WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n3.05
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Trends in Image Object Detection and Segmentation with Deep Neural Network on Natural and Earth Observations

Claudia Hoeser [1], Gongyi Zhang [2, *]

[1] Department of Remote Sensing, Institute of Geography and Geology, University of Extremadura, 10071 Caceres, Spain

[2] Department of Applied Mathematics, The University of Sydney, Sydney, NSW 2006, Australia

*Corresponding Author: Gongyi Zhang

ABSTRACT

Deep learning has a profound impact across multiple scientific domains and has increasingly positioned itself as a versatile tool for addressing new challenges in the field of natural (Earth) observation. However, researchers in Earth observation often face significant entry barriers, primarily due to the complexity and rapid evolution of the field, which is heavily driven by innovations in computer vision. To assist researchers in overcoming these challenges, this paper provides a comprehensive discussion of the development of deep learning, with a particular focus on image segmentation and object detection using convolutional neural networks. The paper begins with key developments when convolutional neural networks set new benchmarks in image recognition, and continues through to the innovations made recently. This paper traces the interconnections among the most influential convolutional architectures and major milestones originating from computer vision, facilitating a clear understanding of how these advances shape modern deep learning models. Additionally, we offer insights into the evolution of popular deep learning frameworks and present a summary of datasets commonly used in natural observation. By exploring well-performing architectures and evaluating their applications on Earth observation datasets, we assess how breakthroughs in computer vision influence future research in natural observation. This paper bridges the gap between theoretical advancements in computer vision and their practical implementation in natural observation, equipping researchers with the knowledge needed to effectively integrate deep learning into their work.

KEYWORDS

Deep learning; Neural networks; Convolutional neural networks; Image segmentation and detection

1. INTRODUCTION

In recent years, deep learning has garnered significant attention in both scientific research and practical applications [1]. The big breakthroughs are not only in computer science fields like face-action recognition [2], multimedia system network [3-6], Large language & vision generation models [7-9], auto-driving navigation [10], but also in Medical [11], Spectroscopy [12, 13], human health [11, 14-16] scientific research [17, 18] etc. Fields. Two primary factors driving this surge are the increased accessibility of data and advancements in computational processing power, especially through the use of graphics processing units. These developments have enabled researchers to demonstrate deep learning concepts that often outperform traditional approaches. The rapid evolution of these insights has also led to their widespread application across various disciplines, fostering a self-reinforcing research environment that profoundly impacts both science and practice.

With big progress achieved in computer vision [19-27] and powerful software in the chain [28, 29], the growing availability of data has been particularly evident in the field of all-kinds observation, where high-resolution optical and multispectral imagery play a crucial role. Recent achievements in reinforcement [30] and social Learning [31-38] makes even great insight in real-world application. With many Earth observation archives now being opened, the volume of available high-resolution remote sensing data is expected to increase significantly in the near future. However, high-resolution imagery has already enabled the transfer of deep learning techniques from computer vision to Earth observation, with applications such as detecting and segmenting vehicles, roads, and buildings from aerial images. These early successes have demonstrated the potential of deep learning for Earth observation, and today, applications are expanding beyond traditional RGB images. The increasing number of deep learning implementations in this field continues to reveal new trends and possibilities for analyzing remotely sensed data.

The importance of deep learning extends far beyond individual disciplines, permeating the broader scientific community with the support from hardware advancements, like millimeter Wave technology [39-43] and Robotics [44]. In 2019, the IEEE Conference on Computer Vision and Pattern Recognition, renowned for its contributions to deep learning research, reached the top ten h5-index rankings for the first time. The involvement of major technology companies such as Google and Facebook have also contributed significantly to the field's growth. These companies not only drive theoretical research but also develop widely adopted frameworks, with Google advancing TensorFlow and Facebook promoting PyTorch. A useful indicator of the field's rapid growth is the increasing number of deep learning-related publications. Both the absolute volume of submissions and their share within the computer science and statistics categories have grown annually, reflecting a broadening interest in deep learning across multiple domains.

One reason for the success of deep learning lies in its ability to represent abstract concepts such as speech and images. These models have outperformed traditional machine learning methods and signal processing techniques in areas like speech recognition and handwritten digit recognition. A pivotal moment came in 2012 when Ciresan [45] and Krizhevsky [46] introduced convolutional neural networks (CNNs) for image recognition. Krizhevsky model, known as AlexNet [46], set a new standard by winning the ImageNet Large Scale Visual Recognition Challenge, a major competition for computer vision tasks. The remarkable performance of AlexNet in 2012 marked the beginning of modern deep learning's rapid development, making it a central point of reference for this paper. Initially rooted in computer science, deep learning has since expanded into numerous other fields. The widespread citation of Krizhevsky work in disciplines beyond computer science illustrates the broad impact of convolutional neural networks. As the number of citations continues to grow, it highlights the increasing relevance of CNNs across various research domains.

