Correl-Net : segmentation des défauts par apprentissage profond dans les films anciens à l'aide de couches de corrélation

Arthur Renaudeau, Travis Seng, Axel Carlier, Jean-Denis Durou

Correl-Net : segmentation des défauts par apprentissage profond dans les films anciens à l'aide de couches de corrélation

Arthur Renaudeau[1] Travis Seng[1] Axel Carlier[1] Jean-Denis Durou[1]

[1] IRIT, UMR CNRS 5505, Université de Toulouse

arthur.renaudeau@irit.fr

Résumé

Le processus de restauration de films anciens implique de nombreuses opérations, dont l'une consiste à identifier les défauts qui ont altéré le film. Cette opération peut être formulée comme un problème de segmentation binaire et résolue à l'aide de réseaux de segmentation de l'état de l'art tels que DeepLab v3+ ou NAS-FPN. Bien qu'elles soient très performantes pour apprendre les caractéristiques spatiales des défauts, ces méthodes ne prennent pas en compte le fait que les défauts constituent aussi des anomalies temporelles. Nous proposons donc une architecture fondée sur l'utilisation de couches de corrélation pour compenser le mouvement et éliminer les faux positifs potentiels, des éléments ressemblant à des défauts mais pouvant être suivis sur plusieurs images et faisant réellement partie de la scène. Nous introduisons également un processus de pré-entraînement auto-supervisé du réseau, qui précède une phase de spécification permettant d'adapter le détecteur à chaque film. Les résultats montrent que notre architecture, plus compacte et moins consommatrice de ressources que les méthodes de l'état de l'art, atteint une précision plus élevée tout en maintenant le rappel à un niveau comparable.

Mots Clés

Segmentation, réseaux de neurones, corrélation.

(a) Image d'origine.

(b) Défauts détectés (en vert).

FIGURE 1 – Détection des défauts (1b) par Correl-Net appliquée à une image de film ancien (1a).

Abstract

The old film restoration process involves many operations, one of which is the ability to identify defects that altered the film. This operation can be formulated as a binary segmentation problem and solved using state-of-the-art segmentation networks such as DeepLab v3+ or NAS-FPN. While being very powerful at describing the spatial characteristics of defects, these methods fail to take into account the fact that defects are also temporal anomalies. We therefore propose an architecture based on the correlation layer previously introduced in FlowNet to compensate motion and eliminate potential false positives, elements that look like defects but can be tracked over multiple frames and are actually part of the scene. We also introduce a self-supervised pre-training process of the network, which precedes a fine-tuning phase to specifically adapt the detector to each film. Results show that our architecture, while being more compact and less resource-consuming than state-of-the-art methods, achieves higher precision while maintaining recall at a comparable level.

Keywords

Segmentation, neural networks, correlation.

1 Introduction

Après plus d'un siècle de cinéma argentique, les bobines de pellicules sont légion, et la question se pose de savoir comment préserver ces œuvres. La préservation et la restauration sont nécessaires à la survie des grands classiques cinématographiques afin que leur état reste le plus proche possible de la version originale. Inévitablement, avec le temps, les interventions humaines et l'usure, les films se détériorent jusqu'à devenir irréparables. Heureusement, des techniques permettent de réduire les dommages subis par ces supports physiques. La restauration peut être définie, selon Usai, comme « un ensemble de procédures techniques, éditoriales et intellectuelles visant à compenser la perte ou la dégradation de l'artefact d'image en mouvement » [23]. Cette définition va au-delà de la simple idée de restaurer le film dans son état original (à condition de pouvoir donner une définition correcte du terme « original »), car toutes les pertes ne peuvent pas,

ou ne doivent pas, être compensées. En tant que praticien lui-même, Busche [2] affirme que « les pertes physiques dans les artefacts du film ne sont préoccupantes qu'en tant que perturbation de l'aspect visuel de l'image », ce qui implique que tous les défauts ne doivent pas être corrigés. La même idée est défendue par Philippot [17] : « Le restaurateur doit faire la distinction entre les lacunes qui doivent être réintégrées et celles qui doivent être laissées intactes ».

Des travaux récents ont considéré la restauration vidéo comme une tâche unique et proposé des solutions tout-en-un [6], mais ils ne permettent pas vraiment à un restaurateur humain de guider le processus, ce qui peut être problématique au regard de l'éthique de la restauration. Nous choisissons plutôt de considérer la restauration comme une approche en deux étapes, dans laquelle les défauts sont d'abord détectés automatiquement, comme cela est illustré sur la FIGURE 1, puis proposés pour validation à un restaurateur, avant d'être corrigés à l'aide de techniques d'inpainting vidéo [25, 18]. Ceci permet au restaurateur d'éliminer les faux positifs, d'identifier les défauts non détectés, mais aussi de conserver une partie des vrais positifs qui ne doivent pas être corrigés selon lui.

La segmentation des défauts est un cas particulier de problème de segmentation, dans le sens où l'aspect temporel est essentiel. Même si les taches ont des formes distinctes, leur origine physique sur la bobine (poussière, détérioration de la bobine, etc.) implique que ces défauts ne peuvent pas apparaître sur des images consécutives. En d'autres termes, elles peuvent être considérées comme des anomalies temporelles, une caractéristique qui doit guider le processus de détection.

Dans cet article, nous nous inspirons de l'architecture FlowNet [4] et introduisons un réseau de neurones convolutif particulièrement adapté à la segmentation des défauts. Avec trois images consécutives en entrée, nous construisons un encodeur dans lequel des caractéristiques sont extraites des trois images séparément, puis comparées à l'aide d'une couche de corrélation. Nous reconstruisons ensuite un masque de défauts à l'aide d'un simple décodeur. Nous montrons que l'utilisation de ces corrélations conduit à de meilleures performances, pour la segmentation des défauts, que les architectures de l'état de l'art. En outre, nous proposons un processus de pré-entraînement auto-supervisé dans lequel des défauts réalistes sont générés artificiellement pendant l'apprentissage, ce qui réduit le besoin de données supervisées. Seul un petit ensemble d'entraînement supervisé est nécessaire pour spécifier le réseau et l'adapter à un film particulier. Cet article est organisé comme suit : nous commençons par passer en revue les travaux voisins dans le paragraphe 2, puis nous détaillons notre méthode dans le paragraphe 3 et nos expériences dans le paragraphe 4.

