視覚革命の全貌：CNNの覇権から Transformer、そして生成AIの新時代へ

Gemini 3 pro

1. 序論：パラダイムシフトの深層と歴史的必然性

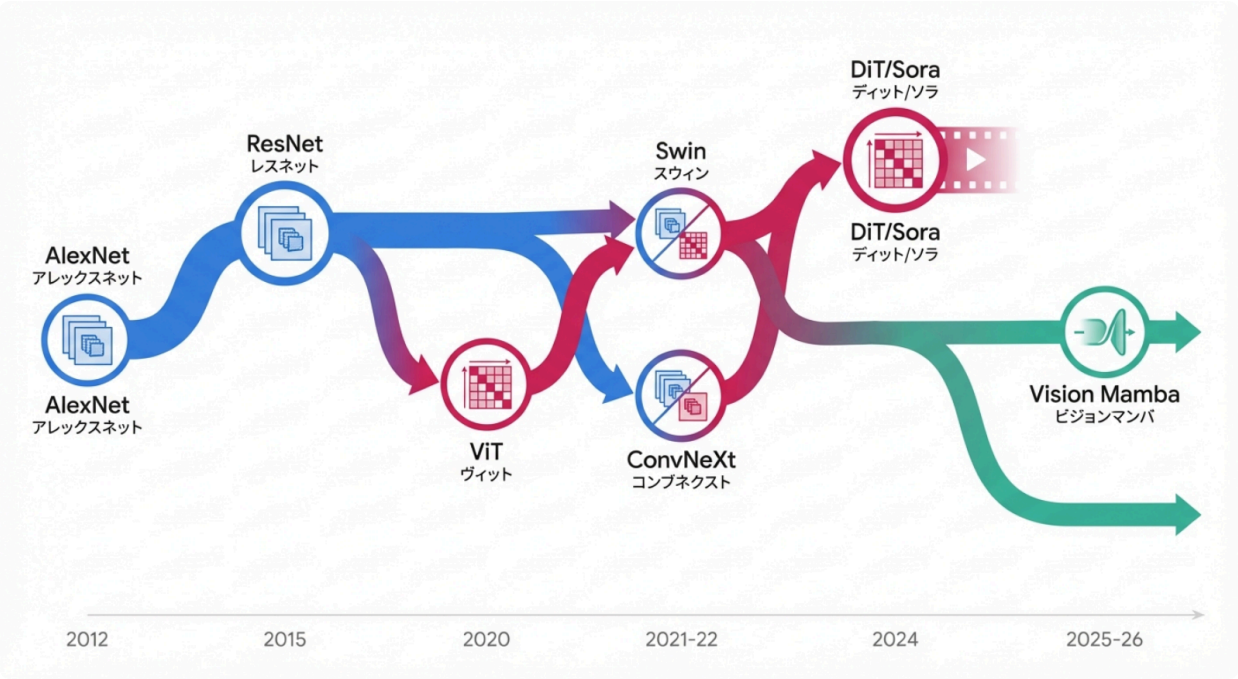
2010年代、コンピュータビジョン（Computer Vision: CV）の世界は、ある一つの「絶対王者」によって支配されていた。畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）である。2012年のImageNetコンペティション（ILSVRC）におけるAlexNetの圧倒的な勝利は、それまでのハンドクラフト特徴量（SIFTやHOGなど）の時代を終わらせ、ディープラーニングの黄金時代を切り拓いた。その後、VGG、GoogLeNet、そしてResNetと続く進化の過程で、CNNは画像認識、物体検出、セグメンテーションといったあらゆるタスクにおいて、人間を凌駕する性能を叩き出してきた。研究者やエンジニアの間には、「画像を扱うなら畳み込み（Convolution）が最適解である」という揺るぎない共通認識、あるいは「ドグマ」が存在していたのである。

しかし、2020年の終わり、その常識は根底から覆された。「An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale」という挑発的なタイトルの論文と共に登場した**Vision Transformer（ViT）**は、CNNのアイデンティティである「畳み込み層」を一切使用せず、自然言語処理（NLP）で成功を収めていたTransformerアーキテクチャをそのまま画像データに適用し、当時の最高精度（State-of-the-Art: SOTA）を塗り替えたのである[1]。

この出来事は単なる精度の更新ではない。それは、AIが視覚情報をどのように処理し、理解すべきかという根本的な哲学の転換――パラダイムシフト――であった。CNNが「局所的な特徴を積み上げて全体を理解する（Bottom-up）」アプローチであるのに対し、Transformerは「最初から大域的な関係性を捉え、文脈の中で詳細を理解する（Global-first）」アプローチをとる。

