IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 January 2024, accepted 7 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376441

SURVEY

A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends

ABOLFAZL YOUNESI[1], MOHSEN ANSARI[1], MOHAMMADAMIN FAZLI[1], ALIREZA EJLALI[1], MUHAMMAD SHAFIQUE[2], (Senior Member, IEEE), AND JÖRG HENKEL[3], (Fellow, IEEE)

[1]Department of Computer Science and Engineering, Sharif University of Technology, Tehran 11365-11155, Iran
[2]eBrainLab, Division of Engineering, New York University (NYU) Abu Dhabi, Abu Dhabi, United Arab Emirates
[3]Department of Computer Science, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany

Corresponding author: Mohsen Ansari (ansari@sharif.edu)

This work was supported in part by New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics (CAIR) funded by Tamkeen under the NYUAD Research Institute under Award CG010, and in part by the Project "eDLAuto: An Automated Framework for Energy-Efficient Embedded Deep Learning in Autonomous Systems" funded by the NYUAD Research Enhancement Fund (REF).

ABSTRACT In today's digital age, Convolutional Neural Networks (CNNs), a subset of Deep Learning (DL), are widely used for various computer vision tasks such as image classification, object detection, and image segmentation. There are numerous types of CNNs designed to meet specific needs and requirements, including 1D, 2D, and 3D CNNs, as well as dilated, grouped, attention, depthwise convolutions, and NAS, among others. Each type of CNN has its unique structure and characteristics, making it suitable for specific tasks. It's crucial to gain a thorough understanding and perform a comparative analysis of these different CNN types to understand their strengths and weaknesses. Furthermore, studying the performance, limitations, and practical applications of each type of CNN can aid in the development of new and improved architectures in the future. We also dive into the platforms and frameworks that researchers utilize for their research or development from various perspectives. Additionally, we explore the main research fields of CNN like 6D vision, generative models, and meta-learning. This survey paper provides a comprehensive examination and comparison of various CNN architectures, highlighting their architectural differences and emphasizing their respective advantages, disadvantages, applications, challenges, and future trends.

INDEX TERMS Deep learning, DNN, CNN, machine learning, vision transformers, GAN, attention, computer vision, LLM, large language model, transformer, dilated convolution, depthwise, NAS, NAT, object detection, 6D vision, vision language model.

I. INTRODUCTION

IN today's world, as technology continues to evolve, deep learning (DL) has become an integral part of our lives [1]. From voice assistants like Siri and Alexa to personalized recommendations on social media platforms, DL algorithms are constantly working behind the scenes to understand our preferences and make our lives easier [2]. With advancements in technology, DL is also being used in various fields such as healthcare, finance, and transportation, revolutionizing the way we approach these industries [3], [4], [5]. As research and development in the field of DL continue to progress, even more innovative applications that will further enhance our daily lives can be expected. DL has ushered in a transformative era in artificial intelligence, empowering machines to assimilate vast datasets and make informed predictions [6], [8]. The development of CNNs has received attention among deep learning's significant advancements. Their impact has been felt in some areas, including generative AI, examining medical images, identifying objects [9], and

The associate editor coordinating the review of this manuscript and approving it for publication was Tyson Brooks.

finding anomalies [10]. CNNs, constituting a feedforward neural network, integrate convolution operations into their architecture [7], [11]. These operations enable CNNs to adeptly capture intricate spatial and hierarchical patterns, rendering them exceptionally well-suited for image analysis tasks [12].

However, CNNs are often burdened by their computational complexity during training and deployment, particularly when operating on resource-constrained devices like mobile phones and wearables [12], [13].

Two principal avenues have emerged to reinforce the energy efficiency of CNNs: Employing Lightweight CNN Architectures: These architectures are deliberately engineered to achieve computational efficiency without compromising accuracy. For instance, the MobileNet family of CNNs has been meticulously tailored for mobile devices and has demonstrated state-of-the-art accuracy across various image classification Applications [13].

Embracing Compression Techniques: These methods facilitate the reduction of CNN model size and consequently diminish the volume of data transfers between devices. A noteworthy example is the TensorFlow Lite framework, which offers a suite of compression techniques tailored for compressing CNN models for mobile devices [14].

The fusion of lightweight CNN architectures and compression techniques yields a substantial boost in the energy efficiency of CNNs. Training and deploying CNNs on resource-constrained devices become feasible, thereby unlocking novel opportunities for employing CNNs in diverse applications like healthcare, agriculture, and environmental monitoring [12], [16].

How different convolutional techniques cater to various AI applications. Convolutions play a fundamental role in contemporary DL architectures and are especially crucial when dealing with data organized in grid-like structures, such as images, audio signals, and sequential data [23]. The convolutional operation entails moving a small filter, also known as a kernel, across the input data, performing element-wise multiplications and aggregations. This process extracts essential features from the input data [24]. The main significance of convolutions lies in their capability to efficiently capture local patterns and spatial relationships within the data. This localization property makes convolutions highly suitable for tasks like image recognition, as objects can be identified based on their local structures. Additionally, convolutions introduce parameter sharing, which results in a significant reduction in the number of trainable parameters, leading to more efficient and scalable models [25].

A. EXISTING SURVEYS

Previous survey papers on CNN architectures such as [118] and [120] provided good overviews of popular architectures from a certain period. However, they lacked a clear Research question and objective, evaluation, and challenges based on their design patterns. They mostly discussed architectures chronologically.

Earlier surveys like [119] and [120] focused on explaining core CNN components and popular architectures up to a certain year. they also lacked research questions and objectives, analysis of datasets, and special types of taxonomy that were not considered complete overviews like large vision models, and large language models, and lack of multipoint of view for challenges.

Previous work discussed the challenges in some specific concepts and applications of CNNs but did not extensively cover the intrinsic taxonomy present in newer CNN architectures. So this caused us to write a survey paper that aims to address the gaps in previous work by proposing a taxonomy to clearly classify CNN architectures based on their intrinsic design patterns rather than release year.

We focus on architectural innovations from 2012 onwards and discuss the recent developments in greater depth than earlier surveys. Discussing the latest trends and challenges provides an updated perspective for researchers.

This comprehensive survey of CNN's history, taxonomy, applications, and challenges is needed to accelerate research progress in this domain further.

In this paper, the key questions we seek to address include:

- How do state-of-the-art CNN models like ResNet, Inception, and MobileNet perform on the target hardware compared to constrained baselines? What are the impacts on accuracy, latency, and memory usage?
- What techniques like pruning, quantization, distillation, and architecture design can help reduce the model size and computational complexity the most while retaining prediction quality?
- How do multi-stage optimization approaches that combine different techniques compare to single methods? Can we achieve better trade-offs between accuracy, latency, and memory?
- For a target application like embedded vision, what are the best practices for benchmarking, tuning, and deploying optimized CNN models considering their unique constraints and specifications?
- Which pruning and quantization techniques work best for our target application and hardware? How does this compare to baselines?

Our overview makes several key contributions to the DL and CV communities:

- **Analyzing multiple types of existing CNNs:** The survey provides a comprehensive and detailed analysis of various DL models and algorithms used in CV Applications.
- **Comparing the CNN models with various parameters and architectures:** The overview offers insights into the performance and efficiency trade-offs.
- **Identifying the strengths and weaknesses of different CNN models:** Aiding researchers in selecting the most suitable model for their specific applications.
- **The overview highlights the challenges and future directions** for further improvement in the fields of DL and computer vision.

TABLE 1. Comparison of existing surveys; +* means conditional cosideration.

Ref.	Year	No. of included studies	Research questions and Objective	Taxonomy	Datasets	Challanges	Comparison of Simulators	Evaluation
[117]	2023	210	-	+*	-	-	+	-
[118]	2021	343	-	+	+	+	+	-
[119]	2022	202	-	+	-	+	-	+
[120]	2020	243	-	+*	+	+*	-	-
Our survey	2024	465	+	+	+	+	+	+

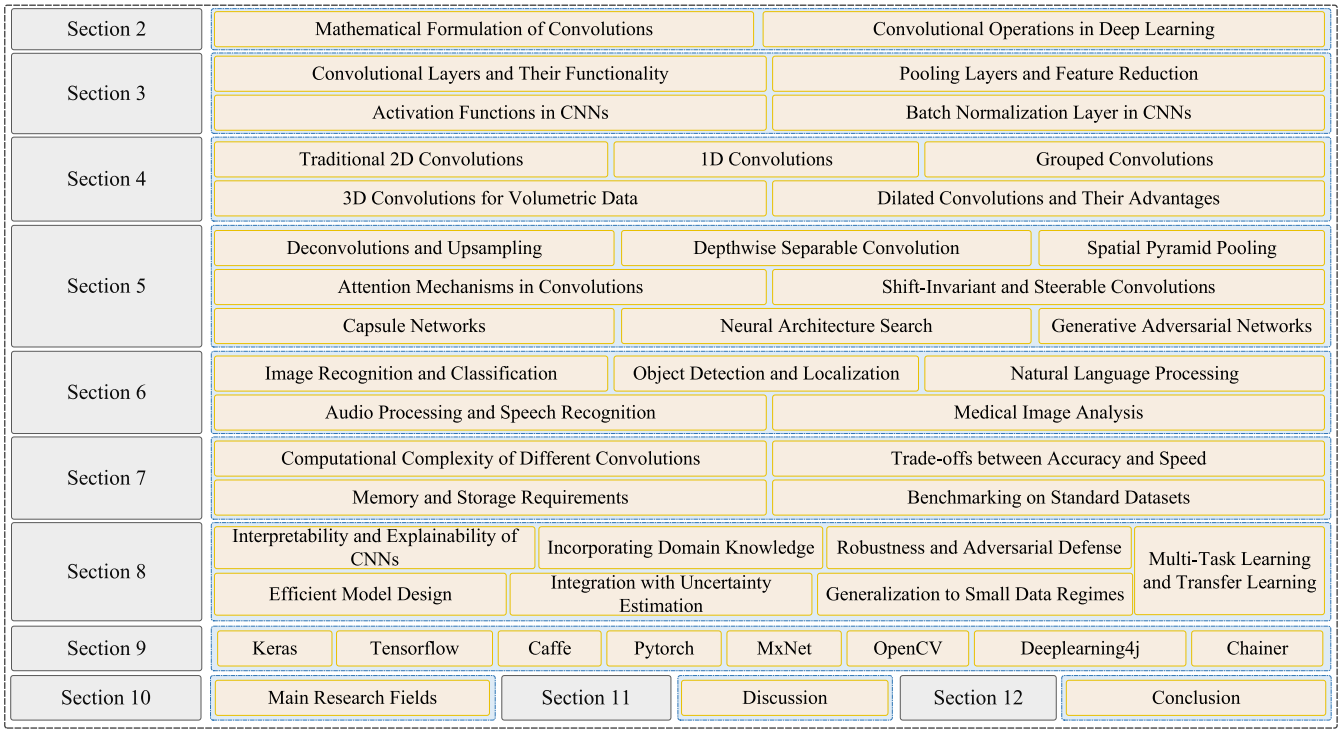

FIGURE 1. Represents the section-by-section structure of the paper that provides a clear and organized framework for presenting the research findings.

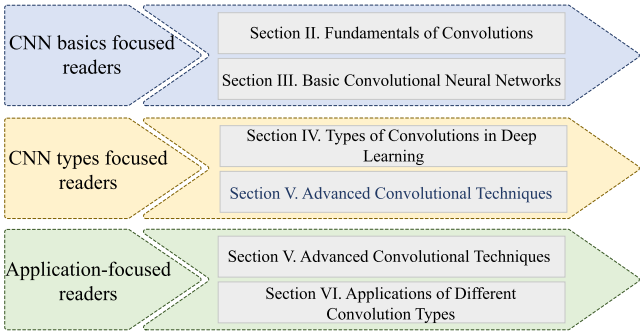

FIGURE 2. A text-based visual reading map that helps individuals navigate and comprehend the paper.

- **Exploring the trends in neural network architecture:** This emphasizes the practical application and exciting nature of the advancements.
- **Comprehensive overview of the Main research fields:** This covers the primary fields of research that are actively pursued by researchers.

The rest of our review paper follows (See Fig. 1): Section II of the paper will delve into the fundamentals of convolutions, elucidating their mathematical formulation, operational mechanics, and the role they play in the architecture of neural networks. Section III describes the basic parts of CNNs. In Section IV, The exploration will cover 2D convolutions, 1D convolutions for sequential data, and 3D convolutions for volumetric data. Section V of the research paper will investigate advanced convolutional techniques that have emerged in recent years. This will encompass topics such as transposed convolutions for upsampling, depthwise separable convolutions for efficiency, spatial pyramid pooling, and attention mechanisms within convolutions. Section VI of the paper will highlight the real-world applications of different convolution types, showcasing their utility in image recognition, object detection, NLP, audio processing, and medical image analysis. In section VII we discuss future trends and some open questions about CNNs. Section VIII is about the performance consideration of CNNs. In Section IX, we are going to talk about the platforms that are mostly used by researchers and developers, and in Section X about

research fields that are popular or trending, then we have a discussion in Section XI. By the end of this research in Section VIII, readers will gain a profound understanding of the importance of convolutions in DL and Fig. 2 represents a reader map to visualize the flow of information within a text. It shows the connections between various sections, assisting readers in comprehending the overall structure of their preferred section following their needs.

II. FUNDAMENTALS OF CONVOLUTIONS

Convolutions form the foundation of crucial mathematical operations used to process data structured in grids, such as images, videos, and time series data [26]. Originally used in signal processing, convolutions were used for analyzing and manipulating signals [27]. In deep learning, convolutions serve as powerful feature extractors, enabling neural networks to efficiently learn from raw data [26], [27]. The essence of a convolution involves the sliding of a small filter, commonly known as a kernel, over the input data. At each position of this sliding operation, the kernel performs element-wise multiplication with the corresponding input values [28]. Through this process, local patterns and relationships within the data are captured, enabling the model to acquire essential features like edges, textures, and shapes.

A. MATHEMATICAL FORMULATION OF CONVOLUTIONS

Mathematically, a 2D convolution between an input matrix (often representing an image) and a kernel can be represented as follows:

$$\text{Output}(i, j) = \sum_{(x,y)} \text{Input}(x, y) \cdot \text{Kernel}(i - x, j - y) \quad (1)$$

Here, Output denotes the resulting feature map, and Input represents the input matrix. The kernel, usually a small square matrix, defines the convolutional filter's weights. The convolution operation is performed by sliding the kernel over the input matrix, and at each position, the element-wise multiplication and summation are computed as described in the formula [29]. For 1D convolutions, the mathematical formulation is similar, with the kernel sliding along a one-dimensional sequence, such as a time series or text data [30].

B. CONVOLUTIONAL OPERATIONS IN DL

Convolutional operations form the core of CNNs, a highly prominent class of DL models widely utilized for various CV applications. Within a CNN, convolutions are typically integrated into specific layers referred to as convolutional layers [31]. These layers are composed of multiple filters, each responsible for detecting distinct patterns in the input data [138], [139], [140], [141], [142], [143], [144], [145]. During the training phase, the model goes through the process of backpropagation and gradient descent to learn the optimal weights of the convolutional filters. This enables the model to automatically discern meaningful patterns within the data. Moreover, CNN architectures (See Fig. 3 and Fig. 4) often incorporate pooling layers following the convolutional layers. As a result of pooling layers, feature maps generated by convolutions are downsampled, reducing computational complexity. Common pooling techniques include max-pooling and average pooling, which we will discuss about them in Section III.B.

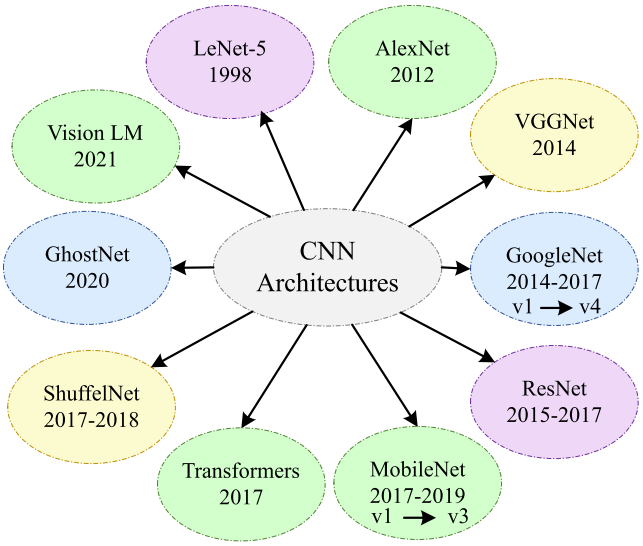

FIGURE 3. A graphical representation of CNN architectures from 1998 to 2023.

C. WAVELETS

Wavelets are an important mathematical tool that has numerous applications in fields such as signal processing and computer graphics. At their core, wavelets rely on convolution to analyze functions or continuous-time signals [104]. By convolving the target function with wavelet basis functions at different scales, wavelets are capable of representing data with varying degrees of resolution [109].

Wavelet analysis uses small waves, called wavelets, as basis functions instead of the sine and cosine functions used in Fourier analysis [105]. Wavelets have the advantage of analyzing properties of data locally in time and frequency instead of globally. This makes them well-suited for tasks such as edge detection, noise removal, and texture identification. The wavelet basis can also be adapted to the input signal or data being analyzed [105], [106].

CNNs naturally lend themselves to wavelet analysis due to their intrinsic use of convolution operations [107], [108]. During training, the convolutional filters within CNNs can learn wavelet-like basis functions tailored to meaningfully represent the given input data distribution at multiple resolutions. By adopting the wavelet bases through gradient descent and backpropagation, CNNs gain an efficient multi-scale representation of patterns in the data [108], [109].

A key characteristic of wavelets is their ability to decompose a signal into different frequency components, with high frequencies corresponding to detailed information and low frequencies corresponding to overall trends [108]. A single-level wavelet decomposition breaks down the

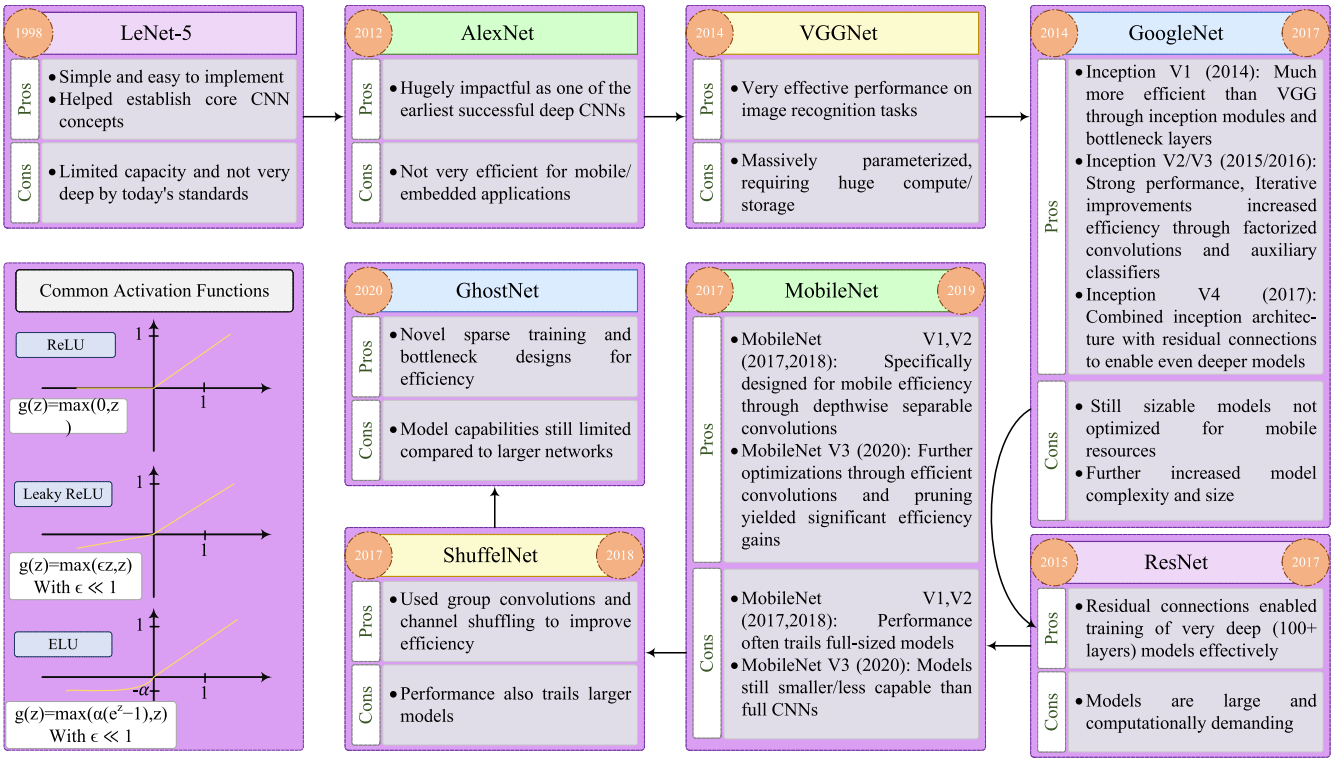

FIGURE 4. The flow of CNN architectures from 1998-2020 with their pros and cons represents that each CNN model is efficient for a specific application.

original signal into approximation and detail coefficients. The approximation contains lower frequency information, while the detail contains higher frequency or detailed information [109].

CNNs can utilize this multi-resolution decomposition property of wavelets by using convolutions to learn wavelet filters at each level [108], [109], [110]. The output of each level becomes the input to the next, with the filters extracting more detailed features at higher levels after the removal of coarse information. This convolutional learning of adapted wavelet bases enables CNNs to hierarchically capture patterns across different scales for improved data representation [110].

In various image processing and computer vision tasks, the use of convolutional wavelets within CNNs has shown promising results. For applications like denoising, super-resolution, and texture synthesis, CNNs equipped with learned wavelet filters have achieved state-of-the-art performance by effectively representing key multi-scale characteristics of visual data [110], [111], [112], [113]. Convolutional wavelets also benefit segmentation, detection, and classification when combined with traditional convolutional filters within CNNs [109]. In summary, wavelets provide a powerful tool for multi-scale analysis that CNNs can leverage through their inherent ability to learn localized basis functions via convolution operations.

III. BASIC CONVOLUTIONAL NEURAL NETWORKS

The CNN architecture typically consists of an initial input layer, followed by several critical components, including

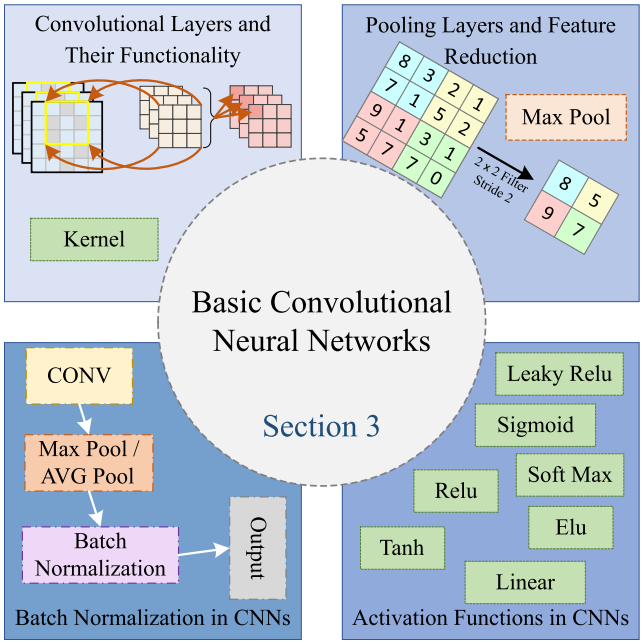

FIGURE 5. A graphical representation of Section III.

convolutional layers, pooling layers, and fully connected layers. This organized structure allows for the systematic processing of raw data, such as images, through a series of layers, which in turn enables the extraction of relevant features and facilitates making predictions.

The convolutional layers hold a central position in this architecture, as they employ learnable filters to process the

input data. This operation is instrumental in detecting diverse patterns and features, thereby enhancing the network's ability to understand the underlying data. Following the convolutional layers, the pooling layers come into play, downsampling the output from the previous layers. This downsampling process reduces the spatial dimensions while retaining crucial information. By focusing on the most significant details, these layers contribute to translational invariance, a valuable aspect in applications like image recognition where object positions may vary.

In Table 2, a comprehensive overview of the core components of basic CNNs is presented (also See Fig. 5), encompassing convolutional layers, pooling layers, and activation functions. The table provides insights into their individual purposes, functionalities, dependencies on input size, parameters, feature maps, translational invariance, computational efficiency, output size, roles in the CNN architecture, and impact on model performance. Analyzing these aspects provides profound insights into the elements that contribute to the effectiveness and performance of CNNs, making it a valuable reference for researchers and practitioners in the field.

A. BACKGROUND OF DEEP LEARNING

Deep learning, a prominent form of machine learning, encompasses the use of neural networks composed of multiple layers to acquire hierarchical representations of data [17]. Taking inspiration from the intricate workings of the human brain, where neurons engage in processing and transmitting information to forge elaborate depictions of the world, DL models, also known as deep neural networks, showcase remarkable prowess in assimilating hierarchical features from raw data. This exceptional ability enables them to discern intricate patterns and achieve remarkable precision in predictions [18].

The roots of DL can be traced back to the nascent endeavors surrounding artificial neural networks in the 1940s. However, the true resurgence and substantial remarkable materialized in the 1980s and 1990s, paving the way for its remarkable revival in the 21st century [19]. Key catalysts driving this resurgence were the strides made in computational power, the vast availability of datasets, and the advent of efficient training algorithms, most notably backpropagation, which played a pivotal role [20]. By harnessing these advancements, DL models attained the ability to process and analyze vast repositories of data, thus acquiring an aptitude for deciphering intricate patterns and making precise predictions.

The convergence of powerful hardware and sophisticated algorithms ushered in an era of remarkable accomplishments across diverse domains. Computer Vision (CV), natural language processing (NLP), and speech recognition (SR), among others, have witnessed remarkable strides through the transformative power of DL [73]. This dynamic discipline's capacity to overcome more difficult problems and promote innovation across various industries is becoming more and more clear as it develops and advances.

B. INTRODUCTION TO CONVOLUTIONAL NEURAL NETWORKS

CNNs, an influential category of DL models, have emerged as a preeminent and extensively utilized algorithm within the realm of DL [21]. Distinctive to CNNs is their capacity to engage in convolution calculations and operate proficiently on intricate structures. This characteristic has propelled CNNs to achieve remarkable breakthroughs in image analysis and feature extraction, bestowing upon them the ability to discern and efficiently classify features in images. Moreover, CNNs are renowned as shift-invariant artificial neural networks, a nomenclature that accentuates their capability to classify input information based on its hierarchical arrangement [22].

The hierarchical architecture of CNNs empowers them to process and extract features from input data in a shift-invariant manner [22]. This implies that CNNs can adeptly recognize and classify objects within images, irrespective of their position or orientation. The realization of this shift-invariant attribute is accomplished through the application of convolutional layers, which employ filters in a sliding window fashion. These filters acquire the ability to detect specific patterns or features at various spatial scales, thereby enabling the network to encapsulate both local and global information. Consequently, CNNs exhibit profound proficiency in extracting meaningful features from images, facilitating a wide array of applications encompassing object detection, image recognition, and even image generation [74].

C. CONVOLUTIONAL LAYERS AND THEIR FUNCTIONALITY

Each convolutional layer comprises multiple filters, also referred to as kernels, which are small windows that slide over the input data [32]. During the training phase, the weights of these filters are learned, and they function as feature extractors, identifying specific patterns, edges, and textures present in the input [33]. When the filters move across the input, they create feature maps that emphasize important parts of the data as regions of interest (ROI). These maps show where specific patterns in the input become active, helping the CNN recognize significant features crucial for later tasks like classification or detection [34].

