SOTA REPORT

VISION ASSISTÉE PAR ORDINATEUR
ET
INTELLIGENCE ARTIFICIELLE

Référence du document:

Document - Rev.- Date – Description	**Par**
SOTA REPORT-Rev.-00-030718- Vision Assistée par Ordinateur et Intelligence Artificielle	Johann VO VAN

SOMMAIRE

1 RÉSUMÉ

L'objet de ce document est de fournir un état de l'art de la Vision Assistée par Ordinateur, en fournissant un panel non exhaustif des dernières méthodes, approches, techniques et algorithmes, impactant déjà notre société, ou susceptible de le faire à court terme.

Dans un premier temps, les techniques générales de Vision Assistée par Ordinateur sont présentées : le traitement numérique des données et ce qui les caractérise. L'Intelligence Artificielle sur laquelle s'appuie cette discipline est introduite par une description des différents apprentissages machine utilisés et les jeux de données associés.

Les principales fonctionnalités de la vision par ordinateur sont ensuite abordées sous la forme d'un état de l'art des applications et travaux de recherche s'y rapportant. Parmi ces travaux, les réseaux de neurones et leurs récentes variantes utilisés dans le cadre de la vision par ordinateur sont présentés : ce qui les définissent, leurs avantages, inconvénients et leurs limitations.

Pour finir, nous aborderons les futurs axes de recherche et les défis à venir de cette discipline, ainsi que les enjeux et risques potentiels de dérives liés à l'usage de l'Intelligence Artificielle.

Mots clés : Computer Vision, Intelligence Artificielle, Apprentissage Machine, réseaux de neurones, jeux de données

2 INTRODUCTION

La ''Computer Vision'' ou Vision assistée par ordinateur est une discipline scientifique visant à donner aux machines la capacité de voir, en leur permettant ainsi d'analyser visuellement leurs environnements et d'interpréter les informations qu'ils contiennent. Ce processus implique l'évaluation et la compréhension d'une image, d'images ou de vidéos. Le terme compréhension contraste avec la définition mécanique de la vision, mettant en avant la complexité du domaine de la Computer vision.

En effet, une véritable compréhension de notre environnement ne peut être obtenue à travers une représentation visuelle seule. En réalité, des signaux visuels sont acheminés à travers le nerf optique jusqu'au cortex visuel primaire, pour être finement interprétés par le cerveau. Les interprétations tirées de ces informations sensorielles mobilisent une grande partie des outils naturels de notre intellect, et modèle la quasi-totalité de nos expériences subjectives. De ce point de vue, la vision n'est qu'une transmission d'images pour une interprétation, tandis que le traitement de ces images est plutôt de l'ordre de la pensée ou du cognitif. Il fait alors appel à une multitude de facultés de notre cerveau. Ainsi, nombreux sont ceux qui partagent l'idée que la Computer Vision ouvre la voie à de future itérations d'Intelligence Artificielle puissante, à travers les nombreux domaines qu'elle impacte.

De nombreux chercheurs considèrent les approches d'Intelligence Artificielle et plus particulièrement celles basées sur les réseaux de neurones à convolution (CNN) comme faisant partie de la Computer Vision. Les variantes de CNN restent le cœur des architectures de nouveaux réseaux de neurones appliqués à des taches de vision. La recherche les intégrant continuellement, telles des briques, appuyées par la puissance des informations ''open-source'' et d'apprentissage profond ou ''Deep learning''. Cependant ce champ fascinant de la science n'en est encore qu'au stade embryonnaire.

Le ''Deep Learning'' introduit des modèles mathématiques contenant de multiples couches de calculs pour apprendre et représenter la donnée à de multiples niveaux d'abstraction. Il singe ainsi la compréhension et la perception d'informations multimodale par notre cerveau en saisissant implicitement les structures intriquées dans un grand volume de données. Le ''Deep

Learning'' est une riche famille de méthodes, incluant les CNN, les modèles hiérarchiques probabilistes, et une variété d'algorithmes d'apprentissages supervisés et non supervisés. Le récent engouement pour ces méthodes est dû au fait qu'elles ont surpassé les méthodes précédentes de l'état de l'art. Cela, dans différentes tâches, et dans des volumes de données importants, complexes et de différentes sources. Un des facteurs les plus importants ayant contribué à l'important essor du ''Deep Learning'' est l'apparition de jeux de donnés volumineux, annotés, et de qualité. Supportés par les calculs parallèles des processeurs graphiques, ces jeux de données permettent une accélération significative de l'apprentissage des modèles.

L'objet de ce document est de fournir un état de l'art de la Computer Vision, en fournissant un panel non exhaustif des dernières méthodes, approches, techniques et algorithmes, impactant déjà notre société, ou susceptible de le faire à court terme.

3 TECHNIQUES GÉNÉRALES DE LA COMPUTER VISION

3.1 TRAITEMENT DE DONNEES NUMERIQUES

La Computer Vision utilise à la base le traitement de données numériques sur des objets d'intérêts en vue d'obtenir une représentation. Le but principal de la représentation de ces données est de décrire quantitativement des objets et des concepts complexes, essentiels pour l'analyse quantitative dont ils feront l'objet. La sélection appropriée des données caractéristiques et le bon choix sur la manière de les visualiser et de les comparer entre elles, correspondent à la formulation d'hypothèses sur les données disponibles, et leur variabilité. Le plus souvent, une segmentation des objets d'intérêt est effectuée. De là, il est considéré que ces objets ont été obtenus, reste à se concentrer sur ce que l'on peut en faire.

3.2 DONNEES CARACTERISTIQUES

Le but du traitement des données numériques caractéristiques est d'envoyer les objets d'intérêt vers un espace vectoriel à plusieurs dimensions. C'est une solution parmi d'autres. Une première possibilité qui mène habituellement à des données caractéristiques interprétables, est de bâtir soit même des objets et un jeu de données caractéristiques associées, spécialement adaptés à la problématique. Une autre stratégie souvent utilisée en conjonction avec la première est de traiter des quantités mathématiques plus complexes qui, tout en étant plus difficile à interpréter, permettront d'exprimer un plus large éventail de comportements des objets étudiés.

Il existe aussi une approche à plusieurs niveaux où les données caractéristiques traitées en provenance d'objets d'intérêts bruts étudiés, sont utilisées pour alimenter un système d'apprentissage supervisé.

Une manière de visualiser les données dans cet espace vectoriel, est un nuage de point à n dimensions. Souvent, certaines données ne sont pas utiles, ou trop bruyantes, ou redondantes. Il est donc très souvent possible de trouver un espace vectoriel aux dimensions inférieures, avec une perte moindre ou contrôlée d'information.

