

Utilisation des images satellites pour améliorer le recensement des logements à Mayotte

Maëlys Bernard
Insee
maelys.bernard@insee.fr

Raya Berova
Insee
raya.berova@insee.fr

Thomas Faria
Insee
thomas.faria@insee.fr

2025-07-17

Abstract Malgré leur disponibilité de plus en plus accrue, les images satellites très haute résolution sont jusqu'ici très peu utilisées au sein de la statistique publique. La manipulation de ces données non conventionnelles n'appartient pas encore aux pratiques usuelles des statisticiens publics, bien qu'elles offrent des opportunités significatives pour améliorer ou créer de nouveaux indicateurs statistiques. Cet article vise justement à démontrer qu'il est possible d'intégrer ces données satellitaires dans les processus statistiques afin d'assurer leur qualité. Mayotte constitue un cas emblématique où les difficultés d'accès au terrain et la forte dynamique de développement urbain complexifient l'organisation des enquêtes cartographiques qui permettent de repérer les logements en amont du recensement. Afin d'orienter les travaux de contrôle de la qualité de ces enquêtes, nous avons exploré l'utilisation des images satellitaires très haute résolution Pléiades (0,5 m de résolution spatiale), en combinaison avec des méthodes d'apprentissage profond, spécifiquement via des modèles de segmentation sémantique. La méthodologie développée consiste en l'entraînement d'un modèle SegFormer, fondé sur une architecture de type Transformer, largement reconnue dans la littérature puisqu'elle est à la base des modèles GPT. Le modèle a été entraîné grâce aux annotations issues du projet CoSIA de l'IGN, qui fournit une référence précise de la couverture du sol. À partir des prédictions du modèle, il est possible de produire des estimations précises des surfaces bâties sur un territoire donné, permettant ainsi d'analyser efficacement leur évolution dans le temps. Les résultats obtenus démontrent une performance élevée dans la détection automatisée des évolutions urbaines dans les DROM. Ces résultats sont disponibles sous la forme d'un tableau de bord interactif permettant aux équipes opérationnelles de cibler efficacement les zones prioritaires pour la réalisation d'enquêtes cartographiques de contrôle sur le terrain.

Après avoir décrit la méthode de prédiction de surface du bâti sur les images satellites, cet article présentera un cas pratique d'utilisation de ces prédictions à travers le contrôle

qualité des enquêtes cartographiques de la population à Mayotte. Depuis 2021, le recensement à Mayotte fonctionne sur un cycle quinquennal comme en métropole et dans les autres DOM. En amont de la collecte auprès des habitants, une enquête cartographique permet de repérer les logements. Le résultat de l'enquête cartographique détermine la base d'immeubles au sein de laquelle est sélectionné l'échantillon¹ de l'enquête annuelle de recensement. Ainsi, si l'enquête cartographique ne couvre pas bien l'intégralité du bâti du territoire à enquêter, les populations et nombre de logements déduits de l'enquête annuelle de recensement seront sous-estimés du fait de ce défaut de couverture. Les données satellites permettent de contrôler la qualité des enquêtes cartographiques en comparant l'évolution de la couverture de bâti avec celle du nombre de logements observé suite aux enquêtes cartographiques au sein de chaque îlot² à enquêter une année donnée. Les îlots sont répartis en cinq groupes de rotation correspondant à l'année à laquelle a lieu l'enquête annuelle de recensement. Afin d'assurer la qualité des enquêtes cartographiques, un deuxième passage sur le terrain a été planifié dans les grandes communes mahoraises pour certains îlots identifiés comme les plus prioritaires. Cet ordre de priorité a été défini par une méthode mobilisant les images satellites pour repérer les plus susceptibles d'être concernés par des manques potentiels dans les enquêtes cartographiques. L'enquête cartographique corrective ciblant ces îlots a permis d'assurer une meilleure couverture de ces enquêtes et d'éviter un risque de sous-estimation des logements.