Given the centrality of imagery data to observation, deep learning has found a natural application in this field. For example, good progress was made by computer vision in surgery of liver Transplantation [47-50]. However, the adoption of these techniques by the Earth observation community lagged three to four years behind their introduction in computer vision. The number of Earth observation studies employing deep learning has more than doubled annually. This growth persisted into 2020, with publications in the first quarter alone accounting for nearly half the total number from 2019. Recent reviews have documented the diverse applications of deep learning in remote sensing, including super-resolution imaging, data fusion, denoising, weather forecasting, scene recognition, classification, and object detection using optical, multispectral, hyperspectral, and synthetic aperture radar sensors. Given the wealth of insights generated in deep learning over the past eight years, this paper aims to provide a detailed examination of the field's progress, particularly within computer vision, from 2012 to late 2019. The review focuses on two core areas: object detection and image segmentation using convolutional neural networks. By tracing the key milestones and their interconnections, we offer Earth observation researchers a roadmap to better understand and leverage these advancements. Whether seeking to adopt deep learning for the first time or selecting

the most suitable models for specific research questions, this review addresses key challenges, including overcoming the high entry barriers associated with deep learning.

This review contributes to addressing one of the field's open questions—how to lower the barriers for adopting deep learning, as highlighted by Ball et al. [51]. The insights presented here provide a solid foundation for understanding the principles of convolutional neural networks. In Part II of this review, we will build on this foundation by exploring practical applications of CNNs within Earth observation, drawing on leading research from top journals in the field.

2. DEEP LEARNING WITH CNNS

In supervised machine learning, algorithms are designed to learn meaningful features from labeled training data to make accurate predictions for unseen inputs [52]. Deep learning models, as a specialized form of machine learning, consist of multiple stacked layers that progressively extract richer and more abstract patterns from input data. As the number of layers increases, the model becomes deeper and capable of capturing increasingly complex representations, which gives rise to the term "deep learning." Machine learning belongs to the broader domain of artificial intelligence, making deep learning a specific area within it rather than synonymous with it. Before delving into different deep learning models, Table 1 offers an overview of essential terminology. This table is organized both thematically and chronologically to help readers navigate the key concepts in deep learning, including topics such as image recognition, image segmentation, and object detection.

A classical deep learning model, the artificial neural network depicted in Figure 1, consists of a sequence of fully connected layers arranged from input to hidden to output layers of artificial neurons [46]. In a fully connected network, each neuron in one layer connects to every neuron in the preceding layer through linear operations, referred to as weights or parameters. These connections transmit values by multiplying each input with its corresponding weight and passing the results to the next layer. Each neuron sums the incoming values and applies a non-linear activation function to transform the data further. This process enables the network to map input data through multiple hidden layers to produce an output, forming a higher-order non-linear function. Images, as records of natural signals, convey information through pixel values and their spatial relationships. Low-level features, such as edges, can be combined into higher-level features with semantic meaning. Analyzing images requires uncovering these distinctive representations by exploring pixel values and their connectivity. A standard artificial neural network may struggle to capture these spatial relationships effectively. In contrast, a convolutional neural network can learn representational features at varying levels of abstraction while accounting for pixel values and their spatial arrangements. Convolutional operations serve as trainable kernel functions, which connect layers in a convolutional neural network similarly to the way linear connections link layers in an artificial neural network. These kernels are locally sensitive, enabling convolutional networks to learn spatial features from input data more effectively.