3.3 L'APPRENTISSAGE MACHINE

L'Apprentissage Machine ou ''Machine learning'', fait référence aux techniques originellement issues de l'intelligence artificielle. Il vise à apprendre des règles à partir de données, permettant typiquement des prédictions à partir de nouveaux points de données. Cela se chevauche clairement avec l'inférence statistique. Le ''Machine Learning'' tend à se concentrer sur les résultats et l'efficacité, parfois au détriment d'une compréhension rigoureuse des méthodes. Les statistiques descriptives décrivent et caractérisent les échantillons, tandis que l'inférence statistique tente de déduire les propriétés de la distribution dont elles proviennent.

Concernant le ‘’Deep Learning’’, les données caractéristiques apprises sont celles qui se différencient le mieux de modèles d’apprentissage. Il est intéressant de noter qu’elles sont intrinsèquement «multi-niveaux». Les données caractéristiques sont considérées comme des variables aléatoires, les valeurs suivent une distribution spécifique, l’ensemble des données étant une variable aléatoire multivariée, et chaque objet à l'étude peut être considéré comme une réalisation particulière de cette variable aléatoire.

L'apprentissage supervisé, part d’un ensemble d'objets modélisés et de leurs données caractéristiques associées et des essais, et de là, en déduit une règle d’apprentissage. Cette règle permettra la prédiction d’un modèle inconnu issu d'un nouvel ensemble de données caractéristiques. Quant à l'apprentissage non supervisé, celui-ci vise à déterminer la structure des données en connaissant uniquement leurs caractéristiques. Cela, en regroupant ensemble des objets similaires. L'algorithme le plus utilisé est le regroupement hiérarchique. Celui-ci utilise les distances entre les points de données pour regrouper les points proches, en commençant par les plus proches, et en bâtissant une arborescence.

Les autres tâches d’apprentissages classiques incluent la régression, des méthodes semi-supervisées, et des méthodes génératives.

4 ETAT DE L’ART DES FONCTIONNALITÉS DE BASE

4.1 LE PROCESSUS DE CLASSIFICATION

La tâche de classification, lorsqu’il s’agit d’image, consiste généralement à annoter l’image. Une fois l’assignation d’un label réalisé, la localisation consiste à définir où l’objet se situe dans ladite image. Dans un problème de classification, la cible représente une classe, et l’entrée associée, une instance de cette classe. Les classes sont en nombre fini, l’espace d’entrée n’est pas nécessairement borné. Généralement, aucune relation entre les classes n’est connue a priori. On utilise des algorithmes de classification lorsque la cible est de nature discrète.

Concernant l’état de l’art de la classification, le concours ImageNet Large Scale Visual Recognition (ILSVRC) évalue les algorithmes pour la détection d’objet et la classification d’image à grande échelle. Les points forts de ce concours ImageNet LSVRC en 2017 [1] :

- Concours de classification de scène: BDAT gagne avec 73% de précision moyenne, avec des modèle Deep de convolution combinés et une méthode d’Attention Adaptative.
- Concours de classification et localisation ImageNet: NUS-Qihoo_DPNs (CLS-LOC) gagne avec 3.41% le top-5 erreur de classification et 6.22% erreur de localisation en utilisant un DPN (Dual Path Network) avec du Fast-RCNN Trimps-Soushen est second de très près dans la Localisation avec 6.49% en utilisant une nouvelle architecture Faster-RCNN, des cascades RPN/RCNN, des entrainement multi-échelles, et des méthodes FPN/Mask et RCNN/Deformable.

4.2 LE PROCESSUS DE DETECTION

La détection d’objet est le processus consistant à détecter des instances d’objet sémantiques d’une classe particulière dans des images et vidéos numériques. Une approche classique pour des systèmes de détection d’objet inclut la création d’un grand jeu de fenêtres candidates qui sont conséquemment classifiées via un CNN.

Un grand nombre de travaux sont basés sur le concept des régions des CNN [2]. Dans ce concept, le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Les neurones de cette région du cerveau sont arrangés de sorte qu'ils correspondent à des régions qui se chevauchent lors du pavage du champ visuel.

Leur fonctionnement consiste en un empilage multicouche de perceptrons, dont le but est de prétraiter de petites quantités d'informations. Ces approches sont précises en termes de détection mais parfois approximatives dans la localisation. Il existe ainsi un nombre importants de méthodes alternatives tentant de les améliorer, notamment en déterminant avec exactitude la position de l'objet [3]. Dans ce but, ces méthodes utilisent souvent conjointement la détection d'objet avec une segmentation sémantique [4-5]. Une vaste majorité de travaux sur la détection d'objet utilisant le ''Deep Learning'' applique des variantes de CNN. Cependant il existe un nombre assez restreint de méthodes de détection utilisant d'autres modèles profonds dont :

- Deep Belief Network pour la reconnaissance d'objet 3D, dans lequel le niveau haut du modèle est une machine de Boltzmann de troisième ordre, entrainé via un algorithme hybride qui combine des gradients génératifs et discriminants [6].

- Méthode d'empilement d'auto-encodeur pour détection multiples. L'objectif d'un auto-encodeur est d'apprendre une représentation d'un ensemble de données, généralement dans le but de réduire la dimension de cet ensemble. Récemment, le concept d'auto-encodeur est devenu plus largement utilisé pour l'apprentissage de modèles génératifs.

Une des tendances majeures de 2018 dans la détection d'objet est une évolution notable vers des systèmes de détection plus rapides et plus efficaces. Notamment à travers des implémentations telles que R-FCN, SSD et YOLO allant toutes vers un partage des calculs sur l'image entière. Ils se différencient ainsi des sous réseaux de neurones couteux associés à des techniques dite Fast R-CNN. Cependant ces dernières techniques restent très efficaces et sont toujours abondamment utilisées pour la détection d'objet.

Les implémentations de détection d'objets majeures ces dernières années sont: « SSD: (Single Shot MultiBox Detector) » [7], « YOLO V3 » [8-9-10], « Réseaux pyramidaux » [11] et « R-FCN: Object Detection via Region-based Fully Convolutional Networks » [12].

Une publication de Huang et al. (2016) [13] compare en détail les performances entre R-FCN, SSD et Faster R-CNN. Ces architectures sont vues comme des "méta architectures" puisqu'elles peuvent être combinées avec différents extracteurs de données caractéristiques tel que Resnet ou Inception. La tendance à implémenter des détections d'objets économiques et efficaces, tout en conservant la précision nécessaire aux applications commerciales, notamment les applications de véhicules autonomes, est également mis en exergue par les algorithmes SqueezeDet [14] et PVANet [15].

