Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19823-3

A systematic literature review of visual feature learning: deep learning techniques, applications, challenges and future directions

Mohammed Abdullahi[1] · Olaide Nathaniel Oyelade[5] · Armand Florentin Donfack Kana[1] · Mustapha Aminu Bagiwa[1] · Fatimah Binta Abdullahi[1] · Sahalu Balarabe Junaidu[1] · Ibrahim Iliyasu[2] · Ajayi Ore-ofe[3] · Haruna Chiroma[4]

Received: 29 February 2024 / Revised: 24 June 2024 / Accepted: 4 July 2024

Abstract
Visual Feature Learning (VFL) is a critical area of research in computer vision that involves the automatic extraction of features and patterns from images and videos. The applications of VFL are vast, including object detection and recognition, facial recognition, scene understanding, medical image analysis, and autonomous vehicles. In this paper, we propose to conduct extensive systematic literature review (SLR) on VFL based on deep learning algorithms. The paper conducted an SLR covering deep learning algorithms such as Convolutional Neural Networks (CNNs), Autoencoders, and Generative Adversarial Networks (GANs) including their variants. The review highlights the importance of VFL in computer vision and the limitations of traditional feature extraction techniques. Furthermore, it provides an in-depth analysis of the strengths and weaknesses of various deep learning algorithms for solving problems in VFL. The discussion of the applications of VFL provides an insight into the impact of VFL on various industries and domains. The review also analyzed the challenges faced by VFL, such as data scarcity and quality, overfitting, generalization, interpretability, and explainability. The discussion of future directions for VFL includes hybrid techniques, unsupervised feature learning, continual learning, attention-based models, and explainable AI. These techniques aim to address the challenges faced by VFL and improve the performance of the models. The systematic literature review concludes that VFL is a rapidly evolving field with the potential to transform many industries and domains. The review highlights the need for further research in VFL and emphasizes the importance of responsible use of VFL models in various applications. The review provides valuable insights for researchers and practitioners in the field of computer vision, who can use these insights to enhance their work and ensure the responsible use of VFL models.

Keywords Visual feature learning · Deep learning · Convolutional neural network · Generative adversarial networks

Extended author information available on the last page of the article

Published online: 20 July 2024

1 Introduction

1.1 Background and motivation

Visual feature learning has gained significant attention in the field of computer vision and image processing. With the rapid advancement of deep learning techniques, visual feature learning has become a crucial component in various applications, including object recognition, image classification, and image generation. Deep learning models, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), have demonstrated remarkable performance in extracting and representing meaningful visual features from raw input data. This systematic literature review aims to provide a comprehensive overview of the current state of visual feature learning techniques, applications, challenges, and future directions.

1.2 Research questions and objectives

The primary research questions addressed in this review are as follows:

RQ1.What are the state-of-the-art deep learning techniques used for visual feature learning?
RQ2.What are the main applications of visual feature learning in computer vision?
RQ3.What are the existing challenges and limitations in visual feature learning?
RQ4.What are the potential future directions for advancing visual feature learning?

The objectives of this review are to systematically analyze the existing literature, synthesize the findings, and provide insights into the current trends and advancements in visual feature learning.

1.3 Brief overview of the related existing reviews

To establish the context for this review, a brief overview of existing reviews on visual feature learning is provided. Previous reviews have primarily focused on specific aspects of visual feature learning, such as deep learning architectures or applications in specific domains. However, a comprehensive review that encompasses a wide range of deep learning techniques, applications, challenges, and future directions is still lacking. This review aims to bridge this gap by providing a holistic view of the field.

We contrast the contribution of this study with some relevant and similar related works focused on the summarization of studies on visual feature learning using deep learning methods. In Table 1, a listing of similar survey or review studies is presented, and a comparison of their approaches are made with the approach presented in our study. Note that most of the approaches used for review studies often follow the methods of systematic literature review, a simple survey, a narrative or literature review, and meta-analysis. In this study we adopted the method of systematic literature review. This selection is motivated from the perspective that is more exhaustive than the traditional literature review method. In addition, the method allows for considering published and archived research studies while addressing some important questions. Also, our study benefited from the method since it supports appropriate search methodology with provision for analysis of data extracted during the search process. All these contradict the traditional literature review

Table 1 A summary and comparative analysis of the proposed review study with similar related works focused on visual feature learning using the approach of machine learning and deep learning

Ref	Year of Publication	Year of Coverage	Number of papers reviewed	Methodology	Description of study	Comparison with our study
[1]	2020	2012–2019	47	Survey	A review of the application of deep learning in medical image classification and segmentation	The study is focused strictly on feature learning on medical image with no wider consideration of other domains
[2]	2022	2014–2021	121	Survey	Human pose, hand and mesh estimation using deep learning: a survey	Summarization of studies which have applied deep learning to estimation of human pose alone
[3]	2020	2012–2020	135	Systematic review	Deep learning for misinformation detection on online social networks: a survey and new perspectives	The study is narrowed to visual feature learning in misinformation detection (MID), whereas our proposed method also provides a review in this area
[4]	2021	2000–2020	150	Systematic review	Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues	A review addressing the question of what deep learning methods are used for detection of objects in a scene where self-driving cars operate. This proposed study improved this by considering reviews on self-driving car itself and car plate number
[5]	2021	2010–2021	123	Systematic review	Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions	Our proposed study provides an insightful review on both the techniques and application of deep learning to selected domains, while this study focused on the evolution of deep learning techniques
[6]	2019	2010–2018	146	Survey	Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey	Limited to self-supervised deep learning models, contrary to the review of all forms of visual feature learning presented in our study
[7]	2023	1988–2023	274	Review	Auto-Encoders in Deep Learning—A Review with New Perspectives	Their work reviews the application of auto-encoders whereas our proposed study covers different techniques for visual feature learning including auto-encoders

Table 1 (continued)

Ref	Year of Publication	Year of Coverage	Number of papers reviewed	Methodology	Description of study	Comparison with our study
[8]	2022	2013–2021	151	Systematic Review	Visual Feature Learning on Video Object and Human Action Detection: A Systematic Review	Only video documents on face detection are investigated in their study, as compared to our study which covers both video and image medium for visual feature learning using deep learning
[9]	2022	2013–2022	73	Survey	A survey: object detection methods from CNN to transformer	The study has shared similarity with our approach, except that our survey expands the domain of application
[10]	2023	1998–2022	164	Review	Deep learning techniques for remote sensing image scene classification: A comprehensive review, current challenges, and future directions	Strictly restricted to surveying studies applying deep learning to remote sensing
[11]	2020	2009–2020	180	Meta-analysis	A State-of-the-art Survey on Deep Learning Approaches in Detection of Architectural Distortion from Digital Mammographic Data	Visual feature learning summarizing the use of deep learning to single abnormality in medical images. Our review considers different abnormalities in medical images, including other domains as well
[12]	2022	2011–2021	142	Systematic Review	Vision-Based Autonomous Vehicle Systems Based on Deep Learning: A Systematic Literature Review	Their study is focused on review on autonomous vehicles, which is a subsection in the domain of application of deep learning reviewed in our study
Current review	2023	2010–2023	190	Systematic Review	A wider systematic review of literatures on visual feature learning cutting across domains including object detection and recognition, facial recognition, scene understanding, medical image analysis, autonomous vehicle, vehicle plate number recognition, and video documents	

method which collects information from studies and discusses the metadata based on the information collected and reveals the current challenge in the domain. The simple survey method is often focused on providing conceptual understanding on a particular subject or domain and advancing discussion to reveal the current state-of-the-art. As a result, most survey papers trail behind literature review papers both in content richness and volume of content. Lastly, the method of meta-analysis has also been promoted for studying related works in research review process. This method draws up and provides statistical information based on the studies which have been reviewed. The approach often applies a specified search protocol on different databases. Therefore, the use of the systematic literature review in this study is suitable for addressing the aim of the study. The comparison outlined on Table 1 therefore demonstrates that the review presented in this study is exhaustive, having wide coverage, outlining the current challenges in the field, and presenting potential future research direction.

The major highlight of this survey is its comprehensive approach to reporting visual feature learning, thereby serving as a differentiator comparatively with the studies [1–11], and [12]. For instance, while [4, 11] and [1, 12] are strictly focused on visual feature learning in autonomous vehicle and abnormality detection in digital mammography leading to breast cancer detection respectively, our study proposes an encompassing report on major techniques and domains of application of visual feature learning including medical and autonomous vehicle. Furthermore, the survey described in [5] is more focused on the technique of deep learning itself. On the contrary, while our study tried to present foundation models based on deep learning, we also advanced the survey to report and discuss different approaches to visual feature learning as applicable to some selected domains. The works of [7] and [9] are more aimed at understanding visual features learning using transformer models. To ensure that our study does not omit this very relevant and evolving encoder-decoder based techniques, we have dedicated a subsection to detailed review and discussion on these approaches as it relates with visual features learning.

This study demonstrates a wider systematic review of literature on visual feature learning as against the self-supervised learning approach which was the focus of the study in [6]. Moreover, our study discusses visual feature learning across domains including object detection and recognition, facial recognition, scene understanding, medical image analysis, autonomous vehicle, vehicle plate number recognition, and video documents. We do not restrict the review to video objects as in [8], but rather elaborated the review to both image and video objects. This shows that discussions and reviews in our study are also applicable to domain specific survey on remote sensing which is reported in [10], because scene understanding is also elaborately reviewed in our study. The survey reported in [2] and [3] are basically aimed at providing readers with the role of deep learning in detection of misinformation on social networks and understanding human pose respectively.

2 Methodology

2.1 Search strategy and selection criteria

A systematic search strategy was employed to identify relevant articles from various academic databases, including IEEE Xplore, ACM Digital Library, PubMed, and Google Scholar. The search terms included combinations of keywords related to visual feature learning, deep learning, convolutional neural networks, generative adversarial networks, and computer vision. The inclusion criteria encompassed studies published in peer-reviewed journals and conference proceedings between a specified time range. The exclusion criteria consisted of non-English publications, duplicate studies, and irrelevant articles.

2.2 Data extraction and analysis process

The data extraction process involved screening the titles and abstracts of the retrieved articles to identify relevant studies. Full-text articles that met the inclusion criteria were then thoroughly reviewed, and pertinent information was extracted. The extracted data included the authors' names, publication year, research objectives, methodology, deep learning techniques employed, application domains, and major findings. The data were organized and analyzed to identify patterns, trends, and key insights related to visual feature learning.

The approach applied for the data collection, filtering and application for the systematic review in this study is presented in Fig. 1 and 2. The illustration follows a five-phase approach including preliminary, database look-up, search filter, consideration of sieved data, and exploration phases. Each phase is a precursor to another lower stage demonstrating a progressive methodology to the systematic literature review process. In phase 1 of the methodology, the review process was planned, scoped and conducted in a manner as to set up the article identification process. During the planning sub-phase, the research questions driving the systematic review process were derived. Furthermore, the scope for the category and type of literature was drawn up to allow for limiting what get considered during the elimination and screening phase. Finally, for phase 1, we applied a step-by-step approach to link-up with the databases for search. This led to the initialization of the phase process which is focused on database look-up. In this phase, all relevant databases were identified and explored for suggestive articles. The resulting samples of articles found yielded about 249 items only after some have been de-identified. Phase three process allows for applying screening to the de-identified items so that some irrelevant articles or items are eliminated from the data provisioned for consideration. The screening process resulted in the elimination of about 59 items so that only 190 samples were sieved out for phase four. The phase ensures that journal articles, conference proceedings and preprint versions are retained, while the exclusion criteria at that phase applied keywords, title and abstract to exclude non-relevant articles. In phase five of the systematic literature review process, the retained data were applied for explorative and analysis task for a thorough review process.

2.3 Quality assessment

A quality assessment was conducted to evaluate the rigor and reliability of the included studies. The assessment criteria considered factors such as study design, sample size, data collection methods, and statistical analysis techniques. Studies that met the predefined quality criteria were given higher weight in the analysis process, ensuring that the findings are based on robust research.

RQ1. What are the state-of-the-art deep learning techniques used for visual feature learning?

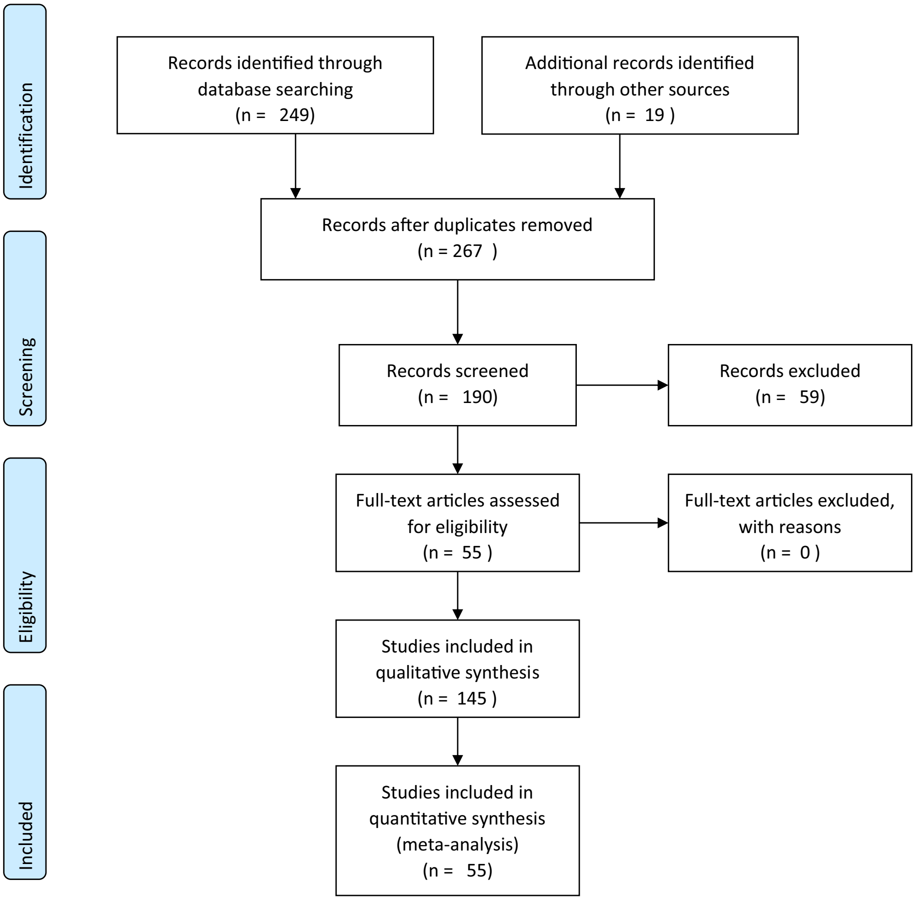
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 249)
Additional records identified through other sources (n = 19)
Records after duplicates removed (n = 267)
Records screened (n = 190)
Records excluded (n = 59)
Full-text articles assessed for eligibility (n = 55)
Full-text articles excluded, with reasons (n = 0)
Studies included in qualitative synthesis (n = 145)
Studies included in quantitative synthesis (meta-analysis) (n = 55)

Fig. 1 A preferred reporting items for systematic reviews and meta-analysis (PRISMA) flow diagram for this study

3 Deep learning techniques for visual feature learning

The evolvement of deep learning techniques has become a motivation for improved methods for visual feature learning. This is necessary considering the feature learning process which largely depends on the neural networks architectural and operational pattern. Their architectures are mostly composed of layers of non-linear neurons which are able to learn by iteratively adjusting the weights and connections of each layer. This architectural layout is motivated from the highly parallel connected networks of non-linear neurons in the human brain. The composition of these neural networks allows for learning patterns represented in visual, audio, textual and other mediums. The process of learning the representation of these patterns allows for feature learning since the features are characteristic of an object which needs to be recognized from a stream of features. It is now established that feature learning has promoted exceptional research breakthroughs in recognizing objects or understanding language. This is largely made possible because of the increasing number of layers of neurons operating on vector representation of features contained in input media so

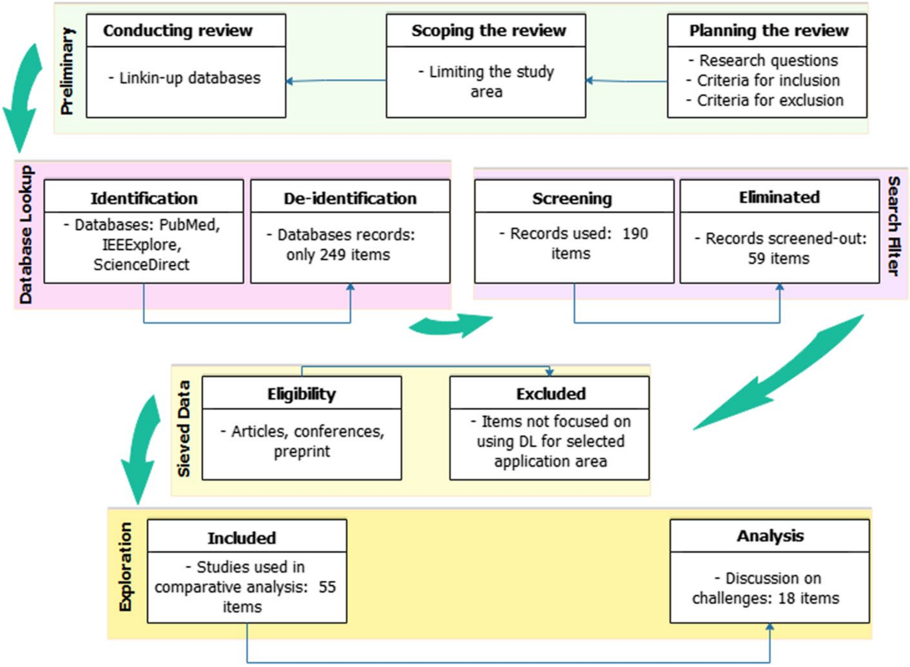
Preliminary
Conducting review
- Linkin-up databases
Scoping the review
- Limiting the study area
Planning the review
- Research questions
- Criteria for inclusion
- Criteria for exclusion
Database Lookup
Identification
- Databases: PubMed, IEEExplore, ScienceDirect
De-identification
- Databases records: only 249 items
Screening
- Records used: 190 items
Eliminated
- Records screened-out: 59 items
Search Filter
Sieved Data
Eligibility
- Articles, conferences, preprint
Excluded
- Items not focused on using DL for selected application area
Exploration
Included
- Studies used in comparative analysis: 55 items
Analysis
- Discussion on challenges: 18 items

Fig. 2 An illustration of the methodology followed to achieved the systematic literature review presented in this study

that an established learning is obtained which follows the stochastic gradient of an objective function [13]. The gradient function estimates and reduces the error value by adjusting the vector representation or weights of the architecture to obtain better performance. The adjustment of the weights follows an iterative approach which depends on large training inputs to optimize the final values of the weights over the iterative time span known as the training time of the neural networks. The final values assigned to the weights of the neurons on each layer of the neural architecture represent the corresponding trained model of the architecture. The trained model can then be applied to test examples that come from a different distribution other than those training inputs used for optimizing the weights of the model. Moreover, the ability of the trained model to generalize well through effective discrimination of features in the test examples often depend on some architectural layout such as the depth and shallowness of the neural network. Typically, it is desirable that deep neural networks should have considerable depth of layers. The deeper the layers, the more the number of parameters to be trained to allow the neural network to learn complex, non-linear relationships between inputs and outputs.

The design, training, and applicability of deep learning methods leverages techniques from several fields of study such as Computer Science, Mathematics, Biology, Physics, Neurology and Neuroscience, Engineering and Cognitive science. These fields have contributed to the evolvement of deep learning using knowledge on graphics, algorithms, system architecture, statistics, mathematical functions, information retrieval, robotics, speech, image processing, biology and cognitive science. This had led to outstanding performance and achievements of using deep learning for visual feature learning. Several deep learning tasks requiring visual feature learning such as image classification[14–18], semantic segmentation[1, 19], image colorization [20, 21], object detection[22], object localization[23, 24], object conceptualization, object recognition [25], virtual assistants [26], face recognition, motion detection [27, 28], surveillance [29, 30], pose estimation [2, 31],structural biology [32], image captioning [33, 34], natural language processing [35–37],news aggregation[38], social networks [3], self-driving cars [4], fraud detection [39, 40], sequential learning [41], and embedding [42], have benefited from the advancement. This advancement demonstrates the progression from the traditional and very error-prone method of visual feature extraction of manual acquisition of features which is now being replaced by unsupervised or semi-supervised learning. Meanwhile, it is worth noting that even the traditional machine learning methods are just a representation of feature learning which is now being replaced by the deep learning techniques discussed in this section. Even though deep learning models take a significant time to train, they remain very outstanding in solving tasks associated with visual feature learning. The success of the use of deep learning reflects on its ability to intelligently learn fine-grained features from raw data, leveraging on their depths for effective learning, and the broad spectrum of problems such models can solve [43]. However, the heterogeneity of real-life problems makes it challenging to adopt a single neural architecture to solve visual feature learning across several problems domains. This implies that each domain of application of deep learning requires a unique and domain-specific trained neural architecture. This section demonstrates the appropriateness of deep learning models which are broadly categorized into discriminative learning, generative learning, and hybrid learning [5] to achieve the task of visual feature learning.

