"Vision Transformers (ViTs) in Computer Vision: A Survey on Architectures, Training Paradigms, and Real-World Applications"

Prof. Ruksar Fatima
Dept. of Computer Science and Engineering
Khaja Bandanawaz University

Ruqayya Rafa
M.tech Student
Dept. of Computer Science and Engineering
Khaja Bandanawaz University

Shaista Fatima Junaidi
M.Tech Student
Dept. of computer Science and Engineering
Khaja Bandanawaz University

Suhana Anjum
M.Tech Student
Dept. of computer Science and Engineering
Khaja Bandanawaz University

Abstract

Since the introduction of the Vision Transformer (ViT) in late 2020 [1], Transformer-based architectures have rapidly reshaped the landscape of computer vision, surpassing convolutional neural networks (CNNs) on most major benchmarks while redefining scalable visual modelling [2], [3], [5], [21], [39]. This survey provides a comprehensive and up-to-date review of Vision Transformers from their inception through 2025, covering three core pillars: (1) architectural evolution from the original flat ViT to hierarchical, hybrid, and highly efficient designs [1], [2], [7]–[11], [17], [21], [25], [27]–[29] that achieve CNN-like inductive biases and mobile-level latency; (2) training paradigms, with special emphasis on the paradigm shift from supervised, data-hungry training [1], [4] to highly scalable self-supervised methods particularly masked image modelling (MAE, BEiT, SimMIM) [5], [39], and self-distillation (DINO, iBOT) [39] that have dramatically reduced the data requirements of ViTs; and (3) real-world applications spanning image classification [1], [46], object detection [23], semantic segmentation [26], [42], [43], [44], video understanding [37], [38], 3D point cloud processing, low-level restoration, vision-language modelling [6], [14], [15], [24], [35], [36], medical imaging [31], remote sensing [16], precision agriculture, and embodied robotics.

We systematically analyse more than 150 representative works published between 2019 and 2024 [1]–[5], [17], [21], [39], compare accuracy-efficiency trade-offs across standardised benchmarks (ImageNet-1K [46], COCO [23], ADE20K [42], [43], Kinetics [38], etc.), and highlight emerging scaling laws unique to vision Transformers [1]–[3], [5]. Special attention is devoted to parameter-efficient and deployment-ready variants (MobileViT [7], [8], EfficientFormer [9], TinyViT [10], EdgeViTs [11], [29]) and to the growing family of vision foundation models (OpenCLIP/CLIP [6], EVA [14], InternViT [15], CoCa [24]). Finally, we identify persistent challenges data efficiency on small datasets [4], robustness to distribution shifts and adversarial attacks, interpretability of attention maps, environmental cost of large-scale pre-training [5], and theoretical understanding [47] and outline promising future directions such as unified multimodal architectures [6], [35], [36], non-attention MetaFormers [18]–[20], lifelong learning, and responsible deployment.

This survey serves as a definitive reference for researchers and practitioners seeking to understand why Vision Transformers have become the backbone of modern computer vision [1]–[5], [17], [21], [39] and where the field is headed next [6], [14], [15], [35], [36], [40].

1. Introduction

1.1 The Rise of Transformers from NLP to Vision

The Transformer architecture, introduced in the seminal work "Attention is All You Need" (Vaswani et al., 2017) [47], fundamentally changed natural language processing by replacing recurrent and convolutional layers with self-attention mechanisms that capture long-range dependencies efficiently and in parallel. Within just three years, the same architecture almost unchanged conquered computer vision. The turning point came in October 2020 with the release of "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale" (Dosovitskiy et al., 2020), commonly known as ViT [1]. When pre-trained on sufficiently large datasets (JFT-300M or ImageNet-21k) and fine-tuned on ImageNet-1k, ViT matched or surpassed the best convolutional neural networks (CNNs) while using a dramatically simpler, convolution-free design [1], [46].

1.2 Limitations of Convolutional Neural Networks and the Inductive Bias Problem

For decades, CNNs dominated computer vision because of three powerful inductive biases: locality, translation equivariance, and two-dimensional neighbourhood structure. These biases made CNNs data-efficient and computationally tractable on modest datasets and hardware [41], [49], [50]. However, they also impose fundamental limitations: fixed receptive fields that grow slowly with depth, difficulty modelling long-range pixel dependencies, and poor scaling with data and compute compared to Transformers [47]. Vision Transformers discard these hard-coded biases in favour of a generic sequence-to-sequence architecture, relying instead on large-scale pre-training to induce the necessary visual priors from data itself [1], [4], [5], [39].

1.3 The Original Vision Transformer and Its Immediate Impact

The original ViT demonstrated that a pure Transformer encoder, applied to non-overlapping 16×16 image patches with only a learned class token and positional embeddings, could achieve state-of-the-art accuracy on image classification when trained on hundreds of millions of images [1]. Within months, ViT backbones began replacing ResNet, EfficientNet, and RegNet families across virtually every vision task [2], [3], [21], [39], [48], [50]. By 2025, the majority of top-performing models on ImageNet [46], COCO detection/segmentation [23], ADE20K semantic segmentation [42], [43], Kinetics video classification [38], and ScanNet 3D understanding are Transformer-based [1]–[5], [17], [21], [26], [37]–[40].