2 État de l'art de la détection

2.1 Avant l'apprentissage profond

Les premiers détecteurs de défauts ont été développés par la BBC pour détecter le bruit impulsionnel ; ils consistent à seuiller les différences absolues entre images consécutives. Cependant, cette méthode donne des résultats limités car elle ne tient pas compte du mouvement. Dans [11], cette limitation est surmontée en introduisant le *Spike Detection Index* (SDI), dans lequel le mouvement est compensé avant de calculer les différences absolues.

Pour détecter spécifiquement les rayures, [9] a introduit une méthode où elles sont modélisées comme des sinusoïdes amorties. La détection est effectuée selon un schéma en deux étapes : un sous-échantillonnage et un filtrage pour rassembler les candidats, suivis d'un affinage bayésien pour éliminer ceux qui ne correspondent pas au modèle de sinusoïde amortie. Par la suite, [16] a étendu [9] en ajoutant une étape de vérification supplémentaire, où les valeurs des pixels voisins autour des rayures candidates sont examinées pour distinguer les rayures réelles des contours. Pour limiter encore plus les faux positifs, les mêmes auteurs réalignent dans [15] les différentes images du masque en utilisant l'estimation du mouvement et éliminent les candidats qui restent verticaux, au contraire des véritables rayures qui vont se déformer (effet de torsion) avec la compensation du mouvement.

Une autre méthode, qui a été utilisée dans [7] pour détecter les rayures, est la fermeture morphologique. La soustraction d'images avant et après fermeture révèle les rayures, qui sont suivies sur plusieurs images à l'aide d'un filtre de Kalman. La même idée est également présente dans [20], mais l'image est d'abord divisée en bandes horizontales afin de traiter séparément le premier plan et l'arrière-plan. La détection est ainsi facilitée dans les zones homogènes.

Concernant les taches, le premier modèle de champ aléatoire de Markov (MRF) de [10] a été réutilisé dans [24], conjointement à un autre MRF qui permet de renforcer la continuité spatiale. De plus, une compensation de mouvement est appliquée pour conserver la cohérence temporelle. En utilisant également la compensation de mouvement, le détecteur ROD [14] compare le pixel courant avec ses voisins temporels pour détecter les anomalies spatio-temporelles. D'autres méthodes, par exemple [26], nécessitent plusieurs étapes pour détecter les taches. Après avoir identifié les candidats en fonction de leurs caractéristiques spatiales, les faux positifs sont éliminés en recherchant les discontinuités temporelles. Des filtres médians sont utilisés dans [27] pour extraire les candidats aux taches où surviennent de brusques changements spatiaux. Ensuite, si ces gradients n'apparaissent que dans l'image courante, et non dans les images précédentes et suivantes, ces candidats sont considérés comme de véritables taches.

2.2 Avec l'apprentissage profond

La première application faisant appel aux réseaux de neurones, présentée par [8], s'attelle à la détection des rayures en utilisant la décomposition *cartoon*-texture. Alors que la détection de la forme (*cartoon*) est effectuée par filtrage, la texture est classée par un réseau de neurones prenant en entrée des images de contours. Une autre application a été proposée dans [21] pour détecter les taches en utilisant la compensation de mouvement, la détection SROD [1] et la classification de tous les pixels aberrants en utilisant un réseau de neurones convolutif. Dans une autre approche en trois étapes, [22] crée un descripteur qui contient la luminance de trois images consécutives, ainsi que les images compensées en mouvement et l'amplitude du flux optique de Lucas-Kanade. Dans une deuxième étape, une détection SDI est effectuée et enfin, dans une troisième étape, son résultat et le descripteur sont transmis en entrée d'un CNN. Pour la détection des taches et des rayures, [28] applique une classification avec une architecture CNN encodeur-décodeur, avec une concaténation dans la partie encodeur. Une moyenne spatiale est effectuée en sortie du réseau avant la dernière convolution, suivie d'un seuillage pour détecter les taches. Les rayures, quant à elles, sont détectées a posteriori par fermeture morphologique de la sortie du réseau.

La détection des défauts peut également être présentée comme un problème de segmentation binaire standard. Les réseaux de neurones convolutifs sont utilisés pour cette tâche depuis [12]. Le problème est formulé comme une classification binaire à chaque emplacement de pixel, et le réseau doit produire une carte de probabilité dense du premier plan. Deux caractéristiques principales des architectures de réseau peuvent être trouvées dans l'état de l'art. L'architecture d'auto-encodeur est souvent au cœur du réseau ; elle consiste en un encodeur qui effectue l'extraction de caractéristiques comme un classifieur standard, ainsi qu'un décodeur qui effectue un sur-échantillonnage de l'espace latent avec les cartes de caractéristiques intermédiaires de l'encodeur. Plusieurs variantes de ce concept ont été proposées, la plus populaire étant U-Net [19] pour la segmentation des lésions cutanées. Un deuxième module commun dans l'architecture des réseaux de segmentation est la pyramide spatiale des caractéristiques, introduite pour combiner les caractéristiques du codeur à différentes échelles afin de reconnaître plus efficacement les objets à différentes échelles. Le *pooling* spatial pyramidal est un exemple de cette technique, dans laquelle plusieurs convolutions de taille de noyau variable sont appliquées simultanément à la carte de caractéristiques pour extraire des informations multi-échelles. Cette idée est utilisée dans DeepLab v3+ [3], l'un des réseaux les plus performants, alliée à une architecture d'auto-encodeur. Certains des travaux les plus récents et les plus performants, comme NAS-FPN [5], ont tenté d'apprendre à combiner de manière optimale plusieurs échelles de la carte de caractéristiques.