Visual feature learning techniques discussed in this section have their architectural design patterned after some distinguishing methods. These methods provide means for differentiating the representational approach to learning visual features in images and video inputs. The discussion of the techniques follows the end-to-end learning, one-shot learning, zero-shot learning, transfer learning, supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning which are illustrated using Fig. 3. The end-to-end learning allows for designing neural network architectures which are capable of learning visual features representational of complex tasks using a single network which combines all subtasks of the main tasks into one operation. This approach differs from the two-stage algorithms which often allow for decomposing subtasks into smaller solutions obtained from some neural networks. Transfer learning supports the process of learning visual features in a particular domain and adapting such learning to extract visual features from another domain with the aim to improve the target task by incorporating information from the source task [44]. On the other hand, one-shot learning and zero-shot learning have some kind of uniqueness since stand to benefit from the transfer learning method. Zero-shot learning depends on some auxiliary information to learn and distinguish the visual features of objects that it has seen from those not seen before. This method of visual feature learning is different from the one-shot learning which works based on evaluating the difference between two samples to learn if they represent the same object. Other forms of learning are supervised learning [45, 46], unsupervised learning [6, 47], semi-supervised learning [48], deep reinforcement learning [49], and hierarchical feature extraction efficient learning algorithms [50, 51].

In Fig. 4, hierarchical illustrations of different neural networks are presented to demonstrate their disposition to visual feature learning process. These networks include the convolutional neural network (CNN), Graph convolution networks (GCNs), recurrent neural network (RNN) generative adversarial (GAN), Radial Basis Function Networks (RBFNs), Multilayer perceptron (MLPs), Self -organizing Maps (SOMs), Deep Belief Networks (DBNs), Restricted Boltzmann Machines (RBMs), Autoencoders, recursive neural networks. The most crucial type of neural network for computer vision is convolutional neural networks.

3.1 Classical convolutional neural networks (CNNs)

The traditional approach to the application of convolutional neural networks (CNNs) to visual feature learning is outstanding and has motivated for designing new techniques. This method of feature learning is the fundamental and most often used technique for extracting video and image features for classification task, and other related tasks. They follow the method of stacking layers of convolution pooling blocks with variety of filter sizes and counts for adapting to the nature of the problem. For instance, when extracting features for characterization of abnormalities in medical images, the nature of the representation of the abnormalities in the samples will influence the choice of values used for depth of layers in the architecture, filter sizes and count for each layer, and the type of regularizers to use. Furthermore, the number of channels obtainable in the input samples has a shared influence on the architectural layout and parameter sizing. A typical CNNs architecture shown in Fig. 5, follows its convolutional-pooling layers and blocks with fully connected layers, possibly a dropout layer for dimensionality reduction to improve performance, and a softmax function in case of multiclass classification.

The challenge of crafting new architectures for every specific problem has motivated for the adoption of benchmark neural architectures through transfer learning or architectural tweaking to arrive at new models. Examples of such benchmark neural network architectures are CiFarNet [53], AlexNet [54], GoogLeNet [55], Inception (Inception v1, Inception-v3, Inception-v4) [56], Xception [57], DenseNet (DenseNet-121, DenseNet-169)

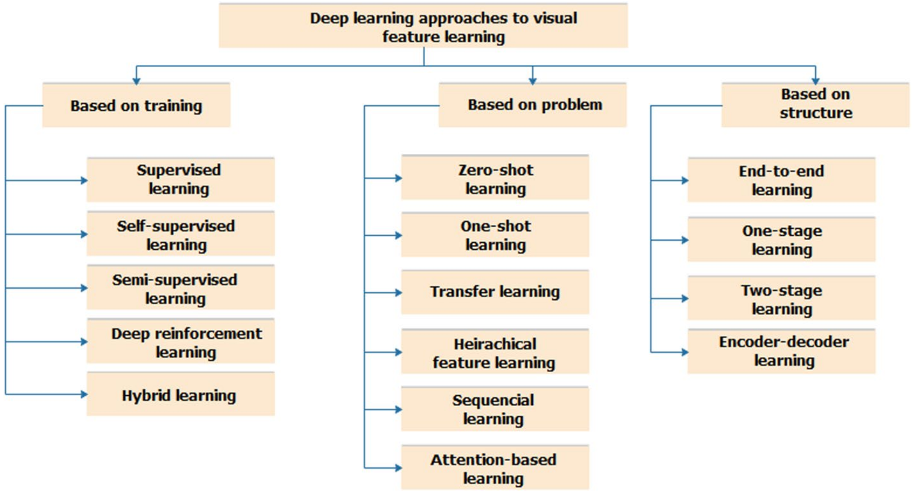

Fig. 3 A diagrammatic illustration on the branching of the various types of visual feature learning approaches

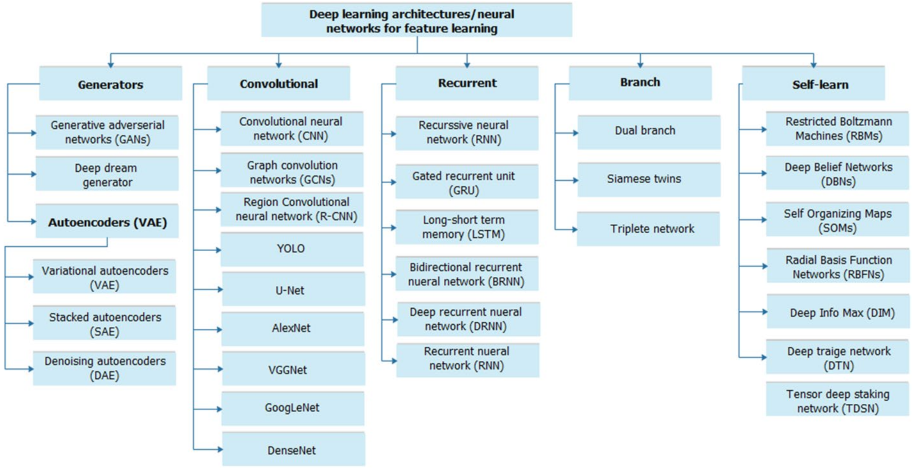
Deep learning architectures/neural networks for feature learning
Generators
Generative adverserial networks (GANs)
Deep dream generator
Autoencoders (VAE)
Variational autoencoders (VAE)
Stacked autoencoders (SAE)
Denoising autoencoders (DAE)
Convolutional
Convolutional neural network (CNN)
Graph convolution networks (GCNs)
Region Convolutional neural network (R-CNN)
YOLO
U-Net
AlexNet
VGGNet
GoogLeNet
DenseNet
Recurrent
Recurssive neural network (RNN)
Gated recurrent unit (GRU)
Long-short term memory (LSTM)
Bidirectional recurrent nueral network (BRNN)
Deep recurrent nueral network (DRNN)
Recurrent nueral network (RNN)
Branch
Dual branch
Siamese twins
Triplete network
Self-learn
Restricted Boltzmann Machines (RBMs)
Deep Belief Networks (DBNs)
Self Organizing Maps (SOMs)
Radial Basis Function Networks (RBFNs)
Deep Info Max (DIM)
Deep traige network (DTN)
Tensor deep staking network (TDSN)

Fig. 4 A categorization of the types of neural networks

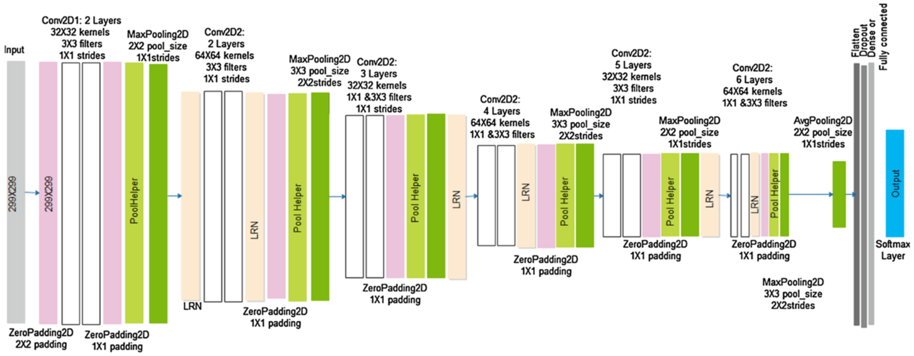
Input
299X299
299X299
Conv2D1: 2 Layers 32X32 kernels 3X3 filters 1X1 strides
MaxPooling2D 2X2 pool_size 1X1strides
PoolHelper
ZeroPadding2D 2X2 padding
ZeroPadding2D 1X1 padding
Conv2D2: 2 Layers 64X64 kernels 3X3 filters 1X1 strides
LRN
LRN
Pool Helper
ZeroPadding2D 1X1 padding
MaxPooling2D 3X3 pool_size 2X2strides
Conv2D2: 3 Layers 32X32 kernels 1X1 &3X3 filters 1X1 strides
Pool Helper
LRN
ZeroPadding2D 1X1 padding
Conv2D2: 4 Layers 64X64 kernels 1X1 &3X3 filters
LRN
Pool Helper
MaxPooling2D 3X3 pool_size 2X2strides
ZeroPadding2D 1X1 padding
Conv2D2: 5 Layers 32X32 kernels 3X3 filters 1X1 strides
MaxPooling2D 2X2 pool_size 1X1strides
Pool Helper
LRN
ZeroPadding2D 1X1 padding
Conv2D2: 6 Layers 64X64 kernels 1X1 &3X3 filters
LRN
Pool Helper
ZeroPadding2D 1X1 padding
MaxPooling2D 3X3 pool_size 2X2strides
AvgPooling2D 2X2 pool_size 1X1strides
Flatten
Dropout
Dense or Fully connected
Output
Softmax Layer

Fig. 5 A typical architectural layout of CNN applied for medical image visual feature learning and classification [52]

[58], MobileNets (MobileNet-V1 MobileNet-V2) [59], ShuffleNet[60], ResNet (ResNet-18, ResNet-50, RedNet-101), ZNet, EfficientNet (EfficientNet-b0) [61], SENet, HighwayNet, FractalNet, VGG (VGG-11, VGG-16, and VGG-19) [62], FunNet [63], SqueezeNet [64], and LeNet (LeNet-5) [65]. These models have served as baseline architectures for applying to solve different problems for visual feature learning. The transfer learning technique has been widely used using the pre-trained models of these architectures. Outcome obtained from such studies have shown that the use of the models for visual feature learning have been impressive.

The AlexNet is one of the major pieces of evidence of research breakthrough in the use of CNN architectures for visual feature learning by promoting the use of graphics processing unit (GPU) and dropout for improving visual feature learning at a significant speed. Classical benchmark neural network architecture is the VGGNet in Fig. 6 which leverages on reduced receptive space for improved visual feature learning and to achieve good classification and localization tasks. ResNet and GoogleNet represents an advanced performing neural architecture which have successfully extracted visual features through kernel size/

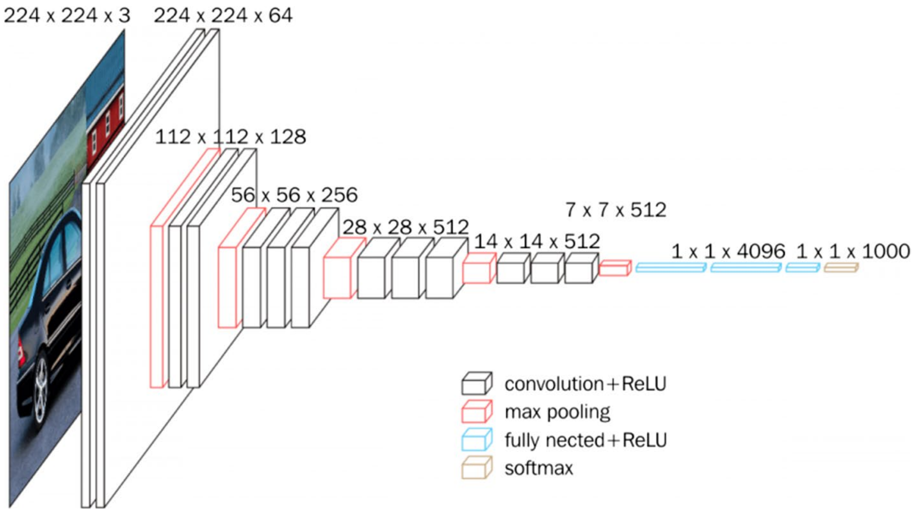

Fig. 6 Architectural layout of the VGG16 deep learning technique [66]

count adjustment, increased depth, learning residual functions, layer connections, and repetition of building transformation blocks. With the GoogleNet layer came the inspiration for the Inception networks which aimed to address the architectural depth issue with the former. Further to this, Inception neural networks motivated for the design of Xception through depth-wise separable convolution (DSC) with the aim of increasing the accuracy of the visual feature learning process.

Another outstanding benchmark classical CNN is the DenseNet which aimed at addressing the problem of large size of parameters in neural networks and the issue of vanishing-gradient to increase accuracy of classification tasks. Recall that the value of classification accuracy is directly proportional to the quality of visual features extracted for the classifier. This it does through the connection of each layer in the architecture with all other layers through a feed-forward approach. The evolving nature of neural networks and their applicability to solving several real-life problems on resource intensive-systems positioned them for deployment on low-resource systems such as mobile devices. This resulted in MobileNet architecture which relies on separable convolution to reduce computational power while sustaining better visual feature learning tasks. SqueezeNet architecrtuee has some shared features with MobileNet based on the need for network compression, though here through reduction of filter size and input channel. Another architectural solution towards reducing computational cost demand of neural networks is the ShuffleNet which uses point-wise group convolution, and a channel shuffle mechanism. Like the SqeezeNet is the SENet which based its network on the concept of squeezing operation. On the other hand, deconvolutional-architectural modeling has been considered to understand how popular convolutional architectures learn visual features. As a result, the ZFNet architecture is a provision in this direction for observing visualization of convolution network. Other related convolution neural architectures with similar performances and network layouts for solving visual feature learning challenge with improved accuracy are the SENet, HighwayNet,andFractalNet.A challenging trade-off often repeated with these neural network architectures is the speed and accuracy. Considering that the need for accuracy largely depends on the quality of visual feature learning process represented in an architecture, it is important that neural networks are focused on increasing accuracy first while trading off speed. It is believed that as high-powered computing resources continue to evolve, the challenge of speed will equally be addressed. In the next subsections, we examine trends in the design of neural architectures with a focus on improving feature learning process.

3.2 Two-stage neural networks (CNNs)

Advancement in neural architecture design has motivated for the novel idea of one-stage and two stage approach for solving the challenge of visual feature learning when solving real-life problems. The two-stage approach to design of neural networks often applies the first stage to learn certain features at one level and used the second stage for learning some other features at another level. For instance, when detecting visual features in chromosome images, the first and second stages of the neural network were designed to learn visual features in dicentric chromosome images from all the images, and to specifically identify the dicentric chromosome images respectively [67]. Another application of two-stage neural network is reported in [68] for applying the first-stage to visual feature learning of lung nodules using U-Net while the second-stage applied CNN with dual pooling structure for another level of visual feature learning. Furthermore, the application of two-stage CNN for improving the scale, type and quality of features through the use of a U-Net with squeeze-excitation block (SE block) for segmentation task, and use of an ensemble neural networks for further visual feature extraction [69]. Also, the first stage of the two-stage neural architectures have been applied for noise removal in images before applying the second stage to feature learning [70]. All these applications of the two-stage algorithms to visual feature learning process demonstrate the important and relevant of continuous evolvement of the CNN architectures in the domain of image processing. Although the two-stage neural network examples mentioned in this paragraph were suitable for solving some peculiar problems, benchmark two-stage neural architectures with wider applicability have been proposed and in use among researchers.

These benchmark two-stage neural architectures adopt the technique of processing image inputs from the perspective of regions of interests situated within the image samples. These region-based two-stage networks first apply one model in the first stage to extract regions of objects while leveraging on the second stage for further feature learning leading to classification and localization tasks. Examples of these region-based models are the R-CNN [71], Fast R-CNN [72], Faster R-CNN [73], Mask R-CNN [74], Feature Pyramid Networks (FPN) [75], Region Proposal Networks (RPN) [76], CarfRCNN, and Cascade R-CNN [77]. The R-CNNis the first attempt to represent such two-stage neural networks, though with some noticeable limitations, the architecture is able to generate regions from an image sample, learn visual feature representation, and the aggregation of the learned features using fully connected layers for further classification task. To address the limitation of R-CNN which is centered on its restriction to separately carry out classification and localization, the Fast R-CNN was proposed to combine these two tasks into one operation and also speed up the processing time. The Fast R-CNN architectures, in addition to the spatial pyramid pooling neural network (SPP-Net), have both been investigated for improving train and test computational cost of R-CNN and as well address the fixed convolution layers problems. Though with this improved performance, the increasing need for better visual feature learning and further elimination of computational bottlenecks, necessitated for Faster R-CNN architecture. Region Proposal Networks (RPN) architecture was built into the Faster R-CNN architecture to solve those limitations of Fast R-CNN. The

RPN model ensures that object proposal task is effectively computed through sharing convolution layers with network dedicated to feature learning and extraction. Like the RPN method is Feature Pyramid Networks (FPN) which can obtain multi-scale features using scale invariance though with large computation time. Meanwhile, the Mask R-CNN, an improvement on Faster R-CNN, is advancement to the R-CNN and other related variants. It achieves visual feature learning by generating pixel-level boundaries which are capable of isolating objects within the image input. This approach for learning visual features through object separation describes instance segmentation. A similar region-based solution to instance segmentation is provided through CarfRCNN [78] which combines CAResNet and CRF module. The CAResNet ensures that the visual feature learning and extraction process is completed with better accuracy, while the CRF module smoothens the pixels for better segmentation tasks. Another improvement to the Faster R-CNN is the Cascade R-CNN which replaces the RPN with a sequence of heads which computes the position for the bounding boxes that regionalize the image input. This is assumed that a refined computation for region boxing will improve visual feature learning within that concentration.

Traditionally, the two-stage neural network architectures enhance the visual feature learning process of the classical CNN by first extracting features at region level, before computing for whole-image level. This advancement in the evolvement and design of such architectures ensures that the accuracy of the feature extraction is improved and fine-grained. As result, the suitability of the architectures in the region generation phase will determine how effective the result of the second phase of the algorithm performs. This region generation phase has resulted in the design of different algorithms such as the RPN and selective search algorithm. The second stage of the two-stage architectures will mostly be focused on classification, localization bounding-box regression, or segmentation. This also requires that a form of visual feature learning process be applied to obtain better accuracy. This therefore demonstrates that the two-stage deep learning technique can yield better accuracy though at a slower processing time. To address this limitation, the one-stage deep learning technique has been proposed.

3.3 One-stage convolutional neural networks (CNNs)

The one-stage deep learning technique for visual feature learning is a group of neural networks representing the state-of-the-art when considering object detection task. They are a solution for addressing the challenge of processing time associated with the two-stage deep learning techniques. In fact studies have investigated and confirmed that though the accuracy of the region-based two-stage method, such as Fast-R-CNN, have 0.70, those of the one-stage algorithm are reputed to have about 0.64 accuracy but with improved inference time of 300 better than the two-stage methods [79]. Interestingly, though the region-based two-stage approach will need to first generate regions before pipelining the regions for classification and bounding-box regression, one-stage method simply attention their architecture to spatial region identification to aid object detection process [80]. As a result, the architectural complexities of single-stage deep learning techniques for visual feature learning are at reduced size compared with those of the two-stage algorithm. The one-stage method accepts an image input and proceeds to learn the visual features to support the task of computing the class probabilities, and then the bounding box coordinates.