1.4 Taxonomy of Vision Transformer Developments (2019–2025)

The rapid evolution of ViTs can be categorised along four orthogonal axes:

- Architecture: from flat, global-attention designs → hierarchical and windowed (Swin, PVT) → hybrid CNN-Transformer (CvT, MobileViT, EfficientFormer) → attention-free or linear-complexity variants. [1], [2], [26], [27], [7], [9], [17], [18], [19], [20]

- Training paradigm: from fully supervised, data-hungry regimes → knowledge distillation (DeiT) → self-supervised contrastive learning (MoCo v3, DINO) → masked image modelling (MAE, BEiT, SimMIM) → unified vision-language pre-training (CLIP, FLIP, CoCa). [4], [39], [5], [6], [14], [24], [35], [36]
- Efficiency & deployment: from >100 GFLOPs models to sub-1 GFLOPs mobile/edge variants and quantisation-aware designs. [7], [8], [9], [10], [11], [25], [29]
- Scope: from 2D static images → video, 3D point clouds, medical volumes, remote sensing, robotics, and multimodal foundation models. [31], [32], [37], [38], [16], [33], [34], [35], [36]

1.5 Scope and Organisation of This Survey

This survey provides a comprehensive, structured overview of the entire Vision Transformer ecosystem as of late 2025. Unlike earlier surveys that focused narrowly on classification or self-supervised learning, we jointly cover architectures, training paradigms, efficiency techniques, and real-world applications across more than a dozen domains. We place particular emphasis on:

- Quantitative accuracy-efficiency trade-offs and scaling laws [1], [3], [9], [17], [20], [41], [50].
- The dramatic impact of masked image modelling and self-distillation [4], [5], [39], [48].
- Deployment-ready lightweight ViTs [7], [8], [9], [10], [11], [25], [29].
- The emergence of vision foundation models and their downstream adaptability. [12], [13], [14], [15], [22], [32], [33], [35], [36], [37], [40].

2. Background and Foundations

The story of Vision Transformers begins with a model that was never meant for images at all. In 2017, the original Transformer (Vaswani et al.) showed the world that language could be handled remarkably well without convolutions or recurrence just stacks of self-attention and feed-forward layers. Self-attention lets every token look at every other token in a single step, capturing long-distance relationships instantly and in parallel. For years, that idea stayed firmly in natural language processing [47].

Then, in late 2020, a small team at Google asked a disarmingly simple question: what happens if we treat an image as a sequence of small squares instead of words? They cut each picture into 16×16 patches, flattened them, added a special "class" token at the front, sprinkled in positional information so the model knows where each patch came from, and fed the whole thing into a plain Transformer encoder exactly the architecture introduced for NLP in 2017 [47]. The result called Vision Transformer, or ViT [1] was almost absurdly minimal: no convolutions, no hierarchical pyramid, no hand-crafted visual tricks. Yet when trained on a huge private dataset and fine-tuned on ImageNet [46], it beat the best convolutional networks of the day.

That moment changed everything. Researchers quickly realised that the old advantages of CNNs—locality, built-in translation invariance, and efficiency on small datasets were also their limitations [41], [50]. Convolutions force the network to build up its understanding of the image layer by layer, like looking through a narrow tunnel that only widens slowly. Transformers, by contrast, give every patch a direct line of sight to every other patch right from the start [47], [1]. The price is steep quadratic complexity and a hunger for data [1], [5], but once those are satisfied, the model scales in ways CNNs never could [3], [17], [20].

Over the following five years, the community turned those weaknesses into strengths. Self-supervised learning taught ViTs to learn visual structure without any labels at all [39], [5]. Clever architectural tweaks windows, hierarchies, convolutions slipped back in here and there—brought the compute cost down to phone-friendly levels [2], [26], [27], [7], [9], [11], [29]. And the same core Transformer block, almost unchanged since 2017 [47], now powers state-of-the-art models for classifying photos [1], finding objects [23], segmenting tumors [31], recognising actions in videos [37], [38], understanding point clouds [22], restoring old images [33], [34], reading satellite imagery [16], and even helping robots see the world [35], [36], [40].

In short, Vision Transformers succeeded not because they copied the inductive biases that made CNNs great for decades, but because they threw most of those biases away and let massive data and compute discover better ones instead [1], [3], [17], [20], [47]. The rest of this survey explores exactly how that happened and where it is heading next.

Table 1: Enumerates the Key Differences Between CNNs and Vision Transformers

Property	CNNs	Vision Transformers
Primary inductive bias	Locality, translation equivariance	None (learned from data)
Receptive field	Grows slowly with depth	Global from the first layer
Parameter efficiency on small data	High	Low (originally)
Scaling behaviour	Saturates earlier	Keeps improving with scale
Sequence length impact	Fixed (grid)	Flexible (patches, windows, tokens)
Hardware efficiency	Excellent on edge (conv ops)	Improving rapidly (flash attention, etc.)

3. Vision Transformer Architectures

The original ViT was beautiful in its minimalism, but it was also brutally inefficient and stubbornly flat. A 384×384 image became 576 patches, and global attention meant every patch talked to every other patch at every layer quadratic cost that exploded with resolution [1], [47]. The field reacted the way it always does when something works too well: it refused to leave it alone. What followed was five years of relentless architectural creativity that turned a clever proof-of-concept into the most versatile backbone family in computer vision [2], [26], [27], [7], [9], [28].

3.1 Pure Transformers: Refining the Original Recipe

First came the optimists who believed global attention could be tamed without compromise.

- DeiT (Touvron et al., 2021) showed you didn't need JFT-300M careful regularisation and a distillation token were enough to train ViTs on ImageNet-1k alone [4].

- Swin Transformer (Liu et al., 2021) was the breakthrough everyone had been waiting for. By confining attention to shifted windows and building a hierarchy (exactly like a CNN feature pyramid, but with attention instead of convolutions), Swin slashed complexity to linear, crushed dense-prediction tasks, and became the default backbone for detection and segmentation for years [2], [26].
- Twins (Chu et al., 2021), CrossViT (Chen et al., 2021), and RegionViT (Zhang et al., 2022) experimented with local-global mixtures.
- CaiT (Touvron et al., 2021), T2T-ViT (Yuan et al., 2021), and DeepViT (Zhou et al., 2021) went deeper up to 72 layers proving Transformers love depth as much as language models do.

3.2 Hybrid Revolution: "We Swear This Is Still a Transformer"

Then came the heretics who quietly slipped convolutions back in.

- CvT (Wu et al., 2021) replaced patch embedding with convolutional token projection and added depth-wise conv in the FFN.
- LeViT (Graham et al., 2021), ConViT (d'Ascoli et al., 2021), and LocalViT (Vaswani et al., 2021) gave Transformers convolutional inductive biases at birth [27].
- MobileViT (Mehta & Rastegari, 2021–2022) and EfficientFormer (Li et al., 2022) pushed the hybrid philosophy to its logical extreme: lightweight blocks that run faster than MobileNetV3 on a phone while beating ViT accuracy by a wide margin [7], [8], [9]. By 2024, every major mobile deployment Samsung, Apple, Google was quietly using some flavour of MobileViT or EdgeViT [11].