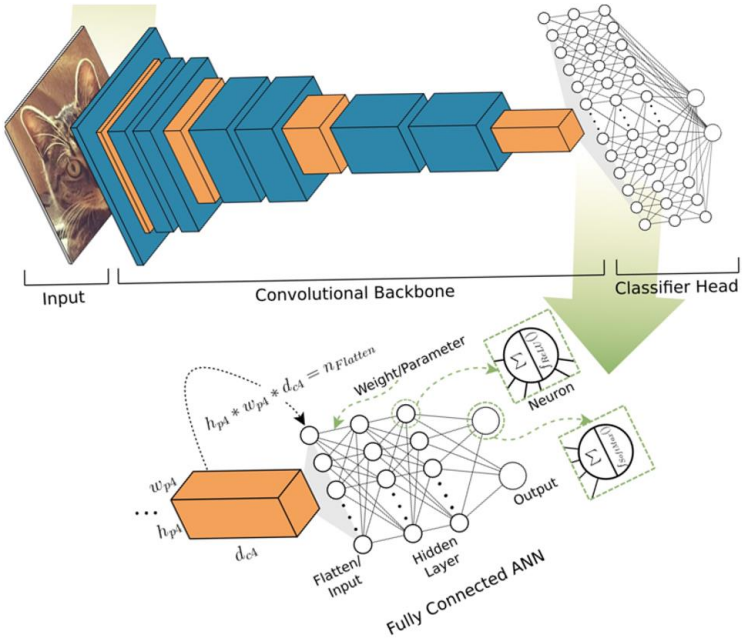

Figure 1. The convolutional backbone is an essential part of the architecture's overall structure. Starting from an input image, it generates feature maps through convolution and reduces their resolution via max pooling operations, resulting in smaller but deeper feature maps. This process continues until the data reaches the classifier head—a fully connected artificial neural network (ANN). The illustration details the transition between the convolutional backbone and the classifier head, as well as the structure of a multi-layer ANN performing classification.

To illustrate the basic functionality of a convolutional neural network, consider a network architecture for image recognition, as shown in Figure 1. This architecture can be divided into three main components: the input, the convolutional backbone, and the classifier head. The input, represented as a two-dimensional array, passes through the convolutional backbone, where a series of convolutional operations, activations, and max-pooling layers extract high-level features. The classifier head, positioned at the end of the backbone, consists of a series of fully connected layers, similar to those in an artificial neural network. It utilizes the features extracted by the backbone to classify the input into predefined output classes and estimate the probability of each class.

3. IMAGE SEGMENTATION

The features extracted by a convolutional backbone contain high-level semantic information, making them useful for predicting the overall class of an image. Image segmentation also relies on these high-level features but aims to classify individual pixels rather than the entire image. This creates a challenge: within a convolutional backbone, feature maps are progressively reduced in resolution as semantic depth increases. Consequently, while these maps carry rich semantic information, they lose precise spatial details, which are essential for accurate pixel-level predictions. This trade-off between feature depth and resolution presents a significant challenge for image segmentation. Furthermore, predicting the correct class for each pixel requires not only spatial accuracy but also an understanding of contextual relationships.

Contextual information for segmentation can span both short and long distances around a pixel [46]. The continuity and size of the semantic segment a pixel belongs to, as well as the density of neighboring segments from other classes or background, play a crucial role in accurate prediction. Therefore, image segmentation is inherently a multi-scale problem, even when working at the pixel level. Advances in the field have focused on addressing this issue by utilizing features from different stages within the network or by preserving and reconstructing high resolution during feature

extraction and prediction. Figure 2 illustrate the applications are shown for all these tasks, which is understood as the prediction of a class for a whole image.

Because image segmentation involves predictions at the pixel level, the benchmark datasets and evaluation metrics differ from those used in traditional image classification tasks like ImageNet. This review employs the PASCAL-VOC [20] test dataset as the primary benchmark. If not specified otherwise, this dataset will be referred to simply as PASCAL-VOC throughout the discussion. PASCAL-VOC was selected due to its established reputation and consistent use over time, allowing for an examination of the evolution of segmentation models since 2014. While newer datasets like Cityscapes offer more challenges, PASCAL-VOC provides valuable insights into long-term developments. The primary performance metric for PASCAL-VOC is the mean Intersection over Union across all classes, which reflects the accuracy of pixel-level predictions.

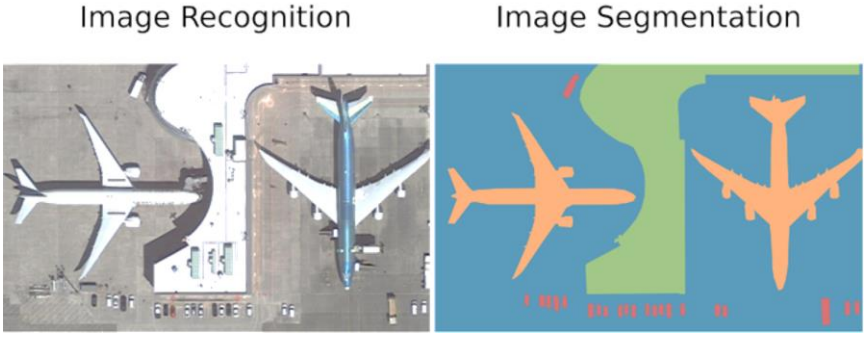

Figure 2. Examples for the tasks of: image recognition, assigns a single label to a whole image; image segmentation, densely classifies each pixel, object detection: locates and classifies specific objects in an image by providing a bounding box; and instance segmentation, provides a segmentation mask for detected objects within a bounding box.

