

The Evolution of Deep Learning Architectures for Visual Recognition: Design Principles, Feature Learning, and Network Strategies

Guoping Xu

Wuhan Institute of Technology

Junjie Zhang

Wuhan Institute of Technology

Jax Luo

Cleveland Clinic

Tianming Du

Aalto University

Chang Li

Hefei University of Technology

Xinglong Wu

Wuhan Institute of Technology

Huabing Zhou

Wuhan Institute of Technology

Huaxin Xu

Wuhan Institute of Technology

Yajun Yu

yajun.yu@fccc.edu

Fox Chase Cancer Center

Research Article

Keywords: Deep Learning, CNNs, Transformer, Architecture, Image Classification, Object Detection, Semantic Segmentation

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11102058/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

The Evolution of Deep Learning Architectures for Visual Recognition: Design Principles, Feature Learning, and Network Strategies

Guoping Xu¹, Junjie Zhang¹, Jax Luo², Tianming Du^{3,4}, Chang Li⁵, Xinglong Wu¹, Huabing
Zhou¹, Huaxin Xu^{1#}, Yajun Yu^{6#}

¹Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering,
Wuhan Institute of Technology, Wuhan, Hubei, 430205, China

²Neurological Institute, Cleveland Clinic, Cleveland, OH, 44195, USA

³ELLIS Institute Finland, Espoo, 02110, Finland

⁴Aalto University, Espoo, 02150, Finland

⁵Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of
Technology, Hefei, Anhui, 230009, China

⁶Department of Radiation Oncology, Fox Chase Cancer Center, Philadelphia, PA 19111, USA

[#]Corresponding Author: huaxin.xu@wit.edu.cn, yajun.yu@fccc.edu

Abstract

Deep learning, particularly deep neural networks, has driven remarkable advances in computer vision over the past two decades, achieving state-of-the-art performance in tasks such as image classification, object detection, and semantic segmentation. Existing survey papers largely concentrate on narrow sub-domains, specific applications, or individual technical components of computer vision. Nevertheless, there is still a shortage of systematic overviews that trace the intertwined evolution of neural network architectures and underlying paradigms which underpin modern deep-learning-driven computer vision.

In this systematic review, we synthesize seminal publications and pivotal methodological advances to trace the architectural and paradigm-driven evolution of deep learning for computer vision. We organize major technical developments into three coherent core dimensions: (1) task-oriented neural network architectures, (2) design principles of effective sub-modules for feature learning, and (3) essential strategies for constructing high-performance deep neural networks. Within each dimension, we examine representative foundational approaches, highlighting their effectiveness, robustness, and generalizability across diverse computer-vision application scenarios.

Beyond retrospective literature synthesis, we deliver a critical discussion of prevailing open challenges, inherent limitations of existing paradigms, and promising emerging research directions. Framed under a unified analytical framework, this survey intends to equip researchers and practitioners with structured insights into how deep-learning architectures and associated paradigms have evolved in computer vision. By consolidating milestone developments, core design principles, and forward-looking research trends, our work seeks to deepen community-wide understanding of this field and stimulate further innovations for both basic theoretical research and practical real-world deployments.

The project page can be found at <https://github.com/apple1986/DevelopmentOfDL>.

Keywords: Deep Learning, CNNs, Transformer, Architecture, Image Classification, Object Detection, Semantic Segmentation

44 1. Introduction

45 “It was the best of times, it was the worst of times” [1]. As the development of artificial
46 intelligence, especially the leap of the deep neural network (dubbed “Deep Learning”), many magic
47 productions have been proposed which we can almost not imagine in the past. The deep convolution
48 neural network (CNN) surpasses human-level performance on ImageNet classification [2] in 2015.
49 The AlphaGo [3] defeated a world champion in the game of Go in 2016. In 2017, Dermatologist-
50 level classification of skin cancer with deep neural networks was reported [4]. The AlphaFold (deep
51 learning-based) was proposed for highly accurate protein structure prediction and achieved
52 remarkable performance [5], which was a difficult thing in the past. Currently, large generation
53 models ignite our enthusiasm to pursue the machine intelligence on par our human’s level, such as
54 GPTs [6-8] for language generation, CLIP [9] and DALL.E [10] for image generation.

55 Although significant progress has been made, an important question remains: what are the
56 fundamental architecture developments that have driven the evolution of computer vision over the
57 past two decades? Many review papers have been published, yet most of them focus on the specific
58 task or topic of computer vision. In [11, 12], the authors reviewed the deep convolutional neural
59 networks for image classification, in which analyzed the short history of deep learning and several
60 symbolic works and improvement for image classification. Several review papers have been
61 published on image classification using deep learning across various modalities, including
62 hyperspectral images [13, 14] and remote sensing [15, 16], medical image [17, 18] and biological
63 image [19, 20]. In addition, numerous review papers have been published on topics such as object
64 detection [21, 22], semantic segmentation [23, 24], image super resolution [25, 26], image denoising
65 [27, 28], transfer learning [29, 30], video object segmentation and tracking [31, 32] and a name of
66 few. These reviews cover various deep neural networks applied to different tasks and integrate prior
67 knowledge specific to different image modalities (see in Table 1). However, they do not fully discuss
68 the architecture evolution and the general principles and strategies for designing or using effective
69 deep neural networks. More specifically, how can we identify, integrate, and even advance beyond
70 the most influential techniques and methodologies for designing effective and efficient deep neural
71 networks across diverse computer vision tasks?

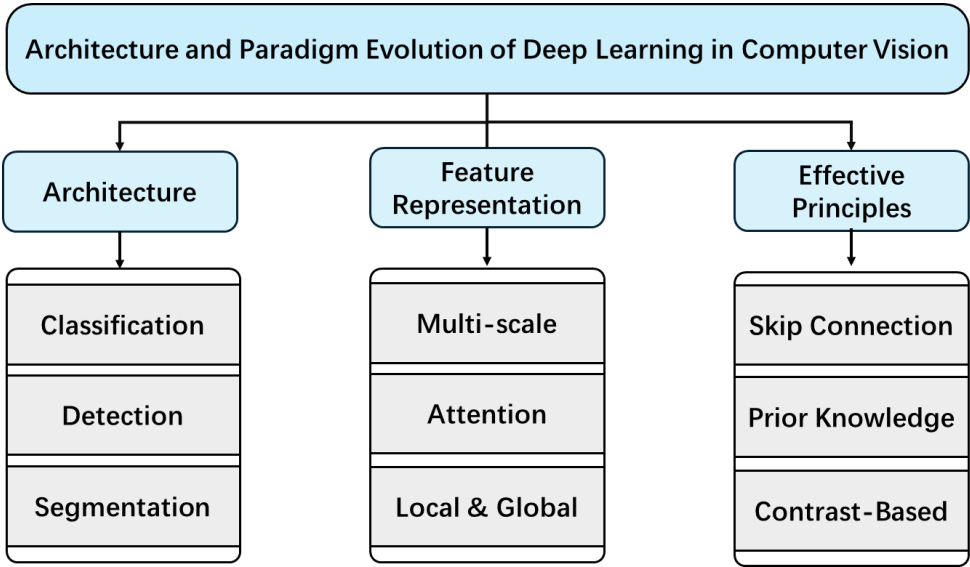
72 Table 1 Comparative analysis of representative survey papers on deep learning for computer
73 vision, categorized by major research topics and their corresponding scopes

Survey	Topic	Description/Scope
MIA[33], Annu. Rev. Biomed. Eng[34]	Annotation efficiency, medical image segmentation	Learning from imperfect annotations, including limited, weak, noisy, and image-level annotations for medical image segmentation.
PR[35], Artif. Intell. Rev[36]	Image augmentation	Comprehensive review of image augmentation techniques for improving deep learning models.
Neural Comput[11], MIA[17]	Image classification	Development and evolution of deep learning methods for image classification.
Artif. Intell. Rev[24], TPAMI[37], MIA[38]	Segmentation	Advances in image segmentation for natural images, medical images, and other imaging modalities.
TKDE[39], TNNLS[40]	Semi-supervised learning	Semi-supervised learning methods for image classification and segmentation, including network architectures, consistency learning, and optimization strategies.
TPAMI [41], Expert Syst. App[42]	Small object detection	Algorithms, benchmarks, and challenges for small object detection.
TNNLS[43], CUSR[44], TIST[45]	Convolution, Transformer, Mamba	Recent advances in convolutional neural networks, Vision Transformers, Mamba, and related architectures for computer vision.
CUSR[46, 47], TPAMI[48]	Efficient learning	Efficient neural architectures, model compression, acceleration algorithms, and benchmark studies.
TPAMI[49], CUSR[50]	Image generation	Deep generative models, including Generative Adversarial Networks (GANs) and diffusion models for image synthesis.
TIST[32], TPAMI[51],	Video analysis	Deep learning methods for video object segmentation, object tracking, and video object detection.
EAAI[52], TCSVT[53]	Essential components and techniques	Fundamental network components and techniques, such as residual learning and Gaussian splatting.

74

75 To address this question, we systematically survey seminal deep learning studies published over
76 the past two decades and distill their core concepts, design philosophies, and underlying
77 methodologies. This review seeks to equip readers with a holistic understanding of the intertwined

78 architectural and paradigm-driven evolutionary trajectory of deep learning in computer vision (CV).
 79 As outlined in Figure 1, we organize the major technical advances into three core dimensions that
 80 structure our manuscript: (1) task-oriented neural network architectures, (2) design principles of
 81 effective sub-modules for feature learning, and (3) essential strategies for constructing high-
 82 performance deep neural networks.
 83 Section 2 is dedicated to task-oriented neural network architectures. Within this section, we
 84 examine backbone and prediction-head designs tailored to representative vision tasks, including
 85 image classification, object detection, semantic segmentation, and image restoration tasks such as
 86 super-resolution and denoising. Section 3 elaborates on the design principles of effective sub-
 87 modules for feature learning. We cover core building blocks including multi-scale feature extraction
 88 and fusion, attention mechanisms, local–global feature modeling, and other practical feature
 89 learning strategies. Section 4 summarizes essential strategies for constructing high-performance
 90 deep neural networks, where we discuss residual learning and skip connections, the integration of
 91 prior knowledge, and contrast-based learning paradigms. Following these retrospective reviews,
 92 Section 5 presents a critical discussion of open challenges and promising future research directions,
 93 covering efficient network architectures, representation learning, world models, universal learning,
 94 and AIGC. Section 6 concludes this survey.



95
 96 Figure 1. Organizational framework of this review, illustrating three core dimensions: task-oriented
 97 network architectures, feature representation design, and effective network construction principles,
 98 along with their respective subtopics.

99 **2. Task-oriented neural network architectures**

100 As the AlexNet [54] achieved the champion of image classification in ImageNet Large scale
101 visual recognition challenge (ILSVRC) in 2012, deep neural networks became widely accepted in
102 computer vision. Unlike conventional approaches that rely on manually designed feature
103 engineering, deep learning models automatically learn hierarchical feature representations from raw
104 data, thereby eliminating labor-intensive handcrafted feature extraction and mitigating the inherent
105 limitations of artificial feature design. Afterwards, many seminal neural network architectures are
106 proposed focusing on specific problems. In this section, we review some fundamental and
107 impressive works in image classification, object detection, semantic segmentation and image super-
108 resolution, which are becoming the first-choice or default baseline for various tasks.

109 **2.1. Image classification**

110 The VGG [55] is a convolutional neural network (CNN) architecture known for its simplicity
111 and depth. It features small 3x3 convolutional filters stacked sequentially to increase the depth of
112 the network, allowing for better feature extraction while maintaining fewer parameters. VGG
113 models, such as VGG-16 and VGG-19, have been widely used in image classification tasks and
114 transfer learning due to their strong performance on large-scale image datasets like ImageNet. Yet,
115 increasing the number of layers does not necessarily lead to improved accuracy in neural networks.
116 In fact, accuracy may plateau and then degrade sharply beyond a certain point [56, 57]. The ResNet
117 [58] was proposed to introduce residual connections, or skip connections, between layers (see Figure
118 2). These residual connections allow the network to effectively "skip" layers, enabling the training
119 of much deeper networks without loss of performance. ResNet's architecture, including models like
120 ResNet-50 and ResNet-101, has become a foundational building block for many tasks in computer
121 vision. Afterwards, the skip connection or residual learning has been the default choice for designing
122 neural networks [52], such as DenseNet [59], Inception-v4 [60], ResNeXt [61], and ConvNeXt [62].
123 Collectively, these advances demonstrate that simply stacking more convolutional layers cannot
124 yield consistent performance gains. Instead, well-designed connection mechanisms for feature
125 propagation and gradient flow are critical for unlocking the representation potential of deep neural
126 networks.

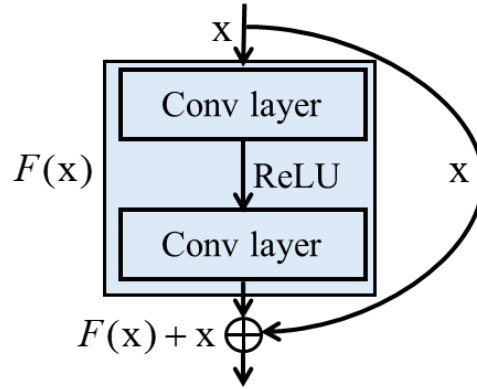


Figure 2. Residual learning block

Despite the strong representation learning capabilities of convolutional neural networks, they still have inherent limitations. One major challenge is their limited ability to capture long-range dependencies, as convolution operations are intrinsically localized [63]. To address this, the Transformer model [64], which relies solely on self-attention mechanisms, was introduced to overcome CNNs' limitations, enabling the capture of global context in natural language processing (NLP). Building on this, the Vision Transformer (ViT) was developed for image classification, allowing the model to learn relationships between all image patches. However, due to ViT's high computational cost, the Swin Transformer [65] was introduced to limit self-attention to non-overlapping local windows while enabling cross-window interactions through shifted window operations. This shifted window mechanism, combined with its hierarchical architecture, provides flexibility to model at various scales and achieves linear computational complexity relative to image size.

Although impressive performance has been achieved in image classification, most state-of-the-art networks still require substantial computational resources and often overlook training and inference efficiency, which is crucial for deployment on edge devices such as smartphones and autonomous vehicles. To this end, many lightweight and efficiency architectures are proposed. In MobileNets [66], it utilizes depthwise separable convolutions to build light-weight deep neural networks. The MobileNetV2 [67] inherits the idea of residual learning and depthwise separable convolutions, presenting an inverted residual structure. It demonstrates that increasing the channel number of feature maps in bottleneck could maintain representational power while trade-offs the accuracy and efficiency. Combination of hardware-aware network architecture search (NAS) and

the NetAdapt algorithm, the MobileNetV3 is present. Besides the MobileNets series, there are other efficient architectures are proposed. The ShuffleNet [68] and ShuffleNet V2 [69] are introduced, which utilizes pointwise group convolution and channel shuffle to reduce computation cost while maintaining accuracy. Due to the success of Transformer, some architectures are present based on Transformer, such as Efficientformer [70], EdgeViTs [71] and EffcientViT [72], or hybrid of CNN and Transformer, like MobileViT [73], Mobile-Former [74] and LeViT [75]. The main purpose of these architectures is to reduce the computation overhead from self-attention between each patch of image of feature maps.