2.3 FlowNet pour compenser le mouvement

Afin de détecter efficacement les défauts dans les films, nous nous inspirons de FlowNet [4], un réseau de neurones proposé à l'origine pour prédire le flux optique à partir d'une paire d'images consécutives dans une vidéo. Les auteurs de FlowNet ont conçu une architecture de type U-Net capable de trouver des correspondances entre les *patches* des deux images en utilisant une couche de corrélation. Cette couche de corrélation ne contient aucun paramètre, et convolue simplement les cartes de caractéristiques calculées séparément pour les deux images. Dans notre cas, la couche de corrélation est très utile pour trouver des régions qui n'ont aucune corrélation avec les *patches* voisins des images précédentes et suivantes de la vidéo, indiquant par là-même une forte probabilité qu'elles constituent des défauts. La couche de corrélation devrait également aider à écarter les éléments de la scène qui ressemblent à des défauts en les faisant correspondre dans plusieurs images consécutives.

3 Correl-Net

Comme cela a déjà été indiqué, nous formulons la détection des défauts comme un problème de segmentation binaire.

3.1 Architecture de réseau

L'architecture détaillée de Correl-Net est représentée sur la FIGURE 2. Correl-Net prend en entrée trois images consécutives de taille 512×512 et produit une seule carte de probabilité de même dimension. Globalement, Correl-Net se fonde sur une architecture U-Net simple, à laquelle ont été apportées des modifications. Premièrement, l'encodeur est composé de trois extracteurs de caractéristiques distincts (un pour chaque image d'entrée) qui partagent les mêmes poids. Deuxièmement, des couches de corrélation sont appliquées entre les cartes de caractéristiques des images i et $i-1$, et celles des images $i+1$ et i, dont la sortie est concaténée avec les cartes de caractéristiques de l'image i. Troisièmement, des connexions saute-couche (*skip connections*) sont utilisées pour connecter l'encodeur et le décodeur, et proviennent de l'extracteur de caractéristiques central dans l'encodeur (celui de l'image i).

Les couches de corrélation prennent la forme de celle décrite dans [4]. Dans notre implémentation, nous utilisons les couches de corrélation de Tensorflow avec les hyperparamètres suivants : taille du noyau égale à 3, déplacement maximum de 10, *stride* d'entrée de taille 1 et *stride* de *patch* de taille 1.

3.2 Détails sur l'apprentissage

La fonction de perte pour l'apprentissage du réseau est l'inverse de l'approximation linéaire du coefficient de Dice (ou F1-score), dont l'expression est la suivante :

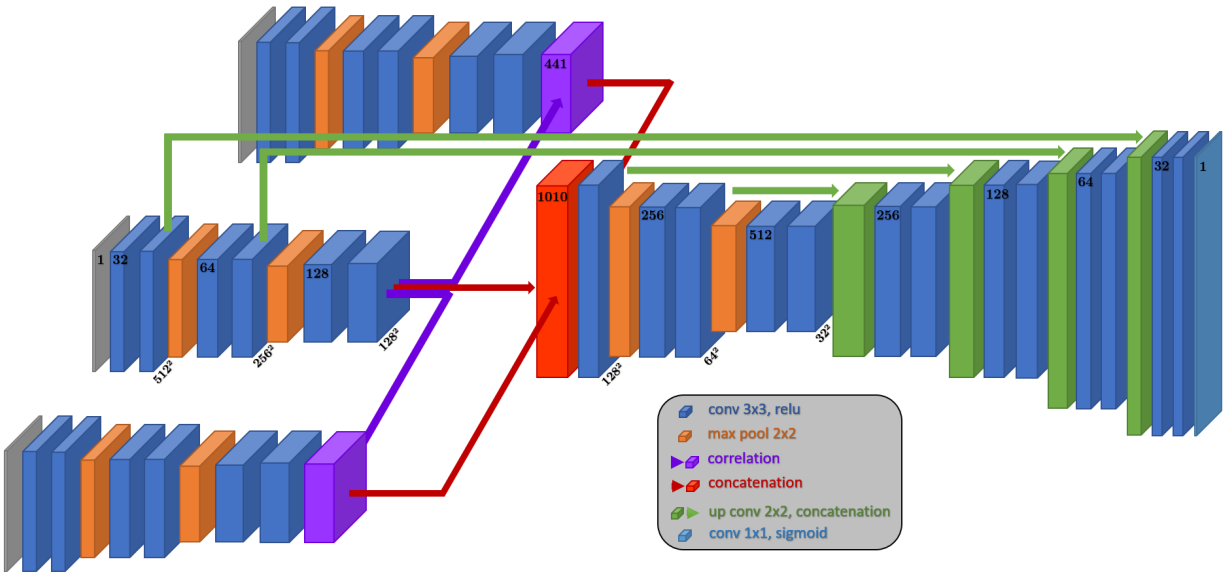


FIGURE 2 – Architecture Correl-Net. L'encodeur est composé de trois branches distinctes qui partagent les mêmes paramètres, une pour chaque image d'entrée (images précédente, courante et suivante). Des couches de corrélation sont utilisées pour comparer les cartes de caractéristiques des branches des images courante et précédente, et des branches des images courante et suivante, dont les sorties sont concaténées avec la carte de caractéristiques de l'image courante. Le décodeur prend la forme de l'architecture classique de type U-Net [19].

$$\text{Perte}\,(y,\hat{y}) = -\frac{2\sum_{i,j} y(i,j)\hat{y}(i,j)}{\sum_{i,j} y(i,j) + \hat{y}(i,j)}$$
$$\approx -\frac{2\times VP}{(VP+FP)+(VP+FN)} \in [-1,0]$$

où $y(i,j) \in \{0,1\}$ et $\hat{y}(i,j) \in [0,1]$ sont respectivement les valeurs du masque de défaut de la vérité terrain et du masque de défaut de sortie du réseau. Les valeurs VP, FP et FN représentent, respectivement, les nombres de pixels comptés comme vrais positifs, faux positifs et faux négatifs. Cette fonction de perte est souvent un bon choix pour la segmentation, dans le cas où il existe un déséquilibre important entre classes [13], ce qui est le cas dans notre application. Le modèle est entraîné avec l'optimiseur Adam, avec un taux d'apprentissage initial de 10^{-5}, et atteint la convergence après 100 époques.