Table des matières

1 Introduction	3
2 Méthodologie	4
2.1 Les modèles de d'apprentissage profond pour l'analyse d'images	4
2.1.a Réseaux de neurones convolutifs	4
2.1.b Segmentation sémantique	8
2.2 Données	9
2.2.a Images satellites	9
2.2.b Annotations	11
2.3 Evaluation	11
3 Résultats	12
3.1 Evaluation des performances	12
3.2 Mise à disposition pour les statisticiens	12
3.2.a Une application interactive	12
3.2.b Vers de nouveaux indicateurs statistiques ?	12
4 Cas d'usage	13
Bibliographie	14

1 Introduction

Reprendre l'abstract soumis pour faire l'intro Reprendre les slides du séminaire à la DIRAG pour le contexte « Momentum de Varsovie » Repartir de l'origine du projet (demande DROM Mayotte/ Guyane) Parler de l'IGN et de leur travaux avec leur dataset FLAIR et leur modèle CoSIA

2 Méthodologie

Nous allons dans cette partie revenir sur les concepts d'apprentissage profond (*deep learning*) afin d'expliquer notre démarche et l'application de ces méthodes à nos problématiques précises. Nous définirons ensuite les données que nous avons utilisées avant de détailler les méthodes d'évaluation que nous avons appliqués à nos algorithmes.

2.1 Les modèles de d'apprentissage profond pour l'analyse d'images

Dans le domaine de l'apprentissage statistique, les réseaux de neurones profonds en sont un sous-domaine. Si les premiers algorithmes de réseaux de neurones ont été développés dès les années 1950, les réseaux de neurones profonds ont eux fait leur révolution dans les années 2010 grâce notamment au développement de nouvelles cartes graphiques (GPU) et d'algorithmes optimisés pour celles-ci. En effet, leur développement a longtemps été freiné par la quantité de données nécessaires pour obtenir de bons résultats impliquant donc des temps de calculs très longs et des coûts d'annotations prohibitifs. La conception des réseaux de neurones profonds est inspirée du fonctionnement du cerveau humain. En effet, l'idée est de représenter les neurones par des fonctions mathématiques très simples qui s'activent si le signal reçu en entrée est suffisamment fort et transmettent alors leur « activation » dans ce cas. Un grand nombre de neurones correctement organisés permet alors de réaliser des opérations plus complexes, résultantes des différentes activations.

Si l'apprentissage profond est devenu la solution de facto pour l'analyse d'images par ordinateur, et notamment les tâches de segmentation sémantique c'est en partie grâce aux travaux fondateurs de Long et al. (2015) qui ont développé les réseaux entièrement convolutifs, une extension de l'architecture des réseaux de neurones convolutif proposé par LeCun et al. (1989).

2.1.a Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) tirent leur inspiration du cortex visuel des animaux et se composent de deux parties (voir Figure 1) :

1. Une première partie composée par des couches de convolutions successives qui permet d'extraire des prédicteurs de l'image en entrée du réseau;
2. Une seconde partie permettant de classifier les pixels de l'image en entrée à partir des prédicteurs obtenus dans la première partie.

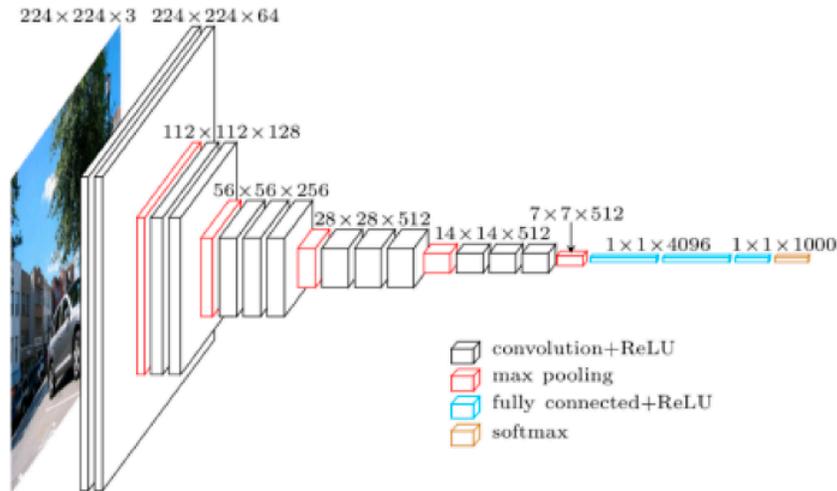


Figure 1. – Réseau de neurones convolutif

Note de lecture : La partie convolutive est représentée par les différents cubes à gauche et la partie classifiante est représentée par les rectangles fins et bleus à droite

Dans la partie convolutive, l'image en entrée du réseau se voit appliquer des filtres appelés convolutions, représentés par des matrices $A = (a_{ij})$ de petite taille appelés noyaux de convolution. L'image en sortie de l'opération de convolution est obtenue à partir de chaque pixel de l'image en entrée en calculant la somme des pixels avoisinant, pondérée par les coefficients a_{ij} du noyau de convolution. Cette opération de convolution est illustrée dans la Figure 2, extraite de l'ouvrage Kim (2017).

$$\begin{array}{|c|c|c|c|} \hline 1 & 1 & 1 & 3 \\ \hline 4 & 6 & 4 & 8 \\ \hline 30 & 0 & 1 & 5 \\ \hline 0 & 2 & 2 & 4 \\ \hline \end{array} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{array}{|c|c|c|} \hline 7 & 5 & 9 \\ \hline 4 & 7 & 9 \\ \hline 32 & 2 & 5 \\ \hline \end{array}$$

Figure 2. – Opération de convolution

Note de lecture : L'image résultante est une image plus petite dont les pixels sont égaux à la somme des pixels de l'image en entrée pondérée par les coefficients du noyau de convolution.

Une convolution vise ainsi à résumer l'information présente dans une région de l'image. Il est d'ailleurs intéressant de remarquer que lorsque le voisinage d'un pixel ressemble à la structure du noyau de convolution, la valeur de sortie est élevée. Ainsi, si le filtre est, par exemple, un détecteur de ligne verticale, il génèrera une grande valeur sur les pixels appartenant à des lignes verticales. Un filtre de convolution permet donc d'obtenir une sorte de carte de caractéristiques de l'image. Une couche convolutive correspond finalement à l'application d'un nombre n_f de filtres différents en parallèle. Ainsi, en sortie d'une couche, il y a autant d'images (carte de caractéristiques) que de filtres appliqués. Ces n_f cartes deviennent les canaux d'entrée de la couche suivante.

Les CNN se composent de couches successives, chacune ayant un rôle spécifique. Les premières couches détectent des motifs simples dans l'image (bords, textures, lignes horizontales ou verticales...) tandis que les couches intermédiaires et profondes, plus spécialisées, combinent ces motifs simples pour repérer des formes de plus en plus complexes (motifs, objets, visages...). La Figure 3 schématise un réseau de neurones convolutif.

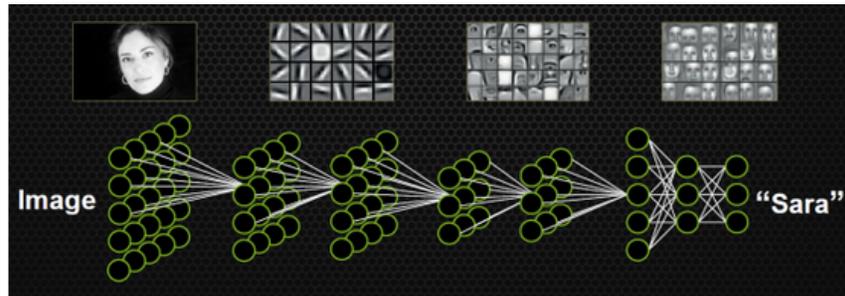


Figure 3. – Représentation d'une convolution

Note de lecture : Les premières couches convolutives reconnaissent les formes simples tandis que les couches les plus profondes à droite reconnaissent des formes plus concrètes.

Après une couche de convolution, la taille des cartes de caractéristiques reste identique à celle de l'image, ce qui peut rapidement entraîner un volume de données important et rendre le modèle coûteux à entraîner. Pour pallier cela, une méthode couramment utilisée est le *max pooling*. Cette opération, représenté par la Figure 4 consiste à diviser chaque carte de caractéristiques en petites régions, puis à ne conserver, pour chacune, que la valeur maximale. Le *max pooling* permet ainsi de réduire la dimension spatiale des données tout en conservant l'information la plus pertinente. Cette réduction favorise également une certaine invariance locale aux petites translations ou déformations de l'image, et contribue à limiter le nombre de paramètres et la complexité de calcul du réseau, tout en mettant en valeur les motifs les plus discriminants. D'autres opérations telles que le padding détaillé dans la Figure 5 permettent malgré tout de contrôler la dimension de l'image en sortie.

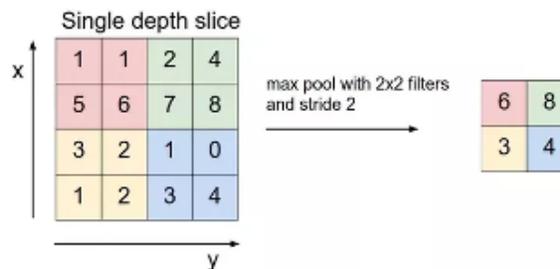


Figure 4. – Représentation du max pooling

Note de lecture : L'image résultante de l'opération de max-pooling est égale au maximum de chaque pixel dans chacune des zones dessinées sur l'image de gauche.

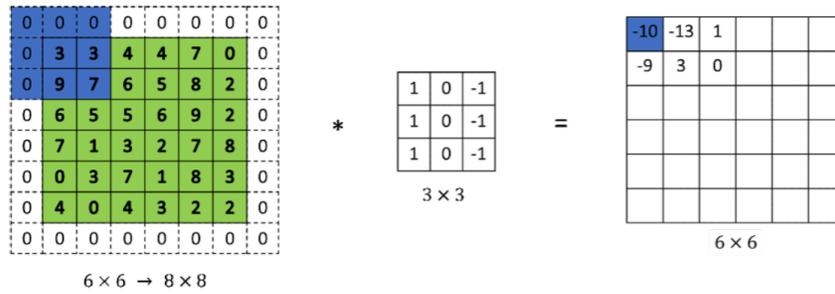


Figure 5. – Représentation du padding

Note de lecture : l'ajout d'une bande de 0 autour de l'image avant d'appliquer la convolution permet de limiter la réduction de dimension de l'image résultante.

Une fonction d'activation non linéaire, comme la ReLU (Rectified Linear Unit), est appliquée systématiquement après chaque couche de convolution (cf. Figure 6). Inspirée du fonctionnement des neurones biologiques, cette opération introduit une non-linéarité dans le réseau. Sans cette étape, l'empilement de couches convolutives ne ferait qu'appliquer une combinaison linéaire de filtres, ce qui limiterait la capacité du modèle à apprendre des relations complexes. L'ajout de fonctions d'activation non linéaires permet ainsi au réseau d'extraire et de modéliser des motifs variés et des structures non linéaires présentes dans les données, ce qui est essentiel pour traiter efficacement des images. L'utilisation de fonctions d'activation non linéaires ne se limite pas aux réseaux convolutifs, c'est justement une composante fondamentale de l'ensemble des architectures d'apprentissage profond.

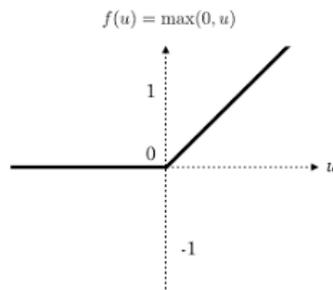


Figure 6. – Fonction d'activation ReLU

Note de lecture : la fonction ReLU « s'active » quand l'argument est positif.

Contrairement aux approches classiques du machine learning, la particularité des réseaux convolutifs réside dans l'automatisation de l'extraction des caractéristiques (*feature extraction*) de l'image. Dans un cadre traditionnel, cette étape est généralement réalisée de manière manuelle ou repose sur l'expertise métier : on choisit alors à l'avance quels prédicteurs utiliser (par exemple, des statistiques sur certaines bandes spectrales ou des filtres définis à la main). À l'inverse, dans un réseau de neurones convolutif, ce sont les poids des noyaux de convolution eux-mêmes qui sont appris automatiquement au cours de l'entraînement. Ces coefficients font partie du vecteur de paramètres du modèle, noté θ . Ainsi, une fois le réseau entraîné, l'ensemble des filtres optimaux θ^* représente la meilleure façon de transformer et d'extraire les informations pertinentes

de l'image. Les cartes de caractéristiques produites par l'enchaînement de ces filtres alimentent ensuite la partie classifiante du réseau. Cette approche permet au modèle de découvrir de manière autonome les motifs les plus discriminants pour la tâche visée, sans intervention manuelle dans la conception des prédicteurs.