2.2 Object detection

As a fundamental task in computer vision, object detection has attracted extensive attention and achieved remarkable progress over the past two decades with the development of deep learning techniques [21]. Compared with image classification, object detection aims not only to classify object categories but also to accurately localize objects with tight bounding boxes (see Figure 3). Depending on whether region proposals (or candidates) are generated, object detection methods can generally be divided into two-stage and one-stage approaches.

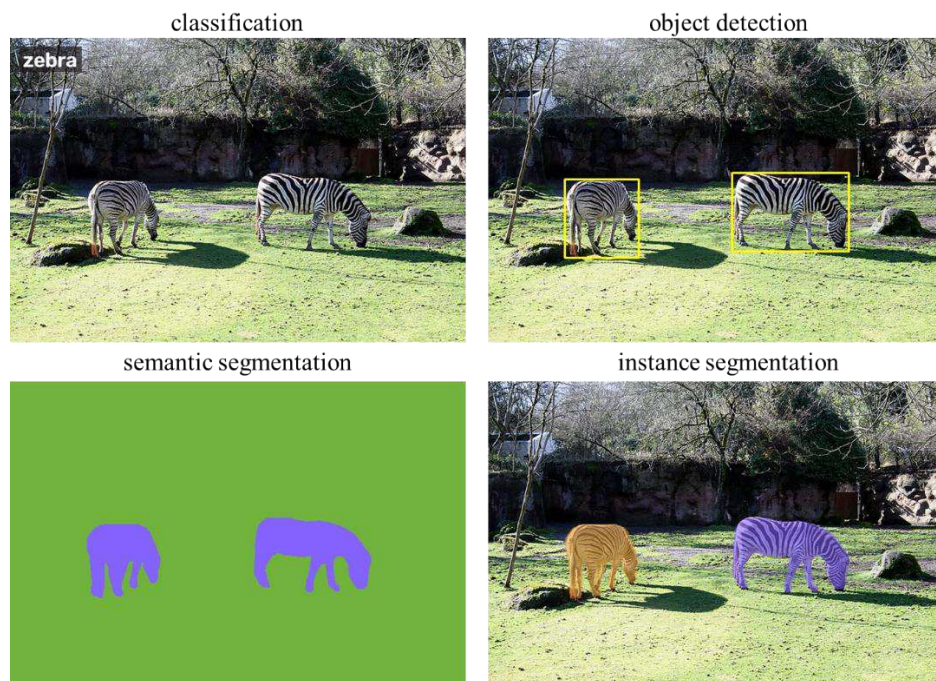


Figure 3. Comparison of representative vision tasks: classification, object detection, semantic segmentation, and instance segmentation

For two-stage methods, the R-CNN series [76-78] represents one of the most influential lines

of research. The pioneering R-CNN introduced bounding box proposal-based object detection using deep convolutional neural networks (DCNNs). Specifically, it first generates approximately 2,000 candidate object proposals using selective search algorithm on inputting image. Each resized proposal is then fed into a CNN for feature extraction, followed by class-specific linear SVMs for classification. However, R-CNN suffers from substantial computational redundancy because each proposal is processed independently.

To address this limitation, Fast R-CNN [77] was proposed to improve computational efficiency by forwarding the entire image through the network only once. Features of all proposals are subsequently extracted directly from the shared feature maps using a Region of Interest (RoI) pooling operation, which produces fixed-size feature representations for classification and bounding-box regression. Building upon this framework, Faster R-CNN [78] further introduces a Region Proposal Network (RPN) to automatically generate region proposals during training. This design enables proposal generation and object detection to be jointly optimized within a unified end-to-end framework, significantly improving both detection accuracy and efficiency.

In contrast to two-stage detection pipelines that first generate sparse region proposals and subsequently refine classification and localization within object proposals, one-stage detectors directly predict object categories and bounding coordinates in a fully end-to-end manner, thereby achieving more concise inference pipelines and higher computational efficiency. As the most representative and pioneering one-stage detection frameworks, YOLO (You Only Look Once) [79] and SSD (Single Shot MultiBox Detector) [80] have laid the foundation for subsequent single-stage detection research. Specifically, the YOLO paradigm reformulates object detection as a unified regression task, which enables simultaneous optimization of bounding box regression and class probability prediction via a single convolutional network. In the original YOLO architecture, input images are resized to a fixed resolution of 448×448 and downsampled through consecutive convolutional layers to produce low-resolution deep feature maps. After a total downsampling stride of 64, the network outputs a 7×7 feature grid (spatial dimension) with rich semantic representations (channel dimension). The extracted feature maps are then flattened and fed into multilayer perceptrons (MLPs) for global feature aggregation, ultimately yielding a fixed-dimensional prediction vector. This output vector is reshaped into a structured prediction tensor with a dimension of $7 \times 7 \times 30$. In this design, the input image is uniformly divided into a 7×7 grid containing 49

independent spatial cells. Each grid cell predicts two bounding boxes, with each box encoded by five parameters representing coordinate offsets, box width and height, and object confidence. Additionally, each grid cell outputs 20 dimensional class probabilities, corresponding to the predefined object categories of the detection dataset.

Following the one-stage object detection paradigm pioneered by YOLO, SSD introduces two key architectural innovations. First, SSD employs a set of predefined default bounding boxes with multiple scales and aspect ratios, and performs bounding-box regression and category prediction for each anchor on different feature maps. Unlike YOLO, which predicts a limited number of bounding boxes for each spatial grid cell, SSD generates a larger number of candidate predictions through dense multi-scale anchor settings, thereby improving detection flexibility and localization capability. Second, SSD effectively exploits multi-scale feature representations from different stages of the backbone network to jointly capture low-level spatial details and high-level semantic information. This design enables the detector to better handle objects of varying sizes, particularly small objects. By combining multi-scale anchor-based prediction with hierarchical feature representations, SSD achieves a favorable trade-off between detection accuracy and computational efficiency. Experimental results in the original study demonstrated that SSD outperformed YOLO in accuracy while maintaining real-time inference speed, with performance comparable to Faster R-CNN under certain benchmark settings.

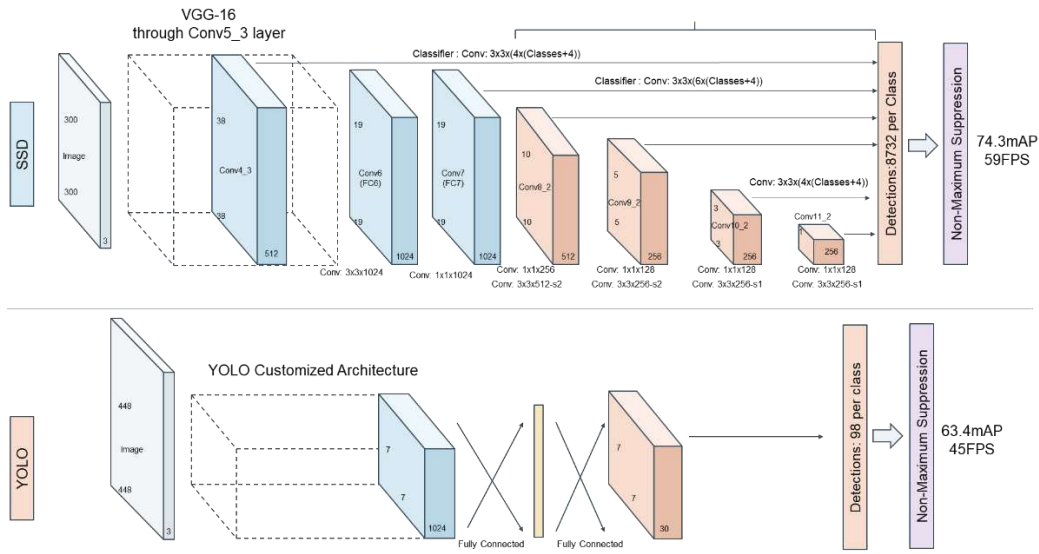


Figure 4. Architecture of SSD and YOLO

After the introduction of YOLOv1 and SSD, a series of increasingly advanced YOLO detectors

have been successively proposed, establishing the YOLO family as one of the most influential paradigms in one-stage real-time object detection. For example, YOLOv2 [81] introduced the novel Darknet-19 backbone, which incorporated batch normalization and global average pooling to improve training stability, convergence speed, and inference efficiency. In addition, anchor boxes and multi-scale training were adopted to significantly enhance localization accuracy and generalization capability. Subsequent versions, ranging from YOLOv3 to YOLOv12, have been released in rapid one-year to two-year iterative cycles, continuously driving the evolution of object detection techniques. These models progressively integrated a wide range of innovations, including more powerful backbone and neck architectures [82, 83], robust data augmentation strategies [83, 84], efficient multi-scale feature fusion mechanisms [83], attention modules [84], anchor-free designs [85], label assignment strategies [86], normalization techniques [83], and advanced optimization paradigms [87] (see Table 2). Such improvements have substantially enhanced both detection accuracy and computational efficiency, enabling YOLO-based detectors to achieve state-of-the-art (SOTA) performance across diverse real-world applications while maintaining real-time inference capability. Furthermore, recent YOLO variants increasingly emphasize lightweight deployment, end-to-end optimization, and adaptability to edge devices, thereby broadening their applicability in autonomous driving, medical image analysis, robotics, remote sensing, and intelligent surveillance systems.

Table 2 The evolution of YOLO series for one-stage object detection

Model	Year	Main contribution
YOLOv1 [79]	2015	First unified one-stage object detector; formulates detection as a single regression problem for real-time performance.
YOLOv2 [81]	2016	Introduced Darknet-19, batch normalization, anchor boxes, and multi-scale training for improved accuracy and speed.
YOLOv3 [82]	2018	Adopted Darknet-53 and feature pyramid networks (FPN) for enhanced multi-scale detection capability.
YOLOv4 [83]	2020	Integrated CSPDarknet53, Mosaic augmentation, CIoU loss, and Bag-of-Freebies/Bag-of-Specials strategies.
YOLOv5 [88]	2020	Improved engineering efficiency and deployment flexibility with adaptive scaling and optimized training pipelines.
YOLOv6 [89]	2022	Designed for industrial applications with efficient decoupled heads and hardware-aware optimization.
YOLOv7 [84]	2022	Introduced trainable Bag-of-Freebies and extended efficient layer

Model	Year	Main contribution
		aggregation networks for higher accuracy-speed tradeoff.
YOLOv8 [85]	2023	Anchor-free architecture with improved backbone and decoupled detection head for enhanced robustness and flexibility.
YOLOv9 [86]	2024	Proposed programmable gradient information and generalized efficient layer aggregation mechanisms.
YOLOv10 [87]	2024	Introduced end-to-end NMS-free detection framework for lower latency and improved efficiency.
YOLOv11 [90]	2024	Further optimized architectural efficiency, feature representation, and lightweight deployment performance.
YOLOv12 [91]	2025	Focuses on next-generation real-time detection with enhanced scalability, efficiency, and cross-domain adaptability.

One limitation of both two-stage and one-stage object detectors lies in their reliance on hand-crafted or pre-defined candidate generation mechanisms. In two-stage methods, such as R-CNN and Fast R-CNN, region proposals are typically generated using external algorithms like Selective Search. Faster R-CNN improves this pipeline by introducing a Region Proposal Network (RPN) that learns to generate proposals directly from feature maps in an end-to-end manner. In contrast, one-stage detectors eliminate explicit proposal generation but rely on dense, predefined anchor boxes of various scales and aspect ratios distributed across feature maps. Both of two-stage or one-stage leads to a large number of candidate bounding boxes, which are subsequently refined and filtered using non-maximum suppression (NMS) to remove redundant detections. However, NMS is a non-differentiable post-processing step, which introduces additional computational overhead and limits full end-to-end optimization.

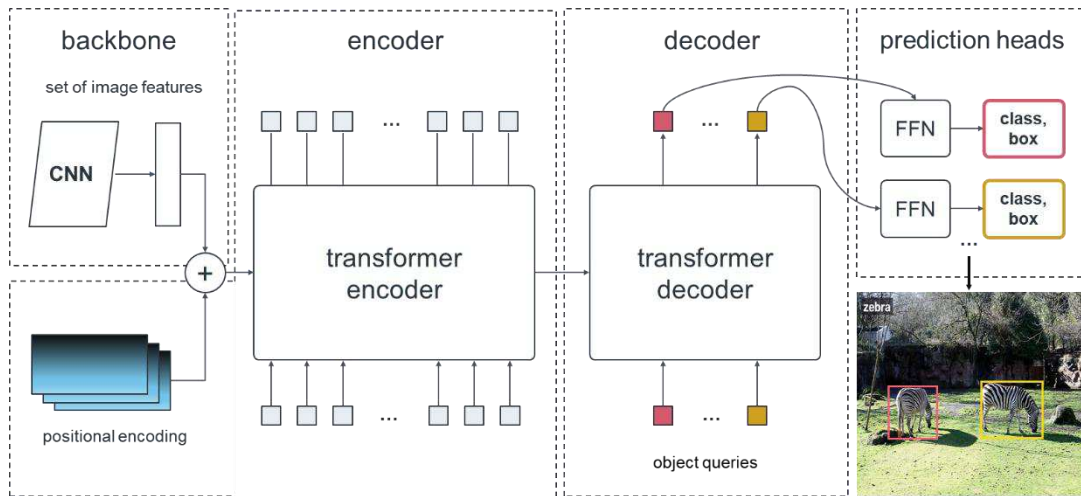


Figure 5. Illustration of DETR

The pursuit of a fully end-to-end object detection framework has motivated the development of transformer-based detectors, which reformulate object detection as a direct set prediction problem and eliminate the need for hand-crafted anchors, region proposal mechanisms, and post-processing non-maximum suppression (NMS). One of the earliest and most representative frameworks is the Detection Transformer (DETR) [92]. DETR adopts an encoder–decoder Transformer architecture to capture global contextual relationships through self-attention mechanisms. Unlike conventional CNN-based detectors, DETR introduces a fixed set of learnable object queries that serve as object-level representations, functioning analogously to the proposal generation process in the Region Proposal Network (RPN) of Faster R-CNN.

Benefiting from the global receptive field of the Transformer, these object queries interact with the encoded image features to generate bounding boxes candidates in parallel. To achieve unique object prediction assignment, DETR employs bipartite matching based on the Hungarian algorithm to establish a one-to-one correspondence between predicted boxes and ground-truth annotations. Consequently, redundant predictions are naturally suppressed during training, eliminating the requirement for heuristic post-processing methods such as NMS. The matched predictions are then optimized using a combination of classification and bounding-box regression losses in an end-to-end manner.

Although the original DETR suffers from relatively slow convergence and suboptimal performance on small-scale objects, its elegant end-to-end formulation and simplified detection pipeline have inspired extensive follow-up research and significantly advanced the development of transformer-based object detection frameworks. Representative extensions include Deformable DETR [93], which introduces sparse attention sampling to accelerate convergence and improve small-object detection; Dynamic DETR [94], which dynamically refines object queries during training; and DN-DETR [95], which incorporates denoising training strategies to stabilize optimization and further speed up convergence.

