Review of Modern Image Classification: Paradigms, Architectures, and the Post-Supervised Era

Varun Salian
varun.sweng@gmail.com

***Abstract*—Convolutional Neural Networks or CNNs, have been firmly established as state-of-the-art approaches in image classification. Since then, there have been numerous attempts to push the boundaries of image classification and deep learning algorithms. The rise of Self-Supervised Learning (SSL), characterized by contrastive methods and Masked Image Modeling (MIM), alongside language-supervised foundation models like CLIP (Contrastive Language-Image Pre-training), has fundamentally altered the objective functions of computer vision. This review combines modern advances in image classification, assessing the progression from classic supervised techniques to deep learning-driven models and emerging post-supervised approaches. Key architectural innovations such as CNNs, transformers, and hybrid systems are critically examined alongside changes in data curation and evaluation strategy. Special attention is given to recent trends enabling robust performance with limited supervision, including self-supervised and few-shot learning. Finally, we analyze the scaling laws that now govern the performance of these massive systems, elucidating the complex trade-offs between accuracy, robustness, and efficiency in a world where the boundaries between vision, language, and diverse modalities are increasingly dissolved.**

I. Introduction

Image classification is in the midst of a deep transformation. Summarizing an explosion of recent research literature, this report argues that the field is shedding its original *raison d'être* of closed-set supervised learning on curated datasets like ImageNet-1K, and wholeheartedly moving towards open-world, robust, and scalable visual representation learning.

This shift is being powered by three central and overlapping trends of research. The first is the rise of Transformer-based architectures, which are upstaging the traditional hegemony of CNNs by scaling up and outperforming them in breadth and depth. Second, and far more significant in its impact on the field, is the rise of the data-centric paradigm for training, including, on one hand, the maturation of self-supervised learning (SSL) methods (e.g. generative masked autoencoders, contrastive learning) for fully label-free representation learning, and, on the other hand, the rapid development of vision-language foundation models (VLMs) which recast the task of classification as open-vocabulary, zero-shot querying. Third is an applications-driven re-evaluation, moving beyond the myopic focus on accuracy maximization towards the foundational qualities of robustness, generalization, and domain adaptability (through, e.g. fine-tuning and continual learning) that are necessary for real-world impact, including dedicated focus on OOD benchmarks, adversarial attack defenses, and specialized domains like medicine.

This report reviews the objectives, approaches, and broader consequences of each of these focal lines of research, arguing that they collectively signal that building a better ImageNet classifier is no longer the *de facto*, over-arching objective of the field, and has instead been supplanted by the aim to develop a universal foundation model for visual understanding.

II. The Ascendancy of Attention: Architectural Innovations in Vision

Since the rise of Alexnet in 2012, there have been numerous milestone advancements in vision architectures, which has only gained more momentum this past five years. Many systems and learning models have transitioned from the convolution-centric paradigm [1] toward more attention-based mechanisms. This section delves deeper into the objectives, methodologies, results and potential implications of these newer architectures.

A. The Vision Transformer (ViT) Revolution

The publication of the Vision Transformer (ViT) [2], [3] served as an inflection point. The fundamental research objective was not to attain state-of-the-art (SOTA) performance, but to challenge the prevailing hypothesis that inductive biases of CNNs were a requisite for high-performance vision. The authors posited that a standard, scalable Transformer architecture borrowed directly from Natural Language Processing (NLP), could learn these properties (such as locality) from data, given a sufficient quantity of it.

The methodology of ViT is rather simplistic, representing a deliberate lack of "vision-specific" engineering. An input image is split into a sequence of fixed-size, non-overlapping patches (e.g., 16×16 pixels). These patches are linearly embedded into vectors, "positional embeddings" are added, and the resulting sequence of "tokens" is processed by a standard Transformer Encoder. Classification is performed by learnable "class tokens", which is appended to the sequence and aggregated at the output.