En sortie d'un réseau de neurones convolutif classique, la partie dite classifiante permet d'associer une catégorie à l'image analysée. Par exemple, on peut vouloir déterminer si une image représente un chien ou un chat. Pour cela, la sortie de la partie convolutive (c'est-à-dire les cartes de caractéristiques extraites par les filtres) est d'abord transformée en un seul vecteur grâce à une opération de mise à plat (*flattening*). Ce vecteur est ensuite traité par un réseau de neurones dense, appelé aussi *fully connected*, qui produit en sortie un vecteur de scores $\mathbf{x} = (x_0, \dots, x_9)$ lorsqu'on cherche à classer l'image parmi 10 classes possibles.

Pour interpréter ces scores comme des probabilités, on applique la fonction *softmax* :

$$\text{Softmax}(x_i) = \frac{\exp x_i}{\sum_{j=0}^9 \exp x_j}.$$

Cette opération convertit les scores en une distribution de probabilité, dont la somme vaut 1, et permet de sélectionner la classe associée à la probabilité la plus élevée comme prédiction finale. La fonction *softmax* est particulièrement utile car elle est infiniment dérivable, ce qui garantit la dérivabilité de tout le modèle f_{θ} par rapport à ses paramètres θ . Cela rend possible l'utilisation de la descente de gradient pour ajuster les poids du réseau au cours de l'apprentissage.

2.1.b Segmentation sémantique

L'objectif premier de notre projet est d'être capable d'identifier des zones de bâtis à partir d'images satellites. Les récents développements en apprentissage profond ont permis d'obtenir de très bons résultats en terme d'analyse d'image par ordinateur sur diverses tâches comme la classification d'images, la détection d'objets, le suivi d'objets ou encore la segmentation sémantique. En l'occurrence, c'est cette dernière méthode qui va nous intéresser puisque les algorithmes de segmentation sémantique consistent à classer les différents pixels d'une image dans diverses classes pré-définies. L'idée est donc détecter automatiquement les pixels qui réfèrent à une zone de bâtis et ainsi être capable de délimiter au mieux les frontières entre les bâtiments et les autres classes.

La segmentation sémantique moderne repose historiquement sur les réseaux de neurones convolutionnels (CN) et utilise généralement des techniques d'apprentissage supervisé nécessitant des jeux de données étiquetés au niveau pixel [6][7]. Les modèles sont entraînés pour apprendre des caractéristiques hiérarchiques complexes permettant de distinguer les différentes classes sémantiques présentes dans les images.

Si l'apprentissage profond est devenu la solution de facto pour les tâches de segmentation sémantique c'est en parti grâce aux travaux fondateurs de Long, Shelhamer, & Darrell (2015) qui

Définition, qu'est ce que c'est ? les modèles originaux (FCN) et Unet Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015

puis l'évolution récente vers les transformers (ViT)

Dans ce contexte on a choisi Segformer. Rappeler qu'on utilise des modèles pré-entraîner qu'on finetune simplement avec les données qu'on a notre dispo (voir partie suivante)

De la segmentation sémantique à la détection de changement => approche temporelle Mentionner les modèles de change détection et dire les différences avec notre approche ->

2.2 Données

2.2.a Images satellites

Une image satellite est une matrice à trois dimensions. Chaque élément de cette image est un **pixel**, qui correspond à une surface au sol (par exemple 10m*10m). Le pixel contient plusieurs valeurs numériques. Ces valeurs expriment l'intensité du rayonnement solaire reflété dans chaque **bande spectrale** pour ce pixel. Une bande spectrale correspond à une portion du spectre électromagnétique, qui peut être par exemple le bleu, le rouge, le vert, le proche infrarouge. Donc, pour une image satellite de 200*200 pixels avec les trois bandes du visible (rouge, vert, bleu), chaque pixel aura trois valeurs différentes représentant l'intensité dans chacune des bandes du visible. Ainsi, un pixel qui représente 10m*10m au sol donnera une couleur : du violet sur l'exemple de la Figure 7.

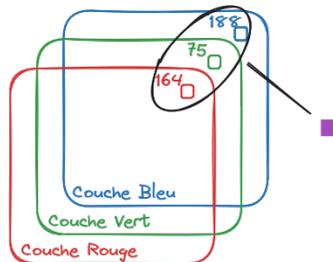


Figure 7. – Exemple d'un pixel d'une image satellite

Il existe de nombreux produits satellitaires disponibles. Nous nous sommes surtout concentrés sur deux d'entre eux : **Pléiades et Sentinel2**.