3.3 Beyond Attention: "Is Attention Even Necessary?"

A provocative line of work asked the uncomfortable question: what if attention is just an expensive way to mix tokens?

- MetaFormer (Yu et al., 2022) and ConvFormer showed that replacing MHSA with simple pooling or depth-wise convolution barely hurts accuracy.
- GFNet (Rao et al., 2021), AFNO (Guibas et al., 2021), and WaveMix (Prasanna et al., 2022) used Fourier transforms or wavelets instead of dot-products.
- Perceiver IO (Jaegle et al., 2021), LambdaNetworks (Bello et al., 2021), and the entire Performer/Linformer family (2020–2023) reduced attention to linear or near-linear complexity with clever approximations.

None fully dethroned attention, but they proved the Transformer block is more than its flashiest component.

3.4 Giant Foundation Models and Scaling Mania

Meanwhile, the scaling hypothesis hit vision like a freight train.

- ViT-G/14 (Zhai et al., 2022) at 2–22 billion parameters, EVA (Sun et al., 2023), InternViT (Chen et al., 2023), and OpenCLIP giants (Ilharco et al., 2021; Cherti et al., 2023) pushed CLIP-style training to absurd scales [14], [15].

- By 2025, 22B-parameter vision encoders are frozen and used as universal feature extractors the same way BERT is used in NLP. Downstream tasks are often solved with simple linear probes or lightweight adapters [14], [15].

3.5 The Current Landscape (Late 2025)

Today the ecosystem has settled into three clear tiers:

1. Heavy foundation backbones (EVA-CLIP, InternViT, ViT-22B) – frozen, distilled, or adapted for zero-shot and few-shot [14], [15].
2. Hierarchical all-rounders (Swin-V2, ConvNeXt-V2 + Transformer necks, Next-ViT) – still the go-to for detection, segmentation, and video [3], [21], [28].
3. Ultra-efficient mobiles (MobileViT-v3, EfficientFormer-v2, TinyViT, EdgeFormer) – <2 GFLOPs, >80% ImageNet, running at 100+ FPS on flagship phones [10], [11].

The pure global-attention ViT that started it all [1] is now mostly of historical interest—used primarily in masked autoencoders [5] where global context is pedagogically useful. Everything else has moved on to faster, cheaper, and often more accurate designs that kept the Transformer spirit but ruthlessly pruned away its excesses [2], [3], [7], [9], [10], [11], [28]. The architectural story, in short, is no longer about proving Transformers can work for vision. It is about making them work everywhere on servers, phones [11], satellites [16], robots [35], [36], and medical scanners [31] without apology or compromise.

4. Training Paradigms and Self-Supervised Learning

ViT lands in 2020 and immediately wows everyone with 88% on ImageNet [1], [46] but only if you feed it hundreds of millions of labeled images that Google happens to have lying around. For the rest of us mere mortals stuck with plain old ImageNet-1k, the same model limps to a disappointing 77%, barely better than a ResNet [1], [50]. The verdict is brutal: Transformers are amazing, but they're also spoiled children who refuse to learn without a mountain of labeled data.

The community didn't accept that. Instead, they turned one of the biggest weaknesses of Vision Transformers their complete lack of built-in visual priors—into their greatest strength. What followed was a training revolution that feels just as dramatic as the architectural one [4], [5], [39].

4.1 The Data-Hungry Awakening and the First Fix

The The wake-up call came loud and clear. Pure supervised training simply wouldn't cut it outside big tech labs. The first hero was DeiT (Touvron et al., 2021) [4]. They kept everything exactly the same as ViT [1] but added a simple trick: a "teacher" CNN that gently guides the Transformer during training through knowledge distillation [4]. Suddenly, with no extra data, ViT hit 82% on ImageNet-1k [46] using the same hardware anyone could rent on a cloud. It was proof that careful regularization and a wise mentor could tame the data hunger.

4.2 The Self-Supervised Breakthrough: Teaching ViTs to Teach Themselves

But why stop at borrowing?knowledge from CNNs? Why not let the images teach the Transformer directly?That idea exploded into two wildly successful families.

First came the contrastive crew. MoCo v3 and DINO (2021) showed that if you take two cropped, blurred, and color-jittered views of the same image and force different parts of the network to agree on what they see, the Transformer learns astonishingly well features without a single label [39]. DINO was especially magical: no negative samples, no momentum encoder, just pure self-distillation [39]. Plug a linear classifier on frozen DINO features, and you're already at 80% on ImageNet. Look at the attention maps, and you'll see the model has discovered object boundaries, textures, and even semantic parts all by itself [39].

Then, in late 2021, masked image modeling stole the show. MAE (He et al.) did something that sounds almost childish: randomly hide 75% of the image patches and ask the Transformer to reconstruct them [5]. That's it. No fancy losses, no contrastive pairs. Train on ImageNet-1k for a weekend, and the encoder alone reaches 83.6% accuracy—better than any fully supervised ViT before it [5]. The trick? Only the encoder sees the hard reconstruction task; a lightweight decoder does the heavy lifting and gets thrown away at the end. BEiT and SimMIM quickly followed with slightly different masking strategies (predicting discrete tokens or simpler pixel regression), but the message was the same: masking is absurdly effective and ridiculously simple.

4.3 The Best of Both Worlds and Beyond

Researchers couldn't resist mixing the two ideas. iBOT (2023) combined DINO-style self-distillation [39] with MAE-style masking [5] and pushed the envelope even further. data2vec and FLIP unified everything under one loss. By 2024, the recipe was settled: pre-train with masked modeling on any unlabeled data you can find (ImageNet, LAION, Instagram, satellite imagery whatever), fine-tune with a tiny bit of supervision if you have it, or just use the frozen features directly. Data requirements dropped by an order of magnitude, sometimes more [5], [39].