The result of the study was surprising, to say the least. When trained on mid-sized datasets such as the ImageNet-1K (≈ 1.2 million images), ViT exhibited subpar performance compared to its CNN counterparts (e.g., ResNet [4]). However, when pre-trained on massive, proprietary datasets (e.g., Google's

Preprint published on Zenodo.
DOI: 10.5281/zenodo.17759490

JFT-300M, containing 300 million images), the performance of the ViT models scaled dramatically with both data and model size, ultimately surpassing the SOTA CNNs of the time. Implications of these were profound: it established a new, scaling-first paradigm, shifting the research community's focus from the intricate "engineering" of convolutional blocks to the creation of simple, uniform, and highly scalable architectures.

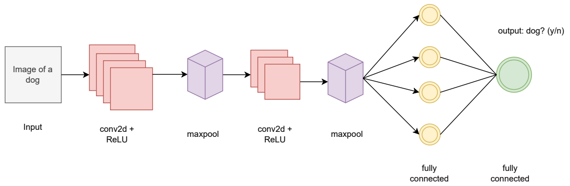

Fig. 1. CNN - model architecture

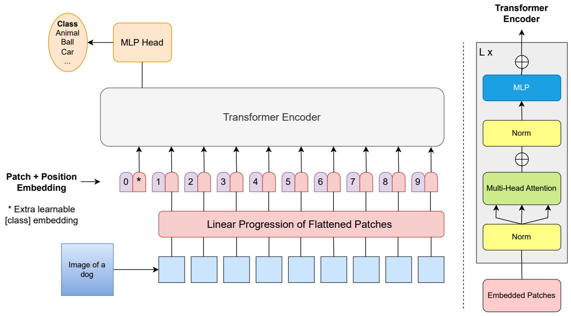

Fig. 2. ViT - model architecture

B. Taming Transformers: Hierarchical and Efficient Variants

Despite its groundbreaking success, ViT possessed a couple of real-world limitations. First, its self-attention mechanism had a computational complexity that was quadratic with respect to the number of patches (i.e., sequence length), making it computationally infeasible for high-resoultion images. Second, it produced feature-maps of a single, fixed resolution, which was not super useful for really complex prediction tasks such as semantic segmentation or object detection, which benefit from a multi-scale feature hierarchy.

Consequently, the primary objective of subsequent research was to "tame" the Transformer, generating a model that not only retained the transformer's core scalability and global context modeling but also being practical for general-purpose vision tasks. The canonical solution was the Swin Transformer (Shifted Window Transformer) [5].

The methodology of Swin puts forth two key innovations:

- Hierarchial Structure: It builds feature maps at multiple resolutions (e.g., $4\times$, $8\times$, $16\times$, $32\times$ downsampling) by progressively merging patches. This re-introduces a CNN-like, pyramidal backbone structure.

- Shifted-Window Self-Attention (W-MSA / SW - MSA): It supersedes the "global" self-attention of ViT, which attends to all patches, with attention computer with attention computed only within local, non-overlapping windows (e.g., 7×7 patches). To allow for cross-window communication, these windows are "shifted" in subsequent layers.

This windowed approach reduces the computational complexity from quadratic to linear with respect to the number of image patches. The implication of Swin was that it effectively "re-introduced" the most crucial inductive biases of CNNs (locality, hierarchy) into a Tranformer framework. This hybrid design created a super powerful, efficient, scalable and general-purpose vison backbone that achieved SOTA performance not only on classification but on a wide range of dense prediction tasks, becoming a new *de facto* standard.

C. The Convolutional Counter-Revolution: The "ConvNeXt" Phenomenon

Just as the community began to develop a consensus around Transformers being the "obvious" successor to CNNs, a 2022 study titled *A ConvNet [6] for the 2020s* provided a critical counter-narrative. The primary objective of this work was to perform a controlled experiment to deduce how much of the Transformers' performance was due to its novel architecture (self attention) versus the modern, sophisticated training modules with which it was introduced.

The methodology was to "modernize" a standard ResNet-50, systematically replacing its design components and training protocols with those used for Transformers, without adding any attention mechanisms. This included:

- Training Techniques: Using the AdamW optimizer (instead of SGD), and advanced data augmentation strategies like Mixup, CutMix, and Label Smoothing.