For example, in a CNN trained to identify cats in images, the filters may learn to recognize the patterns of fur, whiskers, and ears. As the filters convolve across an image of a cat, they generate feature maps that highlight these specific regions of interest. These feature maps indicate the activation of these cat-specific patterns and aid in accurately classifying the image as containing a cat.

D. POOLING LAYERS AND FEATURE REDUCTION

Pooling layers are incorporated following convolutional layers to decrease the spatial dimensions of the feature maps,

TABLE 2. The different aspects of the basic convolutional neural networks.

Aspect	Convolutional Layers	Pooling Layers	Activation Functions	Batch Normalization
Purpose	Feature extraction	Feature reduction	Introduce non-linearity	Training stabilization
Functionality	Detect patterns and textures	Downsample feature maps	Add non-linearity	Normalizing activations
Input size dependency	Depends on input dimensions	Reduces spatial dimensions	Independent of input	Depends on input size
Parameters	Learnable weights (kernels)	No parameters	No parameters	Learnable scaling & shifting parameters
Feature maps	Produce feature maps	No feature maps	No feature maps	No feature maps
Translational invariance	Not inherently invariant	Introduces some invariance	Independent of input	No Translational invariance
Computational efficiency	Computationally intensive	Reduces computation complexity	Low computation cost	Enhanced training stability
Output size	May or may not match the input size	Reduced size	Unchanged	Unchanged
Role in CNN architecture	Central component	Interposed between convolutions	Enable learning complex relationships	Improve convergence, ease of tuning
Influence on model performance	Significantly impacts performance	Affects model efficiency	Crucial for Learning	Significantly impacts performance
Interpretability	Low	Low	Low	Normal
Training complexity	High	Low	Low	Normal
Memory usage	Normal	Low	Low	Normal

thereby reducing the computational complexity of the network [35]. The most frequently utilized pooling techniques in CNNs are max-pooling and average-pooling [37].

Max-pooling entails selecting the maximum value from a small region of the feature map, while average-pooling computes the average value. Pooling offers two primary advantages: first, it effectively reduces the number of parameters in the network, resulting in improved computational efficiency. Second, it introduces a level of translational invariance, signifying that minor spatial translations in the input data do not substantially impact the pooled outputs. This property enhances the CNN's ability to generalize better to variations in the input data.

For example, in image classification applications, after several convolutional and activation layers, a pooling layer can be used to downsample the feature map. This downsampling reduces the spatial resolution of the features, making it more computationally efficient to process and reducing the risk of overfitting. Additionally, because pooling computes either maximum or average values, it can capture the dominant features in an image regardless of their exact location, making the network more robust to slight variations in object position or orientation.

E. ACTIVATION FUNCTIONS IN CNNS

Activation functions play a vital role in CNNs as they are applied to the output of each neuron, introducing nonlinearity to the network and facilitating the learning of complex relationships between input data and their corresponding features. Within CNNs, several commonly used activation functions include Rectified Linear Units (ReLU) [36], which set negative values to zero while preserving positive values unchanged. Variants like Leaky ReLU [36] and Parametric ReLU [39] are also widely employed. The selection of the activation function is of great importance as it directly impacts the network's capacity to learn and make accurate predictions. By introducing nonlinearity, activation functions allow CNN to model intricate patterns and decision boundaries, thereby enhancing its performance across a range of tasks.

For example, in image classification applications, the ReLU activation function has been shown to effectively remove negative pixel values and emphasize positive pixel values, allowing CNN to identify important features and learn discriminative patterns. This enables the CNN to accurately classify different objects in images, such as correctly identifying whether an image contains a cat or a dog.

F. BATCH NORMALIZATION IN CNNS

Batch Normalization is a technique that helps stabilize and accelerate the training of CNNs [78]. It normalizes the activations of each layer by centering and scaling the values using the mean and variance of each mini-batch during training. This process reduces internal covariate shifts, making the optimization process smoother and enabling the use of higher learning rates.

By normalizing activations, Batch Normalization allows for more aggressive learning rates, which leads to faster convergence and improved model generalization. Additionally, it acts as a regularizer, reducing the need for other regularization techniques like dropout.

Overall, Batch Normalization has become a standard component in CNN architectures, contributing to faster training, improved model performance, and increased ease

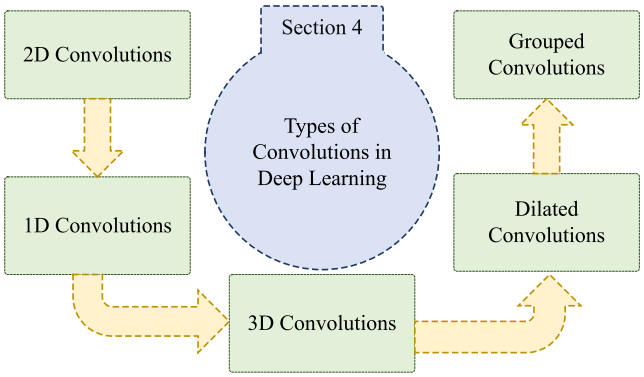

FIGURE 6. An overview of Section IV structure.

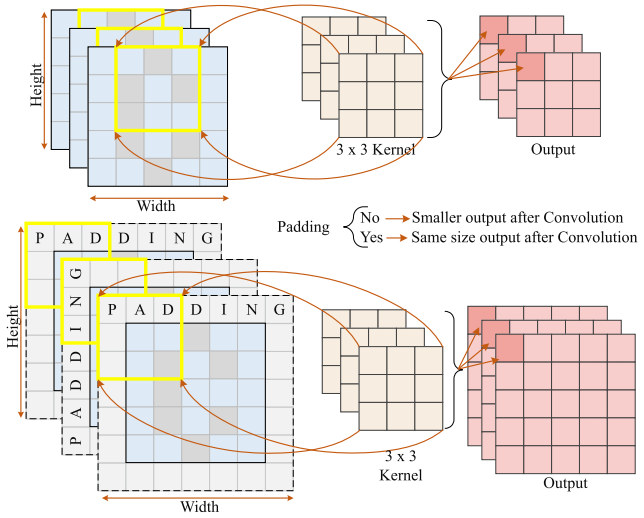

FIGURE 7. The Basic structure of CNN. a) represents CNN without Padding which causes the output image to become smaller. b) represents CNN without Padding which the output image is the same size as the input image.

of hyperparameter tuning. Its widespread adoption has significantly contributed to the success of modern CNNs in various CV and NLP applications. For example, in image classification applications, Batch Normalization helps reduce overfitting by normalizing the input for each mini-batch during training. This ensures that the network learns robust features and avoids relying on specific pixel values or noise in the input data. As a result, the model becomes more generalized and performs better on unseen data.

IV. TYPES OF CONVOLUTION IN DEEP LEARNING

In this section, our goal is to comprehensively explore the different convolution methods (See Fig. 6) commonly used in deep learning models. Table 3 presents a condensed overview of these convolution types, providing important information such as input data type, dimensionality, receptive field, computational cost, primary use case, memory consumption, parallelization capability, consideration of temporal information, and computational efficiency.

It is important to highlight that selecting the appropriate convolutional type relies on the particular task and dataset under consideration. For instance, when working with diverse data types, such as images or text, it may be necessary to employ distinct convolutional types to effectively capture relevant features. Moreover, considering the computational efficiency of each convolutional type becomes important for real-time applications or settings with limited resources.

A. 2D CONVOLUTIONS

2D convolutions (See Fig. 7) serve as the foundational elements in CNNs, particularly for applications related to CV. They are predominantly utilized for processing two-dimensional data, such as images, which can be represented as a grid of pixels. During this convolutional operation, a 2D kernel slides over the input image, enabling the capture of local patterns and the extraction of relevant features [27]. The primary application of 2D convolutions lies in image recognition, wherein the model learns to identify essential patterns, including edges, textures, and object components, thereby facilitating high-level recognition applications [40].

2D convolutions have found use in a variety of fields, including signal processing, CV, and NLP in addition to image recognition. CNNs have completely changed CV processes like object detection, image segmentation, and facial recognition. CNNs can more accurately and efficiently analyze the spatial relationships and hierarchical structures present in images by using 2D convolutions. When learned filters slide across the input image, a CNN can learn to find and locate different objects in images, such as in object detection tasks. This helps the network accurately detect objects even in complicated scenes, as it can identify important patterns of various sizes.

Moreover, CNNs can also be learned to categorize and compare faces by analyzing facial features using 2D convolutions in facial recognition. This makes it possible to create systems like access control and identity verification.

B. 1D CONVOLUTIONS FOR SEQUENTIAL DATA

One-dimensional (1D) convolutions (See Fig. 8) are specially designed for working with sequential data like time series, audio signals, and natural language. Unlike their two-dimensional counterparts, 1D convolutions operate on a single line, allowing them to detect patterns that develop over time [41]. In the field of natural language processing, 1D convolutions are widely used in tasks such as classifying text and analyzing sentiments. They help the model identify complex patterns in sequences of words and understand how these words are related to each other [42]. 1D convolutions have also been successfully applied to audio signal processing applications such as SR and music analysis. By analyzing the temporal patterns of audio signals, these models can extract meaningful features that capture the underlying structure and characteristics of the sound. This has proven to be particularly useful in applications like speaker identification and emotion

TABLE 3. The comparison provides an overview of the characteristics and functionalities of different convolution types.

Convolution Type	2D Convolutions	1D Convolutions	3D Convolutions	Dilated Convolutions	Grouped Convolutions
Input Data Type	Images	Sequential Data (e.g., Text)	Volumetric Data (e.g., Videos)	Images	Images
Dimensionality	2D	1D	3D	1D, 2D	2D
Receptive Field	Local	Local	Volumetric	Local	Local
Computational Cost	Medium	Low	High	Low	High
Main Use Case	Image recognition, Object detection	Text classification, Sentiment analysis	Semantic segmentation, 3D medical imaging	Image Filtering, Image generation	Large-scale CNN architectures
Memory Consumption	Medium	Low	High	Low	Low
Parallelization	Limited	Limited	Limited	Limited	High
Use of Temporal Information	Not applicable	Captures temporal patterns	Captures spatial temporal patterns	Not applicable	Not applicable
Computational Efficiency	Medium	High	Medium	High	High

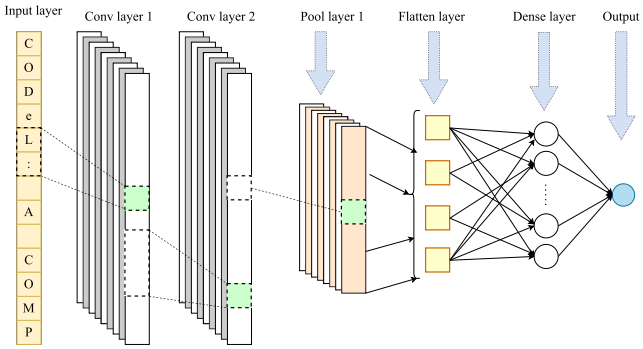

FIGURE 8. An overview to simple one-dimensional (1D) convolution neural network with two convolution layer.

recognition, where the sequential nature of the audio data is sequential.

For example, in speaker identification, 1D convolution can analyze the sequential patterns of an individual's voice and learn to associate certain patterns with specific speakers. This allows the model to accurately identify and differentiate between different speakers in an audio recording. In emotion recognition, 1D convolutions can analyze the temporal changes in pitch, tone, and intensity of an audio signal to classify the emotional state of the speaker, such as happiness, sadness, or anger. This helps in detecting and understanding the underlying emotions conveyed through speech, which can be useful in various applications like customer sentiment analysis, virtual assistants, and mental health monitoring.

C. 3D CONVOLUTIONS FOR VOLUMETRIC DATA

Three-dimensional (3D) convolutions are specifically designed to handle volumetric data, such as 3D medical images or video data [43]. 3D convolutions possess the capability to simultaneously process spatial and temporal dimensions, thereby capturing intricate patterns and distinctive features across all three dimensions. In medical imaging, 3D convolutions are vital in jobs like finding where tumors are. The model uses 3D medical scans to figure out where the important spatial and surrounding details are, which helps accurately locate and describe tumors [44], [45].

The use of 3D convolutions has gone beyond just tumors and is used in various medical imaging tasks like picking out different parts of the body, spotting issues, and classifying diseases. This method lets the model see the whole volume of a medical scan, rather than just individual parts, and consider how different slices are related in space. This comprehensive approach allows the model to effectively capture the overall structure of the target organ or an anomaly, resulting in improved diagnostic accuracy and better patient outcomes.

For instance, in tumor segmentation, 3D convolutions can be used to analyze a series of consecutive medical scans to identify the size and location of tumors over time, allowing doctors to track their growth and plan targeted treatments. This helps improve the accuracy and efficiency of tumor identification, leading to better patient outcomes.

In addition to operating on raw medical images and videos, 3D convolutions can be applied to process point cloud data through voxelization [101]. As point clouds represent 3D geometry as an unordered set of points without connectivity, a common approach is to first discretize the continuous 3D space into regular volumetric grids called voxels. Each voxel is assigned a feature vector, such as the number of points or aggregated point properties within its volume.

Voxelizing the point cloud allows existing 3D convolutional kernel operations to be directly applied. Early works divided the spatial domain into coarse voxels and maxpooled point features inside each voxel [101]. More advanced methods utilize sparse convolutions over fine-grained voxels or use dilated kernels with gaps to control the receptive field size. Multi-scale voxels have also been explored to capture both local and global point features [125], [126].

After 3D convolution and pooling, the extracted voxel features can be decoded back to the original point cloud domain for subsequent 3D fully connected or Transformer layers [129]. Voxel representation serves as an efficient

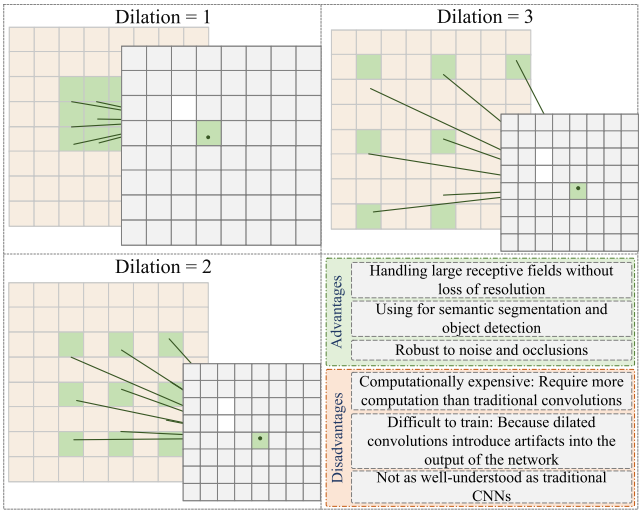

FIGURE 9. Dilation Convolution with multiple dilation rate with 3 × 3 kernel size [74].

intermediary that not only maintains the spatial structure required by CNNs but also allows points of variable density [127], [128] [129]. This two-stage voxel-based approach enables end-to-end training of 3D CNNs for point clouds.

D. DILATED CONVOLUTIONS AND THEIR ADVANTAGES

Dilated convolutions (See Fig. 9), also known as atrous convolutions, are a variant of traditional convolutions that introduce gaps (dilation) between kernel elements. This gap enables for an increased receptive field without increasing the number of parameters, making dilated convolutions more computationally efficient [46]. Dilated convolutions find application in applications like semantic segmentation, where they enable the model to capture broader contextual information without compromising computational efficiency [47].

In semantic segmentation applications, dilated convolutions are particularly useful because they enable the model to capture broader contextual information. By introducing gaps between kernel elements, dilated convolutions increase the receptive field without adding more parameters. This means that the model can understand the surrounding context of each pixel or object in the image without sacrificing computational efficiency. This value is important in applications like semantic segmentation, where accurately identifying and classifying objects within an image is essential.

E. GROUPED CONVOLUTIONS FOR EFFICIENCY

Grouped convolutions (See Fig. 10) involve dividing the input and output channels of a convolutional layer into groups. Within each group, separate convolutions are performed, which are then concatenated to produce the final output. This technique significantly reduces computational cost and memory consumption while promoting model parallelism [48]. Grouped convolutions are commonly used in large-scale

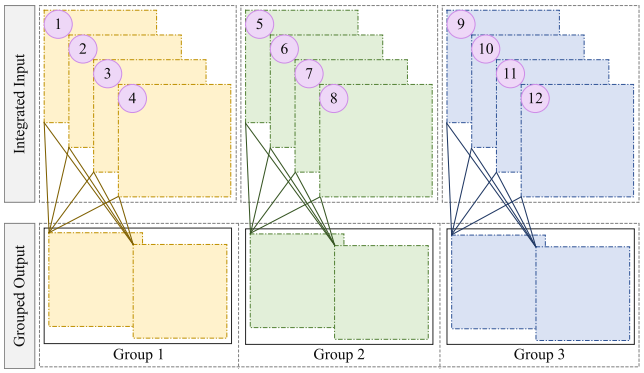

FIGURE 10. Grouped convolution involves dividing the channels of a convolutional layer into 3 groups.

CNN architectures to reduce training time and enhance the scalability of DL models [49].

In addition to reducing computational cost and memory consumption, grouped convolutions also offer other advantages. One of the main benefits is improved model parallelism, which provides for better utilization of parallel computing resources. This is especially important in large-scale CNN architectures where training time can be a bottleneck. By dividing the input and output channels into groups, the convolutions can be performed in parallel, speeding up the entire training process. Furthermore, the scalability of DL models is enhanced with grouped convolutions, making it easier to deal with larger datasets and more complex applications.

For example, in image classification applications, a large-scale CNN architecture such as ResNet can benefit from model parallelism using grouped convolutions. By dividing the input and output channels into groups, different subsets of the model can be trained in parallel on multiple GPUs or distributed systems. This not only reduces the training time but also allows for better resource utilization, eventually improving the scalability of the DL model to handle larger datasets and more complex image recognition applications.

In conclusion, DL offers a diverse range of convolutional techniques to accommodate different data types and applications. From 2D convolutions for image recognition to 1D convolutions for sequential data and 3D convolutions for volumetric data, each convolution type has its unique advantages. Additionally, dilated convolutions and grouped convolutions serve as efficient alternatives, addressing specific challenges in DL models. Understanding the characteristics and applications of these convolution types empowers researchers and practitioners to design efficient and effective models for a wide array of applications.

F. EVOLUTION OF CNN ARCHITECTURES

Since the early origins of CNNs, there has been a rapid evolution in CNN architectures (See Fig. 11) [49] over the past decade to enhance performance and efficiency [51]. Some key developments include:

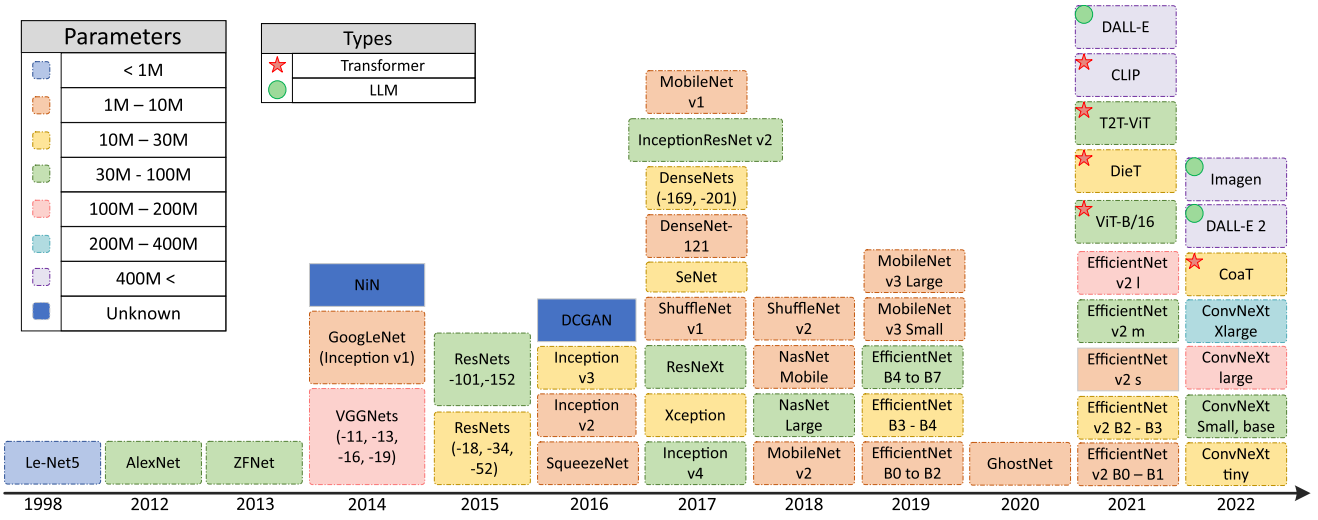

FIGURE 11. The detailed overview of advanced convolutions techniques.

- Inception modules (2014) - The Inception architecture introduced convolutional blocks with multiple filter sizes to capture features at various scales [52]. This improves both accuracy and computational efficiency.
- ResNets (2015) - Residual networks allow the training of much deeper CNNs through shortcut connections that bypass multiple layers [53]. They reduce degradation in very deep models.
- DenseNets (2016) - These connect each layer to all subsequent layers for maximum information flow and feature reuse. This reduces the number of parameters [54].
- MobileNets (2017) - Designed specifically for mobile applications, they use depthwise separable convolutions to minimize model size and latency [55].
- EfficientNets (2019) - By systematically scaling network dimensions, these achieve much better efficiency-accuracy trade-offs [55].

The evolution of CNN architectures (See Fig. 11) has been crucial to their widespread adoption across vision applications.

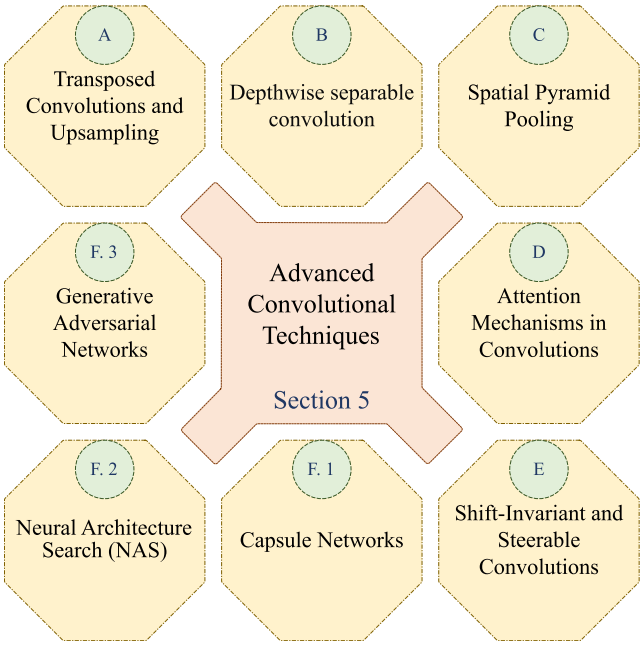

FIGURE 12. The trend of CNNs over time based on the released year and amount of parameters and their types.

V. ADVANCED CONVOLUTIONAL TECHNIQUES

This section provides a detailed overview of advanced convolutional techniques (See Fig. 12). A clear and informative summary of these techniques is available in Table 4. By reviewing this table, readers can gain a better understanding of the state-of-the-art convolutional techniques and their potential uses.

A. TRANSPOSED CONVOLUTIONS AND UPSAMPLING

Transposed convolutions–also referred to as deconvolutions or fractionally stridden convolutions–are sophisticated methods for upsampling feature maps [57]. Transposed convolutions, as opposed to conventional convolutions, increase the feature map size, enabling the model to reconstruct higher-resolution representations from lower-resolution inputs [58]. Traditional convolutions reduce spatial dimensions. In processes like image segmentation [59], image creation [60], and image-to-image translation [61], they are essential. Transposed convolutions employ padding and stride values to regulate the upsampling process and learnable parameters to choose the output size.

Transposed convolution can create artifacts or checkerboard patterns in generated feature maps, due to overlapping receptive fields. To prevent this, stride, padding, and dilation are used to control the output resolution and reduce these

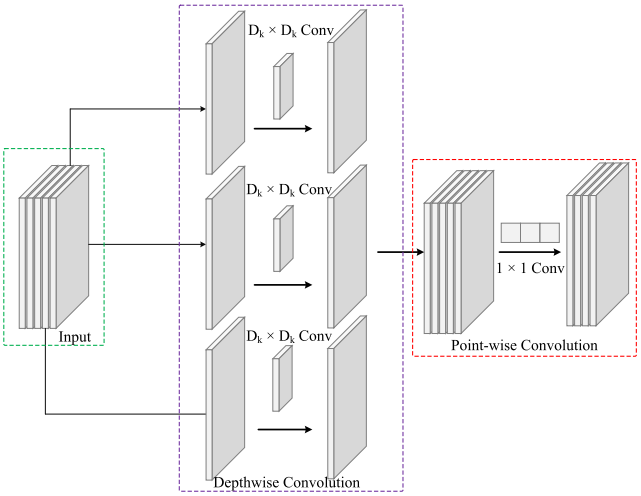

FIGURE 13. The box with purple color represents the depthwise convolution and the box with red color represents pointwise convolution (in pointwise a 1 × 1 convolution is used).

artifacts. In the field of image generation, transposed convolutions are used to upscale low-resolution images into high-resolution ones. To ensure the generated images are free of artifacts or checkerboard patterns, stride, padding, and dilation are adjusted to control the output resolution and enhance the quality of the generated images.

B. DEPTHWISE SEPARABLE CONVOLUTIONS (DSC)

Depthwise separable convolutions (See the purple box in Fig. 13) are an efficient alternative to traditional convolutions, particularly in resource-constrained environments [62], [63]. They split the convolution process into two steps (See Fig. 13) depthwise convolutions [64] and pointwise convolutions [65], [273], [274], [275], [276]. Depthwise convolutions apply a separate kernel to each input channel, capturing spatial patterns independently for each channel. Pointwise convolutions then use 1 × 1 convolutions to combine the output channels from the depthwise step, effectively aggregating the information [66]. Depthwise separable convolutions significantly reduce the number of parameters and computation while maintaining model performance, making them popular in mobile and embedded applications [67].

By decoupling spatial filtering from cross-channel filtering, depthwise convolution achieves higher computational efficiency and is well-suited for resource-constrained environments. MobileNet and Xception are popular CNN architectures that use depthwise convolution to reduce model size and improve inference speed without compromising performance significantly.

C. SPATIAL PYRAMID POOLING (SPP)

Spatial pyramid pooling (SPP) is a technique used to handle inputs of varying sizes and aspect ratios in CNNs [68], [277], [278], [279], [280], [281], [282]. It divides the input feature maps into different regions of interest and applies max-pooling or average-pooling to each region independently. The resulting pooled features are then concatenated to form a fixed-length representation, which is fed into fully connected layers for further processing. SPP enables the CNN to accept input images of different sizes and produces consistent feature maps, making it useful in object detection and image segmentation applications [69].

D. ATTENTION MECHANISMS IN CONVOLUTIONS

Attention mechanisms in convolutions allow the model to focus on relevant parts of the input, emphasizing specific regions during feature extraction [70]. These mechanisms assign weights to different spatial locations based on their importance. Self-attention mechanisms [70], like those used in transformers, have been adapted for use in convolutions. They enable the network to capture long-range dependencies and context, improving the model's ability to recognize complex patterns and relationships.