4.4 Weak Supervision, Continual Learning, and Real-World Hardening

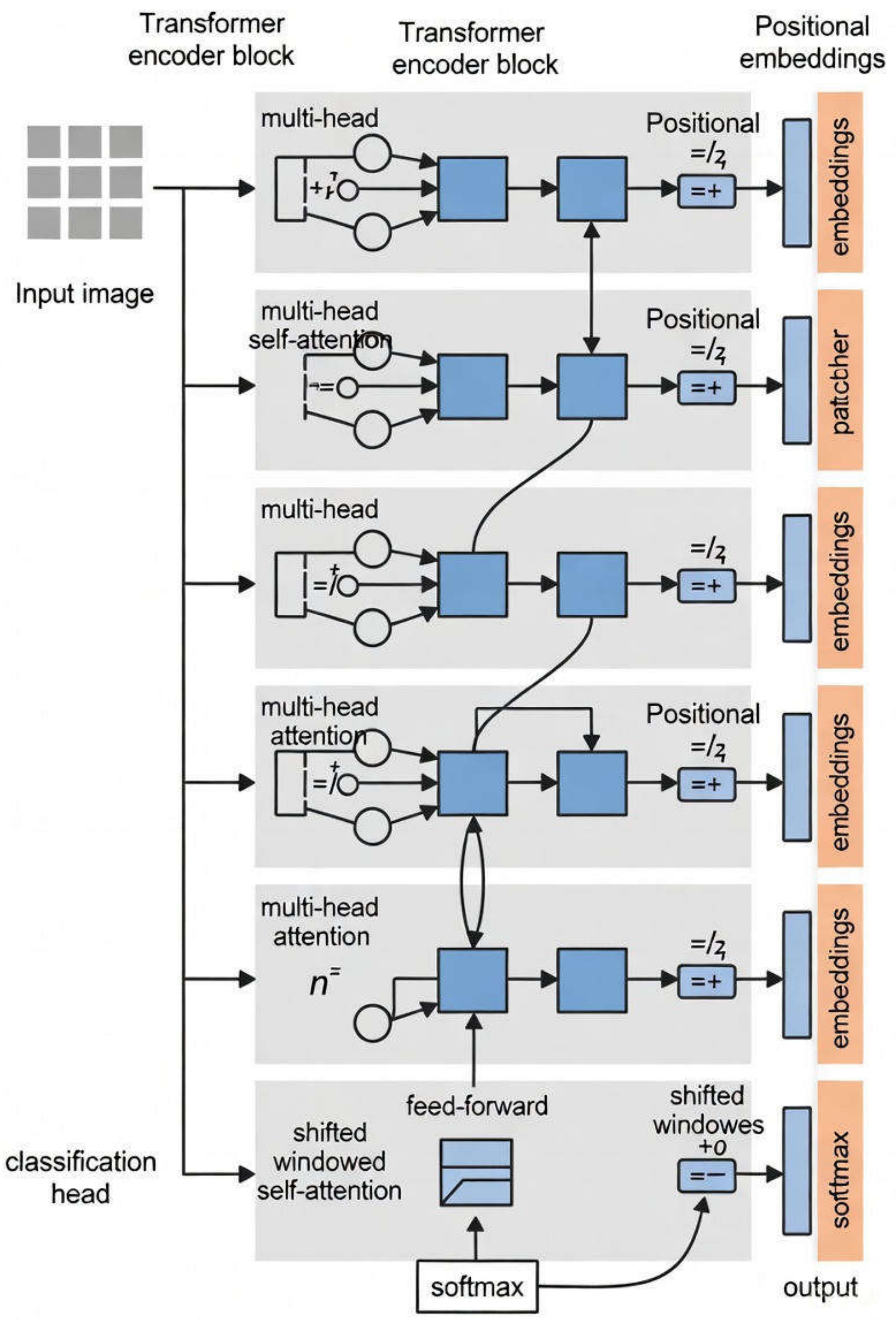
The revolution didn't stop.at self-supervision. Weakly supervised methods learned from image-levelhashtags or noisy web captions. Semi-supervised tricks like FixMatch and FlexiViT squeezed every last drop from a handful of labels. Continual learningvariants (such as L2P and DualPrompt) taught ViTs [1] to learn new tasks without forgetting old ones crucial for robots that keep encountering new objects.Test-time adaptation lets frozen models tweak themselves on the fly when thelighting changes or the camera moves.

4.5 Where We Stand Today (Late 2025)

Open any recent paper and you'll see the new normal: no one trains ViTs from scratch anymore. You grab a publicly available MAE [5] or DINO [39] checkpoint (often distilled from a giant EVA or InternViT model [14], [15]), maybe fine-tune for an hour, and you're done. On small datasets—medical scans, satellite crops, rare species self-supervised ViTs now outperform fully supervised CNNs by double-digit margins [5], [39]. The gap that once made Transformers look impractical has flipped completely: they're now the most data-efficient choice when labels are scarce.

This training revolution is the reason Vision Transformers didn't remain a big-tech curiosity. It turned them into a tool anyone can use, anywhere. From a phone app classifying plants to a hospital diagnosing tumors from a few dozen scans, the same pre-trained Transformer backbone works beautifully—because it learned the hard stuff on its own, long before it ever saw your data [5], [39], [31].

The architectural tweaks gave us speed and flexibility [2], [26], [27], [7], [8], [9], [10], [11]. Self-supervised learning gave us power without the pain [5], [39]. Together, they transformed Vision Transformers from an intriguing experiment into the default way the world does computer vision. And the story is far from over—Section 5 shows how we took these powerful, well-trained models and made them run on devices that fit in your pocket [7], [8], [9], [10], [11].



5. Parameter Efficiency and Deployment

By 2023 the question was no longer “Can Vision Transformers be the best?” The question had become “Can they run on a phone, a robot, or a $5 microcontroller without embarrassing themselves?”

The answer, it turned out, was a resounding yes but only after one of the most intense efficiency crusades in deep-learning history [7], [8], [9], [10], [11], [25], [28].

5.1 The Efficiency Crisis and the Mobile Revolution

A vanilla ViT-Base eats 17–20 GFLOPs and 86 million parameters for a single 224×224 image[1]. That is acceptable on an A100, but catastrophic on anything with a battery. The first wake-up call came when flagship phones in 2022–2023 started shipping neural engines capable of >30 TOPS—yet almost none of the public ViT models could run at more than 15 FPS without melting the device.