Encoder-decoder models provide a framework for tackling the challenges of image segmentation. In this architecture, the encoder acts as the convolutional backbone, extracting feature maps from input data. The decoder complements the encoder by using these final feature maps along with spatially accurate information from earlier encoder layers. Within the decoder, feature maps are unsampled or deconvolved to restore the original resolution. Additionally, skip connections transfer spatial information from corresponding encoder layers to the decoder, ensuring precise pixel-level predictions. Once the input resolution is fully recovered, the model generates a segmentation mask by assigning class labels to each pixel. This paper of image segmentation architectures will first explore the foundational Fully Convolutional Network (FCN) model, followed by a focus on naïve decoders exemplified by the DeepLab family. Finally, it will delve into more advanced encoder-decoder architectures, highlighting how these models address the challenges of high-resolution segmentation.

4. OBJECT DETECTION

While image recognition and image segmentation can be modeled as classification problems, object detection is a multi-task problem. Predicting the object class remains a classification task, whereas predicting the location—defined by a bounding box around each detected object—is a regression task. Therefore, benchmark datasets must contain additional bounding box annotations, and architectures must handle both classification and regression simultaneously. The benchmark dataset used is the object detection test-dev set from the Microsoft Common Objects in Context (MS-COCO) [53]. The performance metric of interest is the mean Average Precision, or in the case of MS-COCO, simply the Average Precision (AP), which is considered equivalent.

Architectures for object detection can be divided into two groups: two-stage detectors and one-stage detectors. In general, two-stage detectors achieve higher Average Precision, as demonstrated by the

progress made from late 2013 to 2019. One-stage detectors are lightweight in terms of parameters and complexity, and are therefore faster, as measured by processed frames per second [54]. More specifically, two-stage detectors are characterized by a first stage that processes class-agnostic region proposals. The object class prediction of those potential regions and the final bounding box regression are then performed in the second stage. On the other hand, one-stage architectures perform class prediction and bounding box regression in a single shot from the input image. One-stage detectors play an important role in the evolution of convolutional neural network-based object detection and are also widely used in applications and research [55, 56]. Therefore, some popular designs are reviewed below. However, focus is placed on two-stage detectors, particularly the Region-based Convolutional Neural Network (R-CNN) family [57].

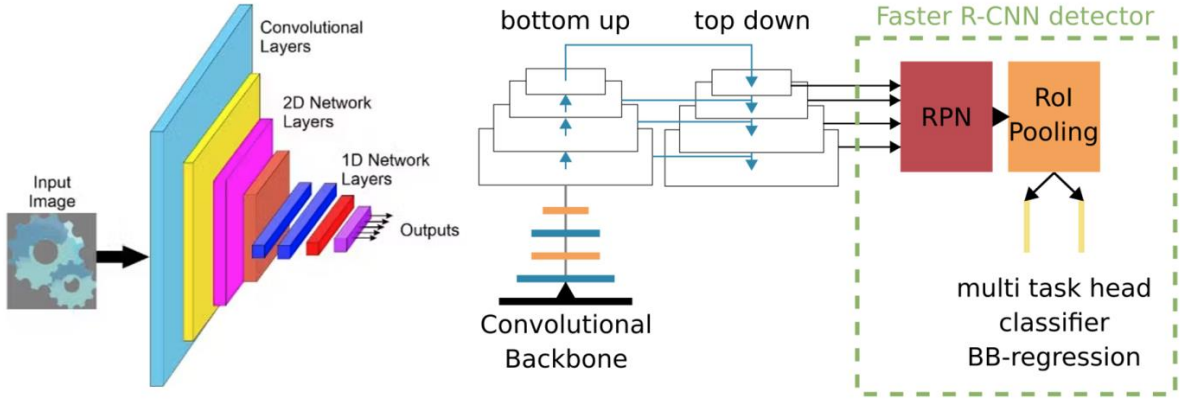

Figure 3. Conceptual overview of Feature Pyramid Network (FPN) and its variants. In the original FPN (left), the last layers of the convolutional backbone (bottom-up path) are paired with a top-down path which uses the bottom-up feature layers to enhance the semantic high-level information with precise localization. Prediction of region proposals is then done on the different layers of the top-down feature pyramid.

An additional change in the overall design of the Spatial Pyramid Pooling Network (SPPNet [58]) made it over one hundred times faster than the original Region-based Convolutional Neural Network (R-CNN) as shown in Figure 3. This speedup was achieved by applying the convolutional backbone along with the adjacent Spatial Pyramid Pooling (SPP) layer to extract features only once on the entire input image, rather than processing each region proposal from selective search separately. As a result, the region proposals derived from selective search are scaled to match the resolution of the shared feature map, allowing corresponding features to be extracted from it. SPPNet with bounding box regression achieved a mean Average Precision of 59.2% on the PASCAL Visual Object Classes (VOC) 2007 dataset, using a fast variant of the Zeiler and Fergus Network (ZFNet) [53] as the convolutional backbone. Although this performance is slightly lower in terms of mean Average Precision compared to R-CNN, SPPNet [53] is 38 times faster.