Reviewing the evolution of object detection reveals several major development trends compared with conventional image classification tasks. First, object detection extends image-level recognition to simultaneous localization and classification, transforming the task from “what” to “where and what,” thereby requiring more sophisticated loss formulations and optimization strategies. Second, the field continuously benefits from advances in network architectures, ranging

from CNN-based backbones to Transformer-based frameworks. Third, object detectors progressively integrate state-of-the-art techniques from the broader computer vision community, such as multi-scale feature learning in SSD and YOLOv3, attention mechanisms and Transformers in DETR, and increasingly efficient optimization and feature representation strategies.

2.3. Semantic segmentation

Semantic segmentation aims to assign a semantic label to each pixel in an image. It plays a crucial role in numerous applications, including autonomous driving, medical image analysis, and image editing. Since the pioneering work of FCN (Fully Convolutional Networks) [96], which introduced end-to-end learning for semantic segmentation, the field has experienced substantial progress. Building upon FCN, the U-Net [97] was proposed as an encoder-decoder architecture with skip connections that transfer feature maps from the encoder to the corresponding decoder stages (See Figure 6). This design effectively combines low-level spatial details with high-level semantic information, enabling more accurate segmentation results. Due to its simplicity, effectiveness, and strong performance in limited-data scenarios, U-Net has become one of the most widely adopted architectures in biomedical image segmentation.

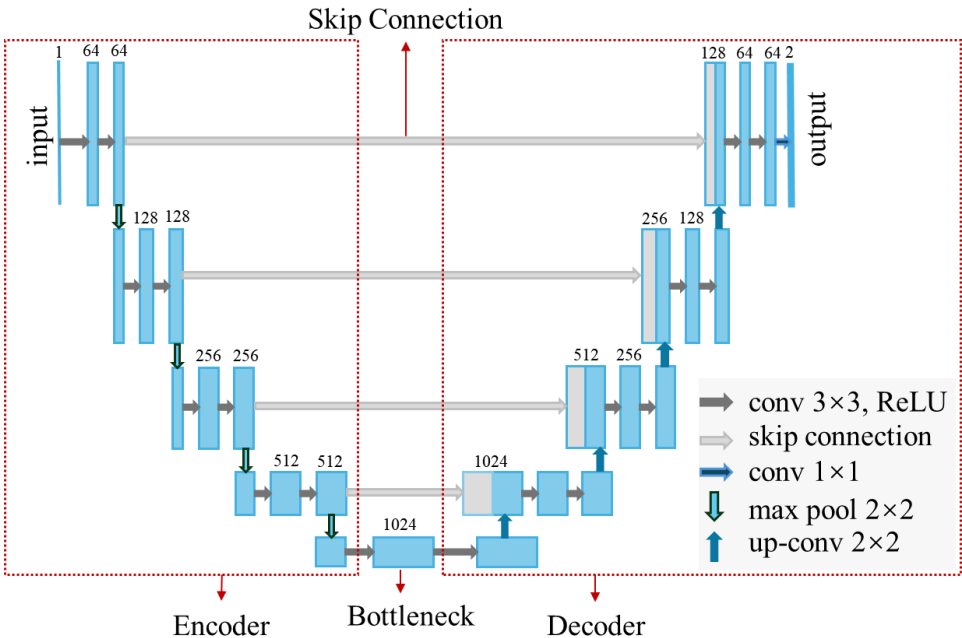


Figure 6. Structure of U-Net

Numerous subsequent studies have extended the U-Net framework by incorporating additional prior knowledge and architectural improvements, such as multi-scale feature fusion, residual

learning, and attention mechanisms. For example, 3D U-Net [98] replaces all 2D operations in U-Net with 3D counterparts to better handle volumetric medical image segmentation. However, the original U-Net directly concatenates encoder and decoder features through skip connections, which may introduce semantic discrepancies between low-level and high-level representations and potentially limit segmentation performance. To address this issue, UNet++ [99] redesigns the skip pathways with nested and dense skip connections to better exploit multi-scale features and reduce the semantic gap between encoder and decoder representations. Furthermore, UNet 3+ [100] introduces full-scale skip connections to aggregate semantic information from multiple feature scales more comprehensively. More recently, nnU-Net [101], built upon the U-Net framework, demonstrated that automatically configuring preprocessing, network architecture, training strategies, and post-processing pipelines for different datasets can achieve highly robust and generalizable segmentation performance across diverse medical imaging tasks.

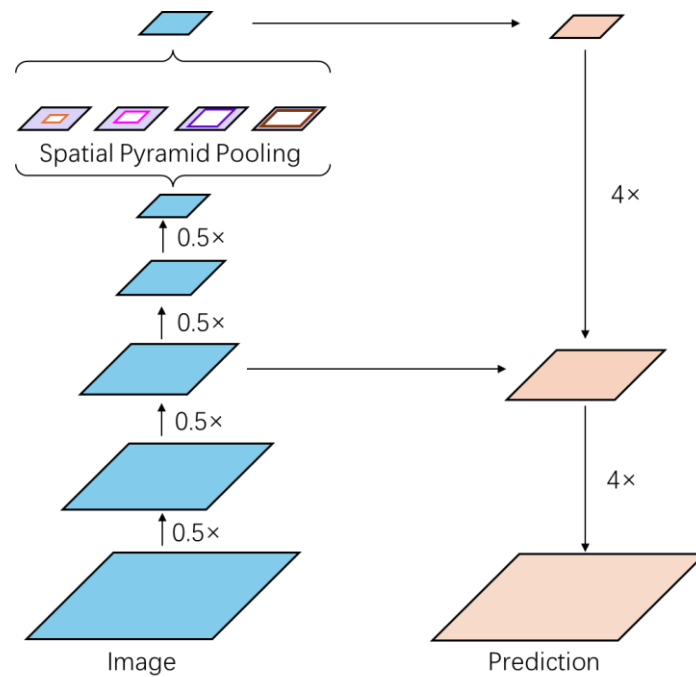


Figure 7. Structure of DeepLabv3+

In addition to the rapid development of U-Net-based architectures for medical image segmentation, several seminal methods have also been proposed for natural image segmentation. Among them, the DeepLab series [102-105] introduced dilated (atrous) convolutions to effectively enlarge the receptive field while preserving spatial resolution. In particular, DeepLabv3+ (see Figure 7) combines an encoder-decoder architecture with Atrous Spatial Pyramid Pooling (ASPP) and

depthwise separable convolutions, leading to significant improvements in segmentation accuracy and computational efficiency. Similarly, PSPNet [106] employs a pyramid pooling module to capture contextual information at multiple spatial scales, thereby enhancing scene understanding. HRNet [107], in contrast, maintains high-resolution representations throughout the entire network and performs repeated multi-scale feature fusion in parallel branches, enabling more precise spatial representation learning.

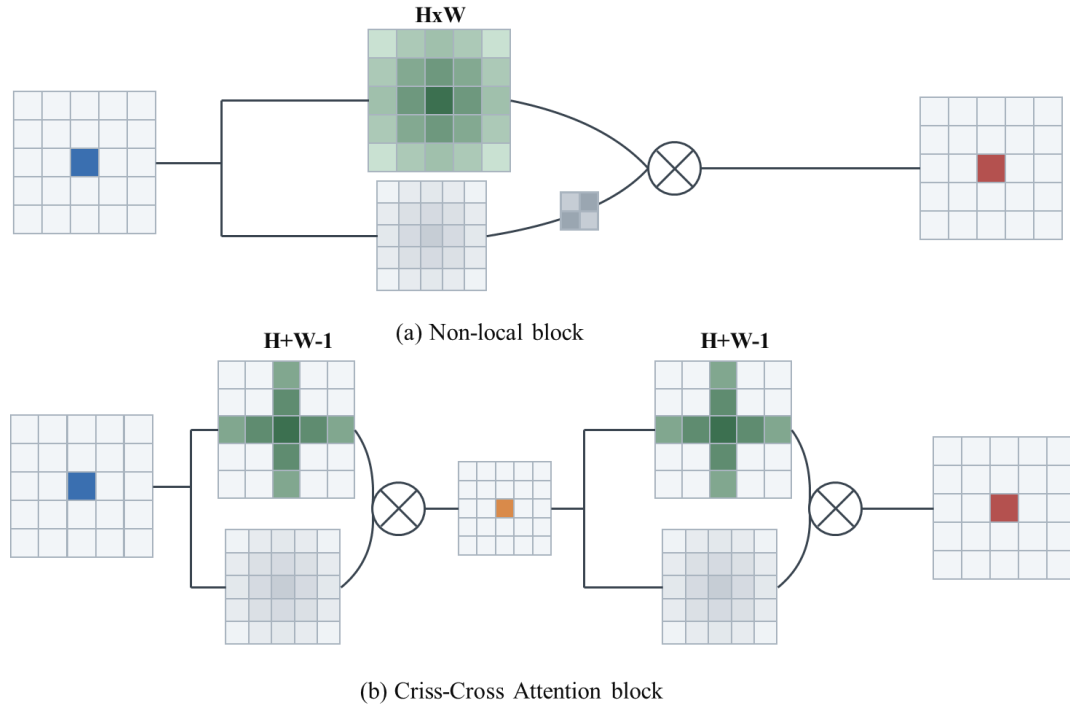


Figure 8. Comparison structure of Non-local block and Criss-Cross Attention block

While methods such as dilated convolutions and pyramid pooling primarily exploit multi-scale contextual information, another line of research focuses on attention mechanisms to model long-range dependencies and contextual relationships. For example, DANet [108] introduces both position attention and channel attention modules to capture spatial and channel-wise contextual dependencies, respectively. Furthermore, Non-Local Networks [109] and CCNet [110] (see Figure 8) leverage self-attention mechanisms to model global contextual interactions, effectively addressing the inherent locality limitation of convolutional operations.

Apart from convolutional neural network (CNN)-based or CNN with attention modules approaches, several representative Transformer-based methods have also been proposed for image segmentation, including SETR [111], Pyramid Vision Transformer (PVT) [112], SwinUNETR [113],

and SegFormer [114]. These methods exploit the strong global modeling capability of Vision Transformers while incorporating multi-scale feature representations to improve segmentation performance. In particular, SETR formulates semantic segmentation as a sequence-to-sequence prediction task using pure Transformers, demonstrating the potential of Transformer architectures for dense prediction tasks.

More recently, the Segment Anything Model (SAM) [115] has emerged as a foundational model for image segmentation due to its remarkable segmentation capability and strong generalization across diverse domains. Owing to its powerful zero-shot and promptable segmentation abilities, SAM has attracted increasing attention in medical imaging applications [116-118], where it has been adapted for various tasks such as organ segmentation, lesion delineation, and interactive medical image analysis.

In summary, image segmentation methods have evolved from conventional CNN-based architectures to more advanced Transformer-based and foundation model-based frameworks, with continuous improvements in contextual modeling, multi-scale representation learning, and generalization capability. In Table 3, we list ten fundamental neural network architectures for image classification, semantic segmentation, object detection and image super-resolution. Beyond the overall macro network paradigms tailored for distinct vision tasks, many performances gains stem from carefully crafted intermediate building blocks. In the next section, we turn our focus to the design principles of effective sub-modules for feature learning, and dissect representative components that empower powerful multi-scale modeling, attention mechanisms, and local-global feature aggregation in deep vision models.

Table 3. Representative seminal neural architectures for computer vision tasks.

Model	Year	Main contribution
VGG	Demonstrated that deeper CNNs with small convolution kernels (3×3) significantly improve representation learning while maintaining a simple architecture.	Image classification, feature extraction
ResNet	Introduced residual learning through skip connections, enabling the effective training of very deep neural networks.	Image classification, detection, segmentation, vision backbone
MobileNetV2	Proposed inverted residual blocks with linear bottlenecks for lightweight and efficient mobile	Mobile vision, edge AI, real-time inference

Model	Year	Main contribution
	vision models.	
U-Net / nnU-Net	Encoder–decoder architecture with skip connections for precise dense prediction; nnU-Net provides a self-configuring framework for medical image segmentation.	Medical image segmentation
DeepLabv3+	Combines atrous convolution and atrous spatial pyramid pooling (ASPP) with an encoder–decoder framework for multi-scale semantic segmentation.	Semantic segmentation
GAN	Introduced adversarial learning through generator–discriminator competition for realistic image synthesis.	Image generation, image translation, super-resolution
Stable Diffusion	Latent diffusion model enabling high-quality and computationally efficient text-to-image generation.	Image synthesis, image editing, content generation
Mask R-CNN	Extended Faster R-CNN by adding a parallel mask prediction branch for instance segmentation.	Instance segmentation, object detection
YOLO Series	Established the one-stage detection paradigm for real-time object detection with successive improvements in speed and accuracy.	Real-time object detection
SAM	Foundation segmentation model trained on over one billion masks, enabling promptable and zero-shot image segmentation.	General-purpose segmentation, interactive annotation, vision foundation models

3. Principles of effective sub-modules for feature learning

3.1. Multiscale features extraction and fusion

Multi-scale (or multi-resolution) analysis has been widely adopted for feature extraction and feature fusion, with a long developmental history spanning from the era of handcrafted feature engineering to modern deep learning–based approaches. Classical handcrafted methods, such as the Scale-Invariant Feature Transform (SIFT) [119], employed Gaussian scale-space representations to achieve scale-invariant feature extraction, while multi-scale sparse representation frameworks further enhanced feature robustness across varying resolutions [120]. In deep learning, multi-scale analysis facilitates the learning of discriminative and scale-invariant representations by capturing complementary information from different spatial resolutions.

Image pyramids and feature pyramids are among the most widely adopted paradigms for multi-scale feature extraction. In early convolutional neural networks (CNNs), such as LeNet [121] and AlexNet [54], pooling operations were introduced to progressively reduce spatial resolution, thereby enabling the extraction of high-level semantic features for image classification tasks. Subsequently, SPP-Net [122] proposed a spatial pyramid pooling layer that aggregates multi-scale features through pooling operations with different grids and strides (see Figure 9). This design eliminates the constraint of fixed input resolutions, boosts the network’s robustness against object deformations, and mitigates overfitting, thereby achieving better recognition performance.

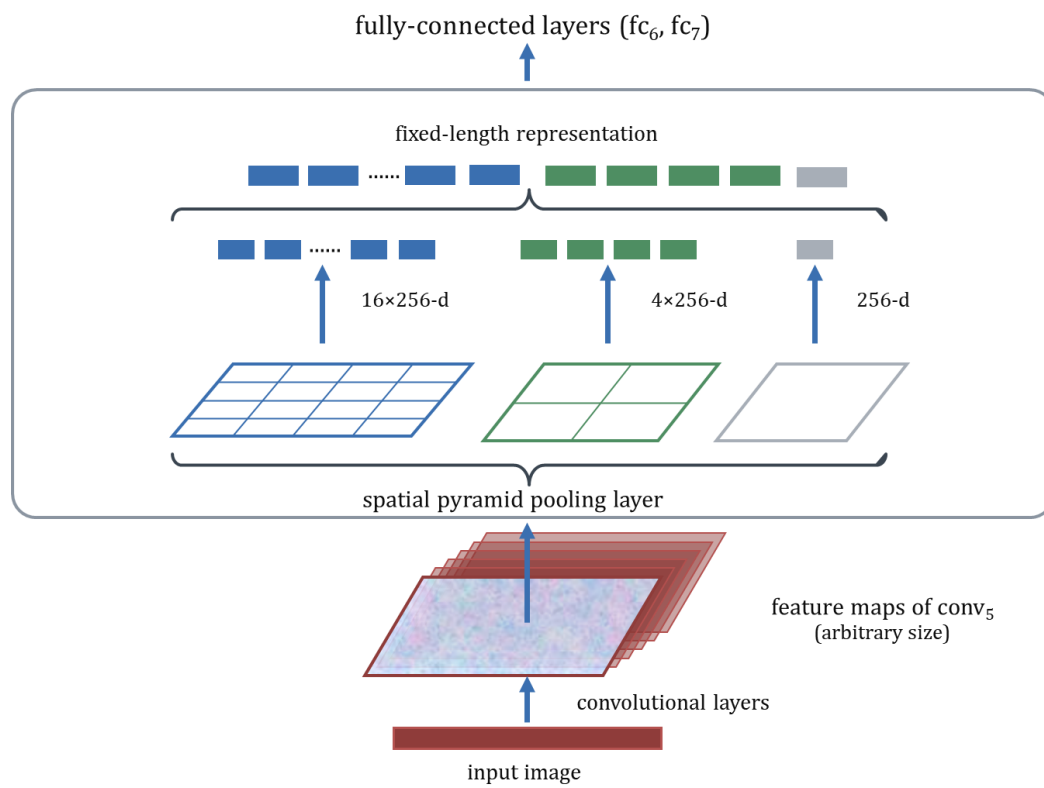


Figure 9. Architecture of SPP-Net

In object detection, the introduction of the Feature Pyramid Network (FPN) [132] marked a major advancement in multi-scale representation learning. FPN constructs a top-down feature fusion pathway with lateral connections, effectively combining high-resolution low-level features with semantically rich high-level features. This architecture significantly improves the detection of objects at different scales, particularly small objects, and has subsequently become a fundamental component in numerous modern detection frameworks. For example, YOLOv6 (see Figure 10) [89] incorporates the Cross Stage Partial (CSP) block [123], which adopts both top-down and bottom-up

feature fusion pathways to facilitate more effective information propagation and feature interaction across different resolutions. Specifically, high-level semantic features from low-resolution layers are fused with spatially detailed features from high-resolution layers, thereby improving feature representation capability while reducing computational redundancy. In EfficientDet [124], the importance of multi-scale feature fusion was further emphasized through the introduction of the weighted bidirectional feature pyramid network (BiFPN, see Figure 11). Unlike conventional FPN or CSP-based fusion approaches that directly perform feature summation across scales while implicitly assuming equal feature importance, BiFPN introduces learnable weights to adaptively model the contributions of features at different resolutions. By repeatedly performing both top-down and bottom-up bidirectional feature fusion, BiFPN achieves more efficient multi-scale information integration and significantly improves the trade-off between detection accuracy and computational efficiency.

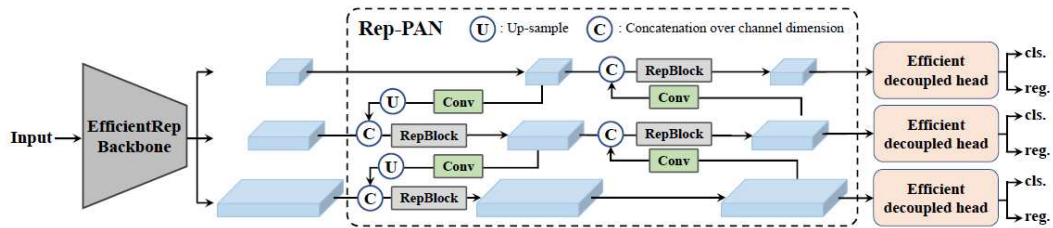


Figure 10. Structure of YOLOv6

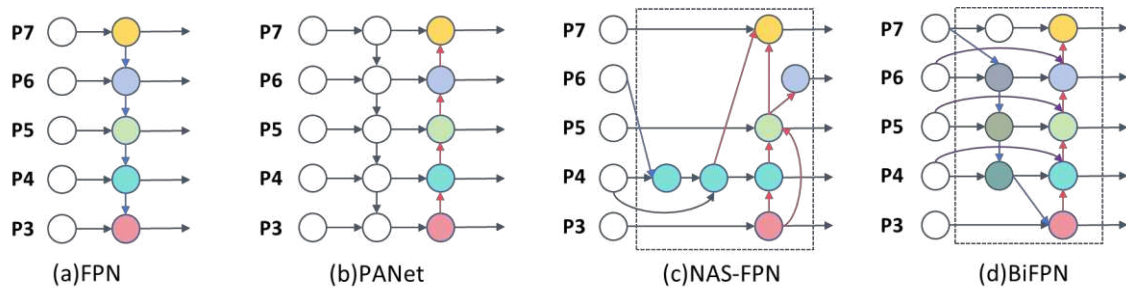


Figure 11. Comparison of classic multi-scale feature fusion modules: (a) FPN, (b) PANet, (c) NAS-FPN, and (d) BiFPN. P3–P7 represent multi-scale pyramid features from the backbone. FPN adopts top-down fusion; PANet introduces an extra bottom-up path; NAS-FPN uses NAS-discovered cross-scale connections; BiFPN implements bidirectional weighted feature fusion.

Following the principle of multi-scale feature extraction and fusion, numerous studies have been proposed for image segmentation. In HRNet [107], multi-scale features from different CNN stages are gradually increased and fused in parallel, enabling high-resolution representations to be maintained throughout the entire network. Unlike backbones such as ResNet or VGG, which

progressively discard high-resolution information through repeated pooling operations, HRNet continuously exchanges information across high- and low-resolution branches. This design has demonstrated strong effectiveness in dense prediction tasks such as semantic segmentation. Similarly, UNet++ [99] adopts the spirit of multi-scale feature learning and fusion with skip connections, enabling multi-scale feature aggregation between the encoder and decoder in an ensemble-like manner.

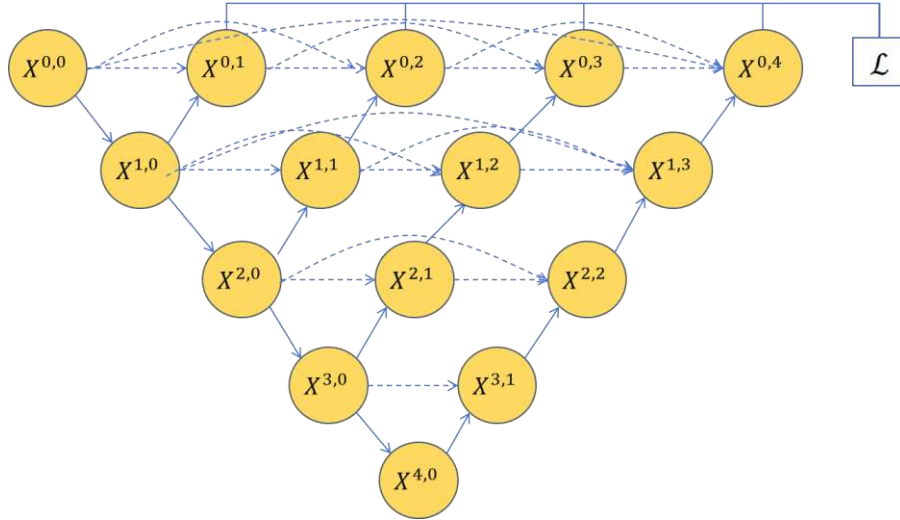


Figure 12. Architecture of UNet++

Different from HRNet and UNet++, which redesign multi-scale representation learning at the architectural level, several studies have revisited multi-scale feature extraction from a module-level perspective. For example, Res2Net [135] introduces a residual-like building block that performs hierarchical feature splitting and fusion within a single module, thereby enhancing the receptive field and enabling finer-grained multi-scale representation learning. Following similar principles, a variety of lightweight and plug-and-play modules have also been proposed to improve multi-scale feature representation while maintaining computational efficiency, such as PSPNet [106], SegNeXt [125], EfficientViT [126] and DeepLab [102].

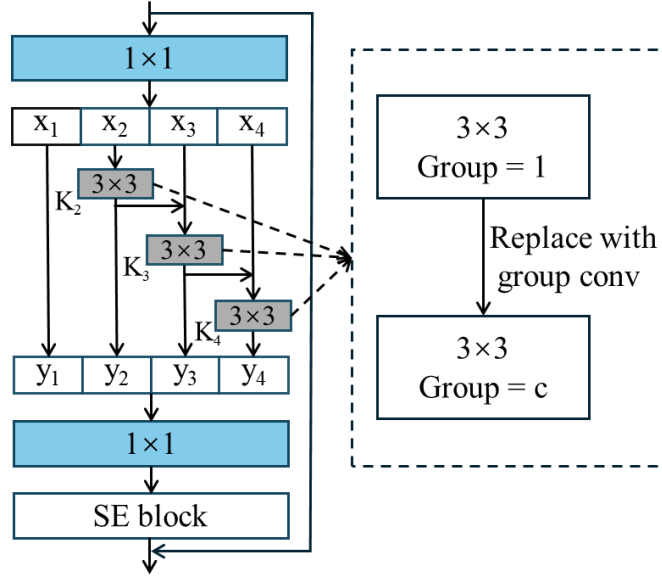


Figure 13. Residual-like building block of Res2Net. This block captures multi-scale visual information by partitioning feature maps into multiple parallel branches and conducting sequential convolutions to build hierarchical feature representations.

In summary, multi-scale feature processing has become an effective and widely adopted strategy in computer vision, as it enables models to capture high-level semantic information while simultaneously preserving fine-grained feature representations. Such capability is particularly important for dense prediction tasks, including image segmentation, where objects and structures may appear at substantially different scales. However, the integration and fusion of multi-scale features often introduce additional computational and memory overhead. Therefore, achieving an appropriate balance between segmentation accuracy and computational efficiency remains an important and ongoing research topic in multi-scale feature learning.

3.2. Attention mechanism

Similar to the human visual system, which naturally focuses more on salient objects and regions of interest, attention mechanisms have been introduced to guide deep learning models toward informative features while suppressing irrelevant background information. For example, Attention U-Net [127] incorporates a series of attention gates to suppress unimportant regions and highlight contextual information related to target structures during segmentation. Extensive studies have demonstrated both its effectiveness and interpretability of attention mechanism, as the visualization of attention maps provides intuitive insights into how the network emphasizes different spatial regions [92, 128, 129]. In this way, attention mechanisms can help models learn more

discriminative representations associated with target objects.

Generally, attention mechanisms can be categorized into three main types: channel attention, spatial attention (with or without channel interaction), and self-attention. Channel attention aims to assign adaptive importance weights to different feature channels. For instance, SENet [130] proposed the “Squeeze-and-Excitation” (SE) block, which models inter-channel dependencies and recalibrates channel-wise feature responses. Specifically, the SE block first performs spatial dimension reduction using global average pooling (the squeeze operation). The resulting channel descriptors are then passed through two fully connected layers with a Rectified Linear Unit activation function to capture channel-wise dependencies and generate adaptive weights. Finally, the learned channel-wise weights are multiplied with the corresponding feature maps to enhance informative channels while suppressing less relevant ones. Following the same spirit, ECA-Net [131] proposed a lightweight and efficient channel attention mechanism that avoids dimensionality reduction while capturing local cross-channel interactions. FcaNet [132] introduced frequency channel attention by incorporating frequency-domain information into channel representation learning, further enhancing the discriminative capability of channel attention mechanisms. SKNet [133] employs multiple convolutional branches with different kernel sizes to capture multi-scale contextual information and adaptively learns channel attention weights for each branch before feature fusion. This dynamic selection mechanism enables the network to effectively integrate multi-scale representations, thereby enhancing the discriminative power and representational capability of the learned features.

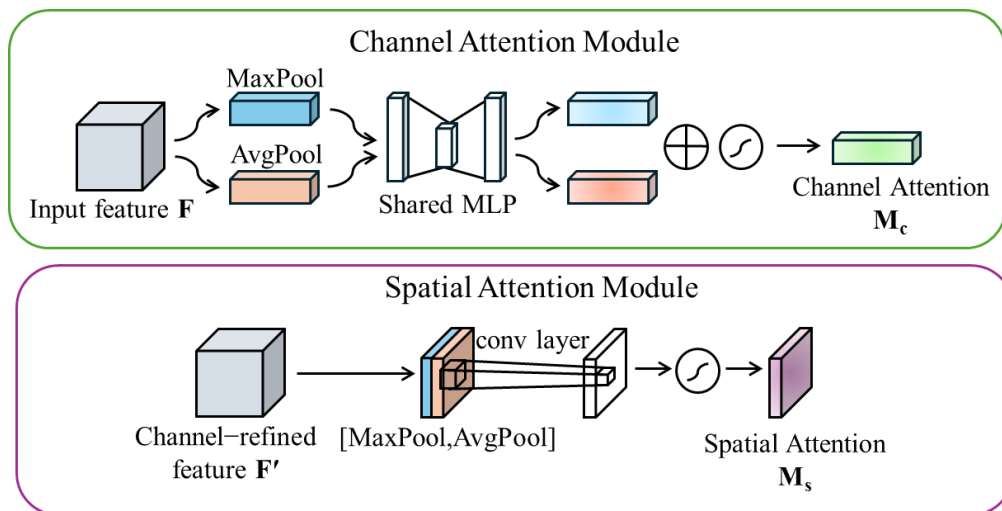


Figure 14. CBAM

Different from channel attention, which learns a one-dimensional channel-wise weighting vector to model the importance of individual feature channels, spatial attention aims to generate a two-dimensional spatial weighting map that emphasizes informative or target regions while suppressing irrelevant background responses. Spatial attention is often combined with channel attention to jointly exploit complementary channel-wise and spatial contextual information, thereby enhancing the representational capacity and discriminative power of learned features. Representative studies in this direction include BAM [134], CBAM [135], and GSoP [136]. Among them, CBAM (see Figure 14) sequentially applies channel attention and spatial attention to intermediate feature maps, enabling the network to adaptively identify both the most discriminative feature channels and the most salient spatial regions. Owing to its lightweight and plug-and-play design, CBAM can be readily integrated into existing CNN architectures to refine feature representations with minimal computational overhead. Overall, these spatial attention mechanisms enhance feature learning by modeling spatial dependencies and selectively emphasizing semantically important regions, thereby improving the discriminative capability of deep feature maps.