COCO 36 (Common Objects in Context) est un autre jeu de données populaire, il est cependant comparativement plus petit et moins précis que d'autre jeux tel que ImageNet. Les résultats de Détection aux deux concours en 2017 d'ILSVRC et de COCO [16] sont :

- ImageNet LSVRC Detection d'Objet sur Video : IC&USYD 81.7%
- ImageNet LSVRC Object Detection d'Objet sur Video avec Poursuite: IC&USYD 64.14%
- COCO 2017 Detection d'Objet (bounding boxes): IC&USYD 81.9%

4.3 LE PROCESSUS DE SEGMENTATION

La Segmentation est un processus central dans la Computer Vision, qui divise une image entière en groupe de pixels qui peuvent ensuite être annotés et classifiés. En sus, la Segmentation Sémantique va plus loin en essayant de comprendre sémantiquement le rôle de chaque pixel dans l'image. La Segmentation par Instance va encore plus loin en segmentant différentes instances de classes. C'est une des difficultés que rencontre les applications de Computer Vision qui sont utilisées dans les suites technologiques de véhicules autonomes.

Parmi les plus importants progrès et développements sur le sujet on peut citer les solutions de FAIR (Facebook Artificial Intelligence Research) qui continue d'améliorer leurs travaux « Deepmask » depuis 2015. Ils génèrent des masques préliminaires sur les objets comme une forme initiale de segmentation. Les travaux et solutions notoires ces dernières années sur le sujet sont : En 2016, FAIR introduit « SharpMask » qui corrige les pertes de détails des masques « Deepmask » [17], « MultiPathNet » [18] et « Video Propagation Networks » [19].

A partir de 2016, les chercheurs ont exploré des configurations de réseaux alternatives pour résoudre ces problèmes d'échelle et de localisation. DeepLab [20] a ainsi obtenu des résultats encourageants pour des taches de Segmentation Sémantiques d'images. Khoreva et al. (2016) [21] se sont appuyés sur les précédents travaux de Deeplab issus courant 2015. Ils proposent une méthode d'entrainement supervisé hebdomadaire qui obtient des résultats comparable à des réseaux entièrement supervisés.

La Computer Vision a ensuite participé à l'amélioration des approches de partage de l'information utiles par les réseaux. Cela, par l'utilisation de réseaux dit "end to end" qui réduisent les besoins de calculs de multiples sous taches omnidirectionnelles pour la classification. Deux publications clés utilisant cette approche sont : « Layers Tiramisu » [22] et « Fully Convolutional Instance-aware Semantic Segmentation » [23].

4.4 LE PROCESSUS DE POURSUITE

Il s'agit du processus de poursuite d'un objet d'intérêt spécifique, ou de multiples objets, dans une scène donnée. Cette technique a traditionnellement des applications dans la vidéo et les interactions dans le monde réel. Les observations sont faites en suivants un objet initialement détecté. Ce processus est crucial par exemple dans les systèmes de véhicules autonomes.

Parmi les derniers travaux importants sur le sujet nous pouvons citer : « Fully-Convolutional Siamese Networks for Object Tracking » [24], « Learning to Track at 100 FPS with Deep Regression Networks » [25], «Deep Motion Features for Visual Tracking » [26], « Virtual Worlds as Proxy for Multi-Object Tracking Analysis » [27] et « Globally Optimal Object Tracking with Fully Convolutional Networks » [28].

4.5 TRAVAUX SUR RÉSOLUTION

Cette technique se réfère au processus d'estimation ou de prédiction d'image de haute résolution à partir d'une image de basse résolution. Celle-ci réalise également la prédiction de particularité d'images à différents agrandissements, ce qu'un cerveau humain peut faire quasiment sans effort. Initialement la Super-Résolution était réalisée par des techniques simples comme les interpolations bicubiques et les méthodes de plus proches voisins.

En termes d'applications commerciales, le besoin de surmonter les contraintes liées aux basses résolutions issues de sources de basse qualité et la réalisation d'images améliorées a conduit la recherche dans ce domaine. Ci-après les avancées sur le sujet ces dernières années : « Neural Enhance » [29], « Real-Time Video Super Resolution », « RAISR: Rapid and Accurate Image Super-Resolution » [30], « SRGAN » [31] et Amortised « MAP Inference for Image Super-resolution » [32].

4.6 LE TRANSFERT DE STYLE

Le transfert de style est un sujet assez intuitif dès lors qu'il est visualisé. Pour cela, prenez une image et imaginez là avec des caractéristiques stylistiques d'une autre image, tel que le style d'un peintre ou artiste célèbre.

Le transfert de Style est une technique ancienne mais converti en réseaux de neurones en 2015 avec la publication de « A Neural Algorithm of Artistic Style ». [33] Depuis, le concept du

Transfert de Style a été étendu par Nikolin et Novak [34] et également appliqué aux vidéos [35], une évolution commune dans la Computer Vision. Le Transfert de Style illustre parfaitement une nouvelle utilisation des réseaux de neurones, descendu dans le domaine publique. Notamment à travers les intégrations ces dernières années de Facebook, ainsi que celle de compagnies comme Prisma [36] et Artomatix [37].

En 2017, Facebook a sorti CAffe2Go [38], leur système de Deep Learning qui s'intègre dans les équipements mobiles. Google a également produit des travaux intéressants recherchant à mélanger de multiples styles pour créer des styles d'image totalement uniques comme dans leur Research blog [39] et leur publication complète sur le sujet [40]. Le Transfert de Style trouve également des applications dans la création de jeux vidéo en facilitant la tâche des artistes au niveau des textures et autres.

4.7 TRAVAUX SUR LA COLORISATION

La colorisation est le processus visant à changer des images monochromes en version multi-couleurs. Initialement cela était fait manuellement par des personnes qui sélectionnaient méticuleusement les couleurs de pixels spécifiques dans chaque image. A partir de 2016 il est devenu possible d'automatiser ce processus en conservant l'apparence réaliste typique des colorisations manuelles. La technique automatisée de colorisation est intéressante dans la mesure où le réseau de neurones assigne les couleurs les plus pertinentes pour une image. Ils se basent alors sur sa compréhension de la localisation de l'objet, sa texture et son environnent. Trois des travaux les plus influents ces dernières années sont :

- Zhang et al ont produit une méthode capable de tromper avec succès des humains sur 32% de leurs essais. Leur méthodologie est comparable à un « Test de Turing de Colorisation » [41].
- Larsson et al ont entièrement automatisé leur système de colorisation en utilisant le Deep Learning pour l'estimation d'Histogram [42].
- Lizuka, Simo-Serra and Ishikawa ont également trouvé un modèle de colorisation basé sur un réseau de neurones. Les résultats de leurs travaux surpassent l'état de l'art [43].