Fast Region-based Convolutional Neural Network (Fast R-CNN) was introduced [59]. Similar to the Spatial Pyramid Pooling Network, it also exploits shared feature maps for the regions proposed by selective search, establishing the use of shared feature maps as the standard approach for two-stage object detection. Another similarity to SPPNet is the Region of Interest (RoI) pooling layer in Fast R-CNN, which connects the convolutional backbone to a RoI-wise classifier and bounding box regression. It performs max pooling by dividing the RoI provided by selective search into a fixed-resolution feature map. A key difference from SPPNet is that Fast R-CNN introduced a multitask head that performs classification and regression within fully connected layers. Apart from the region proposal, feature extraction and object detection are integrated within one model, and no additional support vector machines are necessary. This was made possible by defining a multitask loss function that sums the losses of the regression and classification heads. Because of this combined design, training became much more efficient. However, since selective search is still used for proposing regions, the model is not end-to-end trainable. Using a Visual Geometry Group 16-layer network as

the backbone, Fast R-CNN achieved a mean Average Precision of 66.9% on the PASCAL Visual Object Classes 2007 dataset. An initial performance on the Microsoft Common Objects in Context dataset was also reported, with an Average Precision of 19.7% [60].

The next successor from the R-CNN family is Faster Region-based Convolutional Neural Network (Faster R-CNN) by Ren et al. [59], developed in 2015. With the introduction of the Region Proposal Network (RPN) module, two-stage object detection finally became end-to-end trainable, unified in a single network performing all tasks needed for competitive object detection results. After the convolutional backbone, the Region Proposal Network—a small fully convolutional network—is inserted. At each position of its sliding window, k translation-invariant anchor boxes are investigated, where k is 9 in the proposed method, corresponding to three scales and three aspect ratios.

Since multiple anchors are predicted at the same sliding window position, heavy overlapping and overly noisy object proposals would be the result. To designate a region proposal as valid, it must either have the highest Intersection over Union (IoU) score or an IoU greater than 0.7; proposals with an IoU less than 0.3 are marked as negative examples. Each positive anchor is then regressed to refine the object boundary and is finally used as a Region of Interest. Apart from the Region Proposal Network module, the architecture is the same as that of Fast R-CNN: RoI pooling with adjacent multitask object class predictor and final bounding box regression. Using a Visual Geometry Group 16-layer network (VGG-16) as the backbone, Faster R-CNN achieves an Average Precision of 21.9% [61]. By changing the backbone to a Residual Network with 101 layers (ResNet-101), the Average Precision increases to 27.2% [61].

In 2017, with advances in convolutional neural network architectures for object detection and image segmentation, He et al. [22] presented Mask Region-based Convolutional Neural Network (Mask R-CNN), an end-to-end trainable deep learning model for instance segmentation. Besides the two heads for classification and bounding box regression, a third head that performs instance mask segmentation was added to the architecture. By utilizing the features within a Region of Interest, binary class-specific masks are predicted using the Fully Convolutional Network for image segmentation introduced by Long et al. [62]. Due to the higher spatial precision required for image segmentation, Region of Interest pooling was adapted to become Region of Interest alignment. Thus, the Region of Interest coordinates were represented as floating-point numbers instead of quantizing them to discrete values. To extract the feature values for one Region of Interest bin, values are sampled at four equally spaced points, bilinearly interpolated, and aggregated using max or average pooling to represent the feature value in the bin. Since standard object detection is still possible with instance segmentation architectures, performance is reported on the Microsoft Common Objects in Context object detection task with an Average Precision of 39.8% [22].

With the Region Proposal Network, Feature Pyramid Network, Region of Interest pooling and later Region of Interest alignment, cascading classifier and regression heads, and finally composite backbones, the most important modules and insights for object detection with two-stage detectors were developed in close relation to the R-CNN family. However, those modules are not exclusively designed for the R-CNN architecture, and most of them are not even limited to two-stage detectors. This means that the introduced models and insights for object detection with convolutional neural networks are flexible and form a foundation for highly task-specific architectures. However, this flexibility comes at the price of highly complex models. One-stage detectors, on the other hand, tend to be less complex, by using predefined anchors to extract features after the convolutional backbone and passing them directly to a detector head, resulting in computationally efficient and more streamlined architectures.

5. CONCLUSION AND OUTLOOK

With the availability of advanced architectures and frameworks, applying them to Earth observation data is the next step in bridging the gap between computer vision and applied Earth observation. However, since overhead imaging data differ from natural images, it is uncertain whether the architectures developed for the datasets mentioned earlier will perform well on Earth observation data. The main differences between Earth observation data and natural images used in those datasets are:

Sensor Perspective: In Earth observation data, the sensor typically captures images from an overhead perspective relative to the scene, whereas natural images are taken from a side-looking perspective. Therefore, the same object classes appear differently in each type of data.