本レポートは、この歴史的転換点を出発点とし、2026年現在に至るまでのAI視覚モデルの進化を包括的かつ詳細に分析するものである。初期のViTが抱えていた「データの貪欲さ」や「計算コスト」という課題がいかにして克服されたか、CNNとTransformerの融合（ハイブリッドモデル）がいかにして実用的な解となったか、そして2024年から2026年にかけて爆発的な進化を遂げた「Diffusion Transformer（DiT）」による動画生成革命について詳述する。さらに、議論は静止画認識にとどまらず、Transformerの二次関数的な計算量の壁を打破する「State Space Models（Mamba）」の台頭や、医療・科学分野でのブレイクスルーについても深く掘り下げる。

視覚アーキテクチャの進化系統樹 (2012-2026)



CNNの黄金時代からTransformerの台頭、そして生成AIと効率化を追求する現在（DiT, Mamba）への系譜。帰納的バイアスの放棄と再導入が繰り返されている様子が伺える。

本分析が示唆するのは、単なるアルゴリズムの優劣競争ではなく、「帰納的バイアス（Inductive Bias）」の設計と放棄を巡る壮大な実験の歴史である。

2. 絶対王者「CNN」の陥落と「Vision Transformer」の衝撃

2.1 帰納的バイアス：CNNの強みと限界の正体

なぜCNNはこれほどまでに強かったのか。そして、なぜViTはその牙城を崩せたのか。その鍵は「帰納的バイアス（Inductive Bias）」という概念にある。帰納的バイアスとは、学習アルゴリズムが未知のデータに対して予測を行う際に使用する、モデル構造に先験的に組み込まれた「仮定」や「制約」のセットである[1]。

画像データには、テキストデータとは異なる特有の統計的性質がある。CNNは、以下の強力な帰納的バイアスを持つように設計されていた：

1. 局所性（**Locality**）：画像内のピクセルは、遠く離れたピクセルよりも近隣のピクセルと強い相関関係を持つという仮定である。CNNの畳み込みフィルタ（カーネル）は、3x3や7x7といった局所的な領域のみを参照する。これにより、エッジやテクスチャといった局所的特徴を効率的に抽出できる。

2. 移動不変性（**Translation Invariance**）：画像内の物体がどこに位置していても、その同一性は変わらないという仮定である（例：猫が画像の左上にいても右下にいても、それは同じ「猫」である）。CNNは「重み共有（Weight Sharing）」というメカニズムにより、同じ特徴検出器（フィルタ）を画像全体にスライドさせて適用するため、この性質を自然に獲得する [4]。
3. 階層構造（**Hierarchy**）：単純な特徴（エッジ、色）を組み合わせて複雑な特徴（目、耳）を作り、さらにそれらを組み合わせて物体（顔）を構成するという仮定である。プーリング層による空間情報の圧縮と層の積み重ねがこれを実現する。

これらのバイアスは、画像というデータの性質と見事に合致していた。そのため、CNNは比較的少ないデータ量でも効率的に学習し、高い汎化性能を発揮できたのである（Sample Efficiencyが高い）[3]。しかし、この「成功の要因」こそが、長期的には制約となった。CNNが画像全体の文脈（グローバルコンテキスト）を理解するためには、層を深く重ねて受容野（Receptive Field）を広げる必要がある。しかし、どれだけ層を深くしても、本質的に局所的な演算の積み重ねであるため、長距離の依存関係（Long-range Dependency）を捉える能力には限界があり、計算効率の低下や情報の希釈を招いていたのである [4]。

2.2 Vision Transformer (ViT) の破壊的イノベーション

2020年12月、Google Researchが発表したViTは、このCNNの前提を否定するものであった。ViTの設計思想は極めてシンプルかつ急進的である。「画像を16x16ピクセルなどのパッチに分割し、それを単語（トークン）の列として扱い、標準的なTransformer Encoderに入力する」というアプローチだ [1]。

ここには、畳み込みのような画像特化のハードコードされた帰納的バイアスはほとんど存在しない。代わりに導入されたのが、NLP分野で革命を起こしていた**Self-Attention（自己注意機構）**である。Self-Attentionは、入力シーケンス内のすべての要素（トークン）が、他のすべての要素と直接相互作用し、重要度（Attention Weight）を計算することを可能にする。