E. SHIFT-INVARIANT AND STEERABLE CONVOLUTIONS

Shift-invariant convolutions are designed to be insensitive to small translations in the input data [71], [283], [284], [285]. They ensure that the learned features remain consistent regardless of the object's position within the input image. This property is crucial for object detection applications, where the object's location might vary within the image [27]. Steerable convolutions are filters that can be rotated to different angles, allowing the model to learn orientation-sensitive features in an orientation-invariant manner [286], [287], [288]. These convolutions are often used in applications like text recognition, where the orientation of text can vary.

F. RECENT ADVANCEMENTS AND INNOVATIONS

1) CAPSULE NETWORKS

Capsule Networks, introduced by Geoffrey Hinton and his team, is a revolutionary advancement in CNNs [75]. They aim to address the limitations of traditional CNNs, particularly in handling spatial hierarchies and viewpoint variations [289], [290], [291], [292], [293], [294], [295]. Capsule Networks use capsules as fundamental units, which are groups of neurons that represent various properties of an entity, such as its pose, deformation, and parts.

Capsule Networks offer dynamic routing mechanisms to route information between capsules, allowing them to model complex hierarchical relationships more effectively. This enables the network to recognize objects with various poses and appearances, making Capsule Networks more robust to transformations and occlusions.

2) NEURAL ARCHITECTURE SEARCH FOR CONVOLUTIONS

Neural Architecture Search (NAS) is an automated approach to designing CNN architectures [76], [81]. Instead of relying on human-designed architectures, NAS employs search algorithms and neural networks to discover architectures that perform well on specific applications [76]. This technique

TABLE 4. The comparison provides an overview of the characteristics and functionalities of different convolution types - part 1.

Convolution Technique	Transposed Convolutions	DSC	SPP	Attention Mechanism	Shift-Invariant
Purpose	Upsampling	Parameter Reduction	Handling Varying Input Sizes	Focus on Relevant Features	Invariance
Parameters	Learnable	Learnable	No parameters	Learnable	Learnable
Computational Cost	High	Low	Low	Normal	High
Parameter Efficiency	Low	High	High	Low	Normal
Upsampling	Yes	No	No	No	No
Spatial Handling	Spatially Invariant	Spatially Invariant	Variable regions	Spatially Invariant	Spatially Invariant
Long-range Dependencies	No	No	No	Yes	No
Translation Invariance	Yes	Yes	Yes	Yes	Yes
Rotation Invariance	No	No	No	No	No
Interpretability	Low	Low	Low	Low	Low
Model Size	Large	Small	Small	Small	Large
Versatility	Normal	High	High	High	Normal
Practical Applications	Image Segmentation, Image Super-Resolution, Image Generation	Mobile Vision Applications, Real-time Object Detection	Image Classification, Object Detection, Semantic Segmentation	Image Captioning, Visual Question Answering	Image Recognition, Object Detection, Image Filtering

TABLE 5. The comparison provides an overview of the characteristics and functionalities of different convolution types - part 2.

Convolution Technique	Steerable Convolution	Capsule Networks	NAS	GAN	VIT
Purpose	Efficiency and Invariance	Invariance	Efficiency	Synthesis	Long-range dependencies
Parameters	Learnable	Learnable capsules	Architecture search	Learnable	Learnable
Computational Cost	Low	High	High	High	Higher
Parameter Efficiency	High	Normal	High	Low	Normal
Upsampling	No	No	No	No	No
Spatial Handling	Spatially Invariant	Spatially Invariant	Spatially variant	Spatially Invariant	Spatially Invariant
Long-range Dependencies	No	No	No	No	Yes
Translation Invariance	Yes	Yes	Yes	No	Yes
Rotation Invariance	Yes	Yes	No	No	Yes
Interpretability	Low	Low	Low	Low	High
Model Size	Normal	Normal	Large	Large	Large
Versatility	Low	Low	Low	Low	High
Practical Applications	Image Filtering, Edge Detection, Pattern Recognition	Object Recognition, Image Segmentation, Medical Imaging	Customized CNN Architectures, Resource-Constrained Devices	Image Synthesis, Style Transfer, Data Augmentation	Image recognition, NLP, diverse tasks

has led to the development of state-of-the-art CNNs that outperform hand-crafted models [296], [297], [298], [299], [300], [301], [302], [303], [304], [305], [306].

NAS for convolutions involves exploring various convolutional designs, including different kernel sizes, depths, and connectivity patterns [82]. It evaluates each architecture on a validation set, and through a process of evolution or optimization, identifies the best-performing architecture.

In the scenario of self-autonomous vehicle navigation, NAS for convolutions could be used to design an optimal convolutional neural network architecture specifically tailored for processing and analyzing various types of visual data collected by the vehicle's sensors. By exploring different convolutional designs, such as varying kernel sizes, depths, and connectivity patterns, NAS could identify the most effective architecture for accurately detecting objects and recognizing road signs in real-time. This would ultimately improve the vehicle's ability to navigate autonomously and make informed decisions based on its visual perception.

3) GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) are a class of DL models used for generative applications, such as image synthesis, style transfer, and data augmentation [307], [308], [309], [310], [311], [312], [313]. GANs utilize CNNs as key components to model the generator and discriminator (See Fig. 14) [77], [83], [84]. The generator is a CNN that generates new samples, such as realistic images, while the discriminator is another CNN that aims to distinguish between real and fake samples [77]. These networks are trained adversarially, where the generator's goal is to produce samples that deceive the discriminator, and the discriminator's goal is to become better at distinguishing real from fake [71], [84].

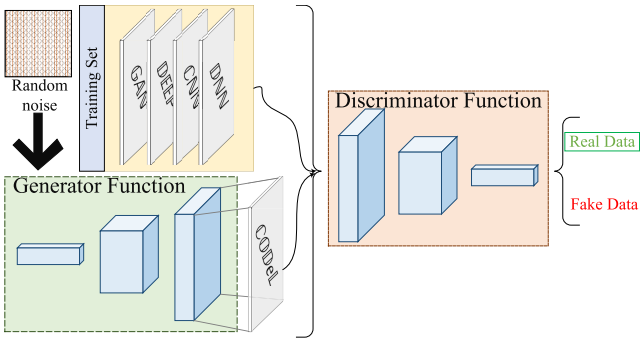

FIGURE 14. A simple GAN architecture represented to detect real and fake data which generator has generated.

GANs with convolution have revolutionized the field of image generation and have produced impressive results in generating high-quality images and realistic textures [263], [264], [265], [266], [267], [268], [269], [270], [271], [272]. They have also been extended to other domains like NLP, audio generation, and video synthesis. This technology has also been applied to other areas such as medical imaging, where GANs have been used to generate high-resolution and accurate images for diagnostic purposes. Additionally, GANs have shown promising results in the field of data augmentation, where they can generate synthetic data to increase the size and diversity of training datasets, improving the performance of machine learning models.

For example, in the field of image generation, GANs with convolutional networks have been used to create realistic images of non-existent landscapes. The generator network creates visually convincing images, while the discriminator network learns to identify any flaws or inconsistencies in these generated images, pushing the generator to improve its output. This adversarial training process ultimately leads to the creation of high-quality and believable images that are indistinguishable from real photographs.

G. VISION TRANSFORMERS AND SELF-ATTENTION MECHANISMS

Through the use of self-attention mechanisms [85], Vision Transformers [242], [243], [244], [245], [246], [247], [248], [249], [250], [251], [252], [253], [254], [255], [256], [257], [258], [259], [260], [261], [262] represent an important evolutionary step away from traditional computer vision architectures [86], [87]. Rather than solely relying on convolutional filters to process visual inputs, as has predominantly been the case, they segment images into distinct finite parts known as patches [87]. Each patch focuses on and extracts features from a different localized region of the photographic scene. This division of images into discrete patches is a major conceptual divergence from how most previous approaches operate.

In conclusion, advanced convolutional techniques have significantly expanded the capabilities of CNNs and revolutionized various fields like CV, image synthesis, and NLP. From transposed convolution for upsampling to capsule networks for handling spatial hierarchies, these innovations have enhanced the efficiency, robustness, and expressiveness of CNNs, making them powerful tools for a wide range of applications. Moreover, recent advancements, such as NAS and GANs, continue to drive progress in the field of DL and hold promise for further breakthroughs in the future.

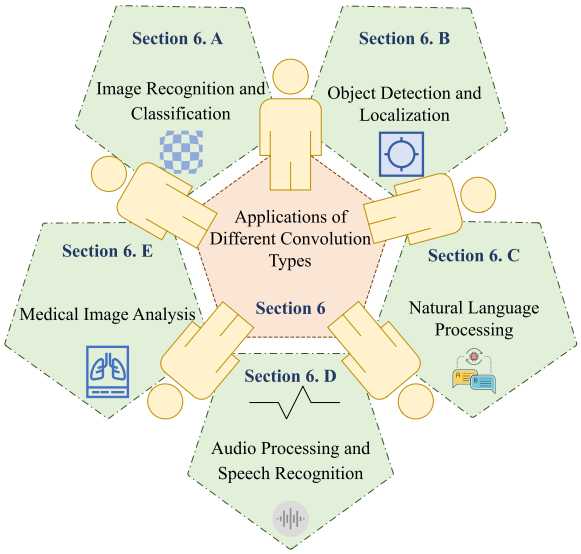

FIGURE 15. The applications of CNN techniques which we have discussed in Section VI.

VI. APPLICATIONS OF DIFFERENT CONVOLUTION TYPE

We provide a thorough overview of the numerous applications of different convolutional types in this section (See Fig. 15). Table 6 provides a brief but comprehensive overview of these applications. Convolutions of various types are used in a variety of contexts, demonstrating the flexibility and strength of CNNs. Convolutional techniques enable machines to understand and interact with complex data, facilitating advancements in a variety of fields and enhancing our daily lives. Examples include image recognition, object detection, NLP, and medical image analysis.

A. IMAGE RECOGNITION AND CLASSIFICATION

There are many uses for CNNs, including image recognition and classification. Traditional 2D convolutions are especially useful in these applications. They make it possible for deep learning models to accurately classify images into various groups and learn crucial features from images. The network's convolutional layers recognize edges, textures, and shapes. The pooling layers reduce the size of the image while preserving the data needed for classification. Image recognition and classification are used for various tasks, including optical character recognition (OCR) [88], [202], [203], [204], [205], [206], [207], [208], [209], classifying different animal species, and recognizing handwritten numbers [88]. In competitions like ImageNet, CNNs have

displayed impressive results, showcasing their abilities for handling wide image classification [89].

B. OBJECT DETECTION AND LOCALIZATION

Multiple objects within an image must be located and identified during object detection [90]. In this application, both conventional 2D convolutions and 3D convolutions are crucial [177], [178], [179], [180], [181], [182], [183], [184], [185], [186], [187], [188], [189], [190]. While 3D convolutions are used for video object detection, 2D convolutions are used to process individual image frames. CNNs can detect objects at different scales and aspect ratios thanks to their region proposal mechanisms and anchor-based methods [191], [192], [193], [194], [195], [196], [197], [198], [199], [200], [201].

Accurate localization of object bounding boxes is made possible by the use of pooling layers and convolutional sliding windows. Robotics, surveillance technology, and autonomous vehicles all use object detection to better understand and interact with their surroundings [91], [92].

C. NATURAL LANGUAGE PROCESSING

For sequential data, such as text processing and sentiment analysis, NLP uses 1D convolutions. 1D convolutions are used in NLP applications to extract pertinent patterns and relationships from sentences, enabling models to understand semantic meaning and context [210], [211], [212], [213], [214]. Sentiment analysis for understanding customer opinions, named entity recognition to extract specific information from text, and text classification to classify news articles or product reviews are examples of NLP applications using 1D convolutions. Applications like machine translation and text summarization have benefited from the successful integration of CNNs and recurrent neural networks (RNNs).

D. AUDIO PROCESSING AND SPEECH RECOGNITION

Audio Processing and Speech Recognition (APSR) benefit from 1D convolutions, which analyze and process sequential audio data such as speech signals or audio waveforms [215], [216], [217], [218], [219], [220], [221]. By extracting temporal patterns and acoustic features, CNNs can learn to recognize spoken words and transcribe audio into text. SR systems, often built upon convolutional and recurrent neural networks, enable voice assistants like Siri and Google Assistant to understand and respond to user commands.

E. MEDICAL IMAGE ANALYSIS

Medical image analysis involves the examination and interpretation of medical images, such as MRI scans, CT scans, and X-rays [92], [222], [223], [224], [225], [226], [227], [228], [229], [230], [231], [232], [233], [234], [235], [236], [237], [238], [239]. In this domain, 3D convolutions and dilated convolutions are frequently used. 3D convolutions process volumetric medical data, allowing CNNs to extract spatial and contextual information for applications like tumor segmentation, organ localization, and disease classification [92], [93]. Dilated convolutions enhance feature extraction and semantic segmentation in medical images, enabling precise identification of abnormal tissues and structures. The applications of convolution types in medical image analysis have led to significant advancements in healthcare, assisting doctors in diagnosis and treatment planning.

VII. FUTURE TRENDS IN CNN

CNNs continue to be a hot topic of research and have achieved remarkable success in various CV applications. Future trends and open research questions in the field of CNNs are emerging as technology develops and DL techniques become increasingly complex.

The investigation of more effective architectures that can achieve comparable performance with fewer parameters and computational resources is one future trend in CNN research. How to make CNNs more interpretable is another unanswered research question, as the reasoning behind CNN decisions is frequently difficult to comprehend due to the internal complexity of these systems. Another crucial area for future research is finding ways to strengthen CNNs and make them less vulnerable to hostile attacks.

One active area of research looks at designing efficient CNN architectures optimized for edge and mobile computing. As CV moves from data centers to cameras, smartphones, and IoT at the network's edge, models need to operate within strict constraints on latency, memory, and power. Techniques including network pruning, compact operators, knowledge distillation, and adaptive quantization help derive lightweight CNN variants suitable for these low-resource scenarios [121]. This focus on efficiency ties into work on improving CNN interpretability.

While today's complex CNNs achieve top accuracy, their decision-making remains poorly understood. Work on saliency mapping, activation clustering, modular CNNs, and other explanatory methods aims to shine light into the ''black box'' and address concerns around reliability, bias, and accountability - important considerations for safety-critical domains like healthcare. New types of CNN modules also aim to expand what these models can represent by incorporating flexible self-attention and capturing non-Euclidean structures.

A particularly compelling avenue involves tackling large-scale vision multimodal (LVM) challenges, which builds upon this work on expanding CNN capabilities. Vast datasets merging diverse visual media with language, audio, and other inputs present unprecedented complexity. However, they also offer unprecedented opportunities to develop general, comprehensive models of multisensory scene understanding.

A. INTERPRETABILITY AND EXPLAINABILITY OF CNNS

The interpretability and explainability of CNNs is a significant open research question. Understanding the

TABLE 6. The compact table highlights the main applications of each convolution type.

Convolution Type	Traditional 2D Convolutions	1D Convolutions	3D Convolutions	Dilated Convolutions	Grouped Convolutions
Image Recognition	Image categorization	Time series analysis	Action recognition	Image segmentation	Real-time recognition
Object Detection	Object detection	Event detection	3D object detection	Semantic segmentation	Efficient detection
NLP	Sentiment analysis	Text classification	Textual entailment	Hierarchical document classification	Parameter reduction
ASPR	Voice activity detection	Speech recognition	Environmental sound classification	Robust speech recognition	Low-latency speech recognition
Medical Image Analysis	Tumor segmentation	ECG signal processing	Brain Tumor Segmentation	Enhanced image segmentation	Faster medical analysis

decision-making process of these models gets harder as CNNs get deeper and more complex. Particularly in critical applications like healthcare and autonomous systems, researchers are investigating ways to interpret and explain CNN predictions. To increase trust and reliability in CNN-based systems, methods such as attention visualization, saliency maps, and attribution methods seek to reveal which areas of the input contribute most to the model's conclusion.

B. INCORPORATING DOMAIN KNOWLEDGE

Incorporating domain knowledge into CNN architectures is another important research direction. While CNNs have shown exceptional generalization abilities, they may not fully exploit domain-specific characteristics. Research focuses on developing architectures that can efficiently utilize domain knowledge or constraints, such as physics-based priors in medical imaging or geometric constraints in robotics, to improve performance and reduce data requirements.

C. ROBUSTNESS AND ADVERSARIAL DEFENSE

Enhancing the robustness of CNNs against adversarial attacks remains a significant challenge. Adversarial attacks involve adding carefully crafted perturbations to inputs, leading to incorrect predictions by the CNN model. Researchers are investigating techniques for adversarial defense, such as adversarial training, robust optimization, and input transformations, to make CNNs more resilient against these attacks.

D. EFFICIENT MODEL DESIGN

When using CNNs on devices with limited resources, such as smartphones and edge devices, efficiency in terms of computation, memory, and power consumption is important [240], [241]. Creating lightweight architectures, knowledge distillation methods, and effective model compression techniques will be future trends in CNN research to decrease the model size and increase inference speed while maintaining accuracy.

Model compression techniques play a crucial role in designing efficient DL models suitable for deployment on resource-constrained edge devices. Several methods (See Table 7) have been proposed to reduce model size and computations without significantly impacting predictive performance. Network pruning and quantization are two widely used compression approaches [102], [103].

Pruning techniques aim to sparsify neural networks by removing redundant connections with minimal impact on functionality [121]. Early methods relied on unstructured pruning where connections were simply set to zero based on their magnitude or importance ranking. However, such arbitrary pruning leads to non-standard sparse matrices thereby preventing hardware acceleration. More recent structured pruning techniques induce channel-wise, filter-wise, or block-wise sparsity to yield compact models amendable to efficient implementations [102], [121], [122], [123].

Filter pruning refers to removing entire convolutional filters, thereby achieving channel-wise sparsity [116], [122]. It has been shown that up to 90% of filters can be removed from VGG16 without accuracy degradation. One method, termed "Pruning-at-Initialization" prunes filters with the lowest sum values at the start of training itself. Alternatively, "One-Shot" prunes filters once based on their first-order Taylor expansion. These filter-level pruning methods lead to uniform sparsity across layers and reduce computation by $\tilde{5}$x.

Another structured approach is to prune blocks of connections rather than individual weights [123]. For example, in "Block Level Pruning", a number of convolution blocks are removed from blocks 1, 2, and 3 of ResNet50, reducing computations without retraining. The block structure ensures layout sparsity, maintaining original convolution block shapes for hardware friendliness. Network slimming is a channel-pruning method that enforces L1-norm regularization during training itself to gradually remove channels with low importance scores.

In unstructured variants, magnitude-based pruning removes weights below a threshold while iterative magnitude pruning alternates between weight updates and pruning based on a dynamic threshold [121], [124]. These maintain sparsity throughout the architecture but induce non-zero filler weights. Lottery ticket hypothesis experiments have demonstrated that dense, randomly-initialized, sub-networks can achieve the accuracy of their original networks if trained in isolation.

Apart from pruning, quantization is another effective technique to compress models (See Table 8). Weight and

TABLE 7. Comparison of pruning technique.

Technique	Sparsity Type	Pruning Granularity	Hardware Friendly	Accuracy Impact	Compression Ratio	Iterative Training	Requires Retraining
Magnitude Pruning	Unstructured	Weight level	No	Medium	2-10x	Yes	No
Filter Pruning	Channel-wise	Filter level	Yes	Low	5-10x	No	Yes
Block Pruning	Block-level	Block level	Yes	Low	2-5x	No	Yes
Network Slimming	Channel-wise	Channel level	Yes	Low	2-5x	Yes	Yes
Lottery Ticket	Unstructured	Weight level	No	Low	2-10x	Yes	Yes
Iterative Magnitude	Unstructured	Weight level	No	Medium	2-5x	Yes	No
Pruning-at-Init	Channel-wise	Filter level	Yes	Low	5-10x	No	No
One-Shot Pruning	Channel-wise	Filter level	Yes	Low	5-10x	No	No

TABLE 8. Comparison of quantization technique.

Technique	Quantization Level	Bit Width	Hardware Friendly	Accuracy Impact	Compression Ratio	Iterative Training	Requires Calibration
Weight Quantization	Weight values	8-bit	Yes	Low	Up to 8x	No	Yes
Activation Quantization	Activations	8-bit	Yes	Low	Up to 8x	No	Yes
Tensor Quantization	Tensors	4-8 bit	Yes	Low	Up to 32x	No	Yes
Tensor Decomposition	Tensors	4-bit	Yes	Medium	Up to 32x	No	No
Huffman Coding	Weights	Variable	No	Low	Up to 10x	No	No
Log Quantization	Activations	1 bit	Yes	Low	Up to 16x	No	No
BNN Quantization	Weights/ Activations	1 bit	Yes	High	Up to 32x	Yes	Yes
Floating Point Quantization	Weights/ Activations	16-bit	Yes	Low	Up to 2x	No	No

activation quantization methods map weights/activations to a small set of discrete values, reducing the number of bits required for representation [114], [115]. For example, 8-bit quantization reduces model size by 4x without accuracy loss for many architectures. Tensor decomposition-based quantization further compresses models by decomposing weight tensors into low-rank approximations.

Some recent works have combined multiple compression approaches in a multi-stage pipeline. One example jointly employs weight quantization, pruning, and Huffman coding on ResNet50, achieving over 10x compression with a minor accuracy drop. Another uses a two-phase pipeline consisting of filtering-based pruning followed by quantization to design efficient MobileNet variants. Such composite methods achieve better accuracy-efficiency tradeoffs than individual techniques alone.

In conclusion, network pruning and quantization offer promising avenues to design compact models for edge and mobile applications. While early methods relied on unstructured sparsing, recent techniques induce structure for hardware friendliness. Looking ahead, continued research on model compression holds the key to facilitating the adoption of deep learning across myriad resource-constrained environments.

E. MULTI-TASK LEARNING AND TRANSFER LEARNING

CNNs are well suited for multi-task learning, in which a single model is trained to carry out several related applications concurrently [161], [162], [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], [174], [175], [176]. The need for large amounts of labeled data for each individual task is being reduced as researchers investigate ways to take advantage of shared representations across applications and enhance generalization by transferring knowledge learned from one task to another [146], [147], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], [159], [160].

F. INTEGRATION WITH UNCERTAINTY ESTIMATION

Understanding model uncertainty is essential for safety-critical applications. Integrating uncertainty estimation into CNNs would allow models to quantify their confidence in predictions and prevent costly errors, which is an area of open research. To improve the uncertainty measures in CNNs, researchers are investigating Bayesian neural networks (BNNs), dropout-based uncertainty estimation, and Bayesian optimization techniques.

G. GENERALIZATION TO SMALL DATA REGIMES

A constant problem in the CNN research area is the generalization to small data regimes, where labeled training data are hard to come by. Essentially using data from related applications or domains, techniques like transfer learning, few-shot learning, and meta-learning work to increase CNNs' capacity to learn from sparse data.

H. EVOLUTION OF LANGUAGE MODELS AND MULTIMODAL LLMS

In recent epochs, the domain of large language models (LLMs) for natural language processing has witnessed a precipitous progression. Prototypes such as BERT, GPT-3, and PaLM have demonstrated exceptional aptitude in language

apprehension and generation, courtesy of self-supervised pretraining on voluminous text corpora [85]. As LLMs expand in magnitude and range, incorporating additional modalities beyond text is a burgeoning field of study. Multimodal LLMs strive to amalgamate language, vision, and other sensory inputs within a singular model architecture. They hold the potential to attain a more holistic understanding of the world by concurrently learning representations across diverse data types [96]. A significant hurdle is the effective fusion of the strengths of CNNs for computer vision and transformer architectures for language modeling.

One strategy involves employing a dual-stream architecture with distinct CNN and transformer encoders interacting via co-attentional transformer layers [97]. The CNN extracts visual features from images, providing contextual information that can guide language generation and comprehension. The transformer architecture models the semantics and syntax of text. Their interaction enables the generation of captions based on image content or the retrieval of pertinent images for textual queries. Alternative methods directly incorporate CNNs within the transformer architecture as visual token encoders that operate with text token encoders [98]. The CNN projections of image patches are appended to text token embeddings as inputs to the transformer layers. This unified architecture allows for end-to-end optimization of parameters for both vision and language tasks. Self-supervised pretraining continues to be vital for multimodal LLMs to learn effective joint representations before downstream task tuning. Contrastive learning objectives that predict associations between modalities have proven highly effective [99]. Models pre-trained on large datasets of image-text pairs have demonstrated robust zero-shot transfer performance on multimodal tasks.

As multimodal LLMs increase in scale, the efficient combination of diverse convolution types and attention mechanisms will be crucial. Compact CNN architectures could help to reduce the cost of computing. Sparse attention and memory compression techniques can assist with scalability.

VIII. PERFORMANCE AND EFFICIENCY CONSIDERATION

Considerations for performance and efficiency (See Figs. 17-20) in CNNs are critical in developing high-performing and resource-efficient models. Researchers can make informed decisions about optimizing their CNN architectures for various applications and deployment scenarios by analyzing computational complexity, trade-offs between accuracy and speed, memory requirements, and benchmarking on standard datasets. For our experiments on the CIFAR-10 dataset, we used an AMD Ryzen 7 4800H processor, 16GB of RAM, and an NVIDIA GeForce GTX 1660 Ti graphics card.

A. COMPUTATIONAL COMPLEXITY OF DIFFERENT CONVOLUTIONS

The computational complexity of different convolutional techniques (See Table 9) is a critical aspect to consider when designing CNNs. It refers to the amount of computation required to perform a convolution operation on input data. The computational complexity is influenced by various factors, including the size of the input data, the size of the convolutional filters, and the number of channels in the feature maps.

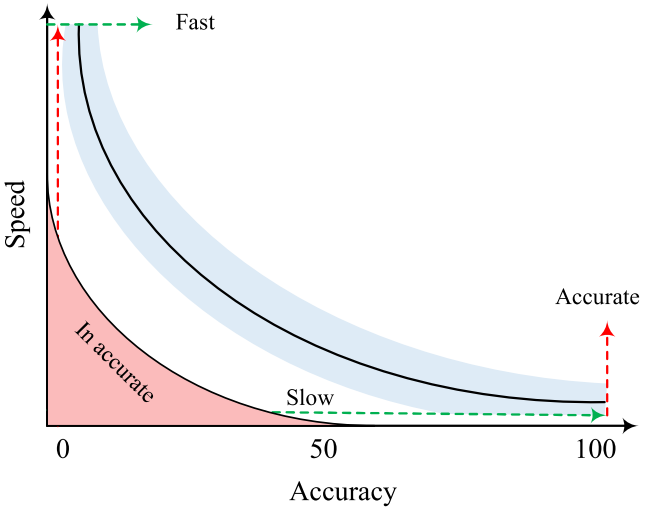

FIGURE 16. The trade-off curve between accuracy and speed of a deep learning model [75].

Traditional convolutional layers, such as the standard convolution and depthwise separable convolution, generally have higher computational complexity compared to other techniques. This is because they involve a large number of convolution operations, especially when dealing with high-resolution images or complex data. On the other hand, techniques like pointwise convolution and transposed convolution tend to have lower computational complexity, making them more suitable for certain resource-constrained applications.

Understanding the computational complexity of different convolution types is crucial for optimizing the performance of CNNs. By selecting convolution techniques that align with the available computational resources, researchers can build efficient models that achieve a good balance between accuracy and speed.

As illustrated in Figs. 17 to 19 the Adam optimizer performed well, as evidenced by key observations ① through ⑥, in both accuracy and loss metrics. Overall, the use of CNN techniques such as VGG, ResNet, and LeNet resulted in improved accuracy and reduced loss.

Also, as depicted in Figure 20, and based on key observation ①,②, and ③, it is evident that the Adam optimizer exhibits less CPU usage in comparison to five other optimizers - RMSprop, Adamax, Adagrad, SGD, and Nadam. This observation holds true when using LeNet-5, VGG16, and ResNet-50. Additionally, the memory usage of the Adam optimizer is among the lowest (See key observation ④).