5 ETAT DE L'ART DES FONCTIONNALITÉS AVANCÉES

5.1 LA RECONNAISSANCE D'ACTIONS HUMAINES

La compréhension de scènes s'inspire de la vision cognitive et nécessite l'association d'au moins trois domaines : la vision par ordinateur, les sciences cognitives et le génie logiciel.

La compréhension de scènes peut atteindre cinq niveaux de fonctionnalités génériques de vision par ordinateur : la détection, la localisation, le suivi, la reconnaissance et la compréhension. Toutefois, les systèmes de compréhension de scènes vont au-delà de la détection de caractéristiques visuelles. La Reconnaissance d'Action exige également des fonctionnalités de vision par ordinateur plus robustes, plus solides et flexibles, en leur donnant une compétence cognitive : la capacité d'apprendre, de s'adapter, d'évaluer les solutions alternatives et de développer de nouvelles stratégies pour l'analyse et l'interprétation.

Récemment sont apparus des algorithmes qui peuvent prédire les suites probables d'interactions en se basant sur seulement quelques trames avant que l'action ait lieu. A cet égard nous constatons de récentes tentatives dans la recherche pour intégrer le contexte dans les algorithmes de décisions. Parmi ces recherches, ci-dessous quels approches notables [44-45]:

- Utilisation combinée de moteurs de reconnaissances et des méthodes apprentissage supervisées et non supervisées, sur des vidéos 2D et 3D, afin de comprendre les

comportements humains dans un grand nombre d'environnements et pour différentes applications [46].

- Détection et reconnaissance d'actions complexes dans des séquences vidéos: en utilisant des "Saliency maps" (forme de segmentation) pour détecter et localiser des évènements, puis du ''Deep learning'' est appliqué aux données pré-entrainées pour identifier les trames les plus importantes correspondant à l'évènement sous-jacent [47].
- Incorporation de vaste marge comme terme de régulation dans des modèles profonds de réseaux de neurones améliorant les performances de généralisation pour la classification [48].
- Utilisation de réseau de neurones comme un modèle combinant l'extraction et la classification pour des activités détaillées, utilisant des fonctionnalités profondes apprises depuis ImageNet et un classificateur à machine à vecteur de support. De par l'adaptabilité du modèle et la disponibilité d'une variété de données de différents capteurs, ce type de fusion de données caractéristiques multimodales devient une stratégie de plus en plus populaire [49].
- Combinaison de données caractéristiques hétérogènes pour des détections d'évènements complexes. En premier lieu, les données caractéristiques les plus informatives pour la reconnaissance d'évènements sont estimées, et ensuite les différentes données caractéristiques sont combinées en utilisant une structure de graph ET/OU [50].
- Fusion multimodale avec une architecture de réseau de convolution et LSTM [51].
- Réseau de DBN utilisant des séquences vidéo et autres informations [52].

Ci-dessous de récentes publications clés sur le sujet: « Long-term Temporal Convolutions for Action Recognition » [53], « Spatiotemporal Residual Networks for Video Action Recognition » [54], « Anticipating Visual Representations from Unlabeled Video » [55].

5.2 LA GESTION D'OBJETS 3D

Dans la Computer Vision, la classification des scènes, objets et activités, et la segmentation d'images en deux dimensions sont à la base de nombreuses nouvelles recherches. Cependant pour que des systèmes interprètent correctement le monde réel et puissent y naviguer, une compréhension en 3D est impérative. Transmettre ces représentations 3D et le savoir qui leur est associé aux systèmes artificiels est un des grands défis à relever pour la Computer vision.

Des travaux sur le sujet apparaissent comme notoires. Ils couvrent un large spectre d'étude allant du stade embryonnaire des premières applications théoriques des futures systèmes d'Intelligence Artificielle Générale (IAG) [56] et de robotiques, jusqu'aux applications d'immersions captivantes de réalité augmentée, virtuelle et mixte qui affectent déjà notre société. Ainsi, il n'est pas déraisonnable de prévoir une croissance exponentielle de ce domaine dans la Computer Vision. Croissance résultant d'applications commerciales lucratives, laissant à penser que bientôt les ordinateurs pourront commencer à raisonner sur le monde plutôt que sur de simples pixels.

Ci-après quelques travaux récents et importants sur le sujet : « OctNet: Learning Deep 3D Representations at High Resolutions » [57], « ObjectNet3D: A Large Scale Database for 3D Object Recognition » [58], « 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction » [59], « 3D Shape Induction from 2D Views of Multiple Objects » [60], et « Unsupervised Learning of 3D Structure from Images » [61].

5.3 L'ESTIMATION DE POSTURE HUMAINE ET LA DETECTION DE POINTS CLES

L'estimation de posture humaine tente de trouver l'orientation et la configuration des différentes parties du corps humain. L'estimation de posture humaine en 2D ou la détection de points clés, se réfère généralement à la localisation des parties du corps. L'estimation de posture 3D va encore plus loin en trouvant l'orientation des parties du corps dans un environnement 3D.

Ensuite, une étape optionnelle de Reconstruction peut intervenir pour estimer ou modeler leurs formes. De très nombreux progrès sont intervenus dans ces sous domaines avancés de la Computer Vision ces dernières années.

L'estimation de posture humaine est une tâche très difficile. Cela, en raison du nombre très vaste de silhouettes et d'apparences humaines, des contextes d'illuminations difficiles, ou d'arrière-plans encombrés. Avant l'ère du ''Deep Learning'', l'estimation de pose était basée sur la détection de partie du corps à travers des structures picturales [62].

Concernant les méthodes de ''Deep Learning'' pour l'estimation de posture, celles-ci peuvent être groupées en deux catégories. Les méthodes holistiques et les modèles « part-based » de type constellations, selon la façon dont les images d'entrée sont traitées. Les méthodes de traitement holistiques tendent à accomplir leurs tâches d'une façon globale. Elles ne définissent pas explicitement un modèle pour chaque partie individuelle et leurs relations spatiales. D'un autre côté, les méthodes de traitement « part-based » se concentrent sur la détection individuelle des parties du corps. Elles sont suivies de modèle graphiques pour incorporer l'information spatiale. Parmi ces travaux de recherches, ci-après quelques approches notables :

- Utilisation de patchs locaux et des patchs d'arrière plans pour entrainer un CNN, afin d'apprendre les probabilités conditionnelles de la présence des parties du corps et leur relations spatiales [63].
- Entrainer de multiples petits CNN pour effectuer des classifications binaires indépendantes des parties du corps, suivi à plus haut niveau d'un modèle spatial pour filtrer les fortes anomalies et imposer une posture globale cohérente [64].
- Utilisation d'un CNN multi résolution pour obtenir une « heat-map » de régression probable pour chaque partie du corps, suivi d'un modèle graphique implicite pour obtenir des articulations cohérentes [65].