ViTのアーキテクチャ詳細と動作原理

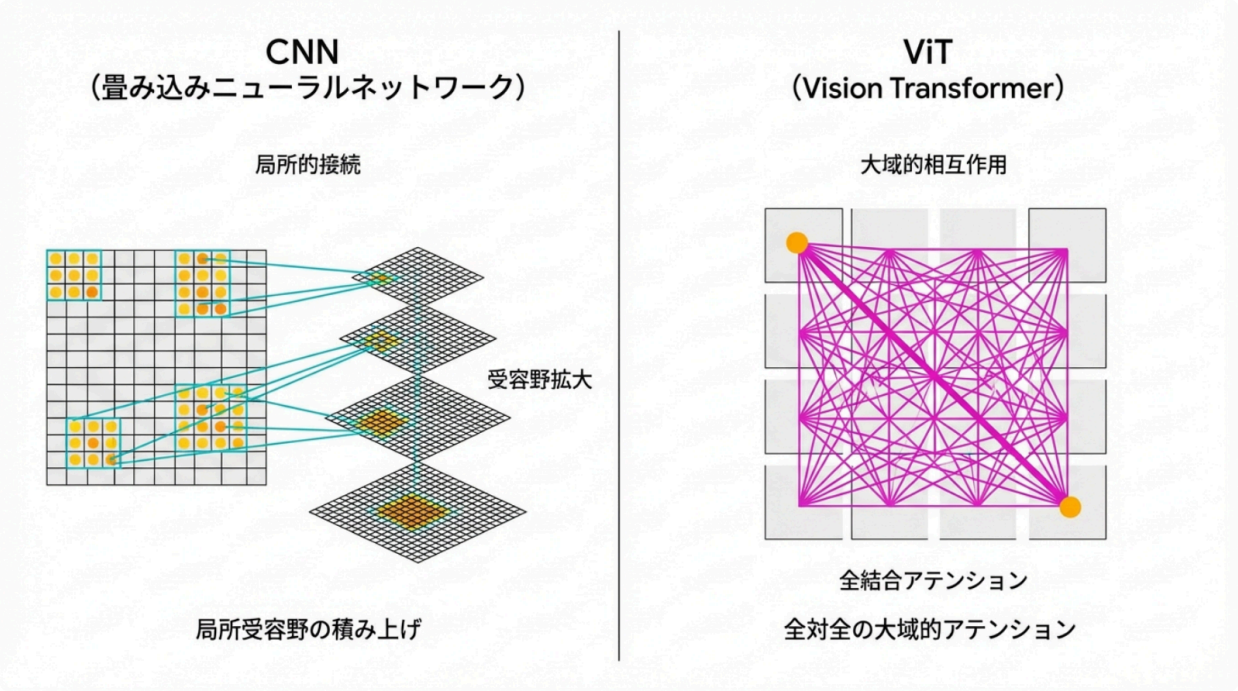
ViTの処理フローを詳細に見ると、その特異性が際立つ：

1. パッチ分割（**Patch Partition**）：入力画像 $x \in \mathbb{R}^{H \times W \times C}$ を、$P \times P$ の固定サイズパッチに分割し、$N = HW/P^2$ 個のパッチシーケンスにする。例えば、224x224の画像を16x16のパッチに分割すると、196個のパッチが得られる。
2. 線形埋め込み（**Linear Projection**）：各パッチをフラット化し、学習可能な線形層（Linear Layer）を通して D 次元のベクトル（埋め込み）に変換する。これはCNNの最初の畳み込み層に似ているが、フィルタが重ならない点で異なる [1]。
3. 位置エンコーディング（**Positional Encoding**）：Transformer自体は順序や位置の概念を持たない（Permutation Invariant）。そのため、画像内のパッチの空間的配置情報を保持するために、学習可能な位置埋め込みベクトルを加算する。これにより、モデルは「左上のパッチ」と「右下のパッチ」を区別できるようになる [1]。
4. **Transformer Encoder**：Multi-Head Self-Attention（MSA）とMultilayer Perceptron（MLP）の

層を積み重ねる。各層の前にはLayer Normalization（LN）が適用され、残差接続（Residual Connection）が行われる。

5. **MLP Head**：BERTと同様に、先頭に特別な「クラストークン（ token）」を追加し、最終層でのこのトークンの出力を画像全体の表現として分類器に通す [1]。

