WMH, respectivement, obtenue par un réglage de 1 à 16 basé sur l'ensemble de validation.

5.5.2. Réglage semi-supervisé

Dans le cadre entièrement supervisé, nous utilisons l'ensemble des données d'entraînement pour pré-entraîner et affiner le réseau. Étant donné que la force de l'autosupervision vient du fait qu'un réseau a besoin d'un volume de données beaucoup plus faible pour être affiné, nous menons également des expériences pour tester cette hypothèse, que nous appelons l'environnement semi-supervisé. Dans ce cadre, le réseau est pré-entraîné sur l'ensemble des données d'entraînement, mais affiné sur une fraction seulement des données d'entraînement. 25 des 100 images étiquetées ont été utilisées à partir de l'ensemble d'entraînement pour affiner le modèle pré-entraîné pour BraTS ; pour WMH, nous n'avons utilisé que 5 images. La même taille de lot est utilisée pour la tâche proxy et la tâche principale que celle utilisée dans le cadre entièrement supervisé.

5.5.3. Paramètres d'augmentation et d'optimisation des données

Des rotations aléatoires, des mises à l'échelle, des retournements et des déformations élastiques sont appliqués aux images 2D d'origine en tant qu'augmentation des données pour toutes les expériences. Conformément à l'article de nnUNet, nous utilisons la méthode SGD optimizer et une politique de taux d'apprentissage "poly" $(1-(\text{epoch}/\text{epoch}_{max})^{0.9})$, où $\text{epoch}_{max}=$ 1000 et pour l'ensemble de données BraTS et 10000 pour WMH, avec le taux d'apprentissage initial 1×10^{-2}, le momentum 0,99, et la décroissance du poids $3\times 10^{-(5)}$ à la fois pour la tâche de substitution et tâche principale. Un arrêt prématuré est appliqué lorsqu'il n'y a pas d'amélioration.
Nous avons également essayé de redémarrer l'optimisation pour la tâche principale après l'optimisation de l'ensemble de validation. Nous avons également essayé de redémarrer l'optimisation pour la tâche principale après avoir op-

de l'initialisation aléatoire, que nous appelons *CNN-restart.* pour les deux ensembles de données afin de permettre une comparaison équitable.

6. Résultats

6.1. Résultats de la segmentation

Le tableau 1 montre les résultats de la segmentation pour les deux ensembles de données dans le cadre entièrement supervisé. La méthode SI proposée pour les patients croisés atteint la meilleure performance moyenne (à l'exception de TC : cœur de la tumeur dans BraTS) dans les deux ensembles de données et montre une amélioration significative par rapport aux autres lignes de base et variantes SI dans quatre des cinq classes (WT : tumeur entière, ET : tumeur améliorante, *All* : WT+ TC+ ET, et WMH). La classe *All* calcule le coefficient de Dice de WT+ TC+ ET ensemble (en concaténant les trois classes mais sans les additionner en une seule classe) et est la plus importante dans BraTS.

Parmi les trois paramètres différents de la tâche d'identification de la source (CSI, WSI et DSI), CSI obtient les meilleurs résultats avec un score de Dice de 0,861 (tous) et de 0,793 dans les ensembles de données BraTS et WMH séparément, ce qui est nettement meilleur que WSI et DSI. WSI et DSI ont des performances similaires dans les deux ensembles de données et ne sont pas significativement différents l'un de l'autre. Cela suggère l'importance de la configuration multi-sources. L'une des raisons pourrait être que, par rapport à WSI et DSI, CSI utilise les données de manière plus efficace lorsque le réseau voit plus d'images sources par époque. Il convient également de noter que la tâche de mélange des pixels présente des performances inférieures à celles de la ligne de base CNN dans quatre des cinq classes (significatives dans les classes TC et ET). Dans la segmentation du noyau tumoral (TC), quatre méthodes (inpainting, intensity shift, super-resolve, et CSI) montrent des améliorations comparables à ligne de base CNN (non significatives entre elles), ce qui indique que l'efficacité des différentes méthodes auto-supervisées peut varier selon les classes et que la segmentation du noyau tumoral est plus difficile à améliorer que les autres classes. Néanmoins, dans l'ensemble, l'ICS proposé peut fournir un meilleur point de départ pour la tâche de segmentation que la plupart des tâches de base d'auto-supervision.