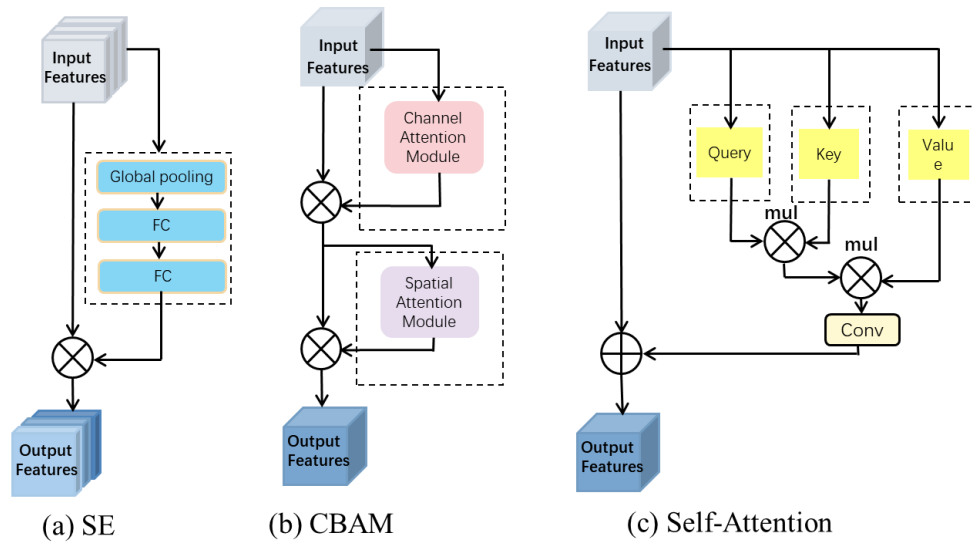


Figure 15. Comparison of classic attention mechanisms: (a) SE module for channel feature recalibration; (b) CBAM, combining sequential channel and spatial attention; (c) Self-Attention, modeling long-range dependencies with query-key-value computation.

In terms of self-attention, it calculates pairwise similarities across all elements in feature maps via the dot product of query and key matrices (See Figure 15). This mechanism empowers the model to capture long-range dependencies and contextual correlations, rendering self-attention highly

effective for tasks that demand thorough contextual reasoning. As a representative exploration of self-attention in computer vision, Non-local Neural Networks [137] introduce a universal non-local block (self-attention-based) to capture long-range dependencies from images and videos. This module can be flexibly embedded into conventional CNN or RNN architectures to aggregate global contextual information from sequential feature representations for diverse vision tasks, including video classification and object detection. Nevertheless, the exhaustive pairwise similarity calculation across all spatial locations leads to substantial computational overhead. Specifically, computing full spatial attention for every pixel generates an attention map of size $hw \times hw$ (suppose the spatial feature map size is h by w), resulting in quadratic computational complexity that severely limits scalability. To alleviate this limitation, non-local blocks are often combined with local convolutional operations or feature downsampling strategies to reduce spatial resolution before attention computation, thereby lowering computational complexity at the expense of fine-grained global modeling capability. To further improve efficiency, CCNet [110] proposes a criss-cross attention mechanism that captures contextual information along the horizontal and vertical directions for each pixel. By recurrently propagating criss-cross attention, the network effectively aggregates global contextual dependencies while reducing the computational cost by approximately 85% compared with the original non-local block, achieving a more favorable balance between efficiency and representational power. Despite its high efficiency in contextual aggregation, CCNet is inherently built upon CNN-based architectures, which inevitably limits its ability to model long-range and complex semantic relationships.

Unlike previous methods such as Non-Local Networks and CCNet, which integrate self-attention modules into convolutional networks, the Vision Transformer (ViT) [128] employs a pure transformer architecture based entirely on self-attention. It demonstrated that such an architecture can achieve performance comparable to or exceeding that of state-of-the-art CNN-based models, especially when trained on large-scale datasets. Specifically, ViT partitions an input image into fixed-size patches (e.g., 16×16 grids) and treats each patch as a visual token, similar to words within a sentence. Each patch is then flattened and projected into a latent embedding through a linear projection layer. Positional embeddings are added to preserve spatial information, and the resulting token sequence is processed by a stack of Transformer encoders, where global dependencies among all patches are modeled through multi-head self-attention. Following the success of ViT, a series of

transformer-based vision models have been proposed [138]. By leveraging self-attention, these models effectively capture long-range dependencies and learn highly representative feature embeddings. However, compared with CNNs, they lack strong inductive biases, such as locality and translation equivariance, making them more dependent on large-scale training data. To alleviate this limitation, self-supervised pre-training approaches, such as DINOv3 and MAE, have been developed and have demonstrated that pre-trained ViT models can serve as powerful feature extractors for a wide range of downstream tasks, including object detection and image segmentation.

3.3. Local and global feature learning based on Convolution and Transformer

The preceding sub-sections have discussed multi-scale feature extraction and attention mechanisms for learning representative feature embeddings. Among them, transformer-based models have demonstrated remarkable representation capability, particularly when pretrained on large-scale datasets. However, their computational complexity grows rapidly with the number of tokens, as self-attention requires pairwise interactions between queries and keys, followed by feature aggregation through fully connected layers (see Figure 16). Consequently, the computational and memory costs become prohibitive for high-resolution images. Moreover, the limited inductive biases of transformers make them heavily reliant on large-scale training data, which is often unavailable in many practical applications, such as autonomous driving and medical image analysis [46, 48].

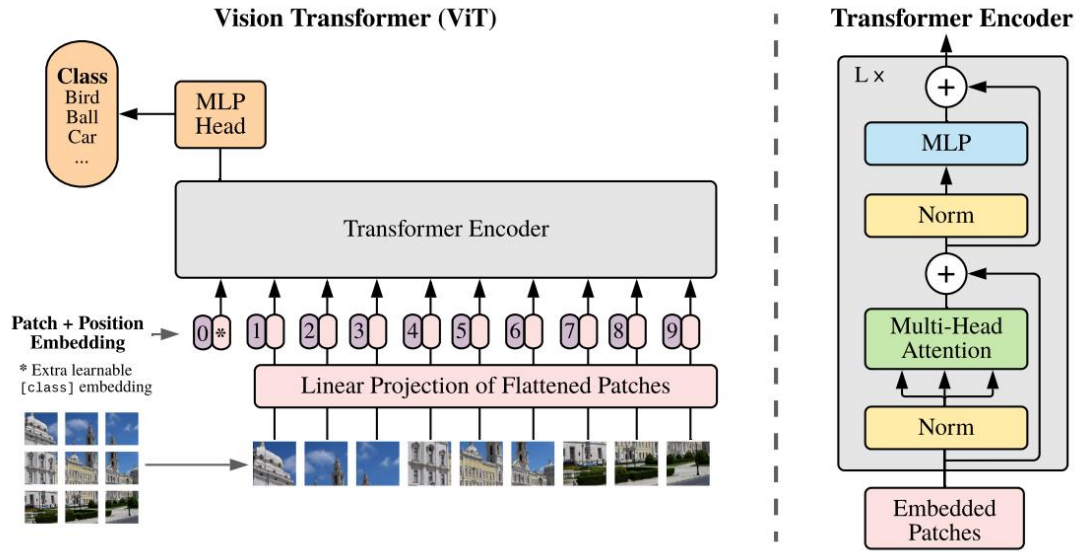


Figure 16. Illustration of ViT. Source: reproduced from [128]

To address these limitations, an effective strategy is to incorporate locality as an inductive bias into the transformer architecture for computer vision. A representative example is the Swin Transformer (SwinT) [65]. Unlike ViT, which partitions an image into non-overlapping patches and performs global self-attention across all patches, SwinT restricts self-attention computation to local windows, substantially reducing computational complexity while preserving local structural information. To facilitate information exchange across neighboring windows, a shifted window mechanism is introduced, enabling cross-window interactions without incurring the cost of global attention (see Figure 17). Furthermore, SwinT adopts a hierarchical architecture in which feature maps are progressively downsampled through multiple stages, producing multi-scale feature representations that are well suited for downstream vision tasks such as detection and segmentation.

Following this line of research, numerous variants of the Swin Transformer have been proposed for different applications. For example, Swin Transformer V2 [139] scales model capacity and input resolution to improve representation learning, Video Swin Transformer [140] adapts the architecture for video recognition by balancing the trade-off between accuracy and efficiency, SparseSwin introduces sparse transformer blocks with a limited number of latent tokens to further reduce computational cost, and Swin-Unet [141] integrates Swin Transformer blocks into the U-Net architecture for medical image segmentation. Collectively, these works incorporate locality into quasi-global self-attention to achieve a better balance between computational efficiency, data efficiency, and model performance.

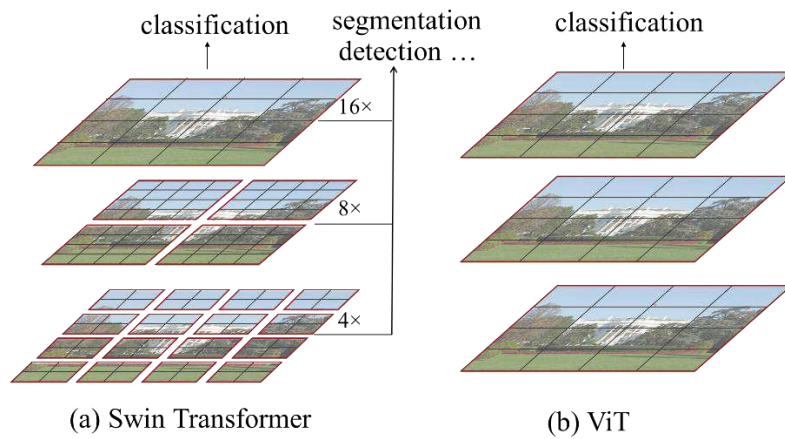


Figure 17. Comparison between ViT and Swin Transformer. ViT performs global self-attention on all image patches, while Swin Transformer implements hierarchical windowed self-attention with shifted windows to capture both local and long-range dependencies at lower computational cost.

Aside from Swin Transformer variants that introduce local modeling via fixed-size windows, another prominent research direction focuses on unifying convolutional and Transformer modules within a single architecture. This hybrid design jointly captures local and global dependencies, striking a favorable balance between segmentation accuracy and computational efficiency. A representative example is LeViT [142], which first employs convolutional layers to generate feature maps with a $16\times$ reduced spatial resolution and subsequently applies a lightweight transformer module, where the query features are spatially subsampled, to model global context. By performing self-attention on low-resolution feature maps, LeViT significantly reduces the computational overhead associated with conventional transformers. Building upon this backbone, LeViT-UNet [63] incorporates multi-stage transformer blocks into a U-Net architecture and fuses multi-scale transformer features with convolutional representations in the decoder, enabling effective integration of global contextual information and local spatial details for medical image segmentation. Other representative architectures, including TransUNet [143], MobileFormer [144], and MobileViT [145], follow a similar hybrid paradigm by coupling convolution and transformer modules in either serial, parallel or alternative configurations. These hybrid designs exploit the strong inductive biases of convolutional networks while leveraging the long-range modeling capability of self-attention, thereby improving computational efficiency and reducing the dependence on large-scale annotated datasets compared with pure transformer-based models.

3.4. Ten effective feature learning methods

The preceding sections have reviewed representative approaches to discriminative feature learning from several complementary perspectives, including multi-scale representation, receptive-field expansion, adaptive feature extraction, attention mechanisms, and local-global feature interaction. Building upon these developments, Table 4 summarizes ten representative strategies that have been widely adopted to enhance feature discriminability in deep neural networks. These strategies capture several fundamental design principles of modern vision architectures, including multi-scale modeling, contextual information aggregation, adaptive spatial sampling, computationally efficient feature extraction, information propagation, attention-based feature reweighting, global dependency modeling, and multi-level supervision. Together, they provide a concise taxonomy of commonly used mechanisms for learning more discriminative and robust visual representations.

Table 4. the effective methods for discriminative features learning

Strategy	Description	Representative works
Multi-scale Pyramid Representation	Constructs hierarchical feature representations at multiple spatial resolutions to capture objects and structures with different scales and receptive fields.	FPN; PSPNet; HRNet
Multi-branch Convolution	Employs parallel convolutional branches with different kernel sizes or receptive fields to capture complementary spatial patterns and contextual information.	Inception/GoogLeNet; Inception-v2/v3
Dilated Convolution	Expands the receptive field by introducing gaps between convolutional kernel elements while preserving spatial resolution, enabling efficient contextual modeling.	DilatedNet; DeepLab
Deformable Convolution	Dynamically learns spatial offsets for convolutional sampling locations, allowing the receptive field to adapt to object geometry and irregular structures.	Deformable ConvNets; Deformable ConvNets v2
Depthwise Separable Convolution	Factorizes standard convolution into depthwise and pointwise operations, substantially reducing computational complexity while retaining effective spatial feature extraction.	Xception; MobileNet

Strategy	Description	Representative works
Residual/Skip Connection	Facilitates information and gradient propagation by directly connecting shallow and deep representations, enabling effective feature reuse and mitigating degradation during deep network optimization.	ResNet; U-Net; DenseNet
Channel Attention	Learns channel-wise importance weights to emphasize informative feature channels and suppress less relevant responses.	SENet; ECA-Net; CBAM
Spatial Attention	Learns spatially varying attention maps to selectively emphasize informative regions and suppress irrelevant background responses.	CBAM; BAM; DANet
Self-Attention	Explicitly models long-range dependencies by computing interactions between distant spatial positions or tokens, enabling global contextual representation.	Non-local Networks; ViT; Swin Transformer
Multi-level Supervision	Introduces auxiliary supervision at intermediate network layers or decoder stages to provide additional optimization signals and encourage discriminative representations at multiple levels.	DSN; U-Net++; deeply supervised segmentation networks

4. Strategies for constructing high-performance deep neural networks

In the preceding sections, we reviewed representative studies across various computer vision tasks and discussed different strategies for feature learning in deep neural networks (DNNs). Based on these developments, this section summarizes several general principles for designing effective neural network architectures for computer vision.

4.1. Residual learning or Skip Connection

Residual learning has become one of the fundamental design principles in image processing, with its origins tracing back to classical image processing methods [146] and its widespread adoption in modern deep learning architectures [52]. In particular, ResNet [58] introduced residual connections through identity skip pathways, enabling networks to learn residual mappings instead of direct transformations. This simple yet effective design significantly alleviates the optimization difficulties associated with training very deep networks by facilitating gradient propagation and

mitigating performance degradation as network depth increases. Owing to its effectiveness, residual learning has evolved into a standard architectural component, serving as a cornerstone not only in CNN-based models such as SENet [130], ResNeXt [147], and ConvNeXt [62], but also in transformer-based architectures like ViT [128], SwinT [65], and DETR [92], where residual connections are extensively employed to stabilize training and improve representation learning.

In our previous survey [52], we systematically reviewed the evolution of residual learning in deep neural networks, covering the diverse forms of skip connections employed across modern neural architectures. In general, residual learning has evolved beyond its vanilla ResNet, and has been integrated into a variety of architectural designs. First, it is widely combined with bottleneck structures, where the intermediate feature dimensions differ from those of the input and output, enabling lightweight and efficient network design, as exemplified by MobileNetV2 [67] and MobileNeXt [148]. Second, residual connections play a crucial role in multi-scale feature propagation and fusion, facilitating information exchange across different resolutions in architectures such as FCN [96] and Res2Net [135]. Third, they are extensively incorporated into multi-branch architectures, allowing complementary feature representations to be learned within a single residual block, as demonstrated by Inception-v4 [60] and ResNeXt [61]. Finally, residual learning has been naturally integrated with attention mechanisms to improve feature representation and optimization stability, as evidenced by recent architectures such as SegNeXt [125] and ViT [128]. These developments highlight the versatility of residual learning as a fundamental design principle that enhances feature representation, optimization efficiency, and architectural flexibility across both convolutional and transformer-based networks.

4.2. Prior knowledge into computer vision

Prior knowledge provides valuable inductive biases that guide model optimization and improve generalization by introducing additional constraints or regularization during training. In general, prior knowledge can be incorporated into deep neural networks through two primary strategies: (1) objective function design and (2) feature normalization or representation learning.

(1). Objective-Function Design

Integrating prior knowledge into the objective function is one of the most direct and effective approaches, as it explicitly influences the optimization process. A common strategy is to assign

class-specific weights in the cross-entropy loss according to the class distribution in the training set, thereby alleviating the adverse effects of severe class imbalance [149]. Beyond static weighting schemes, adaptive loss functions have been developed to further emphasize informative training samples. A representative example is Focal Loss (FL) [150], originally proposed for RetinaNet to address the severe easy and hard object class imbalance encountered in dense object detection. Focal Loss extends the standard cross-entropy loss with a modulating factor that dynamically attenuates the loss weights of easily classified samples and elevates the contribution of hard, misclassified instances. Consequently, the optimization process focuses more on informative samples, leading to improved detection performance. Building upon this idea, Generalized Focal Loss (GFL) [151] further extends the discrete formulation of Focal Loss to continuous probability distributions, enabling the joint optimization of classification confidence and localization quality within a unified framework.

For metric learning tasks, the core optimization target is to shrink feature distances within identical categories while enlarging inter-class feature discrepancies. Triplet loss [152], originally proposed within FaceNet for face recognition and clustering, fulfills this objective by constraining embedding space distances during training. It constructs a triplet unit consisting of one anchor sample, one positive sample sharing the identical label as the anchor, and one negative sample from a distinct category. Explicitly taking the anchor as a reference, the loss enforces embeddings of homogeneous samples to cluster tightly and pushes embeddings of heterogeneous samples far apart in latent space. Compared with pairwise contrastive loss (two-sample loss) that merely computes mean squared error between two individual embeddings, triplet loss exhibits stronger capability to capture fine-grained inter-sample distinctions. It has inspired abundant follow-up research across numerous vision tasks, including person re-identification, content-based image retrieval, and few-shot classification.

$$L_{triplet} = \frac{1}{N} \sum_{i=1}^N \max (\|f(a_i) - f(p_i)\|_2^2 - \|f(a_i) - f(n_i)\|_2^2 + m, 0) \quad (7)$$

In image segmentation, conventional pixel-wise optimization based on cross-entropy loss treats each pixel independently and fails to explicitly account for the spatial structure, shape, and regional continuity of target objects. Moreover, the severe imbalance between foreground and background pixels often biases the optimization toward the dominant class, particularly in medical

image segmentation. To address these limitations, region-based and boundary-aware loss functions have been proposed to incorporate higher-level structural priors into the optimization process. Representative examples include Dice Loss, which directly maximizes region overlap and effectively alleviates class imbalance, and Boundary Loss [153], which explicitly constrains contour alignment to improve boundary delineation. In addition, numerous hybrid loss functions have been developed by combining pixel-wise, region-based, and boundary-aware objectives, enabling models to jointly optimize semantic consistency, shape preservation, and contour accuracy for more robust segmentation performance.

In addition, various forms of prior knowledge, such as sparsity constraints, feature fidelity, low-rank or subspace representations, and structural regularization, have inspired the development of numerous optimization objectives. Representative examples include sparse regularization, feature reconstruction losses, perceptual losses, low-rank constraints, and manifold-preserving objectives, all of which explicitly encode domain-specific priors into the learning process. By guiding optimization toward more meaningful and physically plausible solutions, these regularization strategies improve feature representation, enhance model generalization, and increase robustness, particularly in scenarios with limited or noisy training data. Collectively, these studies demonstrate that incorporating task-specific prior knowledge into the objective function is an effective strategy for improving optimization and enhancing model performance.

Beyond the above-mentioned families, diverse prior assumptions—encompassing sparsity constraints, feature fidelity requirements, low-rank (subspace structure), and topological regularization—have spawned a rich repertoire of optimization objectives. Representative examples include sparse regularization terms, feature-reconstruction losses, perceptual losses, low-rank penalties, and manifold-preserving objectives, all of which bake domain-specific inductive priors into the training loop. By biasing optimization toward semantically meaningful and physically plausible solutions, these regularizers refine feature representations, elevate generalization capacity, and strengthen model robustness, especially under regimes characterized by scarce or noisy training annotations. Taken together, existing literature corroborates that integrating task-aware prior knowledge into loss-function design represents a powerful instrument for stabilizing optimization and elevating downstream model performance.

(2) Normalization and Representation Learning

A second major strategy for incorporating prior knowledge is to regulate the statistical properties and structural characteristics of intermediate feature representations. Normalization techniques were initially developed primarily to stabilize and accelerate network optimization, but they can also be interpreted as imposing implicit priors on feature distributions. By controlling the mean, variance, or other statistical properties of intermediate activations, normalization reduces undesirable distributional variations and facilitates more stable gradient propagation.

Batch Normalization (BN) [154] is one of the most influential normalization techniques and normalizes feature activations using statistics estimated from a mini-batch. By reducing internal distributional variation and improving the conditioning of the optimization problem, BN enables the use of larger learning rates and generally accelerates convergence. However, its reliance on batch-level statistics makes its effectiveness sensitive to batch size and data distribution. This limitation is particularly relevant in applications such as high-resolution medical image segmentation, where GPU memory constraints often necessitate small batch sizes.

To alleviate the dependence on batch statistics, Layer Normalization (LN) [155] normalizes activations across the feature dimension of individual samples rather than across a mini-batch. LN is therefore independent of batch size and has become a fundamental component of Transformer architectures. In contrast, Instance Normalization (IN) [156] computes normalization statistics independently for each sample and channel, making it particularly suitable for tasks in which instance-specific appearance variations should be suppressed. IN has consequently been widely adopted in image generation, style transfer, and domain adaptation. Group Normalization (GN) [157] provides an intermediate formulation by dividing channels into groups and computing normalization statistics within each group. Unlike BN, GN remains effective under small-batch training and has therefore been widely used in dense prediction tasks, including object detection and medical image segmentation. More broadly, normalization strategies can be viewed as mechanisms for introducing assumptions regarding the desirable statistical structure of learned representations. Different normalization schemes impose different invariances and retain or suppress different forms of information. Consequently, the choice of normalization is not merely an implementation detail but can influence the representation capacity, optimization stability, domain robustness, and task-specific information preserved by the network.

Beyond conventional normalization, representation learning with prior knowledge provides

another means by explicitly encouraging learned features to exhibit desirable properties. Examples include feature reconstruction, prototype-based representation learning, contrastive objectives, disentangled representations, and domain-invariant feature learning. These approaches impose constraints on the latent representation space so that semantically related samples exhibit consistent representations while irrelevant variations are suppressed. In this sense, representation-level priors complement objective-level priors: rather than modifying only the final optimization criterion, they directly regulate the intermediate representations from which predictions are generated.

Overall, objective-function design and normalization/representation learning provide two complementary mechanisms for incorporating prior knowledge into deep neural networks. The former primarily constrains what constitutes an optimal solution, whereas the latter regulates how information is represented and propagated throughout the network. This distinction has motivated increasingly sophisticated optimization frameworks that integrate task-specific priors at multiple levels, ranging from the loss function and feature distribution to the network architecture and learned latent space.

4.3. Contrast-Based Deep Learning and Cross-Modal Representation Alignment

Contrastive learning has emerged as a fundamental paradigm for representation learning, particularly in the development of self-supervised and multimodal foundation models. Its central principle is to learn an embedding space in which semantically related samples are encouraged to exhibit similar representations, whereas unrelated samples are separated. Unlike conventional supervised learning, which directly maps an input to a predefined class label, contrastive objectives exploit relationships among samples, views, or modalities as the primary source of supervision. This property enables representation learning from large collections of unlabeled or weakly labeled data and facilitates knowledge transfer across downstream tasks and domains [158, 159].

A major branch of contrastive learning is instance-level and class-level discriminative representation learning, in which positive and negative relationships are explicitly constructed. Early Siamese architectures demonstrated that paired inputs could be embedded into a metric space by optimizing their similarity or dissimilarity, providing an important foundation for subsequent metric and contrastive learning approaches [160]. More recently, self-supervised methods have substantially advanced this paradigm. SimCLR constructs positive pairs from different augmented

views of the same image and treats other samples within a mini-batch as negatives, demonstrating the importance of strong data augmentation, nonlinear projection heads, and sufficiently large batches [158]. MoCo addresses the requirement for large numbers of negative samples through a dynamic dictionary implemented using a queue and a momentum-updated encoder, thereby decoupling the dictionary size from the mini-batch size [159]. These developments established contrastive learning as an effective mechanism for learning transferable visual representations without manual annotations. Supervised Contrastive Learning subsequently incorporated class labels into the contrastive objective, allowing multiple samples from the same class to form positive sets rather than relying on a single positive pair [161].

The contrastive paradigm has subsequently evolved beyond explicit negative sampling. DINO formulates self-supervised learning as a form of self-distillation, where a student network is trained to match the representations produced by a teacher network under different image views [162]. Although DINO does not rely on conventional positive-negative contrastive pairs, it preserves the broader principle of relational consistency across views. Similarly, Barlow Twins learns invariant representations by aligning the cross-correlation matrix of two augmented views with the identity matrix, thereby combining invariance with redundancy reduction without requiring explicit negative samples [163]. These approaches indicate that modern representation learning increasingly extends beyond the classical formulation of “pull positives together and push negatives apart” toward a broader family of comparison- and alignment-based objectives.

A closely related direction is consistency-based learning, which seeks to maintain stable predictions or representations under perturbations, model updates, or different network branches. This principle is particularly important in semi-supervised learning, where unlabeled samples provide supervision through consistency constraints. Mean Teacher, for example, constructs a teacher model using an exponential moving average of the student parameters and encourages the student to match the teacher's predictions [164]. In semantic segmentation, UniMatch further extends weak-to-strong consistency by combining image-level and feature-level perturbations, demonstrating that consistency across diverse perturbation spaces can effectively exploit unlabeled images [165]. Although consistency learning does not necessarily employ an explicit contrastive loss, it shares the fundamental objective of enforcing semantic agreement between multiple views of the same underlying sample. Consequently, contrastive learning and consistency learning can be

785 regarded as complementary mechanisms within a broader comparison-based representation-
786 learning framework.

787 Contrastive objectives have also become central to cross-modal representation alignment,
788 particularly for vision-language learning. CLIP demonstrated that large-scale image-text pairs can
789 be used to learn a shared embedding space by maximizing the correspondence between matched
790 images and textual descriptions while suppressing mismatched pairs [9]. Trained on approximately
791 400 million image-text pairs, CLIP showed that language can provide a flexible supervisory signal
792 for visual representation learning and enable zero-shot transfer to diverse downstream recognition
793 tasks. Subsequent approaches have increasingly integrated alignment with cross-modal interaction.
794 ALBEF first aligns visual and textual representations using contrastive learning and subsequently
795 fuses them through cross-modal attention, while momentum distillation provides additional
796 supervision for noisy web-scale image-text data [166]. BLIP further combines contrastive learning,
797 image-text matching, and language modeling within a unified vision-language pre-training
798 framework, illustrating the transition from pure embedding alignment toward joint multimodal
799 understanding and generation [167].

800 The same principle has been extended to modalities beyond image and text. ImageBind learns
801 a shared embedding space across six modalities, including images, text, audio, depth, thermal
802 signals, and inertial measurements. Importantly, it demonstrates that complete pairwise supervision
803 across all modalities is not necessary; image-paired data can serve as a common semantic bridge for
804 aligning otherwise unpaired modalities [168]. This observation suggests a broader interpretation of
805 multimodal contrastive learning: rather than merely matching two modalities, the objective can
806 establish a common representational geometry in which heterogeneous sensory signals become
807 mutually accessible. Such a formulation is particularly relevant to multimodal foundation models,
808 where scalable alignment across heterogeneous data sources is increasingly important.

809 Cross-modal contrastive learning has also become increasingly influential in medical artificial
810 intelligence, where paired data are substantially more limited and semantically heterogeneous than
811 web-scale image-text data. GLoRIA introduced global-local contrastive learning between
812 radiological images and report text, enabling both global image-report alignment and local
813 correspondence between image regions and textual concepts [169]. MedCLIP further addressed the
814 scarcity of paired medical image-text datasets by decoupling images and reports and introducing a

semantic matching objective designed to reduce false negatives caused by clinically similar but independently acquired samples [170]. These studies illustrate that directly transferring generic contrastive objectives to specialized domains may be insufficient; domain knowledge, semantic granularity, and the structure of cross-modal correspondences must also be considered when constructing positive and negative relationships.

Recent developments have further questioned the necessity of conventional softmax-based contrastive objectives. SigLIP replaces the globally normalized contrastive loss used by CLIP with a pairwise sigmoid loss, allowing image-text pairs to be optimized independently and reducing the dependence on extremely large global batches [171]. This line of research highlights an important evolution of contrastive learning from a specific loss formulation toward a general principle of semantic alignment under scalable and heterogeneous supervision. The objective may involve explicit negative samples, multiple positive samples, self-distillation, redundancy reduction, or pairwise cross-modal matching, depending on the characteristics of the learning problem.

It is therefore useful to distinguish contrastive learning, consistency learning, and other comparison-based objectives, rather than treating them as interchangeable concepts. Contrastive learning explicitly models relative similarity or dissimilarity among representations; consistency learning emphasizes invariance of predictions or representations under transformations or across models; adversarial learning instead optimizes a minimax game between competing networks; and diffusion-based generative learning estimates a denoising or score-matching objective over progressively corrupted data [172, 173]. Although these paradigms all involve comparisons between different states of data or models, their optimization principles and statistical interpretations are fundamentally different. Maintaining this distinction is important for establishing a rigorous taxonomy of modern deep learning methods.

Overall, contrastive and alignment-based learning have evolved from instance-level discrimination into a broader framework for learning relationships among samples, views, models, and modalities. This evolution has contributed to the emergence of self-supervised visual representation learning, semi-supervised consistency learning, vision-language pre-training, and multimodal foundation models. The central methodological shift is from learning solely from explicit labels toward learning from relationships and correspondences embedded in large-scale data. Future progress is likely to depend not only on designing stronger contrastive objectives, but also

on constructing more reliable positive and negative relationships, mitigating false negatives and semantic ambiguity, exploiting domain-specific knowledge, and establishing robust alignment across increasingly heterogeneous modalities.

5. Discussion and future work

(1). Efficient architecture for DNN

As model scale continues to grow, computational efficiency has become a critical consideration in the design of modern deep learning systems. Consequently, a series of efficient architectures, including SqueezeNet [174], ShuffleNet [68], EfficientNet [175], LeViT [75], EdgeViT [71], and EfficientViT [126], have been proposed to reduce computational and memory overhead. Despite their different implementations, these methods generally follow several common principles, including parameter reduction through decomposition and compression, efficient feature representation, lightweight operator design, and the incorporation of inductive biases to reduce the cost of global feature modeling. These strategies collectively enable a better trade-off between model accuracy and computational efficiency. However, most existing approaches are designed for task-specific and relatively lightweight models. As vision foundation models continue to scale in size and capability, the principles governing their efficient design remain less well understood, and a comprehensive theoretical framework for efficiency-oriented foundation model design is still lacking.

Recent advances in large language models (LLMs) have provided valuable insights into the development of efficient large-scale models. Representative examples include the DeepSeek series. DeepSeekMoE [176] adopts a Mixture-of-Experts (MoE) architecture that activates only a subset of experts for each input token, thereby reducing redundant computation while enabling specialized knowledge modeling. Building upon this foundation, DeepSeek-V2 [177] introduces Multi-head Latent Attention (MLA), which employs low-rank representations to compress the key-value cache into a compact latent space, significantly reducing memory consumption during inference. Furthermore, DeepSeek-V3 [178] incorporates multi-token prediction to improve training efficiency and inference throughput. Collectively, these innovations suggest two promising

principles for efficient foundation model design: sparsity, which reduces computation through selective activation and modularized representations, and compression, which minimizes memory and computational costs by learning compact feature representations. These principles may provide valuable guidance for the development of next-generation efficient vision foundation models.

Recent studies in computer vision have begun to explore these directions. For instance, sparse attention mechanisms have been proposed to alleviate the quadratic complexity of full self-attention. A representative example is the bidirectional sparse attention framework in [179], which sparsifies attention from both the query and key-value perspectives to reduce computational overhead while preserving global context modeling. Beyond architectural optimization, lightweight adaptation modules have emerged as an effective strategy for transferring and fine-tuning foundation models. For example, the Segment Anything Model (SAM) and its variants employ lightweight mask decoders on top of powerful pre-trained image encoders, enabling efficient adaptation to downstream segmentation tasks. Similarly, parameter-efficient fine-tuning techniques, such as adapters and Low-Rank Adaptation (LoRA), allow large vision foundation models to be customized for domain-specific applications using only a small fraction of trainable parameters. These developments demonstrate that sparsity and compression are not only practical solutions for improving computational efficiency but also key design principles for scaling vision foundation models while maintaining strong representation capabilities.

Looking forward, efficient vision architectures are likely to evolve along two complementary directions. The first focuses on developing computation-efficient foundation models through sparsity-aware designs and advanced compression techniques, enabling large-scale models to achieve strong representation capabilities with substantially reduced computational and memory requirements. The second aims to build lightweight yet high-performing models that leverage the knowledge embedded in foundation models through techniques such as knowledge distillation, quantization, pruning, and parameter-efficient adaptation. By transferring the capabilities of large foundation models into compact architectures, these approaches can achieve low-latency inference while maintaining competitive performance. Together, these two directions may play a crucial role in advancing efficient and scalable computer vision systems for real-world deployment.

(2). Representative Feature Learning for VLMs

Vision-language models (VLMs) extend visual representation learning by connecting image content with natural-language semantics, supporting tasks such as image-text retrieval, zero-shot recognition, and visual question answering. Learning representative features in this setting requires both semantic alignment across modalities and preservation of the visual evidence needed for downstream tasks. CLIP [9] established a scalable approach by training image and text encoders with a contrastive objective that brings matched image-text embeddings together while separating mismatched pairs. Its ability to recognize categories specified through text demonstrates the transferability of language-supervised visual features. For generative VLMs, BLIP-2 [180] uses a lightweight Querying Transformer to bridge frozen image encoders and frozen large language models, while LLaVA [181] connects a vision encoder to a language model and applies visual instruction tuning. These developments make the visual encoder, cross-modal connector, and training objectives central to the quality of multimodal representations.

Despite these advances, strong semantic alignment does not necessarily imply complete visual understanding. Image-level contrastive objectives can reward recognition of salient concepts without requiring precise representations of their attributes and relationships. The Attribution, Relation, and Order (ARO) benchmark [182] demonstrates that evaluated VLMs can struggle with compositional distinctions even when conventional retrieval performance is strong. For example, recognizing two objects is insufficient to distinguish which object is above the other. Generative VLMs also exhibit object hallucination, producing descriptions of objects absent from the image; the POPE study [183] identifies associations with object frequency and co-occurrence in visual instructions. These findings motivate a closer examination of whether predictions depend on visual evidence or learned linguistic associations. They also suggest that scaling the language component alone cannot be assumed to resolve deficiencies in visual feature extraction and cross-modal grounding.

A promising research direction is to combine language-aligned semantic features with detailed visual representations learned through complementary objectives. DINOv3, although a vision-only model, demonstrates that self-supervised learning can produce transferable features for both image-level and pixel-level tasks. Cambrian-1 [184] examines different vision encoders and introduces a

spatially aware connector for integrating high-resolution features into multimodal language models. Together, these studies motivate investigating how semantic, spatial, and fine-grained information can be retained during cross-modal fusion. Future models could combine global image-text alignment with region-word correspondence, dense visual supervision, and self-supervised feature learning. Composition-aware hard negatives, supported by the findings in [4], can further encourage discrimination between descriptions containing similar objects but different attributes or relations. However, combining objectives or encoders may introduce redundant features and conflicting optimization signals. Determining which information each component contributes, and how to balance their training, remains an important open problem.

Another key direction is to preserve task-relevant information under limited computation. High-resolution images and video produce large numbers of visual tokens, while aggressive compression may discard small objects, spatial relationships, or temporal changes. Building on compact connectors such as those in BLIP-2 and Cambrian-1, future work could investigate adaptive feature selection that allocates more tokens to informative regions or frames while retaining sufficient global context. Its value should be assessed through controlled comparisons of grounding accuracy, token budgets, and inference cost. Evaluation should also distinguish feature quality from the contribution of the language decoder: frozen-feature probes, region-level grounding tasks, compositional tests, and hallucination assessments provide complementary evidence. Testing with altered images or misleading textual prompts could help establish whether answers respond to changes in visual evidence, while evaluation across domains would assess transferability. Ultimately, progress requires representations that remain discriminative and spatially informative after alignment with language, enabling more reliable visual understanding across tasks and environments.

(3). World Models for Visual Prediction and Planning

The architecture of world models has evolved from compact visual state reconstruction to latent dynamics modeling and task-oriented prediction. Early work [185] combined a variational autoencoder, recurrent dynamics, and a controller to compress observations into a tractable latent environment. PlaNet [186] extended this design with a recurrent state-space model that explicitly separates deterministic temporal memory from stochastic latent states, providing a basis for

959 predicting uncertain dynamics and planning in latent space. Later approaches reduced the
960 dependence on visual reconstruction. MuZero [187] learned task-relevant representations and
961 transition dynamics from rewards, values, and policies, while DreamerV3 [188] used imagined
962 latent trajectories to jointly support policy and value learning. TD-MPC2 [189] further adopted a
963 compact decoder-free dynamics model for trajectory optimization. These developments indicate a
964 clear architectural trend: the world model is gradually shifting from a visual reconstruction module
965 to a latent transition model in which the representation is shaped by prediction and planning
966 requirements.

967 Recent work has further changed the role of the visual encoder within this architecture. V-JEPA
968 [190] predicts latent video features instead of reconstructing pixels, reducing the need to model
969 appearance that is difficult to predict and emphasizing spatiotemporal structure. V-JEPA 2 [191]
970 extends this formulation by introducing action-conditioned prediction, separating general visual
971 representation learning from dynamics learning based on interaction data. Genie [192] introduces a
972 latent action space for learning controllable visual transitions, while DIAMOND [193] uses a
973 diffusion model to generate future observations conditioned on previous observations and actions.
974 Compared with earlier latent-state models, these architectures place greater emphasis on what
975 should be predicted and at which representation level, ranging from compact latent states and
976 features to high-fidelity visual observations. This also reveals an unresolved architectural trade-off.
977 Strong compression reduces memory and planning costs but can remove small objects or subtle
978 motion cues, whereas detailed generative models retain more visual information but require
979 substantially greater computation and may accumulate errors over long rollouts. Thus, progress in
980 world models is not simply a matter of increasing model size; it increasingly depends on selecting
981 an appropriate representation and prediction mechanism for the intended planning horizon and task.

982 Several architectural issues remain open. Current models often rely on a relatively fixed latent
983 representation, although different spatial regions and temporal events contribute unequally to future
984 decisions. Hierarchical, object-centric, or adaptive representations could provide a better allocation
985 of model capacity by retaining important details while compressing redundant information. Long-
986 horizon prediction also requires more persistent memory and mechanisms for controlling error
987 accumulation, rather than relying solely on recurrent state updates or short-range temporal
988 prediction. Another limitation is the gap between predicting observed dynamics and predicting the

consequences of interventions: passive video provides information about how scenes change, but does not necessarily identify how particular actions cause those changes. Future architectures therefore need tighter coupling between visual encoding, action representation, transition modeling, and uncertainty estimation. Progress should ultimately be assessed not only by reconstruction or representation quality, but also by long-horizon consistency, action-conditioned prediction accuracy, planning performance, generalization, and computational cost. These requirements suggest that the next stage of world-model development will likely focus on more structured and adaptive latent architectures rather than simply larger visual backbones.

6. Conclusion

In this systematic review, we revisit the two-decade evolutionary trajectory of deep neural networks for computer vision. We organize major technical breakthroughs into three coherent dimensions: task-driven network architectures, feature-learning sub-module designs, and fundamental principles for constructing high-performance models. Representative seminal works are analyzed concerning their effectiveness, robustness and generalizability across vision tasks. Apart from retrospective literature consolidation, we also discuss prevailing bottlenecks of current paradigms and point out potential future research avenues. We hope this survey provides a comprehensive overview for our peers and those interested in applying deep neural networks in their fields, offering insights into the significant strides and ongoing challenges in deep learning.

Acknowledgement

This work is supported by the Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B1212010004), the Hubei Key Laboratory of Intelligent Robot in Wuhan Institute of Technology (No. HBIRL202202), the Open Project Program of the State Key Laboratory of CAD and CG in Zhejiang University (No. A2322), and the Fundamental Research Funds for the Central Universities of China (No. PA2023IISL0095). The author, Guoping Xu, would also like to acknowledge the valuable assistance of Wentao Liao, Xuan Zhang, Xiaxia Wang, and Xuesong Leng.

Reference

- [1] C. Dickens, *A tale of two cities*. Penguin UK, 2007.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," presented at the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

1020 [3] D. Silver and et al., "Mastering the game of Go with deep neural networks and tree search,"
1021 *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

1022 [4] A. Esteva and et al., "Dermatologist-level classification of skin cancer with deep neural
1023 networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

1024 [5] J. Jumper and et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*,
1025 vol. 596, no. 7873, pp. 583–589, 2021.

1026 [6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language
1027 understanding by generative pre-training," 2018.

1028 [7] A. Radford and J. Wu, "Language Models are Unsupervised Multitask Learners," in
1029 "OpenAI Blog," 2019.

1030 [8] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information*
1031 *processing systems*, vol. 33, pp. 1877–1901, 2020.

1032 [9] A. Radford and et al., "Learning transferable visual models from natural language
1033 supervision," presented at the International Conference on Machine Learning (ICML), 2021.

1034 [10] A. Radford *et al.*, "Learning transferable visual models from natural language supervision,"
1035 in *International conference on machine learning*, 2021: PmLR, pp. 8748–8763.

1036 [11] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A
1037 comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

1038 [12] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural
1039 network and its application in image classification: a survey," *Optical Engineering*, vol. 58,
1040 no. 4, p. 040901, 2019.

1041 [13] M. Ahmad and et al., "Hyperspectral image classification—Traditional to deep models: A
1042 survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations*
1043 *and Remote Sensing*, vol. 15, pp. 968–999, 2021.

1044 [14] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for
1045 hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and*
1046 *Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

1047 [15] J. Song, S. Gao, Y. Zhu, and C. Ma, "A survey of remote sensing image classification based
1048 on CNNs," *Big Earth Data*, vol. 3, no. 3, pp. 232–254, 2019.

1049 [16] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image
1050 classification: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 6, 2018.

1051 [17] G. Litjens and et al., "A survey on deep learning in medical image analysis," *Medical Image*
1052 *Analysis*, vol. 42, pp. 60–88, 2017.

1053 [18] W. Wang and et al., *Medical image classification using deep learning* (Deep Learning in
1054 Healthcare: Paradigms and Applications). 2020.

1055 [19] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. d. L. Ferreira, "Deep learning for
1056 biological image classification," *Expert Systems with Applications*, vol. 85, pp. 114–122,
1057 2017.

1058 [20] H. Wang and et al., "Biological image analysis using deep learning-based methods:
1059 literature review," *Digital Medicine*, vol. 4, no. 4, pp. 157–165, 2018.

1060 [21] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey,"
1061 *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

1062 [22] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE*
1063 *Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232,

2019.

[23] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[24] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[25] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[26] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, "Blind image super-resolution: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5461–5480, 2022.

[27] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[28] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, 2020.

[29] F. Zhuang and et al., "A comprehensive survey on transfer learning," presented at the Proceedings of the IEEE, 2020.

[30] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 2, pp. 151–166, 2020.

[31] G. Xu *et al.*, "Segment anything for video: A comprehensive review of video object segmentation and tracking from past to future," *Neurocomputing*, vol. 682, p. 133439, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231226008362>.

[32] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 4, pp. 1–47, 2020.

[33] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical image analysis*, vol. 63, p. 101693, 2020.

[34] J. A. Onofrey *et al.*, "Sparse data-driven learning for effective and efficient biomedical image segmentation," *Annual review of biomedical engineering*, vol. 22, no. 1, pp. 127–153, 2020.

[35] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, vol. 137, p. 109347, 2023.

[36] N. E. Khalifa, M. Loey, and S. Mirjalili, "A comprehensive survey of recent trends in deep learning for digital images augmentation," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2351–2377, 2022.

[37] R. Azad *et al.*, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[38] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Medical image analysis*, vol. 88, p. 102802, 2023.

[39] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE*

1108 *transactions on knowledge and data engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.

1109 [40] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-based semi-supervised learning: A
1110 comprehensive review," *IEEE Transactions on Neural Networks and Learning Systems*, vol.
1111 34, no. 11, pp. 8174–8194, 2022.

1112 [41] G. Cheng *et al.*, "Towards large-scale small object detection: Survey and benchmarks,"
1113 *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 11, pp. 13467–
1114 13488, 2023.

1115 [42] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep
1116 learning methods for small object detection," *Expert Systems with Applications*, vol. 172,
1117 p. 114602, 2021.

1118 [43] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks:
1119 analysis, applications, and prospects," *IEEE transactions on neural networks and learning
1120 systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

1121 [44] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision:
1122 A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

1123 [45] H. Qu *et al.*, "A survey of mamba," *ACM Transactions on Intelligent Systems and
1124 Technology*, 2024.

1125 [46] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM
1126 Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

1127 [47] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller,
1128 faster, and better," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–37, 2023.

1129 [48] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers:
1130 algorithms, techniques, and performance benchmarking," *IEEE transactions on pattern
1131 analysis and machine intelligence*, vol. 46, no. 12, pp. 7682–7700, 2024.

1132 [49] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A
1133 survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp.
1134 10850–10869, 2023.

1135 [50] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants,
1136 applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–49, 2021.

1137 [51] L. Jiao *et al.*, "New generation deep learning for video object detection: A survey," *IEEE
1138 Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3195–3215,
1139 2021.

1140 [52] G. Xu, X. Wang, X. Wu, X. Leng, and Y. Xu, "Development of residual learning in deep
1141 neural networks for computer vision: A survey," *Engineering Applications of Artificial
1142 Intelligence*, vol. 142, p. 109890, 2025.