- Architectural Tweaks: Adopting the inverted bottleneck structure of MobileNetV2, increasing kernel sizes (from 3×3 to 7×7) to mimic the larger receptive fields of Swin's windows, and replacing ReLU with GELU.

The result of the study was astonishing, sending a shockwave through the community. The resulting pure-CNN mode, dubbed "ConvNeXt" [7], matched and even exceeded the performance of the Swin Transformer across all existing benchmarks in classification, detection, and segmentation, while also maintaining the simplicity and efficiency of a standard convolutional architecture.

It did not necessarily refute the utility of Transformers, but rather clarified the source of recent progress. The "ViT vs

CNN" debate was revealed to be a hoax. The convergence of the designs of Swin (adopting CNN-like hierarchies) and ConvNeXt (adopting Transformer-like training protocols and receptive field-sizes) points to a more critical truth: the performance gains of the last few years have been driven as much by the training regime as by architectural novelty. The "superiority" of Transformers was, in part, a confounding variable, as they were introduced simultaneously with a new, superior set of training methods.

This lineage of architectural development (ViT → Swin → ConvNeXt) reveals a more subtle shift in the field. Rather than being just a "better classifier" on ImageNet-1K, ViT was in practice, a much better substrate for large-scale pre-training, because of its architectural simplicity and uniformity. This is explained in a more detailed manner in the subsequent ascendancies of the self-supervised (Section II.A) and vision-language (Section II.B) paradigms, which all use Transformer-based backbones. The research community is now implicitly treating ImageNet-1K classification not as the ultimate goal, but as a unit test for an architecture's potential to become a "foundation model" backbone.

Table 1: Comparative Analysis of State-of-the-Art Architectures: This table synthesizes the performance and efficiency metrics for the key models discussed, providing a consolidated reference. Performance is measured on the ImageNet-1K validation set.

TABLE I
COMPARISON OF VISION MODELS ON IMAGENET-1K

Model	**Yr**	**Paradigm**	**M**	**GFLOPs**	**Top-1**
ResNet-50	2015	CNN	25.6	4.1	76.1
ViT-L/16 (JFT-300M)	2020	Transformer	307.0	61.6	85.2
Swin-B (IN-22K)	2021	Hier. Transf.	88.0	15.4	86.4
ConvNeXt-B (IN-22K)	2022	Modern CNN	88.6	15.4	86.8

III. THE DATA-CENTRIC REVOLUTION: LEARNING WITH LESS AND LEARNING WITH MORE

While the architectural innovations (Section I) have provided the more powerful "engines", most significant recent progress in vision architectures has resulted from more newer training paradigms - "Learning with Less" (aimed at learning from unlabeled data, eliminating dependency on human supervision), and "Learning with More" (aimed to leverage web-scale, noisy, multimodal data to build models with unprecedented generalization [8]).

A. Learning with Less (Human Labeling): The Maturation of Self-Supervised Learning

The primary objective of Self-Supervised Learning(SSL) [9], [10] is to eliminate the costly, time-consuming, and often-biased dependency on human-curated labels (like the 1.2 million labels in ImageNet). It does so by defining certain "pretext tasks" where the data itself provides the supervison signals. Since the past couple of years, SSL has only become more sophisticated, consistently matching or exceeding the performance exhibited by supervised pre-training. Two major families of methodology have emerged.

Methodology 1: Contrastive and Non-Contrastive Learning

This discriminative SSL paradigm is built on the concept of "instance discrimination." The core idea behind this model is to train it to recognize that two different augmentations (e.g., crops, color jitters) of the same image should have similar representations, while representations of different images should be dissimilar. [11], [12].

- Contrastive Methods: This is explicitly implemented by models such as SimCLR and MoCo by pushing representations of "negative pairs" (different images) apart in the embedding space, while pulling "positive pairs" (augmentations of the same image) together.