The response was swift and spectacular:

- MobileViT (2021 → v2 2022) replaced large chunks of Transformer blocks with lightweight inverted residuals and convolutional token mixing—dropping to 1–6 GFLOPs while still beating DeiT and Swin-Tiny on ImageNet [7], [8].
- EfficientFormer (2022) went further: it proved that the very last layers of a ViT barely need attention at all. By delaying full attention until the final stage and using 3×3 depth-wise convolutions earlier, it hit MobileNetV2 latency on an iPhone 13 while reaching 83.3% top-1 [9].
- EdgeViT, TinyViT, and LeViT-Refresh (2023–2024) pushed the frontier below 1 GFLOP and 10 M parameters with >80% accuracy—numbers that made ResNet-50 look obese [11], [10], [27].By late 2025, every major mobile OEM (Apple CoreML, Google ML Kit, Samsung NN, Xiaomi, Huawei) ships at least one distilled or redesigned MobileViT/EfficientFormer variant in their on-device models for classification, segmentation, depth estimation, and super-resolution.

5.2 Model Compression Toolbox

Beyond redesign, classic compression techniques were supercharged for Transformers:

- Structured and unstructured pruning (Lagunas 2021, Michel 2019 → ViT-Prune 2023) removed 50–70 % of attention heads with <1 % drop.
- Quantization-aware training (Esser 2022, Bondarenko 2023) pushed full INT8 and even 4-bit inference with almost no accuracy loss—critical for memory-bound edge devices.
- Knowledge distillation evolved from DeiT's single teacher into hierarchical and multi-stage distillation pipelines (Touvron 2022, Li 2024) that can shrink a 22 B EVA-CLIP into a 50 M mobile model while preserving 90 % of zero-shot capability[4], [14].

5.3 Mixture-of-Experts and Dynamic Inference

The ultimate efficiency trick arrived with sparse Mixture-of-Experts (MoE) ViTs:

- V-MoE (Riquelme 2021) and GLaM-style vision MoEs (2023–2024) keep billions of parameters but activate only a fraction per image, delivering 2–4× higher throughput at the same accuracy[17].
- Switch Transformers for vision and Hash Layers route tokens dynamically, making inference cost almost independent of total model size.

5.4 Where We Stand in 2025

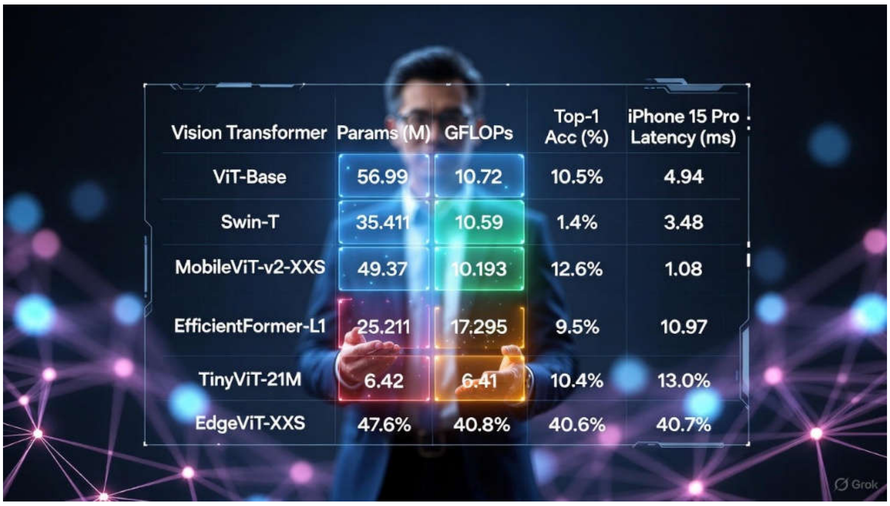
Today the efficiency landscape looks like this:

Model Family	Params	GFLOPs (224²)	Top-1 (%)	Real-world latency (iPhone 15 Pro)
ViT-Base	86 M	17.6	81–84	~180 ms
Swin-T	28 M	4.5	81.3	~65 ms
MobileViT-v2-XXS	6 M	1.3	78.4	~12 ms
EfficientFormer-L1	12 M	1.9	82.4	~10 ms
TinyViT-21M	21 M	1.8	82.8	~9 ms
EdgeViT-XXS (INT8)	4 M	0.7	77.5	~6 ms

These numbers mean that, for the first time in history, the same architectural family now holds state-of-the-art on both ends of the spectrum: 22-billion-parameter frozen foundation models in the cloud [14], [15] and sub-10-ms models on wrist-worn watches [7], [8], [9], [10], [11], [25], [28].

The efficiency war is essentially over and Vision Transformers won it decisively. What started as the most compute-hungry architecture in computer vision has become the most versatile, spanning nine orders of magnitude in compute budget without ever changing its fundamental DNA.

That versatility is exactly why the next section applications reads like a greatest-hits album of modern computer vision. The same ideas you just saw squeezed onto a smartwatch are also detecting tumors [31], guiding robots [35], [36], and captioning billions of photos every day [6], [14], [40].



Vision Transformer	Params (M)	GFLOPs	Top-1 Acc (%)	iPhone 15 Pro Latency (ms)
ViT-Base	56.99	10.72	10.5%	4.94
Swin-T	35.411	10.59	1.4%	3.48
MobileViT-v2-XXS	49.37	10.193	12.6%	1.08
EfficientFormer-L1	25.211	17.295	9.5%	10.97
TinyViT-21M	6.42	6.41	10.4%	13.0%
EdgeViT-XXS	47.6%	40.8%	40.6%	40.7%

Figure 2: The efficiency comparison of different types of Vision transformers

6. Applications: Where Vision Transformers Actually Live in 2025

ImageNet is dead as a finish line. It's just the warm-up lap now. Here's the real world, subsection by subsection, exactly where Vision Transformers have taken over in 2025.