B. TRADE-OFFS BETWEEN ACCURACY AND SPEED

One of the key challenging aspects of designing CNNs is balancing model accuracy and inference speed (see Fig. 16). The

TABLE 9. Comparison on LeNet-5, VGG16, and ResNet-50 with 7 types of optimizers on Cifar-10 dataset, CU: CPU utilization, MU: Memory utilization.

Optimizer Type	CNN Model	Accuracy	Loss	CU	MU
SGD	LeNet-5	0.547	1.277	71	50.7
	VGG16	0.87	0.776	57	55.7
	ResNet-50	0.789	1.1212	63	53.4
Adam	LeNet-5	0.629	1.153	46.2	44.4
	VGG16	0.805	0.821	54.2	51.4
	ResNet-50	0.760	1.016	60.5	51.9
NAdam	LeNet-5	0.624	1.22	58.3	57.6
	VGG16	0.776	1.109	61.1	63.5
	ResNet-50	0.789	0.89	66.4	57.8
RSMProp	LeNet-5	0.605	1.288	50.3	42.9
	VGG16	0.755	22.286	61.2	49.7
	ResNet-50	0.78	1.151	61.7	49.4
Adamax	LeNet-5	0.603	1.132	69.8	56.7
	VGG16	0.8506	0.885	55.8	64.2
	ResNet-50	0.8123	1.002	62.1	56.1
AdaGrad	LeNet-5	0.412	1.65	67.6	44.4
	VGG16	0.822	0.708	55.3	50.3
	ResNet-50	0.75	0.999	62.4	50.6

inference time increases as the complexity of convolutional layers increases to capture more complex features. Using simpler convolutional techniques, on the other hand, may result in lower accuracy. The depth and width of the network, the number of parameters, the choice of convolutional techniques, and the hardware on which the model is deployed all have an impact on the trade-offs between accuracy and speed. For real-time applications or resource-constrained environments, sacrificing some accuracy to achieve faster inference may be necessary.

Model pruning, quantization, and low-rank approximations are commonly used by researchers to reduce the model size (See Section VII -> Subsection D) and improve inference speed without significantly compromising accuracy. Furthermore, attention-based convolutions and other techniques that prioritize important regions of the input can be used to focus computational efforts where they are most needed, improving the balance between accuracy and speed even further.

C. MEMORY AND STORAGE REQUIREMENTS

Memory and storage requirements are crucial considerations in DL, especially when deploying models on edge devices or in cloud environments with limited resources. Convolutional models, particularly those with a large number of layers and parameters, can demand substantial memory and storage resources during training and inference.

Traditional convolutional layers often have higher memory requirements due to the need to store intermediate feature maps and gradients during backpropagation. Depthwise separable convolutions and pointwise convolutions can reduce memory usage by reducing the number of parameters and intermediate feature maps. Memory-efficient CNN design involves strategies like using smaller batch sizes, employing mixed-precision training, and optimizing memory usage during inference. Additionally, model compression techniques, such as knowledge distillation and model quantization, can significantly reduce the size of the model without significant loss in performance.

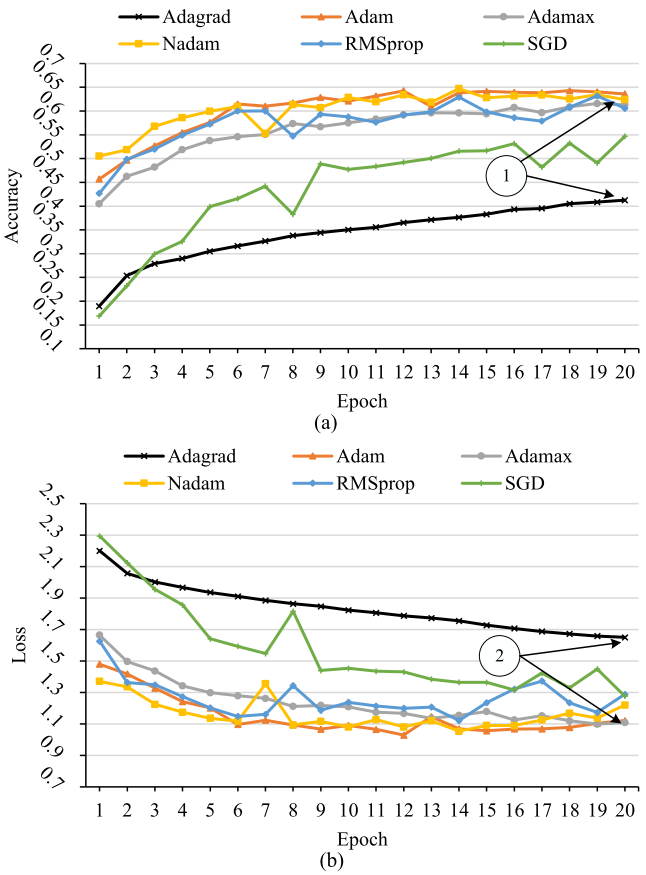

FIGURE 17. Comparison of various optimizers on LeNet-5 with Cifar-10 dataset. a) represents the accuracy of LeNet-5 architecture, b) represents loss of LeNet-5 architecture.

D. BENCHMARKING ON STANDARD DATASETS

Benchmarking convolutional techniques on standard datasets is a crucial step in evaluating their performance and efficiency. Standard datasets, such as ImageNet [95] for image recognition or COCO [94] for object detection, provide a common ground for fair comparison of different models and techniques. By benchmarking convolutional techniques, researchers can objectively assess their effectiveness in various applications and compare their performance with state-of-the-art models. The benchmarks consider metrics like accuracy, inference speed, memory usage, and energy efficiency, allowing for a comprehensive evaluation of the models.

Benchmarking helps the DL community identify the strengths and weaknesses of different convolutional techniques, paving the way for improvements and advancements. It also aids practitioners in selecting the most suitable convolutional techniques for their specific use cases and desired trade-offs between performance and efficiency.

IX. FRAMEWORKS AND LIBRARIES

This section will provide an overview of some of the popular platforms (See Table 10) available for developing deep learning applications. We will compare the frameworks

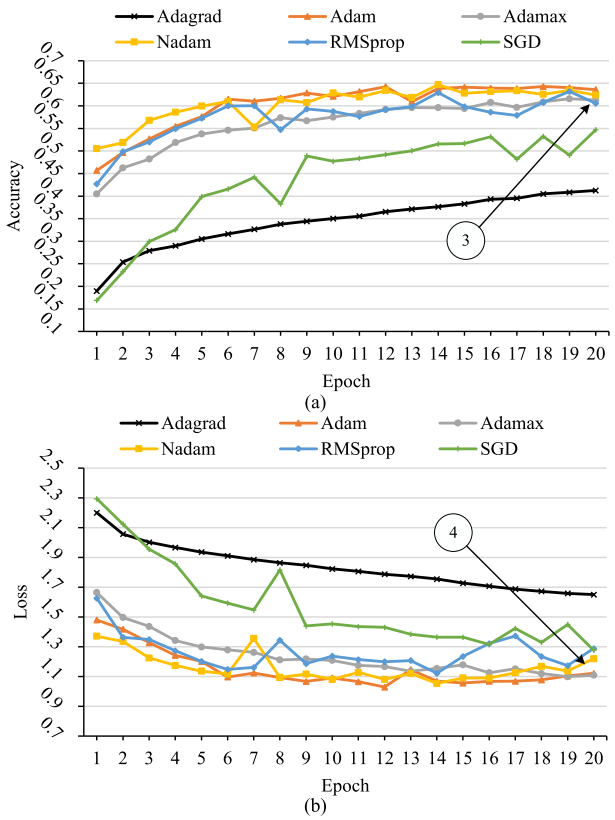

FIGURE 18. Comparison of various optimizers on VGG16 with Cifar-10 dataset. a) represents the accuracy of VGG16 architecture, b) represents loss of VGG16 architecture with various range of optimizers.

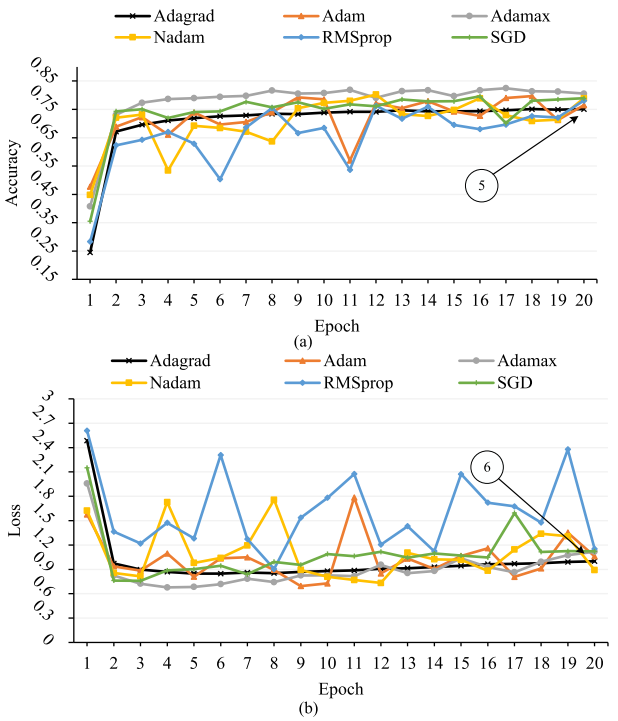

FIGURE 19. Comparison of various optimizers on ResNet-50 with Cifar-10 dataset. a) represents the accuracy of ResNet-50 architecture, b) represents loss of ResNet -50 architecture.

from aspects like their architecture, programming models, supported hardware, and key features. Choosing the right tool is crucial for deep learning success. That's why exploring framework capabilities is key for researchers and engineers

Table 10 provides a comparison of several popular frameworks and libraries used in deep learning. It evaluates key aspects such as the year of release, programming languages supported, license type, model definition approaches, ease of use, speed, and focus or strength of each framework.

A. CAFFE

Caffe was one of the earliest and most influential deep learning frameworks developed specifically for CV tasks [130]. Released in 2013 by the Berkeley Vision and Learning Center (BVLC), Caffe made training convolutional neural networks much faster and more accessible. It has an easy-to-use C++/Python interface and was designed for speed and modularity. Caffe adopted a layered structure that greatly simplified model definition and training. This helped drive wider adoption and enabled researchers to rapidly iterate on vision models. While development has slowed in recent years, Caffe laid important groundwork and is still used for CV research.

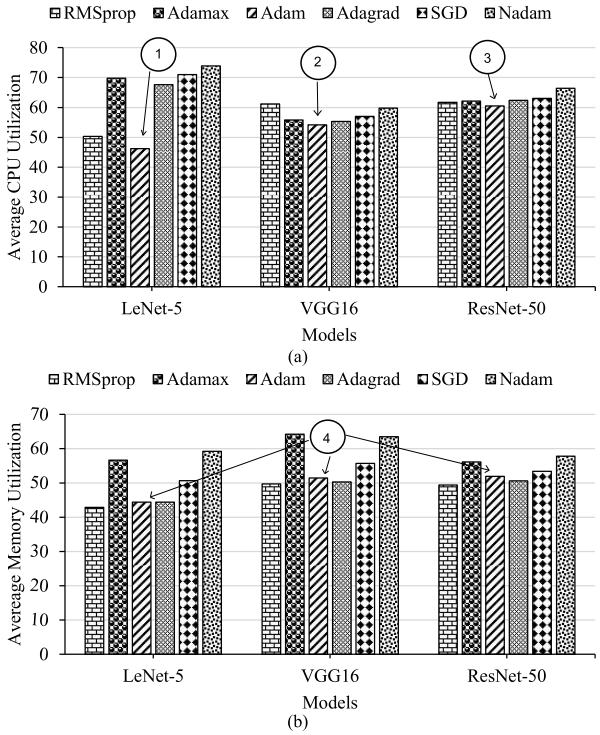

FIGURE 20. The CPU and memory utilization used by each model. a) The average CPU utilization of LeNet-5, VGG16, and ResNet-50 with six types of optimizer (better value recognition depends on use-case), b) The average memory utilization of LeNet-5, VGG16, and ResNet-50 with six types of optimizer (better value recognition depends on usecase).

B. TENSORFLOW

TensorFlow is an end-to-end open-source machine learning platform developed by Google [131]. While not strictly a CV library, it has become one of the most popular and full-featured frameworks for building and training complex

TABLE 10. Comparison of existing popular frameworks and libraries.

Aspect	Caffe	TensorFlow	Keras	PyTorch	OpenCV	Deeplearning4j	MXNet	Chainer
Year released	2013	2015	2015	2016	1999	2014	2015	2015
Programming language	C++/Python	Python, C++	Python	Python	C++, Python, Java	Java, Scala	Python, C++, R, Scala, Perl, Julia	Python
License	BSD 3-Clause	Apache 2.0	MIT	BSD 3-Clause	BSD 3-Clause	Apache 2.0	Apache 2.0	MIT
Model definition	Layered	Graph-based	Sequential & functional	Dynamic computations graphs	N/A	Sequential, compute graphs	Symbolic	Imperative and declarative
Ease of use	Intermediate	Intermediate	High	High	Low	Intermediate	Intermediate	High
Speed	Fast	Fast	Intermediate	Fast	Very fast	Fast	Fast	Fast
Support for computer vision	Very good	Excellent	Good	Excellent	Excellent (library)	Good	Goo	Good
Focus	Research prototyping	Production & research	User-friendly research	Research prototyping	Traditional algorithms	Enterprise production	Distributed training at scale	Intuitive high-level APIs for research
Distributed training	No	Yes	No	No	No	Yes	Yes	No
Model deployment	No	Yes	Yes	Yes	No	Yes	Yes	Limited
Hardware support	CPU, GPU	CPU, GPU, TPU	CPU, GPU	CPU, GPU, TPU	CPU, GPU	CPU, GPU	CPU, GPU, TensorFlow	CPU, GPU
Documentation quality	Good	Excellent	Good	Excellent	Excellent	Good	Good	Good
Community support	Limited	Very active	Very active	Very active	Very active	Active	Active	Active

DL models. TensorFlow has excellent support for CV including pre-trained models, image loading and preprocessing utilities, object detection APIs, and more. Its flexibility has led to it being used for a very wide range of applications from image classification to semantic segmentation. TensorFlow also works seamlessly across CPUs and GPUs and can be easily deployed to production.

C. KERAS

Keras is a high-level deep learning API that runs on top of popular frameworks like TensorFlow and CNTK [132]. Keras was developed with a focus on user-friendliness, modularity and extensibility. It provides excellent abstractions and tools for developing and evaluating deep learning models quickly. For CV, Keras ships with the ImageDataGenertator for real-time data augmentation as well as pre-defined models like VGG16. It also supports popular CV tasks like image segmentation, object detection, and feature extraction through convenient APIs. Keras' simplicity has made it very approachable for developers.

D. PYTORCH

PyTorch is an open-source deep learning platform developed by Facebook's AI Research Lab (FAIR) [133]. In recent years it has emerged as a leading alternative to TensorFlow especially for CV and NLP applications. PyTorch has a strong focus on dynamic neural networks and shares similarities to MATLAB and Numpy. This makes for an intuitive, Pythonic interface that is well-suited to CV prototyping and experimentation. PyTorch also supports GPU/TPU training along with production deployment. It has a growing ecosystem of 3rd party libraries and community support. Like Keras, PyTorch integrates tightly with common CV tasks and datasets.

E. OPENCV

OpenCV (Open Source Computer Vision Library) is a popular CV and machine learning software library [134]. While not specifically designed for deep learning, OpenCV contains many traditional CV algorithms and an extensive collection of image processing functions. These include capabilities like image filtering, morphological operations, feature detection and extraction, object segmentation, and face and gesture recognition among others. OpenCV integrates with deep learning frameworks and is frequently used for simpler CV tasks or as a pre-processing step before feeding data into neural networks.

F. MXNET

MXNet is a flexible, efficient, and scalable deep learning framework [135]. Similar to TensorFlow, it supports a wide variety of programming languages and hardware environments. MXNet excels at distributed training and supports training models containing billions of parameters across hundreds of GPUs. It also includes algorithms for CV like image recognition, object detection, and semantic segmentation. Overall, MXNet strikes a good balance between flexibility, performance, and ease of use making it suitable for large-scale CV problems.

G. CHAINER

Chainer is an open-source deep learning framework created by preferred networks in Japan [136]. It provides straightforward neural network abstraction similar to Keras with imperative and declarative model definitions. Chainer focuses on intuitive high-level APIs combined with low-level performance. It includes CV functionality like image loading, augmentation, pre-trained models, and model export. Chainer

supports GPU and multi-GPU training and deployment. Overall it provides a performant and productive environment for CV development.

H. DEEPLEARNING4J

Deeplearning4j (Dl4j) was launched in 2014 as an open-source deep learning library for Java and Scala on the JVM [137]. It enables large-scale distributed training on GPUs and CPUs. For CV tasks, Deeplearning4j offers tools like image loading, pre-trained models, model import from Keras and ONNX, and the samediff for dynamic model construction. Deeplearning4j focuses on production-ready deployment with capabilities like model serving, online prediction, and on-device inference via Android or iOS apps.

Overall, these libraries and frameworks represent the forefront of open-source tools transforming CV through deep learning. Each offers different strengths and tradeoffs between flexibility, performance, ease of use, and supported features. As CV tasks continue advancing, we can expect these projects to further incorporate state-of-the-art research while also lowering the barrier to development through improved tools and abstractions. CV is sure to remain a major application domain for deep learning innovation in both research and industry.

X. MAIN RESEARCH FIELDS

A. IMAGE CLASSIFICATION

Image classification was one of the earliest successes of CNNs. The seminal AlexNet achieved record-breaking results on the ImageNet challenge in 2012 by drastically improving upon prior techniques. Today, state-of-the-art CNNs for image classification routinely achieve human-level or better accuracy on standardized datasets. Architectures like ResNet, Inception, Xception, and EfficientNets optimize parameters, layer connectivity, and computation to classify thousands of object categories at superhuman performance levels [52], [53], [56], [273]. Beyond static images, video classification CNNs also extract spatial-temporal features to recognize complex activities and events.

B. OBJECT DETECTION

Object detection is another major CV application that relies heavily on convolutional modeling. Two-stage detectors like Faster R-CNN and one-stage detectors like YOLO leverage region proposal networks and anchor boxes trained via priors to simultaneously localize and classify objects within images [314], [315], [316], [317], [318], [319], [320], [321], [322], [323], [324], [325], [326], [327], [328]. Recent works further optimize speed and accuracy, enabling real-time object detection on billions of parameter models. Techniques like mobile object detection address embedded constraints by designing lightweight CNN backbones and feature extractors optimized for on-device inference [329].

C. IMAGE SEGMENTATION

Semantic segmentation tasks require dense pixel-level labeling of image content. FCN and U-Net CNNs employ skip connections and encoder-decoder mirrors to preserve spatial information across resolutions [330], [331], [332], [333], [334], [335], [336], [337], [338], [339], [340], [341], [342], [343], [344], [345]. PSPNet and DeepLab introduce pyramid spatial pooling modules to capture multi-scale contextual cues [346]. GANs and conditional random fields further refine coarse segmentations from CNNs. Advances in medical imaging also apply segmentation CNNs to understand organ structures, localize pathologies, and aid diagnosis.

D. VISION TRANSFORMERS

Vision transformers have also emerged as a compelling alternative to traditional CNNs for CV tasks. Inspired by the success of language models, vision transformers divide images into discrete patches which are embedded and processed with self-attention. This allows them to capture long-range dependencies and multi-scale contextual information more effectively than CNNs. Models like ViT, DeiT, and Visual BERT demonstrate state-of-the-art results in tasks like image classification when pre-trained on large datasets [347], [348], [349], [350], [351], [352], [353], [354]. Research now focuses on optimizing transformer efficiency for real-time CV applications.

E. ONE-SHOT/FEW-SHOT/ZERO-SHOT LEARNING

One-shot and few-shot learning aim to address challenges posed by limited labeled training examples. Through metric learning and prototypical networks that learn robust representations from extensive base classes, models can effectively recognize new concepts from just one or a handful of examples without catastrophic forgetting [355], [356], [357], [358], [359], [360], [361], [362], [363], [364], [365], [366], [367], [368], [369]. This opens up CV to new long-tailed and incremental learning paradigms. Matching networks and prototypical networks efficiently compare test samples to prototype representations of base classes to generalize from limited exposures.

Zero-shot learning emerges as a promising area where CNNs imagine possibilities beyond the limitations of labeled data [370], [371], [372], [373], [374]. Descriptors like attributes or semantic relationships introduce inductive biases facilitating generalization without example. SAE, DeViSE, and contemporary models transfer knowledge by aligning embeddings between seen and unseen categories connected through auxiliary descriptors. Knowledge graphs also provide structural inductive biases through entity and relation modeling.

F. WEAKLY-SUPERVISED LEARNING

Weakly supervised learning techniques also help alleviate dependence on labor-intensive annotations [375], [376],

[377], [378], [379], [380]. Models can be trained end-to-end from weaker input signals like image-level tags or bounding box object locations instead of explicit pixel-level segmentation maps. Multi-instance learning approaches cluster image regions corresponding to each label to iteratively refine local predictions. Expectation-maximization (EM) and multiple instance learning jointly infer labels and recognize discriminative regions, enabling training from cheaper forms of weak supervision.

G. SELF-SUPERVISED/UNSUPERVISED LEARNING

Self-supervised learning has also gained vast attention in CV by enabling pre-training from sheer ubiquity of unlabeled visual data [381], [382], [383], [384], [385], [386], [387], [388], [389], [390]. Pretext tasks like predicting image rotations, solving jigsaw puzzles, or counting pixel colors allow models to learn rich visual representations applicable to downstream tasks. Recent contrastive self-supervised models like SimCLR, SwAV, and MoCo demonstrate that unlabeled pre-training rivals or exceeds supervised pre-training in various vision benchmarks, enabling more data-efficient fine-tuning or transfer to new problems.

H. LIFELONG/CONTINUAL LEARNING

Lifelong and continual learning aim to simulate open-world scenarios where models learn lifelong with non-stationary data distributions [51]. Models must avoid catastrophic forgetting when presented with new classes or shifts in existing class definitions without revisiting historical data [391], [392], [393], [394], [395], [396], [397], [398], [399], [400]. Elastic weight consolidation and incremental moment matching regularization preserve knowledge while accommodating new tasks. Research now explores task-aware architectures, dual-memory systems, and replay buffers that emulate memory reconsolidation to model lifelong visual learning.

I. VISION LANGUAGE MODEL

Vision-language models (VLMs) have also emerged at the intersection of NLP and CV by grounding language in visual contexts. Models fuse multimodal inputs through attention and generate captions conditioned on images, or localize and describe visual entities based on linguistic context. Large pre-trained models such as CLIP, ALIGN, and Oscar demonstrate exciting capabilities like zero-shot classification, question-answering (QA), and visual dialog with potential applications in education, assistive technologies, and more.

J. MEDICAL IMAGE ANALYSIS

Medical imaging epitomizes the necessity of collaboration between deep learning and domain experts. Segmenting organs in volumetric scans, localizing anomalies across imaging modalities, and tracking patients longitudinally all leverage 3D/2D CNNs [225], [227], [228], [230], [401], [402], [403], [404], [405], [406], [407], [408], [409], [410], [411]. Advanced models exploit anatomical priors by enforcing smoothness, and preservation of edges and surfaces in predictions. Self-supervision further enables pre-training from non-private data before fine-tuning target tasks. Model interpretation especially matters here to ensure trust among clinicians [407], [408], [409], [410]. Beyond diagnosis, CNNs can also simulate novel views to aid surgical planning. Efficiency additionally matters for on-device deployment and assisting underserved populations lacking infrastructure.

K. VIDEO UNDERSTANDING

Beyond images, video understanding presents unique challenges in modeling spatial-temporal relationships across consecutive frames. C3D and I3D CNNs introduce 3D convolutions directly learning from video volumes. Advanced techniques in video captioning and action recognition fuse language models and attention to jointly reason about visual content and linguistic semantics over time. Self-supervised learning from large unlabeled video repositories also emerges as a promising pretraining paradigm before fine-tuning downstream tasks.

L. MULTI-TASK LEARNING

Multi-task learning aims to improve generalization by jointly training CNNs on multiple related tasks using shared representations. This has proven successful across numerous applications by leveraging commonalities while mitigating overfitting individual tasks' limited data [412], [413], [414], [415]. For example, YOLO trains object detection alongside other auxiliary predictions like segmentation and counting.

Multi-task CNNs outperform independent models in low-data regimes (See Section VII -> Sub-section G.) by borrowing statistical strength across related problems. Dense captioning localizes objects and describes scenes simultaneously. A single network predicts keypoints, normals, and semantic part segmentation. Deeper tasks benefit substantially from representations learned for more general shallow tasks.

Progressively growing into new problem spaces via related auxiliary objectives also prevents catastrophic forgetting. Self-supervised pre-training establishes features broadly useful across downstream tasks, including those without annotations. Measuring and maximizing modularity in multi-task architectures additionally reduces interference between domains.

Techniques like multi-granularity, multi-level, and heterogeneous multi-task learning further craft diverse objectives to progressively refine semantics captured at differing levels of granularity [416], [417], [418], [419]. Task relations range from independent, and cooperative where tasks improve each other, to completely shared exploiting identical representations. Properly designed, multi-task CNNs deliver state-of-the-art performance while improving generalizability, efficiency, and real-world applicability.

Multi-task models combine CNNs with other modalities like language. For captioning, CNN-RNN fusion grounds generated text within visual contexts. For retrieval, ranking loss trains CNN-LSTM encoders to map semantically aligned vision-text pairs to nearby embeddings. Multi-modal pre-training on enormous unlabeled multimedia collections has proven highly beneficial via self-supervised alignment of domains.

M. 6D VISION

6D vision aims to recover the full 6D pose (3D position, 3D orientation) of objects directly from monocular RGB images. This is a challenging problem due to the loss of depth information when projecting 3D scenes onto 2D images [420], [421], [422], [423], [424], [425]. Early works relied on CAD models and rendered synthetic data which lacked photorealism, while more recent approaches leverage large amounts of real training data.

CNN-based regression networks are commonly used which take images as input and directly predict the 6D pose values. PoseCNN showed this can achieve competitive accuracy to model-based regression if trained end-to-end on real data. Due to the complex, multi-modal nature of the target distribution, losses that ensure consistent predictions under different poses like reprojection or angular are beneficial.

Iterative refinement approaches first detect the object, then iteratively update the pose estimate based on 2D-3D correspondences. DeepIM predicts shape coefficients and refines using PnP. DPOD leverages deep features combined with geometric constraints in a RANSAC framework. Dense representations also help by reasoning about object parts independently.

Multi-view and RGB-D sensors provide additional cues to leverage. MVD helps constrain the problem by training separate networks for each view and fusing results. Using both RGB and depth as input allows Depth-PoseNet to lift 2D predictions to 3D space. Multitask models predicting bounding boxes, keypoints, and poses jointly demonstrate accuracies approaching marker-based motion capture.

N. NEURAL ARCHITECTURE SEARCH

Neural architecture search (NAS) aims to automate the design of neural networks leveraging the power of evolution and reinforcement learning. Rather than relying on human experts to laboriously craft CNN architectures, NAS approaches evolve architectures directly on target datasets and tasks. This has led to state-of-the-art vision models developed without human design choices [426], [427], [428], [429], [430], [431], [432], [433].

Early NAS works explored various search spaces defined by units, operations, and connections between them. Combining concepts like pruning, sharing weights across child models during evolution helped scaling search to larger spaces [292], [296]. Performance predictors further reduced costs by guiding search towards promising regions. Novel methods evolved filters, activation functions, and batch normalization layers for particular domains.