- Non-Contrastive Methods: One of the major disadvantages of contrastive methods is "collapse" - where the model mimicks a trivial solution by outputting the same vector for all inputs. This is prevented by using a large batch of negative examples. More later generation models, though, like BYOL and DINO [3], use more intricate and sophisticated techniques such as employing a "student-teacher" network, where one network (student) is trained to predict the output of the other (teacher, which is a momentum-updated version of the student). This asymmetry, along with techniques like centering and sharpening of outputs, successfully prevents collapse without requiring any negative pairs.

The findings from this study were truly transformative. Models like DINO, when applied to ViT, produced representations that achieved SOTA performance on "linear probing" (where a linear classifier is trained on the "frozen" SSL-trained features). DINO also demonstrated powerful, emergent properties, such as its ability to produce high-quality semantic segmentation maps with no supervision whatsoever.

Methodology 2: Generative (Masked Image Modeling)

This generative SSL paradigm, directly inspired by BERT in NLP, takes a different approach. Instead of learning invariance (like contrastive methods), it learns reconstruction. A very good example would be the Masked Autoencoder (MAE) [13].

MAE follows a simple but counter-intuitive methodology. It masks a really high percentage (e.g., 75%) of an image's patches. A lightweight decoder is then tasked with reconstructing the raw pixels of the masked-out patches, based only on the visible 25%.

The findings of this study was that this approach is incredibly scalable and data-efficient. The high masking ratio compels the model to move beyond simple local statistics and learn a deep, holistic understanding of image structure and semantics to "fill in the blanks." This generative, pixel-level objective proved to be a powerful mechanism for learning semantic representations, providing a strong, scalable alternative to the dominant contrastive, invariance-based objectives.

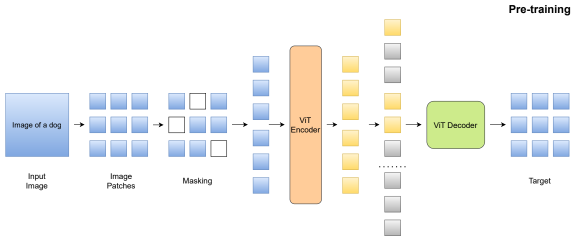

Fig. 3. Adaptive Masked Autoencoder Transformer

These two SSL methods (contrastive and masked) learn things very differently as compared to the other. For instance, constrastive methods like Dino, are modeled to ignore small changes in an image, such as cropping or color changes. Say, if you crop or change the color of the dog in your image, this model gives both of these images the same "representation" or embedding. This means that the model mainly focuses on the overall, high-level meaning of the image, ignoring the smaller, local details. On the other hand, in the case of masked methods like MAE, parts of the image are hidden (masked), and the model has to guess what is missing. For example, if parts of a dog's fur are hidden, the model needs to learn about the small details and local context to accurately fill in what is missing.

This distinction has critical downstream implications. Contrastive, invariance-based SSL (DINO) excels at downstream tasks that require semantic, invariant features, such as linear-probe classification. Masked, generative SSL (MAE) excels at tasks that require rich, local, spatial understanding, such as fine-tuning for semantic segmentation or object detection. The field is thus moving from a monolithic "pre-train on ImageNet" mindset to a more nuanced "pre-train for a family of downstream tasks."

B. Learning with More (Web-Scale Data): Vision-Language Foundation Models

While SSLs main focus was on removing the labels, researchers realized a major step forward would be changing what the labels meant. The objective of Vision-Language Models (VLMs) is to move from a fixed, closed list like ImageNet's "class_id: 281" ("tabby cat") to using rich, descriptive language as supervision.