Tableau 2 : **Résultats du paramétrage semi-supervisé.** Les meilleurs résultats sont indiqués en gras. Chaque expérience est répétée 3 fois avec différentes répartitions aléatoires des données. Pour BraTS, les 100 images d'entraînement sont utilisées pour entraîner la tâche auto-supervisée non étiquetée ; le réglage fin est effectué sur 25 des images d'entraînement en utilisant les étiquettes de segmentation et les 25 images étiquetées sont contenues dans les 100 images non étiquetées. Pour WMH, 5 images sont utilisées pour les données étiquetées et 30 images sont utilisées pour les données non étiquetées ; les 5 images étiquetées sont contenues dans les 30 images non étiquetées. Le score moyen de Dice (écart-type) sur toutes les données de test expérimental est indiqué pour chaque classe individuellement, où WT= tumeur entière, TC= cœur de la tumeur, ET= amélioration de la qualité de vie.

tumeur, *Tous WT*=+ TC+ ET (uniquement pour BraTS), WMH= hyperintensités de la matière blanche. * : significativement meilleure que la ligne de base CNN ($p<$ 0.05). :⋄

significativement moins bonne que la ligne de base CNN ($p<$ 0.05). Les valeurs P sont calculées à l'aide d'un test t bilatéral pour chaque classe. Caractères gras : meilleur et non

significativement différents des meilleurs résultats.

	BraTS				
Méthodes/Classe	WT	TC	ET	*Tous*	WMH
CNN	0.823(0.11)	0.780(0.21)	0.743(0.19)	0.816(0.12)	0.739(0.16)
CNN-restart	0.821(0.13)	0.775(0.22)	0.739(0.19)	0.812(0.13)	0.731(0.16)⋄
Peinture	0.842(0.15)	**0.817**(0.20)*	0.754(0.18)*	0.827(0.15)	**0.761**(0.12)*
Mélange de pixels	0.823(0.17)	0.782(0.23)	0.723(0.21)⋄	0.806(0.17)	0.744(0.15)
Intensité	0.832(0.16)	0.804(0.21)*	0.746(0.19)	0.817(0.16)	0.740(0.15)
Super résolution	0.848(0.15)	**0.819**(0.20)*	**0.760**(0.19)*	0.829(0.14)	0.756(0.13)*
Débruitage SI	0.823(0.16)⋄	0.776(0.22)	0.747(0.20)	0.804(0.16)⋄	0.755(0.13)
SI intra-patient	0.836(0.13)	0.779(0.20)	0.749(0.18)	0.814(0.13)	0.754(0.12)
Patients croisés SI (le nôtre)	**0.855**(0.12)*	**0.811**(0.18)*	**0.764**(0.17)*	**0.837**(0.12)*	**0.783**(0.11)*

6.2. Résultats semi-supervisés

Nous menons des expériences sur les deux ensembles de données dans un cadre semi-supervisé afin d'étudier dans quelle mesure la tâche d'auto-supervision proposée serait utile lorsque seule une petite quantité de données labiles est disponible pour entraîner la tâche de substitution. Les résultats sont présentés dans le tableau 2. Des tendances similaires peuvent être observées dans ces résultats semi-supervisés par rapport aux résultats entièrement supervisés. Comme dans le tableau 1, la méthode CSI proposée obtient les améliorations les plus importantes pour BraTS (à l'exception du cœur de la tumeur) et WMH. Les améliorations sont significatives par rapport à toutes les autres méthodes pour la tumeur entière et pour *toutes les* tumeurs. Dans le cas du WMH, les deux méthodes

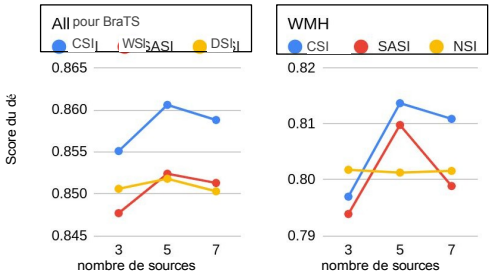
La méthode CSI proposée et l'inpainting sont significativement meilleurs que les autres méthodes. Il convient également de noter que lorsque seules quelques images étiquetées sont disponibles, davantage de méthodes d'autosupervision présentent des améliorations significatives par rapport à la méthode CNN de base.