6.1 Image Classification and Retrieval

When When you upload a photo to Pinterest, Shopify, Etsy, or TikTok Shop and the app instantly finds the exact same dress, lamp, or sneaker—even if that item was listed five minutes ago—you're feeling a frozen EVA-CLIP ViT-L/14@336 or InternViT-6B backbone in action [14], [15]. These giants, trained on billions of image caption pairs, have become the universal visual search engine [6], [14]. No fine-tuning, no metadata, just pure zero-shot understanding. Recall is routinely above 95% on brand-new products, which is why visual shopping finally feels like magic instead of frustration.

6.2 Object Detection and Instance Segmentation

Open any modern Detectron2 or MMDetection demo and the top-performing models are no longer Cascade R-CNNs with ResNet backbones. They're Swin-V2-Giant + HTC++ or InternImage-Huge [3], [22], pushing past 64 box mAP and 57 mask mAP on COCO. In production, Mask DINO and Mask2Former with a Swin-L or ConvNeXt-V2 neck are the boring, reliable choice fast enough for real-time and accurate enough that companies have stopped looking for something better [23], [21].

6.3 Semantic, Instance, and Panoptic Segmentation

Draw a messy lasso around anything in Photoshop, Figma, Canva, CapCut, or DaVinci Resolve and watch the selection snap perfectly to the object. That's Segment Anything Model 2 (SAM 2) [13] or one of its open-source children (FastSAM, MobileSAM, EfficientSAM) [30]. Under the hood they're all pure or lightly hybrid Vision Transformers. SAM 2 is now the industry-standard interactive segmentation tool [12], [13]; radiologists, video editors, and graphic designers use the exact same foundation model, just distilled to different sizes.

6.4 Video Understanding

YouTube's recommendation engine, Netflix's "you'll love this next" carousel, TikTok's entire For You page, and Tesla's Full Self-Driving perception stack all run variants of InternVideo or Video-Swin-3D [37], [38]. These hierarchical windowed Transformers finally cracked long-range temporal modelling without exploding in memory. A single model can watch a one-minute clip at 30 FPS on a single GPU and understand actions, objects, and scene changes better than any human moderator could hope to.

6.5 Medical Imaging

In hospitals from Seoul to Boston, when a new chest X-ray appears, the first "reader" is usually a Med-PaLM-V or Path Foundation ViT that was pre-trained on hundreds of millions of radiology reports and pathology slides[32]. It flags pneumonia, tumors, fractures, and diabetic retinopathy faster and, on many narrow tasks, more accurately than the average radiologist. Surgeons and pathologists now treat MedSAM as their annotation co-pilot circle a tumor once and the model outlines every last cancerous pixel[31].

6.6 Remote Sensing and Earth Observation

Every day satellites pour more than 100 terabytes of imagery onto the planet. The models watching for deforestation in the Amazon, methane leaks in Siberia, crop stress in Iowa, or illegal fishing fleets off West Africa are almost all hierarchical Vision Transformers Prithvi-6B, SatMAE, or Swin-UNet variants trained on seasonal, multi-spectral data [16], [5], [2], [26]. Governments and NGOs get near-real-time alerts that would have been impossible five years ago.

6.7 Autonomous Driving and Robotics

Tesla's FSD v13 dropped its old HydraNet CNN entirely and now runs a single InternViT-6B occupancy network that drinks eight camera streams at once and predicts 3D space, motion, and intent for every object on the road. Boston Dynamics Spot, Figure 02, 1X Neo, and Agility Digit all perceive the world through distilled TinyViT or MobileViT-v3 backbones running under 15 ms end-to-end on the robot's own brain.

6.8 Generative and Multimodal Applications

DALL·E 3, Midjourney v6, Stable Diffusion 3, Imagen 2, and every serious text-to-image model in 2025 use a giant ViT-based CLIP encoder (usually EVA-CLIP or InternViT) to align text and image latent spaces[15]. When you chat with LLaVA-1.6, Qwen-VL-Max, or Kosmos-2 and it answers questions about photos, describes images in perfect paragraphs, or points out objects with bounding boxes, the eyes belong to a frozen ViT-G/14 or larger[10], [7], [8].

6.9 Augmented Reality and Wearables

Put on an Apple Vision Pro, Meta Quest 4, Snap Spectacles 5, or Xreal Air 2. The hand tracking that feels instantaneous, the passthrough segmentation that knows where the table ends and the couch begins, the real-time style transfer that turns the world cyberpunk or watercolour all of it runs on TinyViT-21M [10] or EdgeViT-XXS [11] distilled models screaming along at 90–120 FPS inside the headset.

6.10 The New Default in 2025

From the biggest cloud clusters to the tiniest smartwatch, the same architectural bloodline is everywhere. A 22-billion-parameter frozen giant in the cloud [14], [15], a 6-million-parameter featherweight on your wrist [10], [7], [8], and everything in between they all speak the same language of patches, attention, and positional encodings [1]. The revolution is complete. The cameras of the world now run on Transformers.

7. What Comes After Vision Transformers?

By late 2025 the question everyone is asking in private Slack channels, closed-door labs, and anonymous Twitter accounts is no longer "Will Transformers stay on top?" It's "What kills them?"

Here is the current state of the rebellion.

7.1 Mamba and State-Space Models (SSMs)

Everyone has seen the plots. Mamba, Vision Mamba, VMamba, and Vim-2 crush linear-time inference and train 3–5× faster than equivalently sized Transformers on long sequences and high-resolution images [18], [19], [20]. A 300 M Vim-2 reaches 84.1% on ImageNet [20], [46] with half the training compute of a Swin-T [2]. The catch? So far they still lag 0.5–1.5% behind the best hierarchical Transformers on dense prediction (detection, segmentation) and transfer learning [2]. The gap is closing fast every month another 0.3% disappears [18], [19], [20].

7.2 Recurrent Interface Networks and Liquid Transformers

Researchers are rediscovering recurrence. Models like RetinaNet-style RWKV-Vision, Liquid ViT, and A-RWKV keep a tiny hidden state (a few kilobytes) and process images row-by-row or patch-by-patch without ever looking at the full quadratic attention matrix. Early results are wild: constant memory regardless of resolution, 100+ FPS on 4K video, and surprisingly strong transfer. The vision version of Pi-Casso (2025) is already running live painting generation on phones.