4 Expériences

4.1 Jeu de données

À notre connaissance, il n'existe pas de jeu de données librement accessible de films anciens associés aux masques de segmentation des défauts. Comme nous l'avons expliqué dans l'introduction, la subjectivité d'un restaurateur est un aspect clé du processus de restauration, ce qui signifie qu'il ne peut y avoir de véritable vérité terrain en matière de segmentation des défauts, qui résulterait d'un consensus entre restaurateurs. C'est la raison pour laquelle nous avons décidé de générer des défauts artificiels dont les caractéristiques sont proches de celles des défauts réels, c'est-à-dire avec des tailles, des formes, des couleurs et une transparence aléatoires.

Tous les réseaux ont été entraînés sur l'ensemble de données DAVIS, en utilisant les données *trainVal* comme ensemble d'entraînement, l'ensemble de données *test-dev 2019* comme ensemble de validation et l'ensemble de données *test-challenge 2019* comme ensemble de test. Les défauts ont été générés à la volée pendant l'apprentissage, en utilisant du « bruit fractal », comme cela est décrit sur la FIGURE 3. Le bruit fractal permet, après plusieurs traitements, de produire des taches de différentes tailles dont la forme est très similaire aux défauts réels.

À chaque étape de l'apprentissage, nous divisons les images en niveaux de gris en *patches* de 512×512 pixels. Ensuite, pour chaque *patch*, nous générons aléatoirement un bruit fractal avec les paramètres suivants : forme (512,512), nombre de périodes (8,8), 5 octaves et une lacunarité de 2. Nous obtenons ainsi une matrice de valeurs comprises entre -1 et 1. À partir de cette matrice, nous obtenons alors des défauts foncés en sélectionnant l'intervalle [-0,8 ; -0,65] ainsi que des défauts clairs avec l'intervalle [0,65 ; 0,8]. Comme les défauts réels sont généralement de couleur uniforme, nous appliquons une binarisation, avant de fusionner les défauts à l'image d'entrée. Ce processus est appliqué aux trois canaux du réseau afin d'obtenir trois *patches* consécutifs avec des défauts artificiels différents.

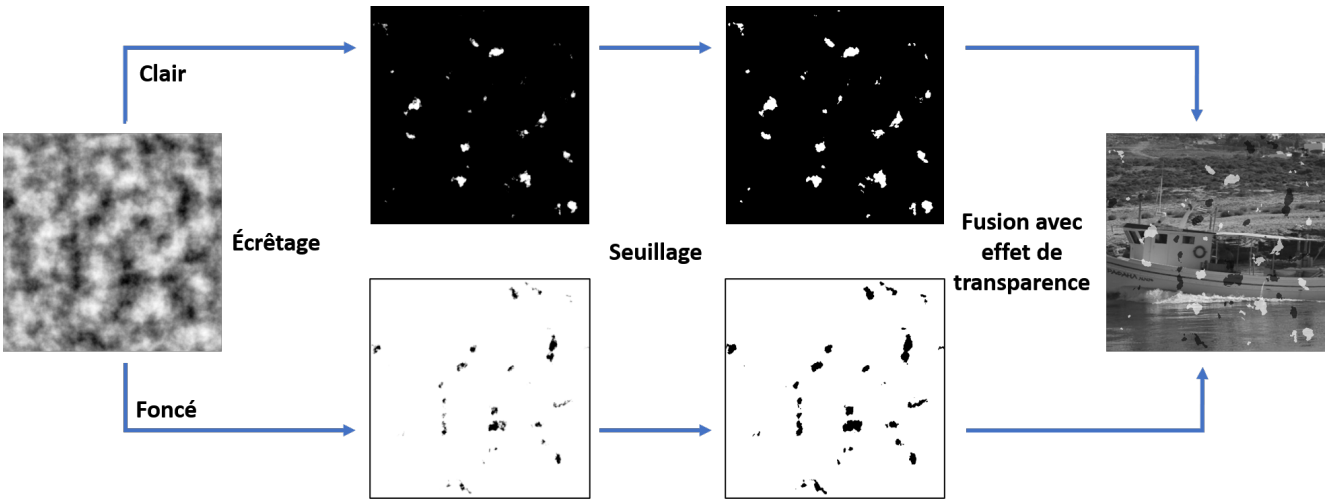


FIGURE 3 – Génération de défauts artificiels. De gauche à droite : génération de bruit fractal ; écrêtage et normalisation pour rassembler les défauts les plus sombres et les plus clairs ; seuillage, puis fusion avec une transparence aléatoire.

4.2 Résultats sur des données synthétiques

Afin d'évaluer quantitativement les performances de notre réseau, nous présentons d'abord les résultats sur le jeu de données de test DAVIS, auquel nous avons ajouté des défauts synthétiques, tel que décrit dans le paragraphe 4.1.

La TABLE 1 présente une comparaison de notre méthode avec plusieurs réseaux de neurones de segmentation classique : un simple U-Net [19], et les architectures plus avancées Deeplab v3+ [3] et NAS-FPN [5], qui ont tous été entraînés et évalués sur les mêmes données que Correl-Net. Les trois images d'entrée consécutives sont regroupées sous la forme d'un tenseur en entrée de ces réseaux.

TABLE 1 – Comparaison des métriques (en pourcentages) pour le jeu de données de test entre cinq réseaux différents.