Les récentes publications clés sur le sujet: « Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. » [66] et « Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. » [67].

5.4 LE PROCESSUS DE RECONSTRUCTION

Les chapitres précédents ont présentés des exemples de Reconstruction mais concentrées sur les objets, et plus spécifiquement leurs formes et leurs postures. Tandis que certaines de ces méthodes sont techniquement des Reconstructions, ce domaine comprend de nombreux autres différents types de Reconstruction. Des Reconstructions de scènes, des Reconstructions simple vue, ou multi-vues, des Reconstructions par le mouvement SFM (Structure From Motion), ou enfin des algorithmes de localisation par rapport à un modèle simultanément reconstruit : SLAM (Simultaneous Reconstruction and Mapping). Les techniques de SFM ou SLAM consistent à localiser un système de perception par rapport à une carte de l'environnement continuellement mise à jour.

Cette approche permet l'exploration d'environnements inconnus. La carte est régulièrement actualisée pour intégrer les nouveaux éléments perçus. La pose du système est calculée en comparant les données observées avec les données déjà présentes dans la carte. Ainsi, les processus de localisation et de mise à jour de la carte sont effectués simultanément de manière dépendante. La localisation utilise le résultat de la cartographie qui a nécessité la connaissance de la pose courante, pour géo-référencer les nouvelles données dans la carte existante.

Les publications suivantes présentent un grand nombre d'approches pour créer des Reconstruction de haute-fidélité en temps réel : « Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera » [68], « Unsupervised CNN for Single View Depth Estimation:

Geometry to the Rescue » [69], « IM2CAD» [70], « Deep Image Homography Estimation » [71], « gvnn: Neural Network Library for Geometric Computer Vision » [72].

Le SLAM est l'une des plus grandes problématiques géométriques. Certains chercheurs commencent à considérer qu'elle pourrait être le prochain problème résolu par le Deep Learning. Le cœur de l'estimation géométrique faisant partie du SLAM est plutôt bien résolu par les approches actuelles. Mais les sémantiques de haut niveau, et les couches basses des composants du système peuvent tous bénéficier du Deep Learning.

En particulier celui-ci peut grandement améliorer la qualité des cartes sémantiques. Comme par exemple dépasser la posture ou le nuage de points, pour une compréhension complète des différents types d'objets ou de régions sur la carte. Comme par exemple une meilleure gestion des objets dynamiques et des changements d'environnement. A plus bas niveau, des améliorations peuvent se faire sur plusieurs composants. Comme notamment la reconnaissance de lieux [73], la détection de fin de boucle, la relocalisation, de meilleurs points descripteurs. En résumé, il est raisonnable de penser que le SLAM ne sera pas complètement remplacé par le ''Deep learning''. Cependant il est très probable que les deux approches deviennent complémentaires.

6 LES RÉSEAUX DE CONVOLUTION

6.1 ARCHITECTURES

Les architectures de réseaux de neurones à convolution sont au premier plan de la Computer Vision. Des avancées dans les architectures, amènent ainsi des améliorations dans la vitesse, la précision et l'entrainement de beaucoup des applications déjà mentionnées dans le présent document.

Le terme de neurone désigne une unité de calcul. Cette unité prend une entrée, c'est à dire un vecteur dont la taille est celle de l'entrée. Elle lui applique ensuite une transformation linéaire dont les variables sont des paramètres du neurone. Après cette transformation linéaire, on applique généralement une fonction dite « d'activation » non linéaire. L'entraînement du neurone se fait par rétro-propagation à partir de la sortie finale du réseau de neurones. Une évaluation de la qualité de la décision finale par rapport à la décision attendue est effectuée. Il s'agit alors de minimiser la distance entre les deux. Pour ce faire, on peut par exemple procéder par descente de gradient. En effet, pour chaque paramètre, on calcule la dérivée partielle du score obtenu. On ajuste alors en conséquence les paramètres du réseau jusqu'à obtenir un résultat satisfaisant.

Les CNN sont inspirés des structures du système visuel animal [74]. Un CNN est constitué de trois type de couches de neurones principales: les couches convolutées, les couches de mises en commun ou POOL, les couches entièrement connectées FC [75]. Chaque couche joue un rôle différent [76]. Les CNN ont été efficaces et ont rencontrés beaucoup de succès dans les applications de Computer Vision, comme la reconnaissance faciale, la détection d'objet et les véhicules autonomes.

Ci-après quelques architectures récentes notables, dont beaucoup s'inspirent du récent succès de Resnets faisant l'objet du chapitre 6 : « Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning » [77], « Densely Connected Convolutional Networks » [78], « FractalNet Ultra-Deep Neural Networks without Residuals » [79], et « Lets keep it simple: using simple architectures to outperform deeper architectures » [80]

Ci-après quelques techniques additionnelles qui complètent les architectures de CNN:
« Swapout: Learning an ensemble of deep architectures » [81] et « SqueezeNet » [82].

6.2 VERS UNE EQUIVARIANCE DES CNN

Les CNN sont invariants en translation, ils peuvent identifier la même caractéristique dans de multiple partie de l'image. Cependant, les CNN typiques ne sont pas invariants en rotation. Si une caractéristique ou l'image entière subissent une rotation, alors les performances du réseau s'amoindrissent. Généralement, les CNN apprennent à quelque peu gérer l'invariance en rotation via une augmentation des données. Le réseau obtient alors des propriétés d'invariance en rotation, sans conception spécifique à cet égard. L'invariance en rotation est fondamentalement limitée dans les réseaux utilisant les techniques actuelles. Les publications ci-après surmontent ces problèmes d'invariance de rotation pour les CNN.

Chaque approche porte une innovation qui lui est propre. Elles améliorent toutes ce problème à travers un usage plus efficace des paramètres, menant à une équivariance de rotation globale : « Harmonic CNNs » [83], « Group Equivariant Convolutional Networks (G-CNNs) » [84], « Exploiting Cyclic Symmetry in Convolutional Neural Networks » [85] et « Steerable CNNs » [86].

7 LES RESEAUX RESIDUELS

Les réseaux résiduels sont devenus très populaires en 2016, suite au succès de ResNet de Microsoft, [87] avec beaucoup de version open sources et des modèles pré-entrainés maintenant disponibles. En 2015, ResNet a gagné la première place dans les catégories d'ImagNet de Détection, Classification, Localisation, et le concours COCO dans la Détection et la Segmentation.

Ce qui différencie l'apprentissage profond des techniques d'apprentissage dites « creuses » est d'utiliser plusieurs couches cachées. Cela augmente remarquablement l'abstraction du problème et permet de mieux traiter les problèmes dits compositionnels. Il est facile en première approximation de comprendre pourquoi cette façon de procéder est intéressante : nous comprenons des choses simples avant de comprendre les choses plus complexes à partir de ce qui a déjà été appris.