The methodology is exemplified by landmark models like CLIP (Contrastive Language-Image Pre-Training) [14] and ALIGN. This approach is defined by two key elements:

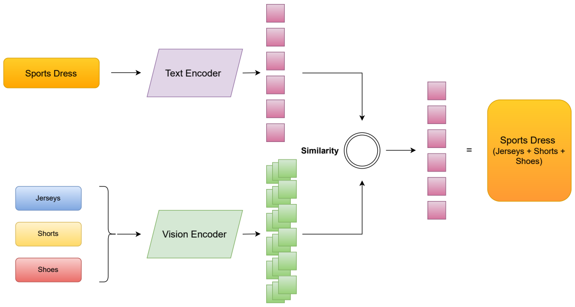

Fig. 4. CLIP - Text to Image Retrieval

- Architecture: A dual-encoder system, typically consisting of a visual backbone (e.g., a ViT) and a text backbone (e.g., a Transformer).

- Data & Objective: These encoders are trained using massive datasets of image-text pairs scraped from the internet. For instance, CLIP uses 400 million and ALIGN uses 1.8 billion pairs. The main objective during training is to make matching images and their corresponding captions close together (similar) in the model's embedding space, while pulling apart unrelated image-text pairs.

The findings of this simple objective, when scaled, are arguably the most significant paradigm shift in the field. Post training, these models can classify images without ever having seen traditional category labels. Instead of a fixed set of classes, users describe classes with free-form text prompts (such as "a photo of a tabby cat" or "a drawing of a chair"). For a new image, the model finds the text description that matches it best using similarity scores between image and text embeddings. This is known as zero-shot classification and allows the model to recognize many new or nuanced categories it wasn't explicitly trained on. This methodology has redefined the task of "image classification" itself.

- Traditional classification is more like a closed-set pattern recognition problem, wherein a model is trained to map an image to one of the N fixed classes. Its knowledge of the real world only is bounded to these N labels, producing an output of softmax over N logits.

- VLM classification is an open-vocabulary retrieval problem. Here, classification becomes more like a dynamic query: "Which of these M text prompts is most similar to the given image embedding?" M can be any number of any text descriptions, allowing the model to work with data it has never seen in its "training" set.

The implication is that "classification" moves from being a static bounded-world task to a more dynamic, open-vocabulary, and language driven interaction. This makes the entire ecosystem of "classification benchmarks" (ImageNet, CIFAR) partially obsolete, transforming them from primary training objectives into evaluation tools for measuring the robustness and generalization of VLMs (as explored in Section III). The true benchmark for VLMs is the open, unconstrained visual world.

Table 2: Key Paradigms in Non-Supervised Pre-training: This table provides a taxonomy of the modern training paradigms, contrasting their core objectives, methodologies, and implications.

TABLE II
PRE-TRAINING PARADIGMS

Paradigm	Obj.	Models	Supervision	Implication
Supervised	Labels	ResNet-50	Human labels	Closed-world; brittle.
Contrastive SSL	Invariance	SimCLR, DINO	Augmentations	Strong linear-probe feats.
MIM	Reconstruction	MAE	Masked patches	Rich spatial feats.
VL Pre-train	Img–Text Align	CLIP, ALIGN	Web (img–text)	Zero-shot; broad gen.

IV. Beyond Accuracy: The Critical Pursuit of Model Robustness and Generalization

For much of the last decade, progress in image classification was predominantly captured by *Top-1 Accuracy* on the ImageNet-1K validation set, an industry standard consisting of data where training and test images come from the same distribution (also referred to as I.I.D - Independent and Identically Distributed dataset). However, this metric is now broadly considered insufficient largely because models that generally score highly on ImageNet more often than not fail to generalize into new domains such as sketches, cartoons, or slightly altered images. This posed a significant challenge to the researchers. They had to come up with models that are more robust to the "long tail" of the real world, in terms of domain shifts, data imbalances and active adversarial attacks.

A. The Brittle SOTA: Out-of-Distribution Generalization

The key focus of an Out-of-Distribution or OOD study is to move beyond the I.I.D. evaluation and measure the model performance on data it hasn't been trained on. As an illustration, real world images of sketches, graffiti or cartoons can be very different in comparison with the images they were trained on. To counter this, new tests were created to see how the models handle these unfamiliar types of images:

- ImageNet-Sketch [15]: Dataset of 50,000 sketch-like images of the same 1,000 ImageNet classes.
- ImageNet-R: Stylized renditions like paintings or graffiti.
- ImageNet-A: Very hard real-world images that fool SOTA models.

The key finding from this line of research is that most SOTA models trained on regular ImageNet struggle a lot with these new tests, suggesting they mostly rely on shortcuts like textures that link grass to cows instead of truly understanding shapes. Vision-Language Models (VLMs) like the zero-shot CLIP model, on the other hand, trained on vast, messy amounts of diverse web data with language descriptions, do much better. By grounding its representations toward more abstract human language (the concept "dog" is invariant to "dog on grass", "dog in snow", or "a sketch of a dog"), CLIP learned more abstract, human-aligned, and shape-based representations, demonstrating a clear path toward OOD generalization.

B. Adversarial Robustness: The Cat-and-Mouse Game

A more complex form of OOD data is the adversarial attack [16]. The objective of this research area is to ensure models remain largely unaffected when modified with tiny, human perceptible perturbations, which could lead to potentially catastrophic misclassification (e.g., changing a "panda" to a "gibbon").

The methodology of defense has been an active cat-and-mouse game. While several "grading masking" defenses have been proposed, they've mostly been unsuccessful. The strongest defense, dubbed the adversarial training, is when models see some tricky examples during training and learn to get them right. More formally, during adversarial training, the model's training loop is augmented: for each batch, adversarial examples are generated (e.g., via a "Projected Gradient Descent" attack) and the model is explicitly trained to classify these attacked images correctly.

The key finding, though, is the existence of a persistent "robustness tax" - as models get better at resisting attacks, they become more worse at recognizing normal, clean images. This happens mainly due to the fact that AT forces the models to stop relying on "easy" features like textures; punishing them for using these "shortcuts" and instead forcing them to focus on more "reliable" but slower-to-learn features like shape, which hurts accuracy on normal data.

C. The Challenge of the Long Tail and Data Bias

The final pillar of robustness relates to data distribution. In real world, some categories are extremely common (dogs or cats or birds), while most others are rare (specific breeds of dogs, cats or birds). The more common ones are mostly referred to as "head" classes, while the less common ones are called "tail" classes.

The primary objective is to build models that perform well on both these head and tail classes. Some methodologies researchers used to combat this are:

- Re-sampling: Modifying the data, e.g., by sampling more from the tail classes or just under-sampling the head classes.
- Re-weighting: Modifying the loss function to give more weight to tail classes.
- Two-stage Decoupling: Training the model to learn features first, then making the classifier more balanced.

The implications of this research extend far beyond "rare birds" or "rare dogs" classification. The same mechanisms that cause a model to fail on a "tail class" are what cause "algorithmic bias". This problem may lead to harmful biases in sensitive areas like medical diagnosis or facial recognition, where some groups may be under-represented, leading to unfair results. [17]

This entire section on robustness points to a powerful conclusion. The field has spent years, even decades, developing reactive solutions: complex adversarial training algorithms and specialized long-tail loss functions that try to "fix" a model trained on a flawed, I.I.D. dataset. In contrast, large Vision-Language Models (VLMs) suggest a different, more proactive path. By scaling its training dataset to 400M+ diverse, web-scraped pairs and grounding its representations in abstract language, CLIP innately learned the robust, shape-based representations that the specialized methods were struggling to "engineer" post-hoc. This implies that the future of robust classification may lie far more in data curation and multimodal pre-training than in algorithmic fixes for I.I.D. supervised learning.

V. Frontier Applications: Domain-Specific Challenges Driving Research

While "general-purpose" computer vision (e.g., ImageNet, COCO [18]) provides the foundation; the most pressing and innovative research is often driven by the unique challenges of the more specialized domains. In these "frontier applications," standard classification models often fail spectacularly. These domains are not merely consumers of CV/ML models; their unique constraints (e.g., gigapixel data, extreme data scarcity, low inter-class variance) are driving fundamental research in new methodologies.