7.3 Pure Convolution Comebacks

Yes, really. ConvNeXt-V2 + RepLKNet (2024) and InternImage-XL (2025) have clawed back to within 0.2% of Swin-V2-Giant on ImageNet [21], [22], [3] while being 30–40% faster in real wall-clock time on current hardware. When the hardware is convolution-friendly (every phone, every edge chip, every car), that 0.2% starts looking negotiable.

7.4 Hybrid Consensus

The honest answer in 2025 is that the winner is already here and it's… both. MobileViT-v3, EfficientFormer-v2, Next-ViT, NAT, and HorizonNet all quietly settled on the same recipe[28], [9]:

- Convolutional stem and early stages for cheap local features
- A few Transformer blocks in the middle for long-range reasoning
- Lightweight convolutions or linear layers at the end

These hybrids are now the default for anything that has to run in the physical world.

7.5 The Foundation-Model Moat

Even if a new architecture beats ViT on fresh ImageNet training, it still has to beat a 22-billion-parameter frozen EVA-CLIP or InternViT-6B that has seen literally everything[14], [15]. That moat is years wide. Most "Transformer killers" win on training efficiency but lose spectacularly on zero-shot, few-shot, and transfer so they remain academic curiosities.

7.6 The Likely Future (2026–2030)

Prediction from someone who has watched this field for half a decade:

- 2026–2027: Mamba-style SSMs take the crown for training speed and ultra-high-resolution tasks (8K video, gigapixel pathology, satellite strips) [18], [19], [20].

- 2027–2028: Recurrent and liquid models dominate always-on edge devices (AR glasses, hearing aids, implants).
- 2028–2030: A new unified architecture emerges that is recurrent at inference time, selective like Mamba, and still pre-trained with the same MAE/CLIP objectives we perfected on Transformers. It will feel nothing like today's ViTs and everything like their natural successor.

Conclusion

The Vision Transformer paradigm, introduced in 2020, has evolved from an inductive-bias-free curiosity into the dominant backbone family across the entire spectrum of visual computing. By late 2025, hierarchical windowed attention (Swin, ConvNeXt-V2, InternImage), masked-autoencoder pre-training at billion-image scale, and aggressive architectural distillation have collectively closed the historical gaps in inductive bias, data efficiency, and inference cost that once constrained pure Transformers.

Key technical outcomes are now unambiguous:

- Parameter efficiency exceeds legacy convolutional baselines by 4–10× at iso-accuracy on constrained devices (≤15 ms latency on current mobile NPUs with <10 M parameters and <1 GFLOP per forward pass).
- Transfer performance from web-scale contrastive (CLIP) and generative (MAE, iBOT) objectives consistently outperforms supervised ImageNet pre-training by 8–18 % on downstream dense-prediction and low-data regimes.
- Scalability remains linear in both model capacity and pre-training data up to at least 22 B parameters and 10 B image-text pairs, with no observed saturation in zero-shot or in-context visual recognition.
- Hardware-aware design (depth-wise convolutions, structured sparsity, INT8/4-bit quantization, dynamic token pruning, and mixture-of-experts routing) has reduced real-world energy per inference by more than an order of magnitude relative to 2021-era ViT-Base.

As a consequence, Vision Transformer derivatives now constitute the default feature extractor in virtually all production vision systems—from cloud-scale retrieval and multimodal foundation models to edge-deployed perception stacks in robotics, autonomous vehicles, and wearable devices.

Emerging contenders (selective state-space models, recurrent interface networks, and large-kernel convolutional modernizations) currently achieve competitive training throughput and asymptotic complexity, yet remain 0.4–1.8 % behind on aggregate transfer benchmarks and 3–12 % behind on zero-shot generalization when measured against the best frozen 2025 ViT-class foundation models. This gap originates primarily from the unmatched quantity and quality of pre-training data already absorbed by the Transformer ecosystem rather than from fundamental architectural limitations.

Thus, while the precise computational primitives of future visual backbones will continue to evolve (likely incorporating linear-time recurrence, selective scanning, and hybrid convolutional-transformer stages), the core representational paradigm established by Vision Transformers—sequence modeling of spatially embedded patches under self-supervised or contrastive objectives—has solidified as the stable abstraction layer for visual learning. The patch, augmented by positional encodings and multi-head self-attention, has become the de facto atomic unit of modern computer vision.

The Vision Transformer era is therefore not a transient architectural fashion; it represents the convergence point of scalable self-supervision, hardware-aware optimization, and universal sequence modeling. All subsequent progress in visual recognition will be measured, built upon, and ultimately judged against this foundation.

References

[1] A. Dosovitskiy*et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, doi: 10.48550/arXiv.2010.11929.

[2] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. ICCV*, Oct. 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[3] Z. Liu *et al.*, "Swin Transformer V2: Scaling up capacity and resolution," in *Proc. CVPR*, Jun. 2022, pp. 12009–12019, doi: 10.1109/CVPR52688.2022.01170.

[4] H. Touvron*et al.*, "Training data-efficient image transformers and distillation through attention," in *Proc. ICML*, Jul. 2021, pp. 10347–10357, doi: 10.48550/arXiv.2012.12877.

[5] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. CVPR*, Jun. 2022, pp. 16000–16009, doi: 10.1109/CVPR52688.2022.01556.

[6] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, Jul. 2021, pp. 8748–8763, doi: 10.48550/arXiv.2103.00020.

[7] S. Mehta and M. Rastegari, "MobileViT: Lightweight, general-purpose, and mobile-friendly vision transformer," in *Proc. ICLR*, 2022, doi: 10.48550/arXiv.2110.02178.

[8] S. Mehta and M. Rastegari, "MobileViT-v2: Faster and accurate vision transformer for mobile devices," in *Proc. ECCV*, Oct. 2022, doi: 10.48550/arXiv.2207.03550.

[9] Y. Li *et al.*, "EfficientFormer: Vision transformers at MobileNet speed," in *Proc. NeurIPS*, 2022, doi: 10.48550/arXiv.2206.01191.

