Amélioration des performances des réseaux de neurones convolutifs en localisation indoor par augmentation des données

Andrea Daou, Jean-baptiste Pothin, Paul Honeine, Abdelaziz Bensrhair

Amélioration des performances des réseaux de neurones convolutifs en localisation indoor par augmentation des données

Improving Convolutional Neural Networks Performance in Indoor Localization by Data Augmentation

Andrea DAOU[1,2] Jean-Baptiste POTHIN[2] Paul HONEINE[1] Abdelaziz BENSRHAIR[3]

[1] LITIS Lab, Université de Rouen Normandie, Saint-Etienne-du-Rouvray, France
[2] Département de recherche et développement, DataHertz, Troyes, France
[3] LITIS Lab, INSA Rouen Normandie, Saint-Etienne-du-Rouvray, France

andrea.daou@univ-rouen.fr

Résumé

Les réseaux de neurones convolutifs (CNN) offrent des performances remarquables dans la détection et la reconnaissance d'objets, en grande partie grâce aux jeux de données volumineux et de qualité existants. Ces performances se détériorent en localisation indoor à cause de la faible quantité de données disponibles. Une autre limite des CNN est leur robustesse réduite aux transformations géométriques comme la rotation et le changement d'échelle. Pour pallier ces défauts, nous analysons l'effet de l'augmentation des données sur les performances du système de classification, en ajoutant des images modifiées pour tenir compte des changements de point de vue représentatifs. Pour compenser les temps de calcul plus longs, nous utilisons un modèle CNN pré-entraîné et appliquons un apprentissage par transfert. Les résultats obtenus sur les jeux de données d'images de scènes MIT Indoor67 et Scene15 montrent l'intérêt de l'approche proposée.

Mots Clef

Apprentissage profond, augmentation de données, classification, CNN, localisation, vision par ordinateur

Abstract

Convolutional Neural Networks (CNNs) offer remarkable performance in object detection and recognition tasks, mainly thanks to large-scale high-quality datasets. This performance deteriorates in indoor localization because of the small amount of data available. Another limitation of CNNs is their reduced robustness to geometric transformations, such as rotation and scaling. To overcome these shortcomings, we analyze the effect of data augmentation on the performance of the classification system by adding modified images to account for representative perspective changes. To compensate for the long computing delays, we use a pre-trained CNN model and apply transfer learning. The results obtained on the MIT Indoor67 and Scene15 datasets demonstrate the relevance of the proposed method.

Keywords

Deep learning, data augmentation, classification, CNN, localization, computer vision

1 Introduction

La géolocalisation d'un bien ou d'une personne fait partie de notre quotidien. Un système de localisation indoor permet de trouver la position d'une personne ou d'un objet dans un espace intérieur. L'utilisation courante de systèmes de positionnement par satellites (e.g. GPS et Galileo) en zones externes n'est pas applicable dans les zones internes à cause de la faible intensité du signal et la précision réduite dans un environnement clos et encombré [1]. Deux approches principales existent pour répondre au problème de localisation indoor : les systèmes qui nécessitent une infrastructure émetteurs/récepteurs, et les systèmes sans infrastructures pouvant fonctionner de façon (quasi)-autonome. Les systèmes avec infrastructure utilisent pour la majorité des technologies radiofréquences (e.g. RFID, Bluetooth, Wifi) [2, 3]. La principale limite réside dans le coût souvent important lié à l'installation et à la maintenance de l'infrastructure dédiée, ainsi que la forte sensibilité aux conditions environnementales qui altèrent les ondes, comme la nature des obstructions (murs, mobiliers, êtres humains, etc.). Les systèmes sans infrastructure visent à surmonter ces inconvénients. Dans ce cadre, les méthodes de vision par caméra embarquée ont montré un grand intérêt au cours des dernières années [4]. En effet, les méthodes de suivi basées sur la reconnaissance visuelle des caractéristiques des images ne nécessitent qu'un dispositif d'acquisition d'images.

Au cours des deux dernières décennies, la détection de caractéristiques d'images a été impulsée par les approches

basées sur des caractéristiques locales prédéfinies de type SIFT (*Scale-Invariant Feature Transform*) [5] et SURF (*Speeded-Up Robust Features*) [6], notamment en localisation indoor [7]. Pour améliorer les performances des systèmes de détection et classification, les chercheurs se sont récemment intéressés à remplacer ces méthodes traditionnelles de détection par des réseaux de neurones à architecture profonde, dont le fer de lance est les réseaux de neurones convolutionnels (CNN pour *Convolutional Neural Networks*) [8]. Bien que les CNN aient abouti à des performances prometteuses dans la détection et la reconnaissance d'objets, leur faible invariance aux rotations et aux changements d'échelle restent des défis importants à affronter [9]. D'autre part, la qualité et la quantité de données disponibles constituent des facteurs clés contribuant au succès de ces réseaux [10].