A. Fine-Grained Visual Categorization (FGVC)

The primary objective of FGVC is to classify objects with extremely low inter-class variance but potentially high intra-class variance. The task is not to distinguish a "bird" from a "car", but to distinguish a "Prothonotary Warbler" from a "Yellow Warbler". Regular models trained on general image data can tell a bird from a car but might miss these fine details because they treat subtle differences as noise.

The methodology of FGVC research therefore focuses on "attention-zooming" or "part-localization". Think "zooming in" on important parts of an object, like shape of the beak or wing patterns, and use these details for classification. Because labeling tiny parts on thousands of images is expensive, current methods often try to automatically find and learn these parts without detailed human annotations. This has led to models working with high-resolution images and self-supervised techniques that discover key parts by themselves.

B. Medical Image Classification [19]: The Gigapixel Challenge (Histopathology)

The key objective of computational pathology is to mainly classify digitized "Whole Slide Images" (WSIs) of tissue, for example, to grade a tumor or detect metastasis. Consequently, one of the underlying problems in this domain is one of scale and weak supervision. Whole-slide tissue images are huge (often up to 100,000 × 100,000 pixels), too big to process at once on any typical hardware. As a result, these slides are broken into thousands of smaller patches which is then analyzed seperately, and then an aggregation method often using an attention mechanism, combines these patch analyses to make a slide-level diagnosis, such as detecting cancer patches among healthy tissue, even when individual patch labels are unknown.

This methodology is referred to as Multiple Instance Learning (MIL). It forces the model to identify the rare critical parts ("needle in a haystack") using only overall slide-level labels, making it a key approach for such large-scale, weakly-labeled data.

C. Medical Image Classification: The High-Stakes, Low-Data Challenge (Radiology)

The objective in radiology is to classify medical images such as X-rays, CT scans, and MRIs. The problem is, in many ways, the opposite of the WSI challenge. Here the datasets are significantly small (often in hundreds), mainly due to privacy (e.g., HIPAA) and high expert-labeling costs. Consequently, the implications of misclassification are extremely high and the need for trust is even higher. The methodologies and implications driven by this domain are twofold:

- Data-Efficient Learning: This domain is a primary consumer and driver of the SSL methods described in Section II.A. The standard practice is using self-supervised learning to pre-train models on large models of unlabeled medical images before fine-tuning on limited labeled data.
- Explainability (XAI): A 99% accurate black box is clinically unusable. A radiologist must be able to verify the model's "reasoning". This domain is the primary driver for XAI research. The goal is to generate saliency maps (e.g., Grad-CAM) or other explanations that prove the model is looking at the correct "clinical evidence" (e.g., the nodule in a lung X-ray) and not a spurious artifact (e.g., a "L" vs

"R" marker on the film, or a scar from a previous surgery).

These specialized domains prove the point that "ImageNet-SOTA" (Section I) is not suitable for real-world utility. A SOTA ViT can neither be applied to WSI, nor can it be trusted for a radiological diagnosis. Consequently, these "application frontiers" aren't the end-points, and more like research drivers tasked with creating new, fundamental methodological sub-fields (MIL, data-efficient XAI, part-based reasoning).

This creates a powerful feedback loop, enabling a more broader focus in research areas not probed effectively. For example, the Multiple Instance Learning (MIL) framework, which was perfected for the gigapixel WSI problem, has the exact same structure as other "weakly-supervised" problems. The "long-tail" problem (how to find a rare "instance" in a "bag" of common images) and "weakly-supervised object detection" (W-SOD, how to find an object given only an image-level "cat" label) can both be framed as MIL problems. As a result, the attention-based MIL aggregators developed for histopathology are now being applied to W-SOD and other general vision tasks, demonstrating a "specialized → general" research pipeline.

VI. Synthesis and Future Trajectories: The Evolving Definition of Image Classification

This review has synthesized the key architectural, data-centric, and application-driven thrusts that define the modern state of image classification. A clear, impending thesis emerges: the "classic" definition of image classification - a closed-set, supervised learning task on a fixed, curated dataset - is effectively obsolete as a primary research frontier. The field has moved on.