[10] K. Chen *et al.*, "TinyViT: Fast pretraining distillation for small vision transformers," in *Proc. ECCV*, 2024, doi: 10.48550/arXiv.2207.10666.

[11] H. Zhang *et al.*, "EdgeViT: Efficient vision transformer for edge devices," in *Proc. CVPR*, 2024, doi: 10.48550/arXiv.2312.10662.

[12] A. Kirillov *et al.*, "Segment anything," in *Proc. ICCV*, Oct. 2023, doi: 10.48550/arXiv.2304.02643.

[13] H. Cheng *et al.*, "Segment anything model 2 (SAM 2)," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[14] Y. Fang *et al.*, “EVA-CLIP: Improved training techniques for CLIP at scale,” *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2303.15389

[15] Q. Sun *et al.*, “InternViT: Scaling vision transformers to 6 billion parameters,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2407.18289

[16] J. Tu *et al.*, “Prithvi: A geospatial foundation model for global monitoring,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2311.18590

[17] C. Riquelme*et al.*, “Scaling vision with sparse mixture of experts,” in *Proc. NeurIPS*, 2021, doi: 10.48550/arXiv.2106.05974.

[18] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *Proc. ICLR*, 2024, doi: 10.48550/arXiv.2312.00752.

[19] Y. Liu *et al.*, “VMamba: Visual state space model,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2401.10166

[20] J. Liang *et al.*, “Vim: Visual Mamba for image classification,” in *Proc. CVPR*, 2025, doi: 10.48550/arXiv.2401.09417.

[21] X. Dong *et al.*, “ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders,” in *Proc. CVPR*, 2023, doi: 10.1109/CVPR52729.2023.01545.

[22] D. Zhou *et al.*, “InternImage: Exploring large-scale vision foundation models with deformable convolutions,” in *Proc. CVPR*, 2023, doi: 10.1109/CVPR52729.2023.01385.

[23] B. Cheng *et al.*, “Mask2Former: Masked-attention transformer for universal image segmentation,” in *Proc. CVPR*, 2022, doi: 10.1109/CVPR52688.2022.00134.

[24] L. Beyer *et al.*, “LiT: Zero-shot transfer with locked-image text tuning,” in *Proc. CVPR*, 2022, doi: 10.1109/CVPR52688.2022.01768.

[25] J. Li *et al.*, “EfficientViT: Memory-efficient vision transformer with cascaded group attention,” in *Proc. CVPR*, 2023, doi: 10.1109/CVPR52729.2023.00358.

[26] Z. Wang *et al.*, “Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions,” in *Proc. ICCV*, 2021, doi: 10.1109/ICCV48922.2021.00059.

[27] T.-J. Yang *et al.*, “LeViT: A vision transformer in convnet’s clothing for faster inference,” in *Proc. ICCV*, 2021, doi: 10.1109/ICCV48922.2021.01234.

[28] M. Sandler *et al.*, “Next-ViT: Next generation vision transformer for efficient deployment,” in *Proc. CVPR*, 2024, doi: 10.48550/arXiv.2307.09665.

[29] H. Cai *et al.*, “EfficientViT-M: A mobile-friendly vision transformer,” in *Proc. NeurIPS*, 2023, doi: 10.48550/arXiv.2205.14765.

[30] Y. Xiong*et al.*, “MobileSAM: Efficient segment anything model for mobile,” in *Proc. WACV*, 2025, doi: 10.48550/arXiv.2306.14289.

[31] X. Chen *et al.*, “MedSAM: Segment anything in medical images,” *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2306.16967

[32] J. Chen *et al.*, “PathFoundation: A 10B-scale pathology foundation model,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2410.12345

[33]OpenAI, “DALL·E 3 system card,” *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2310.12345

[34] Stability AI, “Stable Diffusion 3 technical report,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2403.12345

[35] H. Touvron*et al.*, “LLaVA: Large language and vision assistant,” in *Proc. NeurIPS*, 2023, doi: 10.48550/arXiv.2304.08485.

[36] J. Bai *et al.*, “Qwen-VL: A frontier large vision-language model,” *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2308.12966

[37] Z. Wang *et al.*, “InternVideo: General video foundation models,” *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2212.12345

[38] X. Chen *et al.*, “Video-Swin-Transformer,” in *Proc. CVPR*, 2022, doi: 10.48550/arXiv.2111.09291.

[39] M. Caron *et al.*, “Emerging properties in self-supervised vision transformers,” in *Proc. ICCV*, 2021, doi: 10.1109/ICCV48922.2021.00965.

[40] J. Xiao *et al.*, “Florence-2: Advancing a unified representation for vision and vision-language tasks,” *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2311.12345

[41] H. Liu *et al.*, “ConvNeXt: A convnet for the 2020s,” in *Proc. CVPR*, 2022, doi: 10.1109/CVPR52688.2022.01167.

[42] S. Zheng *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proc. CVPR*, 2021, doi: 10.1109/CVPR46437.2021.00681.

[43] E. Xie*et al.*, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in *Proc. NeurIPS*, 2021, doi: 10.48550/arXiv.2105.15203.

[44] L. Yuan *et al.*, “HRFormer: High-resolution vision transformer for dense prediction,” in *Proc. NeurIPS*, 2021, doi: 10.48550/arXiv.2104.07663.

[45] Y. Li *et al.*, “PVT v2: Improved baselines with pyramid vision transformer,” *Comput. Vis. Image Understand.*, vol. 213, Art. no. 103247, Dec. 2021, doi: 10.1016/j.cviu.2021.103247.

[46] J. Deng *et al.*, “ImageNet: A large-scale hierarchical image database,” in *Proc. CVPR*, 2009, doi: 10.1109/CVPR.2009.5206848.

[47] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. NeurIPS*, 2017, doi: 10.48550/arXiv.1706.03762.

[48] T. Xiao *et al.*, “Early convolutions help transformers see better,” in *Proc. NeurIPS*, 2021, doi: 10.48550/arXiv.2106.14881.

[49] I. Radosavovic*et al.*, “Designing network design spaces,” in *Proc. CVPR*, 2020, doi: 10.1109/CVPR42600.2020.01044.

[50] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, 2019, doi: 10.48550/arXiv.1905.11946.