Néanmoins, cela a longtemps posé des problèmes d'évanouissement ou d'explosion du gradient sur les réseaux profonds. En effet, la dérivée partielle sur les paramètres des couches les plus basses est calculée multiplicativement à partir de celle sur des paramètres des couches supérieures. Il est donc difficile d'ajuster ces couches bas-niveau, ce qui limite l'intérêt même de l'apprentissage profond.

Les Réseaux Résiduels sont souvent conceptualisés comme un ensemble de réseaux peu profonds. D'une certaine façon, ils contrecarrent la nature hiérarchique des réseaux profonds. Ils utilisent en effet des connections de raccourcis parallèles à leur couches convolutives. Ces raccourcis atténuent les problèmes d'évanouissement ou d'explosion du gradient. Ils permettent ainsi une rétro-propagation plus facile des gradients à travers leurs couches du réseau.

Ci-après quelques théories et améliorations notables apportées à l'apprentissage résiduel : « Wide Residual Networks » [88], « Deep Networks with Stochastic Depth » [89], « Learning Identity Mappings with Residual Gates » [90], « Residual Networks Behave Like Ensembles of Relatively Shallow Networks » [91], « Identity Mappings in Deep Residual Networks » [92] et « Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks » [93].

8 LES JEUX DE DONNÉES

Par essence, la richesse des jeux de données a toujours eu une importance significative dans le Machine Learning. Les algorithmes sont toujours évalués sur des jeux de données afin de montrer leur faisabilité en pratique. Le succès ou l'échec des algorithmes peut parfois s'expliquer par rapport aux caractéristiques des données. Ces jeux de données peuvent être

synthétiques. Ils sont alors générés de façon automatique ou proviennent d'applications réelles offrant ainsi un gage d'applicabilité.

A partir de 2016 les jeux de données traditionnels tels que ImageNet [94], COCO 170, CIFAR [95] et MNIST [96], ont été complété par de nombreux autres. On note également la recrudescence des jeux de données synthétiques. Ci-après une sélection subjective des plus importants nouveaux jeux de données apparus ces dernières années : « Places2 » [97], « SceneNet RGB-D » [98], « CMPlaces, MS-Celeb-1M » [99], « Open Images » [100], et « YouTube-8M » [101].

9 CONCLUSION

L'engouement pour le ''Deep Learning'' ces dernières années est pour beaucoup dû aux énormes progrès qu'il a permis de réaliser dans la Computer Vision. Toutes les catégories clés du ''Deep Learning'' pour la Computer Vision ont été utilisées pour finalement atteindre des taux de performances importants dans un large spectre de tâches liées à la compréhension visuelle. Notamment la détection d'objet, la poursuite, la reconnaissance faciale, la reconnaissance d'action et d'activité humaine, l'estimation de posture, la récupération d'image, et la segmentation sémantique. Cependant chaque catégorie de méthode présente des avantages et désavantages distincts. Les CNN ont l'unique capacité d'apprentissage automatique des données caractéristiques sur un jeu de données annotées. Ils sont également invariants aux transformations, ce qui est une grande qualité pour certaines applications de Computer Vision. A défaut, ils sont totalement dépendants de l'existence de données annotées, contrairement à d'autres approches qui peuvent apprendre de manière non supervisée.

Malgré les résultats prometteurs et parfois impressionnants qui ont documenté la littérature existante, de nombreux défis restent d'actualité comme le travail préparatoire théorique qui expliquerait clairement la façon de définir la sélection optimale des types de modèles et de structure pour une tache donnée, ou pour comprendre en profondeur les raisons pour laquelle une architecture définie ou un algorithme est efficace ou non pour une tache donnée. Ces questions primordiales continueront de susciter l'intérêt de la communauté de recherche du Machine Learning dans les années à venir. D'autres défis relatifs à l'analyse des signaux visuels font l'objet de programmes de recherche, comme le passage à l'échelle, la transition entre images fixes et vidéo, la multi-modalité, l'introduction de connaissances a priori.

Le taux d'erreur n'est pas le seul paramètre à être optimisé avec acharnement, les chercheurs travaillant continuellement à l'amélioration de la vitesse, l'efficacité des algorithmes et leur capacité à généraliser à d'autres taches et problèmes dans des approches totalement nouvelles. Certains chercheurs passent au premier plan avec des approches comme le « One-Shot Learning », les modèles génératifs, l'apprentissage par transfert et récemment l'apprentissage évolutif.

Ces derniers avancements ne manquent pas d'agiter le spectre, même lointain de l'Intelligence Artificielle Générale. Les inquiétudes et débats sur l'Intelligence Artificielle sont nombreux, et pourtant il est raisonnable de penser que le risque de perte de contrôle par les humains ne présente pas de caractère critique dans un avenir prévisible. Il n'en reste pas moins intéressant de considérer certains risques potentiels à court terme, dont la communauté de chercheur est consciente, et sur lesquels elle ne manque pas de communiquer. Les risques notables à court termes sont notamment, les bugs dans les logiciels, les cyberattaques, donner la capacité aux systèmes d'IA de comprendre ce que veulent les utilisateurs au lieu d'interpréter littéralement leurs ordres, l'autonomie partagée, et enfin les impacts socio-économiques de l'IA.