The popular YOLO (You Only Look Once) [81, 82], RetinaNet [83], CenterNet [84], LADNet [85],Singe Shot MultiBox Detector (SSD) [86], and Deconvolution Single Shot Detector (DSSD) [87],have dominated the one-stage CNN deep learning techniques for applicability to visual feature learning. These architectures have also been evolved in a manner as to improve on the general limitation of one-stage algorithms which is reduced accuracy. In fact, recent versions of YOLO such as the YOLOv4 [88], have significantly improved in their performance accuracy while also sustaining the desirable inference or processing time. Another interesting neural network is the Capsule networks (CapsNets) which aim to address the underlying problems of CNNs. Most importantly, CapsNets are adapted to ensure that the pose representations in image or video inputs are preserved and accurately detected. The CapsNets can maintain the spatial hierarchical structures in input samples, such that even existing deformation in those samples can be identified. Several studies have leveraged on the success of CapsNets to propose further enhancements that allow for addressing limitations such as complex capsule routing algorithms [89].

The challenge of visual feature learning, especially when addressing the task of object detection, classification and localization, is multi-faceted. Traditional use of normal deep learning techniques struggle with the problem of achieving better accuracy when datasets are small demanding high computational power when datasets sizes are increased sufficiently for improved accuracy. They also suffer from the challenge of stabilizing their performance in a multi-scale input approach since such deep learning techniques often require fixed size input. All these challenges, including the sensitive case of underperforming when input sizes are lower, underpins the necessity for better visual feature learning deep learning techniques. As result, most of these problems are sufficiently addressed using the one-stage deep learning techniques with emphasis on the YOLO methods. This method of deep learning technique for learning visual features in image or video inputs relies on the same single-CNN pipeline for extracting regions, though using a rather lightweight neural network for the feature learning and extraction, although are still being evolved for detecting smaller objects. The YOLOv1, YOLOv2, YOLOv3 and YOLOv4 all demonstrate a progressive evolvement and improvement on this popular single-stage deep learning technique for visual feature learning. For instance, the YOLOv2 has better performance on visual feature learning resulting in fine-grained features being extracted. Similarly, the YOLOv4 has added features such as the Spatial Pyramid Pooling (SPP) component, DropBlock regularization to improve the task of visual feature learning. In addition to this addition, the recent version of YOLO has added smoothening of labels to support the process of visual feature learning.

3.4 Autoencoders

Autoencoders are another category of deep learning technique being applied for visual feature learning in a manner demonstrating data compression and decompression algorithm. The motivation for the use of compression pattern for the data is to allow for reduction of dimensionality of the data for an effective visual feature learning process. This approach is completely different from those of classical CNN, one-stage and two-stage CNN techniques. Their representational pattern follows the traditional feed forward neural networks which requires that input be the same as the output. This implies that the deep learning techniques benefit from the backpropagation algorithm approach to feature learning by first downscaling the dimension of input data so that only relevant features are retained, before mapping back to the input space [90]. This visual feature process is achieved under unsupervised manner since the inputs are not associated with labels. The autoencoder technique, by following the unsupervised learning approach, set some target values to be equal to the input data to generate different visual feature representation. A basic representation of autoencoder architecture, which have motivated for the design of improved models such as the convolutional variational autoencoder-based as applied for visual feature learning, with the study conforming that the quality of visual features extracted by the method better than the vanilla architecture [91]. One other suitability of autoencoder architectures is their capability to apply their visual feature learning process to reconstruct data through denoising, sparsity, and under-completeness [92]. Interestingly, different usage of the autoencoder architecture have resulted in more complex representation of the architecture when considered as basic building blocks which can be stacked to form hierarchical deeper network [7].

The stacking of the autoencoders benefits from using repeated encoder and decoder which are task with down-sampling and up-sampling the input data. The encoder receives the *N*-dimensional input data of features vector form and transforms into another M-dimensional vector feature vector. This output is passed to the decoder transforms it back to *N*-dimensional feature vector, and by so doing, the model can maintain a measure of similarity between input and output. A completely trained autoencoder provides a functional detached encoder for application to visual feature learning. In addition to the stacked autoencoder are other variants such as the denoising auto-encoder (DAE) [93], (VAE) [94], a sparse auto-encoder (SAE) [95], a sparse auto-encoder (SAE) [95], Adversarial Auto-Encoder (AAE) [96], sequence-to-sequence auto-encoder (SSA) [97], and many others. All these variants of the autoencoder address the challenge of visual feature learning and extraction by building on the foundational blocks for improved and fine-grained feature set. For instance, the DAE is suitable for removing noise from distorted samples so that the feature learning process on such image sample will yield high quality feature set. Similarly, the SAE type of autoencoder handles noise in a more efficient manner compared with DAE, and as a result proved relevant for use solving classification problem considering their possibility to yield better visual features. VAE represents an improvement to the architectural layout of autoencoders with provision to visual feature learning through understanding of the parameters of a probability distribution of the input sample. The CAE and AAE variants of autoencoders were targeted at improving the reconstruction mechanism through an enhancement to the reconstruction cost function. These variants of autoencoders have been largely applied to visual feature learning, except for the SSA which is suitable for textual feature learning seeing that it is built on the popular sequential model, Recurrent Neural Network (RNN). The derivation of these entire architectural enhancements to the basic autoencoder design has situated the use of the model in solving different real-life problems which are based on visual feature learning. For instance, issues relating with image denoising, recommendation system, reduction of dimension of input data, compression algorithm, and decompression algorithm, image processing, image restoration, object detection and recognition, image retrieval, pattern recognition, abnormality recognition, and data generation. However, other deep learning techniques have been proposed for solving data generation problems. We consider their architectural formation in the next subsection.

3.5 Generative adversarial networks (GANs)

Generative adversarial networks (GANs) are a special type of deep learning technique which supports visual feature learning. The technique follows an adversarial approach by pitching two neural networks against each for a min–max learning process. The participating networks play the roles of generator and discriminator accordingly [98]. The generator adapts to learning visual features in original image samples and then attempts to generate

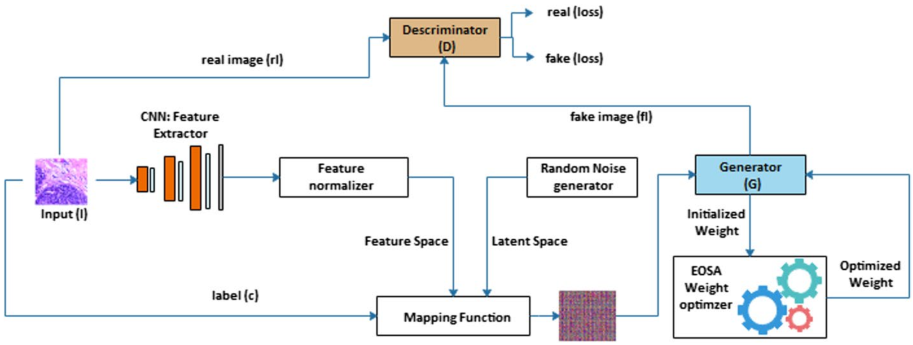

Fig. 7 An illustration of the architecture of GAN based on architectural enhancement which improves the latent space of generator for better visual feature learning [100]

similar representation. The generator depends on the discriminator to validate and monitor the visual feature learning and representational process. Interestingly, visual feature learning tasks has been enriched with the combination of the autoencoder and GAN architecture for a hybrid, variation autoencoder generative adversarial network (VAE-GAN), can eliminate the challenge of stabilizing GAN's learning process.

Like other deep learning techniques, GAN-based method to visual feature learning have evolved based improvement on objective function, addition of conditional mechanism, and based on architectural redesign. The conditional generative adversarial network (cGAN), information maximizing generative adversarial network (InfoGAN) and auxiliary classifier generative adversarial network (AC-GAN) form the condition-based GANs. The GANs which improved on the latent space of the generator includes the Ebola optimization search algorithm generative adversarial network (EOSA-GAN) [99] in Fig. 7, bidirectional generative adversarial network (BiGAN), and semi-supervised generative adversarial network (SGAN), while those based on strictly based on architecture enhancement are big generative adversarial network (BigGAN), cycle generative adversarial network (CycleGAN), self-attention generative adversarial network (SAGAN), boundary equilibrium generative adversarial network (BEGAN), deep convolutional generative adversarial network (DCGAN), Laplacian pyramid generative adversarial network (LAPGAN), and the VAE-GAN. Considering that GAN training largely depends on optimization of a min–max function, recent advances in the field have investigated evolving new architectures based on the objective function of the visual feature learning process. For instance, the least square generative adversarial network (LSGAN), energy-based generative adversarial network (EBGAN), categorical generative adversarial network (catGAN), geometric generative adversarial network (GGAN), and Wasserstein generative adversarial network (WGAN) belong to this category. Because of the significance of the nature of the visual feature representation in images, some feature-specific GANs such as the MammoGAN [45] and ArchGAN [98] have been researched to generate samples based on the abnormalities represented by the features in medical images. All these types of GANs have a wide range of application in several fields of study for solving visual feature learning problems associated with image synthetization or generation, image resolution enhancement, picture to picture translation (pix2pix), image inpainting, and image coloration.

3.6 Transformers and attention deep learning techniques

The application of transformers and attention mechanism signals the beginning of a new approach to deep learning based visual feature learning. Although the concept of transformers started with textual feature learning or natural language processing, it was later advanced to address the similar challenge in visual feature learning. Attention mechanisms have served to ensure that similar operation obtained when using convolutional functions are made available to transformers using this mechanism by mapping and encoding dependencies. Traditionally, transformers were designed to support the feature learning processing existing in long-range and short-range dependencies in both textual and visual mediums. For the visual medium such as an image, this has allowed for understanding images as sequence of patches which makes it possible for transformers to process them. The result of this impressive visual feature learning has been further applied or extended as in graph-based transformer (GT) [100], pre-trained vision transformer (ViT) on images [101, 102], off-the-shelf ViT-based transformer [103]. Although attention model alone has proven relevant to visual feature learning [104], their combination with the architecture of transformers has transformer-based neural network to outperform similar deep learning techniques in learning of global and long-range semantic information. The attention mechanism sits between encoder and decoder of a transformer for effective performance like that of convolutional operation for effective visual feature learning [105]. Several architectural adjustments have been made to transformers to obtain better variants such as the Swin Transformer demonstrating hierarchical transformer representation [106], detection transformer (DETR) [107], and the popular ViT [108].

Table 2 presents a comparative outline of all the deep learning techniques relevant for application to addressing the task of visual feature learning in digital images. This comparison allows for seeing relevant architectural types in their respective categories, and understanding what the advantages and disadvantages of neural networks are described in each category.

RQ2.What are the main applications of visual feature learning in computer vision?

4 Applications of visual feature learning

The application of deep learning techniques to addressing real-life problems is widening with becoming ubiquitous. Real-life problems such as image classification, object detection and localization, image segmentation, video understanding and summarization, image synthetization and reconstruction, and image translation have motivated some state-of-the-art systems and solutions which are now supporting several processes with divers' application areas. These areas of applications include computer vision, scene understanding, vehicular navigation and self-driving cars, medical image analysis, surveillance and ambient security, network abnormality detection, cybersecurity, video surveillance, robotics and several other domains. Interestingly, the nature and complexity of the applicability of a solution have inspired new architectural composition of neural networks. The evolvement of neural networks over the years, especially as it relates with visual feature learning, have demonstrated outstanding performance which is now made the field of artificial intelligence the media booze. In this section, we review studies which have applied deep learning techniques in some outstanding manner to solve some real-life problems. To organize the discussion, the study on the literature with respect to visual feature learning for this section is categorized into object detection and recognition, facial recognition, scene understanding, medical image analysis, autonomous vehicle, and some other areas of applications. The aim is to present state-of-the-art in the application of these techniques, and to further inspire new models which can address some other complex problems.

4.1 Object detection and recognition

The challenge of object detection and recognition is demanding and represents an index problem for solving most feature learning problems in computer vision. The problem of object detection (OD) presents a different formalism when compared with object recognition (OR) [109]. In most cases, achieving the task of object recognition involves first addressing the problem of object detection since there is first needed to identify an object before drawing up a semantic similarity to complete OR task. The complexity of OD is therefore less compared with that of OR since the former is contained in the latter. OD therefore attempts to extract and learn features suggestive distinguishing the presence of an object with the aim of assigning label to the object. This is first preceded with training a deep learning model with a set of labels to allow for learning features associated with each label. The restriction of the training process of OD to a set of predefined labels situates it as a form of closed vocabulary detection. A very similar computer vision problem which has shared similarity with the OD is the zero-shot object detection (ZSOD) which does not require predefined and restricted labels to achieve feature learning process. The ZSOD relies on an approach which maps some learned visual object to a label derivable from computing the semantic distances between known and unknown labels positioning it to learn with very few sample sizes. Generally, the challenging task of OD often requires that instances of some objects be identified in a natural image scene [110]. On the other hand, the task of OR often leads to addressing the problem of localizing the recognized object in a natural scene. This is referred to as object localization (OL) considering the demand of identifying the position of the object detected and recognized, so that the localized object is distinctively bounded using a rectangular box with an assigned label on it. Another computer vision problem which advances on the techniques of OD, OR and OL is the task of image segmentation which partition, rather than simply identify, an object or part of an object from an entire image. The approach leverages on difference in a cluster of pixels based on their intensity and texture to segment an object or some lines, curves, component and shape from an object.

The application of deep learning techniques to solving object detection can be categorized into basic CNN models, a two-stage algorithm, one-stage algorithm, and the method of transformer [111]. These approaches evolved from the basic use of CNN models to the current use of transformers. The use of ShuffleNetV2 which is a kind of lightweight object detection algorithm has been reported in the literature [112]. The use of ShuffleNetV2 was in a reduced form to obtain L-Net by scaling down the architectural layout of ShuffleNetV2 so that 5×5 depth is replaced with a 3×3 depth convolution and reduced the number of input channels. In addition to the scale-down and modification to ShuffleNetV2, the pooling operation was replaced using pyramid pooling method and an attention mechanism added to the L-Net architecture. Another architecture which have been applied to the task of object detection is a point-wise separable (PWS) and depth-wise separable (DWS) neural network as combined with spatial feature pyramid network (SFPN) [113].The aim of this combination is to ensure that visual feature learning process engages the spatial and depth of the image for an effective performance in the object detection task. This approach

Table 2 Comparison of six (6) categories of deep learning techniques suitable for visual feature learning in digital images and videos

Visual feature learning category	Deep learning technique	Advantage	Disadvantage
Classical CNN	CiFarNet [53], AlexNet [54], GoogLeNet [55], Inception [56], Xception[57], DenseNet[58], MobileNets[59], ShuffleNet [60], ResNet ZNet, EfficientNet) [61], VGG [62], FunNet [63], SENet, HighwayNet, FractalNet, SqueezeNet [64], and LeNet [65]	First to promote the use of graphics processing unit (GPU) and dropout for improved visual feature learning Increased the depth of convolutional layers	Mostly characterized with very high architectural layouts resulting in high number of training parameter size, and computational cost
Two-stage CNN	R-CNN [71], Fast R-CNN [72], Faster R-CNN [73], Mask R-CNN [74], FPN [75], RPN [76], CarfRCNN, and Cascade R-CNN [77]	Capable of sharing visual featuring learning tasks, segmentation, classification and even localization tasks across the two-stages of neural convolutional operations Each layer is appropriately fine-tuned to effectively achieve a task Two-stage detectors usually reach better accuracy	Slow in the training and testing phase thereby discouraging their real-life applicability to problem solving
One-stage CNN	YOLO [81, 82], RetinaNet [83], CenterNet [84], LADNet [85], SSD [86], and DSSD [87]	Faster than the two-stage and classical deep learning techniques for feature learning	Suffers from competitive object detection accuracy
Autoencoders	DAE [93], VAE [94], SAE [95], SAE [95], AAE [96], and SSA [97]	Autoencoders have the advantage of reducing the dimension of data and as a result reduce computation time Represent unsupervised learning approach with less challenging when being trained Autoencoders are suitable for learning visual features that are abstract and at higher-level	Suffer from unperfected decoding algorithm Depends on large, clean unlabeled data for effective training
Generators	cGAN, InfoGAN, AC-GAN, EOSA-GAN, BiGAN, SGAN, BigGAN, CycleGAN, SAGAN, BEGAN, DCGAN, LAPGAN, and VAE-GAN. LSGAN, EBGAN, catGAN, GGAN, WGAN MammoGAN, and ArchGAN	Applicable for visual feature learning and the synthetization of the feature into new image samples	GANs are very difficult to train and often suffer from mode collapse which is an indication of failure to achieve good visual feature learning
Transformers	Swin Transformer [106], ViT[108], DETR [107], Transformer [105]	Capable of learning techniques in learning of global and long-range semantic information	It is very challenging to control the operation of the attention mechanism of transformers

has benefited from the use of transformer encoder–decoder and feature fusion models to speed up the process of visual feature learning. Similarly, a popular network, namely the YOLO has been combined with a Dense Prediction Simplified (DPRS) [114] for a better object detection task and for performance enhancement. The study investigated the benefit of adding backbone network to the aid DPRS architecture through fusion of its output with that of the backbone network. A scale-down version of YOLO version 4 has also been proposed to achieve the task of object detection yielding the SlimYOLOv4 architecture [115]. The slimming down follows replacement of CSPDarknet53 with MobileNetV2 and replaced convolution operations with depth-wise over-parameterized depth-wise convolutional layer (DO-DConv) and depth-wise separable convolution (DSC).

We have also noted a scale-down version of the neural network used in solving object detection such as the YOLO and SSD. YOLOv3-tiny network is presented as a scale-down version, called Micro-YOLO + , to ensure that the trained models are useful on embedded devices characterized with restricted resources and low power requirements [116]. The scale-down version added modules of depth-wise separable convolution (DSConv), the mobile inverted bottleneck convolution with squeeze and excitation block (MBConv) and the ghost module (GConv). The increasing evolvements of such lightweight models are now the focus of research in the application of neural networks to object detection. Another recent similar approach showed that lightweight framework named multi-level feature reused detector (MFRDet) [117] which can use visual features obtained through deep and shallow feature maps. The motivating factor for the evolvement of lightweight object detection neural network as we have observed in most studies is to promote improved accuracy, precision, fine-grained visual feature learning and allow for usability to mobile or embedded devices. Another object detector with recognition facility based on the SqueezeNet architecture has been observed to perform the task better when supported by metaheuristic algorithms such as the African Adam Optimization-based and African Vultures Optimization Algorithm (AVOA) [118]. The optimized SqueezeNet model was further supported with Generative Adversarial Network (GAN) to aid the process of detection of the object, while Resnet was adapted for visual feature learning.

The basic neural network approach to solving object detection was automated through the adopted neural architecture search (NAS) to evolve from the handcrafted approach to designing CNN architectures for solving such problems. The NAS method has been effectively deployed to searching for the feature pyramid network and used a reinforcement learning model to locate a prediction head [119]. A novel application of the YOLO network has also been demonstrated for searching for missing items in a video document [120] extracted from CCTV to solve the challenge of applicability of deep learning to surveillance problem. This also has applicability to video summarization based on some objects or human actions detected in the video as supported by using-temporal-information detection [8]. Contrary to popular use of YOLO which represents a single-staged algorithm, the region-based object detection method which is a two-stage method is vastly and competitively applied in literature. The Mask R-CNN method has reported to effectively identify and locate objects in images to support text-to-speech synthetization for vocalizing objects' names and their positions [121]. Another interesting neural network architecture in solving object detection is the transformer which has now had a wide acceptability in the field of natural language processing [9]. Whereas the supervised deep learning method is replete, the unsupervised deep learning models sued for contextual visual feature learning have also been researched [122]. The aim of such an approach is to allow understanding the semantics between objects in a scene, observe semantic similarity and co-occurrences through the unsupervised method.