Recent efforts evolve entire sections or blocks, expanding applicable search spaces. Single-path one-shot approaches drastically sped up search without compromising quality. ProxylessNAS found efficient mobile architectures directly on target devices. NAS approaches also discover non-CNN models suiting problems beyond CV.

Once identified, the best architectures can be trained from scratch to further improve upon proxy accuracies predicted during the search. Late phase evolution also enhances architectures initially identified, while architecture parameters themselves may evolve. Overall, NAS technologies continuously push forward state-of-the-art for vision tasks given diverse data, constraints, or objectives.

O. NEURAL ARCHITECTURE TRANSFORMER

Neural architecture transformers (NAT) replace CNNs' fixed topology with self-attention, replacing convolutional filtering with axial self-attention [429], [434]. This increased flexibility allows modeling long-range pixel dependencies crucial for vision tasks like segmentation. Vil-BERT introduced a multi-stage training procedure enabling pre-trained models to learn visual representations as well as natural language tasks.

Early works divided input images into uniform patches processed independently by attention layers. More sophisticated designs aim to capture visual locality through hierarchical patch divisions better. Rotary positional embeddings and attention patterns help encode translation equivariance. Architectures like CoAtNet cascade blocks with increased resolution, improving accuracy and interpretability.

Multi-scale vision transformers (MViT) incorporate prior convolutional inductive biases in hybrid models jointly benefiting from attention and translation equivariance. Combining vision transformers with convolutional networks particularly benefits medical image segmentation leveraging anatomical priors. Swin Transformers introduces a shifted window mechanism to focus computation locally across higher-resolution feature maps.

Though still an emerging direction, neural architecture transformers open new pathways for CV by bringing the full generality of self-attention to bear on visual problems. Their continued development will surely impact future CV research by unlocking novel representational abilities. Alongside NAS, they hold promise for pushing boundaries through data-driven discovery operating directly within much broader algorithmic search spaces.

P. GENERATIVE MODELS

Generative models have made large strides in the area of CV through techniques like GANs and diffusion models [435], [436], [437]. GANs pair a generator network against a discriminator network in an adversarial training procedure. This drives the generator to synthesize increasingly realistic fake images that can fool the discriminator.

GANs have produced impressive results generating photos that are near-indistinguishable from real images. Applications include image-to-image translation, super-resolution, and manipulating image attributes like style [437], [438], [439]. However, GAN training remains tricky to stabilize. Issues like mode collapse require careful architecture and hyperparameter choices.

Diffusion models provide an alternative generative framework gaining popularity. They utilize denoising diffusion probabilistic models (DDPMs) which gradually corrupt data with Gaussian noise before reversing the process [435], [436], [437], [439], [440], [441]. During generation, the model adds noise to a blank canvas and then predicts the noise-reduced output iteratively. This diffusion process proves more stable than adversarial training.

Sampling from DDPMs follows an ancestral sampling approach regressing the noise at each step conditioned on the previous denoised output. Advanced techniques like score-based sampling further improve sample quality by maximizing the model's density rather than following ancestral noise. Generative diffusion models (GDMs) also maximize a denoising score objective specifically for a generation [441].

Diffusion models have proven highly effective at synthesizing crisp, detailed images across varied datasets. Large-scale vision diffusion models (LVMs) like DALL-E 2 and DALL-E 3 demonstrate unparalleled capabilities of generating images from text prompts, and can even fuse language and vision to answer trivia questions about synthetic images.

By generating synthetic training data, generative models also benefit downstream classification, detection, and segmentation tasks through data augmentation. As generative diffusion models continue advancing, they will surely establish new frontiers in CV domains ranging from image editing to scientific discovery through computational experimentation.

Q. META LEARNING

Meta-learning, also known as learning to learn, aims to develop models that can rapidly adapt to new tasks and environments using only a few training examples. This is achieved by learning inductive biases about learning itself on a variety of related tasks during a meta-training phase. These biases are then leveraged during meta-test time on novel tasks [442], [443].

In CV, meta-learning enables CNNs to generalize beyond the restrictions of limited labeled examples through fast adaptation. Model-agnostic meta-learning (MAML) trains initial model parameters such that a few gradient steps fine-tune into new tasks. This learns efficient parameter initialization rather than solutions for any specific task [442], [443], [444], [445], [446], [447], [448], [449].

Metric-based approaches represent classes using prototypes that summarize inter/intra-class relationships independent of tasks [442], [443], [444]. Matching networks compare new examples to prototypes, providing fast adaptation through learned metric space similarities. Meta-Dataset consolidates many few-shot image classification datasets, advancing state-of-the-art and evaluation protocols in this challenging zero/few-shot regime [442], [443], [444], [449].

Self-supervised auxiliary tasks like prediction, rotation, and context modeling further enhance generalization when used alongside supervised meta-learning objectives. Temporal ensemble models aggregate diverse predictions over time from a generator network, improving robustness to noise and outliers. Reinforcement meta-learning successfully trains visuomotor policies for robotic control from only a handful of demonstrations.

R. FEDERATED LEARNING

Federated learning (FL) enables distributed training across decentralized edge devices without exchanging private user data like images, videos, or medical scans [81]. It aims to collaboratively learn a shared global model tailored to non-IID user distributions through coordinated local updates. This paradigm attracts increased interest due to growing concerns around data privacy and security.

FL trains a centralized CNN model through an iterative process where devices download the latest parameters, contribute updates computed over shards of local data, and then push weights back. A parameter server aggregates updates to globally improve the model. A key challenge arises from heterogeneity in non-IID data distributions, devices, and unreliable network connectivity. FedVision applies FL to object detection directly over fragmented client videos.

Techniques like personalized, multi-task, and meta-learning help address statistical heterogeneity in FL. Continual learning aspects prevent catastrophic forgetting when populations change over disseminated rounds. Differentially private algorithms and secure aggregation schemes ensure strong privacy in collaborative updates, advancing FL under stringent privacy constraints beyond vision to sensitive domains like healthcare.

XI. DISCUSSION

We have methodically explored the various CNN variations that have become more and more popular in recent years across a wide range of application sectors through this thorough survey. Our goal in this discussion part is to summarize the most significant findings from our evaluation of the literature and offer an analytical viewpoint on significant problems regarding the development and prospects of this area of study.

Convolutional layers are well-suited for grid-like data types, like images because they have proven highly capable of capturing spatial relationships and extracting hierarchical patterns. At the core of CNNs, commonly used for computer vision tasks such as object identification and image classification, remain traditional 2D convolutions.

However, as the field has evolved, additional specialized convolution approaches have emerged to handle different data modalities more effectively. One notable application of 1D convolutions is in sequential data domains including time series analysis and natural language processing. Their ability to capture temporal dependencies has enabled state-of-the-art accuracy on various language and audio processing problems. Likewise, 3D convolutions allow CNNs to effectively model volumetric medical images and video inputs by accounting for both spatial and temporal dimensions.

While basic convolution varieties such as 2D and 3D continue powering many top models, more efficient variants have also been developed. Dilated convolutions utilize dilations to widen receptive fields without loss of resolution, aiding high-level semantic tasks such as segmentation. Grouped convolutions offer a means of factorizing convolutions to dramatically reduce computation and memory usage, enabling large, deep architectures. However, their representational abilities may remain limited compared to standard convolutions for advanced analysis. Depthwise separable convolutions, as used in MobileNets, have achieved tremendous success in deploying efficient CNNs on embedded and mobile devices via their channel-wise decomposition.

In addition to novel convolution designs, the field is witnessing increasingly innovative integration of concepts from parallel research areas. For example, vision transformer models incorporate attention mechanisms to replace convolutional building blocks entirely, achieving strong results, especially on large datasets. Techniques like capsule networks aim to overcome CNN limitations through dynamic routing between feature vectors. Generative models such as Pix2Pix employ convolutional decoders to generate high-fidelity images from semantic maps or sketches. Advances in self-supervised learning provide alternative pretraining paradigms bypassing the need for vast annotated datasets.

Further combining of deep learning techniques seems poised to yield fruitful synergies. For instance, incorporating attention into convolutional pipelines could endow them with the benefits of both approaches. Moreover, self-supervised mechanisms may help the unsupervised discovery of interpretable convolutional filters well-suited to specific domains. Despite remarkable achievements, open challenges remain regarding robustness, sparse data scenarios, model interpretability, and trustworthiness. Future progress relies on close collaboration between academia and industry to define real-world needs and expand deep learning's positive societal impact.

Some convolution types have proven more enduring than others based on their flexibility and ability to adaptively fit diverse applications. While LeNet certainly played an instrumental pioneering role, more recent architectures better capture inherent data properties through principled network designs and optimizations. Meanwhile, innovation continues on all fronts, suggesting no single solution has emerged as definitive. Success hinges on judiciously combining innovations tailored to particular contexts rather than wholesale replacement of existing paradigms.

A promising outlook envisions continued refinement of core CNN building blocks and their harmonious integration with new algorithmic concepts from self-supervised learning, attention mechanisms, and generative models. In conclusion, this survey highlights both the remarkable advances of convolutional neural networks to date and their vast unrealized potential through the future intersection of ideas across deep learning's constantly evolving landscape.

XII. CONCLUSION

In this comprehensive study of different convolution types in deep learning, we have gained valuable insights into these techniques' diverse applications and strengths. CNNs have proven to be highly effective in various domains, ranging from image recognition to natural language processing. We compared various types of CNNS in various aspects, allowing us to understand their unique characteristics and advantages for specific tasks. Overall, this study emphasizes the importance of convolution in deep learning and its potential for future advances and improvements in artificial intelligence. Furthermore, the findings suggest that CNNs' versatility makes them suitable for various applications beyond traditional computer vision tasks. Furthermore, the study emphasizes the importance of additional research and development to optimize and refine these techniques for specific domains and tasks.

REFERENCES

[1] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[2] G. Hong, A. Folcarelli, J. Less, C. Wang, N. Erbasi, and S. Lin, "Voice assistants and cancer screening: A comparison of Alexa, Siri, Google Assistant, and Cortana," *Ann. Family Med.*, vol. 19, no. 5, pp. 447–449, Sep. 2021, doi: 10.1370/afm.2713.

[3] A. Kumar, S. Gadag, and U. Y. Nayak, "The beginning of a new era: Artificial intelligence in healthcare," *Adv. Pharmaceutical Bull.*, vol. 11, no. 3, pp. 414–425, Jul. 2020, doi: 10.34172/apb.2021.049.

[4] J. B. Heaton and N. Polson, "Deep learning for finance: Deep portfolios," *SSRN Electron. J.*, vol. 33, no. 1, pp. 3–12, Sep. 2016, doi: 10.2139/ssrn.2838013.

[5] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020, doi: 10.1109/TITS.2019.2929020.

[6] M. Ghaderzadeh and F. Asadi, "Deep learning in the detection and diagnosis of COVID-19 using radiology modalities: A systematic review," *J. Healthcare Eng.*, vol. 2021, pp. 1–10, Mar. 2021, doi: 10.1155/2021/6677314.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[9] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annu. Rev. Neurosci.*, vol. 19, no. 1, pp. 577–621, Mar. 1996, doi: 10.1146/annurev.ne.19.030196.003045.

[10] H. Zhou, X. Yu, A. Alhaskawi, Y. Dong, Z. Wang, Q. Jin, X. Hu, Z. Liu, V. G. Kota, M. H. A. H. Abdulla, S. H. A. Ezzi, B. Qi, J. Li, B. Wang, J. Fang, and H. Lu, "A deep learning approach for medical waste classification," *Sci. Rep.*, vol. 12, no. 1, p. 2159, Feb. 2022, doi: 10.1038/s41598-022-06146-2.

[11] S. Yang, A. Jin, and Y. Xu, "Recognition of oil and gas reservoir space based on deep learning," *E3S Web Conf.*, vol. 267, p. 01038, Jan. 2021, doi: 10.1051/e3sconf/202126701038.
[12] D. Jimenez-Carretero, V. Abrishami, L. Fernández-de-Manuel, I. Palacios, A. Quílez-Álvarez, A. Díez-Sánchez, M. A. del Pozo, and M. C. Montoya, "Tox_(R)CNN: Deep learning-based nuclei profiling tool for drug toxicity screening," *PLOS Comput. Biol.*, vol. 14, no. 11, Nov. 2018, Art. no. e1006238, doi: 10.1371/journal.pcbi.1006238.
[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[14] *TensorFlow Lite*. Accessed: Mar. 12, 2024. [Online]. Available: https://www.tensorflow.org/lite/
[15] T. Tran and R. Kavuluru, "An end-to-end deep learning architecture for extracting protein–protein interactions affected by genetic mutations," *Database*, vol. 2018, pp. 1–13, Jan. 2018, doi: 10.1093/database/bay092.
[16] A. Kamilaris and F. X. Prenafeta-Boldú, "A review of the use of convolutional neural networks in agriculture," *J. Agricult. Sci.*, vol. 156, no. 3, pp. 312–322, Apr. 2018, doi: 10.1017/s0021859618000436.
[17] S. Klos, J. Patalas-Maliszewska, and D. Tront, "A model for the intelligent supervision of production for Industry 4.0," *J. Phys., Conf. Ser.*, vol. 2198, no. 1, May 2022, Art. no. 012005.
[18] R. Nakajo, S. Murata, H. Arie, and T. Ogata, "Acquisition of viewpoint transformation and action mappings via sequence to sequence imitative learning by deep neural networks," *Frontiers Neurorobotics*, vol. 12, p. 46, Jul. 2018, doi: 10.3389/fnbot.2018.00046.
[19] V. D. Veksler, B. E. Hoffman, and N. Buchler, "Symbolic deep networks: A psychologically inspired lightweight and efficient approach to deep learning," *Topics Cogn. Sci.*, vol. 14, no. 4, pp. 702–717, Oct. 2021.
[20] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, and V. B. Kolachalama, "Detection of dementia on voice recordings using deep learning: A Framingham heart study," *Alzheimer's Res. Therapy*, vol. 13, no. 1, p. 146, Aug. 2021, doi: 10.1186/s13195-021-00888-3.
[21] C.-J. Lai, P.-F. Pai, M. Marvin, H.-H. Hung, S.-H. Wang, and D.-N. Chen, "The use of convolutional neural networks and digital camera images in cataract detection," *Electronics*, vol. 11, no. 6, p. 887, Mar. 2022, doi: 10.3390/electronics11060887.
[22] H. Maher Ahmed and M. Younis Kashmola, "A proposed architecture for convolutional neural networks to detect skin cancers," *IAES Int. J. Artif. Intell.*, vol. 11, no. 2, p. 485, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp485-493.
[23] D. Shen, X. Jiang, and L. Teng, "A novel gauss-laplace operator based on multi-scale convolution for dance motion image enhancement," *ICST Trans. Scalable Inf. Syst.*, vol. 9, Jul. 2018, Art. no. 172439, doi: 10.4108/eai.17-12-2021.172439.
[24] S. Sun, B. Hu, Z. Yu, and X. Song, "A stochastic max pooling strategy for convolutional neural network trained by noisy samples," *Int. J. Comput. Commun. Control*, vol. 15, no. 1, Feb. 2020, doi: 10.15837/ijccc.2020.1.3712.
[25] Y. Zhang, J. Hou, Q. Wang, A. Hou, and Y. Liu, "Application of transfer learning and feature fusion algorithms to improve the identification and prediction efficiency of premature ovarian failure," *J. Healthcare Eng.*, vol. 2022, pp. 1–10, Mar. 2022, doi: 10.1155/2022/3269692.
[26] A. Younesi, R. Afrouzian, and Y. Seyfari, "A transfer learning approach with convolutional neural network for face mask detection," 2023, *arXiv:2310.18928*.
[27] S. Ajala, H. Jalajamony, and R. Fernandez, "Deep-learning based estimation of dielectrophoretic force," *Micromachines*, vol. 13, no. 1, p. 41, Dec. 2021, doi: 10.3390/mi13010041.
[28] W.-J. Liang, H. Zhang, G.-F. Zhang, and H.-X. Cao, "Rice blast disease recognition using a deep convolutional neural network," *Sci. Rep.*, vol. 9, no. 1, Feb. 2019, doi: 10.1038/s41598-019-38966-0.
[29] G. Yang, X. Liang, S. Deng, and X. Chen, "Principal component research of the teaching model based on multimodal neural network algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, Jun. 2022, doi: 10.1155/2022/5888299.
[30] J. Rong, Y. Chen, and J. Yang, "CNN-LSTM hybrid model for kinematic feature analysis and parabolic radian prediction in basketball videos," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Sep. 2021, doi: 10.1155/2021/7844472.
[31] A. Tiwari, M. Silver, and A. Karnieli, "A deep learning approach for automatic identification of ancient agricultural water harvesting systems," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, Apr. 2023, Art. no. 103270, doi: 10.1016/j.jag.2023.103270.
[32] I. Taha Ahmed, B. Tareq Hammad, and N. Jamil, "A comparative analysis of image copy-move forgery detection algorithms based on hand and machine-crafted features," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, p. 1177, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1177-1190.
[33] E. Setyati, S. Az, S. P. Hudiono, and F. Kurniawan, "CNN based face recognition system for patients with down and William syndrome," *Knowl. Eng. Data Sci.*, vol. 4, no. 2, p. 138, Dec. 2021, doi: 10.17977/um018v4i22021p138-144.
[34] G. Hu, K. Wang, and L. Liu, "Underwater acoustic target recognition based on depthwise separable convolution neural networks," *Sensors*, vol. 21, no. 4, p. 1429, Feb. 2021, doi: 10.3390/s21041429.
[35] Z. Jiang, Z. Dong, L. Wang, and W. Jiang, "Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–12, Aug. 2021, doi: 10.1155/2021/7529893.
[36] L. Wang, W. Zhou, Q. Chang, J. Chen, and X. Zhou, "Deep ensemble detection of congestive heart failure using short-term RR intervals," *IEEE Access*, vol. 7, pp. 69559–69574, 2019, doi: 10.1109/ACCESS.2019.2912226.
[37] A. Zafar, M. Aamir, N. M. Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, "A comparison of pooling methods for convolutional neural networks," *Appl. Sci.*, vol. 12, no. 17, p. 8643, Aug. 2022, doi: 10.3390/app12178643.
[38] J. Zhu, J. Jang-Jaccard, and P. A. Watters, "Multi-loss Siamese neural network with batch normalization layer for malware detection," *IEEE Access*, vol. 8, pp. 171542–171550, 2020, doi: 10.1109/ACCESS.2020.3024991.
[39] A. Maniatopoulos and N. Mitianoudis, "Learnable leaky ReLU (LeLeLU): An alternative accuracy-optimized activation function," *Information*, vol. 12, no. 12, p. 513, Dec. 2021, doi: 10.3390/info12120513.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.
[41] L. E. van Dyck, R. Kwitt, S. J. Denzler, and W. R. Gruber, "Comparing object recognition in humans and deep convolutional neural networks—An eye tracking study," *Frontiers Neurosci.*, vol. 15, Oct. 2021, Art. no. 750639, doi: 10.3389/fnins.2021.750639.
[42] Z. Wu, Y. Guo, W. Lin, S. Yu, and Y. Ji, "A weighted deep representation learning model for imbalanced fault diagnosis in cyber-physical systems," *Sensors*, vol. 18, no. 4, p. 1096, Apr. 2018, doi: 10.3390/s18041096.
[43] H. Xia, C. Ding, and Y. Liu, "Sentiment analysis model based on self-attention and character-level embedding," *IEEE Access*, vol. 8, pp. 184614–184620, 2020, doi: 10.1109/ACCESS.2020.3029694.
[44] Z. Zhao, X. Xie, W. Liu, and Q. Pan, "A hybrid-3D convolutional network for video compressive sensing," *IEEE Access*, vol. 8, pp. 20503–20513, 2020, doi: 10.1109/ACCESS.2020.2969290.
[45] Z. Liu, L. Tong, L. Chen, Z. Jiang, F. Zhou, Q. Zhang, X. Zhang, Y. Jin, and H. Zhou, "Deep learning based brain tumor segmentation: A survey," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 1001–1026, Jul. 2022, doi: 10.1007/s40747-022-00815-5.
[46] H. Chen, C. Hu, F. Lee, C. Lin, W. Yao, L. Chen, and Q. Chen, "A supervised video hashing method based on a deep 3D convolutional neural network for large-scale video retrieval," *Sensors*, vol. 21, no. 9, p. 3094, Apr. 2021, doi: 10.3390/s21093094.
[47] G. Li, S. Jiang, I. Yun, J. Kim, and J. Kim, "Depth-wise asymmetric bottleneck with point-wise aggregation decoder for real-time semantic segmentation in urban scenes," *IEEE Access*, vol. 8, pp. 27495–27506, 2020, doi: 10.1109/ACCESS.2020.2971760.
[48] Z. Lu, Z. Yu, P. Yali, L. Shigang, W. Xiaojun, L. Gang, and R. Yuan, "Fast single image super-resolution via dilated residual networks," *IEEE Access*, vol. 7, pp. 109729–109738, 2019, doi: 10.1109/ACCESS.2018.2865613.
[49] R. Kolaghassi, M. K. Al-Hares, and K. Sirlantzis, "Systematic review of intelligent algorithms in gait analysis and prediction for lower limb robotic systems," *IEEE Access*, vol. 9, pp. 113788–113812, 2021, doi: 10.1109/ACCESS.2021.3104464.
[50] M. Capra, B. Bussolino, A. Marchisio, G. Masera, M. Martina, and M. Shafique, "Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead," *IEEE Access*, vol. 8, pp. 225134–225180, 2020, doi: 10.1109/ACCESS.2020.3039858.

[51] K. Shaheen, M. A. Hanif, O. Hasan, and M. Shafique, "Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks," *J. Intell. Robotic Syst.*, vol. 105, no. 1, p. 9, Apr. 2022, doi: 10.1007/s10846-022-01603-6.

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[55] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient CNNs for mobile vision applications," 2017, *arXiv:1704.04861*.

[56] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for CNNs," in *Proc. ICML*, 2019, pp. 6105–6114.

[57] B. Jin, L. Cruz, and N. Gonçalves, "Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123649–123661, 2020, doi: 10.1109/ACCESS.2020.3005687.

[58] Y. Yang, D. Kim, and B. T. Oh, "Deep convolutional grid warping network for joint depth map upsampling," *IEEE Access*, vol. 8, pp. 147580–147590, 2020, doi: 10.1109/ACCESS.2020.3015209.

[59] P. Lang, X. Fu, C. Feng, J. Dong, R. Qin, and M. Martorella, "LW-CMDANet: A novel attention network for SAR automatic target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6615–6630, 2022, doi: 10.1109/JSTARS.2022.3195074.

[60] L. Li, J. Wang, and X. Li, "Efficiency analysis of machine learning intelligent investment based on K-means algorithm," *IEEE Access*, vol. 8, pp. 147463–147470, 2020, doi: 10.1109/ACCESS.2020.3011366.

[61] H. Tomosada, T. Kudo, T. Fujisawa, and M. Ikehara, "GAN-based image deblurring using DCT loss with customized datasets," *IEEE Access*, vol. 9, pp. 135224–135233, 2021, doi: 10.1109/ACCESS.2021.3116194.

[62] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, "Pretraining is all you need for image-to-image translation," 2022, *arXiv:2205.12952*.

[63] B. Kaddar, H. Fizazi, M. Hernández-Cabronero, V. Sanchez, and J. Serra-Sagristà, "DivNet: Efficient convolutional neural network via multilevel hierarchical architecture design," *IEEE Access*, vol. 9, pp. 105892–105901, 2021, doi: 10.1109/ACCESS.2021.3099952.

[64] R. Yu and J. Sun, "Learning polynomial-based separable convolution for 3D point cloud analysis," *Sensors*, vol. 21, no. 12, p. 4211, Jun. 2021, doi: 10.3390/s21124211.

[65] A. Kara, "A hybrid prognostic approach based on deep learning for the degradation prediction of machinery," *Sakarya Univ. J. Comput. Inf. Sci.*, vol. 4, no. 2, pp. 216–226, Aug. 2021, doi: 10.35377/saucis.04.02.912154.

[66] X. Zhang, J. Zhou, W. Sun, and S. Kumar Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Comput., Mater. Continua*, vol. 72, no. 1, pp. 1123–1137, Jan. 2022, doi: 10.32604/cmc.2022.024589.

[67] F. Sultonov, J.-H. Park, S. Yun, D.-W. Lim, and J.-M. Kang, "Mixer U-Net: An improved automatic road extraction from UAV imagery," *Appl. Sci.*, vol. 12, no. 4, p. 1953, Feb. 2022, doi: 10.3390/app12041953.

[68] J. Shao, C. Qu, J. Li, and S. Peng, "A lightweight convolutional neural network based on visual attention for SAR image target classification," *Sensors*, vol. 18, no. 9, p. 3039, Sep. 2018, doi: 10.3390/s18093039.

[69] D. Al-Alimi, Y. Shao, R. Feng, M. A. A. Al-Qaness, M. A. Elaziz, and S. Kim, "Multi-scale geospatial object detection based on shallow-deep feature extraction," *Remote Sens.*, vol. 11, no. 21, p. 2525, Oct. 2019, doi: 10.3390/rs11212525.

[70] X. Yang, A. Chen, G. Zhou, J. Wang, W. Chen, Y. Gao, and R. Jiang, "Instance segmentation and classification method for plant leaf images based on ISC-MRCNN and APS-DCCNN," *IEEE Access*, vol. 8, pp. 151555–151573, 2020, doi: 10.1109/ACCESS.2020.3017560.

[71] Q. Liu, L. Han, R. Tan, H. Fan, W. Li, H. Zhu, B. Du, and S. Liu, "Hybrid attention based residual network for pansharpening," *Remote Sens.*, vol. 13, no. 10, p. 1962, May 2021, doi: 10.3390/rs13101962.

[72] A. Ju and Z. Wang, "Convolutional block attention module based on visual mechanism for robot image edge detection," *ICST Trans. Scalable Inf. Syst.*, vol. 9, Jul. 2018, Art. no. 172214, doi: 10.4108/eai.19-11-2021.172214.

[73] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020.

[74] Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, and S.-C. Zhu, "Interpretable CNNs for object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3416–3431, Oct. 2021, doi: 10.1109/TPAMI.2020.2982882.

[75] M. K. Patrick, A. Felix Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks—A survey," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 1, pp. 1295–1310, Jan. 2022, doi: 10.1016/j.jksuci.2019.09.014.

[76] C. White, M. Safari, R. Sukthanker, B. Ru, T. Elsken, A. Zela, D. Dey, and F. Hutter, "Neural architecture search: Insights from 1000 papers," 2023, *arXiv:2301.08727*.

[77] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[78] X. Yuan, Z. Feng, M. Norton, and X. Li, "Generalized batch normalization: Towards accelerating deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 1682–1689, doi: 10.1609/aaai.v33i01.33011682.

[79] Y. Wang, G. Wang, C. Chen, and Z. Pan, "Multi-scale dilated convolution of convolutional neural network for image denoising," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19945–19960, Feb. 2019, doi: 10.1007/s11042-019-7377-y.

[80] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," 2016, *arXiv:1611.10012*.

[81] H. Zhu, H. Zhang, and Y. Jin, "From federated learning to federated neural architecture search: A survey," *Complex Intell. Syst.*, vol. 7, no. 2, pp. 639–657, Jan. 2021, doi: 10.1007/s40747-020-00247-z.

[82] O. N. Oyelade and A. E. Ezugwu, "A bioinspired neural architecture search based convolutional neural network for breast cancer detection using histopathology images," *Sci. Rep.*, vol. 11, no. 1, Oct. 2021, doi: 10.1038/s41598-021-98978-7.