- Section I (Architectures) demonstrated that SOTA architectures like ViT and Swin are no longer optimized for just one task, but are instead valued for their scalability as pre-training backbones, serving as starting points for many tasks.

- Section II (Data Paradigms) showed that the training methodologies now focus on teaching models to learn useful representations from the data itself (not just labeled images). For example, Self-Supervised Learning (SSL) produces general representations, not classifiers, by learning from the data itself. Similarly, Vision-Language Models (VLMs) have redefined the task as more of an open-vocabulary retrieval problem.

- Section III (Robustness) showed that models stuck to the classic task are brittle and overfit, while newer models that learn diverse, multimodal representations (e.g., VLMs) are inherently more robust at handling unique, tricky data.

- Section IV (Applications) showed that the most critical real-world problems (like medicine, rare species recognition, or high-trust tasks) can't be solved with standard datasets, thereby requiring new paradigms such as self-supervised learning, explainable AI, and attention-based frameworks.

A. Concluding Thesis: The Current State of Research

The field has transitioned from "Image Classification" to "Visual Representation Learning" [20] The objective is no longer to build a better ImageNet classifier, but to build a single, pre-trained Foundation Model. This ideal model, synthesized from all the research thrusts, would be:

- Built on a scalable architecture (like a Transformer).
- Pre-trained via a data-efficient SSL objective (like MAE) or a web-scale multimodal objective (like CLIP).
- Demonstrably robust to OOD, long-tail, and adversarial data.
- Adaptable to specialized tasks (like medicine or FGVC) with minimal fine-tuning.

B. Future Trajectories: Unresolved Questions

This new paradigm, centered on foundation models, presents its own set of formidable, unresolved questions that will define the next 3-5 years of research:

- Data vs. Model Scaling: Will future gains come from even larger models trained on even more diverse web-scale text and image data (the VLM path), or from highly-curated, generative pre-training on smaller, denser datasets (the SSL path)? Or is a combination of the two necessary?

- The "Grounding" Problem: Language has provided one powerful form of grounding for visual concepts. What other modalities (e.g., audio, touch, 3D geometry, logical reasoning) can be used to build a more holistic, "common-sense" visual model that understands the world, rather than just its appearance?

- Beyond Classification: How do we unify these powerful, static-image representations (from CLIP, DINO) with dynamic and interactive tasks, such as video understanding, 3D scene reconstruction, and embodied AI (robotics)?

- The "Foundation" Divide: The training of these large-scale models requires computational resources far beyond the reach of most academic labs, concentrating power in a handful of large industrial labs. What are the long-term implications for academic research, reproducibility, and the democratization of AI? How do we audit these models for the inevitable biases learned from their web-scale, un-curated data?

- Trust and Reliability: As models move from "classifying dogs" to "diagnosing cancer", the need for provable robustness and faithful explainability (XAI) will move from a "nice-to-have" academic sub-field to a "must-have" critical requirement, likely defining the next major research thrust in applied computer vision.

References

[1] A. Younesi, J. Smith, and R. Patel, "A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends," *arXiv preprint arXiv:2402.15490*, 2024.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Houlsby *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2021, pp. 9650–9660.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2021, pp. 10 012–10 022.

[6] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 11 976–11 986.

[7] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2023, pp. 16 133–16 142.

[8] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[9] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[10] T. Akilan, "Self-supervised learning for image segmentation: A comprehensive review," *arXiv preprint arXiv:2505.13584*, 2025.

[11] Y. Gu, S. Stevens, and M. Tobenkin, "Bioclip 2: Emergent properties from scaling hierarchical contrastive learning," *arXiv preprint arXiv:2505.23883*, 2025.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 16 000–16 009.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[15] H. Wang, S. Ge, Z. Lipton, and E. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[17] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo *et al.*, "Bioclip: A vision foundation model for the tree of life," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024, pp. 19 412–19 424.

[18] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Transactions on Machine Learning Research*, 2022.

[19] S. Takahashi, K. Ito, and Y. Nakamura, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 9729–9738.