Réseau	F1-score	Rappel	Précision
U-Net [19]	96,60	99,37	93,99
Deeplab v3+ [3]	94,77	97,47	92,22
NAS-FPN [5]	**98,72**	**99,52**	**97,92**
Tri-Net	98,23	**99,50**	97,00
Correl-Net	**98,90**	99,47	**98,32**

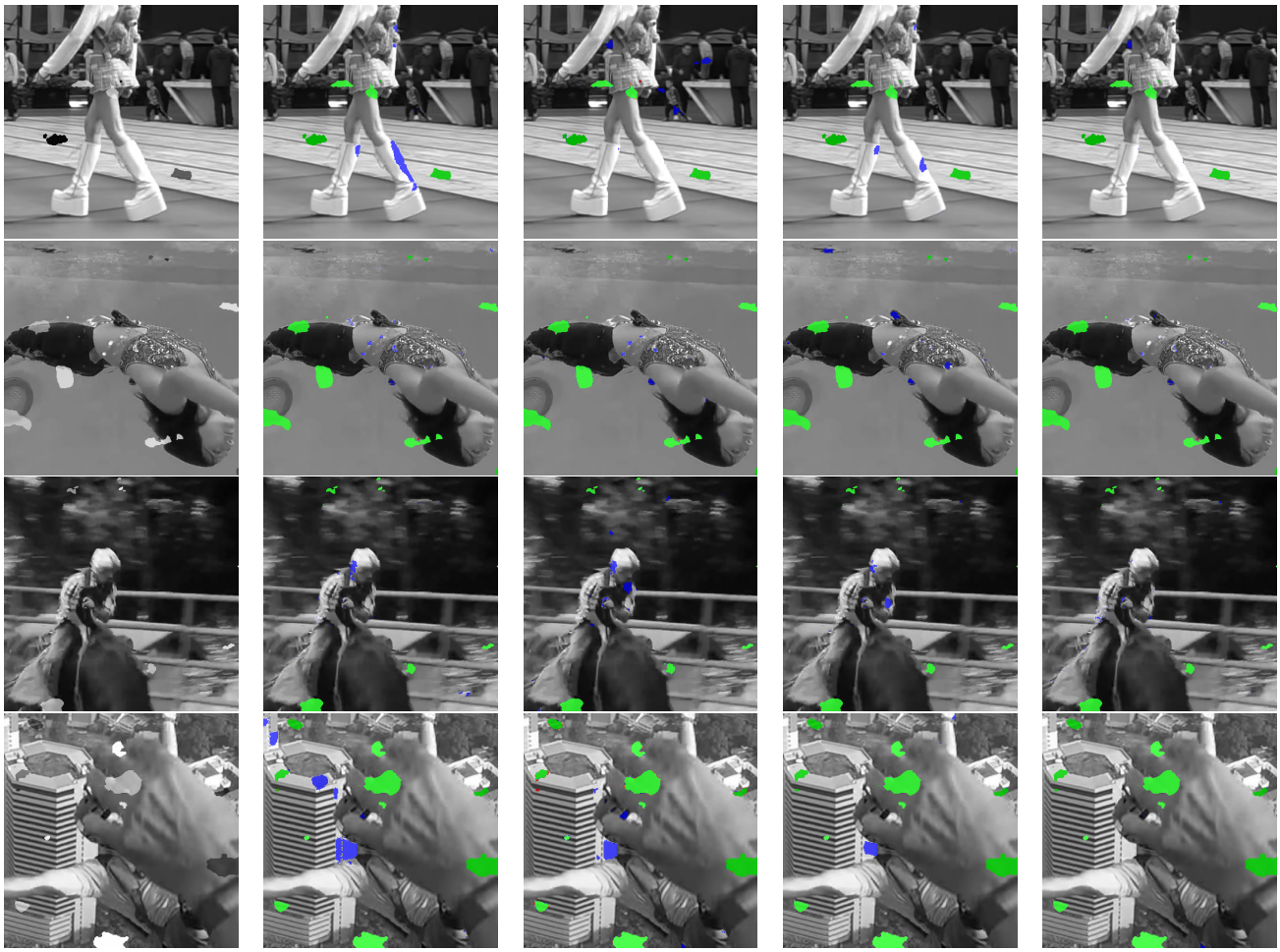
Correl-Net et NAS-FPN obtiennent de meilleures performances globales que U-Net et Deeplab v3+. Il est notable que, de manière inattendue, U-Net obtient de meilleures performances que Deeplab v3+, ce qui est principalement dû à la conception de ce dernier réseau. En effet, la dernière opération du décodeur Deeplab v3+ consiste en un redimensionnement $\times 4$ (en utilisant l'interpolation bilinéaire) pour récupérer une carte de segmentation de sortie de même taille que l'image initiale, sans aucune convolution supplémentaire. Par conséquent, les contours des défauts détectés sont moins précis qu'avec les autres réseaux, ce qui est particulièrement problématique pour les petits défauts.

Correl-Net est légèrement plus performant que NAS-FPN vis-à-vis du F1-score. Cette différence est principalement due à la meilleure précision obtenue par Correl-Net (98,32 contre 97,92). Même si cette différence est faible, elle présente un intérêt particulier par rapport à la tâche puisque, comme indiqué précédemment, les restaurateurs sont soucieux de préserver l'état original du film, ce qui implique d'éliminer les faux positifs manuellement.

La FIGURE 4 montre les défauts détectés sur différentes images de l'ensemble de test, avec un code couleur qui distingue les vrais positifs (vert), les faux positifs (bleu) et les faux négatifs (rouge). Un problème commun à tous les réseaux est que les éléments de l'arrière-plan peuvent être détectés comme des défauts lorsqu'ils sont situés dans des régions qui présentent un mouvement important (voir l'exemple de la séquence « Rodeo » sur la troisième ligne) ou des occultations (par exemple, la séquence « Skydiving-jumping » sur la quatrième ligne). Les objets texturés dans les scènes ont également tendance à provoquer des mauvaises détections (par exemple, le torse de la femme dans la séquence « Mermaid » sur la deuxième ligne). Dans tous ces cas, la couche de corrélation de Correl-Net permet de distinguer les défauts réels des artefacts de mouvement et de texture qui ressemblent à des défauts. Tous les résultats quantitatifs associés à ces séquences sont également reportés dans la TABLE 2.