Most enhancement process are often focused on improving the convolutional depth, adding context-aware backbone neural networks, inventing new loss functions and activation mapping strategy, classical object detection architectures, training with latest and large-scale datasets, and using informative evaluation metrics [123]. The evolving process is motivated by the architectural layout of some classical networks such as the VGG-Net, ResNet, ResNeXt, LeNet, SENet, DenseNet, Xception, AlexNet, ZFNet, GoogleNet, PNAS, and ENAS. As a result, new neural models aimed at object detection tasks are now able to achieve fine-grained image classification and object detection so that high intra-class and low inter-class variance challenges are addressed [124]. Whereas the classical neural network approaches to solving object detection will mostly be focused on two-staged object detectors, recent algorithms are now evolving into the one-stage method. The one-stage methods can use scale-down neural network to achieve excellent visual feature learning on spatial region of images in search of objects. Whereas the two-stage algorithms approach the object detection problem on image samples through region selection rather than the spatial region [125]. The challenging task of the single-stage algorithm is to compete using their accuracy whereas the two-staged method often yields better accuracy though they suffer from increased inference time to the advantage of the single-stage method. The YOLO and its associated versions remain typical examples of the effective single-staged object detection algorithms, while Fast-RCNN demonstrates the two-staged methods. In fact, a study [126] have investigated the performances of these single-staged, two-staged and basic neural network models aimed at object detection by experimenting with YOLO, Faster Region based Convolutional Neural Networks (Faster R-CNN), and Single Shot Detection (SSD). Findings from the study demonstrated that the domain of applicability of the object detectors often influence performance with the YOLO considered to mostly outperforms SSD and Faster R-CNN. The wide range of applicability of deep learning models aimed at object detection and recognition have further demonstrated the need for focused research efforts to benefit the challenge of image and video classification, face parsing and recognition, scene labeling, semantic segmentation, saliency detection [127, 128], counting of objects, and object monitoring [111].

Most importantly with object detection is the approach of using saliency in identifying segments within the image scene. In salient object detection, detection is more focused on very noticeable segments existing within the scene owing to distinguishing features. Such features might be based on very high-level features or even low-level features. For instance, the low-level features reveal a concentration segment within the image scene which requires some attention, leading to detection of objects situated in that region. On the other hand, the high-level features are used to also identify objects, but objects which might represent shape and major outlines demonstrating a complete object. Several studies have investigated and reported these two categories of salient object detections. For example, a study has contrasted between supervised and non-supervised salient object detection. In the study, authors showed that unsupervised salient object detection though presents a more useful methodology to object detection but suffers from effective performance in a complex scene [129]. As a result, a Belief Capsule Network (BCNet) architecture was proposed to address this limitation. Similarly, both the issues of the need for guided multi-level contextual extraction and integration from scenes and need for aligning relationship between Contrast and part-whole in image scenes, have been addressed using salient object detection approaches [130, 131]. These solutions to contrast and part-whole relation holds relevance in scene understanding to aid self-driving vehicles in especially unfavorable weather conditions. Moreover, even the demand for effectively processing complex scenes do have a wider scope of applications and use. This therefore emphasizes the need for research in salient object detection methods.

4.2 Facial recognition

Face parsing and recognition represent another interesting and challenging task when using deep learning models. Research has long shifted from the classical method of handcrafting the feature extraction method to the use of convolutional procedures for extracting the face features. Face parsing task the neural network on the need to accurately segment facial parts at the pixel level [132]. This is contrary to the task of face recognition which attempts to match image samples of video frame of faces to an established database. These task often have to deal with the difficult problem distortions, angle differentiation, and displacements to positively impact the outcome of the recognition and parsing procedure.However, the applicability of both traditional and GAN-based data augmentation have improved the performance of deep learning model to overcome the issue varying angles, pose, expression, age, occlusion, disguise, lighting, distortions, and displacement so that the training process benefits from a diverse input source to improve generalization. The resulting trained model has proven relevant to solving problems associated with mobile payment, environmental surveillance, real time attendance, safe city, communication, identity authentication, criminal investigation and other fields. Datasets such as the Labeled Face in the Wild (LFW), YouTube Faces (YTF), VGGFace1, VGGFace2, ORL, IITK, CVL, AR, CASIA-Face-V5, FERET, CAS-PEAL, WIDER FACE, FDDB, FFHQ, 100 K-Faces DFFD, and CASIA-WebFace, have surfaced as a benchmark of image samples for training the face recognition models.

The evolvement of the design of neural architectures used for face recognition have transformed from the basic deep learning models such as neural computing model to overcome the laborious process of manually extracting features and by so doing, improve processing time [133, 134], to more advance neural models. However, this advancement was creeped through component-wise improvement of the neural architectures. For instance, activation function and loss function techniques have been advocated as means for eliminating model errors in deep learning. This has led to the improving the use of loss function such as the softmax to hybrid method such as the angularly discriminative features which combine multiplicative angular and additive cosine margin. The aim is to ensure that increment of the large margin cosine, and as well increase accuracy. This performance improvement demonstrates the suitability of deep learning to addressing face recognition problem as against the use of other techniques such as deep infrared method which aim to significantly reduce the factors that adversely affect the accuracy of recognition accuracy by eliminating visible spectrum problem [135]. Transfer Learning method have also demonstrated some positive outcome when improving the performance of basic deep learning models such as CNN on face recognition problem [136]. Benchmark neural networks such as the GoogLeNet and FaceNet were pre-trained to fine-tune the last layer of CNN architecture for improved the performance of the face recognition task using deep learning. The method relied on Euclidean embedding and clustering techniques to match faces. One other challenging task in face recognition with neural network is handling frontal and profile faces in a manner. Tree Structured Part Model, supported by Scale Invariant Feature Transform for feature learning [137], has been recently used to address this challenge by leveraging on the facial landmark points through an extraction of regions-of-interests (ROIs). The visual features extracted were classified for recognition of faces using Support Vector Machine (SVM).

Although GAN-based methods have been applied for augmentation purposes, the adversarial deep learning architectures are very relevant in detecting non-existent faces which originate from GAN model. The study in [138] addressed identification of deep fake images using deep learning model named fisherface which relied on Local Binary Pattern Histogram (FF-LBPH) for dimensionality reduction for detecting of deep fake face image. More recent studies have demonstrated that using face-selective neurons approach to simulate neural network model for face recognition will be impressive. This approach requires that neural model attenuates to the visual experience in getting well-trained for face recognition task [139]. This visual experience was simulated into neural network through a hierarchical deep neural network model of the ventral visual stream and combined with a mechanism capable of face-selectivity. Similarly, a report on the use of pre-trained face recognition deep neural network (DNN) combined with diverse array of stimuli for selectivity of faces for identification [140]. The simulation of stimuli was based on the monkey and human single-neuron recordings which were compared with artificial units with real primate neurons. This task was aimed at linking artificial and primate neural systems for solving face recognition problems. Other studies have also considered rats, in addition to the popular simulation of visual faculties of humans and monkeys, to improve the performance of deep learning models [141]. Stimuli were collected from those previously adapted to deep learning models and combined with unseen stimuli to derive contrast features which will support facial recognition process.

Continuous improvement of deep learning models for face recognition is aimed for addressing the reoccurring problem of occlusion, blurring, and partially occluded faces in uncontrolled conditions. For instance, a single-stage face detector based on RetinaNet [142] has reported improved accuracy while at the same time addressing those reoccurring problems mentioned earlier. Similarly, deep learning and the popular image processing library, namely OpenCV have been reported to be capable of solving face recognition problem [143]. Also, the Face mesh method has been used for investigating the applicability of deep learning technique to solving face recognition problem [144]. Although several deep learning methods remain relevant to addressing face recognition problems, most still leverage on component-wise enhancement of convolutional-pooling layers for effective performance. A study [145] observed that barely using Max Pooling instead of the popular average pooling in convolutional operations resulted in better face recognition task. Most deep learning methods often rely on regular image spectra to collect images for training. But in very adverse conditions, images collected through such regular spectra losses quality thereby necessitating the use of infrared spectra for collection of mid-wave infrared (MWIR) images which can be applied for face recognition task [146]. This medium of image sample increases the degree of difficulty related with the recognition process. However, a study [146], has demonstrated that a dual branch deep learning model can be adopted to solve this problem. Combining the techniques of transfer learning and data augmentation through traditional and GAN-based methods, the study demonstrated that face recognition problem using MVIR is possible.

All these different deep learning techniques demonstrate the suitability of the approach for addressing face recognition problem which is an important biometric modality and verification system. The reviewed of content in literature demonstrates that most approaches have often relied on the two-stage algorithms with some benefiting from the use of face authentication, facial recognition, face clustering, transfer learning and GAN-based approaches to improving performance. The quality of the visual feature learning process often affects the face detection face alignment, and face identification which are the phases of face recognition problem. One very interesting solution noted with solving face recognition problem is the nature inspired visual patterns recognition observed in human and animals such as rat and monkey. The extraction and simulation of stimuli from these organisms have fast tracked the process of applying deep learning models to solving the problem.

4.3 Scene understanding

Scene understanding is another challenge of computer vision considering the demand to segment and recognize objects within a scene represented as image or video. This is more challenging than the regular object detection problem due to the complexity of techniques needed to isolate objects, curves, lines and other shapes to achieve scene understanding. Interestingly, the evolvement of deep learning has allowed for solving scene understanding problems in manner that now attracts applicability of the method in complex scenery. This evolution has allowed machine perception to represent complex abstractions in data, and then encode into a representation for learning process. This therefore has positioned deep learning architectures has potential solution to the challenge of detection of objects and shapes to enable scene perception. In addition to the popular use of CN N for scene understanding task, we noted that complex neural network architectures such as the autoencoders, generative adversarial networks, transformers, and few-shot learning have demonstrated impressive performance in scene understand. The result of this advancement in the use of deep learning has supported research in autonomous vehicles and remote sensing. Scene represented in remote sensing requires classifying the scenes according to some classes or categories. The complexity of scenes as represented in images and videos presents a difficult task for computer vision in classification of scenes in remote sensing images. When scene understanding is also looked at from the perspective of autonomous driving, we see another level of complexity considering the nature of objects, persons, instances of shapes, curves and lines which need to be isolated and classified to achieve a complete understanding of the scene.

This therefore demonstrates the relevance of an effective visual feature learning process to support an accurate scene understanding task. Both autonomous vehicles and remote sensing are examples of applicability of scene understanding which have now received extensive research efforts. These research efforts have often investigated developments on methods that promote the enablement of algorithm driven and data-driven cars or remote sensing. The underlying motivation for applying deep learning techniques to these domains is to promote autonomous systems with possibility of scalable, feasible, accurate and reliable processes with little or no human intervention [4]. One prominent deep learning technique which has featured well in scene understanding is semantic segmentation which is required to partition a scene semantically to achieve overall scene understanding. This issue of segmentation has a wider applicability involving object detection, full scene semantic segmentation, instance segmentation, and lane line segmentation when applied to autonomous driving [147]. Second to this technique is the use of method for understanding actions as demonstrated in videos or images through action recognition [148]. As a result, deep learning methods needs to continuously evolve to enhance their effectiveness to solving scene understanding to obtain better accuracy in classifying remote sensing images, autonomous driving, and other domains which rely on the solution [10].

In furthering research on scene understanding, understand complicated social scenes, an end-to-end deep learning model [149] has been proposed to solve a smaller problem in scene understanding known as addressee recognition. This allows for deriving visual features to enable deciphering social scenes to locate places such as cafeterias, hospitals, barracks, hotel, schools and social outdoors meeting points. The study was aided using a newly curated dataset for scene understanding namely the addressee recognition in visual scenes with utterances (ARVSU) which contains variations of images in visual scenes with utterance and a corresponding addressee. The end-to-end deep learning model was supported by a multi-modal deep learning technique which trains us to understand different human cues such as eye gazes and transcripts of an utterance. The trained model will support the end-to-end deep learning model to improve the prediction on the conversational addressee based on a speaker's transcripts of utterance for addressee recognition. Another scene understanding problem which has benefited from the use of deep learning techniques is enhancing reasoning in robots. For instance, when studying the inclination of robots when encountering partial occlusion of objects and the stability of object configurations, both incremental inductive deep learning technique and non-monotonic logical reasoning with [150]. This has been combined with incomplete commonsense domain knowledge to guide the construction of deep network models. This approach that combines deep learning with logical reasoning is a deviation from the norm that promotes only CNN technique as a means for scene understanding through classification of images [151].

The use of multimodal neural networks for solving scene understanding problems is rarely being investigated. This is challenging because the conglomerating neural networks need to jointly learn visual features of different tasks at the same time [152]. The neural model proposed for the multi-tasking multimodal architecture allows for increased inference speeds through the delay of feature computation in the shared layers of the neural model. In addition, the multi-tasking neural architecture improves performance of scene understanding since associated tasks share complementary information, and act as a regularizer for one another. Another similar architectural reengineering method for improving the task of visual feature learning for scene understanding is based on automated branched multi-task networks [152]. The architectural improvement supports enabling convolutional operations in the deeper layers to gradually grow more task specific. The approach relies on predictions of other tasks at multiple scales to improve the prediction of a single task within the multi-tasking neural network. However, multi-tasking neural architecture such as this often suffers from the challenge of balancing the multi-task learning optimization problem.

The context of visual scene understanding presents a very challenging task for visual feature learning considering several instances of objects which may exist in and scene. This will require that for the scene understanding task to be adequately addressed, object detection in the scene, localization of the detected objects, and estimation of the distance existing among detected/localized objects remains an attractive area of research in the domain. Another interesting subject of research in this issue of scene understanding is observing the trajectory of objects detected and localized. We consider that as neural network architectures evolve, most of these problems will be equally addressed by advancing such models from the traditional silo-like solution which usually trains a separate neural network for each individual task. The multimodal and multi-tasking neural networks are also some interesting neural networks suitable for scene understanding and which bring with them several advantages. However, they often suffer from the challenge of striking a balance between optimizing the training among all contributing tasks. We consider that increasing the variety of image samples in publicly accessible datasets, such as the PLACES2 and ARVSU, for scene understanding will also contribute to the advancement of the evolving deep learning techniques in the field.

4.4 Medical image analysis

The task of feature learning in digital medical images represents another dimension of complexity and high dimensional operation considering the nature of abnormalities in such samples. Several research efforts have been focused on using deep learning models, especially the convolutional neural network, for feature learning. The problem of disease classification and localization of abnormality in the image samples have for long been the interest of the research community. However, recently, issues like image augmentation [45, 52, 98, 99] and multimodal classification have dominated most research outcomes. Another challenging problem which has not received research focus is understanding the history and trajectory of the progress of a disease in an individual through medical images using deep learning models. We anticipate that this will now be given significant attention considering the challenge of early detection of some disease with high treatment option when isolated at their early stage. Although the issue of learning features of abnormality presenting very subtle [11] and irregular characteristics remain challenging especially for cancer cases, we understand that continuous evolvement of deep learning model present a good solution to this problem. The evolution have followed from the basic artificial neural networks and have now progressed to convolutional structures, recurrent networks, attention models, and now transformer [153]. Image segmentation task is another reoccurring task when handling medical image analysis using deep learning model. This is required to help scale-down the search space represented in image pixel to enable region-based search or create regions of interests (ROIs) to assist in addressing visual feature learning and classification. The first dimension to this task is to ensure that very rich and fine-grain visual features are extracted so that discriminative features are not filtered out but are extracted for use by the classifier. The second dimension to the contribution of segmentation using deep learning is that region-based features are a cluster of similar features which present the classifier with concentrated features with little difference distance values. Such reduced distance values enable classifier easily and accurately determines what class to label the ROIs or even the whole image. The use of classifiers has been advanced from the traditional machine learning approaches such as ANN, SVM, Decision trees, and KNN, to the more advanced methods such as the softmax, sigmoid and many more. In this section, we examine advances in the use of deep learning models to solve medical image analysis with emphasis on the visual feature learning process.

The application of machine learning to addressing image classification problems in medical image analysis has received significant attention with the outstanding performance of the deep learning models. This is motivated by the difficulty encountered by human experts in isolating abnormalities in medical images. To buttress this further, recently, a study [11] has confirmed that the complex and subtle characteristics of some abnormalities as represented in digital mammography is making it difficult for radiologists to detect early signs of breast cancer. This therefore is a pointer to grave human omission when disease a not detected in their early stages due to the complexity and presentation of the abnormalities in medical images. To address this problem, convolution neural network (CNN) method has been widely applied to detect and classify abnormalities in medical images. For example, the work in [52] approaches the challenge by designing a new CNN architecture suitable for extracting features from multi-view forms of MLO and CC in digital mammography. The study also leverages the technique of image augmentation to improve classification accuracy. Although this approach has now been largely reported in literature, a domain-based challenge of effectively combining hyperparatemers of the CNN soon presented as an obstacle to performance enhancement. This challenge has been addressed in [16] using a random-grid model for selecting an optimal combination of hyperparameters for CNN. Using a combination of heterogeneous sources of digital mammography, the study showed that the approach demonstrates a significant performance improvement. In another related work, the use of wavelet-based function to replace the classical rectifier linear unit (RELU) based function as activation function of CNN architectures has been motivated in [46]. This is necessary to ensure that discriminant features are not missed out during the convolutional-subsampling operations with the aim that classification accuracy will be improved by the method. In addition to using CNN models to address breast cancer disease detection, other fields of medicine have also received attention from researchers in using the model to improve detection rates. In [154], authors proposed a novel CNN architecture namely CovFrameNet for the detection features suggestive of the presence of corona virus in chest X-ray. The aim of the study was to support the early detection of COVID19 and as a result, reduce the contagion and spread rate of the disease. Still on the issue of the COVID19 pandemic, a novel machine learning framework has been proposed for deployment in smart cities in aiding detection, prevention, and contact tracing in the spread of COVID-19 and provide a better understanding of the disease [155]. In another work, deep learning model has been applied to surveillance task by training the models designed as CNN form to detect abnormal intrusion in a home setting. The CNN model was trained to characterize homeowners from strangers so that anomaly intrusion triggers an alarm [156].

The application of deep learning models to generative tasks in image synthetization problem has also been considered a new direction to the use of machine learning medicine. This has necessitated the design of generative adversarial networks (GAN) as replacement models to human in the loop in image annotation and segmentation tasks. A study has proposed the use of GANs to generate regions of interest (ROIs) characterized with architectural distortion, asymmetry, mass, and microcalcification abnormalities in digital mammography. The aim of the study is to allow for synthesized images to augment the challenge of data insufficiency and privacy to provide CNN models with enough data for training. The resulting GAN model was named ROImammoGAN [45]. In a related work [98], a single-abnormality focused GAN model namely ArchGAN has been proposed strictly for synthetization of medical images with architectural distortion abnormality. This is motivated by the repeated unavailability of this form of abnormality in digital mammography datasets that are publicly available. In another study, it has been demonstrated that optimizing the weights and biases of neural architectures composing GAN model can improve the accuracy and reduce the loss function values for the image synthetization process [99]. The study also proposed the use of a mapping function to enrich the input latent space of the generator, with more features from the domain of the problems being addressed.

While the issue of image classification and synthetization appeared to have been sufficiently addressed using deep learning, localization of abnormalities in medical images remained unaddressed especially with breast cancer. As a result, a study has focused on the approach of dual-branch CNN framework for solving the problem. The aim of the approach was to apply one branch of the framework to detection and classification and another branch to localization. A combination of whole-image based CNN (WCNN) and region-based CNN (RCNN) architectures were systematically arranged to achieve the proposed framework [157]. The outstanding performance of the nature-inspired optimization method has motivated the use of the algorithms in optimizing the performance of deep learning methods. This optimization could take the approach of hyperparameter selection in an optimal manner, or the optimal combination of the building block of the deep learning models. In [17], the authors have investigated the performance of five optimization algorithms CNN including genetic algorithm (GA), whale optimization algorithm (WOA), multiverse optimizer (MVO), satin bower optimization (SBO), and life choice-based optimization (LCBO). The aim of the study was to optimally select and combine hyperparameters to observe which of the algorithms outperforms others in ensuring the classification accuracy of deep learning models are increased and improved. Similarly, authors in [158] have applied EOSA, GA, SBO, LCBO, WOA, and MVO to also investigate what category of optimizers are suitable for solving combinatorial problems in CNN hyperparameters. The study was aimed at learning visual features in image samples characterized with abnormities of lung cancer. In a related work, the building blocks of CNN architectures have been evolved in a method that is popularly known as neural architecture search. This evolvement was aided using EOSA optimization algorithm. The study [159], also implemented a block-based stochastic categorical-to-binary (BSCB) algorithm to support the EOSA method in building the candidate CNN architectures. On the other hand, a new variant of the EOSA metaheuristic algorithm was proposed based on the human immunity response to the Ebola disease. The new algorithm which is named the immunity-based Ebola optimization search algorithm (IEOSA) was applied to solve the problem of reducing features in CNN. The aim of the study is to ensure that extracted features from the layers of CNN are filtered using IEOSA so that non-discriminant features are dropped to eliminate bottle-necking the classifier [160].

