Proceedings of CONF-SEML 2026 Symposium: Multimodal Data Acquisition: Applications in Physiological and Behavioral Research
DOI: 10.54254/2755-2721/2026.AD32137

A Survey on the Applications of Convolutional Neural Networks in Computer Vision

Junwei Zhang

Jinan University - University of Birmingham Joint Institute, Jinan University, Guangzhou, China
zhangjun1267@qq.com

Abstract. Convolutional Neural Networks (CNNs) have become the dominant paradigm in computer vision since the ImageNet breakthrough, establishing a solid engineering foundation in both cloud and edge scenarios. Addressing the current status where existing reviews often focus on a single dimension and lack an integrated engineering perspective, this paper systematically reviews the architectural evolution, key components, and application practices of modern CNNs, primarily focusing on image classification. First, the paper elucidates the design principles of key components, including convolution and receptive fields (RF) , normalization and activation functions , and attention mechanisms. It then traces the evolution path of modern hybrid backbones, from AlexNet/ResNet to MobileNet/EfficientNet, and further to ConvNeXt , following a timeline of "deepening and residualization – lightweight and automated scaling – large kernels and Transformer fusion". Second, by integrating typical application scenarios such as object detection, semantic segmentation, and super-resolution , this review distills reusable training recipes and efficiency optimization strategies involving the synergistic use of pruning, quantization, and distillation. It also provides a checklist for evaluation and deployment geared toward actual hardware. Finally, the article analyzes challenges such as long-tail categories, cross-domain distribution shift, and on-device computational constraints, and looks forward to future trends in self-supervised learning, hardware-aware design, and model robustness optimization.

Keywords: Convolutional Neural Networks, Image Classification, Lightweight, Model Efficiency, Attention, Robustness

1. Introduction

While Convolutional Neural Networks (CNNs) remain a cornerstone of computer vision—particularly for resource-constrained and edge scenarios—existing literature often lacks an integrated engineering perspective that bridges the gap between theoretical design and practical deployment. To address this fragmentation, this paper systematically reviews the evolution of modern CNN components and backbones under unified evaluation standards, distilling reusable training recipes and synergistic efficiency strategies such as pruning, quantization, and distillation. By providing hardware-oriented deployment checklists and extending insights to downstream tasks,

DOI: 10.54254/2755-2721/2026.AD32137

we offer a comprehensive, actionable guide that connects architectural principles, historical evolution, and industrial applications to cutting-edge challenges in robustness and efficiency.

2. CNN architecture design

2.1. Basic concepts of CNN

A Convolutional Neural Network (CNN) is a class of feedforward neural networks specifically designed to process data with a grid structure (e.g., 2D images, 3D volumetric data). Compared to traditional fully connected networks, CNNs leverage local connections and parameter sharing to effectively reduce the model's parameter size and enhance robustness against geometric transformations such as translation, thereby becoming the dominant architecture in computer vision tasks [1,2]. A typical CNN generally consists of several stacked modules: Convolutional layers, non-linear Activation, Normalization, Downsampling/Pooling, and a Classification or Regression Head. These modules progressively extract features from low-level textures, mid-level structures, and high-level semantic features, ultimately completing tasks like classification, detection, or segmentation.

Given an input image $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C_{\text{in}}}$,the convolutional layer generates the$k$-th output feature map $Y^{(k)} \in R^{H'\times W'}$ through a set of learnable convolutional kernels $W^{(k)} \in R^{K_h\times K_w\times C_{in}}$ and biases $b^{(k)}$. Without loss of generality, the convolution operation at a single output location(i,j) can be expressed as:

$$y_{i,j}^{(k)} = b^{(k)} + \sum_{c=1}^{C_{in}} \sum_{u=1}^{K_h} \sum_{v=1}^{K_w} w_{u,v,c}^{(k)} \cdot x_{i+u} - \Delta_h, j+v-\Delta_w, c$$

where the summation index Δ_h, Δ_w is determined by the convolutional stride and padding method. Unlike a fully connected layer which connects to all input dimensions, the convolutional kernel only interacts with the input within a local receptive field, and the same kernel shares weights across the spatial dimension. This mechanism significantly reduces the number of parameters and computational cost while modeling local structures. This property endows CNNs with good translation equivariance: when the input undergoes a small shift, the response in the feature map shifts accordingly, providing a foundation for subsequent achievement of translation invariance through pooling and data augmentation.

2.2. Core components of modern CNN architectures

Modern CNN backbones rely on the synergistic integration of several fundamental components to balance representation power, efficiency, and architectural scaling.

Receptive Field (RF) and Convolution: Spatial perception is dictated by the centrally-biased effective RF , which networks expand via stacked small kernels, large kernels, or dilated convolutions. Depthwise operations primarily drive RF growth in separable architectures , while detail-sensitive tasks utilize multi-scale aggregation (e.g., ASPP) to capture remote dependencies without premature downsampling.

Normalization and Activation: To stabilize gradient flow and mitigate covariate shift, techniques like Batch Normalization (BN) are typically utilized. These are commonly paired with activations like ReLU or GELU , whereas pre-activation structures facilitate the training of deeper layers.

DOI: 10.54254/2755-2721/2026.AD32137

Spatial Resampling and Fusion: Downsampling via strided convolutions or pooling expands the RF and extracts semantics , often requiring anti-aliasing filters to ensure smooth spectral transitions. Conversely, upsampling (e.g., transposed convolution, Pixel Shuffle) and pyramidal feature fusion (e.g., FPN, BiFPN) restore spatial resolution and blend cross-scale semantics. These processes necessitate stringent alignment to mitigate checkerboard artifacts.

Topological Connectivity: Residual connections facilitate deep network optimization via layer-wise identity mapping$y = (F(x) + x)$, establishing themselves as the high-throughput default in engineering practice. Alternatively, dense connections maximize parameter efficiency through layer-wise concatenation , albeit at the cost of increased memory bandwidth pressure [3,4].

Attention and Recalibration: Attention mechanisms dynamically scale feature maps to amplify salient spatial or channel information. Lightweight attention variants optimize mobile deployments , while integration with Multi-Head Self-Attention (MHSA) establishes a robust foundation for modern Convolutional-Transformer hybrid frameworks, yielding superior global modeling capabilities

3. Evolutionary lineage of architectures

3.1. Deepening and residualization: from AlexNet/VGG to ResNet (2012–2016)

This phase was mainly driven by "deeper networks and more stable training": AlexNet and VGG achieved a large receptive field by stacking small-kernel convolutions without significantly increasing parameters, and employed Local Response Normalization/BatchNorm to mitigate gradient vanishing and Internal Covariate Shift [1,5]. The Inception series captured multi-scale information within the same layer using multi-branch convolutions [6]. ResNet effectively solved the deep layer degradation problem by introducing residual learning with identity mapping, making networks with hundreds or even thousands of layers possible, and giving rise to pyramidal fusion paradigms like FPN in detection/segmentation [3,7]. Overall, this period established the "Conv–Norm–Act–Residual" infrastructure paradigm, but the model's computation and memory footprint increased linearly with depth, limiting deployment on mobile and edge devices. The evolution of these architectures is depicted in Fig. 1.

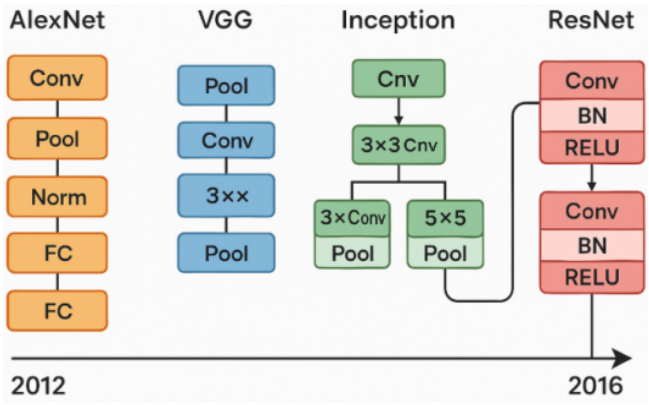

Figure 1. Evolution of representative CNN architectures from AlexNet/VGG through Inception to ResNet (2012–2016)

DOI: 10.54254/2755-2721/2026.AD32137

3.2. Lightweighting and automated scaling: separable convolutions and efficient scaling (2017–2020)

The second stage focused on an efficiency revolution of "lower computational cost for the same accuracy." The Xception and MobileNet series drastically reduced FLOPs by decoupling spatial and channel operations using depthwise separable convolutions [8,9]. ShuffleNet improved cross-channel information flow through group convolutions and channel shuffling. Automated search frameworks like NASNet/MNASNet jointly explored depth, width, and resolution(Fig.2). EfficientNet systematically scaled up networks using a compound scaling coefficient and introduced channel attention (SE/ECA), achieving an excellent accuracy-efficiency trade-off across multiple tasks [10]. During this period, models became easier to deploy for mobile and large-scale online services, but they were sensitive to the maturity of the operator library, quantization, and parallel friendliness. Cross-hardware migration often required targeted parameter tuning and enhanced distillation.

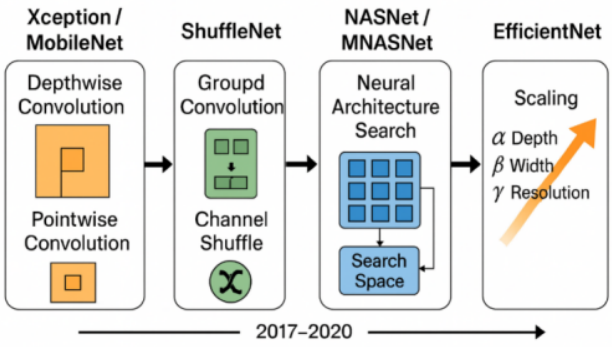

Figure 2. Evolution of lightweight and automated CNN architectures from Xception/MobileNet through ShuffleNet and NASNet/MNASNet to EfficientNet (2017–2020)

3.3. Fusion with transformer: ConvNeXt/RepLKNet×ViT/Swin (2021–Present)

The latest phase strikes a balance between "convolutional modernization" and "self-attention fusion": ConvNeXt absorbed the training and normalization experience of the Transformer with a larger kernel size, Layer Normalization, and a simplified post-ResNet design [11]. RepLKNet and others directly expanded the global receptive field with ultra-large kernel convolutions. RepVGG enhanced expressiveness while maintaining inference-friendliness through training-time re-parameterization. Hierarchical Transformers (such as Swin) and decoding paradigms like DETR/Mask2Former introduced global relationship modeling into detection/segmentation heads, and self-supervised pre-training (MAE/MoCo) further unlocked the advantages of big data. The result is that modern CNNs are approaching or even matching pure Transformers in classification/detection/segmentation, while retaining the advantages of convolution in parallelism and memory access(refer to Fig.3). However, both large kernels and self-attention introduce bandwidth/cache pressure, requiring collaborative design with hardware-friendly re-parameterization, operator fusion, quantization, and distillation to achieve the optimal accuracy–latency–power balance.

DOI: 10.54254/2755-2721/2026.AD32137

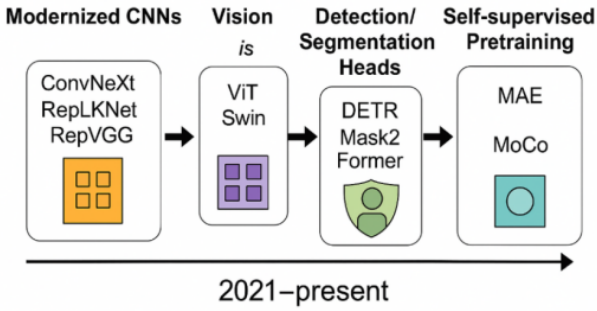

Figure 3. Fusion of modern CNNs and vision transformers in backbones, detection/segmentation heads, and self-supervised pretraining (2021–present)

4. Typical applications

4.1. Image classification and metric learning

Supervised classification serves as the "universal feature extractor" and the entry point for large-scale pre-training: Modern backbones (ResNet/ConvNeXt), combined with BN/LN, residual blocks/large kernels, and regularization, form a stable optimization loop [1]. During the training phase, strong data augmentation (Rand Augment/AutoAugment, Mixup/CutMix, Label Smoothing) is employed, along with coordinated scheduling of learning rate and weight decay (Cosine, WD Decoupling). Small-sample and cross-domain robustness is improved through distillation and self-supervised pre-training (MoCo/SimCLR/MAE). For fine-grained, long-tail, and multi-label tasks, category re-weighting and Focal/Asymmetric Loss are introduced, along with fine-grained local alignment (Part/Attention) and open-vocabulary text guidance (CLIP-like models) to mitigate inter-class similarity and data imbalance.

In Metric Learning, embedding learning based on contrastive/angular constraints (Triplet, ArcFace/CosFace, Circle Loss) is used for face recognition, retrieval, and ReID (Re-identification). From an engineering perspective, deployability is prioritized: a trade-off between throughput, latency, and recall is achieved through vector normalization, low-dimensional embeddings (128/256). Approximate Nearest Neighbor (ANN) search (FAISS/ScaNN), and quantization distillation. Fig. 4illustrates typical pipelines for these tasks. Evaluation is based on metrics such as Top-k/mAP/ROC-AUC and actual recall-latency curves.

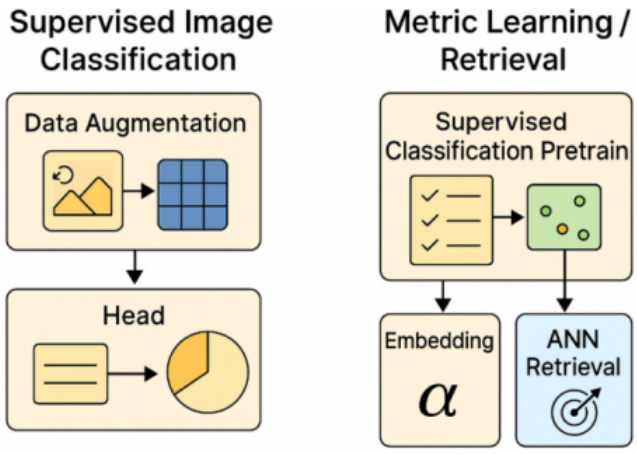

Figure 4. Typical pipelines of supervised image classification and metric learning–based retrieval

DOI: 10.54254/2755-2721/2026.AD32137

4.2. Object detection

Detection focuses on the joint prediction of "location + category." Paradigms coexist, evolving from Two-Stage methods (Faster/Mask R-CNN: RPN proposals + refinement) to One-Stage methods (YOLO/RetinaNet: dense regression) [12-14]. Multi-scale pyramids (FPN/BiFPN/PAN) and more robust regression objectives (GIoU/DIoU/CIoU, Distribution Focal Loss) significantly improve localization quality [7]. Anchor-free methods (FCOS/CenterNet/YOLOv8 family) reduce anchor box hyperparameters by regressing key points/center-ness, enhancing adaptability to small objects and crowded scenes. High-resolution or long-range surveillance environments often integrate large kernels/dilated convolutions and lightweight attention mechanisms to enhance detail perception.

4.3. Semantic segmentation and instance

Semantic segmentation aims for pixel-level classification. After the foundational work of FCN, effective receptive field was expanded through dilated convolutions and multi-scale aggregation (ASPP, PSP) [15]. Encoder–Decoder structures(U-Net/DeepLab/HRNet+OCR)use skip connections/ high-resolution branches to balance detail preservation and semantic information [16,17]. Real-time segmentation favors lightweight backbones (MobileNet/PP-Lite/SegFormer-B0) and shallow decoders [9]. Instance/Panoptic segmentation distinguishes instances on top of semantic segmentation. Detection-based methods (Mask R-CNN/CondInst/SOLOv2) run parallel to fully convolutional grouping/center voting approaches (CenterMask/PolarMask). Panoptic segmentation (PanopticFPN/Mask2Former) unifies "things/countable and stuff/uncountable" objects.

4.4. Super-resolution and image restoration

This category of tasks emphasizes high fidelity and artifact-free reconstruction: Residual and dense connections (EDSR/RCAN/RDN) provide stronger representations, while attention and multi-scale feature fusion (Non-Local/Channel Attention, SwinIR) enhance texture details [18,19]. Sub-pixel rearrangement (Pixel Shuffle) and transposed/content-adaptive upsampling (CARAFE/Deformable) balance efficiency and quality [19,20]. Real-world degradations are often closer to the scene (camera ISP, compression noise, motion blur). Generalization outside the training domain is often improved by degradation modeling combined with synthetic-to-real joint training, stage-wise learning, and uncertainty modeling.For perceptual optimization, VGG perceptual loss/adversarial loss (GAN) improves subjective quality but requires controlling ringing and over-sharpening artifacts, forming a trade-off with numerical metrics like PSNR/SSIM [18-20].

5. Conclusion and future directions

This paper has systematically reviewed the architectural evolution, key components, and diverse applications of modern Convolutional Neural Networks (CNNs), underscoring their continued relevance and parallel-friendliness in edge deployment despite the rise of Vision Transformers. However, deploying robust models in open-world scenarios still faces critical bottlenecks such as long-tail distributions, cross-domain shifts, and strict on-device computational constraints. To address these challenges and drive the development of green, highly reliable visual systems, future research must co-optimize across multiple dimensions. This includes leveraging self-supervised learning and generative augmentation for data efficiency, integrating hardware-aware optimization techniques like quantization-distillation and Neural Architecture Search (NAS), and notably, fusing

DOI: 10.54254/2755-2721/2026.AD32137

large-kernel convolutions with self-attention mechanisms—such as in CNN-Transformer hybrid architectures—to boost global modeling and interactive capabilities.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks, " in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105. 1111

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition, " Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition, " in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[6] C. Szegedy et al., "Going deeper with convolutions, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, and B. Hariharan, "Feature pyramid networks for object detection, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.

[8] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications, " arXiv preprint arXiv: 1704.04861, 2017.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks, " in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[11] Z. Liu et al., "A ConvNet for the 2020s, " in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976–11986.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 580–587.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks, " in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 91–99.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779–788.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3431–3440.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation, " in Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241.

[17] 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 Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[18] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1132–1140.

[19] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks, " in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), 2018, pp. 63–79.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution, " in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 184–199.