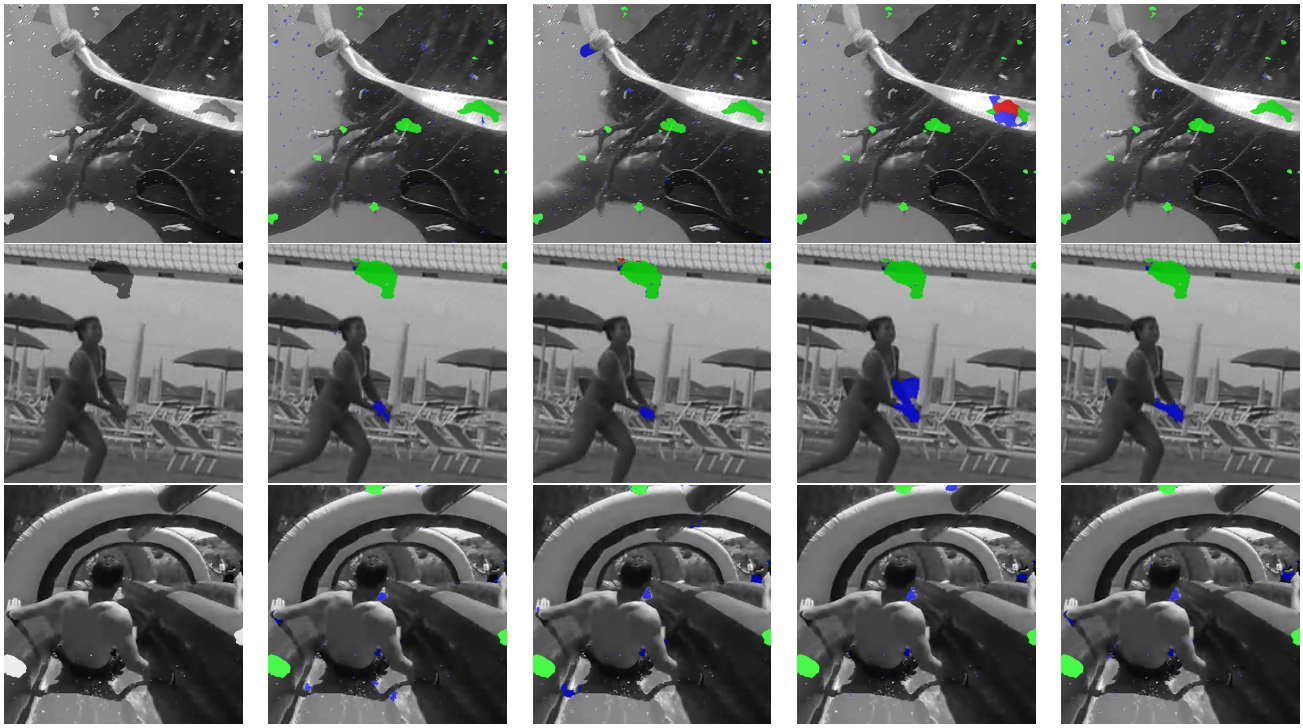
TABLE 2 – Comparaison des métriques pour les séquences « Luggage », « Mermaid », « Rodeo » et « Skydiving-jumping ».

Réseau \ Scène		Luggage	Mermaid	Rodeo	Skydiving-jumping
U-Net [19]	**F1-score**	98,71	99,06	98,25	93,84
	Rappel	**99,60**	99,48	99,54	99,22
	Précision	97,83	98,64	96,99	89,01
Deeplab v3+ [3]	**F1-score**	95,41	96,86	93,03	92,01
	Rappel	97,89	97,97	98,00	96,11
	Précision	93,04	95,77	88,54	88,24
NAS-FPN [5]	**F1-score**	**99,50**	**99,08**	**98,70**	**99,28**
	Rappel	**99,64**	**99,60**	**99,62**	**99,58**
	Précision	**99,36**	**98,58**	**97,79**	**98,98**
Correl-Net [Nous]	**F1-score**	**99,47**	**99,33**	**98,81**	**99,06**
	Rappel	**99,60**	**99,57**	**99,56**	**99,50**
	Précision	**99,34**	**99,09**	**98,08**	**98,62**



(a) Image en entrée. (b) U-Net. (c) Deeplab v3+. (d) NAS-FPN. (e) Correl-Net.

FIGURE 4 – Comparaison visuelle des détections de défauts : les vrais positifs apparaissent en vert, les faux positifs en bleu et les faux négatifs en rouge. De haut en bas, zooms sur les images extraites des séquences « Luggage », « Mermaid », « Rodeo » et « Skydiving-jumping ».



(a) Image en entrée. (b) U-Net. (c) Deeplab v3+. (d) NAS-FPN. (e) Correl-Net.

FIGURE 5 – Comparaison visuelle des détections de défauts défaillantes : les vrais positifs apparaissent en vert, les faux positifs en bleu et les faux négatifs en rouge. De haut en bas, zooms sur les images extraites des séquences « Sea-turtle », « Volleyball-beach » et « Water-slide ».

4.3 Étude d'ablation

Correl-Net présente deux différences fondamentales avec les réseaux de segmentation qui lui sont comparés dans la TABLE 1 : (*i*) les images d'entrée sont traitées dans des branches séparées de l'encodeur avec Correl-Net, alors qu'elles sont empilées et traitées dans une seule branche avec U-Net, Deeplab v3+ et NAS-FPN, et (*ii*) l'utilisation de couches de corrélation est spécifique à Correl-Net.

Afin de quantifier l'impact de ces deux conceptions architecturales, nous définissons un réseau intermédiaire que nous appelons Tri-Net. Tri-Net partage la même architecture globale que Correl-Net (voir FIGURE 2) à l'exception des couches de corrélation qui ne sont pas incluses dans Tri-Net. Les cartes de caractéristiques des trois branches distinctes sont simplement concaténées, sans aucun calcul de corrélation.