[83] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3648–3657, doi: 10.1109/CVPR.2019.00377.

[84] A. Singh, V. Jaiswal, G. Joshi, A. Sanjeeve, S. Gite, and K. Kotecha, "Neural style transfer: A critical review," *IEEE Access*, vol. 9, pp. 131583–131613, 2021, doi: 10.1109/ACCESS.2021.3112996.

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[86] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[87] A. Khan, Z. Rauf, A. Sohail, A. Rehman, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and its CNN-transformer based variants," 2023, *arXiv:2305.09880*.

[88] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[89] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.

[90] M. Hnewa and H. Radha, "Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 53–67, Jan. 2021, doi: 10.1109/MSP.2020.2984801.

[91] M. Elhoseny, "Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems," *Circuits, Syst., Signal Process.*, vol. 39, no. 2, pp. 611–630, Aug. 2019, doi: 10.1007/s00034-019-01234-7.

[92] S. Thakur and A. Kumar, "X-ray and CT-scan-based automated detection and classification of COVID-19 using convolutional neural networks (CNN)," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102920, doi: 10.1016/j.bspc.2021.102920.

[93] F. Ramzan, M. U. G. Khan, S. Iqbal, T. Saba, and A. Rehman, "Volumetric segmentation of brain regions from MRI scans using 3D convolutional neural networks," *IEEE Access*, vol. 8, pp. 103697–103709, 2020, doi: 10.1109/ACCESS.2020.2998901.

[94] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.

[95] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[96] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie, "UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3042–3051, doi: 10.1109/cvpr52688.2022.00305.

[97] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.

[98] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[99] J.-B. Alayrac. A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. D. Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-Supervised MultiModal versatile networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jun. 2020, pp. 25–37. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf

[100] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics: Polymers of Hexadromicon*, vol. 3, 2nd ed., J. Peters, Eds. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64. [Online]. Available: http://www.bookref.com

[101] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," 2017, *arXiv:1711.06396*.

[102] Y. He and L. Xiao, "Structured pruning for deep convolutional neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 28, 2023, doi: 10.1109/TPAMI.2023.3334614.

[103] I. Oguntola, S. Olubeko, and C. Sweeney, "SlimNets: An exploration of deep model compression and acceleration," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Waltham, MA, USA, Sep. 2018, pp. 1–6, doi: 10.1109/HPEC.2018.8547604.

[104] T. Guo, T. Zhang, E. Lim, M. López-Benítez, F. Ma, and L. Yu, "A review of wavelet analysis and its applications: Challenges and opportunities," *IEEE Access*, vol. 10, pp. 58869–58903, 2022, doi: 10.1109/ACCESS.2022.3179517.

[105] G. Othman and D. Q. Zeebaree, "The applications of discrete wavelet transform in image processing: A review," *J. Soft Comput. Data Mining*, vol. 1, no. 2, pp. 31–43, 2020.

[106] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *J. Artif. Intell. Syst.*, vol. 2, no. 1, pp. 53–79, 2020, doi: 10.33969/ais.2020.21005.

[107] T. Williams and R. Li, "Advanced image classification using wavelets and convolutional neural networks," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Anaheim, CA, USA, Dec. 2016, pp. 233–239, doi: 10.1109/ICMLA.2016.0046.

[108] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level wavelet convolutional neural networks," *IEEE Access*, vol. 7, pp. 74973–74985, 2019, doi: 10.1109/ACCESS.2019.2921451.

[109] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," 2018, *arXiv:1805.08620*.

[110] W. Zhang, K. Jiang, L. Wang, N. Meng, Y. Zhou, Y. Li, H. Hu, X. Chen, and B. Jiang, "A wavelet-based asymmetric convolution network for single image super-resolution," *IEEE Access*, vol. 9, pp. 28976–28986, 2021, doi: 10.1109/ACCESS.2021.3058648.

[111] Y. Wu, P. Qian, and X. Zhang, "Two-level wavelet-based convolutional neural network for image deblurring," *IEEE Access*, vol. 9, pp. 45853–45863, 2021, doi: 10.1109/ACCESS.2021.3067055.

[112] Z. Tao, T. Wei, and J. Li, "Wavelet multi-level attention capsule network for texture classification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1215–1219, 2021, doi: 10.1109/LSP.2021.3088052.

[113] M. C. Kim, J. H. Park, and M. H. Sunwoo, "Multilevel feature extraction using wavelet attention for deep joint demosaicking and denoising," *IEEE Access*, vol. 10, pp. 77099–77109, 2022, doi: 10.1109/ACCESS.2022.3192451.

[114] Z. Xie, Z. Wen, J. Liu, Z. Liu, X. Wu, and M. Tan, "Deep transferring quantization," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, Jan. 2020, pp. 625–642, doi: 10.1007/978-3-030-58598-3.

[115] H. Li, X. Wu, F. Lv, D. Liao, T. H. Li, Y. Zhang, B. Han, and M. Tan, "Hard sample matters a lot in zero-shot quantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24417–24426, doi: 10.1109/cvpr52729.2023.02339.

[116] M. Lin, R. Ji, Y. Wang, Y. Zhang, B. Zhang, Y. Tian, and L. Shao, "HRank: Filter pruning using high-rank feature map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1526–1535, doi: 10.1109/CVPR42600.2020.00160.

[117] M. Krichen, "Convolutional neural networks: A survey," *Computers*, vol. 12, no. 8, p. 151, Jul. 2023, doi: 10.3390/computers12080151.

[118] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[119] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[120] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/s10462-020-09825-6.

[121] S. Xu, A. Huang, L. Chen, and B. Zhang, "Convolutional neural network pruning: A survey," in *Proc. 39th Chin. Control Conf. (CCC)*, Shenyang, China, Jul. 2020, pp. 7458–7463, doi: 10.23919/CCC50068.2020.9189610.

[122] U. Kulkarni, S. S. Hallad, A. Patil, T. Bhujannavar, S. Kulkarni, and S. M. Meena, "A survey on filter pruning techniques for optimization of deep neural networks," in *Proc. 6th Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud) (I-SMAC)*, Dharan, Nepal, Nov. 2022, pp. 610–617, doi: 10.1109/I-SMAC55078.2022.9987264.

[123] S. Vadera and S. Ameen, "Methods for pruning deep neural networks," *IEEE Access*, vol. 10, pp. 63280–63300, 2022, doi: 10.1109/ACCESS.2022.3182659.

[124] A. R. Aswani, R. Chithra, and A. P. James, "Unstructured weight pruning in variability-aware memristive crossbar neural networks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Austin, TX, USA, May 2022, pp. 3458–3462, doi: 10.1109/ISCAS48785.2022.9937284.

[125] A. Liang, H. Zhang, H. Hua, and W. Chen, "To drop or to select: Reduce the negative effects of disturbance features for point cloud classification from an interpretable perspective," *IEEE Access*, vol. 11, pp. 36184–36202, 2023, doi: 10.1109/ACCESS.2023.3266340.

[126] Y. Peng, M. Chang, Q. Wang, Y. Qian, Y. Zhang, M. Wei, and X. Liao, "Sparse-to-dense multi-encoder shape completion of unstructured point cloud," *IEEE Access*, vol. 8, pp. 30969–30978, 2020, doi: 10.1109/ACCESS.2020.2973003.

[127] A. Zhang, S. Li, J. Wu, S. Li, and B. Zhang, "Exploring semantic information extraction from different data forms in 3D point cloud semantic segmentation," *IEEE Access*, vol. 11, pp. 61929–61949, 2023, doi: 10.1109/ACCESS.2023.3287940.

[128] Y. Wang, X. Tang, and C. Yue, "Enhancing the local graph semantic feature for 3D point cloud classification and segmentation," *IEEE Access*, vol. 10, pp. 74620–74628, 2022, doi: 10.1109/ACCESS.2022.3190966.

[129] J. Zeng, D. Wang, and P. Chen, "A survey on transformers for point cloud processing: An updated overview," *IEEE Access*, vol. 10, pp. 86510–86527, 2022, doi: 10.1109/ACCESS.2022.3198999.

[130] Berkeley Vision. (2012). *Caffe | Deep Learning Framework*. [Online]. Available: https://caffe.berkeleyvision.org/

[131] PyTorch Organisation. (2023). *PyTorch*. [Online]. Available: https://pytorch.org/

[132] TensorFlow. (2019). *TensorFlow*. [Online]. Available: https://www.tensorflow.org/

[133] Keras. (2019). *Home—Keras Documentation*. [Online]. Available: https://keras.io/

[134] OpenCV. (2019). *OpenCV Library*. [Online]. Available: https://opencv.org/

[135] GitHub. (Jan. 9, 2024). *Apache/MXNet*. Accessed: Jan. 09, 2024. [Online]. Available: https://github.com/apache/mxnet

[136] Chainer. *Chainer: A Flexible Framework for Neural Networks*. Accessed: Mar. 12, 2024. [Online]. Available: https://chainer.org/

[137] *Eclipse DeepLearning4J*. Accessed: Mar. 12, 2024. [Online]. Available: https://deeplearning4j.konduit.ai/

[138] F. Ullah, I. Ullah, R. U. Khan, S. Khan, K. Khan, and G. Pau, "Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3878–3916, 2024, doi: 10.1109/JSTARS.2024.3353551.

[139] V. A. Ashwath, O. K. Sikha, and R. Benitez, "TS-CNN: A three-tier self-interpretable CNN for multi-region medical image classification," *IEEE Access*, vol. 11, pp. 78402–78418, 2023, doi: 10.1109/ACCESS.2023.3299850.

[140] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019, doi: 10.1109/LGRS.2019.2911322.

[141] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of image degradation and degradation removal to CNN-based image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1239–1253, Apr. 2021, doi: 10.1109/TPAMI.2019.2950923.

[142] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, "A deep multi-modal CNN for multi-instance multi-label image classification," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 6025–6038, Dec. 2018, doi: 10.1109/TIP.2018.2864920.

[143] C. Shi, L. Fang, and H. Shen, "Convolutional neural networks with class-driven loss for multiscale VHR remote sensing image classification," *IEEE Access*, vol. 8, pp. 149162–149175, 2020, doi: 10.1109/ACCESS.2020.3014975.

[144] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2536–2551, 2023, doi: 10.1109/TIP.2023.3270104.

[145] S. Khan, M. Sajjad, T. Hussain, A. Ullah, and A. S. Imran, "A review on traditional machine learning and deep learning models for WBCs classification in blood smear images," *IEEE Access*, vol. 9, pp. 10657–10673, 2021, doi: 10.1109/ACCESS.2020.3048172.

[146] S. A. H. Minoofam, A. Bastanfard, and M. R. Keyvanpour, "TRCLA: A transfer learning approach to reduce negative transfer for cellular learning automata," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 5, pp. 2480–2489, May 2023, doi: 10.1109/TNNLS.2021.3106705.

[147] J. Hao, "Deep learning-based medical image analysis with explainable transfer learning," in *Proc. Int. Conf. Comput. Eng. Distance Learn. (CEDL)*, Shanghai, China, Jun. 2023, pp. 106–109, doi: 10.1109/CEDL60560.2023.00029.

[148] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.

[149] E. Chalmers, E. B. Contreras, B. Robertson, A. Luczak, and A. Gruber, "Learning to predict consequences as a method of knowledge transfer in reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2259–2270, Jun. 2018, doi: 10.1109/TNNLS.2017.2690910.

[150] T. V. Phan, S. Sultana, T. G. Nguyen, and T. Bauschert, "*Q*-TRANSFER: A novel framework for efficient deep transfer learning in networking," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Fukuoka, Japan, Feb. 2020, pp. 146–151, doi: 10.1109/ICAIIC48513.2020.9065240.

[151] T. T. Chungath, A. M. Nambiar, and A. Mittal, "Transfer learning and few-shot learning based deep neural network models for underwater sonar image classification with a few samples," *IEEE J. Ocean. Eng.*, vol. 49, no. 1, pp. 294–310, Jan. 2024, doi: 10.1109/JOE.2022.3221127.

[152] A. M. Nagib, H. Abou-Zeid, and H. S. Hassanein, "Safe and accelerated deep reinforcement learning-based O-RAN slicing: A hybrid transfer learning approach," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, pp. 310–325, Feb. 2024, doi: 10.1109/JSAC.2023.3336191.

[153] M. Biehler, Y. Sun, S. Kode, J. Li, and J. Shi, "PLURAL: 3D point cloud transfer learning via contrastive learning with augmentations," *IEEE Trans. Autom. Sci. Eng.*, early access, 2004, doi: 10.1109/TASE.2023.3345807.

[154] H. Li, Z. Wang, C. Lan, P. Wu, and N. Zeng, "A novel dynamic multiobjective optimization algorithm with non-inductive transfer learning based on multi-strategy adaptive selection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2024, doi: 10.1109/TNNLS.2023.3295461.

[155] H. Chen, H. Luo, B. Huang, B. Jiang, and O. Kaynak, "Transfer learning-motivated intelligent fault diagnosis designs: A survey, insights, and perspectives," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 2969–2983, Mar. 2024, doi: 10.1109/TNNLS.2023.3290974.

[156] H. S. Mputu, A. Abdel-Mawgood, A. Shimada, and M. S. Sayed, "Tomato quality classification based on transfer learning feature extraction and machine learning algorithm classifiers," *IEEE Access*, vol. 12, pp. 8283–8295, 2024, doi: 10.1109/ACCESS.2024.3352745.

[157] T. Zhou, F. Zhang, K. Shao, Z. Dai, K. Li, W. Huang, W. Wang, B. Wang, D. Li, W. Liu, and J. Hao, "Cooperative multi-agent transfer learning with coalition pattern decomposition," *IEEE Trans. Games*, early access, 2023, doi: 10.1109/TG.2023.3272386.

[158] L. He, Q. Wei, M. Gong, X. Yang, and J. Wei, "Transfer learning-based center-of-mass positioning methods for cultural relics," *IEEE Access*, vol. 12, pp. 7911–7926, 2024, doi: 10.1109/ACCESS.2023.3349017.

[159] M. S. Azari, F. Flammini, S. Santini, and M. Caporuscio, "A systematic literature review on transfer learning for predictive maintenance in Industry 4.0," *IEEE Access*, vol. 11, pp. 12887–12910, 2023, doi: 10.1109/ACCESS.2023.3239784.

[160] D. Onita, "Active learning based on transfer learning techniques for text classification," *IEEE Access*, vol. 11, pp. 28751–28761, 2023, doi: 10.1109/ACCESS.2023.3260771.

[161] Q. Li, Z. Yang, L. Luo, L. Wang, Y. Zhang, H. Lin, J. Wang, L. Yang, K. Xu, and Y. Zhang, "A multi-task learning based approach to biomedical entity relation extraction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Madrid, Spain, Dec. 2018, pp. 680–682, doi: 10.1109/BIBM.2018.8621284.

[162] H. Li and J. Qi, "A multi-task learning and data augmentation-based pose estimation algorithm," in *Proc. 8th Int. Conf. Inf. Syst. Eng. (ICISE)*, Dalian, China, 2023, pp. 358–361, doi: 10.1109/ICISE60366.2023.00082.

[163] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020, doi: 10.1109/ACCESS.2020.2989428.

[164] S. Liu, F. Yang, F. Kang, and J. Yang, "A multi-task learning method for weakly supervised sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 8802–8806, doi: 10.1109/ICASSP43922.2022.9746947.

[165] X. Ouyang, S. Xu, C. Zhang, P. Zhou, Y. Yang, G. Liu, and X. Li, "A 3D-CNN and LSTM based multi-task learning architecture for action recognition," *IEEE Access*, vol. 7, pp. 40757–40770, 2019, doi: 10.1109/ACCESS.2019.2906654.

[166] L. Yunxiang and Z. Kexin, "Design of efficient speech emotion recognition based on multi task learning," *IEEE Access*, vol. 11, pp. 5528–5537, 2023, doi: 10.1109/ACCESS.2023.3237268.

[167] Q. Zhou and Q. Zhao, "Flexible clustered multi-task learning by learning representative tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 266–278, Feb. 2016, doi: 10.1109/TPAMI.2015.2452911.

[168] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1070–1083, Jun. 2016, doi: 10.1109/TPAMI.2015.2477843.

[169] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017, doi: 10.1109/TPAMI.2016.2537337.

[170] Q. Chen, W. Liu, and X. Yu, "A viewpoint aware multi-task learning framework for fine-grained vehicle recognition," *IEEE Access*, vol. 8, pp. 171912–171923, 2020, doi: 10.1109/ACCESS.2020.3024658.

[171] G. Buroni, B. Lebichot, and G. Bontempi, "AST-MTL: An attention-based multi-task learning strategy for traffic forecasting," *IEEE Access*, vol. 9, pp. 77359–77370, 2021, doi: 10.1109/ACCESS.2021.3083412.

[172] C. Ding, Z. Lu, S. Wang, R. Cheng, and V. N. Boddeti, "Mitigating task interference in multi-task learning via explicit task routing with non-learnable primitives," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7756–7765, doi: 10.1109/cvpr52729.2023.00749.

[173] W. Choi and S. Im, "Dynamic neural network for multi-task learning searching across diverse network topologies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 3779–3788, doi: 10.1109/cvpr52729.2023.00368.

[174] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019, doi: 10.1109/TPAMI.2017.2781233.

[175] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018, doi: 10.1109/TPAMI.2017.2738004.

[176] G. He, Y. Huo, M. He, H. Zhang, and J. Fan, "A novel orthogonality loss for deep hierarchical multi-task learning," *IEEE Access*, vol. 8, pp. 67735–67744, 2020, doi: 10.1109/ACCESS.2020.2985991.

[177] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[178] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.

[179] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016, doi: 10.1109/TPAMI.2015.2437384.

[180] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.

[181] Z. Wu, J. Wen, Y. Xu, J. Yang, X. Li, and D. Zhang, "Enhanced spatial feature learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 961–972, Jan. 2024, doi: 10.1109/TNNLS.2022.3178180.

[182] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/ACCESS.2019.2939201.

[183] J. Kang, S. Tariq, H. Oh, and S. S. Woo, "A survey of deep learning-based object detection methods and datasets for overhead imagery," *IEEE Access*, vol. 10, pp. 20118–20134, 2022, doi: 10.1109/ACCESS.2022.3149052.

[184] S. Hoque, Md. Y. Arafat, S. Xu, A. Maiti, and Y. Wei, "A comprehensive review on 3D object detection and 6D pose estimation with deep learning," *IEEE Access*, vol. 9, pp. 143746–143770, 2021, doi: 10.1109/ACCESS.2021.3114399.

[185] Y.-L. Li and S. Wang, "HAR-Net: Joint learning of hybrid attention for single-stage object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3092–3103, 2020, doi: 10.1109/TIP.2019.2957850.

[186] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3D LiDAR-based video object detection for autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2068–2078, Apr. 2022, doi: 10.1109/TCSVT.2021.3082763.

[187] H. Ibrahem, A. D. A. Salem, and H.-S. Kang, "Real-time weakly supervised object detection using center-of-features localization," *IEEE Access*, vol. 9, pp. 38742–38756, 2021, doi: 10.1109/ACCESS.2021.3064372.

[188] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023, doi: 10.1109/ACCESS.2023.3266093.

[189] H. Wang, Q. Wang, H. Zhang, Q. Hu, and W. Zuo, "CrabNet: Fully task-specific feature learning for one-stage object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2962–2974, 2022, doi: 10.1109/TIP.2022.3162099.

[190] T. Gao, H. Pan, and H. Gao, "Monocular 3D object detection with sequential feature association and depth hint augmentation," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 2, pp. 240–250, Jun. 2022, doi: 10.1109/TIV.2022.3143954.

[191] Z. Zhang and T. D. Bui, "Attention-based selection strategy for weakly supervised object localization," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 10305–10311, doi: 10.1109/ICPR48806.2021.9412173.

[192] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou, and S. Cui, "Shallow feature matters for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 5989–5997, doi: 10.1109/CVPR46437.2021.00593.

[193] L. Zhu, Q. She, Q. Chen, Y. You, B. Wang, and Y. Lu, "Weakly supervised object localization as domain adaption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 14617–14626, doi: 10.1109/CVPR52688.2022.01423.

[194] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 2866–2875, doi: 10.1109/ICCV48922.2021.00288.

[195] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, and C. Xu, "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 11637–11646, doi: 10.1109/CVPR46437.2021.01147.

[196] G. Guo, J. Han, F. Wan, and D. Zhang, "Strengthen learning tolerance for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 7399–7408, doi: 10.1109/CVPR46437.2021.00732.

[197] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, and L. Shen, "Online refinement of low-level feature based activation map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 132–141, doi: 10.1109/ICCV48922.2021.00020.

[198] J. Xu, J. Hou, Y. Zhang, R. Feng, R. W. Zhao, T. Zhang, X. Lu, and S. Gao, "CREAM: Weakly supervised object localization via class RE-activation mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 9427–9436, doi: 10.1109/CVPR52688.2022.00922.

[199] Z. Min, B. Zhuang, S. Schulter, B. Liu, E. Dunn, and M. Chandraker, "NeurOCS: Neural NOCS supervision for monocular 3D object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 21404–21414, doi: 10.1109/cvpr52729.2023.02050.

[200] X. Yu, P. Chen, D. Wu, N. Hassan, G. Li, J. Yan, H. Shi, Q. Ye, and Z. Han, "Object localization under single coarse point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4858–4867, doi: 10.1109/cvpr52688.2022.00482.

[201] V. Gaudillière, L. Pauly, A. Rathinam, A. G. Sanchez, M. A. Musallam, and D. Aouada, "3D-aware object localization using Gaussian implicit occupancy function," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Detroit, MI, USA, Oct. 2023, pp. 5858–5863, doi: 10.1109/iros55552.2023.10342399.

[202] G. Lv, Y. Sun, F. Nian, M. Zhu, W. Tang, and Z. Hu, "COME: Clip-OCR and master object for text image captioning," *Image Vis. Comput.*, vol. 136, Aug. 2023, Art. no. 104751, doi: 10.1016/j.imavis.2023.104751.

[203] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "OCR binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397, Feb. 2007, doi: 10.1016/j.patcog.2006.04.043.

[204] I. Bazzi, R. Schwartz, and J. Makhoul, "An omnifont open-vocabulary OCR system for English and Arabic," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 6, pp. 495–504, Jun. 1999, doi: 10.1109/34.771314.

[205] J. Park, V. Govindaraju, and S. N. Srihari, "OCR in a hierarchical feature space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 400–407, Apr. 2000, doi: 10.1109/34.845383.

[206] G. Nagy, S. Seth, and K. Einspahr, "Decoding substitution ciphers by means of word matching with application to OCR," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 710–715, Sep. 1987, doi: 10.1109/TPAMI.1987.4767969.

[207] A. Y. Sugiyono, K. Adrio, K. Tanuwijaya, and K. M. Suryaningrum, "Extracting information from vehicle registration plate using OCR tesseract," *Proc. Comput. Sci.*, vol. 227, pp. 932–938, Jan. 2023, doi: 10.1016/j.procs.2023.10.600.

[208] C. C. Paglinawan, M. Hannah M. Caliolio, and J. B. Frias, "Medicine classification using YOLOv4 and tesseract OCR," in *Proc. 15th Int. Conf. Comput. Autom. Eng. (ICCAE)*, Sydney, NSW, Australia, Mar. 2023, pp. 260–263, doi: 10.1109/ICCAE56788.2023.10111387.

[209] T. Thapliyal, S. Bhatt, V. Rawat, and S. Maurya, "Automatic license plate recognition (ALPR) using YOLOv5 model and tesseract OCR engine," in *Proc. 1st Int. Conf. Adv. Electr., Electron. Comput. Intell. (ICAEECI)*, Tiruchengode, India, Oct. 2023, pp. 1–5, doi: 10.1109/icaeeci58247.2023.10370919.

[210] S. K. Ladi, G. K. Panda, R. Dash, and P. K. Ladi, "A pioneering approach of hyperspectral image classification employing the cooperative efforts of 3D, 2D and depthwise separable-1D convolutions," in *Proc. IEEE 2nd Int. Symp. Sustain. Energy, Signal Process. Cyber Secur. (iSSSC)*, Odisha, India, Dec. 2022, pp. 1–6, doi: 10.1109/iSSSC56467.2022.10051566.

[211] K. J. Han, R. Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1D convolutions," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 54–61, doi: 10.1109/ASRU46091.2019.9003730.

[212] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, and M. Gabbouj, "1-D convolutional neural networks for signal processing applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8360–8364, doi: 10.1109/ICASSP.2019.8682194.

[213] W. Huang and G. Liu, "Hierarchical text classification based on the end-to-end MCHA-BERT," in *Proc. IEEE 3rd Int. Conf. Frontiers Technol. Inf. Comput. (ICFTIC)*, Greenville, SC, USA, Nov. 2021, pp. 301–306, doi: 10.1109/ICFTIC54370.2021.9647279.

[214] J. Abdelnour, J. Rouat, and G. Salvi, "NAAQA: A neural architecture for acoustic question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4997–5009, Apr. 2023, doi: 10.1109/TPAMI.2022.3194311.

[215] J. Rämö and V. Välimäki, "Optimizing a high-order graphic equalizer for audio processing," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 301–305, Mar. 2014, doi: 10.1109/LSP.2014.2301557.

[216] Y. Yamazaki, C. Premachandra, and C. J. Perea, "Audio-processing-based human detection at disaster sites with unmanned aerial vehicle," *IEEE Access*, vol. 8, pp. 101398–101405, 2020, doi: 10.1109/ACCESS.2020.2998776.

[217] K. Kumar, R. Pandey, S. S. Bhattacharjee, and N. V. George, "Exponential hyperbolic cosine robust adaptive filters for audio signal processing," *IEEE Signal Process. Lett.*, vol. 28, pp. 1410–1414, 2021, doi: 10.1109/LSP.2021.3093862.

[218] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Frequency-domain adaptive filtering: From real to hypercomplex signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7745–7749, doi: 10.1109/ICASSP.2019.8683403.

[219] W. Fan, K. Chen, J. Lu, and J. Tao, "Effective improvement of under-modeling frequency-domain Kalman filter," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 342–346, Feb. 2019, doi: 10.1109/LSP.2019.2890965.

[220] É. Bavu, A. Ramamonjy, H. Pujol, and A. Garcia, "TimeScaleNet: A multiresolution approach for raw audio recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5686–5690, doi: 10.1109/ICASSP.2019.8682378.

[221] Ç. Bilen, A. Ozerov, and P. Pérez, "Solving time-domain audio inverse problems using nonnegative tensor factorization," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5604–5617, Nov. 2018, doi: 10.1109/TSP.2018.2869113.

[222] W. Cai, L. Xie, W. Yang, Y. Li, Y. Gao, and T. Wang, "DFTNet: Dual-path feature transfer network for weakly supervised medical image segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 4, pp. 2530–2540, Jul./Aug. 2023, doi: 10.1109/TCBB.2022.3198284.

[223] X. Dai, T. Ma, H. Cai, and Y. Wen, "Unsupervised hierarchical translation-based model for multi-modal medical image registration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 1261–1265, doi: 10.1109/ICASSP43922.2022.9746324.

[224] L. Xie, W. Cai, and Y. Gao, "DMCGNet: A novel network for medical image segmentation with dense self-mimic and channel grouping mechanism," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5013–5024, Oct. 2022, doi: 10.1109/JBHI.2022.3192277.

[225] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019, doi: 10.1109/TMI.2019.2903562.

[226] X. Bing, W. Zhang, L. Zheng, and Y. Zhang, "Medical image super resolution using improved generative adversarial networks," *IEEE Access*, vol. 7, pp. 145030–145038, 2019, doi: 10.1109/ACCESS.2019.2944862.