(12 résultats⋆ dans le tableau 2 contre 4 résultats⋆ dans le tableau 1). Cela montre les avantages généraux de l'apprentissage par caractéristiques dans l'auto-apprentissage.

par rapport à la ligne de base du CNN.

Les variantes SI WSI et DSI affichent toujours des performances proches les unes des autres dans la plupart des classes et sont nettement moins performantes que CSI. Comme dans le cas de la configuration entièrement supervisée, la tâche de mélange des pixels ne présente pas d'amélioration par rapport à la ligne de base CNN dans la plupart des classes. Il convient de noter que les performances de l'ICS en mode semi-supervisé (0,837 dans BraTS et 0,783 dans WMH) sont très comparables aux résultats de la ligne de base CNN en mode entièrement supervisé (0,846 dans BraTS et 0,775 dans WMH), qui ont nécessité 4 fois plus d'images d'entraînement. L'inpainting et la super-résolution montrent de meilleures performances que la ligne de base CNN, mais toujours moins bonnes que le CSI (significatives dans BraTS). La méthode proposée présente des améliorations de performance plus importantes dans l'ensemble de données WMH où beaucoup moins de données étiquetées sont utilisées par rapport à l'ensemble de données BraTS (5 étiquetées contre 25 étiquetées et avec 4,4 % contre 3,1 % d'améliorations Dice par rapport à la ligne de base CNN). Cela

6.3. Influence du nombre de sources



montre que dans une situation pratique d'imagerie médicale où les étiquettes de segmentation sont rares, une tâche d'auto-supervision bien conçue peut encore préserver des performances considérables avec suffisamment de données non étiquetées.

Fig. 5 : **Influence du nombre de sources fusionnées.** Les résultats sont obtenus par une exécution indépendante sur les ensembles de données BraTS et WMH en utilisant les mêmes données dans une répartition aléatoire des données avec un réglage entièrement supervisé, similaire au tableau 1. Le nombre de sources 5 est utilisé pour les expériences des tableaux 1 et 2. Meilleure visualisation en couleur.

Nous menons des expériences pour étudier l'influence du nombre d'images utilisées dans la tâche SI proposée. N= 3, 5, 7 et $N\tilde{}$= 2, 3, 4 sources (par exemple, dans l'équation (12), N= 5, $N\tilde{}$= 3) sont testées pour générer $M\tilde{}$ =2 images fusionnées en entrée du réseau. Les expériences sont des exécutions indépendantes sur les ensembles de données BraTS et WMH dans un cadre entièrement supervisé. Il convient de noter que les hyperparamètres N et $N\tilde{}$ sont réglés sur l'ensemble de validation pour toutes les expériences. Les résultats sont présentés à la figure 5. Nous pouvons constater que le paramètre N= 5, $N\tilde{}$= 3 sources permet d'obtenir les meilleures performances dans la tâche principale de segmentation pour CSI et WSI, tandis que pour DSI, l'effet est beaucoup plus faible. Un nombre insuffisant de sources peut rendre trop facile la reconstruction du signal cible, ce qui peut donner lieu à des caractéristiques triviales, tandis qu'un trop grand nombre de sources peut rendre trop difficile la reconnaissance de la cible, ce qui donne lieu à des caractéristiques arbitraires.

6.4. Comparaison avec Mixup

Le tableau 3 présente les résultats de la comparaison entre l'approche proposée et Mixup dans un cadre entièrement supervisé. Ici, l'identification de la source des patients croisés (CSI) basée sur l'auto-supervision de la pré-supervision a été utilisée pour l'identification de la source des patients croisés.

utilisent toutes la méthode des

Tableau 3 : **Comparaison entre le SI inter-patients et le Mixup dans un environnement entièrement super- visé.** Les meilleurs résultats sont indiqués en gras. Chaque expérience est répétée trois fois avec des répartitions aléatoires des données, et les scores Dice moyens sont indiqués. * : significativement meilleur que le mélange CNN+ ($p<$ 0.05). En gras : Résultat Dice moyen le plus élevé.

Méthodes/Classe	BraTS WT	WMH
CNN	0.866(0.11)	0.774(0.11)
CNN+ confusion	0.875(0.11)	0.801(0.11)
Patients croisés SI	0.878(0.09)	0.793(0.11)
Mélange de patients croisés SI +	**0.886**(0.10)*	**0.803**(0.12)

Les résultats montrent que Mixup améliore à la fois l'approche base et notre approche proposée, avec une amélioration relative plus importante pour la détection des tumeurs dans l'ensemble de données BraTS. Les résultats montrent que Mixup améliore à la fois 'approche de base et notre approche proposée, avec une amélioration relative plus importante pour la détection des tumeurs dans l'ensemble de données BraTS.

7. Discussion

Dans cet article, nous proposons une nouvelle tâche d'autosupervision appelée identification de la source (SI) qui s'inspire du problème de la séparation aveugle de la source, et nous étudions l'ambiguïté de la tâche dans le problème de l'identification de la source pour les réseaux neuronaux. Contrairement à la plupart des tâches d'autosupervision basées sur la reconstruction qui se concentrent sur la restauration du contenu de l'image à partir d'une seule image source, la tâche proposée permet au réseau de voir plusieurs images de mélanges et d'apprendre à séparer l'image source des autres et à la reconstruire. Les expériences montrent que la méthode proposée surpasse les méthodes de base dans les deux ensembles de données, y compris la ligne de base CNN, restart-CNN avec le taux d'apprentissage initial, et les méthodes auto-supervisées couramment utilisées dans la peinture, le shuf- fle de pixels, le décalage d'intensité, la super-résolution et le débruitage. La méthode proposée présente les améliorations les plus importantes dans le cadre semi-supervisé lorsque très peu de données étiquetées et de nombreuses données non étiquetées sont disponibles, ce qui est un scénario courant dans les applications d'imagerie médicale.

7.1. Comparaison avec d'autres méthodes d'autosupervision

L'une des principales différences entre la tâche SI proposée et les tâches d'autosupervision basées sur la reconstruction existantes est que la tâche SI apprend les caractéristiques non seulement de la partie restante de la même image déformée, mais aussi d'autres images du même domaine. En distinguant chaque image des autres, des caractéristiques dis- criminatives potentiellement utiles peuvent être apprises lors de la reconstruction de l'image déformée. Ces caractéristiques peuvent mieux capturer la connaissance générale du domaine, par exemple la connaissance de l'anatomie et de la pathologie, en voyant et en comparant les images de différents patients en même temps. Une bonne compréhension de l'anatomie et de la pathologie chez différents individus est nécessaire pour résoudre avec succès l'identification et la reconstruction d'une seule image. Les caractéristiques apprises par l'IS peuvent donc constituer un meilleur point de départ pour l'optimisation de la tâche en aval que les caractéristiques apprises par les tâches d'autosupervision précédentes telles que l'inpainting, le pixel shuffling, l'intensity shift, la super-résolution et le denoising.

Dans cet article, nous nous concentrons sur la comparaison entre les méthodes auto-supervisées basées sur la reconstruction, qui

l'image synthétique déformée en tant qu'entrée et l'image cible originale en tant que vérité de base. Nous considérons les méthodes basées sur la prédiction du contexte, telles que la prédiction de l'emplacement des tuiles (Doersch et al., 2015), la résolution d'énigmes (Zhuang et al., 2019), l'apprentissage contrastif (Becker et Hinton, 1992), comme une autre catégorie de tâches autosupervisées. Ces méthodes optimisent une tâche prédéfinie de clas- sification/régression basée sur les informations d'une seule image (Doersch et al., 2015 ; Zhuang et al., 2019) ou de différentes images (Becker et Hinton, 1992), et elles n'entraînent donc généralement pas de décodeur pertinent (dense). Au contraire, les méthodes basées sur la reconstruction nécessitent intrinsèquement un décodeur dense pour l'apprentissage de caractéristiques concrètes et à haute résolution et la sortie de prédictions denses par pixel, ce qui peut donner lieu à un modèle qui s'adapte mieux aux tâches de prédiction denses comme la segmentation.

7.2. Appliquer l'IS à l'aide de données non étiquetées avec moins de surajustement

L'apprentissage auto-supervisé permet d'utiliser des données non étiquetées sans annotations supplémentaires de la part d'experts et d'effectuer un préapprentissage avec des données étiquetées et non étiquetées avant l'apprentissage entièrement supervisé. La qualité des caractéristiques apprises dans le cadre de tâches autosupervisées est généralement évaluée pour des tâches en aval telles que la segmentation. Dans nos expériences, des améliorations plus importantes sont observées dans le cadre semi-supervisé par rapport au cadre entièrement supervisé, en particulier pour l'ensemble de données WMH. Nos résultats montrent qu'avec la même quantité de données non étiquetées, l'IS proposé peut apprendre plus de caractéristiques utiles à partir des données non étiquetées qu'avec d'autres tâches auto-supervisées. L'une des raisons pourrait être que la tâche SI proposée souffre moins du problème de surajustement que les méthodes traditionnelles telles que l'inpainting et la super-résolution. Par exemple, étant donné les données non étiquetées, le modèle peut essayer de résoudre la tâche d'inpainting ou de super-résolution en mémorisant les images d'entrée et en restaurant le contenu manquant lorsque la capacité du modèle est suffisante, ce qui peut entraîner l'apprentissage de caractéristiques triviales. En revanche, tâche de super-résolution prend en compte des combinaisons d'images beaucoup plus nombreuses différentes pour une même quantité de données non étiquetées (lorsque N= 5 pour 100 images, le nombre de combinaisons d'images possibles serait le coefficient binomial

$C(100, 5)\times 5\approx 3.8\times 10^{8}$), ce qui rend le modèle plus difficile à comprendre.

de mémoriser et de s'adapter à une image particulière, mais il doit trouver des solutions à ses problèmes.

une manière plus générale de résoudre la tâche SI, par exemple en apprenant la connaissance de l'anatomie, ce qui peut être non trivial et utile pour des tâches en aval telles que la segmentation.

7.3. Application à d'autres tâches de prédiction de densité

Dans cet article, nous appliquons la méthode SI proposée à la segmentation, une tâche de prédiction dense. Les caractéristiques SI pré-entraînées peuvent également être transférées à d'autres tâches de prédiction dense en imagerie médicale, comme par exemple l'estimation de la profondeur (Liu et al., 2019), le recalage d'images (Balakrishnan et al., 2018) et la détection basée sur des cartes de distance (van Wijnen et al., 2019). De plus, ces tâches peuvent également bénéficier des caractéristiques inter-sources apprises dans méthode SI. Par exemple, un bon modèle d'enregistrement d'images peut nécessiter non seulement les alignements entre les modèles locaux à travers différentes modalités (au sein d'un patient), mais aussi la connaissance de l'anatomie générale à travers différents patients pour contraindre les transformations possibles. Avec une conception appropriée de l'ensemble de données de substitution

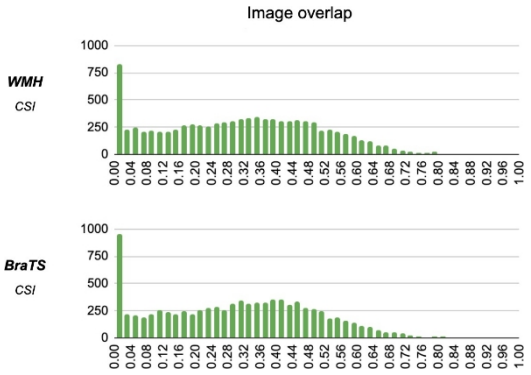


Fig. 6 : **Distribution du degré de mélange d'images échantillonnées aléatoirement.** Nous observons que cette distribution est presque uniforme dans les deux ensembles de données utilisés.

et le cadre du SI, les scénarios potentiels d'application de la méthode proposée peuvent être considérablement élargis.

7.4. Limites

Il a été étudié dans la littérature que la performance des approches d'auto-supervision diffère de manière significative en fonction de la difficulté de la tâche de pré-entraînement et de son lien avec la tâche principale (Su et al., 2020 ; You et al., 2020 ; Kayal et al., 2020). Par exemple, les performances de la peinture en tant que tâche d'autosupervision sont affectées lorsque la taille de la zone masquée est trop grande ou trop petite. Si la zone masquée est trop grande, la tâche de pré-entraînement sera trop difficile à résoudre ; si elle est trop petite, elle sera très facile. Cela affecterait la qualité des caractéristiques apprises et donc l'efficacité du réseau pour la tâche principale. De même, pour notre approche, la performance du réseau est déterminée par le degré de séparabilité des images mélangées et par la quantité d'informations que le réseau doit apprendre pour les séparer.

Nous testons indirectement la première hypothèse à la section 3.2, , où il est démontré que des images très similaires seraient extrêmement difficiles séparer. Pour déterminer si cela constitue un problème pratique dans notre cas, nous concevons une expérience statistique simple. Tout d'abord, des paires de coupes 2D sont échantillonnées au hasard à partir de différentes images d'un ensemble de données (BraTS ou WMH), et l'augmentation des données leur est appliquée comme décrit dans la section 5.5.3. Ensuite, le masque cérébral est extrait des images résultantes, via un simple seuillage basé sur l'intensité, et le chevauchement des masques cérébraux correspondants est mesuré à l'aide de la similarité de Jaccard. Enfin, la distribution u- tion des similitudes mesurées est tracée et illustrée dans la Figure
6. Comme nous pouvons l'observer, les similitudes varient presque uniformément de très faibles (presque 0) à modérément élevées (0,75), indiquant que pour nos ensembles de données, le réseau recevrait un large éventail d'images mélangées pour l'apprentissage. Comme mentionné dans la section 5.2, nous excluons toutes les images mixtes avec 0 similarité (pas de chevauchement du) pour éviter que le réseau n'apprenne des caractéristiques triviales. Ainsi, pour nos expériences, nous n'avons pas besoin d'un contrôle supplémentaire du degré de mélange des images.

La deuxième hypothèse porte sur la quantité d'informations que le réseau doit apprendre pour identifier la source à partir de les images mélangées. Dans la section 6.3, nous démontrons empiriquement l'effet du nombre de sources fusionnées sur la performance finale. On constate qu'un nombre trop faible ou trop élevé de sources fusionnées nuit à l'efficacité du réseau.

L'approche que nous proposons est sensible à ces deux degrés de liberté et, bien que nous disposions de suffisamment de preuves empiriques pour les ensembles de données en question, des tests supplémentaires sont nécessaires pour formuler un commentaire général sur la sensibilité de notre méthode à ces deux facteurs.

8. Conclusion

Nous proposons une nouvelle tâche d'auto-supervision appelée identification de la source, qui s'inspire du problème classique de séparation aveugle de la source. La tâche proposée consiste à identifier et à séparer une image source cible des mélanges avec d'autres images de l'ensemble de données, ce qui nécessite des caractéristiques qui sont également pertinentes pour la tâche de segmentation en aval. Sur deux tâches de segmentation d'IRM cérébrale, la méthode proposée fournit un modèle pré-entraîné significativement meilleur pour la segmentation par rapport à d'autres lignes de base d'auto-supervision, y compris l'inpainting, le mélange local de pixels, le décalage d'intensité non-linéaire et la super-résolution dans des contextes entièrement supervisés et semi-supervisés. La méthode proposée peut être généralisée à d'autres applications de prédiction dense.

Remerciements

Les auteurs souhaitent remercier Gerda Bortsova et Hoel Kervadec pour leurs suggestions constructives, ainsi que les organisateurs des défis BraTS 2018 et WMH 2017 pour la mise à disposition des ensembles de données publiques. Ce travail a été partiellement financé par le Conseil chinois des bourses d'études (dossier n° 201706170040).