Several other medical images, in addition to mammography and histopathology, have become very useful for image analysis leading to disease detection through visual feature learning. For instance, computed tomography (CT) image samples are very supportive in characterization of disease using deep learning models. The use of convolutional autoencoder deep learning method has been applied to visual feature learning on medical images with focus on lung cancer images [161]. The study demonstrated that the unsupervised approach is suitable for improving the feature learning task for lung nodule through a small number of unlabeled data. Another use of the CT images was reported in [162] for sample-efficient deep learning methods to help detect abnormalities representing and leading toCOVID-19. The neural network benefited from the transfer learning approach to improve classification accuracy and eliminate overfitting the model. An attribution framework using adaptive path-based gradient integration techniques have been proposed to understand how deep learning achieve its inference leading to classification solution with CT, mammography, histopathology, magnetic resonance imaging (MRI), and many other medium [163].

The use of deep learning methods on medical image analysis tasks supports achieving image data interpretation, finding semantic and logical relation among features, and achieving visualization [164]. The aim is to evolve and adopt deep learning techniques to improve computer aided systems developed for visual feature learning and detection of disease through medical images. Success achieved in this direction has now further motivated for the multimodal approach which combined both textual and visual feature learning for disease diagnosis. Another dimension needing consideration to further buttress on the success of deep learning techniques to image analysis it to seek to improve compositional components of their architectures so that visual feature learning tasks will return better performance when interpreting medical images [165].

4.5 Autonomous vehicles

Autonomous driving has generated extensive research in building autonomous vehicles. This often relies strongly on advances in computer vision and reinforcement learning. The motivation for concentrating research in this direction is to ensure mass production of such autonomous vehicles (AV) and reduce human interference. However, research in this area is characterized with challenges which require a high degree of visual feature learning of concepts and object with the aim of achieving better performances in object detection and semantic segmentation which supports autonomous vehicles process. It is understood that exceptional performance of AV system depends largely on the vehicle's navigation system which in turn requires perfecting the semantic segmentation and object recognition tasks. In addition, the sensor technologies contribute to the navigation system and are significant contributory measure to obtaining accurate outcome capable of yielding safe condition for autonomous driving. Interestingly, we see the techniques of deep learning presenting state-of-the-art solutions to these problems with the aim issues like perception and decision-making which are the component of the AV system, will benefit.

Deep learning methods are suitable for ensuring that all relevant information is gathered using visual feature learning approach so that the perception of AV systems are to correct decision-making for controlling AV under different situations. The concept of perception therefore affects a significant aspect of the AV system including the detection of vehicles, traffic control objects and their associated signal systems, road network infrastructure, detection of the pedestrian infrastructure and persons, localization of signs and object, and other important entities within the environment. It is assumed that this will further help determine path and motion for the AV. Unfortunately, there are more challenging current issues associated with the use of deep learning in AV systems since even the accuracy of some data sources using sensors depends on environmental or weather conditions [166]. In addition, the increasing cost of deploring sufficient sensors and combining all input from the sensors may result in a measure of uncertainty which can adversely affect AV systems [12]. The dependency of AV system on these sensors therefore requires research attention in advancing deep learning techniques and augmented reality-based head-up display (AR-HUD) holds better promise for the system. Interestingly, literature is replete with studies demonstrating this advancement of the use of deep learning have progressed from the basic handcrafted image local features extraction [167] to the now widely proposed Zero-Shot Learning (ZSL). ZSL benefits from using only a few data samples to fully generalize a model since it depends on trained models and other auxiliary information for visual feature learning [168]. Another example of ZSL is the few-shot learning (FSL) which represents an extreme case of ZSL. Both FSL and ZSL are useful for automating object detection and semantic segmentation which are subtasks in AV system [169]. In this section, we understudy some related works to understand the current state-of-the-art in the domain of AV system.

The most challenging task of AV requires that modular operations be carried out to address each contributing problem and subtasks of the system. However, such component-wise approaches often lead to situations where a failed component will adversely impair the accuracy and safety of the autonomous systems. To address these problems, an end-to-end deep learning method has been proposed to allow for single neural network solution for handling all tasks in AV system [170]. Similarly, another end-to-end deep learning method has been proposed which applies CNN architectures for feature learning and extraction, and then fuses the features with other outputs from camera [171]. The fused features are contributory to the improvement of the performance of the motor controllers and drive actuators. A reinforcement deep learning methods was combined with the CNN architecture for achieving an optimized fully autotomized self-driving vehicle. Road assets and barriers represent important sets of objects needing to be recognized when improving the safety conditions of AV system. To address this also, transfer learning, which depends on pre-trained benchmark models such as the inception v3, DenseNet 121, and VGG 19 have been proposed for deployment in identifying these objects [172] and compared approach with those not using transfer learning. Furthermore, recent studies in deep learning aimed at solving AV system problems are now beginning to identify and isolate peculiar challenges for addressing them in detail. Although several subtasks represent the AV systems, issues like the object detection, semantic segmentation and localization of detected object present some interesting fronts. Research efforts have been particularly focused on handling these core functionalities of the system. Visual odometers through the use of deep learning have been channeled towards solving the matter of object localization [173]. The aim is to ensure that object correctly localized provides support to the popular use of global positioning system (GPS) since the visual feature learning of deep learning is suitable for achieving such visioning.

Accurate perception in autonomous vehicles is fundamental to effective semantic segmentation. The issue is semantic segmentation allows for visual signal processing to drill down to pixel level representation for understanding the environment of the AV system. A hybrid method has been proposed which combines CNN, feature pyramid networks and bottleneck residual blocks, and autoendcoders [174] to achieve better semantic segmentation. The evolvement of the use of deep learning models in the scene understanding for AV system follows from handcrafted image processing to rule-based logic systems, and then machine learning. Although CNN architectures reoccur in the deep learning approach, the need to use their ensemble as regression models for autonomous driving has been reported [175]. The study compared the proposed method which stacks two CNN ensemble to the performances of Nvidia proposed CNN, MobileNet-V2 as regression model and Ensemble-M. Drawing inspiration and text-based support through human in the loop for improving the performance of automating self-driving cars have been proposed in [176]. The study noted that by representing features obtained through visual feature learning, in the form of natural language, the human factor can assist the AV system through helpful input. The method uses the approach to achieve semantic segmentation and object-centric RoI pooling. It is assumed that this will support the process of predicting the correct controls and choosing appropriate action to an input. While some of these methods represents the current state-of-the-art, we noted that issues such as semantic segmentation, object recognition, visual feature learning, object detection, visual feature matching and fusion, and 3D information acquisition remain research areas in AV system with potential to benefit from advances in deep learning techniques [177].

Addressing the issue of barriers in the environment of self-driving cars is an important component of the AV system. This is because these types of objects are often found at entry points to housing structures and gates. This increasing call for applying machine learning to estimate cost of increasing barrier heights [178], has further shown the importance deploying deep learning techniques to understanding them in scenes to avoid crashing into them by autonomous vehicles. A hybrid of CNN and long short-term memory (LSTM) a type of recurrent neural network (RNN), have been proposed to address this problem. By understanding the contextual positioning of objects through spatial autocorrelation effect along linear road network paths, barriers can be isolated for avoidance of crashes [179]. On the other hand, the use of visual feature learning using machine learning methods for understanding traffic signals have been investigated [180]. The approach combines deep reinforcement learning (DRL) and transfer learning to design a full ring-and-barrier style controller. In a related work, deep learning model have been proposed for understanding traffic signals in construction site by first recognizing traffic control devices, and then attempts to and as well the construction barricade using YOLOv3 algorithm [181]. Another applicability of visual features is learning to vehicle and traffic control systems in compilation of important statistical information on highways. Using a bounding-box algorithm which has underlying deep learning technique combined with a shaking filter, and Euclidean distance-based similarity measure, passing vehicles were box-bounded for the statistical information gathering [182].

In addition to barrier, a reoccurring issue with AV system is recognition of vehicle plate number using deep learning methods. The need for sourcing input from closed-circuit television (CCTV) to achieve recognition using deep learning model for increase accuracy. A super-resolution generative adversarial network (SRGAN) model has been proposed for use with perspective distortion correction algorithm (PDCA) for ensuring that distorted inputs from CCTV does not impair the recognition operation [183]. Also, deep learning technique based on super-resolution (SR) method and automatic license-plate recognition has been used [184] to identify plate number characters in low-quality video frames. Also, addressing low-quality image obtained through poorly illuminated, and cross angled number plates, morphological transformation, Gaussian smoothing and thresholding were used with K-nearest neighbor (KNN) algorithm for plate number recognition [185]. Similarly, to foster the use of deep learning models in Intelligent Transport System (ITS), FasterRegion-based Convolutional Neural Network algorithm (Faster R-CNN) has been used with morphological methods for plate number recognition. The methods first identify vehicles in traffic, and then narrow closely to identification of plate number region for the recognition task [186]. The characters on the plate are recognized using look-up table (LUT) classifier using adaptive boosting with modified census transform (MCT). When aiming at plate number recognition, increasing accuracy and reducing processing time is very important before a passing car leaves the scene. Using the concept of ROIs, the whole image is first segmented so that visual feature is done using the histogram of oriented gradients method [187] and trained on neural network. Both morphological and edge processing algorithms have also been proposed for plate number recognition in parking spaces [188]. In another work, a single-shot detector- (SSD-) deep learning model have been used with deep convolutional network and long short-term memory (LSTM) [189] to recognize plate numbers in difficult situations such as real-time applications having artificial noise and different light and weather conditions. Similarly, CNN and deep bidirectional LSTM (DB-LSTM) network have been combined for visual feature learning for plate number recognition from video frames of moving vehicles [190].

The need to consider building on the hardware and software aspects of autonomous vehicles and plate number recognition is challenging though it requires research attention for better performance. Availability of datasets in addition to the CamVid dataset, the Berkeley DeepDrive eXplanation (BDD-X) dataset, publicly available for research holds promise for using deep learning architectures for visual feature learning.

4.6 Other applications

Visual features using deep learning techniques are also applicable to other domains in addition to the focused domains provided in previous subsections. Their usefulness for understanding global and local features in videos and images presents an interesting opportunity for solving both classification and segmentation problems. For instance, a deep learning method called VICRegL [191] has been proposed for understanding image representations. The method to simultaneously learn both global and local features for achieving linear classification and segmentation transfer tasks. The VICRegL consists of a dual branches CNN which feeds on same image sample with different distorted degree. Similarly, visual feature learning on both video and images have been researched for learning features from large-scale unlabeled data using self-supervised deep learning methods [192]. While the usefulness of deep learning models for visual feature learning is now widely successful for abstraction of complex representation, interpretability of these models is considered black-box. As a result, Invertible neural networks have been proposed to learn invariance in data and transform them into expressive semantic concepts for extracting meaning [193]. This further allows for update this representation in a semantically reasonable manner for explaining the block-box effect of neural networks. Table 3 provides a detailed summary of all the studies reviewed under this section considering the applications of visual feature learning.

In the following section, we highlight the current challenges of visual feature learning to reveal possible future directions for research in this evolvement of deep learning neural networks.

RQ3.What are the existing challenges and limitations in visual feature learning?

5 Challenges of visual feature learning

Due to the growing usage of image classification and retrieval, together with the proliferation of enormous scale image data, how to utilize data-driven approaches to learn the feature representation of images and how to use high-dimensional vectors to represent the discriminative features of images have become significant study areas [194]. Feature expression in images using traditional approaches is primarily manual and depends on prior data thus, hindering the accuracy of these extracted visual features. The precision of image classification and video understanding tasks has dramatically increased today as a result of the development of Deep learning (DL) model, which have greatly enhanced the ability to differentiate image features under supervisory signals. However, as there is more unlabeled image data in the real world, it is impractical to annotate every sample precisely. DL models and clustering techniques have been integrated in some earlier attempts to narrow the vast distinction in performance between various features. More research is being done on this mechanism, which is currently being applied to more discriminative visual feature learning. Researchers all over the world are now racing to use the concept of visual features to solve more and more problems. However, there are several challenges faced by the proposed approaches to visual feature learning. These challenges are depicted in Fig. 8 and discussed in the preceding sub sections.

5.1 Data scarcity and quality

The generation of data in modern industrial applications is growing as a result of improved technological sensors' accessibility, availability, and cost-effectiveness. In truth, huge amounts of data are frequently assumed to be available using modern techniques of data analysis. Many DL models need massive amounts of data for accurate inference. This is true for numerous use cases [195], including machine problem diagnosis [196] and preventative maintenance [195]. With regard to understanding visual features, this is typically not the case. When DL models are being trained for feature interpretation, data scarcity is a significant barrier. To attain great performance, DL needs a lot of data. Unfortunately, many applications that depend on visual understanding only have scant or insufficient training data for DL frameworks. In many of these application sectors, such as marketing, computer vision, and medical science, there is only a small amount of data available for training DL models because obtaining fresh data is either impractical or requires more resources. Furthermore, for these models to avoid the over fitting issue, a lot of data must be collected. Therefore, more research on data generator techniques is required to produce sample data in situations where the real data is either unavailable or needs to be kept confidential due to personally identifiable information (PII) or compliance problems.

5.2 Overfitting and generalization

Overfitting is one of the most frequent and harmful problems you may incur with your deep learning models. An overfitted model is produced when a model performs well on training data but poorly on new, untested data. In other words, the model identified patterns that are specific to the training set and irrelevant in other data sets. Overfitting is a prevalent error in visual feature learning algorithms as well where a model tries to match the training data and ends up retaining both the data patterns, the noise and random shifts. Since these models do not generalize effectively and do not perform well in the presence of unknown data instances, their intended purpose is typically not accomplished completely. Some techniques have been put out as ways to prevent overfitting in regular training, using explicit regularization and data augmentation for Deep learning [197].

5.2.1 Explicit Regularization

Adding an explicit regularization component to the loss and penalizing the complexity of the model parameters is a traditional technique for avoiding overfitting. It makes sense to reduce the number of features if overfitting arises as a result of an overly complex model. However, in certain instances, one may be unsure as to which features to keep and regularize in the model. If such happens, regularization techniques like Lasso and L1 can be helpful. Although this technique is indeed helpful but too much regularization on image and video data can be detrimental. We incur the danger of underfitting, where the model performs badly on training data and is unable to describe the relationship between the input data and output class labels (since the model capacity is excessively restricted).

Table 3 A summary table on the application of deep learning techniques to the areas of object detection and recognition, facial recognition, scene understanding, medical image analysis, autonomous vehicles and other areas of application

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[112]	ShuffleNetV2 and L-Net: A lightweight object detection	Object detection	Reduced complexity of the neural network Extract a better discriminative image feature	Losses the full benefit of ShuffleNetV2 architecture
[113]	Combination of point-wise separable (PWS), depth-wise separable (DWS), and spatial feature pyramid network (SFPN) neural architectures. Transformer encoder–decoder and feature fusion models were also added to the framework	Object detection	Reduced computation time for object detection Capable of extracting low-level and spatial-level features	Increased level of component combination
[114]	Dense PRediction Simplified (DPRS) was based on the popular YOLO architecture	Object detection	Performance enhancement through fusion with backbone network	Drawback from ineffective fusion method
[115]	Optimized YOLOv4 leading to what is known as SlimYOLOv4: Added MobileNetV2, DCS, DO-DConv, and ReLU6	Object detection	Reduces computation and improves network performance Improve the numerical resolution, speed and accuracy	Suffers from the omission of Leaky ReLU as used in YOLOv4
[116]	Micro-YOLO + : YOLOv3-tiny network with lightweight convolutional layers featuring DSConv, MBConv, and GConv	Object detection	Balance the neural network model size so that devices with restricted resources and low power requirements can run the model Suitable for network accuracy, and inference speed	Oversimplified for embedded devices so that the trained model is strictly useful on such devices

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[117]	MFRDet: a lightweight framework capable of reusing learned features in deep and shallow feature maps	Object detection	Aimed to outperform two-stage algorithms based on accuracy Avoids the use of very deep convolution neural networks Improves precision through reusability of information from deep and shallow feature maps	The shallowness of the neural network positions it to underperform when learning visual features
[118]	African Adam Optimization-based SqueezeNet with GAN assisted approach	Object detection and recognition (IODR)	Improved the accuracy, precision, and recall metrics	Suffers from the challenge of two-staged algorithms
[120]	YOLO technique for identification of, and search for items in video	Object detection and search in video	Improvement of sparsity compared with similar approach	Weak backbone network for visual feature learning
[121]	Mask R-CNN model combined with a text-to-speech generator for object detection and position localization	Object detection and localization	Improve the accuracy of the detection and localization of objects in images	The region based method have a general deficiency which puts them behind the single-stage approach
[119]	NAS applied to architectural evolvement to search for feature pyramid network which can detect objects in image samples	Object detection	Make a balance between performance and efficiency	Approach follows the basic CNN method to object detection
[122]	Unsupervised neural network model for contextual visual feature learning	Object detection	Benefits from the little or no supervision or labeling while addressing the face recognition problem	Demand that model be adequately trained for effective visual feature learning
[133]	Basic neural computing model	Face recognition	Overcome subjective factors of handcrafted feature extraction which is often time-consuming and laborious	The simplicity of the neural model does not achieve fine-grained visual feature learning process
[137]	Tree Structured Part Model which relies on Scale Invariant Feature Transform for visual feature learning	Face recognition	Suitable for handling frontal and profile faces	Machine learning algorithms such as Support Vector Machine was used for classification task

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[138]	Deep learning method: fisherface was supported by Local Binary Pattern Histogram (FF-LBPH)	Detection of deep fake face image	Capable of detecting deep fakes to eliminate circulation of fake facial representations	Tendency of filtering out discriminant features through the use of FF-LBPH
[136]	Transfer Learning method using GoogLeNet and FaceNet, clustering technique	Face verification and recognition	Obtained good performance despite small size of training set	Selection for face matching using the clustering technique suffers from the limitation of the clustering method applied
[139]	Hierarchical deep neural network model of simulating ventral visual stream	Face recognition	Leverage on face-selective neurons to simulate neural network model	Uses random feed forward connections in untrained deep neural networks for the visual selectivity
[140]	Pre-trained face recognition deep neural network (DNN) with a diverse array of stimuli	Face recognition	Links between artificial and primate neural systems	Unperfected simulation of the artificial units impaired on performance
[141]	deep neural networks was trained from stimuli extracted from trained animal rat	Face recognition	Previously collected stimuli and unseen stimuli were used to aid generalization	Unperfected simulation of the artificial units impaired on performance
[142]	RetinaNet baseline, a single-stage face detector	Face recognition	Addresses occlusion, blurring, and partially occluded faces problems	Method is based on a single-staged algorithm
[143]	Deep learning model and OpenCV library	Face recognition	Simple and easy to implement	Visual feature extraction are not fine-grained
[144]	Face mesh	Face recognition	Capable of detecting and recognizing the face under varying conditions Suitable for handling non-frontal images of males and females of all ages and races	Attenuating the model to varying illumination and background is challenging
[145]	Average-pooling layers were replaced by max-pooling layers in convolutional-pooling blocks of CNN	Face recognition	Simple approach to deep learning model enhancement on face recognition	Less significant network enhancement with little performance improvement

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[146]	A dual branch deep neural network using MWIR with GAN-based network using transfer learning	Face recognition	Achieves face recognition using MWIR images	Complex of neural networks used for the face recognition task
[149]	multi-modal end-to-end deep learning model	Scene understanding	Trains on a rich dataset named ARVSU	The use of a speaker's transcripts of utterance for addressee recognition impairs the scene understanding process
[150]	Incremental inductive deep learning technique and non-monotonic logical reasoning	Scene understanding	Learn using small number of training examples	Combination of logic and deep learning remain is currently challenging leaving possibility for errors
[151]	Basic CNN architecture	Scene understanding	Simplicity of neural network applied	Vague approach description to the visual feature learning process
[152]	multimodal and multi-tasking neural networks	Scene understanding	Memory footprint is substantially reduced Increased inference speeds	Performance improvement is dependent on if the associated tasks share complementary information, or act as a regularizer for one another Have the challenge of optimizing multiple objectives in parallel since one or more tasks could start to dominate the weight
[152]	Neural network architecture for jointly tackling multiple dense prediction tasks and a branched multi-task networks	Scene understanding	Supports deeper layers of the neural network gradually grow more task-specific	Achievement of task similarity was at the detriment of network complexity
[52]	Used CNN for extracting features from multi-view forms of MLO and CC in digital mammography	Medical image analysis for disease feature learning and classification	Leverage the technique of transformational image augmentation	The CNN architecture is underrepresented