La TABLE 1 fournit un aperçu quantitatif de l'impact de ces changements sur les architectures de réseau. Il est intéressant de noter que le rappel reste relativement stable entre U-Net (99,37), Tri-Net (99,50) et Correl-Net (99,47). Le passage d'une seule branche (U-net) à trois branches distinctes (Tri-Net) apporte déjà une amélioration significative de la précision (97,00 contre 93,99). Cela pourrait être dû au fait que la comparaison entre les cartes de caractéristiques des différentes images est effectuée à une résolution inférieure, ce qui permet de compenser les mouvements des petits objets dans la scène.

L'utilisation de couches de corrélation permet d'obtenir une meilleure précision en compensant davantage le mouvement. Le choix de l'hyperparamètre dans la couche de corrélation a un impact important sur le nombre d'opérations en virgule flottante (flops) de notre réseau. En effet, le choix d'une taille de *patch* et d'un déplacement maximal plus grands pourrait potentiellement permettre d'atteindre une plus grande précision pour Correl-Net, mais au prix d'un nombre de flops significativement plus élevé.

4.4 Limitations

Malgré les performances remarquables de Correl-Net, certains cas d'échec sont illustrés sur la FIGURE 5. La première ligne provient d'une séquence sous-marine, dans laquelle les bulles du coin supérieur gauche sont identifiées comme des défauts par tous les réseaux, y compris Correl-Net. La deuxième ligne est issue d'une séquence de sport dans laquelle une grande région (mains et bras du joueur) se déplace rapidement et est détectée comme un défaut. La dernière ligne est un cas compliqué, représentant des mouvements rapides dans une scène avec des variations de lumière et des reflets sur l'eau. De nombreux faux positifs sont détectés par tous les réseaux, y compris Correl-Net.

(a) Image centrale d'un groupe de trois images défectueuses.

(b) Prédiction après avoir appris sur les données synthétiques.

(c) Prédiction après avoir appris sur le film.

(d) Prédiction après avoir effectué une spécification sur le film.

FIGURE 6 – Résultats lors du transfert vers un nouveau jeu de données : les prédictions sont appliquées à l'image centrale (6a). Alors que la prédiction après pré-entraînement sur des données synthétiques ne parvient pas à obtenir un bon rappel (6b), de nombreux faux positifs sont évités lors de la spécification sur le nouveau jeu de données (6d) par rapport à un entraînement direct sur ce nouveau jeu de données sans pré-entraînement (6c) (voir la jambe du cycliste ou les cailloux sur la route).

4.5 Spécification sur des défauts réels

Dans ce paragraphe, nous considérons un autre ensemble de données provenant d'un film ancien qui a été restauré manuellement par un expert indépendant. Nous obtenons la vérité terrain des masques de défauts en comparant l'image restaurée manuellement avec l'image originale. Malheureusement, ces masques de défauts peuvent difficilement être considérés comme une vérité terrain, car dans de nombreux cas, le processus de restauration a altéré des pixels qui ne se trouvent pas sur les défauts réels mais dans des régions voisines. Nous pouvons néanmoins utiliser ces données pour une analyse qualitative de Correl-Net.

Notre pré-entraînement auto-supervisé sur des données vidéo augmentées de défauts synthétiques permet à Correl-Net de bien apprendre ce qu'est une anomalie temporelle. Cependant, lorsque le même réseau est utilisé pour détecter des défauts dans un film ancien, les résultats ne sont pas satisfaisants (voir FIGURE 6b). Le rappel est très faible, ce qui est probablement dû aux raisons suivantes : le grain de l'image est différent de celui du jeu de données DAVIS, la couleur et la forme des défauts sont distribuées différemment des défauts synthétiques.

Ce jeu de données contient 3000 images, que nous avons divisées en ensembles d'apprentissage (80% des images), de validation (10%) et de test (10%). Lors de l'entraînement de Correl-Net sur ces images uniquement (sans pré-entraînement auto-supervisé), le rappel s'améliore à nouveau, mais la précision reste très faible (voir FIGURE 6c), ce qui est dû à la mauvaise qualité des masques de vérité terrain.

D'autre part, pour l'adapter à un nouveau film, nous proposons d'affiner un Correl-Net qui a été pré-entraîné sur des données synthétiques. La FIGURE 6d montre les résultats que nous obtenons après une seule époque de spécification sur ce nouveau jeu de données, avec un taux d'apprentissage de 10^{-6}. Nous obtenons bien moins de faux positifs par rapport à l'entraînement direct sur le film.

5 Conclusion

Dans cet article, nous présentons Correl-Net, une architecture de réseau de neurones conçue pour détecter efficacement les défauts dans les vidéos. Correl-Net utilise des couches de corrélation précédemment introduites dans FlowNet [4] pour discriminer avec précision les défauts réels des artefacts dus au mouvement de la caméra. Nous montrons que Correl-Net obtient des résultats légèrement meilleurs que le réseau de segmentation NAS-FPN [5]. Plus précisément, il atteint une plus grande précision grâce à la couche de corrélation, ce qui est souhaitable pour un restaurateur. Bien que Correl-Net soit entraîné sur un grand ensemble de données vidéo augmenté de défauts synthétiques de manière auto-supervisée, nous montrons également qu'il peut bénéficier d'un processus de spécification pour être plus efficace lorsqu'il est appliqué à un nouveau film.

En plus des résultats présentés dans les paragraphes précédents, il est important de noter que Correl-Net est un réseau beaucoup plus simple (9 millions de paramètres, 189 Gflops) que NAS-FPN (485 millions de paramètres, 666 Gflops), ce qui signifie que Correl-Net a besoin de moins de mémoire et de moins de ressources pour fonctionner.

Bien que Correl-Net présente encore quelques limites, par exemple pour des objets en mouvement rapide ou des reflets lumineux, nous pensons qu'il constitue une aide significative pour le travail du restaurateur et une étape importante vers une restauration semi-automatique efficace des films anciens. Dans un futur proche, nous prévoyons de mener une étude auprès d'experts en restauration, afin de valider la qualité de notre détection de défauts sur une plus grande variété de films, ainsi que de comparer notre détection de défauts aux outils commerciaux utilisés par les restaurateurs. Nous voulons évaluer si le passage à Correl-Net permettrait effectivement d'aider les restaurateurs, par exemple en diminuant la quantité d'édition manuelle nécessaire à la correction des défauts détectés.