[227] M. Z. Khan, M. K. Gajendran, Y. Lee, and M. A. Khan, "Deep neural architectures for medical image semantic segmentation: Review," *IEEE Access*, vol. 9, pp. 83002–83024, 2021, doi: 10.1109/ACCESS.2021.3086530.

[228] J. Duan, S. Mao, J. Jin, Z. Zhou, L. Chen, and C. L. P. Chen, "A novel GA-based optimized approach for regional multimodal medical image fusion with superpixel segmentation," *IEEE Access*, vol. 9, pp. 96353–96366, 2021, doi: 10.1109/ACCESS.2021.3094972.

[229] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019, doi: 10.1109/ACCESS.2019.2908991.

[230] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022, doi: 10.1109/TMI.2022.3161829.

[231] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, and L. Ma, "MISSU: 3D medical image segmentation via self-distilling TransUNet," *IEEE Trans. Med. Imag.*, vol. 42, no. 9, pp. 2740–2750, Sep. 2023, doi: 10.1109/TMI.2023.3264433.

[232] Y. Zhao, K. Lu, J. Xue, S. Wang, and J. Lu, "Semi-supervised medical image segmentation with voxel stability and reliability constraints," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 8, pp. 3912–3923, Aug. 2023, doi: 10.1109/JBHI.2023.3273609.

[233] L. Wang, J. Zhang, Y. Liu, J. Mi, and J. Zhang, "Multimodal medical image fusion based on Gabor representation combination of multi-CNN and fuzzy neural network," *IEEE Access*, vol. 9, pp. 67634–67647, 2021, doi: 10.1109/ACCESS.2021.3075953.

[234] N. Hilmizen, A. Bustamam, and D. Sarwinda, "The multimodal deep learning for diagnosing COVID-19 pneumonia from chest CT-scan and X-ray images," in *Proc. 3rd Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Yogyakarta, Indonesia, Dec. 2020, pp. 26–31, doi: 10.1109/ISRITI51436.2020.9315478.

[235] T. Anwar and S. Zakir, "Deep learning based diagnosis of COVID-19 using chest CT-scan images," in *Proc. IEEE 23rd Int. Multi-topic Conf. (INMIC)*, Bahawalpur, Pakistan, Nov. 2020, pp. 1–5, doi: 10.1109/INMIC50486.2020.9318212.

[236] Y. F. Riti, H. A. Nugroho, S. Wibirama, B. Windarta, and L. Choridah, "Feature extraction for lesion margin characteristic classification from CT scan lungs image," in *Proc. 1st Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Aug. 2016, pp. 54–58, doi: 10.1109/ICITISEE.2016.7803047.

[237] A. Seum, A. H. Raj, S. Sakib, and T. Hossain, "A comparative study of CNN transfer learning classification algorithms with segmentation for COVID-19 detection from CT scan images," in *Proc. 11th Int. Conf. Electr. Comput. Eng. (ICECE)*, Dhaka, Bangladesh, Dec. 2020, pp. 234–237, doi: 10.1109/ICECE51571.2020.9393129.

[238] N. Vani and D. Vinod, "A comparative analysis on random forest algorithm over K-means for identifying the brain tumor anomalies using novel CT scan with MRI scan," in *Proc. Int. Conf. Bus. Anal. Technol. Secur. (ICBATS)*, Dubai, United Arab Emirates, Feb. 2022, pp. 1–6, doi: 10.1109/ICBATS54253.2022.9759036.

[239] A. Hoque, A. K. M. A. Farabi, F. Ahmed, and Md. Z. Islam, "Automated detection of lung cancer using CT scan images," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Dhaka, Bangladesh, Jun. 2020, pp. 1030–1033, doi: 10.1109/TENSYMP50017.2020.9230861.

[240] W. Cao, J. Zhang, C. Cai, Q. Chen, Y. Zhao, Y. Lou, W. Jiang, and G. Gui, "CNN-based intelligent safety surveillance in green IoT applications," *China Commun.*, vol. 18, no. 1, pp. 108–119, Jan. 2021, doi: 10.23919/JCC.2021.01.010.

[241] J. Xu, W. Zhou, Z. Fu, H. Zhou, and L. Li, "A survey on green deep learning," 2021, *arXiv:2111.05193*.

[242] Y. Rao, Z. Liu, W. Zhao, J. Zhou, and J. Lu, "Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10883–10897, Sep. 2023, doi: 10.1109/TPAMI.2023.3263826.

[243] Z. Li, M. Chen, J. Xiao, and Q. Gu, "PSAQ-ViT v2: Toward accurate and general data-free quantization for vision transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024, doi: 10.1109/TNNLS.2023.3301007.

[244] R. Garcia-Martin and R. Sanchez-Reillo, "Vision transformers for vein biometric recognition," *IEEE Access*, vol. 11, pp. 22060–22080, 2023, doi: 10.1109/ACCESS.2023.3252009.

[245] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and A. Alqahtani, "Mel-MViTv2: Enhanced speech emotion recognition with mel spectrogram and improved multiscale vision transformers," *IEEE Access*, vol. 11, pp. 108571–108579, 2023, doi: 10.1109/ACCESS.2023.3321122.

[246] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, Y.-G. Jiang, and S.-N. Lim, "AdaViT: Adaptive vision transformers for efficient image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12299–12308, doi: 10.1109/CVPR52688.2022.01199.

[247] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 6252–6261, doi: 10.1109/CVPR52729.2023.00605.

[248] J.-N. Chen, S. Sun, J. He, P. Torr, A. Yuille, and S. Bai, "TransMix: Attend to mix for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12125–12134, doi: 10.1109/CVPR52688.2022.01182.

[249] Y. Tang, K. Han, Y. Wang, C. Xu, J. Guo, C. Xu, and D. Tao, "Patch slimming for efficient vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12155–12164, doi: 10.1109/CVPR52688.2022.01185.

[250] A. Hatamizadeh, H. Yin, H. Roth, W. Li, J. Kautz, D. Xu, and P. Molchanov, "GradViT: Gradient inversion of vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10011–10020, doi: 10.1109/CVPR52688.2022.00978.

[251] Y. He, Z. Lou, L. Zhang, J. Liu, W. Wu, H. Zhou, and B. Zhuang, "BiViT: Extremely compressed binary vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 5628–5640, doi: 10.1109/iccv51070.2023.00520.

[252] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, and J. Ren, "Rethinking vision transformers for MobileNet size and speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 16843–16854, doi: 10.1109/iccv51070.2023.01549.

[253] T. Sun, C. Zhang, Y. Ji, and Z. Hu, "MSnet: Multi-head self-attention network for distantly supervised relation extraction," *IEEE Access*, vol. 7, pp. 54472–54482, 2019, doi: 10.1109/ACCESS.2019.2913316.

[254] Z. Xie, G. Zheng, L. Miao, and W. Huang, "STGL-GCN: Spatial–temporal mixing of global and local self-attention graph convolutional networks for human action recognition," *IEEE Access*, vol. 11, pp. 16526–16532, 2023, doi: 10.1109/ACCESS.2023.3246127.

[255] C.-X. Zhang, Y.-L. Zhang, and X.-Y. Gao, "Multi-head self-attention gated-dilated convolutional neural network for word sense disambiguation," *IEEE Access*, vol. 11, pp. 14202–14210, 2023, doi: 10.1109/ACCESS.2023.3243574.

[256] K. Zhao, W. Guo, F. Qin, and X. Wang, "D3-SACNN: DGA domain detection with self-attention convolutional network," *IEEE Access*, vol. 10, pp. 69250–69263, 2022, doi: 10.1109/ACCESS.2021.3127913.

[257] F. Zhang, A. Panahi, and G. Gao, "FsaNet: Frequency self-attention for semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 4757–4772, 2023, doi: 10.1109/TIP.2023.3305090.

[258] B. Kelenyi and L. Tamas, "D3GATTEN: Dense 3D geometric features extraction and pose estimation using self-attention," *IEEE Access*, vol. 11, pp. 7947–7958, 2023, doi: 10.1109/ACCESS.2023.3238901.

[259] Y. Wang, Y. Xie, X. Ji, Z. Liu, and X. Liu, "RacPixGAN: An enhanced sketch-to-face synthesis GAN based on residual modules, multi-head self-attention mechanisms, and CLIP loss," in *Proc. 4th Int. Conf. Electron. Commun. Artif. Intell. (ICECAI)*, Guangzhou, China, May 2023, pp. 336–342, doi: 10.1109/ICECAI58670.2023.10176715.

[260] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, "Lite vision transformer with enhanced self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11988–11998, doi: 10.1109/CVPR52688.2022.01169.

[261] S. Park and Y. S. Choi, "Image super-resolution using dilated window transformer," *IEEE Access*, vol. 11, pp. 60028–60039, 2023, doi: 10.1109/ACCESS.2023.3284539.

[262] F. Wang, X. Wang, D. Lv, L. Zhou, and G. Shi, "Separable self-attention mechanism for point cloud local and global feature modeling," *IEEE Access*, vol. 10, pp. 129823–129831, 2022, doi: 10.1109/ACCESS.2022.3228044.

[263] S. Lamba, A. Baliyan, and V. Kukreja, "GAN based image augmentation for increased CNN performance in paddy leaf disease classification," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Greater Noida, India, Apr. 2022, pp. 2054–2059, doi: 10.1109/ICACITE53722.2022.9823799.

[264] B. Sandhiya, R. Priyatharshini, B. Ramya, S. Monish, and G. R. S. Raja, "Reconstruction, identification and classification of brain tumor using GAN and faster regional-CNN," in *Proc. 3rd Int. Conf. Signal Process. Commun. (ICPSC)*, Coimbatore, India, May 2021, pp. 238–242, doi: 10.1109/ICSPC51351.2021.9451747.

[265] N. Sasipriyaa, P. Natesan, R. S. Mohana, P. Arvindkumar, R. S. A. Prakadis, and K. A. Surya, "Recognizing handwritten offline Tamil character using VAE-GAN & CNN," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Coimbatore, India, Jan. 2023, pp. 1–7, doi: 10.1109/ICCCI56745.2023.10128520.

[266] S. K. Singh, M. H. Anisi, S. Clough, T. Blyth, and D. Jarchi, "CNN-BiLSTM based GAN for anamoly detection from multivariate time series data," in *Proc. 24th Int. Conf. Digit. Signal Process. (DSP)*, Rhodes, Greece, Jun. 2023, pp. 1–4, doi: 10.1109/DSP58604.2023.10167937.

[267] Y. Fu, T. Sun, X. Jiang, K. Xu, and P. He, "Robust GAN-face detection based on dual-channel CNN network," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Suzhou, China, Oct. 2019, pp. 1–5, doi: 10.1109/CISP-BMEI48845.2019.8965991.

[268] A. Akram, N. Wang, X. Gao, and J. Li, "Integrating GAN with CNN for face sketch synthesis," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2018, pp. 1483–1487, doi: 10.1109/CompComm.2018.8780648.

[269] C.-H. Rhee and C. H. Lee, "Estimating physically-based reflectance parameters from a single image with GAN-guided CNN," *IEEE Access*, vol. 10, pp. 13259–13269, 2022, doi: 10.1109/ACCESS.2022.3147483.

[270] C. Mao, L. Huang, Y. Xiao, F. He, and Y. Liu, "Target recognition of SAR image based on CN-GAN and CNN in complex environment," *IEEE Access*, vol. 9, pp. 39608–39617, 2021, doi: 10.1109/ACCESS.2021.3064362.

[271] J. Wang, M. Xu, X. Deng, L. Shen, and Y. Song, "MW-GAN+ for perceptual quality enhancement on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4224–4237, Jul. 2022, doi: 10.1109/TCSVT.2021.3128275.

[272] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 79–88, doi: 10.1109/CVPR.2018.00016.

[273] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[274] S. Ma, W. Liu, W. Cai, Z. Shang, and G. Liu, "Lightweight deep residual CNN for fault diagnosis of rotating machinery based on depthwise separable convolutions," *IEEE Access*, vol. 7, pp. 57023–57036, 2019, doi: 10.1109/ACCESS.2019.2912072.

[275] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved MobileNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 14588–14597, doi: 10.1109/CVPR42600.2020.01461.

[276] A. Batool and Y.-C. Byun, "Lightweight EfficientNetB3 model based on depthwise separable convolutions for enhancing classification of leukemia white blood cell images," *IEEE Access*, vol. 11, pp. 37203–37215, 2023, doi: 10.1109/ACCESS.2023.3266511.

[277] N. A. Mohamed, M. A. Zulkifley, and S. R. Abdani, "Spatial pyramid pooling with atrous convolutional for MobileNet," in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Batu Pahat, Malaysia, Sep. 2020, pp. 333–336, doi: 10.1109/SCOReD50371.2020.9250928.

[278] S. Sriram, R. Vinayakumar, V. Sowmya, M. Alazab, and K. P. Soman, "Multi-scale learning based malware variant detection using spatial pyramid pooling network," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Canada, Jul. 2020, pp. 740–745, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162661.

[279] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.

[280] E. Prasetyo, N. Suciati, and C. Fatichah, "YOLOv4-tiny and spatial pyramid pooling for detecting head and tail of fish," in *Proc. Int. Conf. Artif. Intell. Comput. Sci. Technol. (ICAICST)*, Yogyakarta, Indonesia, Jun. 2021, pp. 157–161, doi: 10.1109/ICAICST53116.2021.9497822.

[281] Y. Tian, F. Chen, H. Wang, and S. Zhang, "Real-time semantic segmentation network based on lite reduced atrous spatial pyramid pooling module group," in *Proc. 5th Int. Conf. Control, Robot. Cybern. (CRC)*, Wuhan, China, Oct. 2020, pp. 139–143, doi: 10.1109/CRC51253.2020.9253492.

[282] A. Qayyum, I. Ahmad, W. Mumtaz, M. O. Alassafi, R. Alghamdi, and M. Mazher, "Automatic segmentation using a hybrid dense network integrated with an 3D-atrous spatial pyramid pooling module for computed tomography (CT) imaging," *IEEE Access*, vol. 8, pp. 169794–169803, 2020, doi: 10.1109/ACCESS.2020.3024277.

[283] C.-M. Pun and M.-C. Lee, "Extraction of shift invariant wavelet features for classification of images with different sizes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1228–1233, Sep. 2004, doi: 10.1109/TPAMI.2004.67.

[284] L. Liang and H. Liu, "Dual-tree cosine-modulated filter bank with linear-phase individual filters: An alternative shift-invariant and directional-selective transform," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5168–5180, Dec. 2013, doi: 10.1109/TIP.2013.2283146.

[285] L.-D. Kuang, Q.-H. Lin, X.-F. Gong, F. Cong, Y.-P. Wang, and V. D. Calhoun, "Shift-invariant canonical polyadic decomposition of complex-valued multi-subject fMRI data with a phase sparsity constraint," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 844–853, Apr. 2020, doi: 10.1109/TMI.2019.2936046.

[286] G. Papari, P. Campisi, and N. Petkov, "New families of Fourier eigenfunctions for steerable filtering," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 2931–2943, Jun. 2012, doi: 10.1109/TIP.2011.2179060.

[287] T. Zhao and T. Blu, "The Fourier-argand representation: An optimal basis of steerable patterns," *IEEE Trans. Image Process.*, vol. 29, pp. 6357–6371, 2020, doi: 10.1109/TIP.2020.2990483.

[288] A. Depeursinge, Z. Püspöki, J. P. Ward, and M. Unser, "Steerable wavelet machines (SWM): Learning moving frames for texture classification," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1626–1636, Apr. 2017, doi: 10.1109/TIP.2017.2655438.

[289] A. M. Alhassan and W. M. N. W. Zainon, "BAT algorithm with fuzzy C-ordered means (BAFCOM) clustering segmentation and enhanced capsule networks (ECN) for brain cancer MRI images classification," *IEEE Access*, vol. 8, pp. 201741–201751, 2020, doi: 10.1109/ACCESS.2020.3035803.

[290] P. Haridas, G. Chennupati, N. Santhi, P. Romero, and S. Eidenbenz, "Code characterization with graph convolutions and capsule networks," *IEEE Access*, vol. 8, pp. 136307–136315, 2020, doi: 10.1109/ACCESS.2020.3011909.

[291] B. Kakillioglu, A. Ren, Y. Wang, and S. Velipasalar, "3D capsule networks for object classification with weight pruning," *IEEE Access*, vol. 8, pp. 27393–27405, 2020, doi: 10.1109/ACCESS.2020.2971950.

[292] A. Marchisio, V. Mrazek, A. Massa, B. Bussolino, M. Martina, and M. Shafique, "RoHNAS: A neural architecture search framework with conjoint optimization for adversarial robustness and hardware efficiency of convolutional and capsule networks," *IEEE Access*, vol. 10, pp. 109043–109055, 2022, doi: 10.1109/ACCESS.2022.3214312.

[293] Y. Zhao, T. Birdal, H. Deng, and F. Tombari, "3D point capsule networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1009–1018, doi: 10.1109/CVPR.2019.00110.

[294] M.-H. Ha and O. T. Chen, "Deep neural networks using capsule networks and skeleton-based attentions for action recognition," *IEEE Access*, vol. 9, pp. 6164–6178, 2021, doi: 10.1109/ACCESS.2020.3048741.

[295] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne, and R. Rodrigo, "DeepCaps: Going deeper with capsule networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10717–10725, doi: 10.1109/CVPR.2019.01098.

[296] Z. Yan, X. Dai, P. Zhang, Y. Tian, B. Wu, and M. Feiszli, "FP-NAS: Fast probabilistic neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 15134–15143, doi: 10.1109/CVPR46437.2021.01489.

[297] M. Zhang, H. Li, S. Pan, X. Chang, and S. Su, "Overcoming multi-model forgetting in one-shot NAS with diversity maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7806–7815, doi: 10.1109/CVPR42600.2020.00783.

[298] X. Li, C. Lin, C. Li, M. Sun, W. Wu, J. Yan, and W. Ouyang, "Improving one-shot NAS by suppressing the posterior fading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 13833–13842, doi: 10.1109/CVPR42600.2020.01385.

[299] Z. Li, T. Xi, J. Deng, G. Zhang, S. Wen, and R. He, "GP-NAS: Gaussian process based neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11930–11939, doi: 10.1109/CVPR42600.2020.01195.

[300] C. Jiang, H. Xu, W. Zhang, X. Liang, and Z. Li, "SP-NAS: Serial-to-parallel backbone search for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11860–11869, doi: 10.1109/CVPR42600.2020.01188.

[301] Y. Gao, H. Bai, Z. Jie, J. Ma, K. Jia, and W. Liu, "MTL-NAS: Task-agnostic neural architecture search towards general-purpose multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Aug. 2020, pp. 11540–11549, doi: 10.1109/CVPR42600.2020.01156.

[302] Y. Liu, Y. Sun, B. Xue, M. Zhang, G. G. Yen, and K. C. Tan, "A survey on evolutionary neural architecture search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 550–570, Feb. 2023, doi: 10.1109/TNNLS.2021.3100554.

[303] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, "Neural architecture search benchmarks: Insights and survey," *IEEE Access*, vol. 11, pp. 25217–25236, 2023, doi: 10.1109/ACCESS.2023.3253818.

[304] Y. Weng, T. Zhou, L. Liu, and C. Xia, "Automatic convolutional neural architecture search for image classification under different scenes," *IEEE Access*, vol. 7, pp. 38495–38506, 2019, doi: 10.1109/ACCESS.2019.2906369.

[305] X. Zheng, R. Ji, Y. Chen, Q. Wang, B. Zhang, J. Chen, Q. Ye, F. Huang, and Y. Tian, "MIGO-NAS: Towards fast and generalizable neural architecture search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 2936–2952, Sep. 2021, doi: 10.1109/TPAMI.2021.3065138.

[306] C. Wei, C. Niu, Y. Tang, Y. Wang, H. Hu, and J. Liang, "NPENAS: Neural predictor guided evolution for neural architecture search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8441–8455, Nov. 2023, doi: 10.1109/TNNLS.2022.3151160.

[307] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric GAN for unpaired image-to-image translation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881–5896, Dec. 2019, doi: 10.1109/TIP.2019.2922854.

[308] C. D. Prakash and L. J. Karam, "It GAN do better: GAN-based detection of objects on images with varying quality," *IEEE Trans. Image Process.*, vol. 30, pp. 9220–9230, 2021, doi: 10.1109/TIP.2021.3124155.

[309] Y. Huang, F. Zheng, R. Cong, W. Huang, M. R. Scott, and L. Shao, "MCMT-GAN: Multi-task coherent modality transferable GAN for 3D brain image synthesis," *IEEE Trans. Image Process.*, vol. 29, pp. 8187–8198, 2020, doi: 10.1109/TIP.2020.3011557.

[310] T. Chen, Y. Zhang, X. Huo, S. Wu, Y. Xu, and H. S. Wong, "SphericGAN: Semi-supervised hyper-spherical generative adversarial networks for fine-grained image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 9991–10000, doi: 10.1109/CVPR52688.2022.00976.

[311] Z. Liu, M. Li, Y. Zhang, C. Wang, Q. Zhang, J. Wang, and Y. Nie, "Fine-grained face swapping via regional GAN inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 8578–8587, doi: 10.1109/cvpr52729.2023.00829.

[312] Y. Dong, W. Tan, D. Tao, L. Zheng, and X. Li, "CartoonLoss-GAN: Learning surface and coloring of images for cartoonization," *IEEE Trans. Image Process.*, vol. 31, pp. 485–498, 2022, doi: 10.1109/TIP.2021.3130539.

[313] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (GANs): A survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019, doi: 10.1109/ACCESS.2019.2905015.

[314] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[315] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.

[316] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.

[317] Y. Lu, L. Zhang, and W. Xie, "YOLO-compact: An efficient YOLO network for single category real-time object detection," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Hefei, China, Aug. 2020, pp. 1931–1936, doi: 10.1109/CCDC49329.2020.9164580.

[318] T.-H. Wu, T.-W. Wang, and Y.-Q. Liu, "Real-time vehicle and distance detection based on improved YOLO v5 network," in *Proc. 3rd World Symp. Artif. Intell. (WSAI)*, Guangzhou, China, Jun. 2021, pp. 24–28, doi: 10.1109/WSAI51899.2021.9486316.

[319] H. Yu, Y. Li, and D. Zhang, "An improved YOLO v3 small-scale ship target detection algorithm," in *Proc. 6th Int. Conf. Smart Grid Electr. Autom. (ICSGEA)*, Kunming, China, May 2021, pp. 560–563, doi: 10.1109/ICSGEA53208.2021.00132.

[320] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on YOLO network model," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Changchun, China, May 2018, pp. 1547–1551, doi: 10.1109/ICMA.2018.8484698.

[321] J. -H. Kim, N. Kim, and C. S. Won, "High-speed drone detection based on YOLO-V8," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Oct. 2023, pp. 1–2, doi: 10.1109/ICASSP49357.2023.10095516.

[322] Z. Li, Z. Liu, and X. Wang, "On-board real-time pedestrian detection for micro unmanned aerial vehicles based on YOLO-v8," in *Proc. 2nd Int. Conf. Mach. Learn., Cloud Comput. Intell. Mining (MLCCIM)*, Jiuzhaigou, China, Jul. 2023, pp. 250–255, doi: 10.1109/mlccim60412.2023.00042.

[323] N. Madhasu and S. D. Pande, "Chrometect GAYO: Classification and colorization using PIX2PIX and YOLOV8," in *Proc. 7th Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions (CSITSS)*, Bangalore, India, Nov. 2023, pp. 1–6, doi: 10.1109/csitss60515.2023.10384657.

[324] Y. Wang, H. Wang, and Z. Xin, "Efficient detection model of steel strip surface defects based on YOLO-V7," *IEEE Access*, vol. 10, pp. 133936–133944, 2022, doi: 10.1109/ACCESS.2022.3230894.

[325] F. Fang, L. Li, H. Zhu, and J.-H. Lim, "Combining faster R-CNN and model-driven clustering for elongated object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 2052–2065, 2020, doi: 10.1109/TIP.2019.2947792.

[326] X. Chen, C. Lian, H. H. Deng, T. Kuang, H.-Y. Lin, D. Xiao, J. Gateno, D. Shen, J. J. Xia, and P.-T. Yap, "Fast and accurate craniomaxillofacial landmark detection via 3D faster R-CNN," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3867–3878, Dec. 2021, doi: 10.1109/TMI.2021.3099509.

[327] S.-L. Chen, T.-Y. Chen, Y.-C. Mao, S.-Y. Lin, Y.-Y. Huang, C.-A. Chen, Y.-J. Lin, M.-H. Chuang, and P. A. R. Abu, "Detection of various dental conditions on dental panoramic radiography using faster R-CNN," *IEEE Access*, vol. 11, pp. 127388–127401, 2023, doi: 10.1109/ACCESS.2023.3332269.

[328] F. Selamet, S. Cakar, and M. Kotan, "Automatic detection and classification of defective areas on metal parts by using adaptive fusion of faster R-CNN and shape from shading," *IEEE Access*, vol. 10, pp. 126030–126038, 2022, doi: 10.1109/ACCESS.2022.3224037.

[329] A. Omid-Zohoor, C. Young, D. Ta, and B. Murmann, "Toward always-on mobile object detection: Energy versus performance trade-offs for embedded HOG feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1102–1115, May 2018, doi: 10.1109/TCSVT.2017.2653187.

[330] Y. Guo and B. Yang, "A survey of semantic segmentation methods in traffic scenarios," in *Proc. Int. Conf. Mach. Learn., Cloud Comput. Intell. Mining (MLCCIM)*, Xiamen, China, Aug. 2022, pp. 452–457, doi: 10.1109/MLCCIM55934.2022.00083.

[331] X. Xu, S. Huang, and H. Lai, "Lightweight semantic segmentation network leveraging class-aware contextual information," *IEEE Access*, vol. 11, pp. 144722–144734, 2023, doi: 10.1109/ACCESS.2023.3345790.

[332] S. Abdigapporov, S. Miraliev, V. Kakani, and H. Kim, "Joint multiclass object detection and semantic segmentation for autonomous driving," *IEEE Access*, vol. 11, pp. 37637–37649, 2023, doi: 10.1109/ACCESS.2023.3266284.

[333] Y. Zheng, Y. Xu, S. Qiu, W. Li, G. Zhong, and M. Sarem, "A novel semantic segmentation algorithm for RGB-D images based on non-symmetry and anti-packing pattern representation model," *IEEE Access*, vol. 11, pp. 36290–36299, 2023, doi: 10.1109/ACCESS.2023.3266251.

[334] Z. Cao, T. Zhang, W. Diao, Y. Zhang, X. Lyu, K. Fu, and X. Sun, "Meta-Seg: A generalized meta-learning framework for multi-class few-shot semantic segmentation," *IEEE Access*, vol. 7, pp. 166109–166121, 2019, doi: 10.1109/ACCESS.2019.2953465.

[335] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, J. M. Alvarez, and S. Gould, "Incorporating network built-in priors in weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1382–1396, Jun. 2018, doi: 10.1109/TPAMI.2017.2713785.

[336] P. Anilkumar, P. Venugopal, P. K. R. Maddikunta, T. R. Gadekallu, A. Al-Rasheed, M. Abbas, and B. O. Soufiene, "An adaptive DeepLabv3+ for semantic segmentation of aerial images using improved golden eagle optimization algorithm," *IEEE Access*, vol. 11, pp. 106688–106705, 2023, doi: 10.1109/ACCESS.2023.3318867.

[337] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021, doi: 10.1109/TIP.2021.3109518.