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[16]	Optimizing the combinatorial problem of hyperparatemers of the CNN	Medical image analysis for disease feature learning and classification	A simple approach was used of representing different CNN parameters	Random-grid model for selecting an optimal combination of hyperparameters is very basic when compared with the metaheuristic method
[46]	Uses wavelet-based function to replace the classical rectifier linear unit (ReLU) based function as activation function of CNN	Medical image analysis for disease feature learning and classification	Improved classification accuracy	Wavelet-based function is more complex and time consuming compared with ReLU function
[154]	CovFrameNet for the detection features suggestive of the presence of corona virus in chest X-ray	Medical image analysis for disease feature learning and classification	Improve the feature learning for aiding the detection of COVID-19	Imbalance dataset suggested model overfitting
[155]	Machine learning framework for contact tracing for Covid-19 in smart city	Medical image analysis for disease feature learning and classification	A hybrid of machine and deep learning models for improved performance	Method proposed was not experimentally evaluated
[156]	CNN architecture combined with IoT devices for feature learning to detect abnormal intrusion	Medical image analysis for disease feature learning and classification	Promotes interfacing neural networks with inputs from IoTs sensors	Limited human cues, gesture, and positions were represented in the dataset used for training the model
[45]	ROImammoGAN: region-based generative adversarial neural networks for data augmentation	Medical image analysis for image synthetization	Synthesized images are rich in region-based and class-label-based features	The GAN was trained multiple times as the number of class-labels
[98]	ArchGAN: whole-image-based generative adversarial neural networks for data augmentation	Medical image analysis for image synthetization	Training complexity often associated with GANs was overcome	Samples only with architectural distortion can be generated using the trained GAN
[99]	EOSA-GAN: metaheuristic optimized generative adversarial neural networks for data augmentation	Medical image analysis for image synthetization	Generator component of the GAN has it weight and biases optimized for better synthetization	Increased training time due to the use of metaheuristic algorithm
		Medical image analysis with GAN-based data augmentation		

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[157]	Whole-image based CNN (WCNN) and region-based CNN (RCNN) architectures were systematically arranged to achieve the proposed framework	Medical image analysis for disease feature learning and classification and localization	Dual branch neural network for richer visual feature learning	Fusion layer depends on the softmax probability distribution
[17]	Hybrid CNN with the GA, WOA, MVO, SBO, and LCBO metaheuristic algorithms	Medical image analysis for disease feature learning and classification	Successfully evaluated category-based optimizer with CNN for hyperparameter selection	Hyperparameter selection and combination has now been adequately addressed
[159]	Neural architecture search (NAS) using metaheuristic algorithm and CNN based on a block-based stochastic categorical-to-binary (BSCB) algorithm	Medical image analysis for disease feature learning and classification	Evolved problem-specific CNN architecture for feature learning and classification using histopathology images	Formalism for CNN components during optimization purpose requires better algorithmic approach
[160]	Filters out non-discriminant visual features from CNN using IEOSA	Medical image analysis for disease feature learning and classification	Visual feature optimization avoids bottlenecking classifier	Dimensionality reduction using PCA for image features introduces error
[158]	Hybrid CNN with the EOSA, GA, WOA, MVO, SBO, and LCBO metaheuristic algorithms	Medical image analysis for disease feature learning and classification	Optimizes deep learning model for visual feature learning for lung cancer	Several level of image preprocessing were added which may unnecessary filter out relevant features
[161]	Convolutional autoencoder deep learning framework	Medical image analysis for disease feature learning and classification	Achieved better visual feature learning with small samples	The Convolutional autoencoder is suboptimal compared to recent state-of-the-art architectures
[162]	Sample-efficient deep learning with transfer learning methods	Medical image analysis for disease feature learning and classification	high diagnosis accuracy Trained with only a few CT images	Possibility of trained model suffering from non-generalization
[163]	Deep learning based attribution framework using adaptive path-based gradient integration techniques	Medical image analysis for disease feature learning and classification	Promotes understanding the input-prediction correlative structures for deep learning model	Applicability to identifying bio-markers can be challenging
[170]	End-to-end deep learning architecture	Autonomous driving	Applies a single pipeline to learn sub-tasks of AV systems	Lacking in capacity in complexity of visual feature learning for AV system

Table 3 (continued)

Ref	Deep learning technique for feature learning	Category of application of method	Advantage	Disadvantage
[172]	Using pre-trained inception v3, denseNet 121, and VGG 19 for roadside barrier identification	Autonomous driving	Estimating, detecting, and localization of traffic barrier through images	Failed to gather sufficient observation on roadside barriers
[173]	Visual odometer combined with the use of deep learning	Autonomous driving	Outperformed and override the possible reliance of GPS	Object detection is not adequately handled
[171]	Supervised CNN and reinforcement deep learning methods for end-to-end AVS	Autonomous driving	Single-algorithm for complete AV system subtasks	The model is limited by complexity of modules combined which were not isolated for focused accuracy
[174]	Hybrid of CNN, feature pyramid networks and bottleneck residual blocks, and autoendcoders	Autonomous driving	Support understanding environment for self-driving system Accurate real-time visual signal processing	Visual feature can be impaired by the neural network in the hybrids when fusion method is not properly handled
[175]	Ensemble of CNN for regression models for autonomous-driving	Autonomous driving	Improved performance compared with single CNN architecture	Approach for visual feature fusion mechanism is not adequately reported
[176]	Combines visual feature learning using CNN and attention mechanism for semantic segmentation and summarization visual observations in NL	Autonomous driving	Approach is grounded in semantic representation	Interpretable visual explanations depend only on visualizing object-centric attention maps
[179]	hybrid of CNN and RNN/LSTM	Barriers detection for Autonomous driving	Contextual Scene understanding	Jointly trained two different models for visual feature learning
[186]	Faster R-CNN have been used with morphological technique	Vehicle plate number recognition	detect all the vehicles recognize plates numbers	Inaccuracy of morphological operations can crop out sensitive figures
[189]	SSD-deep learning model, deep convolutional network, and LSTM	Vehicle plate number recognition	End-to-end method a single-shot detector of plate numbers	Using SSD on real-life video frames requires optimization

5.2.2 Data Augmentation for Deep learning

Data augmentation is one of the solutions to the issue of limited data. It is a widely used method for improving an over fitted data model's generalizability. In order to increase the amount of data, guided techniques are used to add transformed copies of already existing data. This generates newly synthesized data from the existing data. However, data augmentation raises the danger of over fitting the model by creating extra training data and exposing the model to different data versions. Data augmentation is seen by researchers as a breakthrough in terms of improved accuracy and the amount of time needed to enhance existing data for training DL models. Along with its benefits, data augmentation still has challenges to overcome, such as a lack of mechanisms for assessing the quality of enhanced datasets and the need for fresh research to generate synthetic or new data in order to support advanced applications. Another difficulty is determining the best approach for data augmentation since, in many cases, biased data from the augmented dataset is present.

5.3 Interpretability and explainability

Since the 1970s, Artificial Intelligence has evolved, fundamentally reinventing decision-making systems. With a focus on accuracy, learning can now be done directly from data rather than human knowledge, thanks to the development of machine learning. But accuracy is no longer sufficient to assess the power of deep learning models for visual feature learning. Other criterions include interpretability and explainability. Interpretability describes how precisely a machine learning model can connect a cause to an effect. The ability to speak with or explain in human words is aided by interpretability. The degrees of interpretability may differ significantly in the field of visual feature learning. Some techniques for visual feature learning are easy for humans to understand, thus they are friendly. Others require ad-hoc techniques to acquire an interpretation since they are too complex to comprehend. Explainability has to do with how well the parameters, which are frequently concealed in Deep learning models, can defend the outcomes. There are ways to get the interpretation or the explanation, though. There is not, however, a single technique that can be used to every visual learning model without risk and with reliable results. Additionally, once a method is chosen, it is crucial to use it carefully, largely because certain methods lack consistency, which is another challenge to visual feature learning models. This is because most of the methods for interpretation and explanation frequently depend on a particular dataset, region, or portion of the data space for both interpretation and explanation to be valid. Furthermore, some interpretation techniques fail to take into account the relationships between the features or only provide one counterfactual explanation when several others may have been offered. Due to these difficulties, it is crucial that researchers concentrate their efforts on creating interpretation and explanation methods that do not rely on the aforementioned difficulties.

5.4 Ethical considerations

There are ethical and legal concerns with regards to visual feature learning. This can be divided in to privacy, morality and the introduction of potential biases as shown in Fig. 9.

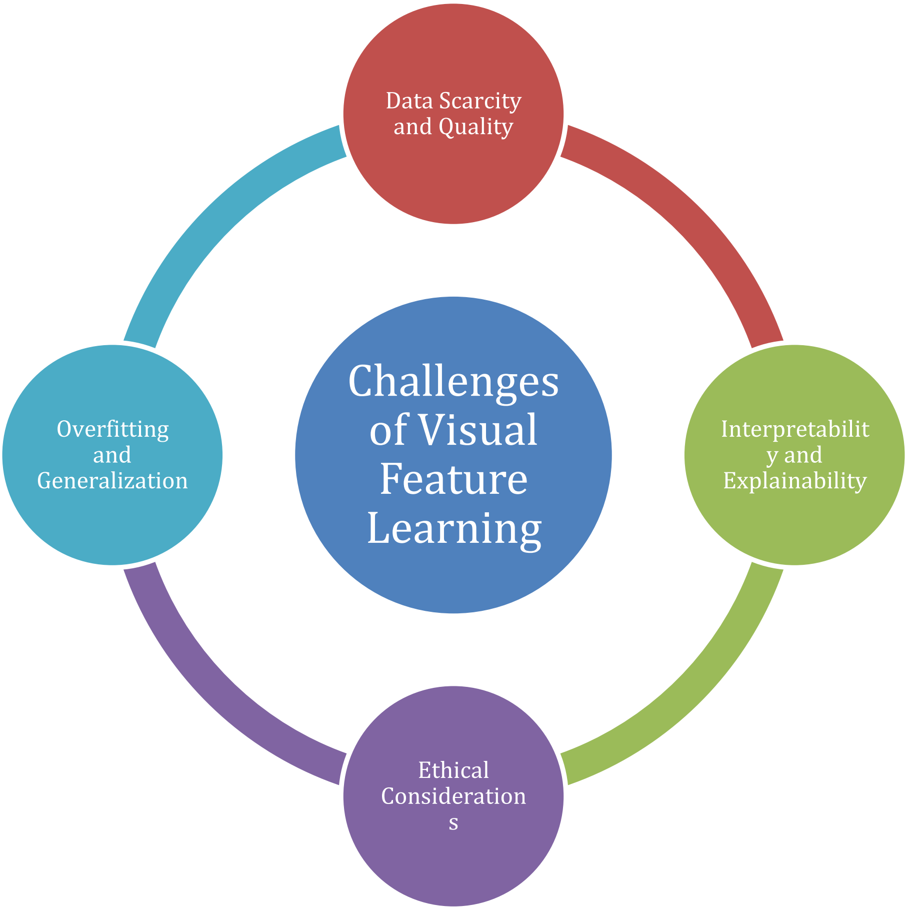
Data Scarcity and Quality
Overfitting and Generalization
Challenges of Visual Feature Learning
Interpretabilit y and Explainability
Ethical Consideration s

Fig. 8 Challenges of Visual Feature Learning

5.4.1 Privacy

The recently developed machine learning techniques, such as visual feature learning models, are a major factor behind the revolutionization of numerous applications such as object and pattern recognition. Large volumes of personal data on people are frequently gathered for visual feature learning purposes, whether through social media platforms, healthcare systems, or other means. Although this data can be very helpful for developing models and making predictions, it also raises important concerns about security and privacy. Who is the owner of this data, and who has access to it? How do we make sure that private data is shielded from bad actors? Furthermore, privacy has become a major issue in this era of artificial intelligence based on machine learning. It is significant to emphasize that because machine learning can be both a friend and an enemy, the problem of privacy preservation in the context of visual feature learning is very different from that in standard data privacy protection. As most existing solutions mainly target privacy issues during the machine learning process, the research on privacy preservation and machine learning is still in its infancy. As visual feature learning progresses, these intricate ethical issues must be resolved. [198].

5.4.2 Morality

Globally, the rapid development of artificial intelligence (AI) in terms of visual feature learning models has opened an array of possibilities, from making medical diagnosis easier to enhancing image and video understanding. But these swift developments also give rise to serious concerns about morality [199].

5.4.3 Biases

The danger of maintaining and aggravating biases is one of the most critical problems with visual feature learning. Algorithms for visual feature learning can only be as objective as the data they are trained on, and if that data reflects biases, the algorithms will as well. Anything because of this could be unfair. The continual issue of eliminating bias in visual feature learning models is a challenge that must be resolved.

RQ4.What are the potential future directions for advancing visual feature learning?

6 Future directions of visual feature learning

Visual Feature Learning is a fundamental aspect of deep learning that has revolutionized computer vision tasks by enabling machines to extract meaningful representations from raw visual data. From the discoveries achieved by Convolutional Neural Networks [21] to the advancements in pre-training, transfer learning [1], autoencoders [200], GANs, and self-supervised learning [200], the field has witnessed tremendous growth. Visual feature learning techniques have enabled machines to understand and interpret visual data, which surpasses human performance in several domains. As research continues to evolve, visual feature learning will remain a key area, which will shape the future of deep learning and expand the frontiers of computer vision applications. As the field continues to evolve, here are some future directions to consider:

6.1 Hybrid techniques

Hybrid techniques in visual feature learning have demonstrated remarkable success in leveraging the strengths of different approaches to enhance the quality and effectiveness of feature extraction [23]. While deep learning has achieved significant breakthroughs in visual feature learning, here are some future direction of hybrid techniques in visual feature learning are discussed.

Integration of Deep Learning with Classic Computer Vision Methods Classic computer vision techniques still hold value in certain scenarios by integrating deep learning models with classic methods such as handcrafted features, image segmentation, or geometric transformations. The fusion of these approaches can combine the power of deep learning's representation learning with the interpretability and efficiency of classic techniques [23].

Multi-Modal Fusion VFL can benefit from merging information across multiple modalities such as text, images and audio or sensor data. Hybrid techniques can effectively integrate information from different modalities to learn more comprehensive and robust visual representations. Thus, enabling models to leverage on the complementary strengths of different modalities and enhance performance in tasks such as image captioning, video understanding, or multi-modal retrieval [23].

Transfer Learning and Domain Adaptation Transfer learning and domain adaptation play vital roles in real-world applications where the target domain differs from training domain; here hybrid techniques can explore method that will leverage both labeled and unlabeled data from different domain to learn transferable visual features. Thus, combining domain adaptation algorithms, unsupervised pre-training and fine-tuning strategies to effectively adapt models to new domains with limited labeled data [1, 20].

6.2 Unsupervised feature learning

Unsupervised feature learning plays a crucial role in visual feature learning by allowing models to extract meaningful representations from unlabeled data [20]. As the field continues to advance, there are several exciting future directions and potential advancements in unsupervised feature learning that can further improve representation learning and enhance the performance of visual feature learning models. Here are some key areas to consider:

Contrastive learning has emerged as a powerful technique in unsupervised feature learning. Future research can explore new approaches and advancements in contrastive learning methods, such as better sampling strategies, more effective similarity metrics, or improved negative sample generation. In addition, self-supervised learning formulates unsupervised

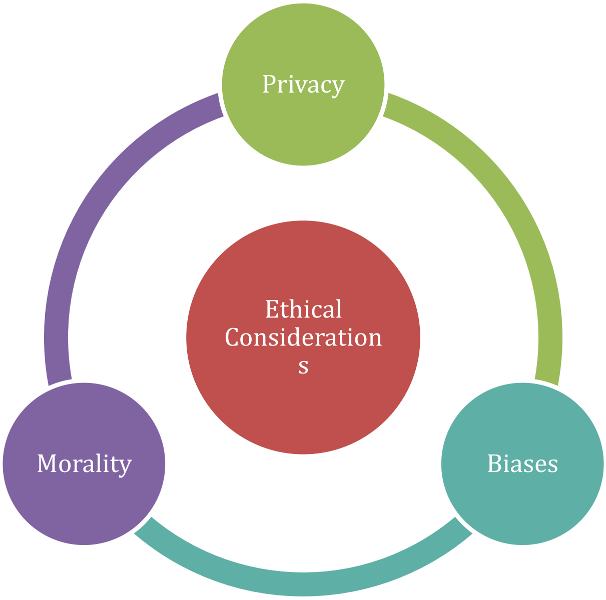

Fig. 9 Ethical Considerations to Visual Feature Learning

learning tasks from data itself can be further explored. This includes designing novel pretext tasks that capture high level semantics or incorporate domain specific knowledge for improve self-supervised learning [201].

Generative Models for Unsupervised Learning Generative models, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), have shown promise in unsupervised feature learning. Future research can focus on developing advanced generative models that can capture complex data distributions and learn disentangled representations. These involved exploring improved training techniques, regularization methods and evaluation metrics for generative models. In addition, leveraging generative models can enable unsupervised learning with the ability to generate new samples and perform data augmentation for improved visual feature learning [200].

Hierarchical and Multi-Level Representations Future research can explore the learning of hierarchical and multi-level representations in unsupervised feature learning. This involves capturing representations at different levels of abstraction, from low-level visual features to high-level semantics. Hierarchical representations can enable models to capture complex dependencies and abstract concepts, leading to more robust and expressive visual features. Techniques such as autoencoders with multiple layers or self-supervised learning with progressive tasks can be explored to learn hierarchical and multi-level representations [200].

6.3 Continual learning

Continual learning also known as lifelong learning or incremental learning, focuses on enabling deep learning models to learn continuously from a stream of data, by retaining knowledge and adapting to new tasks or concepts without catastrophic forgetting [202]. In the context of visual feature learning, continual learning is particularly important to ensure models can adapt and improve over time. Here are some future directions:

Catastrophic Forgetting Mitigation One of the main challenges in continual learning is mitigating catastrophic forgetting, where the model forgets previously learned knowledge when exposed to new data. Future research can explore techniques that effectively address this issue. This includes regularization techniques (e.g., elastic weight consolidation), rehearsal-based approaches (e.g., generative replay), or architectural modifications (e.g., using separate modules for different tasks) to alleviate forgetting and preserve previously learned representations [203].

Adaptive Model Capacity The Future research can focus on developing models with adaptive capacity to dynamically allocate resources for new tasks while leveraging existing knowledge. This involves designing architectures that can expand or compress their capacity based on task complexity or novelty. It allows models to efficiently allocate resources and avoid unnecessary growth [203].

Task-Incremental Learning Task-incremental learning refers to the ability of models to learn new tasks while retaining knowledge of previously learned tasks. Future research can explore methods that allow models to learn from new visual tasks incrementally without the need for extensive retraining on the entire dataset. This includes techniques such as parameter isolation, task-specific gating mechanisms, or dynamically managing task-specific memory to enable efficient and scalable continual learning [204].

6.4 Attention-based models

Attention-based models have shown significant success in various visual tasks by allowing models to focus on relevant regions or features in an input image [205]. Here are some key areas to consider.

Fine-grained Attention and Localization Future research can focus on developing attention mechanisms that enable more fine-grained localization and selection of relevant regions or features within an image. This involves exploring techniques that can capture more precise spatial information and attend to smaller or more specific regions of interest by enhancing the fine-grained attention capabilities of models, attention-based methods can improve localization accuracy and provide more detailed insights into the visual features contributing to model decision making [206, 207].

Hierarchical Attention and Multiscale Processing Attention-based models can benefit from hierarchical attention mechanisms that operate at different scales and levels of abstraction. Future research can explore methods that incorporate multiscale processing and attention, allowing models to attend to features at different levels of granularity. This includes designing architectures that can dynamically allocate attention across multiple layers or levels,enabling the model to capture both local details and global context simultaneously [208].