Le présent article examine ces problématiques rencontrées dans un système de localisation par vision embarquée, où les prises d'images par un utilisateur sont loin d'être parfaites pour plusieurs raisons. La variation de la distance entre la caméra du dispositif porté et les objets dans l'environnement concerné aboutit à un changement d'échelle, étant donné que les images de scènes en intérieur sont composées de plusieurs objets constituants l'identité du lieu. De plus, la capture d'images par un smartphone peut être exposée à une rotation ; normalement entre le cadre portrait et le cadre paysage. Or, le pouvoir de généralisation des CNN diminue considérablement lorsqu'ils rencontrent des données avec des transformations sévères, ce qui rend souvent leur fonctionnement inefficace en inférence. Par suite, une propriété souhaitable est d'avoir un système capable de surpasser ces défis et prédire correctement l'emplacement de la personne.

Pour faire face à ces défis, nous analysons dans cet article l'effet de l'augmentation des données sur les performances du système de classification, où le jeu de données est augmenté par des transformations géométriques qui prennent compte des changements de point de vue représentatifs, notamment le changement d'échelle et la rotation. Pour compenser les temps de calcul plus long à cause de l'augmentation de données, nous utilisons un modèle CNN pré-entraîné (AlexNet, pré-entraîné sur le jeu de données ImageNet [11]) et appliquons un apprentissage par transfert. Afin d'analyser l'effet de l'augmentation des données sur les performances du système de classification, nous considérons deux jeux de données d'images de scènes en intérieur très connus : MIT Indoor67 constitué de 67 catégories groupant plusieurs environnements intérieurs [12] et Scene15 avec ses 5 catégories de scènes en intérieur [13]. Les résultats expérimentaux obtenus sur ces jeux de données d'images de scènes en intérieur montrent l'intérêt de l'approche proposée.

La suite de l'article est organisée comme suit. Dans la Section 2, nous fournissons plus de détails sur les défis des CNN rencontrés lors de la classification de scènes en intérieur, plus précisément le manque de données de qualité, l'invariance à l'échelle ainsi qu'à la rotation. Dans la Section 3, nous expliquons le système proposé et la méthode choisie pour surmonter les problématiques déjà évoquées. La Section 4 décrit les différentes expérimentations réalisées pour ensuite analyser les résultats dans la Section 5. Finalement, la Section 6 conclut l'article.

2 Les défis des CNN pour la classification de scènes en intérieur

Un jeu de données volumineux et de qualité est un facteur clé dans les problèmes d'apprentissage profond, plus précisément l'apprentissage supervisé. Pour réaliser des classifications d'images en se basant sur les CNN pré-entraînés, grâce à un apprentissage par transfert, ou par un apprentissage complet, un jeu de données approprié est primordial.
D'autre part, les CNN entraînés avec des images parfaitement capturées ne peuvent pas réaliser de bonnes classifications lorsqu'ils sont confrontés à des images ayant subi des transformations géométriques comme la rotation ou le changement d'échelle puisqu'ils ont une invariance limitée à ces transformations.
Dans la suite, nous présentons l'état de l'art des jeux de données de scènes, et décrivons les défis d'invariance à l'échelle et à la rotation.

2.1 Jeu de données de scènes

Un jeu de données d'apprentissage doit contenir un nombre suffisant d'images annotées des catégories du sujet de classification étudié. Les catégories d'images de scènes en intérieur couvrent un large éventail de domaines comme les lieux de travail (bureaux, hôpitaux, salles de cours, etc.), les magasins, les pièces intérieures des maisons, les endroits de loisirs et d'autres. D'autre part, les catégories d'images de scènes en extérieur couvrent l'ensemble des environnements urbains comme les ponts, les gratte-ciels, les autoroutes, les campus, les stations, l'extérieur des cathédrales et des monuments historiques ainsi que les environnements naturels comme les forêts, les lacs, les grottes, les montagnes, les parcs, etc.
Généralement, la majorité des jeux de données d'images de scènes disponibles comportent des images d'environnements externes et internes, comme MIT Indoor67 [12], Scene15 [13], SUN [14], SUN Attribute [15] et Places [16]. Les images contenues dans ces jeux de données ne tiennent pas compte des différents points de vue qui peuvent être rencontrés dans la vie réelle lors d'une capture d'image.

2.2 Invariance à l'échelle

Même si les CNN ont atteint des performances très proches de celles des humains dans diverses missions de vision par ordinateur, leur capacité à tolérer les variations d'échelle est limitée [17].
Les couches de max-pooling contribuent à la résilience à une légère déformation, ainsi qu'à un changement à petite échelle [17]. En effet, une couche de max-pooling

FIGURE 1 – Exemples de catégories d'images dans le jeu de données MIT Indoor67.

rassemble les valeurs maximales des caractéristiques présentes dans chaque région de la carte générée après une couche de convolution. Ainsi, les opérations qui suivent sont effectuées sur des caractéristiques regroupées au lieu d'un positionnement précis. Cela rend le modèle plus robuste aux variations de position et d'échelle des caractéristiques dans l'image d'entrée. Cependant, en raison de la taille de la matrice de pooling typiquement faible, les CNN ne sont pas réellement invariants aux grandes transformations des données d'entrée [18]. Plusieurs chercheurs ont évoqué cette problématique rencontrée dans les tâches de détection d'objet, en proposant des modifications sur l'architecture du CNN [19, 20].

2.3 Invariance à la rotation

Les CNN présentent une robustesse limitée à la rotation [21]. Plusieurs approches existent pour cibler cet aspect, dont l'utilisation d'un bloc de prétraitement des données pour corriger l'orientation de l'image capturée avant de réaliser la classification. Cela peut être fait en utilisant des méthodes traditionnelles pour prédire l'angle de rotation [22] et puis entamer la correction ; une autre approche repose sur un modèle CNN supplémentaire en amont pour prédire l'angle d'orientation de l'image, soit par classification [23], soit par régression [24].

3 Système proposé

Pour réduire les temps de calcul, nous considérons un modèle CNN pré-entraîné sur le jeu de données de qualité. Le modèle choisi est le CNN AlexNet [25]. Il est constitué de huit couches (5 couches convolutionnelles, certaines suivies de max-pooling, puis 3 couches entièrement connectées), et utilise la fonction d'activation non saturante ReLU. AlexNet est pré-entraîné sur le jeu de données ImageNet [11]. Ce dernier est composé de plus de 15 millions d'images haute résolution étiquetées dans plus de 22 000 catégories d'objets. Le modèle AlexNet pré-entraîné sur ImageNet est une référence de l'état de l'art depuis qu'il a remporté de loin le concours ImageNet Large Scale Visual Recognition Challenge. Bien que de nouveaux modèles CNN plus profonds sont plus performants, AlexNet reste intéressant par sa faible taille pour un traitement embarqué sur caméra portée.

Nous opérons un apprentissage par transfert pour adapter le modèle AlexNet, pré-entraîné sur les catégories d'objets d'ImageNet, au problème de classification de scènes en intérieur. Pour ce faire, nous considérons un jeu d'images de scènes en intérieur, comme MIT Indoor67 constitué de 67 catégories d'environnements intérieurs. Nous décrivons dans la suite la méthode d'augmentation des données, qui consiste à ajouter des images modifiées pour tenir compte des changements de point de vue représentatifs.

Augmentation de données

Comme déjà mentionné, deux défis sont visés : le manque de données de scènes intérieures ainsi que l'invariance [26, 27]. L'approche utilisée dans cet article pour contourner ces problématiques est la méthode d'augmentation des données [28]. Cette méthode s'est avérée bénéfique pour l'entraînement des modèles d'apprentissage automatique à architecture profonde puisqu'elle aide à réduire le surapprentissage et améliore la généralisation.

Il existe plusieurs méthodes d'augmentation de données, dont l'application de transformations géométriques comme le recadrage, le changement d'échelle, la rotation, l'effet miroir et d'autres sur les images du jeu de données d'apprentissage. Cette approche permet au modèle CNN d'apprendre des caractéristiques d'images plus diversifiées et par suite pouvoir prédire correctement la catégorie de l'image capturée. Comme nous nous intéressons à l'aspect d'invariance en rotation et en échelle, nous nous contentons d'appliquer ces deux types de transformations aux images. Les augmentations de rotation se font en faisant pivoter l'image à droite ou à gauche autour d'un axe d'un angle entre 1° et 359°.

Les changements d'échelle se font en multipliant les di-

mensions de l'image par un facteur, qui permet ainsi soit d'agrandir l'échelle de l'image (facteur supérieur à 1), soit de la rétrécir (facteur inférieur à 1). Notons que les CNN prennent une taille fixe d'image ainsi qu'un nombre défini de canaux en entrée et donc toutes les images du jeu de données sont prétraitées [1], puis le changement d'échelle est appliqué. Par suite, si l'image est agrandie, des parties des bords seront éliminées et, si l'image est rétrécie, un contour noir sera automatiquement ajouté.

4 Expérimentations

4.1 Jeux de données évalués

L'analyse de l'augmentation des données sur les performances du système de classification est réalisée sur MIT Indoor67 et les 5 catégories de scènes en intérieur dans Scene15.

Un facteur provoquant le ralentissement d'amélioration des systèmes de classification des images de scènes en intérieur est le manque de données de qualité. Pour cela, les auteurs de [12] ont créé MIT Indoor67 [2], un jeu de données contenant 67 catégories groupant plusieurs environnements intérieurs, dont quelques exemples sont représentés dans la FIGURE 1. Ce jeu de données contient en tout 15 620 images en couleur, divisées d'une façon non équitable entre les catégories. Il existe au moins 100 images par catégorie et la dimension minimale des images est de 200 pixels.

Scene15 [3] est un jeu de données qui contient 15 catégories d'images de scènes dont 5 en intérieur et 10 en extérieur [13]. Ces images sont en représentation de niveaux de gris [29]. Chaque catégorie comprend 200 à 400 images, et la taille des images est en moyenne égale à 300 × 250 pixels.

4.2 Configuration

Les implémentations sont réalisées à l'aide de MATLAB R2019a sur un processeur Intel Core i7- 6700 CPU. Pour chaque époque d'apprentissage, une combinaison aléatoire de transformations est appliquée aux images dans le mini-lot de données d'apprentissage. Une perturbation aléatoire de ces transformations choisies est exécutée lors de chaque époque, de sorte que chaque époque utilise un ensemble de données légèrement différent. Notons que les images après transformations ne sont pas stockées en mémoire et que le nombre réel d'images d'entraînement à chaque époque ne change pas.

Nous réalisons plusieurs essais en combinant les approches suivantes d'entraînement du modèle CNN (initialement pré-entraîné sur ImageNet) avec celles de test :

— Phase d'entraînement du modèle CNN avec le jeu de données, avec ou sans augmentation.

1. Les images du jeu de données de scènes en intérieur sont redimensionnées à 227 × 227 pixels et subissent un prétraitement de couleur si nécessaire (passage de représentation niveaux de gris à une représentation couleur) pour respecter la dimension et le nombre de canaux égal à 3 acceptés par la couche d'entrée du CNN AlexNet utilisé.

2. MIT Indoor67 dataset : http ://web.mit.edu/torralba/www/indoor.html

3. 15-Scene image dataset : https ://figshare.com/articles/dataset/15-Scene_Image_Dataset/7007177

FIGURE 2 – Exemples d'images après augmentation de données.

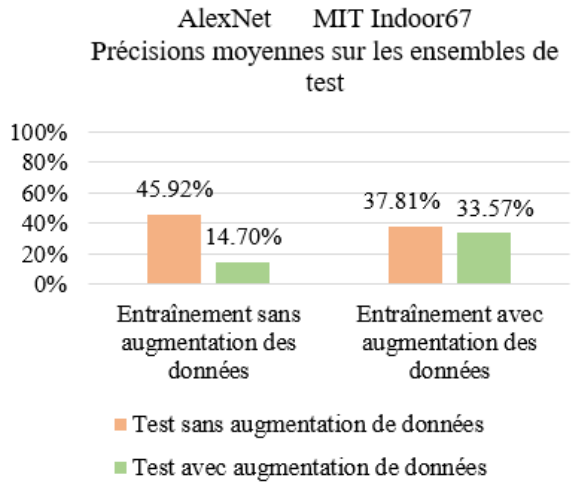

FIGURE 3 – Précisions moyennes des ensembles de test pour le modèle AlexNet entraîné et testé avec les variations du jeu de données MIT Indoor67, avec et sans augmentation de données.

— Phase de test avec le jeu de données, avec et sans augmentation.

Nous utilisons 100 images pour chaque catégorie du jeu de données MIT Indoor67 avec 5 divisions aléatoires comprenant 70% des images pour la phase d'apprentissage et 30% pour la phase de test. Pour Scene15, nous nous contentons des 5 catégories de scènes en intérieur avec 210 images pour chacune avec 5 divisions aléatoires comprenant 70% des images pour la phase d'entraînement et 30% pour la phase de test.

La rotation normale d'images de scène dans un système de localisation par vision embarquée est prise entre -90° et 90° pour représenter la variation entre le cadre portrait et le cadre paysage. L'intervalle des facteurs d'échelle est choisi égal à [0,5 2]. Durant l'apprentissage, l'angle de rotation et le facteur d'échelle sont choisis aléatoirement à partir d'une distribution uniforme continue dans les intervalles déjà spécifiés. La FIGURE 2 présente quelques exemples d'images après l'application de l'augmentation de données.

5 Résultats et discussion

Les FIGURES 3 et 4 montrent les résultats en termes de précisions moyennes de test pour chacune des combinaisons du modèle CNN avec les jeux de données MIT Indoor67 et Scene15, respectivement.

Avec MIT Indoor67, nous pouvons observer que, pour les réseaux entraînés par des jeux de données avec augmentation, la chute de précision de test est d'environ 4 points de pourcentage entre l'ensemble de test sans augmentation de données et l'ensemble de test avec augmentation de données. Tandis que la chute de performances pour les réseaux entraînés avec l'ensemble de données sans augmentation est d'environ 31 points de pourcentage entre les deux ensembles de test. Précisons que les mêmes paramètres (algorithme d'optimisation, taux d'apprentissage et taille du mini-lot d'images) sont choisis durant les différents apprentissages du CNN.

Bien que la moyenne de précision de test avec des réseaux entraînés à l'aide de jeux de données avec augmentation soit inférieure à celle avec des images sans augmentation, ces réseaux présentent une meilleure généralisation puisqu'ils ont de meilleures performances face à des images non habituelles (avec transformations). Cela fait qu'avec une augmentation de données si l'image est capturée après une rotation du dispositif de sa position normale ou avec une échelle différente que celle lors de la création du jeu de données d'apprentissage, le système de localisation sera plus apte à reconnaître correctement l'emplacement de la personne. Notons que le temps d'entraînement des CNN avec un jeu de données qui a subi une augmentation de données est plus long, ce qui est normal.

Les mêmes déductions peuvent être faites suite à l'entraînement du réseau AlexNet avec les 5 catégories d'images de scènes en intérieur du jeu de données Scene15. Pour les réseaux entraînés par des jeux de données avec augmentation, la chute de précision de test est d'environ 7 points de pourcentage entre l'ensemble de test sans augmentation de données et l'ensemble de test avec augmentation de données. La chute de performances pour les réseaux entraînés avec l'ensemble de données sans augmentation est d'environ 32 points de pourcentage.

6 Conclusion

Dans cet article, nous avons considéré le problème de localisation indoor par vision par ordinateur. Nous avons analysé l'effet de l'augmentation des données sur les performances en phase de test pour le modèle AlexNet avec deux jeux de données différents, en conservant les mêmes paramètres d'apprentissage dans chaque cas. Même si l'augmentation des données nécessite plus de temps d'entraînement, cette méthode aide à atteindre de meilleures performances avec des images de test qui ont subi des transformations géométriques. Non seulement les taux de précision recueillis décrivent un bon modèle mais sa capacité à avoir une bonne généralisation est aussi un critère important.

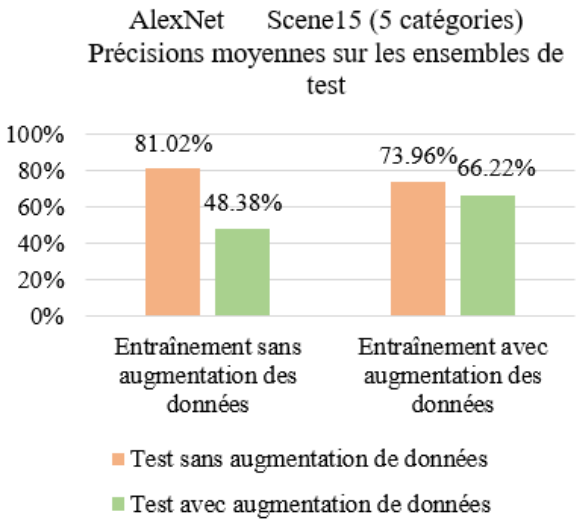

FIGURE 4 – Précisions moyennes des ensembles de test pour le modèle AlexNet entraîné et testé avec les variations du jeu de données Scene15, avec et sans augmentation des données.