Références

[1] J. BIEMOND, P. M. B. van ROOSMALEN et R. L. LAGENDIJK : Improved Blotch Detection by Postprocessing. *In Proceedings of the ICASSP*, vol. 6, p. 3101–3104, 1999.

[2] A. BUSCHE : Just Another Form of Ideology ? Ethical and Methodological Principles in Film Restoration. *The Moving Image*, 6(2):1–29, 2006.

[3] L.-C. CHEN, Y. ZHU, G. PAPANDREOU, F. SCHROFF et H. ADAM : Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *In Proceedings of ECCV*, p. 801–818, 2018.

[4] A. DOSOVITSKIY, P. FISCHER, E. ILG, P. HAUSSER, C. HAZIRBAS, V. GOLKOV, P. VAN DER SMAGT, D. CREMERS et T. BROX : Flownet: Learning Optical Flow with Convolutional Networks. *In Proceedings of ICCV*, p. 2758–2766, 2015.

[5] G. GHIASI, T.-Y. LIN et Q. V. LE : NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. *In Proceedings of CVPR*, p. 7036–7045, 2019.

[6] S. IIZUKA et E. SIMO-SERRA : Deepremaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. *ACM Transactions on Graphics*, 38(6):1–13, 2019.

[7] L. JOYEUX, O. BUISSON, B. BESSERER et S. BOUKIR : Detection and Removal of Line Scratches in Motion Picture Films. *In Proceedings of CVPR*, vol. 1, p. 548–553, 1999.

[8] K.-T. KIM et E. Y. KIM : Automatic Film Line Scratch Removal System Based on Spatial Information. *In Proceedings of ISCE*, p. 1–5, 2007.

[9] A. C. KOKARAM : Detection and Removal of Line Scratches in Degraded Motion Picture Sequences. *In Proceedings of ESPC*, p. 1–4, 1996.

[10] A. C. KOKARAM, R. D. MORRIS, W. J. FITZGERALD et P. J. RAYNER : Detection of Missing Data in Image Sequences. *IEEE TIP*, 4(11):1496–1508, 1995.

[11] A. C. KOKARAM et P. J. RAYNER : System for the Removal of Impulsive Noise in Image Sequences. *In Proceedings of VCIP*, vol. 1818, p. 322–331, 1992.

[12] J. LONG, E. SHELHAMER et T. DARRELL : Fully Convolutional Networks for Semantic Segmentation. *In Proceedings of CVPR*, p. 3431–3440, 2015.

[13] F. MILLETARI, N. NAVAB et S.-A. AHMADI : V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *In Proceedings of 3DV*, p. 565–571, 2016.

[14] M. J. NADENAU et S. K. MITRA : Blotch and Scratch Detection in Image Sequences Based on Rank Ordered Differences. *In Time-Varying Image Processing and Moving Object Recognition, 4*, p. 27–35. Elsevier, 1997.

[15] A. NEWSON, A. ALMANSA, Y. GOUSSEAU et P. PÉREZ : Temporal Filtering of Line Scratch Detections in Degraded Films. *In Proceedings of ICIP*, p. 4088–4092, 2013.

[16] A. NEWSON, P. PÉREZ, A. ALMANSA et Y. GOUSSEAU : Adaptive line scratch detection in degraded films. *In Proceedings of CVMP*, p. 66–74, 2012.

[17] P. PHILIPPOT : Historic Preservation: Philosophy, Criteria, Guidelines. *In Preservation and Conservation: Principles and Practices. Proceedings of the North American Regional Conference*, p. 367–382, 1976.

[18] A. RENAUDEAU, F. LAUZE, F. PIERRE, J.-F. AUJOL et J.-D. DUROU : Alternate Structural-Textural Video Inpainting for Spot Defects Correction in Movies. *In Proceedings of SSVM*, p. 104–116, 2019.

[19] O. RONNEBERGER, P. FISCHER et T. BROX : U-Net: Convolutional Networks for Biomedical Image Segmentation. *In Proceedings of MICCAI*, p. 234–241, 2015.

[20] T. K. SHIH, L. H. LIN et W. LEE : Detection and Removal of Long Scratch Lines in Aged Films. *In Proceedings of ICME*, p. 477–480, 2006.

[21] R. SIZYAKIN, N. GAPON, I. SHRAIFEL, S. TOKAREVA et D. BEZUGLOV : Defect Detection on Videos using Neural Network. *In Proceedings of the International Scientific-Technical Conference "Dynamic of Technical Systems"*, vol. 132 de *MATEC Web of Conferences*, p. 05014, 2017.

[22] R. SIZYAKIN, V. VORONIN, N. GAPON, M. PISMENSKOVA et A. NADYKTO : A Blotch Detection Method for Archive Video Restoration using a Neural Network. *In Proceedings of ICMV*, vol. 11041, p. 110410W, 2019.

[23] P. C. USAI : *Silent Cinema: A Guide to Study, Research and Curatorship*. Bloomsbury Publishing, 2019.

[24] X. WANG et M. MIRMEHDI : Archive Film Defect Detection and Removal: an Automatic Restoration Framework. *IEEE TIP*, 21(8):3757–3769, 2012.

[25] R. XU, X. LI, B. ZHOU et C. C. LOY : Deep Flow-Guided Video Inpainting. *In Proceedings of CVPR*, p. 3723–3732, 2019.

[26] Z. XU, H. R. WU, X. YU et B. QIU : Features-Based Spatial and Temporal Blotch Detection for Archive Video Restoration. *Journal of Signal Processing Systems*, 81(2):213–226, 2015.

[27] H. YOUS et A. SERIR : Blotch Detection in Archived Video Based on Regions Matching. *In Proceedings of ISIVC*, p. 379–383, 2016.

[28] H. YOUS, A. SERIR et S. YOUS : CNN-Based Method for Blotches and Scratches Detection in Archived Videos. *Journal of VCIR*, 59:486–500, 2019.