10 ANNEXES

10.1 SOURCES ET RÉFÉRENCES

Item	Auteurs	Description	URL
1	NA	Challenges LSVRC 2017 results	http://image-net.org/challenges/LSVRC/2017/results
2	NA	Challenges LSVRC 2017 results	http://image-net.org/challenges/LSVRC/2017/results
3	NA	Challenges LSVRC 2017 results	http://image-net.org/challenges/LSVRC/2017/results
4	R. Girshick, J. Donahue, T. Darrell, and J. Malik	Rich feature hierarchies for accurate object detection and semantic segmentation	https://arxiv.org/pdf/1311.2524.pdf
5	J. Hosang, R. Benenson, and B. Schiele	How good are detection proposals, really?	https://arxiv.org/pdf/1406.6962.pdf
6	B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik	Simultaneous detection and segmentation	https://arxiv.org/pdf/1407.1808.pdf
7	Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler	SegDeepM:Exploiting segmentation and context in deep neural networksfor object detection	https://arxiv.org/pdf/1502.04275.pdf
8	V. Nair and G. E. Hinton	3D object recognition with deep belief nets	https://pdfs.semanticscholar.org/4fdd/812505a362c6e7e6b1857f1d9be699d1b112.pdf
9	Liu et al	SSD: Single Shot MultiBox Detector	https://arxiv.org/pdf/1512.02325v5.pdf
10	Redmon, J. Farhadi	YOLO9000: Better, Faster, Stronger	https://arxiv.org/pdf/1612.08242v1.pdf
11	Redmon et al	You Only Look Once: Unified, Real-Time Object Detection	https://arxiv.org/pdf/1506.02640v5.pdf
12	Redmon	YOLO: Real-Time Object Detection	https://pjreddie.com/darknet/yolo/
13	Lin et al	Feature Pyramid Networks for Object Detection	https://arxiv.org/pdf/1612.03144v1.pdf
14	Dai et al	R-FCN: Object Detection via Region-based Fully Convolutional Networks	https://arxiv.org/pdf/1605.06409v2.pdf
15	Kevin Liang	Modern Convolutional Object Detectors, Faster R-CNN, R-FCN, SSD	http://people.ee.duke.edu/~lcarin/Kevin9.29.2017.pdf
16	Wu et al	SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving.	https://arxiv.org/pdf/1612.01051v2.pdf
17	Hong et al	PVANet: Lightweight Deep Neural Networks for Real-time Object Detection	https://arxiv.org/pdf/1611.08588v2.pdf
18	NA	ImageNet Challenges LSVRC 2017 results	http://image-net.org/challenges/LSVRC/2017/results
19	Pinheiro et al	Learning to Refine Object Segments.	https://arxiv.org/pdf/1603.08695v2.pdf
20	Zagoruyko	A MultiPath Network for Object Detection	https://arxiv.org/pdf/1604.02135v2.pdf
21	Jampani et al	Video Propagation Networks.	https://arxiv.org/pdf/1612.05478v2.pdf
22	Chen et al	DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.	https://arxiv.org/pdf/1606.00915v1.pdf
23	Khoreva et al	Simple Does It: Weakly Supervised Instance and Semantic Segmentation.	https://arxiv.org/pdf/1603.07485v2.pdf
24	Jégou et al	The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation.	https://arxiv.org/pdf/1611.09326v2.pdf
25	Li et al	Fully Convolutional Instance-aware Semantic Segmentation.	https://arxiv.org/pdf/1611.07709v1.pdf
26	Bertinetto et al	Fully-Convolutional Siamese Networks for Object Tracking.	https://arxiv.org/pdf/1606.09549v2.pdf
27	Held et al	Learning to Track at 100 FPS with Deep Regression Networks.	https://arxiv.org/pdf/1604.01802v2.pdf
28	Gladh et al	Deep Motion Features for Visual Tracking.	https://arxiv.org/pdf/1612.06615v1.pdf
29	Gaidon et al	Virtual Worlds as Proxy for Multi-Object Tracking Analysis.	https://arxiv.org/pdf/1605.06457v1.pdf
30	Lee et al	Globally Optimal Object Tracking with Fully Convolutional Networks.	https://arxiv.org/pdf/1612.08274v1.pdf
31	Champandard, A.J	Neural Enhance (latest commit 30/11/2016).	https://github.com/alexjc/neural-enhance
32	Romano et al	RAISR: Rapid and Accurate Image Super Resolution.	https://arxiv.org/pdf/1606.01299v3.pdf
33	Ledig et al	Photo-Realistic Single Image Super-Resolution Using a Generative Adversarila Network.	https://arxiv.org/pdf/1609.04802v3.pdf
34	Sønderby et al	Amortised MAP Inference for Image Super-resolution.	https://arxiv.org/pdf/1610.04490v1.pdf
35	Gatys et al	A Neural Algorithm of Artistic Style.	https://arxiv.org/pdf/1508.06576v2.pdf
36	Nikulin & Novak	Exploring the Neural Algorithm of Artistic Style.	https://arxiv.org/pdf/1602.07188v2.pdf
37	Ruder et al	Artistic style transfer for videos.	https://arxiv.org/pdf/1604.08610v2.pdf
38		Prisma. 2018.	https://prisma-ai.com/
39		Artomatix. 2018.	https://artomatix.com/
40	ia and Vajda	Delivering real-time AI in the palm of your hand.	https://code.fb.com/android/delivering-real-time-ai-in-the-palm-of-your-hand/
41	Dumoulin et al	Supercharging Style Transfer.	https://ai.googleblog.com/2016/10/supercharging-style-transfer.html
42	Dumoulin et al	A Learned Representation For Artistic Style.	https://arxiv.org/pdf/1610.07629v5.pdf
43	Zhang et al	Colorful Image Colorization.	https://arxiv.org/pdf/1603.08511v5.pdf
44	Larsson et al	Learning Representations for Automatic Colorization.	https://arxiv.org/pdf/1603.06668v2.pdf
45	Lizuka, Simo-Serra and Ishikawa	Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification.	http://hi.cs.waseda.ac.jp/~iizuka/projects/colorization/en/
46	A. S. Voulodimos, D. I. Kosmopoulos, N. D. Doulamis, and T. A. Varvarigou	A top-down event-driven approach for concurrent activity recognition	https://www.researchgate.net/profile/Athanasios_Voulodimos/publication/257627039_A_top-down_event-driven_approach_for_concurrent_activity_recognition/links/5aa2fe6c0f7e9badd9a6784b/A-top-down-event-driven-approach-for-concurrent-activity-recognition.pdf?origin=publication_detail
47	A. S. Voulodimos, N. D. Doulamis, D. I. Kosmopoulos, and T. A. Varvarigou	Improving multi-camera activity recognition by employing neural network based readjustment	http://culturetechlab.culture.uwg.gr/sites/default/files/downloads/AAI10.pdf
48	Bertrand Braunschweig et al	Livre Blanc Intelligence Artificielle, les défis actuels et l'action d'Inria	https://www.inria.fr/content/download/103897/1529370/.../AI_livre-blanc_n01.pdf
49	C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann	DevNet: A Deep Event Network for multimedia event detection and evidence recounting	https://pdfs.semanticscholar.org/3de0/50d1707524512eeab99780df3cbdee09670c.pdf
50	L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang	A deep structured model with radius-margin bound for 3D human activity recognition	https://arxiv.org/pdf/1512.01642.pdf
51	S. Cao and R. Nevatia	Exploring deep learning based solutions in fne grained activity recognition in the wild	https://projet.liris.cnrs.fr/imagine/pub/proceedings/ICPR-2016/media/files/1333.pdf
52	K. Tang, B. Yao, L. Fei-Fei, and D. Koller	Combining the right features for complex event recognition	http://vision.stanford.edu/pdf/iccv13-andor.pdf