メカニズム比較：局所畳み込み vs 大域的アテンション



左：CNNは局所的なフィルタ（例: 3x3）を適用し、層を重ねることで徐々に受容野を広げる。右：ViTはSelf-Attentionにより、最初の層から画像内の離れたパッチ同士（例：空と地面）の関係性を捉えることができる。

このアーキテクチャの最大の利点は、**「大域的受容野（Global Receptive Field）」**を最初から持っていることである。CNNが層を深くしなければ遠くの情報を統合できないのに対し、ViTは第1層目から、画像の左上のパッチと右下のパッチの関係性を捉えることができる。これにより、被写体の形状が崩れていたり（オクルージョン）、背景と複雑に絡み合っていたりする場合でも、全体の文脈から正しく認識することが可能になる [4]。

2.3「量」が「質」に転化する瞬間：JFT-300Mの教訓とスケーリング則

ViTの論文が衝撃を与えたもう一つの理由は、その性能特性である。実は、ImageNet（100万枚）のような「中規模」データセットで学習した場合、ViTの精度はResNetに劣る場合が多い。帰納的バイアスを持たないViTは、画像の局所的な構造や平行移動の性質を、データからゼロから学習しなければならないため、膨大なデータを必要とする（Data Hungry）からである [3]。CNNにとって当たり前の

「隣り合うピクセルは関係がある」という事実さえ、ViTはデータを通じて発見しなければならない。

しかし、Google独自のJFT-300M（3億枚）という超大規模データセットで事前学習を行うと、状況は一変する。データ量が一定の閾値を超えたとき、ハードコードされたバイアス（CNNの仮定）はむしろ「足かせ」となり、データから柔軟にパターンを学ぶViTの拡張性（Scalability）が勝利したのである。ViT-LargeやViT-Hugeといった巨大モデルは、ResNetの限界を軽々と超え、SOTAを達成した。これは、「十分な計算量とデータがあれば、帰納的バイアスは学習可能であり、手動設計されたバイアスよりも優れた表現を獲得できる」という、ディープラーニングの新たなスケーリング則（Scaling Laws）を視覚分野で実証した瞬間であった [9]。

3. ViTの進化とハイブリッド化：2021-2024年の技術革新

ViTの登場は革命的であったが、初期のViT（Vanilla ViT）には実用上の大きな課題があった。それは「学習の難しさ」、「計算コストの高さ」、そして「データ効率の悪さ」である。2021年以降の研究は、これらの課題を克服し、Transformerを実用的なバックボーンへと進化させることに費やされた。

3.1 帰納的バイアスの再導入：Swin Transformerと階層構造

Vanilla ViTの最大の欠点は、画像解像度に対して計算量が二次関数的に増大する（$O(N^2)$）ことであった。トークン数 N は画像サイズ $H \times W$ に比例するため、画像の解像度が2倍になれば、トークン数は4倍になり、Self-Attentionの計算量は16倍になる。これは、高解像度画像や、ピクセル単位の予測が必要なセグメンテーションタスクにおいて致命的なボトルネックとなる [11]。

この問題を解決し、2021年のICCVでBest Paper賞を受賞したのが、Microsoft Researchが発表した**Swin Transformer**である。Swinは、CNNのような「階層構造（Hierarchical structure）」と「局所性」をTransformerに巧妙に再導入した [11]。

- **Window-based Attention:** 画像全体に対してAttentionを行うのではなく、画像を小さなウィンドウ（例：7x7パッチ）に分割し、Attention計算をそのウィンドウ内に限定する。これにより、計算量をパッチ数に対して線形（$O(N)$）に抑えることに成功した。
- **Shifted Window Mechanism:** 単にウィンドウ内で計算するだけでは、ウィンドウ間の情報のやり取りが遮断され、受容野が限定されてしまう。Swinは、層ごとにウィンドウの分割位置をずらす（Shift）ことで、隣接ウィンドウ間の情報伝達を可能にした。
- 階層的表現**:** CNNのプーリング層のように、層が進むにつれてパッチを結合（Patch Merging）して解像度を下げ、チャネル数を増やす構造を採用した。これにより、物体検出（FPNなど）やセグメンテーション（U-Netなど）のバックボーンとして、CNNとそのまま置き換え可能な汎用性を獲得した。

Swin Transformerは、ImageNetだけでなく、COCO（物体検出）やADE20K（セグメンテーション）においても当時のSOTAを記録し、Transformerが分類以外のタスクでもCNNを凌駕できることを証明した [11]。

3.2 データの壁を越える：Masked Autoencoders (MAE)

ViTの「データ貪欲性」に対する回答として登場したのが、自己教師あり学習（Self-Supervised Learning）、特に**Masked Autoencoders (MAE)**である（2022年、Meta AI）[10]。NLPにおけるBERTの成功に触発されたこの手法は、視覚モデルの学習方法を一変させた。

MAEのアプローチは極めてシンプルである：

1. 入力画像をパッチに分割し、その大部分（例：75%）をランダムに隠す（マスクする）。
2. 残りの見えるパッチ（25%）だけをエンコーダーに入力し、潜在表現を得る。
3. 軽量なデコーダーを用いて、潜在表現とマスクトークンから、元の画像の画素値を復元する。

このタスクは非常に難易度が高い。画像の75%が欠落した状態で元の画素を復元するためには、モデルは「物体の形状」、「テクスチャの連続性」、「シーンの文脈的整合性」を深く理解していなければならない。NLP（BERT）ではマスク率は15%程度であるが、画像は冗長性が高いため、75%という高いマスク率がむしろ効果的な表現学習を促すことが発見された [10]。

MAEの利点は二つある。第一に、学習効率である。マスクされた75%のパッチはエンコーダーに入力されないため、学習時の計算量とメモリ消費を劇的に削減できる。第二に、表現能力である。ラベルなしデータを用いた事前学習だけで強力な特徴表現を獲得でき、少数のラベル付きデータでのファインチューニングでSOTAを達成した。これにより、ViTは「ラベル付き巨大データセット」への依存から解放され、より広範なドメインへの適用が可能となった。

3.3 ハイブリッドモデルの完成形：CNNは死なず

一方、CNN側も黙ってはいなかった。Transformerの成功要因を分析し、それをCNNに取り入れる試みが進んだ。その代表例が**ConvNeXt**である（2022年）。ConvNeXtは、ResNetの構造を出発点とし、ViTの設計思想（パッチ化に近いカーネルサイズ、GELU活性化関数、LayerNorm、AdamWオプティマイザなど）を徹底的に模倣して改良されたCNNである [14]。

ConvNeXtは、Self-Attentionを一切使用しない純粋なCNNでありながら、Swin Transformerと同等以上の精度を達成した。これは、「Transformerの強みの一部は、Attention機構そのものよりも、学習レシピやアーキテクチャの細部（マクロな設計）にある」ことを示唆している。また、Googleの**CoAtNet**や**BoTNet**のように、浅い層ではCNNを用いて局所的な特徴（エッジやテクスチャ）を効率的に抽出し、深い層ではTransformerを用いて大域的な推論を行う「ハイブリッドアーキテクチャ」も、収束の速さと精度の高さを両立する現実解として定着した [6]。

4. 2025-2026年の最前線：ImageNetと物体検出の現在地

2026年現在、静止画認識におけるSOTA争いは、単なるアーキテクチャの優劣から、学習レシピ、データスケーリング、そしてマルチモーダル統合の競争へと移行している。