Références

Amari, S.i., Cichocki, A., Yang, H.H., et al, 1996. A new learning algorithm for blind signal separation, in : Advances in neural information processing systems, Morgan Kaufmann Publishers. pp. 757-763.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Frey- mann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 170117.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shino- hara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al, 2018. Identifier les meilleurs algorithmes d'apprentissage automatique pour la segmentation des tumeurs cérébrales, l'évaluation de la progression et la prédiction de la survie globale dans le défi brats. arXiv preprint arXiv:1811.02629 .

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. Un modèle d'apprentissage non supervisé pour l'enregistrement d'images médicales déformables, in : Proceedings of the IEEE conference on computer vision and pattern recog- nition, pp. 9252-9260.

Becker, S., Hinton, G.E., 1992. A self-organizing neural network that discovers surfaces in random-dot stereograms (Un réseau neuronal auto-organisé qui découvre des surfaces dans des stéréogrammes à points aléatoires). Nature 355, 161-163.

Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural computation 7, 1129- 1159.

Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C., 2000. Image inpaint- ing, in : Proceedings of the 27th Annual Conference on Computer Graph-and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co., USA. p. 417-424. URL: https://doi.org/10.1145/344779.344972, doi:10.1145/344779.344972.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 .

Chandna, P., Miron, M., Janer, J., Go'mez, E., 2017. Monoaural audio source separation using deep convolutional neural networks, in : International con- ference on latent variable analysis and signal separation, Springer. pp. 258- 266.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis 58, 101539.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. Un cadre simple pour l'apprentissage contrastif des représentations visuelles. URL : https://arxiv.org/abs/2002.05709.

Choi, S., Cichocki, A., Park, H.M., Lee, S.Y., 2005. Blind source separation and independent component analysis : A review. Neural Information Processing- Letters and Reviews 6, 1-57.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert : Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in : International Conference on Computer Vision (ICCV).

Drude, L., Haeb-Umbach, R., 2019. Intégration des réseaux neuronaux et des modèles spatiaux proba- bilistes pour la séparation des sources acoustiques aveugles. IEEE Journal of Selected Topics in Signal Processing 13, 815-826.

Drude, L., von Neumann, T., Haeb-Umbach, R., 2018. Deep attractor net- works for speaker re-identification and blind source separation, in : 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 11-15.

Eaton-Rosen, Z., Bragman, F.J.S., Ourselin, S., Cardoso, M.J., 2018. Impro v- ing data augmentation for medical image segmentation. Medical Imag- ing with Deep Learning URL : https://openreview.net/forum?id= rkBBChjiG.

Feng, C., Vanderbilt, C., Fuchs, T., 2021. Nuc2vec : Learning representations of nuclei in histopathology images with contrastive loss, in : Medical Imaging with Deep Learning.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv:1512.03385.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep cluster - ing : Discriminative embeddings for segmentation and separation, in : 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 31-35.

Hyvarinen, A., 2001. Blind source separation by nonstationarity of variance : A cumulant-based approach. IEEE transactions on neural networks 12, 1471- 1474.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Ko¨hler, G., Norajitra, T., Wirkert, S.J., Maier-Hein, K.H., 2018. nnu-net : Self-adapting framework for u-net-based medical image segmentation. CoRR abs/1809.10486. URL : http://arxiv.org/abs/1809.10486, arXiv:1809.10486.

Isomura, T., Toyoizumi, T., 2016. A local learning rule for independent com- ponent analysis. Scientific reports 6, 1-17.

Jiao, J., Cai, Y., Alsharid, M., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Self-supervised contrastive video-speech representation learning for ultrasound, in : International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 534-543.

Jing, L., Tian, Y., 2019. Self-supervised visual feature learning with deep neural networks : A survey. CoRR abs/1902.06162. URL : http://dblp.uni-trier.de/db/journals/corr/corr190 2. html#abs-1902-06162.

Kayal, S., Chen, S., de Bruijne, M., 2020. Region-of-interest guided su- pervoxel inpainting for self-supervision, in : Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I, Springer. pp. 500-509. URL: https://doi.org/10.1007/978-3-030-59710-8_49, doi:10.1007/978-3-030-59710-8_49.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S., 2015. Skip-thought vectors. arXiv preprint arXiv:1506.06726 .