53	R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic	Multiview fusion for activity recognition using deep neural networks	http://community.wvu.edu/~vkkulathumani/article-jei.pdf
54	H. Yalcin	Human activity recognition using deep belief networks	https://zapdf.com/human-activity-recognition-using-deep-belief-networks.html
55	Varol et al	Long-term Temporal Convolutions for Action Recognition.	https://arxiv.org/pdf/1604.04494v1.pdf
56	Feichtenhofer et al	Spatiotemporal Residual Networks for Video Action Recognition	https://arxiv.org/pdf/1611.02155v1.pdf
57	Vondrick et al	Anticipating Visual Representations from Unlabeled Video.	https://arxiv.org/pdf/1504.08023v2.pdf
58		Artificial General Intelligence	
59	Riegler et al	OctNet: Learning Deep 3D Representations at High Resolutions.	https://arxiv.org/pdf/1611.05009v3.pdf
60	Xiang et al	ObjectNet3D: A Large Scale Database for 3D Object Recognition.	http://cvgl.stanford.edu/projects/objectnet3d/
61	Choy et al	3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction.	https://arxiv.org/pdf/1604.00449v1.pdf
62	Gadelha et al	3D Shape Induction from 2D Views of Multiple Objects.	https://arxiv.org/pdf/1612.05872v1.pdf
63	Rezende et al	Unsupervised Learning of 3D Structure from Images.	https://arxiv.org/pdf/1607.00662v1.pdf
64	P. F. Felzenszwalb and D. P. Huttenlocher	Pictorial structures for object recognition	http://www.cs.cornell.edu/~dph/papers/pictorial-structures.pdf
65	X. Chen and A. L. Yuille	Articulated pose estimation by a graphical model with image dependent pairwise relations	https://papers.nips.cc/paper/5291-articulated-pose-estimation-by-a-graphical-model-with-image-dependent-pairwise-relations.pdf
66	A. Jain, J. Tompson, and M. Andriluka	Learning human pose estimation features with convolutional networks	https://cims.nyu.edu/~tompson/others/iclr2014_paper.pdf
67	J. J. Tompson, A. Jain, Y. LeCun et al	Joint training of a convolutional network and a graphical model for human pose estimation	https://www.robots.ox.ac.uk/~vgg/rg/papers/tompson2014.pdf
68	Cao et al	Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields	https://arxiv.org/pdf/1611.08050v1.pdf
69	Bogo et al	Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image.	https://arxiv.org/pdf/1607.08128v1.pdf
70	Kim et al	Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera.	https://www.doc.ic.ac.uk/~ajd/Publications/kim_etal_eccv2016.pdf
71	Garg et al	Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue.	https://arxiv.org/pdf/1603.04992v2.pdf
72	Izadinia et al	IM2CAD.	https://arxiv.org/pdf/1608.05137v1.pdf
73	DeTone et al	Deep Image Homography Estimation.	https://arxiv.org/pdf/1606.03798v1.pdf
74	Handa et al	gvnn: Neural Network Library for Geometric Computer Vision.	https://arxiv.org/pdf/1607.07405v3.pdf
75	Chaoyang Zhu	Place recognition: An Overview of Vision Perspective	https://arxiv.org/pdf/1707.03470.pdf
76	D. H. Hubel and T. N. Wiesel	Receptive felds, binocular interaction, and functional architecture in the cat's visual cortex	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/pdf/jphysiol01247-0121.pdf
77		Réseau de Neuronnes convolutif	(https://fr.wikipedia.org/wiki/Réseau_neuronal_convolutif),
78		Réseaux de neuronnes	http://www.statsoft.fr/concepts-statistiques/reseaux-de-neurones-automatises/reseaux-de-neurones-automatises.htm
79	Szegedy et al	Rethinking the Inception Architecture for Computer Vision	https://arxiv.org/pdf/1512.00567v3.pdf
80	Huang et al	Densely Connected Convolutional Networks.	https://arxiv.org/pdf/1608.06993v3.pdf
81	Larsson et al	FractalNet: Ultra-Deep Neural Networks without Residuals.	https://arxiv.org/pdf/1605.07648v2.pdf
82	Hossein HasanPour et al	Lets keep it simple: using simple architectures to outperform deeper architectures	https://arxiv.org/vc/arxiv/papers/1608/1608.06037v3.pdf
83	Singh et al	Swapout: Learning an ensemble of deep architectures.	https://arxiv.org/pdf/1605.06465v1.pdf
84	Iandola et al	SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size.	https://arxiv.org/pdf/1602.07360v4.pdf
85	Worrall et al	Harmonic Networks: Deep Translation and Rotation Equivariance.	https://arxiv.org/pdf/1612.04642v1.pdf
86	Cohen & Welling	Group Equivariant Convolutional Networks.	https://arxiv.org/pdf/1602.07576v3.pdf
87	Dieleman et al	Exploiting Cyclic Symmetry in Convolutional Neural Networks.	https://arxiv.org/pdf/1602.02660v2.pdf
88	Cohen & Welling	Steerable CNNs.	https://arxiv.org/pdf/1612.08498v1.pdf
89	He et al	Deep Residual Learning for Image Recognition.	https://arxiv.org/pdf/1512.03385v1.pdf
90	Zagoruyko, S. and Komodakis, N	Wide Residual Networks.	https://arxiv.org/pdf/1605.07146v3.pdf
91	Huang et al	Deep Networks with Stochastic Depth.	https://arxiv.org/pdf/1603.09382v3.pdf
92	Savarese et al	Learning Identity Mappings with Residual Gates.	https://arxiv.org/pdf/1611.01260v2.pdf
93	Veit, Wilber and Belongie	Residual Networks Behave Like Ensembles of Relatively Shallow Networks.	https://arxiv.org/pdf/1605.06431v2.pdf
94	He at al	Identity Mappings in Deep Residual Networks.	https://arxiv.org/pdf/1603.05027v3.pdf
95	Abdi, M., Nahavandi, S	Multi-Residual Networks: Improving the Speed and Accuracy of Residual Networks.	https://arxiv.org/pdf/1609.05672v3.pdf
96		ImageNet. 2017. Homepage	http://image-net.org/index
97		CIFARs. 2017. The CIFAR-10 dataset	https://www.cs.toronto.edu/~kriz/cifar.html
98		MNIST. 2017. THE MNIST DATABASE of handwritten digits.	http://yann.lecun.com/exdb/mnist/
99	Zhou et al	Places2	http://places2.csail.mit.edu/
100	McCormac et al	SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectorieswith Ground Truth	https://arxiv.org/pdf/1612.05079v3.pdf
101	Guo et al	MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition	https://arxiv.org/pdf/1607.08221v1.pdf
102	Open Images	Open Images Dataset	https://github.com/openimages/dataset
103	Abu-El-Haija et al	YouTube-8M: A Large-Scale Video Classification Benchmark.	https://arxiv.org/pdf/1609.08675v1.pdf