[338] A. Sohail, N. A. Nawaz, A. A. Shah, S. Rasheed, S. Ilyas, and M. K. Ehsan, "A systematic literature review on machine learning and deep learning methods for semantic segmentation," *IEEE Access*, vol. 10, pp. 134557–134570, 2022, doi: 10.1109/ACCESS.2022.3230983.

[339] W. Ji, J. Li, C. Bian, Z. Zhou, J. Zhao, A. Yuille, and L. Cheng, "Multispectral video semantic segmentation: A benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 1094–1104, doi: 10.1109/cvpr52729.2023.00112.

[340] J. Zhang, X. Zhao, Z. Chen, and Z. Lu, "A review of deep learning-based semantic segmentation for point cloud," *IEEE Access*, vol. 7, pp. 179118–179133, 2019, doi: 10.1109/ACCESS.2019.2958671.

[341] B. Woo and M. Lee, "Comparison of tissue segmentation performance between 2D U-Net and 3D U-Net on brain MR images," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2021, pp. 1–4, doi: 10.1109/ICEIC51217.2021.9369797.

[342] P. Harsh, R. Chakraborty, S. Tripathi, and K. Sharma, "Attention U-Net architecture for dental image segmentation," in *Proc. Int. Conf. Intell. Technol. (CONIT)*, Hubli, India, Jun. 2021, pp. 1–5, doi: 10.1109/CONIT51480.2021.9498422.

[343] G. B. Kande, L. Ravi, N. Kande, M. R. Nalluri, H. Kotb, K. M. AboRas, A. Yousef, Y. Y. Ghadi, and A. Sasikumar, "MSR U-net: An improved U-Net model for retinal blood vessel segmentation," *IEEE Access*, vol. 12, pp. 534–551, 2024, doi: 10.1109/ACCESS.2023.3347196.

[344] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, doi: 10.1109/ACCESS.2021.3086020.

[345] Y.-J. Huang, Q. Dou, Z. X. Wang, L. Z. Liu, L. S. Wang, H. Chen, P. A. Heng, and R. H. Xu, "HL-FCN: Hybrid loss guided FCN for colorectal cancer segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, May 2018, pp. 195–198, doi: 10.1109/ISBI.2018.8363553.

[346] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016, *arXiv:1612.01105*.

[347] H. Tang, "Vision question answering system based on RoBERTa and ViT model," in *Proc. Int. Conf. Image Process., Comput. Vis. Mach. Learn. (ICICML)*, Xi'an, China, Oct. 2022, pp. 258–261, doi: 10.1109/ICICML57342.2022.10009711.

[348] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A ViT backbone for diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 22669–22679, doi: 10.1109/cvpr52729.2023.02171.

[349] Z. Li and Q. Gu, "I-ViT: Integer-only quantization for efficient vision transformer inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 17019–17029, doi: 10.1109/iccv51070.2023.01565.

[350] H. Dong, C. Chen, J. Wang, F. Shen, and Y. Pang, "ViT-SAPS: Detail-aware transformer for mechanical assembly semantic segmentation," *IEEE Access*, vol. 11, pp. 41467–41479, 2023, doi: 10.1109/ACCESS.2023.3270807.

[351] L. Zou, Z. Huang, N. Gu, and G. Wang, "6D-ViT: Category-level 6D object pose estimation via transformer-based instance representation learning," *IEEE Trans. Image Process.*, vol. 31, pp. 6907–6921, 2022, doi: 10.1109/TIP.2022.3216980.

[352] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[353] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," 2021, *arXiv:2103.00112*.

[354] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[355] Z. Hu, Z. Gan, W. Li, J. Z. Wen, D. Zhou, and X. Wang, "Two-stage model-agnostic meta-learning with noise mechanism for one-shot imitation," *IEEE Access*, vol. 8, pp. 182720–182730, 2020, doi: 10.1109/ACCESS.2020.3029220.

[356] L. Xie, Y. Yang, Z. Fu, and S. M. Naqvi, "One-shot medical action recognition with a cross-attention mechanism and dynamic time warping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10097186.

[357] C. Huang, Y. Dang, P. Chen, X. Yang, and K.-T. Cheng, "One-shot imitation drone filming of human motion videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5335–5348, Sep. 2022, doi: 10.1109/TPAMI.2021.3067359.

[358] S. K. Biswas and P. Milanfar, "One shot detection with Laplacian object and fast matrix cosine similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 546–562, Mar. 2016, doi: 10.1109/TPAMI.2015.2453950.

[359] S.-Y. Huang and W.-T. Chu, "Searching by generating: Flexible and efficient one-shot NAS with architecture generator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 983–992, doi: 10.1109/CVPR46437.2021.00104.

[360] Y. Chen, T. Huang, Y. Niu, X. Ke, and Y. Lin, "Pose-guided spatial alignment and key frame selection for one-shot video-based person re-identification," *IEEE Access*, vol. 7, pp. 78991–79004, 2019, doi: 10.1109/ACCESS.2019.2922679.

[361] S. K. Roy, P. Kar, M. E. Paoletti, J. M. Haut, R. Pastor-Vargas, and A. Robles-Gómez, "SiCoDeF^2 Net: Siamese convolution deconvolution feature fusion network for one-shot classification," *IEEE Access*, vol. 9, pp. 118419–118434, 2021, doi: 10.1109/ACCESS.2021.3107626.

[362] Y. Song, T. Wang, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," 2022, *arXiv:2205.06743*.

[363] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," 2020, *arXiv:1904.05046*.

[364] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, Jan. 2019, doi: 10.1145/3293318.

[365] R. Xu, L. Xing, S. Shao, L. Zhao, B. Liu, W. Liu, and Y. Zhou, "GCT: Graph co-training for semi-supervised few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8674–8687, Dec. 2022, doi: 10.1109/TCSVT.2022.3196550.

[366] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4650–4666, Apr. 2023, doi: 10.1109/TPAMI.2022.3193587.

[367] L. Zhu and Y. Yang, "Label independent memory for semi-supervised few-shot video classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 273–285, Jan. 2022, doi: 10.1109/TPAMI.2020.3007511.

[368] Z. Yang, C. Zhang, R. Li, Y. Xu, and G. Lin, "Efficient few-shot object detection via knowledge inheritance," *IEEE Trans. Image Process.*, vol. 32, pp. 321–334, 2023, doi: 10.1109/TIP.2022.3228162.

[369] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022, doi: 10.1109/TPAMI.2020.3013717.

[370] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1530–1544, Ma.r 2024, doi: 10.1109/TPAMI.2021.3140070.

[371] S. Chen, Z. Hong, G. Xie, Q. Peng, X. You, W. Ding, and L. Shao, "GNDAN: Graph navigated dual attention network for zero-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2022, doi: 10.1109/TNNLS.2022.3155602.

[372] P. Ma, Z. He, W. Ran, and H. Lu, "A transferable generative framework for multi-label zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2024, doi: 10.1109/TCSVT.2023.3324648.

[373] L. Feng and C. Zhao, "Transfer increment for generalized zero-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2506–2520, Jun. 2021, doi: 10.1109/TNNLS.2020.3006322.

[374] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015, doi: 10.1109/TPAMI.2015.2408354.

[375] C. Gong, J. Yang, J. You, and M. Sugiyama, "Centroid estimation with guaranteed efficiency: A general framework for weakly supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2841–2855, Jun. 2022, doi: 10.1109/TPAMI.2020.3044997.

[376] D. Ienco, Y. J. E. Gbodjo, R. Gaetano, and R. Interdonato, "Weakly supervised learning for land cover mapping of satellite image time series via attention-based CNN," *IEEE Access*, vol. 8, pp. 179547–179560, 2020, doi: 10.1109/ACCESS.2020.3024133.

[377] J. Han, Y. Yang, D. Zhang, D. Huang, D. Xu, and F. De La Torre, "Weakly-supervised learning of category-specific 3D object shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1423–1437, Apr. 2021, doi: 10.1109/TPAMI.2019.2949562.

[378] Y. Feng, L. Wang, and M. Zhang, "Weakly-supervised learning of a deep convolutional neural networks for semantic segmentation," *IEEE Access*, vol. 7, pp. 91009–91018, 2019, doi: 10.1109/ACCESS.2019.2926972.

[379] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 601–614, Mar. 2012, doi: 10.1109/TPAMI.2011.158.

[380] R. Cong, Q. Qin, C. Zhang, Q. Jiang, S. Wang, Y. Zhao, and S. Kwong, "A weakly supervised learning framework for salient object detection via hybrid labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 534–548, Feb. 2023, doi: 10.1109/TCSVT.2022.3205182.

[381] X. Zhou, A. Sun, Y. Liu, J. Zhang, and C. Miao, "SelfCF: A simple framework for self-supervised collaborative filtering," *ACM Trans. Recommender Syst.*, vol. 1, no. 2, pp. 1–25, Jun. 2023, doi: 10.1145/3591469.

[382] L. Tian, Z. Tu, D. Zhang, J. Liu, B. Li, and J. Yuan, "Unsupervised learning of optical flow with CNN-based non-local filtering," *IEEE Trans. Image Process.*, vol. 29, pp. 8429–8442, 2020, doi: 10.1109/TIP.2020.3013168.

[383] Y. A. D. Djilali, T. Krishna, K. McGuinness, and N. E. O'Connor, "Rethinking 360° image visual attention modelling with unsupervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 15394–15404, doi: 10.1109/ICCV48922.2021.01513.

[384] J. Xu, Z. Zhang, and X. Hu, "Extracting semantic knowledge from GANs with unsupervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9654–9668, Aug. 2023, doi: 10.1109/TPAMI.2023.3262140.

[385] Y. Shan, H. S. Sawhney, and R. Kumar, "Unsupervised learning of discriminative edge measures for vehicle matching between nonoverlapping cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 700–711, Apr. 2008, doi: 10.1109/TPAMI.2007.70728.

[386] W. Kim, A. Kanezaki, and M. Tanaka, "Unsupervised learning of image segmentation based on differentiable feature clustering," *IEEE Trans. Image Process.*, vol. 29, pp. 8055–8068, 2020, doi: 10.1109/TIP.2020.3011269.

[387] K. Song, J. Xie, S. Zhang, and Z. Luo, "Multi-mode online knowledge distillation for self-supervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 11848–11857, doi: 10.1109/CVPR52729.2023.01140.

[388] S. Zare and H. Van Nguyen, "Evaluating and improving domain invariance in contrastive self-supervised learning by extrapolating the loss function," *IEEE Access*, vol. 11, pp. 137758–137768, 2023, doi: 10.1109/ACCESS.2023.3339775.

[389] R. Li, C. Zhang, G. Lin, Z. Wang, and C. Shen, "RigidFlow: Self-supervised scene flow learning on point clouds by local rigidity prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16938–16947, doi: 10.1109/CVPR52688.2022.01645.

[390] F. D. Pup and M. Atzori, "Applications of self-supervised learning to biomedical signals: A survey," *IEEE Access*, vol. 11, pp. 144180–144203, 2023, doi: 10.1109/ACCESS.2023.3344531.

[391] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," 2023, *arXiv:2302.00487*.

[392] J. He and F. Zhu, "Out-of-distribution detection in unsupervised continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 3849–3854, doi: 10.1109/CVPRW56347.2022.00430.

[393] B. Kwon and T. Kim, "Toward an online continual learning architecture for intrusion detection of video surveillance," *IEEE Access*, vol. 10, pp. 89732–89744, 2022, doi: 10.1109/ACCESS.2022.3201139.

[394] Z. Cai, O. Sener, and V. Koltun, "Online continual learning with natural distribution shifts: An empirical study with visual data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 8261–8270, doi: 10.1109/ICCV48922.2021.00817.

[395] Y. Zhao, D. Saxena, and J. Cao, "AdaptCL: Adaptive continual learning for tackling heterogeneity in sequential datasets," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023, doi: 10.1109/TNNLS.2023.3341841.
[396] X. Li and W. Wang, "GopGAN: Gradients orthogonal projection generative adversarial network with continual learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 215–227, Jan. 2023, doi: 10.1109/TNNLS.2021.3093319.
[397] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022, doi: 10.1109/TPAMI.2021.3057446.
[398] L. Wang, B. Lei, Q. Li, H. Su, J. Zhu, and Y. Zhong, "Triple-memory networks: A brain-inspired method for continual learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 1925–1934, May 2022, doi: 10.1109/TNNLS.2021.3111019.
[399] M. Xue, H. Zhang, J. Song, and M. Song, "Meta-attention for ViT-backed continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 150–159, doi: 10.1109/CVPR52688.2022.00025.
[400] X. Wang, L. Yao, X. Wang, H.-Y. Paik, and S. Wang, "Uncertainty estimation with neural processes for meta-continual learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 6887–6897, Oct. 2023, doi: 10.1109/TNNLS.2022.3215633.
[401] B. Sistaninejhad, H. Rasi, and P. Nayeri, "A review paper about deep learning for medical image analysis," *Comput. Math. Methods Med.*, vol. 2023, pp. 1–10, May 2023, doi: 10.1155/2023/7091301.
[402] G. Papanastasiou, N. Dikaios, J. Huang, C. Wang, and G. Yang, "Is attention all you need in medical image analysis? A review," 2023, *arXiv:2307.12775*.
[403] Z. M. C. Baum, Y. Hu, and D. C. Barratt, "Meta-learning initializations for interactive medical image registration," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 823–833, Mar. 2023, doi: 10.1109/TMI.2022.3218147.
[404] M. T. Irshad and H. U. Rehman, "Gradient compass-based adaptive multimodal medical image fusion," *IEEE Access*, vol. 9, pp. 22662–22670, 2021, doi: 10.1109/ACCESS.2021.3054843.
[405] J. S. Duncan and N. Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 85–106, Jan. 2000, doi: 10.1109/34.824822.
[406] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018, doi: 10.1109/ACCESS.2017.2788044.
[407] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.
[408] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, p. 226, Oct. 2018, doi: 10.1007/s10916-018-1088-1.
[409] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.
[410] J. Jiang, P. Trundle, and J. Ren, "Medical image analysis with artificial neural networks," *Computerized Med. Imag. Graph.*, vol. 34, no. 8, pp. 617–631, Dec. 2010, doi: 10.1016/j.compmedimag.2010.07.003.
[411] S. Ye, T. Wang, M. Ding, and X. Zhang, "F-DARTS: Foveated differentiable architecture search based multimodal medical image fusion," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3348–3361, Nov. 2023, doi: 10.1109/TMI.2023.3283517.
[412] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1871–1880, doi: 10.1109/CVPR.2019.00197.
[413] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022, doi: 10.1109/TKDE.2021.3070203.
[414] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3994–4003, doi: 10.1109/CVPR.2016.433.
[415] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022, doi: 10.1109/TPAMI.2021.3054719.
[416] F. Zhao, Y. Li, L. Bai, Z. Tian, and X. Wang, "Semi-supervised multi-granularity CNNs for text classification: An application in human-car interaction," *IEEE Access*, vol. 8, pp. 68000–68012, 2020, doi: 10.1109/ACCESS.2020.2985098.
[417] H. Chen, Y. Wang, and Q. Hu, "Multi-granularity regularized re-balancing for class incremental learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7263–7277, Jul. 2023, doi: 10.1109/TKDE.2022.3188335.
[418] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," *IEEE Trans. Image Process.*, vol. 29, pp. 5542–5556, 2020, doi: 10.1109/TIP.2020.2984883.
[419] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov. 2017, doi: 10.1109/TPAMI.2016.2640292.
[420] C. Sahin, G. Garcia-Hernando, J. Sock, and T.-K. Kim, "A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators," *Image Vis. Comput.*, vol. 96, Apr. 2020, Art. no. 103898, doi: 10.1016/j.imavis.2020.103898.
[421] T. Hodan, D. Baráth, and J. Matas, "EPOS: Estimating 6D pose of objects with symmetries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11700–11709, doi: 10.1109/CVPR42600.2020.01172.
[422] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1530–1538, doi: 10.1109/ICCV.2017.169.
[423] T. Vaudrey, A. Wedel, C. Rabe, J. Klappstein, and R. Klette, "Evaluation of moving object segmentation comparing 6D-vision and monocular motion constraints," in *Proc. 23rd Int. Conf. Image Vis. Comput.*, Christchurch, New Zealand, Nov. 2008, pp. 1–6, doi: 10.1109/IVCNZ.2008.4762126.
[424] Z. He, W. Feng, X. Zhao, and Y. Lv, "6D pose estimation of objects: Recent technologies and challenges," *Appl. Sci.*, vol. 11, no. 1, p. 228, Dec. 2020, doi: 10.3390/app11010228.
[425] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6D pose estimation," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 657–678, Nov. 2019, doi: 10.1007/s11263-019-01250-9.
[426] T. Elsken, J. Hendrik Metzen, and F. Hutter, "Neural architecture search: A survey," 2018, *arXiv:1808.05377*.
[427] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley, "Neural architecture search without training," 2020, *arXiv:2006.04647*.
[428] L. Sekanina, "Neural architecture search and hardware accelerator co-search: A survey," *IEEE Access*, vol. 9, pp. 151337–151362, 2021, doi: 10.1109/ACCESS.2021.3126685.
[429] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani, "Neural architecture search for transformers: A survey," *IEEE Access*, vol. 10, pp. 108374–108412, 2022, doi: 10.1109/ACCESS.2022.3212767.
[430] K. G. Mills, M. Salameh, D. Niu, F. X. Han, S. S. C. Rezaei, H. Yao, W. Lu, S. Lian, and S. Jui, "Exploring neural architecture search space via deep deterministic sampling," *IEEE Access*, vol. 9, pp. 110962–110974, 2021, doi: 10.1109/ACCESS.2021.3101975.
[431] Z. Ding, Y. Chen, N. Li, D. Zhao, Z. Sun, and C. L. P. Chen, "BNAS: Efficient neural architecture search using broad scalable architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 5004–5018, Sep. 2022, doi: 10.1109/TNNLS.2021.3067028.
[432] Z. Ma, Z. Zhou, Y. Liu, Y. Lei, and H. Yan, "Auto-ORVNet: Orientation-boosted volumetric neural architecture search for 3D shape classification," *IEEE Access*, vol. 8, pp. 12942–12954, 2020, doi: 10.1109/ACCESS.2019.2961715.
[433] X. Zhang, Z. Huang, N. Wang, S. Xiang, and C. Pan, "You only search once: Single shot neural architecture search via direct sparse optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 2891–2904, Sep. 2021, doi: 10.1109/TPAMI.2020.3020300.
[434] Y. Guo, Y. Zheng, M. Tan, Q. Chen, Z. Li, J. Chen, P. Zhao, and J. Huang, "Towards accurate and compact architectures via neural architecture transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6501–6516, Oct. 2022, doi: 10.1109/TPAMI.2021.3086914.

[435] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," 2022, *arXiv:2209.02646*.
[436] W. Mao, B. Han, and Z. Wang, "Sketchffusion: Sketch-guided image editing with diffusion model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, Oct. 2023, pp. 790–794, doi: 10.1109/icip49359.2023.10222365.
[437] X. P. Ooi and C. Seng Chan, "LLDE: Enhancing low-light images with diffusion model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, Oct. 2023, pp. 1305–1309, doi: 10.1109/icip49359.2023.10222446.
[438] T. Roque, L. Risser, V. Kersemans, S. Smart, D. Allen, P. Kinchesh, S. Gilchrist, A. L. Gomes, J. A. Schnabel, and M. A. Chappell, "A DCE-MRI driven 3-D reaction-diffusion model of solid tumor growth," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 724–732, Mar. 2018, doi: 10.1109/TMI.2017.2779811.
[439] D. Kim, E. Lee, D. Yoo, and H. Lee, "Fine-grained human hair segmentation using a text-to-image diffusion model," *IEEE Access*, vol. 12, pp. 13912–13922, 2024, doi: 10.1109/ACCESS.2024.3355542.
[440] A. Karnewar, A. Vedaldi, D. Novotny, and N. J. Mitra, "HOLODIFFUSION: Training a 3D diffusion model using 2D images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 18423–18433, doi: 10.1109/CVPR52729.2023.01767.
[441] W. Ran, W. Yuan, and R. Shibasaki, "Few-shot depth completion using denoising diffusion probabilistic model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 6559–6567, doi: 10.1109/cvprw59228.2023.00697.
[442] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022, doi: 10.1109/TPAMI.2021.3079209.
[443] I. Khan, X. Zhang, M. Rehman, and R. Ali, "A literature survey and empirical study of meta-learning for classifier selection," *IEEE Access*, vol. 8, pp. 10262–10281, 2020, doi: 10.1109/ACCESS.2020.2964726.
[444] P. Zhang, C. Liu, X. Chang, Y. Li, and M. Li, "Metric-based meta-learning model for few-shot PolSAR image terrain classification," in *Proc. CIE Int. Conf. Radar (Radar)*, Hainan, China, Dec. 2021, pp. 2529–2533, doi: 10.1109/Radar53847.2021.10027883.
[445] X. Zhang, D. Meng, H. Gouk, and T. Hospedales, "Shallow Bayesian meta learning for real-world few-shot recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 631–640, doi: 10.1109/ICCV48922.2021.00069.
[446] K. Gao, B. Liu, X. Yu, and A. Yu, "Unsupervised meta learning with multiview constraints for hyperspectral image small sample set classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3449–3462, 2022, doi: 10.1109/TIP.2022.3169689.
[447] H. Cho, Y. Cho, J. Yu, and J. Kim, "Camera distortion-aware 3D human pose estimation in video with optimization-based meta-learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11149–11158, doi: 10.1109/ICCV48922.2021.01098.
[448] L. Zhang, Z. Liu, W. Zhang, and D. Zhang, "Style uncertainty based self-paced meta learning for generalizable person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 2107–2119, 2023, doi: 10.1109/TIP.2023.3263112.
[449] H. Coskun, M. Z. Zia, B. Tekin, F. Bogo, N. Navab, F. Tombari, and H. S. Sawhney, "Domain-specific priors and meta learning for few-shot first-person action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6659–6673, Jun. 2023, doi: 10.1109/TPAMI.2021.3058606.
[450] Y. Deng, T. Han, and N. Ansari, "FedVision: Federated video analytics with edge computing," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 62–72, 2020, doi: 10.1109/OJCS.2020.2996184.
[451] K. Doshi and Y. Yilmaz, "Privacy-preserving video understanding via transformer-based federated learning," in *Proc. IEEE Conf. Dependable Secure Comput. (DSC)*, Tampa, FL, USA, Nov. 2023, pp. 1–8, doi: 10.1109/dsc61021.2023.10354099.
[452] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Federated learning-powered visual object detection for safety monitoring," *AI Mag.*, vol. 42, no. 2, pp. 19–27, Jun. 2021, doi: 10.1609/aimag.v42i2.15095.
[453] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Fedvision: An online visual object detection platform powered by federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13172–13179, doi: 10.1609/aaai.v34i08.7021.
[454] K. Zhou and X. Eric Wang, "FedVLN: Privacy-preserving federated vision-and-language navigation," 2022, *arXiv:2203.14936*.
[455] H. Li, K. Yin, X. Ji, Y. Liu, T. Huang, and G. Yin, "Improved YOLOV3 surveillance device object detection method based on federated learning," in *Proc. 4th Int. Conf. Data-Driven Optim. Complex Syst. (DOCS)*, Chengdu, China, Oct. 2022, pp. 1–6, doi: 10.1109/DOCS55193.2022.9967481.
[456] M. Alazab, R. M. S. Priya, M. Parimala, P. K. R. Maddikunta, T. R. Gadekallu, and Q.-V. Pham, "Federated learning for cybersecurity: Concepts, challenges, and future directions," *IEEE Trans. Ind. Informat.*, vol. 18, no. 5, pp. 3501–3509, May 2022, doi: 10.1109/TII.2021.3119038.
[457] P. K. Mandal, C. Leo, and C. Hurley, "Horizontal federated computer vision," 2023, *arXiv:2401.00390*.

ABOLFAZL YOUNESI received the B.Sc. degree in computer engineering from Tabriz University, Tabriz, Iran, in 2021. He is currently pursuing the M.Sc. degree in computer engineering with the Sharif University of Technology (SUT), Tehran, Iran. He is also a member of the Embedded Systems Research Laboratory (ESR-LAB), Department of Computer Engineering, Sharif University of Technology. His research interests include the Internet of Things (IoT) and cyber-physical systems (CPS), distributed systems, machine learning, and computer vision.

MOHSEN ANSARI received the Ph.D. degree in computer engineering from the Sharif University of Technology, Tehran, Iran, in 2021. He is currently an Assistant Professor in computer engineering with the Sharif University of Technology. He was a Visiting Researcher with the Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Germany, from 2019 to 2021. His research interests include cyber-physical systems, embedded machine learning, edge, fog, cloud computing, and thermal and low-power design of CPSs.

MOHAMMADAMIN FAZLI received the B.Sc. degree in hardware engineering and the M.Sc. and Ph.D. degrees in software engineering from the Sharif University of Technology, Tehran, Iran, in 2009, 2011, and 2015, respectively. He is currently an Assistant Professor with the Sharif University of Technology, where he is also a Research and Development Supervisor with the Intelligent Information Solutions Center. His current research interests include game theory, combinatorial optimization, computational business and economics, graphs and combinatorics, complex networks, and dynamical systems.

ALIREZA EJLALI received the Ph.D. degree in computer engineering from the Sharif University of Technology (SUT), Tehran, Iran, in 2006. He is currently an Associate Professor in computer engineering with SUT. From 2005 to 2006, he was a Visiting Researcher with the Electronic Systems Design Group, University of Southampton, Southampton, U.K. He is also the Director of the Embedded Systems Research Laboratory, Department of Computer Engineering, Sharif University of Technology. His research interests include low-power design, real-time, and fault-tolerant embedded systems.

MUHAMMAD SHAFIQUE (Senior Member, IEEE) received the Ph.D. degree in computer science from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2011. He was a Full Professor with the Institute of Computer Engineering, TU Wien, Vienna, Austria, from October 2016 to August 2020. Since September 2020, he has been with the Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. He is currently a Global Network Faculty Member of the NYU Tandon School of Engineering, Brooklyn, NY, USA. His research interests include system-level design for brain-inspired computing, AI/machine learning hardware, wearables, autonomous systems, energy-efficient and robust computing, the IoT, and smart CPS. He received the 2015 ACM/SIGDA Outstanding New Faculty Award, the AI 2000 Chip Technology Most Influential Scholar Award, in 2020, 2022, and 2023, six gold medals, and several best paper awards and nominations. He has given several keynotes, talks, and tutorials and organized special sessions at premier venues. He has served as the PC chair, the general chair, the track chair, and a PC member for several conferences.

JÖRG HENKEL (Fellow, IEEE) received the Diploma and Ph.D. degrees (summa cum laude) from the Technical University of Braunschweig. He is currently the Chair Professor in embedded systems with Karlsruhe Institute of Technology (KIT). Before that, he was a Research Staff Member with the NEC Laboratories, Princeton, NJ, USA. His research interests include co-design for embedded hardware/software systems with respect to power security and means of embedded machine learning. He is also the Vice President of Publications at IEEE CEDA and a fellow of the ACM. He has led several conferences as the General Chair, including ICCAD and ESWeek and is also the General of DAC'60. He serves as a steering committee chair/member for leading conferences and journals for embedded and cyber-physical systems. He has coordinated the DFG Program SPP 1500 "Dependable Embedded Systems" and is a Site Coordinator of the DFG TR89 Collaborative Research Center on "Invasive Computing." He is also the Chairman of the IEEE Computer Society, Germany Chapter. He has received six best paper awards throughout his career from, among others, ICCAD, ESWeek, and DATE. For two consecutive terms each, he served as the Editor-in-Chief for the *ACM Transactions on Embedded Computing Systems* and *IEEE Design & Test* magazine.