Kuijf, H.J., Biesbroek, J.M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al. 2019. Évaluation normalisée de la segmentation automatique des hyperintensités de la matière blanche ; résultats du défi de segmentation wmh. IEEE transactions on medical imaging .

Larsson, G., Maire, M., Shakhnarovich, G., 2017. Colorization as a proxy task for visual understanding, in : 2017 IEEE Conference on Computer Vi- sion and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21-26 juillet 2017, IEEE Computer Society. pp. 840-849. URL : https://doi.org/ 10.1109/CVPR.2017.96, doi:10.1109/CVPR.2017.96.

Ledig, C., Theis, L., Husza'r, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic sin- gle image super-resolution using a generative adversarial network, in : 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105-114. doi:10.1109/CVPR.2017.19.

Li, H., Yang, X., Liang, J., Shi, W., Chen, C., Dou, H., Li, R., Gao, R., Zhou, G., Fang, J., et al, 2020. Contrastive rendering for ultrasound image seg- mentation, in : International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 563-572.

Li, J., Zhao, G., Tao, Y., Zhai, P., Chen, H., He, H., Cai, T., 2021. Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19. Pattern Recognition 114, 107848.

Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M., 2019. Estimation de la profondeur dense en endoscopie monoculaire avec des méthodes d'apprentissage auto-supervisées. IEEE transactions on medical imaging 39, 1438-1447.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. 2014. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34, 1993-2024.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Mortenson, M.E., 1999. Mathematics for Computer Graphics Applications : An Introduction to the Mathematics and Geometry of CAD/Cam, Geometric Modeling, Scientific Visualizati. 2e édition, Industrial Press, Inc, États-Unis.

Noroozi, M., Favaro, P., 2016. Apprentissage non supervisé de représentations visuelles par la résolution de puzzles, in : ECCV.

Pathak, D., Kra¨henbu¨hl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Con- text encoders : Feature learning by inpainting, in : 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27-30 juin 2016, pp. 2536-2544.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net : Convolutional networks for biomedical image segmentation, in : International Conference on Medi- cal image computing and computer-assisted intervention, Springer. pp. 234- 241.

Shurrab, S., Duwairi, R., 2021. Méthodes d'apprentissage auto-supervisé et applications dans l'analyse de l'imagerie médicale : A survey. arXiv:2109.08685.

Su, J.C., Maji, S., Hariharan, B., 2020. When does self-supervision improve few-shot learning, in : European Conference on Computer Vision, Springer. pp. 645-666.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., Lin, C.W., 2020. Apprentissage profond sur le débruitage d'images : An overview. Neural Networks .

van Wijnen, K.M.H., Dubost, F., Yilmaz, P., Ikram, M.A., Niessen, W.J., Adams, H., Vernooij, M.W., de Bruijne, M., 2019. Automated lesion detec- tion by regressing intensity-based distance with a neural network, in : Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, Springer International Publishing, Cham. pp. 234-242.

You, Y., Chen, T., Wang, Z., Shen, Y., 2020. When does self-supervision help graph convolutional networks, in : International Conference on Machine Learning, PMLR. pp. 10871-10880.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup : Beyond empirical risk minimization. International Conference on Learning Repre- sentations URL : https://openreview.net/forum?id=r1Ddp1-Rb.

Zhao, C., Dewey, B.E., Pham, D.L., Calabresi, P.A., Reich, D.S., Prince, J.L., 2020. Smore : A self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. IEEE transactions on medical imaging .

Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Got- way, M.B., Liang, J., 2019. Modèles de genèse : Generic autodidactic models for 3d medical image analysis, in : Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, Springer International Publishing, Cham. pp. 384-393.

Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y., 2019. Self- supervised feature learning for 3d medical images by playing a rubik's cube, in : Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P., Khan, A.R. (Eds.), Medical Image Computing and Computer Assisted

Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part IV, Springer. pp. 420-428. URL : https://doi.org/10.1007/978-3-030-32251-9_46, doi:10. 1007/978-3-030-32251-9_46.