Data Channels: Computer vision often utilizes three-channel RGB images, whereas Earth observation data frequently consist of multichannel image stacks with more than three channels. This difference must be considered, especially when transferring models from computer vision to Earth observation applications.

Sensor and Platform Variability: Input data in computer vision models usually come from the same sensor and platform. In Earth observation, both the sensor and platform can vary, necessitating the incorporation of data fusion into the model.

Object Orientation: Objects in overhead images lack a general orientation, meaning that objects of the same class can appear at any rotation within 360 degrees. This must be accounted for in the training data, the architecture, or both. In contrast, natural images often have a more defined top and bottom, resulting in a general orientation of objects.

Object Position and Resolution: In natural images, objects of interest tend to be centered and captured in high resolution. In Earth observation data, objects can be off-nadir or located at the edges of images with coarse resolution.

Object Density and Heterogeneity: Objects or classes in Earth observation data tend to be more densely packed and heterogeneous than in natural images.

Because of these characteristics, the tasks of image recognition, image segmentation, and object detection are more challenging with observation data. To address this problem, deep learning models trained on computer vision datasets are fine-tuned on Earth observation datasets. These Earth observation datasets are often smaller, so it is common practice to refine models that have already learned to hierarchically extract features from imagery data on larger computer vision datasets. This refinement of models originally trained on other datasets is known as transfer learning. Optimized parameters for a computer vision task are adapted to an Earth observation task, effectively transferring learned skills to a different context. However, even for transfer learning, specialized Earth observation datasets must be created. Reflecting on the differences and similarities between data and architectures in computer vision and Earth observation, we assert that advances in computer vision must be adapted to suit Earth observation applications. Therefore, a thorough understanding of deep learning concepts is crucial for assessing and modifying models appropriately. With the extensive introduction to convolutional neural networks (CNNs) provided, we have established a foundation to closely examine the application of deep learning in natural observation research.

REFERENCES

[1] J. Huo, S.F. Quan, J. Roveda, A. Li, Coupling analysis of heart rate variability and cortical arousal using a deep learning algorithm, Plos one 18(4) (2023) e0284167.

[2] R. Li, S. Sun, M. Elhoseiny, P. Torr, OxfordTVG-HIC: Can Machine Make Humorous Captions from Images?, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20293-20303.

[3] X. Chen, W. Liu, X. Liu, Y. Zhang, T. Mei, A cross-modality and progressive person search system, Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4550-4552.

[4] L.V. Jospin, H. Laga, F. Boussaid, W. Buntine, M. Bennamoun, Hands-on Bayesian neural networks—A tutorial for deep learning users, IEEE Computational Intelligence Magazine 17(2) (2022) 29-48.

[5] X. Chen, X. Liu, K. Liu, W. Liu, T. Mei, A baseline framework for part-level action parsing and action recognition, arXiv preprint arXiv:2110.03368 (2021).

[6] X. Chen, X. Liu, W. Liu, K. Liu, D. Wu, Y. Zhang, T. Mei, Part-level Action Parsing via a Pose-guided Coarse-to-Fine Framework, 2022 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2022, pp. 419-423.

[7] M. Qu, X. Chen, W. Liu, A. Li, Y. Zhao, ChatVTG: Video Temporal Grounding via Chat with Video Dialogue Large Language Models, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1847-1856.

[8] L. Yang, Z. Zhang, J. Han, B. Zeng, R. Li, P. Torr, W. Zhang, Semantic Score Distillation Sampling for Compositional Text-to-3D Generation, arXiv preprint arXiv:2410.09009 (2024).

[9] Z. Gui, S. Sun, R. Li, J. Yuan, Z. An, K. Roth, A. Prabhu, P. Torr, kNN-CLIP: Retrieval Enables Training-Free Segmentation on Continually Expanding Large Vocabularies, arXiv preprint arXiv:2404.09447 (2024).

[10] M. Yin, T. Li, H. Lei, Y. Hu, S. Rangan, Q. Zhu, Zero-Shot Wireless Indoor Navigation through Physics-Informed Reinforcement Learning, 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 5111-5118.

[11] J. Huo, H. Li, J. Roveda, S.F. Quan, A. Li, A Multi-task Deep Learning Algorithm for Sleep Stage Scoring and Sleep Arousal Detection, Authorea Preprints (2023).

[12] H. Guo, A.B. Tikhomirov, A. Mitchell, I.P.J. Alwayn, H. Zeng, K.C. Hewitt, Real-time assessment of liver fat content using a filter-based Raman system operating under ambient light through lock-in amplification, Biomedical Optics Express 13(10) (2022) 5231-5245.

[13] H. Guo, B.L. Gala-Lopez, I.P. Alwayn, K.C. Hewitt, Liver discard rate due to conservative estimations of steatosis: an inference-based approach, medRxiv (2023) 2023.12. 04.23299406.

[14] J. Huo, Machine Learning Application in Sleep Disorder Analysis, The University of Arizona, 2023.

[15] C. Ding, T. Yao, C. Wu, J. Ni, Deep Learning for Personalized Electrocardiogram Diagnosis: A Review, arXiv preprint arXiv:2409.07975 (2024).

[16] J. Huo, S.F. Quan, J. Roveda, A. Li, BASH-GN: a new machine learning–derived questionnaire for screening obstructive sleep apnea, Sleep and Breathing 27(2) (2023) 449-457.

[17] J. Huo, Y. Wang, N. Wang, W. Gao, J. Zhou, Y. Cao, Data-driven design and optimization of ultra-tunable acoustic metamaterials, Smart Materials and Structures 32(5) (2023) 05LT01.

[18] J. Huo, Y. Wang, Y. Cao, 3D computational study of arc splitting during power interruption: the influence of metal vapor enhanced radiation on arc dynamics, Journal of Physics D: Applied Physics 54(8) (2020) 085502.

[19] Z. Hu, Y. Sun, Y. Yang, Switch to generalize: Domain-switch learning for cross-domain few-shot classification, International Conference on Learning Representations, 2022.

[20] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303-338.

[21] Z. Hu, Y. Sun, Y. Yang, Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation, The Eleventh International Conference on Learning Representations, 2023.

[22] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961-2969.

[23] X. Chen, X. Liu, W. Liu, X.-P. Zhang, Y. Zhang, T. Mei, Explainable person re-identification with attribute-guided metric distillation, Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11813-11822.

[24] Z. Hu, Y. Sun, Y. Yang, J. Zhou, Divide-and-regroup clustering for domain adaptive person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 980-988.

[25] X. Chen, W. Liu, X. Liu, Y. Zhang, J. Han, T. Mei, MAPLE: Masked pseudo-labeling autoencoder for semi-supervised point cloud action recognition, Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 708-718.

[26] Z. Hu, Y. Sun, J. Wang, Y. Yang, DAC-DETR: Divide the attention layers and conquer, Advances in Neural Information Processing Systems 36 (2024).

[27] X. Chen, W. Liu, Q. Bao, X. Liu, Q. Yang, R. Dai, T. Mei, Motion Capture from Inertial and Vision Sensors, arXiv preprint arXiv:2407.16341 (2024).

[28] Z. Hu, J. Ye, Y. Zhang, X. Wang, Seeing is Not Always Believing: An Empirical Analysis of Fake Evidence Generators, 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P), IEEE, 2024, pp. 560-579.

[29] Y. Zhang, Z. Hu, X. Wang, Y. Hong, Y. Nan, X. Wang, J. Cheng, L. Xing, Navigating the Privacy Compliance Maze: Understanding Risks with {Privacy-Configurable} Mobile {SDKs}, 33rd USENIX Security Symposium (USENIX Security 24), 2024, pp. 6543-6560.

[30] B. Song, X. Wang, P. Sun, A. Boukerche, Robust COVID-19 vaccination control in a multi-city dynamic transmission network: A novel reinforcement learning-based approach, Journal of Network and Computer Applications 219 (2023) 103715.

[31] Z. Shou, X. Chen, Y. Fu, X. Di, Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment, Transportation Research Part C: Emerging Technologies 137 (2022) 103560.

[32] S. Liu, Y. Wang, X. Chen, Y. Fu, X. Di, SMART-eFlo: An integrated SUMO-gym framework for multi-agent reinforcement learning in electric fleet management problem, 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2022, pp. 3026-3031.

[33] X. Chen, S. Liu, X. Di, A hybrid framework of reinforcement learning and physics-informed deep learning for spatiotemporal mean field games, In Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, ACM DIgital Library, 2023.

[34] F. Zhou, C. Zhang, X. Chen, X. Di, Graphon Mean Field Games with A Representative Player: Analysis and Learning Algorithm, arXiv preprint arXiv:2405.08005 (2024).

[35] X. Chen, S. Liu, X. Di, Learning Dual Mean Field Games on Graphs, ECAI, 2023, pp. 421-428.

[36] S. Liu, X. Chen, X. Di, Scalable Learning for Spatiotemporal Mean Field Games Using Physics-Informed Neural Operator, Mathematics 12(6) (2024) 803.

[37] X. Chen, Z. Li, X. Di, Social learning in Markov games: Empowering autonomous driving, 2022 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2022, pp. 478-483.

[38] X. Chen, X. Di, Z. Li, Social Learning for Sequential Driving Dilemmas, Games 14(3) (2023) 41.