1143 [53] Y. Bao *et al.*, "3d gaussian splatting: Survey, technologies, challenges, and opportunities,"
1144 *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 7, pp. 6832–
1145 6852, 2025.

1146 [54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep
1147 convolutional neural networks," *Advances in neural information processing systems*, vol.
1148 25, 2012.

1149 [55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image
1150 recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*,
1151 2015: Computational and Biological Learning Society.

1152 [56] K. He and J. Sun, "Convolutional Neural Networks at Constrained Time Cost," presented
1153 at the CVPR, 2015.

1154 [57] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway Networks," presented at the arXiv
1155 preprint arXiv:1505.00387, 2015.

1156 [58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in
1157 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp.
1158 770–778.

1159 [59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected
1160 convolutional networks," in *Proceedings of the IEEE conference on computer vision and
1161 pattern recognition*, 2017, pp. 4700–4708.

1162 [60] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the
1163 impact of residual connections on learning," in *Proceedings of the AAAI conference on
1164 artificial intelligence*, 2017, vol. 31, no. 1.

1165 [61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for
1166 deep neural networks," in *Proceedings of the IEEE conference on computer vision and
1167 pattern recognition*, 2017, pp. 1492–1500.

1168 [62] H. Liu and et al., "A ConvNet for the 2020s " presented at the CVPR, 2022.

1169 [63] G. Xu, X. Zhang, X. He, and X. Wu, "Levit-unet: Make faster encoders with transformer for
1170 medical image segmentation," in *Chinese Conference on Pattern Recognition and
1171 Computer Vision (PRCV)*, 2023: Springer, pp. 42–53.

1172 [64] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing
1173 systems*, vol. 30, 2017.

1174 [65] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in
1175 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp.
1176 10012–10022.

1177 [66] A. G. Howard, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision
1178 Applications," presented at the arXiv preprint arXiv:1704.04861, 2017.

1179 [67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted
1180 residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer
1181 vision and pattern recognition*, 2018, pp. 4510–4520.

1182 [68] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional
1183 neural network for mobile devices," in *Proceedings of the IEEE conference on computer
1184 vision and pattern recognition*, 2018, pp. 6848–6856.

1185 [69] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient
1186 cnn architecture design," in *Proceedings of the European conference on computer vision
1187 (ECCV)*, 2018, pp. 116–131.

1188 [70] H. Liu and et al., "EfficientFormer: Vision Transformers at MobileNet Speed," presented at
1189 the NeurIPS, 2022.

1190 [71] J. Pan and et al., "EdgeViTs: Competing Light-Weight CNNs on Mobile Devices with Vision
1191 Transformers," presented at the ECCV, 2022.

1192 [72] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory Efficient
1193 Vision Transformer with Cascaded Group Attention," presented at the CVPR, 2023.

1194 [73] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-
1195 friendly Vision Transformer," presented at the ICLR, 2022.

1196 [74] Y. Chen and et al., "Mobile-Former: Bridging MobileNet and Transformer," presented at
1197 the CVPR, 2022.

1198 [75] B. Graham and et al., "LeViT: A Vision Transformer in ConvNet's Clothing for Faster
1199 Inference," presented at the ICCV, 2021.

1200 [76] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object
1201 detection and semantic segmentation," in *Proceedings of the IEEE conference on*
1202 *computer vision and pattern recognition*, 2014, pp. 580–587.

1203 [77] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer*
1204 *vision*, 2015, pp. 1440–1448.

1205 [78] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection
1206 with region proposal networks," *IEEE transactions on pattern analysis and machine*
1207 *intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

1208 [79] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time
1209 object detection," in *Proceedings of the IEEE conference on computer vision and pattern*
1210 *recognition*, 2016, pp. 779–788.

1211 [80] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer*
1212 *vision*, 2016: Springer, pp. 21–37.

1213 [81] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE*
1214 *conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

1215 [82] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint*
1216 *arXiv:1804.02767*, 2018.

1217 [83] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of
1218 Object Detection," *arXiv preprint arXiv:2004.10934*, 2020.

1219 [84] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets
1220 New State-of-the-Art for Real-Time Object Detectors," presented at the Proceedings of
1221 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

1222 [85] M. Yaseen, "What is YOLOv9: An in-depth exploration of the internal features of the next-
1223 generation object detector," *arXiv preprint arXiv:2409.07813*, 2024.

1224 [86] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn
1225 Using Programmable Gradient Information," *arXiv preprint arXiv:2402.13616*, 2024.

1226 [87] A. Wang, H. Chen, and L. Liu, "YOLOv10: Real-Time End-to-End Object Detection," *arXiv*
1227 *preprint arXiv:2405.14458*, 2024.

1228 [88] R. Khanam and M. Hussain, "What is YOLOv5: A deep look into the internal features of
1229 the popular object detector," *arXiv preprint arXiv:2407.20892*, 2024.

1230 [89] C. Li *et al.*, "YOLOv6: A Single-Stage Object Detection Framework for Industrial
1231 Applications," *arXiv preprint arXiv:2209.02976*, 2022.

1232 [90] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements,"
1233 *arXiv preprint arXiv:2410.17725*, 2024.

1234 [91] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors,"
1235 *Advances in neural information processing systems*, vol. 38, pp. 78433–78457, 2026.

1236 [92] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end
1237 object detection with transformers," in *European conference on computer vision*, 2020:
1238 Springer, pp. 213–229.

1239 [93] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable

Transformers for End-to-End Object Detection," in *International Conference on Learning Representations*.

[94] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic detr: End-to-end object detection with dynamic attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2988–2997.

[95] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13619–13627.

[96] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[97] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, 2015: Springer, pp. 234–241.

[98] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, 2016: Springer, pp. 424–432.

[99] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[100] H. Huang *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2020: IEEE, pp. 1055–1059.

[101] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[102] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[103] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[104] C. Liu *et al.*, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 82–92.

[105] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European conference on computer vision*, 2020: Springer, pp. 108–126.

[106] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[107] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE*

1284 *transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364,
1285 2020.

1286 [108] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proceedings of the*
1287 *IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

1288 [109] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local u-nets for biomedical image
1289 segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol.
1290 34, no. 04, pp. 6315–6322.

1291 [110] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention
1292 for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on*
1293 *computer vision*, 2019, pp. 603–612.

1294 [111] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence
1295 perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer*
1296 *vision and pattern recognition*, 2021, pp. 6881–6890.

1297 [112] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction
1298 without convolutions," in *Proceedings of the IEEE/CVF international conference on*
1299 *computer vision*, 2021, pp. 568–578.

1300 [113] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image
1301 analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern*
1302 *recognition*, 2022, pp. 20730–20740.

1303 [114] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and
1304 efficient design for semantic segmentation with transformers," *Advances in neural*
1305 *information processing systems*, vol. 34, pp. 12077–12090, 2021.

1306 [115] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international*
1307 *conference on computer vision*, 2023, pp. 4015–4026.

1308 [116] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images,"
1309 *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

1310 [117] Y. Huang *et al.*, "Segment anything model for medical images?," *Medical Image Analysis*,
1311 vol. 92, p. 103061, 2024.

1312 [118] G. Xu, X. Qian, H. C. Shao, J. Luo, W. Lu, and Y. Zhang, "A segment anything model-guided
1313 and match-based semi-supervised segmentation framework for medical imaging," *Med*
1314 *Phys*, 2025.

1315 [119] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the*
1316 *seventh IEEE international conference on computer vision*, 1999, vol. 2: IEEE, pp. 1150–
1317 1157.

1318 [120] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet
1319 representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11,
1320 no. 7, pp. 674–693, 2002.

1321 [121] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to
1322 document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

1323 [122] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional
1324 networks for visual recognition," *IEEE transactions on pattern analysis and machine*
1325 *intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

1326 [123] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A
1327 new backbone that can enhance learning capability of CNN," in *Proceedings of the*

1328 *IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp.
1329 390–391.

1330 [124] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in
1331 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,
1332 2020, pp. 10781–10790.

1333 [125] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking
1334 convolutional attention design for semantic segmentation," *Advances in Neural
1335 Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

1336 [126] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "Efficientvit: Lightweight multi-scale attention for
1337 high-resolution dense prediction," in *Proceedings of the IEEE/CVF international
1338 conference on computer vision*, 2023, pp. 17302–17313.

1339 [127] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," presented at
1340 the Medical Imaging with Deep Learning, 2018.

1341 [128] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image
1342 Recognition at Scale," in *International Conference on Learning Representations*, 2020.

1343 [129] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint
1344 arXiv:1804.03999*, 2018.

1345 [130] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE
1346 conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

1347 [131] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention
1348 for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on
1349 computer vision and pattern recognition*, 2020, pp. 11534–11542.

1350 [132] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in
1351 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–
1352 792.

1353 [133] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the
1354 IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

1355 [134] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv
1356 preprint arXiv:1807.06514*, 2018.

1357 [135] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module,"
1358 in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

1359 [136] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks,"
1360 in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*,
1361 2019, pp. 3024–3033.

1362 [137] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of
1363 the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

1364 [138] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and
1365 machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

1366 [139] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of
1367 the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12009–
1368 12019.

1369 [140] Z. Liu *et al.*, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on
1370 computer vision and pattern recognition*, 2022, pp. 3202–3211.

1371 [141] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in

1372 *European conference on computer vision*, 2022: Springer, pp. 205–218.

1373 [142] B. Graham *et al.*, "Levit: a vision transformer in convnet's clothing for faster inference," in
1374 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp.
1375 12259–12269.

1376 [143] J. Chen *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image
1377 segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p.
1378 103280, 2024.

1379 [144] Y. Chen *et al.*, "Mobile-former: Bridging mobilenet and transformer," in *Proceedings of*
1380 *the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5270–
1381 5279.

1382 [145] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-
1383 friendly Vision Transformer," in *International Conference on Learning Representations*.

1384 [146] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in
1385 *Readings in computer vision*: Elsevier, 1987, pp. 671–679.

1386 [147] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for
1387 Deep Neural Networks," presented at the CVPR, 2017.

1388 [148] D. Zhou, Q. Hou, Y. Chen, J. Feng, and S. Yan, "Rethinking bottleneck structure for efficient
1389 mobile network design," in *European conference on computer vision*, 2020: Springer, pp.
1390 680–697.

1391 [149] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE*
1392 *international conference on computer vision*, 2015, pp. 1395–1403.

1393 [150] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection,"
1394 in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–
1395 2988.

1396 [151] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized focal loss: Towards efficient
1397 representation learning for dense object detection," *IEEE transactions on pattern analysis*
1398 *and machine intelligence*, vol. 45, no. 3, pp. 3139–3153, 2022.

1399 [152] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face
1400 recognition and clustering," in *Proceedings of the IEEE conference on computer vision*
1401 *and pattern recognition*, 2015, pp. 815–823.

1402 [153] J. Ma *et al.*, "Loss odyssey in medical image segmentation," *Medical image analysis*, vol.
1403 71, p. 102035, 2021.

1404 [154] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by
1405 reducing internal covariate shift," in *International conference on machine learning*, 2015:
1406 pmlr, pp. 448–456.

1407 [155] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*,
1408 2016.

1409 [156] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient
1410 for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

1411 [157] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on*
1412 *computer vision (ECCV)*, 2018, pp. 3–19.

1413 [158] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive
1414 learning of visual representations," in *International conference on machine learning*, 2020:
1415 Pmlr, pp. 1597–1607.

- 1416 [159] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual
1417 representation learning," in *Proceedings of the IEEE/CVF conference on computer vision
1418 and pattern recognition*, 2020, pp. 9729–9738.
- 1419 [160] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image
1420 recognition," in *ICML deep learning workshop*, 2015, vol. 2, no. 1: Lille, pp. 1–30.
- 1421 [161] P. Khosla *et al.*, "Supervised contrastive learning," *Advances in neural information
1422 processing systems*, vol. 33, pp. 18661–18673, 2020.
- 1423 [162] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *2021
1424 IEEE/CVF international conference on computer vision (ICCV)*, 2021: IEEE, pp. 9630–9640.
- 1425 [163] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning
1426 via redundancy reduction," in *International conference on machine learning*, 2021: PMLR,
1427 pp. 12310–12320.
- 1428 [164] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged
1429 consistency targets improve semi-supervised deep learning results," *Advances in neural
1430 information processing systems*, vol. 30, 2017.
- 1431 [165] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in
1432 semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on
1433 Computer Vision and Pattern Recognition*, 2023, pp. 7236–7246.
- 1434 [166] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision
1435 and language representation learning with momentum distillation," *Advances in neural
1436 information processing systems*, vol. 34, pp. 9694–9705, 2021.
- 1437 [167] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for
1438 unified vision-language understanding and generation," in *International conference on
1439 machine learning*, 2022: PMLR, pp. 12888–12900.
- 1440 [168] R. Girdhar *et al.*, "ImageBind one embedding space to bind them all," in *2023 IEEE/CVF
1441 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023: IEEE, pp. 15180–
1442 15190.
- 1443 [169] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local
1444 representation learning framework for label-efficient medical image recognition," in *2021
1445 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: IEEE, pp. 3922–3931.
- 1446 [170] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired
1447 medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods
1448 in Natural Language Processing*, 2022, pp. 3876–3887.
- 1449 [171] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-
1450 training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023:
1451 IEEE, pp. 11941–11952.
- 1452 [172] I. J. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information
1453 processing systems*, vol. 27, 2014.
- 1454 [173] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural
1455 information processing systems*, vol. 33, pp. 6840–6851, 2020.
- 1456 [174] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet:
1457 AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size," *arXiv preprint
1458 arXiv:1602.07360*, 2016.
- 1459 [175] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural

networks," in *International conference on machine learning*, 2019: PMLR, pp. 6105–6114.

[176] D. Dai *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1280–1297.

[177] A. Liu *et al.*, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.

[178] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[179] C. Zhan, W. Li, C. Shen, J. Zhang, S. Wu, and H. Zhang, "Bidirectional sparse attention for faster video diffusion training," *arXiv preprint arXiv:2509.01085*, 2025.

[180] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, 2023: PmLR, pp. 19730–19742.

[181] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[182] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?," *arXiv preprint arXiv:2210.01936*, 2022.

[183] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 292–305.

[184] S. Tong *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 87310–87356, 2024.

[185] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, vol. 2, no. 3, p. 440, 2018.

[186] D. Hafner *et al.*, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*, 2019: PMLR, pp. 2555–2565.

[187] J. Schrittwieser *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[188] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse control tasks through world models," *Nature*, vol. 640, no. 8059, pp. 647–653, 2025.

[189] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," in *International Conference on Learning Representations*, 2024, vol. 2024, pp. 47376–47405.

[190] A. Bardes *et al.*, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint arXiv:2404.08471*, 2024.

[191] M. Assran *et al.*, "V-jepa 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.

[192] J. Bruce *et al.*, "Genie: Generative interactive environments," in *Forty-first international conference on machine learning*, 2024.

[193] E. Alonso *et al.*, "Diffusion for world modeling: Visual details matter in atari," *Advances in Neural Information Processing Systems*, vol. 37, pp. 58757–58791, 2024.