Les futurs travaux se concentreront sur la proposition d'un nouveau CNN qui cherche à surmonter les défis d'invariance à l'échelle et à la rotation soulevés dans cet article. En outre, nous proposerons un système de localisation en intérieur par l'intégration de ces travaux sur la classification de scènes.

Références

[1] J. Kunhoth, A. Karkar, S. Al-Maadeed, and A. Al-Ali, "Indoor positioning and wayfinding systems : a survey," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, p. 18, 2020.

[2] S. Kammoun, J.-B. Pothin, and J.-C. Cousin, "Beacon Placement using Simulated Annealing for Indoor Localization," in *iswcs2014*, (Barcelone, Spain), Aug. 2014.

[3] D. AlShamaa, F. Chehade, P. Honeine, and A. Chkeir, "Fusion of multiple mobility and observation models for indoor zoning-based sensor tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, pp. 4315–4326, Dec. 2020.

[4] F. Hu, *Emerging techniques in vision-based indoor localization*. PhD thesis, PhD Computer Science, University of New York, New York, USA, Feb. 2015.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *The proceedings of the seventh IEEE international conference on Computer vision, 1999.*, vol. 2, pp. 1150–1157, IEEE, 1999.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3,

pp. 346–359, 2008. Similarity Matching in Computer Vision and Multimedia.

[7] S. Konlambigue, P. Honeine, J.-B. Pothin, and A. Bensrhair, "Performance Evaluation of State-of-the-art Filtering Criteria Applied to SIFT Features," in *19th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, (Ajman, United Arab Emirates), Dec. 2019.

[8] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN : A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[9] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years : A survey," *arXiv preprint arXiv :1905.05055*, 2019.

[10] C. Luo, X. Li, L. Wang, J. He, D. Li, and J. Zhou, "How does the data set affect CNN-based image classification performance ?," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, pp. 361–366, 2018.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet : A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[12] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 413–420, 2009.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2169–2178, 2006.

[14] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database : Large-scale scene recognition from abbey to zoo," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

[15] G. Patterson and J. Hays, "Sun attribute database : Discovering, annotating, and recognizing scene attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, 2012.

[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places : A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[17] Y. Xu, T. Xiao, J. Zhang, K. Yang, and Z. Zhang, "Scale-invariant convolutional neural networks," *arXiv preprint arXiv :1411.6369*, 2014.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, 2015.

[19] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn : Object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 379–387, Curran Associates Inc., 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[21] R. Takahashi, T. Matsubara, and K. Uehara, "Scale-invariant recognition by weight-shared CNNs in parallel," in *Proceedings of the Ninth Asian Conference on Machine Learning* (M.-L. Zhang and Y.-K. Noh, eds.), vol. 77 of *Proceedings of Machine Learning Research*, pp. 295–310, PMLR, 15–17 Nov 2017.

[22] C. Hollitt and A. S. Deeb, "Determining image orientation using the hough and fourier transforms," in *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, IVCNZ '12, (New York, NY, USA), p. 346–351, ACM, 2012.

[23] K. Swami, P. P. Deshpande, G. Khandelwal, and A. Vijayvargiya, "Why my photos look sideways or upside down ? detecting canonical orientation of images using convolutional neural networks," in *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 495–500, 2017.

[24] S. Maji and S. Bose, "Deep image orientation angle detection," *arXiv preprint arXiv :2007.06709*, 2020.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, p. 84–90, May 2017.

[26] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor scene classification," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4829–4841, 2016.

[27] X. Shen, X. Tian, A. He, S. Sun, and D. Tao, "Transform-invariant convolutional neural networks for image classification and search," in *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, (New York, NY, USA), p. 1345–1354, Association for Computing Machinery, 2016.

[28] F. Quiroga, F. Ronchetti, L. Lanzarini, and A. F. Bariviera, "Revisiting data augmentation for rotational invariance in convolutional neural networks," in *Modelling and Simulation in Management Sciences* (J. C. Ferrer-Comalat, S. Linares-Mustarós, J. M. Merigó, and J. Kacprzyk, eds.), (Cham), pp. 127–141, 2020.

[29] H. Patel and H. Mewada, "Analysis of machine learning based scene classification algorithms and quantitative evaluation," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 7811–7819, 2018.