[39] M. Yin, A.K. Veldanda, A. Trivedi, J. Zhang, K. Pfeiffer, Y. Hu, S. Garg, E. Erkip, L. Righetti, S. Rangan, Millimeter wave wireless assisted robot navigation with link state classification, IEEE Open Journal of the Communications Society 3 (2022) 493-507.

[40] V. Semkin, M. Yin, Y. Hu, M. Mezzavilla, S. Rangan, Drone detection and classification based on radar cross section signatures, 2020 International Symposium on Antennas and Propagation (ISAP), IEEE, 2021, pp. 223-224.

[41] M. Yin, Millimeter Wave Wireless Assisted Indoor Robot Navigation, New York University Tandon School of Engineering, 2024.

[42] K. Pfeiffer, Y. Jia, M. Yin, A.K. Veldanda, Y. Hu, A. Trivedi, J. Zhang, S. Garg, E. Erkip, S. Rangan, Path planning under uncertainty to localize mmWave sources, 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 3461-3467.

[43] Y. Hu, M. Yin, W. Xia, S. Rangan, M. Mezzavilla, Multi-frequency channel modeling for millimeter wave and thz wireless communication via generative adversarial networks, 2022 56th Asilomar Conference on Signals, Systems, and Computers, IEEE, 2022, pp. 670-676.

[44] S. Cao, J. Xiao, Human-Robot Complementary Collaboration for Flexible and Precision Assembly, 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 12971-12977.

[45] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3642-3649.

[46] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60(6) (2017) 84-90.

[47] H. Guo, A.E. Stueck, J.B. Doppenberg, Y.S. Chae, A.B. Tikhomirov, H. Zeng, M.A. Engelse, B.L. Gala-Lopez, A. Mahadevan-Jansen, I.P. Alwayn, Evaluation of minimum-to-severe global and macrovesicular steatosis in human liver specimens: a portable ambient light-compatible spectroscopic probe, medRxiv (2023) 2023.12. 04.23299259.

[48] H. Guo, V.S. Zions, B.A. Law, K.C. Hewitt, Potential of Raman-Reflectance Combination in Quantifying Liver Steatosis and Fat Droplet Size: Evidence From Monte Carlo Simulations and Phantom Studies, Journal of Biophotonics (2024) e202400156.

[49] H. Guo, A.E. Stueck, J.B. Doppenberg, Y.S. Chae, A.B. Tikhomirov, H. Zeng, B.L. Gala-Lopez, A. Mahadevan-Jansen, M.A. Engelse, I.P. Alwayn, Assessment of liver steatosis using an ambient light-compatible Raman system: enhancing specificity with supplementary reflectance information, Biomedical Vibrational Spectroscopy 2024: Advances in Research and Industry, SPIE, 2024, p. PC128390B.

[50] H. Guo, A.E. Stueck, A.B. Tikhomirov, H. Zeng, I.P. Alwayn, B.L. Gala-Lopez, A. Mahadevan-Jansen, A.K. Locke, K.C. Hewitt, Evaluation of Steatosis in Human Liver Specimens Using an Ambient Light-compatible Raman Spectroscopy Approach, Bio-Optics: Design and Application, Optica Publishing Group, 2023, p. JTu4B. 26.

[51] J.E. Ball, D.T. Anderson, C.S. Chan, Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community, Journal of applied remote sensing 11(4) (2017) 042609-042609.

[52] E. Imani, G. Zhang, R. Li, J. Luo, P. Poupart, P.H. Torr, Y. Pan, Label Alignment Regularization for Distribution Shift, Journal of Machine Learning Research 25(247) (2024) 1-32.

[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740-755.

[54] K. Huang, X. Chen, X. Di, Q. Du, Dynamic driving and routing games for autonomous vehicles on networks: A mean field game approach, Transportation Research Part C: Emerging Technologies 128 (2021) 103189.

[55] Y. Fu, A. Jain, X. Di, X. Chen, Z. Mo, DriveGenVLM: Real-world Video Generation for Vision Language Model based Autonomous Driving, arXiv preprint arXiv:2408.16647 (2024).

[56] X. Chen, F. Yongjie, S. Liu, X. Di, Physics-informed neural operator for coupled forward-backward partial differential equations, 1st Workshop on the Synergy of Scientific and Machine Learning Modeling@ ICML2023, 2023.

[57] S. Sun, R. Li, P. Torr, X. Gu, S. Li, Clip as rnn: Segment countless visual concepts without training endeavor, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13171-13182.

[58] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence 37(9) (2015) 1904-1916.

[59] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE transactions on pattern analysis and machine intelligence 39(6) (2016) 1137-1149.

[60] R. Girshick, Fast r-cnn, arXiv preprint arXiv:1504.08083 (2015).

[61] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026-1034.

[62] G. McCracken, The long interview, Sage publications1988.