Figure 8. – Mamoudzou, Mayotte (2024)

Les images satellites **Pléiades** constituent une ressource précieuse dans notre cas d’usage. Ce sont des images à très haute résolution, spécifiquement conçues pour l’observation fine des territoires. Elles offrent trois bandes spectrales dans le visible (**rouge, vert, bleu**) auxquelles s’ajoute une quatrième bande dans le proche infrarouge (NIR), bien que cette dernière ne soit pas disponible dans nos données actuelles. Leur résolution spatiale remarquable de **0,5m** par pixel permet une détection très précise des objets au sol, ce qui est essentiel pour nos traitements d’analyse fine.

Ces images peuvent être obtenues soit via les **archives gratuites** (sous conditions d’accord), soit par acquisition à la demande, un service payant encadré par une licence Airbus©, avec des délais d’environ 6 à 8 mois par département.

Dans le cadre du cyclone Chido qui a frappé Mayotte en décembre 2024, un plan d’urgence a permis l’accès gratuit à une mosaïque d’images Pléiades post-cyclone couvrant l’île. Une **mosaïque** est une image composite constituée d’assemblages de prises de vues réalisées à différentes dates. Elle permet d’obtenir une couverture complète, avec un minimum de zones nuageuses et une meilleure homogénéité visuelle. Ce travail de reconstitution, mené par **l’Institut National de l’information géographique et forestière (IGN)**, est crucial pour garantir des données exploitables, notamment dans le cadre de l’entraînement de modèles d’analyse automatique.

Mais Pléiades n’est pas notre seule source d’imagerie satellite. Les satellites **Sentinel-2**, développés par l’Agence spatiale européenne (ESA) dans le cadre du programme Copernicus, offrent une alternative open source, particulièrement intéressante pour les analyses à large échelle ou à forte fréquence temporelle. Contrairement à Pléiades, Sentinel-2 capte **treize bandes spectrales**, réparties entre le visible, le proche infrarouge (NIR) et l’infrarouge à ondes courtes (SWIR), ce qui

ouvre la voie à une multitude d'indices et d'analyses thématiques (comme la détection de la végétation, de l'humidité, ou des matériaux).

En revanche, la résolution spatiale de Sentinel-2 est moindre : selon les bandes, elle varie entre **10 m**, 20 m et 60 m, ce qui rend l'observation de petits objets ou de détails fins plus difficile. Néanmoins, ces images ont l'avantage d'être acquises automatiquement **tous les cinq jours**, garantissant une fréquence de revisite élevée et une mise à disposition régulière des données, y compris en période de crise.

Ainsi, les images Pléiades semblent être les plus adaptés pour de la détection de bâtiment dans les DROM.

- définition
- produits
- acquisition/livraison (partenariat IGN a promouvoir)

2.2.b Annotations

- Rappeler que c'est le plus important pour avoir de bons modèles
- Nous n'avons rien annoté manuellement => full automatique
- Rappel du coût d'annotation pour en avoir de qualité (en ETP ?)
- RIL (Insee),
- BDTOPO (better)
- CoSIA

peut être rappeler la difficulté d'avoir des couples qui sont iso temporel

2.3 Evaluation

- IOU, Loss, zone de test manuel (calculer la taille des zones de test)
- inférence (parler moyenne mobile) -> mise à dispo d'une API

3 Résultats

3.1 Evaluation des performances

Metriques et images

inférence à grande échelle et robustesse des résultats (inférence avec fenetre mouvante) parler des ressources computationnelles requises

3.2 Mise à disposition pour les statisticiens

3.2.a Une application interactive

Pour les enquêteurs

3.2.b Vers de nouveaux indicateurs statistiques ?

Fichiers parquet d'évolution de bâti

4 Cas d'usage

Bibliographie

- Kim, P. (2017). Convolutional Neural Network. In P. Kim (éd.), *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence* (p. 121-147). Apress. https://doi.org/10.1007/978-1-4842-2845-6_6
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541-551. <https://doi.org/10.1162/neco.1989.1.4.541>
- Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully Convolutional Networks for Semantic Segmentation*. <https://arxiv.org/abs/1411.4038>