Explainability and Interpretability Attention-based models have the potential to provide interpretability and explainability by highlighting the regions or features that contribute most to the model's predictions. Future research can focus on developing attention mechanisms that do not only improve performance but enhance model interpretability. This involves designing attention models that can provide clear and meaningful explanations for their decisions, allowing users to understand and trust the model's reasoning process [209].

6.5 Explainable AI

Explainable AI (XAI) is aimed at providing humanly understandable explanations for the decisions made by AI models. In the context of visual feature learning, the ability to explain the reasoning behind the model's predictions is crucial for building trust, improving transparency, and facilitating collaboration between humans and AI systems [210]. Here are some future directions.

Contextual Explanations Visual feature learning often relies on the context surrounding the visual information. Future research can explore methods that provide contextual explanations, considering the relationships between objects, scenes, or events in an imageThis involves developing techniques that can capture and explain the context-aware reasoning of AI models, allowing users to understand how the model's decisions are influenced by the overall context [210].

Multimodal Explanations Visual feature learning is not limited to images alone but often it involves multimodal data, such as text, audio, or sensor inputs. Future research can focus on developing methods that provide explanations that integrate information from multiple modalities. This includes techniques that can generate explanations to connect visual features with textual descriptions, audio cues, or other modalities enabling a more comprehensive and holistic understanding of the model's decisions [201].

Domain-specific Explanations Visual feature learning is applied in various domains, each with its own unique characteristics and requirements. Future research can explore domain-specific explanations that are tailored to the needs and constraints of specific applications. This involves developing techniques that can provide explanations relevant to specific domains, such as healthcare, autonomous driving, or industrial inspections, ensuring that the explanations are aligned with the domain-specific constraints, regulations, and user expectations [211].

7 Discussion and conclusions

7.1 Summary of findings

The systematic literature review on visual feature learning has revealed several key findings. Firstly, deep learning techniques, particularly CNNs and GANs, have emerged as state-of-the-art methods for visual feature learning. These models have demonstrated impressive performance in extracting and representing meaningful features from raw input data. Various architectures and training strategies have been proposed to enhance the learning capabilities of these models.

Secondly, visual feature learning has found widespread applications in computer vision. These include object recognition, image classification, image generation, video analysis, and more. Deep learning-based visual feature learning approaches have shown remarkable results in improving the accuracy and efficiency of these applications. The extracted features provide rich representations that capture relevant information, leading to better performance in various visual tasks.

Thirdly, the review identified several challenges and limitations in visual feature learning. One major challenge is the need for large labeled datasets for training deep learning models. Data scarcity and annotation efforts can hinder the performance and generalizability of visual feature learning algorithms. Additionally, the interpretability of learned features and the robustness against adversarial attacks remain areas of concern.

7.2 Implications for research and practice

The findings of this review have significant implications for both research and practice. For researchers, the review provides a comprehensive understanding of the state-of-the-art deep learning techniques for visual feature learning. It highlights the potential areas of improvement and guides future research directions. Researchers can explore novel architectures, training strategies, and data augmentation techniques to address the challenges identified.

For practitioners, the review underscores the importance of visual feature learning in various computer vision applications. It provides insights into the most effective deep learning models and architectures for extracting meaningful visual features. Practitioners can leverage these findings to enhance the performance and efficiency of their visual tasks, leading to improved accuracy and better user experiences.

7.3 Limitations and future research directions

While this systematic literature review offers valuable insights, it is not without limitations. Firstly, the search strategy may have omitted some relevant studies, although efforts were made to minimize this possibility. Additionally, the review focused primarily on deep learning techniques and may have missed potential contributions from other approaches to visual feature learning.

Future research should address these limitations and explore promising avenues. To overcome the data scarcity challenge, researchers can investigate techniques such as transfer learning and domain adaptation to leverage pre-trained models on related tasks or domains. Moreover, advancements in explainable AI and interpretability techniques can help unravel the black box nature of deep learning models and provide insights into the learned visual features.

Furthermore, the integration of visual feature learning with other modalities, such as textual or sensor data, could lead to more comprehensive and robust representations. Exploring the combination of multiple modalities and their fusion can open up new opportunities for visual feature learning in multimodal settings.

Finally, this systematic literature review on visual feature learning provides a comprehensive overview of deep learning techniques, applications, challenges, and future directions. It highlights the significance of visual feature learning in computer vision and presents key findings that can inform future research and guide practical applications. The review identifies the state-of-the-art techniques, emphasizes the challenges faced, and suggests potential avenues for advancing the field of visual feature learning.

Acknowledgements This work was supported by Tertiary Education Fund (TETFund) under National Research Fund (NRF), Nigeria with reference number TETF/DR&D/CE/NRF-2021/SETI/SAE/00111/VOL.1.

Author contribution All authors have contributed in measure to the completion of the whole study.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors have declared that no competing interests exist.

References

1. Cai L, Gao J, Zhao D (2020) A review of the application of deep learning in medical image classification and segmentation. Ann Transl Med 8(11):713
2. Toshpulatov M, Lee W, Lee S, Roudsari AH (2022) Human pose, hand and mesh estimation using deep learning: a survey. J Supercomput 78:7616–7654

3. Islam M, Liu S, Wang X, Xu G (2020)"Deep learning for misinformation detection on online social networks: a survey and new perspectives," Soc Netw Anal Min. 82.
4. Gupta A, Anpalagan A, Guan L, Khwaja AS (2021) Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array 10:100057
5. Sarker IH (2021) Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science 2:420
6. Jing L, Tian Y (2022) Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. IEEE Trans Pattern Anal Mach Intell 43(11):4037–4058
7. Chen S, Guo W (2023) Auto-Encoders in Deep Learning—A Review with New Perspectives. Mathematics 11(8):1777
8. Li D, Wang R, Chen P, Xie C, Zhou Q, Jia X (2022) Visual Feature Learning on Video Object and Human Action Detection: A Systematic Review. Micromachines 13(1):72
9. Arkin E, Yadikar N, Xu X. et al. (2022) "A survey: object detection methods from CNN to transformer." Multimed Tools Appl.
10. Kumari M, Kaul A (2023) "Deep learning techniques for remote sensing image scene classification: a comprehensive review, current challenges, and future directions," Concurr Comput. pp. 1–126.
11. Oyelade ON, Ezugwu AE (2020) A state-of-the-art survey on deep learning approaches in detection of architectural distortion from digital mammographic data. IEEE Access 8(2020):148644–148676
12. Pavel MI, Tan SY, Abdullah A (2022) Vision-based autonomous vehicle systems based on deep learning: a systematic literature review. Appl Sci 12(6831):1–51
13. Bengio Y, Lecun Y, Hinton G (2021) Deep Learning for AI. Commun ACM 64(7):58–65
14. Koh DM, Papanikolaou N, Bick U, Illing R, Kahn CE Jr, Kalpathi-Cramer J, Matos C, Martí-Bonmatí L, Miles A, Mun SK, Napel S, Rockall A, Sala E, Strickland N, Prior F (2022) Artificial intelligence and machine learning in cancer imaging. Communications Medicine 2:2022
15. Bharath M, Reddy S, Rana P (2021) "Biomedical image classification using deep convolutional neural networks – overview," in IOP Conf. Series: Materials Science and Engineering.
16. Oyelade ON, Ezugwu AE (2022) A comparative performance study of random-grid model for hyperparameters selection in detection of abnormalities in digital breast images. Concurrency and Computation: Practice and Experience 34(13):e6914
17. Oyelade ON, Ezugwu AE (2021), "Characterization of abnormalities in breast cancer images using nature-inspired metaheuristic optimized convolutional neural networks model," Concurr Comput Pract Exp, Wiley. https://doi.org/10.1002/cpe.6629
18. Yaqub M, Jinchao F, Arshid K, Ahmed S, Zhang W, Nawaz MZ, Mahmood T (2022) "Deep learning-based image reconstruction for different medical imaging modalities," Comput Math Methods Med. 2022.
19. Sehar U, Naseem ML (2022) How deep learning is empowering semantic segmentation: traditional and deep learning techniques for semantic segmentation: a comparison. Multimed Tools Appl 81:30519–30544
20. Wang Y, Yan WQ (2022) Colorizing Grayscale CT images of human lungs using deep learning methods. Multimed Tools Appl 81:37805–37819
21. Fan J, Xie W, Ge T (2022) "Automatic gray image coloring method based on convolutional network," Comput Intell Neurosci, vol. 2022.
22. Sharma T, Debaque B, Duclos N, Chehri A, Kinder B, Fortier P (2022) Deep Learning-Based Object Detection and Scene Perception under Bad Weather Conditions. Electronics 11(4):563
23. Ouadiay FZ, BouftaihHamza H, Houssine B, Himmi BM (2018) "Simultaneous Object Detection and Localization using Convolutional Neural Networks," in ISCV'18.
24. Long Y, Zhai X, Wan Q, Tan X (2022) Object Localization in Weakly Labeled Remote Sensing Images Based on Deep Convolutional Features. Remote Sens 14(13):3230
25. Li Z, Caro JO, Rusak E, Brendel W, Bethge M, Anselmi F, Patel AB, Tolias AS, Pitkow X (2023) Robust deep learning object recognition models rely on low frequency information in natural images. PLoS Comput Biol. 19:3
26. Vishnu R, Prakash NK (2021) "Mobile application-based virtual assistant using deep learning," in Soft computing and signal processing: advances in intelligent systems and computing.
27. Zhang L (2022) Applying Deep Learning-Based Human Motion Recognition System in Sports Competition. Front Neurorobot 16:2022
28. Song Y, Taylor W, Ge Y, Usman M, Imran MA, Abbasi QH (2022) Evaluation of deep learning models in contactless human motion detection system for next generation healthcare. Scientific Reports 12:21592
29. Iqbal MJ, Iqbal MM, Ahmad I, Alassafi MO, Alfakeeh AS, Alhomoud A (2021) "Real-Time Surveillance Using Deep Learning," Security and Communication Networks.

30. Singh A, Bhatt S, Nayak V, Shah M (2023) Automation of surveillance systems using deep learning and facial recognition. Int J Syst Assur Eng Manag 14(1):236–245
31. Marks M, Jin Q, Sturman O, Ziegler Lv, Kollmorgen S, Behrens Wvd, Mante V, Bohacek J, Yanik MF (2022) Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. Nature Machine Intelligence 4:331–340
32. Aithani L, Alcaide E, Bartunov S, Cooper CD, Doré AS, Lane TJ, Maclean F, Rucktooa P, Shaw RA, Skerratt SE (2023) Advancing structural biology through breakthroughs in AI. Current Opinion in Structural Biology 80:102601
33. Nogueira TdC, Vinhal CDN, Júnior GdC, Ullmann MRD, Marques TC (2023) A reference-based model using deep learning for image captioning. Multimed Syst 29:1665–1681
34. Chun P-J, Yamane T, Maemura Y (2022) A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. Comput Aided Civ Infrastruct Eng 37(11):1387–1401
35. Kłosowski P (2018) "Deep Learning for Natural Language Processing and Language Modelling," in 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan.
36. Oyelade ON, Ezugwu AE (2020) A case-based reasoning framework for early detection and diagnosis of novel coronavirus. Inform Med Unlocked 20:100395
37. Lombo X, Oyelade O, Ezugwu AE (2022), "Crime Detection and Analysis from Social Media Messages Using Machine Learning and Natural Language Processing Technique," in ICCSA 2022: Computational Science and Its Applications – ICCSA 2022 Workshops.
38. Probierza B, Stefański P, Kozak J (2021) "Rapid detection of fake news based on machine learning methods," in 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.
39. Sanober S, Alam I, Pande S, Arslan F, Rane KP, Singh BK, Khamparia A, Shabaz M (2021) "An enhanced secure deep learning algorithm for fraud detection in wireless communication," Wireless Communications and Mobile Computing.
40. Roy A, Sun J, Mahoney R, Alonzi L, Adams S, Beling P (2018) "Deep learning detecting fraud in credit card transactions," in 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville
41. Liang Q, Zeng Y, Xu B (2020) Temporal-Sequential Learning With a Brain-Inspired Spiking Neural Network and Its Application to Musical Memory. Front Comput Neurosci 14:2020
42. Asudani DS, Nagwani NK, Singh P (2023) "Impact of word embedding models on text analytics in deep learning environment: a review," Artif Intell Rev.
43. Santosh K, Das N, Ghosh S (2021) "Chapter1: Introduction," in Deep learning models for medical imaging: a volume in primers in biomedical imaging devices and systems.
44. Subasi A, Panigrahi SS, Patil BS, Canbaz MA, Klén R (2022) "Chapter 8 - Advanced pattern recognition tools for disease diagnosis," in Intelligent Data-Centric Systems, pp 195–229.
45. Oyelade ON, Almutari MS, Ezugwu AE, Chiroma H (2022)"A generative adversarial network for synthetization of regions of interest based on digital mammograms," Sci Rep.
46. Oyelade ON, Ezugwu AE (2022) "A novel wavelet decomposition and wavelet transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram," Sci Rep.
47. Chen X, Jin Z, Wang Q, Yang W, Liao Q, Meng H (2022) Unsupervised visual feature learning based on similarity guidance. Neurocomputing 490(14):358–369
48. Mandapati S, Kadry S, Kumar RL, Sutham K, Thinnukool O (2023) Deep learning model construction for a semi-supervised classification with feature learning. Complex Intel Syst 9:3011–3021
49. Liu D-R, Li H-L, Wang D (2015) Feature selection and feature learning for high-dimensional batch reinforcement learning: A survey. Int J Autom Comput 12:229–242
50. Qian J, Zhao R, Wei J, Luo X, Xue Y (2019) Feature extraction method based on point pair hierarchical clustering. Connect Sci 32(3):223–238
51. Hu J, Li S, Hu J, Yang G (2018) A hierarchical feature extraction model for multi-label mechanical patent classification. Sustainability 10(1):219
52. Oyelade ON, Ezugwu AE (2021) A deep learning model using data augmentation for detection of architectural distortion in whole and patches of images. Biomed Signal Process Control 65:102366
53. Liu S, Deng W (2015) "Very deep convolutional neural network based image classification using small training sample size," in 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR).

54. Krizhevsky A, Sutskever I, Hinton GE (2012) "ImageNet Classification with Deep Convolutional Neural Networks," in NeurIPS 2012.
55. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE.
56. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem.
57. Chollet F (2017) "Xception: Deep Learning with Depthwise Separable Convolutions," in CVPR 2017.
58. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) "Densely connected convolutional networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
59. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) "Mobilenets: efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861.
60. Ma N, Zhang X, Zheng HT, Sun J (2018) "Shufflenet V2: practical guidelines for efficient CNN architecture design," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11218 LNCS.
61. Tan M, Le Q (2019) "EfficientNet: Rethinking model scaling for convolutional neural networks," in 36th International Conference on Machine Learning. ICML.
62. Simonyan K, Zisserman A (2015) "Very deep convolutional networks for large-scale image recognition, 3rd International Conference on Learning Representations," in ICLR 2015 - Conference Track Proceedings. 2015.
63. Anju Thomas PMH, Gopi VP (2022) "Chapter 7 - FunNet: a deep learning network for the detection of age-related macular degeneration," in Edge-of-Things in Personalized Healthcare Support Systems Cognitive Data Science in Sustainable Computing. pp 157–172
64. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) "SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <0.5MB Model Size," arXiv:1602.07360, p. 1–13
65. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) "Gradient-based learning applied to document recognition," in Proceedings of the IEEE.
66. Simonyan K, Zisserman A (2015) "Very Deep Convolutional Networks for Large-Scale Image Recognition," in ICLR 2015.
67. Shen X, Ma T, Li C, Wen Z, Zheng J (2023) High-precision automatic identification method for dicentric chromosome images using two-stage convolutional neural network. Sci Rep 13:2124
68. Cao H, Liu H, Song E, Ma G, Xu X, Jin R, Liu T, Hung C-C (2020) A Two-Stage Convolutional Neural Networks for Lung Nodule Detection. IEEE J Biomed Health Inform 24(7):2006–2015
69. Ding J, Song J, Li J, Tang J, Guo F (2022) Two-Stage Deep Neural Network via Ensemble Learning for Melanoma Classification. Front Bioeng Biotechnol 9:2022
70. Nguyen NHT, Perry S, Bone D, Le HT, Nguyen TT (2021) Two-stage convolutional neural network for road crack detection and segmentation. Expert Syst Appl 186:115718
71. Girshick R, Donahue J, Darrell T, Malik J (2014) "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR 2014.
72. Girshick R (2015) "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV).
73. Ren S, He K, Girshick R, Sun J (2015) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Advances in Neural Information Processing Systems 28 (NIPS 2015).
74. He K, Gkioxari G, Dollár P, Girshick R (2017) "Mask R-CNN," in IEEE International Conference on Computer Vision (ICCV).
75. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) "Feature Pyramid Networks for Object Detection," arXiv:1612.03144.
76. Li W, Song A (2021) "UFO RPN: a region proposal network for ultra fast object detection," in AI 2021: Advances in Artificial Intelligence.
77. Cai Z, Vasconcelos N (2021) Cascade R-CNN: High Quality Object Detection and Instance Segmentation. IEEE Trans Pattern Anal Mach Intell 43(5):1483–1498
78. Chen Y, Zhao P, Chen J (2021) "CarfRCNN: a two-stage effective model for instance segmentation," in CIMIA 2021.
79. Diwan T, Anirudh G, Tembhurne JV (2023) Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl. 82:9243–9275

80. Soviany P, Ionescu RT (2018) "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," in 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC).
81. Redmon J, Farhadi A (2017) "YOLO9000: better, faster, stronger," in In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017.
82. Redmon J, Divvala S, Girshick R, Farhadi A (2016) "You only look once: unified, real-time object detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
83. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision.
84. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) "CenterNet: Keypoint Triplets for Object Detection," in IEEE/CVF International Conference on Computer Vision (ICCV).
85. Yang J, Jiang J, Fang Y, Sun J (2021) LADNet: an ultra-lightweight and efficient dilated residual network with light-attention module. IEEE Access 9:41373–41382
86. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) "SSD: single shot multibox detector," in European Conf Comput Vis 2016.
87. Fu CY, Liu W, Ranga A, Tyagi A, Berg AC (2017) "Dssd: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659.
88. Bochkovskiy A, Wang CY, Liao HY (2020) "YOLOv4: optimal speed and accuracy of object detection.," arXiv preprint arXiv:2004.10934.
89. Liu Y, Cheng D, Zhang D, Xu S, Han J (2024) "Capsule networks with residual pose routing," IEEE Trans Neural Netw Learn Syst.
90. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y (2011) "Contracting auto-encoders." in In Proceedings of the International Conference on Machine Learning (ICML), Bellevue, WA, USA.
91. Zilvan V, Ramdan A, Heryana A, Krisnandi D, Suryawati E, Yuwana RS, Budiarianto R, Kusumo S, Pardede HF (2022) Convolutional variational autoencoder-based feature learning for automatic tea clone recognition. J King Saud Univ Comput Inf Sci 34(6):3332–3342
92. Arul VH (2021) "5 - Deep learning methods for data classification," in Artificial Intelligence in Data Mining Theories and Applications. pp. 87–108
93. Vincent P, Larochelle H, Bengio Y, Manzagol P (2008) "Extracting and composing robust features with denoising autoencoders." In: Helsinki, Finland, Proceedings of the 25th International Conference on Machine Learning.
94. Kingma D, Welling M (2013) "Auto-encoding variational bayes.," arXiv 2013, arXiv:1312.6114.
95. Ng A (2011) "Sparse autoencoder," in CS294A Lect. Notes 2011.
96. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) "Adversarial autoencoders.," arXiv 2015, arXiv:1511.05644.
97. Chung Y, Wu C, Shen C, Lee H, Lee L (2016) "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder.," arXiv 2016, arXiv:1603.00982.
98. Oyelade ON, Ezugwu AE (2021) "ArchGAN: A generative adversarial network for architectural distortion abnormalities in digital mammograms," in 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town.
99. Oyelade ON, Ezugwu AE (2023) "EOSA-GAN: feature enriched latent space optimized adversarial networks for synthesization of histopathology images using ebola optimization search algorithm," Biomed Signal Process Control.
100. Zheng Y, Gindra RH, Green EJ, Burks EJ, Betke M, Beane JE, Kolachalama VB (2022) A Graph-Transformer for Whole Slide Image Classification. IEEE Trans Med Imaging 41(11):3003–3015
101. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) "An image is worth 16x16 words: transformers for image recognition at scale," arXiv:2010.11929.
102. Willemink MJ, Roth HR, Sandfort V (2022) Toward foundational deep learning models for medical imaging in the new era of transformer networks. Radiol Artif Intell 4:6
103. Uparkar O, Bharti J, Pateriya R, Gupta RK, Sharma A (2023) Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images. Procedia Comput Sci 218(2023):2338–2349
104. Li X, Zhao B, Lu X (2017) "MAM-RNN: Multi-level Attention Model Based RNN for Video Captioning," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).
105. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) "Attention Is All You Need," in NeurIPS 2017.

106. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2022) "Swin transformer: hierarchical vision transformer using shifted windows," in In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
107. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) "End-to-end object detection with transformers," arXiv:2005.12872.
108. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2021) "An image is worth 16x16 words: transformers for image recognition at scale.," in In Proceedings of the ICLR.
109. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep Learning for Generic Object Detection: A Survey. Int J Comput Vis 128:261–318
110. Zhang HHX (2019) Recent progresses on object detection: a brief review. Multimed Tools Appl 78:27809–27847
111. Kaur JSW (2022) Tools, techniques, datasets and application areas for object detection in an image: a review. Multimed Tools Appl 81:38297–38351. https://doi.org/10.1007/s11042-022-13153-y
112. Han J, Yang Y (2021) L-Net: lightweight and fast object detector-based ShuffleNetV2. J Real-Time Image Proc 18:2527–2538
113. Junayed MS, Islam MB, Imani H, Aydin T (2022) PDS-Net: A novel point and depth-wise separable convolution for real-time object detection. Int J Multimed Info Retr 11:171–188
114. Balakrishna S, Mustapha A (2022) "Progress in multi-object detection models: a comprehensive survey." Multimed Tools Appl.
115. Ding P, Qian H, Chu S (2022) SlimYOLOv4: lightweight object detector based on YOLOv4. J Real-Time Image Proc 19:487–498
116. Hu L, Zhang Y, Zhao Y, Wu T, Li Y (2022) Micro-YOLO+: Searching Optimal Methods for Compressing Object Detection Model Based on Speed, Size, Cost, and Accuracy. SN Comput Sci 3:391
117. Wei L, Cui W, Hu Z, Sun H, Hou S (2021) A single-shot multi-level feature reused neural network for object detection. Vis Comput 37:133–142
118. Anandh N, Gopinath MP (2023) "Resnet features and optimization enabled deep learning for indoor object detection and object recognition," Cybernet Syst.
119. Wang N, Gao Y, Chen H, Wang P, Tian Z, Shen C, Zhang Y (2021) NAS-FCOS: Efficient Search for Object Detection Architectures. Int J Comput Vis 129:3299–3312
120. Dharmik RC, Chavhan SY, Sathe SR (2022) Deep learning based missing object detection and person identification: an application for smart CCTV. 3C Tecnología. Glosas de Innovación Aplicadas a La Pyme 11:51–57
121. Alzahrani N, Al-Baity HH (2023) Object recognition system for the visually impaired: a deep learning approach using arabic annotation. Electronics 12(3):541
122. Ke TW, Kim DJ, Yu SX, Gou L, Ren L (2022) "Contextual visual feature learning for zero-shot recognition of human-object interactions," in 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
123. Xiao Y, Tian Z, Yu J, Zhang Y et al (2020) A review of object detection based on deep learning. Multimed Tools Appl 79:23729–23791
124. Dhillon A, Verma G (2020) Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 9:85–112
125. Diwan T, Anirudh G, Tembhurne J (2023) Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl 82:9243–9275
126. Srivastava S, Divekar AV, Anilkumar C, Naik I, Kulkarni V, Pattabiraman V (2021) Comparative analysis of deep learning image detection algorithms. J Big Data 8:66
127. Wang X (2016) Deep Learning in Object Recognition, Detection, and Segmentation.
128. Zhao ZQ, Zheng P, Xu S-t, Wu X (2017) "Object detection with deep learning: a review," IEEE Transactions on Neural Networks and Learning Systems.
129. Liu Y, Dong X, Zhang D, Xu S (2024) Deep unsupervised part-whole relational visual saliency. Neurocomputing 563:1
130. Liu Y, Han J, Zhang Q, Shan C (2019) Deep salient object detection with contextual information guidance. IEEE Trans Image Process 29:360–374
131. Liu Y, Zhou L, Wu G, Xu S, Han J (2023) "TCGNet: type-correlation guidance for salient object detection." IEEE Trans Intell Transp Syst.
132. Iqbal M, Sameem MSI, Naqvi N, Kanwal S, Ye Z (2019) A deep learning approach for face recognition based on angularly discriminative features. Pattern Recogn Lett 129:414–419
133. Yu C, Pei H (2021) Face recognition framework based on effective computing and adversarial neural network and its implementation in machine vision for social robots. Comput Elect Eng 92:107128

134. Harikrishnan J, Sudarsan A, Sadashiv A, Ajai RA (2019) "Vision-Face Recognition Attendance Monitoring System for Surveillance using Deep Learning Technology and Computer Vision," in 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN).
135. Mahouachi D, Akhloufi MA (2023) Recent advances in infrared face analysis and recognition with deep learning. AI 4:199–233
136. Rajagopalan A (2022) Real-Time Deep Learning-Based Face Recognition System. Culminat Proj Elect Eng 8:1–85
137. Umer S, Dhara BC, Chanda B (2019) Face Recognition Using Fusion of Feature Learning Techniques. Measurement 146:4
138. Suganthi ST, Ayoobkhan MU, Bacanin N, Venkatachalam K, Štěpán H, Pavel T (2022) Deep learning model for deep fake face recognition and detection. PeerJ Comput Sci. 8:1–20
139. Baek S, Song M, Jang J, Kim G, Paik S-B (2021) Face detection in untrained deep neural networks. Nat Commun 12(7328):1–15
140. Wang J, Cao R, Brandmeir NJ, Li X, Wang S (2022) Face identity coding in the deep neural network and primate brain. Commun Biol 5(611):1–16
141. Schnell AE, Vinken K, de Beeck HO (2023) The importance of contrast features in rat vision. Sci Rep 13(459):1–13
142. Mamieva D, Abdusalomov AB, Mukhiddinov M, Whangbo TK (2023) Improved face detection method via learning small faces on hard images based on a deep learning approach. Sensors 23(502):1–16
143. F. R. a. I. u. D. L. Approach (2020) "KH Teoh; RC Ismail; SZM Naziri; R Hussin; MNM Isa; MSSM Basir," in 5th International Conference on Electronic Design (ICED) 2020.
144. Hangaragi S, Singh T, Neelima N (2023) "Face detection and recognition using face mesh and deep neural networ," in International Conference on Machine Learning and Data Engineering.
145. Shamrat FJM, Jubair MA, Billah MM, Chakraborty S, Alauddin M, Ranjan R (2021) "A deep learning approach for face detection using max pooling," in Proceedings of the Fifth International Conference on Trends in Electronics and Informatics (ICOEI). IEEE Xplore Part Number:CFP21J32-ART; ISBN:978–1–6654–1571–2.
146. Mokalla SR (2020) "Deep learning based face detection and recognition in mwir and visible bands," Graduate Theses, Dissertations, and Problem Reports, West Virginia University.
147. Guo Z, Huang Y, Hu X, Wei H, Zhao B (2021) A survey on deep learning based approaches for scene understanding in autonomous driving. Electronics 10(471):1–29
148. Husain F, Dellen B, Torras C (2016) "Scene understanding using deep learning," in Handbook on Neural Computation, pp. 1–11.
149. Minh TL, Shimizu N, Miyazaki T, Shinoda K (2018) "Deep learning based multi-modal addressee recognition in visual scenes with utterances," in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).
150. Sridharan M, Mota T (2020) "Commonsense reasoning to guide deep learning for scene understanding," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Sister Conferences Best Papers Track.
151. Masood S, Ahsan U, Munawwar F, Rizvi DR, Ahmed M (2020) "Scene recognition from image using convolutional neural network," in International Conference on Computational Intelligence and Data Science (ICCIDS 2019).
152. Vandenhende S (2022) "Multi-task learning for visual scene understanding," Doctoral thesis, KU Leuven – Faculty of Engineering Science.
153. Anaya-Isaza A, Mera-Jiménez L, Zequera-Diaz M (2021) An overview of deep learning in medical imaging. Inform Med Unlocked 26:100723
154. Oyelade O, Ezugwu A, Chiroma H (2021) CovFrameNet: an enhanced deep learning framework for COVID-19 detection. Ieee Access 9:77905–77919
155. Ezugwu A, Hashem IAT, Oyelade ON, Almutari M, Al-Garadi M, Abdullahi I, Otegbeye O, Shukla A, Chiroma H (2021) "A machine learning solution framework for combatting COVID-19 in smart cities from multiple dimensions," BioMed Res Int.
156. Taiwo O, Ezugwu AE, Oyelade ON, Almutairi MS (2022) "Enhanced intelligent smart home control and security system based on deep learning model," Wirel Commun Mob Com. pp. 1–22.
157. Oyelade ON, Ezugwu AE, Venter HS, Mirjalili S, Gandomi AH (2022) "Abnormality classification and localization using dual-branch whole-region-based CNN model with histopathological images," Comp Biol Med.

158. Mohamed TIA, Oyelade ON, Ezugwu AE (2023) "Ebola Optimization Search Algorithm (EOSA) with application to deep learning model for early detection and classification of Lung Cancer on CT images," PLoS One.
159. Oyelade ON, Ezugwu AE (2021) "A bioinspired neural architecture search based convolutional neural network for breast cancer detection using histopathology images. Sci Rep Nat 11:19940
160. Oyelade ON, Ezugwu AE (2022) Immunity-based Ebola optimization search algorithm for minimization of feature extraction with reduction in digital mammography using CNN models. Sci Rep 12:17916
161. Chen M, Shi X, Zhang Y, Wu D, Guizani M (2018) "Deep feature learning for medical image analysis with convolutional autoencoder neural network," IEEE Trans Big Data.
162. He X, Yang X, Zhang S, Zhao J, Zhang Y, Xing E, Xie P (2020) "Sample-efficient deep learning for COVID-19 diagnosis based on CT Scans," IEEE Trans Med Imaging. pp. 1–10.
163. Brimaa Y, Atemkeng M (2022) "What do Deep Neural Networks Learn in Medical Images?," arXiv:2208.00953v1.
164. Gao Z, Lou L, Wang M, Sun Z, Chen X, Zhang X, Pan Z, Hao H, Zhang Y, Quan S, Yin S, Lin C, Shen X (2022) Application of machine learning in intelligent medical image diagnosis and construction of intelligent service process. Comput Intell Neurosci 2022:1–14
165. Ursuleanu TF, Luca AR, Gheorghe L, Grigorovici R, Iancu S, Hlusneac M, Preda C, Grigorovici A (2021) Deep learning application for analyzing of constituents and their correlations in the interpretations of medical images. Diagnostics 11(1373):1–48
166. Jiang Y, Hsiao T (2021) "Deep Learning in Perception of Autonomous Vehicles," in Proceedings of the 2021 International Conference on Public Art and Human Development (ICPAHD 2021).
167. Fujiyoshi H, Hirakawa T, Yamashita T (2019) Deep learning-based image recognition for autonomous driving. IATSS Res 43:244–252
168. Rezaei M, Shahidi M (2020) Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: a review. Intell Based Med 3:4
169. Jebamikyous H-H, Kashef R (2022) Autonomous Vehicles Perception (AVP) Using Deep Learning: Modeling, Assessment, and Challenges. IEEE Access 10(2022):10523–10535
170. Tampuu A, Matiisen T, Semikin M, Fishman D, Muhammad N (2021) "A survey of end-to-end driving: architectures and training methods," IEEE Transactions on Neural Networks and Learning System: arXiv:2003.06404v2, pp. 1–23.
171. Silva VAT (2020) "End to end self-driving using convolutional neural networks," University of Western Australia.
172. Rezapour M, Ksaibati K (2021) Convolutional neural network for roadside barriers detection: transfer learning versus non-transfer learning. Signals 2:72–86
173. Bag S (2017) "Deep learning localization for self-driving cars," Rochester Institute of Technology.
174. Sellat Q, Bisoy S, Priyadarshini R, Vidyarthi A, Kautish S, Barik RK (2022) Intelligent Semantic Segmentation for Self-Driving Vehicles Using Deep Learning. Comput Intell Neurosci 2022:1–10
175. Gupta M, Upadhyay V, Kumar P, Al-Turjman F (2021) "Deep learning implementation of autonomous driving using ensemble-m in simulated environment," Res Sq. pp. 1–19.
176. Kim J, Rohrbach A, Akata Z, Moon S, Misu T, Chen YT, Darrell T, Canny J (2021) Applied AI Letters. pp. 1-13.
177. Fan R, Wang L, Bocus MJ, Pitas I (2023) Computer stereo vision for autonomous driving: theory and algorithms. Recent Advances in Computer Vision Applications Using Parallel Processing, Studies in Computational Intelligence. vol. 1073 pp. 41-70
178. Rezapour M, Ksaibati K (2022) Cost–benefit analysis of traffic barrier geometric optimization, a hurdle machine learning-based technique. Engineering Reports 4(e12435):1–13
179. Sainju AM, Jiang Z (2020) Mapping road safety features from streetview imagery: a deep learning approach. ACM/IMS Trans Data Sci 1(3):1–20
180. Muresan M, Pan G, Fu L (2021) Multi-intersection control with deep reinforcement learning and ring-and-barrier controllers. Transport Res Record 2675(4):308–319
181. Seo S, Chen D, Kim K, Kang K, Koo D, Chae M, Park HK (2022) "Temporary traffic control device detection for road construction projects using deep learning application," in Conference.
182. Azimjonov J, Özmen A (2022) Vision-based vehicle tracking on highway traffic using bounding-box features to extract statistical information. Comput Electr Eng 97:1–13
183. Kim T-G, Yun B-J, Kim T-H, Lee J-Y, Park K-H, Jeong Y, Kim HD (2021) Recognition of vehicle license plates based on image processing. Appl Sci 11(6292):1–12
184. Anisha PR (2021) Automatic license-plate recognition using image segmentation & processing. Turk J Com Math Educ 12:3459–3472

185. Ravi Kiran Varma P, Ganta S, Hari Krishna B, Svsrk P (2020) "A novel method for Indian vehicle registration number plate detection and recognition using image processing techniques," in International Conference on Computational Intelligence and Data Science (ICCIDS 2019).
186. Khan K, Imran A, Rehman HZU, Fazil A, Zakwan M, Mahmood Z (2021) Performance enhancement method for multiple license plate recognition in challenging environments. EURASIP J Image Video Proces 30:2021
187. Islam KT, Raj RG, Islam SMS, Wijewickrema S, Hossain S, Razmovski T, O'Leary S (2020) A vision-based machine learning method for barrier access control using vehicle license plate authentication. Sensors 20(3578):1–18
188. Surekha P, Gurudath P, Prithvi R, Ananth VR (2018) "Automatic license plate recognition using image processing and neural network," ICTACT J Image Video Process 2018.
189. Pirgazi J, Kallehbasti MMP, Sorkhi AG (2022) "An end-to-end deep learning approach for plate recognition in intelligent transportation systems," Wirel Commun Mob Com. pp. 1–13.
190. Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW (2018) Action recognition in video sequences using deep bi-directional LSTM With CNN features. IEEE Access 6(2018):1155–1166
191. Bardes A, Ponce J, LeCun Y (2022) "VICRegL: Self-supervised learning of local visual features," in 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
192. Jing L, Tian Y (2017) "Self-supervised visual feature learning with deep neural networks: a survey," IEEE, pp. 1–8.
193. Rombach R, Esser P, Blattmann A, Ommer B (2022) "Invertible neural networks for understanding semantics of invariances of CNN representations," in Deep neural networks and data for automated driving:robustness, uncertainty quantification, and insights towards safety, Cham, Switzerland, Springer Nature.
194. Chen X, Jin Z, Wang Q, Yang W, Liao Q, Meng H (2022) Unsupervised visual feature learning based on similarity guidance. Neurocomputing 490:358–369
195. Sharp M, Ak R, Hedberg T (2018) A survey of the advancing use and development of machine learning in smart manufacturing. J Manuf Syst 48:170–179
196. Carvalho TP, Soares FA, Vita R, Francisco RD, Basto JP, Alcalá SG (2019) A systematic literature review of machine learning methods applied to predictive maintenance. Comput Ind Eng 137:106024
197. Lei Y, Yang B, Jiang X, Jia F, Li N, Nandi AK (2020) Applications of machine learning to machine fault diagnosis: a review and roadmap. Mech Syst Signal Process 138:106587
198. Rice L, Wong E, Kolter Z (2020) Overfitting in adversarially robust deep learning. Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research vol. 119, pp. 8093–8104 Available from https://proceedings.mlr.press/v119/rice20a.html.
199. Dimiccoli M, Marín J, Thomaz E (2018) Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. Proc ACM Interact Mob Wearable Ubiquitous Technol 1(4):1–18
200. Saranya A, Subhashini R (2023) A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Dec Analyt J 7:100230
201. Lei C (2021) Unsupervised Learning: Deep Generative Model. In: Deep Learning and Practice with MindSpore. Cognitive Intelligence and Robotics. Springer, Singapore. https://doi.org/10.1007/978-981-16-2233-5_9
202. Zheda M, Ruiwen L, Jihwan J, David Q, Hyunwoo K (2022) Online continual learning in image classification: an empirical survey. J Neurocomput 469:28–51. https://doi.org/10.1016/j.neucom.2021.10.021
203. Ting C, Simon K, Mohammad N, Geoffrey H (2020) "A Simple Framework for Contrastive Learning of Visual Representations. "Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria.
204. Patricia N, Caputo B (2014) "Learning to learn, from transfer learning to domain adaptation: a unifying perspective," IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, vol. 2014, pp. 1442-1449. https://doi.org/10.1109/CVPR.2014.187
205. Guangting W, Yizhou Z, Chong L, Wenxuan X, Wenjun Z, Zhiwei X (2021) "Unsupervised visual representation learning by tracking patches in video" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2021, pp. 2563–2572
206. Li D, Wang R, Chen P, Xie C, Zhou Q, Jia X (2022) Visual feature learning on video object and human action detection: a systematic review. Micromachines 13:72. https://doi.org/10.3390/mi13010072
207. Yue L, Aixi Z, Zhiyuan C, Tianrui H, Si L (2022) Progressive language-customized visual feature learning for one-stage visual grounding. IEEE Transac Image Process 31:4266–4277

208. Junaid SB, Imam AA, Balogun AO, De Silva LC, Surakat YA, Kumar G, Abdulkarim M, Shuaibu AN, Garba A (2022) Recent advancements in emerging technologies for healthcare management systems: a survey. Healthcare 10:1940. https://doi.org/10.3390/healthcare10101940
209. Gulsum A, Bo S (2022) A survey of visual analytics for Explainable Artificial Intelligence methods. Compute Graph 102(2022):502–520. https://doi.org/10.1016/j.cag.2021.09.002
210. Minh D, Wang HX, Li YF et al (2022) Explainable artificial intelligence: a comprehensive review Artif. Intell Rev 55(2022):3503–3568. https://doi.org/10.1007/s10462-021-10088
211. Gopalan R, Ruonan L, Chellappa R (2011) "Domain adaptation for object recognition: an unsupervised approach," International Conference on Computer Vision, Barcelona, Spain, vol. 2011, pp. 999–1006 https://doi.org/10.1109/ICCV.2011.6126344.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mohammed Abdullahi[1] **· Olaide Nathaniel Oyelade**[5] **· Armand Florentin Donfack Kana**[1] **· Mustapha Aminu Bagiwa**[1] **· Fatimah Binta Abdullahi**[1] **· Sahalu Balarabe Junaidu**[1] **· Ibrahim Iliyasu**[2] **· Ajayi Ore-ofe**[3] **· Haruna Chiroma**[4]

✉ Mohammed Abdullahi
abdullahilwafu@abu.edu.ng

[1] Department of Computer Science, Ahmadu Bello University Zaria, Zaria, Nigeria

[2] Department of Mechanical Engineering, Ahmadu Bello University, Zaria, Nigeria

[3] Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria

[4] College of Computer Science and Engineering, University of Hafr Al Batin, 31991 Hafar Al-Batin, Saudi Arabia

[5] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK