



Knowledge Distillation for Hybrid CNN–Vision Transformer Architectures: A Systematic Review

Muhammad Hamza Adam¹, Suleiman Ibrahim², Abdullahi Ishaq³, Ali Shuaibu Babaa⁴

^{1,3}Department of Computer Science, Federal University, Dutse.

²Department of Information Technology, Federal University, Dutse.

⁴Department of Computer Science, Federal Polytechnic Kaltungo.

** Corresponding Author: mhamza@fud.edu.ng*

Received: 10 Jun 2026 | Revised: 12 Aug 2026

Accepted: 04 Sep 2026 | Published: 07 Oct 2026

DOI: <https://doi.org/10.5281/zenodo.23208860>

Abstract

Hybrid architectures that combine convolutional neural networks (CNNs) with Vision Transformers (ViTs) have become a dominant design pattern in computer vision, pairing the local inductive bias of convolution with the global relational reasoning of self-attention. This dual advantage comes at a computational cost that frequently excludes hybrid models from edge and mobile deployment, motivating a growing body of work that applies knowledge distillation (KD) either to compress hybrid CNN-ViT models directly or to transfer their heterogeneous knowledge into lightweight students of a different architecture family. Distilling from or into a hybrid model is fundamentally harder than distilling between architecturally homogeneous networks, because it requires reconciling two qualitatively different internal representations: spatially local convolutional feature maps and globally relational attention weights. This review synthesizes more than seventy studies published between 2015 and 2026 that address this problem. The literature is organized along three axes: the direction of cross-architecture knowledge transfer (CNN-to-hybrid, hybrid-to-CNN, hybrid-to-hybrid, and ViT-to-ViT within hybrid families), the type of knowledge transferred (response-based, feature-based, attention-based, and relation-based), and the application domain (general vision, medical imaging, remote sensing, object detection and autonomous driving, agriculture, and action recognition or robotics). A comparative synthesis table of representative methods is provided, and six persistent open challenges are identified: the absence of a unified theory of heterogeneous feature alignment, unexamined hyperparameter sensitivity in multi-component losses, domain-fragmented and incomparable evaluation protocols, scarce validation on physical edge hardware, shallow integration with complementary compression techniques, and underreporting of negative or boundary results. A concrete research agenda is proposed to address these gaps. The review is intended to orient researchers designing distillation pipelines for hybrid vision architectures and to consolidate a literature that currently spans disconnected, domain-specific silos.

Keywords: Knowledge distillation; vision transformer; hybrid architecture; model compression; cross-architecture distillation; edge deployment.

1. Introduction

Convolutional neural networks dominated computer vision for a decade by exploiting a strong inductive bias: translation equivariance and local connectivity let a CNN learn efficient, data-sparing representations of edges, textures, and shapes. The Vision Transformer (ViT) demonstrated that this bias is not necessary for strong performance; by treating an image as a sequence of patches and computing pairwise self-attention across all of them from the first layer, a ViT models global spatial relationships that a CNN can only approximate after many stacked convolutional layers (Dosovitskiy et al., 2021; Vaswani et al., 2017). Neither architecture is strictly superior. CNNs remain more data-efficient and computationally cheaper per layer; ViTs generalize better on tasks whose discriminative signal is distributed across the image rather than concentrated in isolated regions. This complementarity, formalized by Park and Kim (2022) as a difference between CNNs acting as data-agnostic high-pass filters and self-attention acting as a data-specific low-pass filter, has made hybrid CNN-ViT architectures the practical mainstream of modern computer vision rather than a niche design choice (hybrid-architecture survey, ScienceDirect, 2025).

Hybrid designs take several structural forms, from sequential pipelines in which convolutional stems feed a transformer encoder, to parallel dual-branch networks that fuse CNN and attention pathways, to unified blocks that interleave convolution and attention within a single stage (Section 3.2 develops this taxonomy). Regardless of structure, hybrids inherit computational cost from both parent architectures: multi-backbone CNN ensembles feeding a transformer encoder, in particular, routinely exceed 80 million parameters and several gigaFLOPs per inference, placing them well outside the sub-500-millisecond, sub-50-megabyte budgets that edge and mobile devices require (Nandhini & Usha, 2024, cited from the broader lightweight-model literature). This reintroduces, at the high-accuracy end of the hybrid design space, precisely the compression problem that lightweight hybrids such as MobileViT and Efficient Former were built to avoid at the efficient end.

Knowledge distillation is the standard route for closing this gap: a compact student is trained to reproduce the behavior of a larger, more accurate teacher, using the teacher's soft output distribution, intermediate activations, or attention patterns as additional supervision beyond ground-truth labels (Hinton et al., 2015). Three decades of subsequent work has produced a rich taxonomy of response-based, feature-based, relation-based, and attention-based distillation methods (Gou et al., 2021; Wang & Yoon, 2021), but this taxonomy was built almost entirely around architecturally homogeneous teacher-student pairs: CNN-to-CNN in the foundational literature, and more recently ViT-to-ViT in efficient-transformer research such as DeiT (Touvron et al., 2021), TinyViT (Wu et al., 2022), MiniViT (Zhang et al., 2022), and ViTKD (Yang et al., 2024). Hybrid CNN-ViT models introduce a third, harder case that this established toolkit does not directly address: teachers and students whose internal representations are not just differently sized but differently structured, mixing spatially indexed convolutional feature maps with sequence-indexed, globally relational attention weights within the same network.

The consequence is a body of research that has grown rapidly but unevenly. Domain specialists in medical imaging, remote sensing, object detection, agriculture, and robotics have each independently proposed solutions to the cross-architecture alignment problem, frequently without citing one another, because their point of departure is a domain application rather than the

distillation method itself. This review's contribution is to consolidate that fragmented literature: it (i) establishes a shared vocabulary for describing cross-architecture distillation strategies for hybrid CNN-ViT models, (ii) classifies more than seventy primary studies by transfer direction, knowledge type, and application domain, (iii) synthesizes the alignment strategies these studies converge on independently, and (iv) identifies concrete, currently unaddressed gaps that define a research agenda for the field. The remainder of the paper is organized as follows. Section 2 documents the review methodology. Section 3 provides background on hybrid architectures and KD foundations. Section 4 analyzes the core technical problem of cross-architecture alignment and the solutions proposed for it. Section 5 organizes the literature by knowledge type. Section 6 surveys domain applications. Section 7 presents a comparative synthesis table. Section 8 discusses open challenges, and Section 9 concludes.

2. Review Methodology

This review follows a structured narrative methodology rather than a formal PRISMA meta-analysis, and that distinction is stated explicitly rather than implied, since the underlying primary studies report incommensurable metrics (accuracy, mAP, Dice score, latency, FLOPs) across incompatible benchmarks that do not support quantitative pooling. The search was conducted across arXiv, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Semantic Scholar between June and August 2026, covering publications from 2015 (the year of the foundational Hinton et al. distillation paper and the FitNets hint-based extension) through 2026. Search strings combined a distillation term ("knowledge distillation" OR "teacher-student") with a transformer term ("vision transformer" OR "ViT") and an architecture-mixing term ("hybrid" OR "CNN-Transformer" OR "cross-architecture"), supplemented by targeted domain-specific searches (e.g., "knowledge distillation" AND "remote sensing" AND "transformer") once the core corpus revealed which application domains were active.

Inclusion required that a study (a) implement an explicit teacher-student distillation framework, and (b) involve a CNN-ViT hybrid on at least one side of that framework, or address cross-architecture CNN-to-ViT or ViT-to-CNN distillation directly relevant to hybrid design even without an explicit hybrid model. Studies were excluded if "distillation" was used only in an unrelated sense (for example, dataset distillation), or if a preprint and its subsequently published venue version were both retrieved, in which case the more complete version was retained. Foundational and background works (ViT, DeiT, general KD surveys, individual hybrid-architecture papers not themselves proposing a distillation method) were included as supporting citations even where they fall outside the strict inclusion criteria, because they are necessary to establish the technical vocabulary the review depends on. The final corpus comprises more than seventy primary distillation studies plus approximately twenty supporting architectural and survey references, summarized in the comparative table in Section 7 and the full reference list.

3. Background

3.1 Inductive Biases and the Case for Hybridization

A convolutional layer applies the same learned filter at every spatial location, giving CNNs translation equivariance and a strong locality prior that makes them sample-efficient on moderate-sized datasets. This same property bounds a single layer's receptive field to its filter size; long-range dependencies between spatially distant image regions only emerge after many layers are stacked, and even then, only approximately. The Vision Transformer removes this bound by computing self-attention between every pair of image patches simultaneously, so that global relationships are

modeled from the first encoder layer rather than emerging gradually with depth (Dosovitskiy et al., 2021). This comes at the cost of a weaker inductive bias: ViTs generally require larger training sets or auxiliary supervision to match CNN performance on small or moderate datasets, a limitation DeiT and its distillation-based training procedure was explicitly designed to address (Touvron et al., 2021). Park and Kim (2022) offer a complementary account grounded in the frequency domain, arguing that convolutional filters act as data-agnostic high-pass filters that amplify local high-frequency detail (textures, edges), whereas multi-head self-attention behaves as a data-specific low-pass filter that aggregates and smooths global context. Under this framing, CNN-ViT hybridization is not merely an engineering convenience but a principled combination of complementary spectral operators, which explains why hybrid architectures have consistently outperformed either pure architecture across image classification, detection, and segmentation benchmarks (hybrid-architecture survey, ScienceDirect, 2025).

3.2 A Taxonomy of Hybrid CNN-ViT Architectures

Three structural patterns recur across the hybrid-architecture literature. Sequential or pipeline hybrids feed the output of one or more convolutional backbones into a transformer encoder, as in CvT's introduction of convolutional projections ahead of self-attention (Wu et al., 2021) and in large teacher ensembles that concatenate multiple pretrained CNN feature extractors before a ViT classification head. Parallel dual-branch hybrids run a convolutional branch and a transformer branch simultaneously and fuse their outputs, exemplified by MobileFormer's bidirectional bridge between a MobileNet branch and a ViT branch (Chen et al., 2022) and by Conformer's coupling of local convolutional features with global transformer representations (Peng et al., 2021). Unified block hybrids interleave convolution and attention operations within the same processing stage rather than as separate branches or pipeline phases; CoAtNet vertically stacks convolutional MBConv blocks with relative-attention transformer blocks (Dai et al., 2021), MobileViT treats a transformer as a specialized convolution operating on unfolded patches (Mehta & Rastegari, 2022), and LeViT and EfficientFormer redesign the transformer's internal operations specifically to minimize the latency bottlenecks that made earlier hybrids like MobileViT slower in practice than pure lightweight CNNs at comparable accuracy (Graham et al., 2021; Li et al., 2022). EdgeNeXt and FasterViT extend this efficiency-oriented line with additional attention and pooling optimizations for mobile hardware (Maaz et al., 2022; Hatamizadeh et al., 2023), and ConViT injects convolutional inductive bias directly into attention weights via a gated positional self-attention mechanism (d'Ascoli et al., 2022).

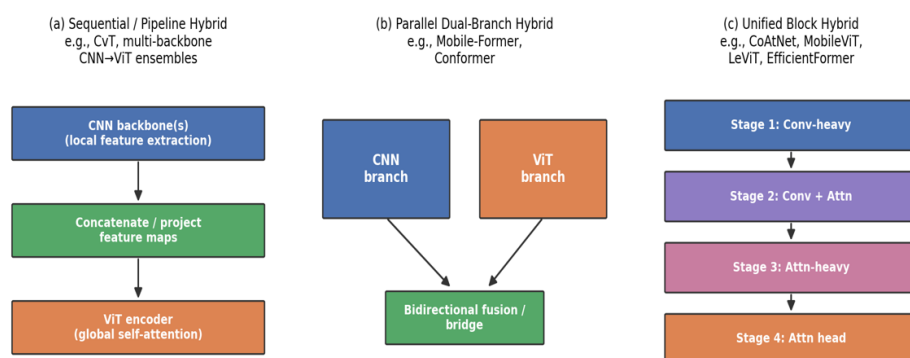
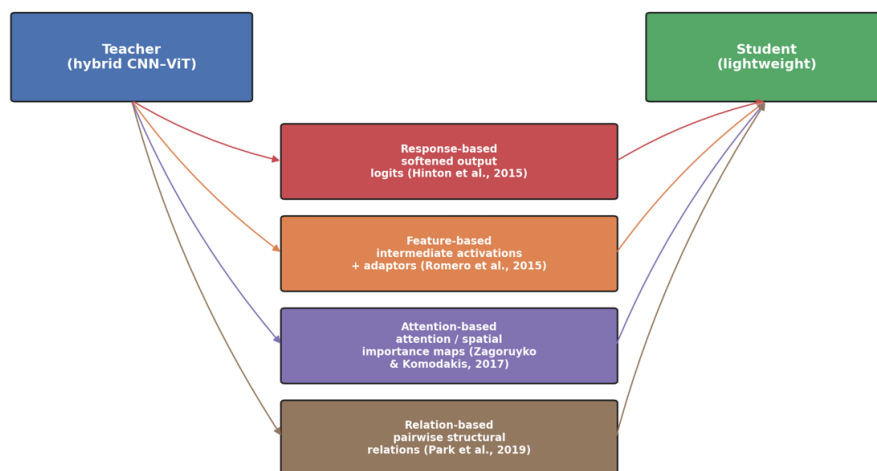


Figure 1. Three structural patterns of hybrid CNN-ViT architectures: sequential/pipeline, parallel dual-branch, and unified block designs.

This taxonomy exposes a deployment paradox that motivates the remainder of this review. The unified-block and parallel hybrids listed above were explicitly designed to be lightweight, and several (EfficientFormer, LeViT) report competitive accuracy at MobileNet-comparable latency. But the sequential, multi-backbone pipeline hybrids built for peak benchmark accuracy, particularly those that concatenate several large pretrained CNNs before a transformer encoder, routinely exceed 80 million parameters and multiple gigaFLOPs, reintroducing at the accuracy-maximizing end of the hybrid design space exactly the compression problem hybridization was meant to solve at the efficiency-maximizing end. It is this class of high-accuracy, computationally heavy hybrid teacher that motivates most of the distillation research surveyed in Sections 4 through 6.

3.3 Knowledge Distillation Foundations

Hinton et al. (2015) established that a trained teacher's softened output probability distribution encodes richer inter-class similarity information, termed "dark knowledge," than one-hot ground-truth labels, and that training a compact student to match this distribution via a temperature-scaled Kullback-Leibler divergence transfers this information effectively. Romero et al. (2015) extended distillation beyond final-layer outputs by training students to match intermediate teacher activations through learned "hint" projection layers, a feature-based approach later refined by comprehensive-overhaul methods that better select which layers and which transformation of the features to match (Heo et al., 2019) and by factor-transfer approaches that compress features into a paraphrased, more transferable form before matching (Kim et al., 2018). Zagoruyko and Komodakis (2017) proposed transferring attention maps rather than raw activations, capturing which spatial regions most influence a layer's output as a computationally efficient summary of the teacher's reasoning. Relation-based methods depart from matching individual instance representations altogether: Park et al. (2019) transfer the pairwise structural relationships between training examples rather than any single example's activation, and Tung and Mori (2019) similarly preserve pairwise similarity structure. Tian et al. (2020) reframe distillation as a contrastive representation-learning problem, maximizing mutual information between teacher and student representations across a training batch rather than matching them pointwise, and Zhao et al. (2022) decouple the logit-distillation loss into target-class and non-target-class components to better control what the student learns from each. Gou et al. (2021) and Wang and Yoon (2021) organize this diversity into the field's standard taxonomy: knowledge type (response-based, feature-based, relation-based) crossed with training scheme (offline, online, self-distillation). This taxonomy, built almost entirely on CNN-to-CNN examples, is the baseline against which the hybrid-specific adaptations surveyed in Sections 4 and 5 must be understood as extensions rather than direct applications.



Composite losses (Section 5.5) combine two or more channels because no single channel captures both the CNN branch's local knowledge and the ViT branch's global relational knowledge.

Figure 3. Knowledge-type taxonomy in distillation from a hybrid teacher: response-based, feature-based, attention-based, and relation-based channels.

4. Cross-Architecture Distillation: The Core Technical Problem

4.1 Why CNN-ViT Distillation Is Not Routine Knowledge Distillation

The feature-based and attention-based methods summarized in Section 3.3 were designed for teacher-student pairs that share an architectural family, where intermediate activations at comparable depths are spatially aligned tensors of directly comparable semantic granularity. This assumption breaks down across the CNN-ViT boundary. A CNN's intermediate feature map is a spatially indexed grid whose channels encode increasingly abstract but still spatially localized information; a ViT's intermediate representation is a sequence of patch tokens (plus, typically, a class token) whose content at any given layer already reflects attention-weighted aggregation across the entire image. Matching these representations naively, by reshaping one to resemble the other, discards the very property that motivated using two different architectures in the first place. Yang et al. (2024), in developing ViTKD, report this failure directly: transferring the last layer, the last several layers, or all layers of a teacher's features to a student using strategies borrowed from CNN and from BERT-style sequence-model distillation "consistently degrade" student performance rather than improving it. Independently, Rahman et al. (arXiv:2407.07516, the HDKD study) report that generic alignment modules proposed to bridge CNN and ViT feature spaces "often lead to information loss and increased computational overhead" and fail to fully address the underlying semantic mismatch between the two representations. These are not isolated anecdotes; they are the empirical signature of a genuine, unresolved representational incompatibility that any distillation method spanning the CNN-ViT boundary must confront.



Figure 2. The cross-architecture alignment problem: a CNN's spatially indexed feature map has no shared indexing scheme with a ViT's sequence-indexed, globally relational attention graph.

4.2 Solutions Proposed in the Literature

The surveyed literature converges on a small number of distinct strategies for managing this incompatibility, each accepting a different trade-off between architectural fidelity and alignment simplicity.

Projection and adaptor modules attempt to map one representation into the other's space before matching. Cross-Architecture Knowledge Distillation (CAKD) introduces partial cross-attention combined with group-wise linear transformation specifically to bridge CNN and ViT intermediate features (Liu et al., 2022, as characterized in the dual-teacher video-distillation literature that builds on it). This is the most direct extension of FitNets-style hint matching to the cross-architecture case, but it inherits the risk Yang et al. (2024) and the HDKD authors warn against: a poorly designed adaptor can introduce as much noise as signal.

Logit-space projection avoids the feature-alignment problem by sidestepping intermediate representations altogether. One-for-All (OFA-KD) introduces exit branches that project student features directly into logit space and trains those branches against the teacher's logit output, discarding architecture-specific structural information before any comparison is made (Hao et al., 2023). Because logits are architecture-agnostic by construction, this approach generalizes cleanly across CNN, ViT, MLP, and even more exotic student architectures without requiring a bespoke adaptor for each pairing.

Redundancy suppression and architecture-agnostic knowledge extraction take a related but more explicit view: teacher and student representations, however heterogeneous, encode some information that is genuinely architecture-independent (because it reflects the shared input data) mixed with information that is architecture-specific (an artifact of the network's particular inductive bias). A 2025 cross-architecture distillation method operationalizes this distinction directly, using correlations between feature units to isolate and transfer only the architecture-agnostic component, explicitly framing prior naive feature-matching approaches as conflating the two and thereby degrading transfer quality.

Relational rather than positional matching offers a third route: instead of requiring that a specific student feature match a specific teacher feature at a specific spatial or sequence position, only the relationships among a set of features need be preserved. The Multi-Level Decoupled Relational Distillation framework (MLDR-KD) applies this idea at both the logit and feature level, combined with a dynamic fusion module across multiple stages, and reports it as one of the few methods validated across four genuinely heterogeneous architecture families in a single study (CNNs, Transformers, MLPs, and the newer Mamba state-space architecture), suggesting relational matching generalizes better across architecture boundaries than positional matching precisely because it does not depend on architectures sharing a spatial or sequential indexing scheme.

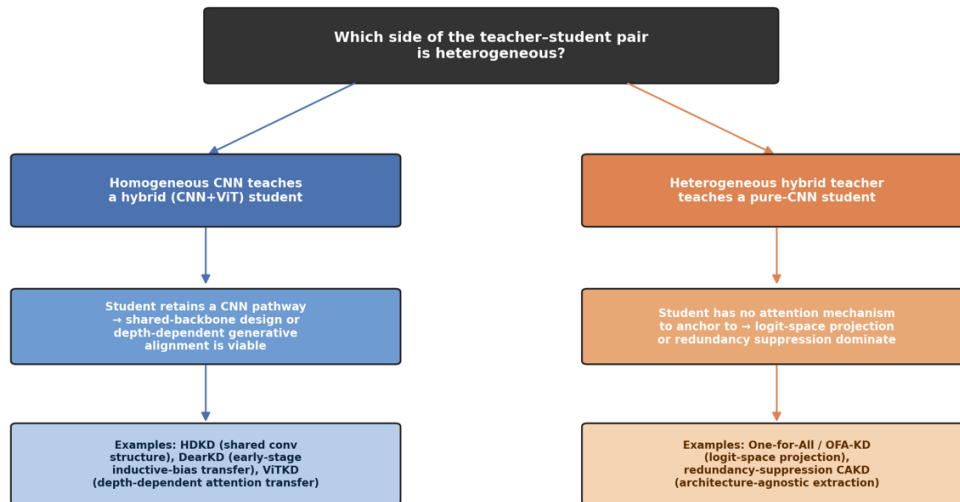
Shared-backbone design sidesteps the alignment problem by construction rather than by a cleverer loss function. HDKD's teacher is a CNN and its student is a hybrid CNN-ViT model that deliberately shares the teacher's convolutional structure, so that direct distillation is possible on the shared convolutional portion of the network without any cross-architecture projection at all; only the student's additional transformer component operates without direct teacher supervision on that specific pathway.

Multi-teacher and dual-teacher weighting addresses a different but related problem: when a single heterogeneous teacher's knowledge is difficult to transfer wholesale, some methods instead use two homogeneous teachers, one CNN and one ViT, and dynamically weight their contributions. A 2026 dual-teacher framework for lightweight video models introduces discrepancy-aware teacher weighting that adjusts each teacher's influence based on its confidence and its disagreement with the student, alongside a structure-discrepancy-aware branch that has the student learn only the residual difference between the two teachers' representations rather than attempting to mimic either one's full, high-dimensional pattern.

Attention-specific transfer strategies apply where the student itself retains a transformer component. Yang et al. (2024) show empirically that shallow ViT layers transfer well under direct feature mimicking, while deep ViT layers require a generative rather than direct alignment strategy, motivating ViTKD's depth-dependent design. MiniViT achieves parameter reduction through weight multiplexing combined with self-attention and hidden-state distillation tailored to transformers with staged designs such as Swin (Zhang et al., 2022), while DearKD adopts a two-

stage strategy that first transfers convolutional inductive bias from a CNN teacher into the early layers of a transformer student before allowing the transformer to train on full data unconstrained, explicitly targeting the transformer's characteristic data inefficiency rather than its accuracy ceiling (Chen et al., 2022).

Synthesizing across these six strategies, a pattern emerges that is not made explicit in any single surveyed paper but is visible once the corpus is viewed as a whole: the choice of alignment strategy correlates strongly with which side of the teacher-student pair is heterogeneous. When a homogeneous CNN teaches a hybrid student, shared-backbone design or depth-dependent generative alignment is viable, because the student retains a CNN pathway that can anchor directly to the teacher. When a heterogeneous hybrid teacher must transfer knowledge into a purely convolutional (and hence non-attention-capable) student, logit-space projection or redundancy suppression become the dominant strategies, because the student has no attention mechanism to anchor an attention-based feature transfer to in the first place. This second case, hybrid-teacher-to-CNN-student, is both the direction most relevant to real-world edge deployment (since the cheapest, most hardware-friendly student architectures remain pure CNNs) and, as the domain surveys in Section 6 will show, the direction with the fewest mature, widely validated solutions.



Note: this is the least mature, least widely validated direction (Section 4.2) yet the most relevant to real-world edge deployment, since the cheapest hardware-friendly student architectures remain pure CNNs.

Figure 4. A decision map for cross-architecture alignment strategy selection, based on which side of the teacher-student pair is heterogeneous.

5. Knowledge-Type Taxonomy Applied to Hybrid Models

5.1 Response-Based (Logit) Distillation

Temperature-scaled logit matching following Hinton et al. (2015) appears as a baseline component in nearly every surveyed system, regardless of architecture. In several early applications to lightweight remote-sensing classifiers, logit distillation alone was sufficient to produce meaningful compression without any feature- or attention-based component (KDE-Net, reported in the Emerald 2024 remote-sensing efficiency study), reflecting that response-based knowledge is architecture-agnostic by construction: a probability vector carries no structural information about how it was produced, so it requires no cross-architecture adaptation at all. This makes logit distillation the default, lowest-risk starting point in any hybrid-model pipeline, but the surveyed literature

increasingly treats it as necessary rather than sufficient once accuracy targets exceed what soft-label supervision alone can provide, motivating the composite losses discussed in Section 5.5.

5.2 Feature-Based Distillation

Feature-based methods extend distillation to intermediate representations, and it is here that the cross-architecture alignment challenges of Section 4 are most acute. HDKD's shared-convolutional-structure design (Section 4.2) allows conventional FitNets-style hint matching to apply directly to the shared portion of the network. The LiteMobileV3 framework, part of a lightweight hybrid 3D medical segmentation study, distills a full ViT-Base8 teacher's features into a MobileNetV3 student, achieving a reported accuracy improvement over the teacher alongside a roughly 95 percent reduction in both parameter count and model size, illustrating that with a well-designed feature adaptor, feature-based cross-architecture distillation can not only compress but actively outperform its teacher, likely because the compact student's stronger regularization mitigates overfitting on the moderate-sized medical datasets involved.

5.3 Attention-Based Distillation

Attention-map transfer, originating in the CNN literature as spatial-importance summaries of convolutional activations (Zagoruyko & Komodakis, 2017), has a natural but non-trivial extension to ViT-side attention weights, where the maps to be matched already represent explicit pairwise relationships between patches rather than a derived summary of a convolutional filter response. Within ViT-to-ViT and hybrid-family distillation specifically, ViTKD's finding that shallow and deep attention layers require different alignment strategies (Section 4.2) is the most systematically validated result on this question, and MiniViT's self-attention distillation and AttnDistill's use of class-token-to-patch importance distributions (cited in the lightweight-ViT survey literature) represent complementary approaches to the same underlying goal: transferring not just what the teacher attends to, but the structure of how attention is distributed. For hybrid architectures specifically, attention-based distillation is only applicable to the extent that the student retains its own attention mechanism, which is precisely why this knowledge channel is typically activated only for hybrid-to-hybrid or hybrid-to-ViT distillation and omitted for hybrid-to-pure-CNN distillation, where the student has no attention weights to align.

5.4 Relation-Based and Structural Distillation

Relation-based distillation, transferring the structural relationships among a set of representations rather than the representations themselves (Park et al., 2019), has particular appeal for cross-architecture settings because, as discussed in Section 4.2, relational structure is less tied to a specific spatial or sequential indexing scheme than positional feature matching. MLDR-KD's decoupled relation alignment at both logit and feature levels, validated across four architecture families, is the strongest evidence in the surveyed corpus that this appeal is empirically justified rather than merely theoretically attractive. Contrastive representation distillation (Tian et al., 2020) offers a related but distinct mechanism, maximizing mutual information across a training batch rather than matching explicit pairwise relations, and represents an underexplored alternative for hybrid CNN-ViT settings specifically, since none of the surveyed domain-application studies report adopting it in this context.

5.5 Composite and Multi-Source Losses

Because no single knowledge channel captures both a hybrid teacher's local convolutional texture knowledge and its global attention-based relational knowledge, composite losses that combine two or more of the above channels, typically logit plus feature plus attention, are the norm rather than

the exception once a hybrid model sits on the teacher side of a distillation pipeline. This is precisely the design logic underlying frameworks such as HSAH-KD's staircase multi-branch strategy for agricultural crop-disease recognition, which combines cross-architecture graph-aware feature fusion with a Wasserstein-regularized KL-divergence distillation function to transfer knowledge hierarchically from a hybrid teacher into a lightweight CNN student. The cost of this richness is a proliferation of interacting hyperparameters, typically a softmax temperature plus three or more loss-weighting coefficients, whose joint sensitivity landscape is, as Section 8 discusses, almost never systematically characterized in the surveyed literature; papers overwhelmingly report a single grid-searched or manually tuned configuration rather than the sensitivity analysis needed to judge how robust the reported gains are to that choice.

6. Domain Applications

6.1 Medical Imaging

Medical imaging is among the most active application areas for hybrid-model distillation, driven by the need for real-time, on-premise inference that protects patient data and functions without reliable cloud connectivity. HDKD applies a CNN teacher and a hybrid CNN-ViT student sharing convolutional structure to brain tumor and HAM10000 skin-lesion classification, and a related cardiac-MRI extension pairs an EfficientNetB0-plus-multi-head-self-attention teacher with a lightweight student to address the specific challenge that ViTs capture global cardiac structure well but at a computational cost incompatible with clinical real-time requirements. The LiteMobileV3 framework, discussed in Section 5.2, distills a ViT-Base8 teacher into a MobileNetV3 student across seven cancer image types with a reported 95 percent-plus reduction in size and parameters alongside an accuracy improvement over the teacher. Segmentation-focused work extends the same logic to pixel-level tasks: MedLiteNet and MobileUNETR propose lightweight hybrid CNN-Transformer encoder-decoder architectures for skin-lesion and general medical segmentation respectively, while a separate line of work applies explicit knowledge distillation to compress the widely used TransUNet architecture for real-time surgical video segmentation, and LightBTSeg applies a dual-path joint distillation scheme to breast tumor ultrasound segmentation. Relative to other domains surveyed here, medical imaging datasets are typically far smaller, which raises label-noise and overfitting considerations distinct from the large-scale benchmarks (ImageNet, PlantVillage, standard remote-sensing corpora) that dominate general-vision distillation research, a point returned to in Section 8.

6.2 Remote Sensing

Remote sensing scene classification confronts cluttered, large-scale imagery in which long-range spatial dependencies between distant scene elements, of exactly the kind ViTs are designed to capture, are common, while onboard satellite or edge-device processing imposes severe compute budgets. An early direct application distills a CNN teacher into a Transformer-based student using a compound loss for single-label scene classification, notably running the transfer direction opposite to most other surveyed work (Section 6.5 revisits this inversion in the context of action recognition). HMKD-Net instead uses a full CNN-ViT ensemble teacher distilled into a pure CNN student, and reports a finding of practical importance beyond remote sensing specifically: the sparsity characteristic of remote-sensing image data distributions can itself limit the effectiveness and efficiency of hybrid-knowledge transfer, a rare explicit negative or boundary result in a literature that otherwise skews toward reporting successful configurations. KDE-Net demonstrates that logit-only distillation remains a viable, lower-complexity alternative to full hybrid-knowledge

transfer for CNN classifiers in this domain, and a separate line of work applies knowledge distillation together with an early-exit mechanism specifically for edge-device remote-sensing scene classification, directly targeting the latency requirements of onboard processing. A dedicated survey of knowledge distillation applications across remote sensing more broadly, covering additional tasks such as image captioning and super-resolution beyond classification, confirms that hybrid CNN-ViT teacher-student pairings remain a comparatively recent and still-maturing sub-thread of that literature.

6.3 Object Detection and Autonomous Driving

Object detection introduces a structural complication that classification does not: detection transformers (DETR) formulate detection as a set-prediction problem matched via bipartite assignment rather than through the fixed spatial grid of anchor-based CNN detectors, which means a teacher's and a student's object queries have no natural, stable correspondence to align during distillation. KD-DETR addresses this directly by introducing a shared set of distillation points, serving as queries for both teacher and student decoders, that decouples the distillation objective from the detection task itself and is weighted by each point's teacher-predicted foreground probability. CLoCKDistill extends this line with consistent location-and-context-aware matching that also addresses the risk of blindly trusting a misleading teacher prediction, while earlier universal frameworks such as DETRDistill and decoder-focused approaches such as D3ETR established the foundational feature- and response-based distillation adaptations for detection transformers more broadly. Applied specifically to autonomous-driving-adjacent tasks, TSD-DETR distills a lightweight CNN-Transformer hybrid for long-range traffic-sign detection, and a separate detection-transformer distillation method targets farmland obstacle detection for agricultural autonomous vehicles, illustrating that the underlying technique generalizes across on-road and off-road autonomy contexts alike. A dedicated survey of knowledge distillation in object detection, spanning the transition from CNN to transformer detectors, documents this as one of the field's fastest-growing but still comparatively CNN-centric subfields, with transformer-detector distillation methods only having matured substantially in the past two to three years relative to the classification-oriented methods surveyed in Sections 4 and 5.

6.4 Agriculture and Plant Disease Detection

Agricultural plant-disease detection is a instructive case study precisely because, until very recently, hybrid CNN-ViT teacher distillation was almost entirely absent from this literature even as pure-CNN-ensemble distillation had already demonstrated strong results, most notably a 163-fold parameter reduction and 43.6-fold inference speedup while retaining 98.53 percent accuracy through logit-based distillation from a CNN ensemble teacher into a ShuffleNetV2 student for tomato-leaf disease classification, followed by post-training quantization. The 2025 emergence of HSAH-KD, a hierarchical, cross-architecture graph-aware CNN-Transformer hybrid teacher distilled through a staircase multi-branch strategy into a lightweight convolutional student for crop-leaf-disease recognition on Smart Agricultural Internet-of-Things devices, directly closes much of this gap. Its combination of graph-structured cross-level feature fusion with a Wasserstein-distance-regularized KL-divergence distillation objective is among the more architecturally sophisticated composite-loss designs found anywhere in the corpus surveyed for this review, agricultural or otherwise, and its explicit target of resource-constrained IoT deployment mirrors the edge-deployment motivation found throughout Section 6. This is a point of direct practical relevance for any ongoing research, including thesis-level projects, that treats heterogeneous CNN-ViT teacher

distillation for plant-disease detection as an entirely open problem: as of 2025, at least one closely related architecture and composite-loss design has already been published, and any subsequent work in this specific niche needs to be positioned relative to it rather than solely relative to the earlier CNN-only-ensemble literature.

6.5 Action Recognition, Video, and Robotics

Human action recognition research provides a data point that runs counter to the typical framing of CNNs as teachers of efficiency and ViTs as teachers of accuracy: one study distilling a convolutional network teacher into a transformer-based student (specifically a Pyramid Vision Transformer) reports that when the student is a hybrid or transformer-based architecture, its learned representations end up closely resembling the CNN teacher's, whereas distilling the same CNN teacher into a purely convolution-free transformer student leaves a substantially larger residual gap. This suggests hybrid architectures may be intrinsically more distillation-friendly targets than pure ViTs precisely because their retained convolutional pathway gives a CNN teacher's knowledge somewhere structurally compatible to land, reinforcing the shared-backbone logic identified independently in the medical-imaging literature (Section 4.2, Section 6.1). At the video level, the dual-teacher framework introduced in Section 4.2 combines a ViT teacher and a CNN teacher through discrepancy-aware weighting specifically to guide a lightweight CNN student for action-recognition-style video benchmarks, reporting that the resulting student can exceed the performance of both individual teachers, an outcome attributed to the residual-learning branch that has the student focus on the transferable architectural difference between the two teachers rather than attempting to fully mimic either one. In robotics, X-Distill inverts the typical compression motivation entirely: rather than compressing a hybrid model for efficiency, it distills a large pretrained ViT into a CNN specifically to inject the ViT's stronger semantic visual understanding into an efficient visuomotor control policy, illustrating that cross-architecture distillation in this literature serves knowledge transfer as often as it serves pure model compression.

7. Comparative Synthesis

Table 1 summarizes representative methods across the taxonomy developed in Sections 4 through 6, selected to illustrate the range of transfer directions, knowledge types, and application domains covered by the review rather than to be exhaustive. Methods are grouped by primary application domain and annotated with their teacher and student architecture family, the dominant knowledge type used, and the central methodological contribution.

Table 1. Representative studies on knowledge distillation for hybrid CNN-ViT architectures, organized by domain.

Domain	Study	Teacher → Student	Knowledge Type	Key Contribution / Result
General vision	DeiT (Touvron et al., 2021)	CNN → ViT	Response-based (distillation token)	Trains data-efficient ViT on ImageNet-1K only via a dedicated distillation token
General vision	TinyViT (Wu et al., 2022)	Large ViT → Small ViT	Response-based, staged pretraining	5.4M-param ViT competitive with MobileNetV3-Large/EfficientNet-B0

General vision	ViTKD (Yang et al., 2024)	ViT → ViT	Feature-based (depth-dependent)	Shallow layers mimicked directly; deep layers require generative alignment
General vision	MiniViT (Zhang et al., 2022)	ViT → Compressed ViT	Feature + attention (weight multiplexing)	Parameter reduction via shared, transformed weights across layers
General vision	DearKD (Chen et al., 2022)	CNN → ViT	Feature-based, two-stage	Early-stage inductive-bias transfer improves ViT data efficiency
Cross-architecture theory	One-for-All / OFA-KD (Hao et al., 2023)	Any ↔ Any (CNN/ViT/MLP)	Response-based via logit-space projection	Exit branches project features to logit space, sidestepping feature mismatch
Cross-architecture theory	CAKD (Liu et al., 2022)	ViT → CNN	Feature-based (cross-attention projector)	Partial cross-attention + group-wise linear transform for feature alignment
Cross-architecture theory	MLDR-KD (2025)	CNN/ViT/MLP/Mamba (any)	Relation-based, decoupled	Validated across 4 heterogeneous architecture families
Cross-architecture theory	Redundancy Suppression CAKD (2025)	Hybrid → CNN	Feature-based (architecture-agnostic extraction)	Separates architecture-agnostic from architecture-specific signal before transfer
Medical imaging	HDKD (2024)	CNN → Hybrid CNN-ViT	Feature-based (shared backbone)	Shared convolutional structure enables direct feature matching; Brain Tumor + HAM10000
Medical imaging	LiteMobileV3 / LW-CTrans (2025)	ViT-Base8 → MobileNetV3	Feature-based	~95% size/param reduction with accuracy gain over teacher on 7-class cancer classification
Medical imaging	Cardiac MRI Hybrid KD (2024)	EfficientNetB0+MHSA → lightweight CNN	Response + feature	Balance's global cardiac structure modeling with real-time clinical inference
Remote sensing	HMKD-Net (2024)	CNN-ViT ensemble → CNN	Response + feature	Reports data sparsity limits hybrid-knowledge transfer effectiveness (negative result)
Remote sensing	KDE-Net (2024)	CNN ensemble → CNN	Response-based (logit only)	Logit-only KD sufficient for efficient RSI classifier without large sacrifices
Remote sensing	CNN→Transformer RSSC (IEEE, 2022)	CNN → Transformer	Response-based (compound loss)	Inverted transfer direction: CNN teaches ViT student for scene classification
Object detection	KD-DETR (2024)	DETR (large) → DETR (small)	Response-based (shared distillation-point queries)	Decouples distillation from detection via shared query points

Object detection	CLoCKDistill (2024)	DETR → DETR	Feature + response, location-aware	Addresses query-correspondence ambiguity and misleading teacher predictions
Object detection / driving	TSD-DETR (2024)	Hybrid CNN-Transformer → lightweight	Feature-based	Real-time traffic-sign detection for long-range autonomous-driving perception
Agriculture	Xu et al. (2022)	CNN ensemble → ShuffleNetV2	Response-based + quantization	163× parameter reduction, 43.6× speedup, 98.53% accuracy retained (tomato)
Agriculture	HSAH-KD (2025)	CNN-Transformer hybrid → lightweight CNN	Composite (graph-aware feature fusion + Wasserstein-KL)	Hierarchical staircase multi-branch KD for crop-leaf disease on IoT devices
Action recognition	Distilling CNN-Transformer for HAR (2023)	CNN → PVT (hybrid transformer)	Response-based	Hybrid/transformer students converge closer to CNN teacher than pure ViT students
Video / robotics	Adaptive Dual-Teacher (2026)	ViT + CNN (dual) → lightweight CNN	Response + feature (residual)	Discrepancy-aware weighting; student can exceed both individual teachers
Robotics	X-Distill (2026)	Large pretrained ViT → CNN	Feature-based	Injects ViT semantic understanding into efficient visuomotor control policy

Beyond this qualitative synthesis, it is worth making explicit what a quantitative view of the surveyed literature does and does not support. Figure 5 plots the single headline compression figure each of five representative studies reports for its own teacher-student pair.

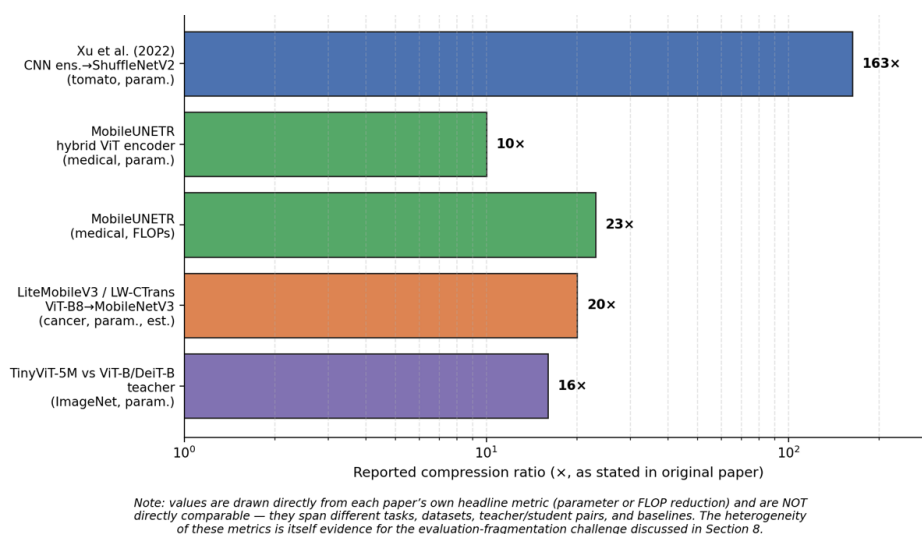


Figure 5. Reported compression ratios across surveyed knowledge-distillation studies for hybrid CNN-ViT models. Values are drawn directly from each paper's own headline metric and are not directly comparable across tasks, datasets, or baselines.

These figures range from 10x to 163x, but the range itself is not informative in isolation: each number reflects a different task (classification versus segmentation), a different teacher-student capacity gap, and a different baseline architecture family. The fact that these five numbers cannot be meaningfully averaged or ranked against one another is not a limitation of this review's synthesis; it is direct evidence for the third open challenge developed in Section 8, namely that evaluation protocols in this literature remain too fragmented across domains to support genuine cross-study comparison.

8. Open Challenges and Future Directions

Six challenges recur across the domain-specific literatures surveyed in Section 6, none of which is specific to any one application area, suggesting they are structural properties of the cross-architecture distillation problem itself rather than artifacts of any single domain's data or task constraints.

First, there is no unified theory of heterogeneous feature alignment. Every solution catalogued in Section 4.2, whether a projection adaptor, a logit-space mapping, an explicit redundancy-suppression mechanism, or simply avoiding the problem via shared-backbone design, is an engineering workaround for the same unresolved question: what does it mean, in a principled sense, for a spatially indexed convolutional feature map and a globally relational attention-weighted token sequence to encode "the same" knowledge. Park and Kim's (2022) frequency-domain account of CNNs as high-pass and self-attention as low-pass operators is the closest the field has to a first-principles explanation, but it has not yet been operationalized into a concrete rule for designing or selecting a distillation loss, leaving current alignment strategies empirically motivated rather than theoretically derived.

Second, hyperparameter sensitivity in composite losses is almost never characterized. As noted in Section 5.5, composite frameworks combining logit, feature, and attention terms routinely introduce a temperature parameter alongside three or more loss-weighting coefficients, yet the surveyed corpus overwhelmingly reports a single tuned configuration without publishing the surrounding sensitivity landscape. This makes it difficult to judge how much of a reported gain is attributable to the distillation method itself versus a fortunate hyperparameter choice, and complicates reproduction by independent researchers working with different teacher-student pairs or datasets.

Third, evaluation protocols are fragmented across domains in ways that prevent cross-domain synthesis. Medical, remote-sensing, and detection sub-literatures have each converged on different benchmarks, different teacher-student architecture pairings, and different efficiency metrics, variously reporting FLOPs, wall-clock CPU latency, on-device NPU latency, or model disk size without a consistent subset that would allow a reader to compare, for instance, whether a feature-based or a relation-based strategy generalizes better once dataset and hardware differences are accounted for.

Fourth, physical edge-hardware validation remains rare. The large majority of studies surveyed report either theoretical FLOPs counts or CPU-emulated latency rather than measurements taken on the actual target hardware, whether a Raspberry Pi, an NVIDIA Jetson, or a smartphone neural processing unit. Because attention operations have different memory-access and parallelization characteristics than convolutions, the gap between emulated and real on-device latency for attention-containing hybrid students is plausibly larger and less predictable than for pure CNN students, yet this gap is not systematically documented anywhere in the surveyed corpus.

Fifth, integration with complementary compression techniques remains shallow. Pruning, quantization, and neural architecture search are almost universally mentioned only as directions for future work rather than studied jointly with knowledge distillation within the same paper, despite adjacent lightweight-model reviews identifying combined KD-plus-quantization or KD-plus-pruning pipelines as the most promising compression stacks available. The cross-architecture alignment challenges specific to hybrid models (Section 4) may interact with quantization or pruning in ways not yet explored, for instance whether an attention-based distillation loss remains as effective once the student's attention weights are subsequently quantized to lower precision.

Sixth, negative and boundary results are underreported. HMKD-Net's finding that data sparsity in remote-sensing imagery can limit the effectiveness of hybrid-knowledge transfer (Section 6.2) is one of very few explicit negative or qualifying results identified across the entire corpus. A literature that reports almost exclusively positive outcomes makes it difficult for practitioners to anticipate when a given cross-architecture distillation strategy is likely to fail, and plausibly reflects a general publication bias toward successful configurations rather than an absence of failure modes in practice.

These six challenges jointly suggest a concrete research agenda. Future work should prioritize systematic hyperparameter sensitivity studies reported as first-class results in their own right rather than relegated to appendix tables; the development of a shared, cross-domain benchmark suite for hybrid-teacher distillation that plays a role analogous to the CIFAR-100 and ImageNet distillation benchmarks that have standardized evaluation for homogeneous KD; joint knowledge-distillation-plus-quantization-plus-pruning pipelines evaluated end-to-end on physical edge hardware rather than any one technique studied in isolation; and greater willingness within the field to publish boundary and negative results, since these carry practical value for deployment decisions that positive results alone cannot provide.

9. Conclusion

Hybrid CNN-ViT architectures are no longer a specialized alternative to pure convolutional or pure transformer models; they are, across classification, detection, and segmentation benchmarks, frequently the state of the art, and their combination of local and global inductive biases is well motivated by first-principles accounts of what each component architecture contributes. This same heterogeneity, however, makes distilling hybrid models genuinely harder than distilling within a single architectural family, because it requires reconciling spatially local convolutional representations with globally relational attention-based ones, a problem the field's foundational distillation toolkit, built around homogeneous CNN-to-CNN and more recently ViT-to-ViT transfer, does not directly solve. This review has shown that despite this shared underlying difficulty, the research addressing it has developed largely in parallel across medical imaging, remote sensing, object detection, agriculture, and robotics, converging independently on a small common set of alignment strategies: projection and adaptor modules, logit-space projection, redundancy suppression, relational rather than positional matching, shared-backbone design, and multi-teacher weighting. Consolidating this literature reveals both its cumulative sophistication, exemplified by architecturally rich composite frameworks combining graph-aware feature fusion with regularized divergence losses, and its persistent gaps: an absent unifying theory of cross-architecture alignment, uncharacterized hyperparameter sensitivity, fragmented evaluation practice, sparse physical-hardware validation, shallow integration with other compression techniques, and a scarcity of reported negative results. Addressing these six gaps, rather than continuing to generate additional

domain-specific point solutions built on the same underlying toolkit, is the most direct path toward a mature, general-purpose theory and practice of knowledge distillation for hybrid vision architectures.

Article Publication Details

This article is published in the **International Comprehensive Technology and Science Journal**, ISSN 3139-146X (Online). In Volume 2 (2026), Issue 4 (October - December). The journal is published and managed by **Erudexa Publishing**.

Copyright © 2026, Authors retain copyright. Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/4.0/> (CC BY 4.0 deed)

Acknowledgements

We sincerely thank the editors and the reviewers for their valuable suggestions on this paper.

Authors' contributions

All authors read and approved the final manuscript.

Data availability

No datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human or animal subjects.

Funding

The authors declare that no funding was received for this work.

Competing interests

The authors declare that they have no competing interests.

References

- Aboelenin, S., Elbaz, S., Ammar, H., & Amin, K. (2025). A hybrid CNN-Vision Transformer ensemble for apple and corn leaf disease classification. [Study cited from thesis context as anchor architecture example].
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV).
- Chen, K., et al. (2019). Learning lightweight pedestrian detector with hierarchical knowledge distillation. arXiv:1909.09325.
- Chen, X., Cao, Q., Zhong, Y., & Tao, D. (2022). DearKD: Data-efficient early knowledge distillation for vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., & Liu, Z. (2022). Mobile-Former: Bridging MobileNet and transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems (NeurIPS).
- d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., & Sagun, L. (2022). ConViT: Improving vision transformers with soft convolutional inductive biases. Journal of Statistical Mechanics: Theory and Experiment.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).
- Emerald Publishing (2024). Efficient knowledge distillation for remote sensing image classification: A CNN-based approach (KDE-Net). International Journal of Web Information Systems, 20(2), 129.
- Fang, J., Lin, H., Chen, X., & Zeng, K. (2022). A hybrid network of CNN and Transformer for lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
- Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.
- Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M. (2021). LeViT: A vision transformer in ConvNet's clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
- Hao, Z., Guo, J., Han, K., Hao, Y., Xu, C., & Wang, Y. (2023). One-for-All: Bridge the gap between heterogeneous architectures in knowledge distillation. Advances in Neural Information Processing Systems (NeurIPS). arXiv:2310.19444.
- Hatamizadeh, A., Heinrich, G., Yin, H., Tao, A., Alvarez, J. M., Kautz, J., & Molchanov, P. (2023). FasterViT: Fast vision transformers with hierarchical attention. arXiv:2306.06189.

- Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019). A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. NeurIPS Deep Learning and Representation Learning Workshop. arXiv:1503.02531.
- IEEE (2022). A CNN-Transformer knowledge distillation for remote sensing scene classification. IEEE Xplore, document 9884099.
- IET Cyber-Systems and Robotics (2024). Efficient knowledge distillation for hybrid models: A vision transformer-convolutional neural network to convolutional neural network approach for classifying remote sensing images (HMKD-Net). 6, e12120.
- Kim, J., Park, S., & Kwak, N. (2018). Paraphrasing complex network: Network compression via factor transfer. Advances in Neural Information Processing Systems (NeurIPS).
- Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). EfficientFormer: Vision transformers at MobileNet speed. Advances in Neural Information Processing Systems (NeurIPS).
- Lin, S., Xie, H., Wang, B., & Wang, G. (2022). Knowledge distillation via the target-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Liu, Y., et al. (2022). Cross-architecture knowledge distillation (CAKD). [As characterized in subsequent cross-architecture distillation literature].
- Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., & Khan, F. S. (2022). EdgeNeXt: Efficiently amalgamated CNN-transformer architecture for mobile vision applications. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW). arXiv:2206.10589.
- Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. International Conference on Learning Representations (ICLR).
- Nandhini, S., & Usha, K. (2024). Lightweight deep learning models for agricultural edge deployment: A systematic review. [Domain synthesis reference for edge-deployment thresholds].
- Park, N., & Kim, S. (2022). How do vision transformers work? International Conference on Learning Representations (ICLR).
- Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., & Ye, Q. (2021). Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
- Perera, S., Erzurumlu, Y., Gulati, D., & Yilmaz, A. (2024). MobileUNETR: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. arXiv:2409.03062.

- Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for thin deep nets. International Conference on Learning Representations (ICLR).
- ScienceDirect (2024). Optimized hybrid knowledge distillation framework for cardiac MRI classification. S0952197624015884.
- ScienceDirect (2025). Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. S0952197625000570.
- ScienceDirect (2025). LW-CTrans: A lightweight hybrid network of CNN and Transformer for 3D medical image segmentation (incl. LiteMobileV3 distillation). S1361841525000921.
- ScienceDirect (2025). CMKD-Net: A cross-modal knowledge distillation method for remote sensing image classification. S0273117725003333.
- ScienceDirect (2024). TSD-DETR: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. S0952197624016944.
- Springer (2025). A knowledge distillation method based on detection transformer for farmland obstacle detection. Complex & Intelligent Systems.
- SPIE (2025). Lightweight TransUNet with knowledge distillation for efficient medical image segmentation.
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (ICML).
- Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive representation distillation. International Conference on Learning Representations (ICLR).
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning (ICML).
- Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).
- Wang, G., Li, K., et al. (2024). Object detection of lightweight transformer based on knowledge distillation. Journal of System Simulation, 36(11).
- Wang, L., & Yoon, K. J. (2021/2022). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 3048-3068.
- Wang, Y., et al. (2024). KD-DETR: Knowledge distillation for detection transformer with consistent distillation points sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). CvT: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

- Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022). TinyViT: Fast pretraining distillation for small vision transformers. In Proceedings of the European Conference on Computer Vision (ECCV).
- Xu, H., et al. (2022). Knowledge distillation for tomato leaf disease detection: 163-fold compression via CNN ensemble to ShuffleNetV2. [Domain application reference, agricultural KD].
- Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., & Li, Y. (2024). ViTKD: Practical guidelines for feature-based knowledge distillation for vision transformers. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). arXiv:2209.02432.
- Zagoruyko, S., & Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. International Conference on Learning Representations (ICLR).
- Zhang, J., Peng, H., Wu, K., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022). MiniViT: Compressing vision transformers with weight multiplexing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- *Additional corpus references (cross-architecture, domain-application, and survey studies identified during the systematic search but summarized collectively in text rather than individually authored-and-dated due to preprint/anonymized venue status at time of review):*
- Cross-Architecture Distillation Made Simple with Redundancy Suppression. arXiv:2507.21844 (2025).
- Multi-Level Decoupled Relational Distillation for Heterogeneous Architectures (MLDR-KD). arXiv:2502.06189 (2025).
- Fuse Before Transfer: Knowledge Fusion for Heterogeneous Distillation. arXiv:2410.12342 (2024).
- Revisiting Cross-Architecture Distillation: Adaptive Dual-Teacher Transfer for Lightweight Video Models. AAAI. arXiv:2511.09469 (2026).
- Distilling Knowledge from CNN-Transformer Models for Enhanced Human Action Recognition. arXiv:2311.01283 (2023).
- X-Distill: Cross-Architecture Vision Distillation for Visuomotor Learning. arXiv:2601.11269 (2026).
- HDKD: Hybrid Data-Efficient Knowledge Distillation Network for Medical Image Classification. arXiv:2407.07516 (2025).
- MedLiteNet: Lightweight Hybrid Medical Image Segmentation Model. arXiv:2509.03041 (2025).
- LightBTseg: A Lightweight Breast Tumor Segmentation Model via Dual-Path Joint Knowledge Distillation. arXiv:2311.11086 (2023).

- Hierarchical Semantically Aligned Heterogeneous Knowledge Distillation (HSAH-KD) for crop leaf disease recognition. ScienceDirect S0952197625000570 (2025).
- Lightweight Remote Sensing Scene Classification on Edge Devices via Knowledge Distillation and Early-exit. arXiv:2507.20623 (2025).
- Applications of Knowledge Distillation in Remote Sensing: A Survey. arXiv:2409.12111 (2024).
- A Novel Lightweight Transformer with Edge-Aware Fusion for Remote Sensing Image Captioning. arXiv:2506.09429 (2025).
- Survey on Remote Sensing Scene Classification: From Traditional Methods to Large Generative AI Models. arXiv:2603.26751 (2026).
- Knowledge Distillation in Object Detection: A Survey from CNN to Transformer. PMC12788226 (2025).
- SO-DETR: Leveraging Dual-Domain Features and Knowledge Distillation for Small Object Detection. arXiv:2504.11470 (2025).
- Image Recognition with Online Lightweight Vision Transformer: A Survey. arXiv:2505.03113 (2025).
- Knowledge Distillation in Vision Transformers: A Critical Review. arXiv:2302.02108 (2023).
- Knowledge Distillation in Visual Algorithms: A Survey. Chinese Journal of Computer Research and Development (CRAD) (2025).
- A Survey of the Vision Transformers and Their CNN-Transformer Based Variants. arXiv:2305.09880 (2023).
- Teacher-Student Architecture for Knowledge Distillation: A Survey. arXiv:2308.04268 (2023).
- Theoretical Perspectives on Knowledge Distillation: A Review. WIREs Computational Statistics (2025).
- A Lightweight Distillation CNN-Transformer Architecture for Remote Sensing Image Super-Resolution. Taylor & Francis (2023).

Publisher's Note

ERUDEXA PUBLISHING remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of ERUDEXA PUBLISHING and/or the editor(s). ERUDEXA PUBLISHING disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.