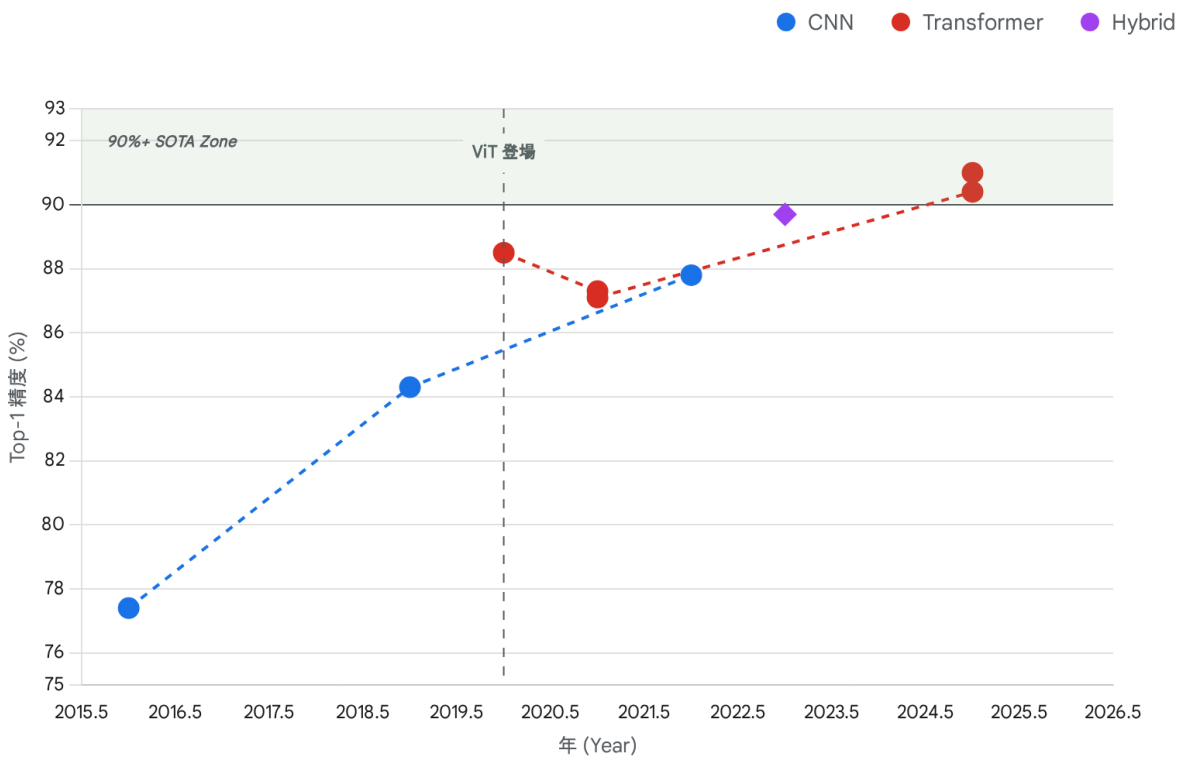
4.1 ImageNet Top-1 精度の到達点

最新のリーダーボード（2025-2026年）において、ImageNet（1K）のTop-1精度は、かつての夢の数

字であった90%の壁を超え、91.0%以上に達している[14]。

- **CoCa (Contrastive Captioners):** Googleが開発したCoCaは、画像単体での学習ではなく、画像とテキストのペアを用いたマルチモーダル学習（CLIPのようなContrastive Learningと、画像キャプション生成のGenerative Lossの組み合わせ）を活用している。ファインチューニングによりImageNetで91.0%を達成。これは、純粋な画像分類モデルというよりも、視覚と言語を統合した基盤モデル（Foundation Model）が、単一タスクにおいても最強であることを示している[14]。
- **DaViT (Dual Attention Vision Transformers):** 空間方向（Spatial）とチャネル方向（Channel）のAttentionを分離して処理することで、計算効率を高めつつ大域的コンテキストを捉えるモデル。DaViT-Giantは90.4%を記録している[14]。
- **ConvNeXt V2:** MAEの事前学習手法をCNN向けにアレンジしたFCMAE（Fully Convolutional Masked Autoencoder）を用いることで、CNNベースながら88%超の精度を実現。CNNが依然としてトップティアで競争力を持つことを証明している[14]。

ImageNet (Top-1) 精度の変遷：CNNからTransformer、そして基盤モデルへ



2020年のViT登場以降、精度向上曲線が再加速したことがわかる。2025-26年はCoCaやDaViTなどの大規模モデルが90%超の領域に到達している。一方でConvNeXt V2などCNN勢も巻き返しており、純粋な優劣よりも「学習手法（自己教師あり学習など）」の重要性が増している。

Data sources: NVIDIA Blog (VOLO), HiringNet (2025 SOTA), HyperAI, ResearchGate

4.2 物体検出：DETRの遺産とYOLOの逆襲

物体検出（Object Detection）においてもパラダイムシフトが起きた。かつてはRegion Proposal Network (RPN)、Anchor Box、NMS（Non-Maximum Suppression）といった手動設計のヒューリスティクスに依存していたが、Transformerベースの**DETR (DEtection TRansformer)** は、物体検出を「セット予測問題（Set Prediction Problem）」として定式化し、End-to-Endでの学習を可能にした。DETRは、画像内の物体間の関係性をSelf-Attentionでモデル化することで、重複検出を抑制し、NMSなしで高精度な検出を実現した [16]。

2025年のCOCOデータセットリーダーボードでは、以下の二極化が見られる [17]：

- **RF-DETR (Real-time Flow DETR):** 従来のDETRは収束が遅く、推論コストが高いことが欠点だったが、Roboflowなどが開発したRF-DETRは、YOLOに匹敵する速度と、それを上回る精度（

60.6 mAP）を達成。Transformerによるリアルタイム検出が産業レベルで実用化された。
- **YOLOv12:** 一方で、CNNベースのYOLO（You Only Look Once）も進化を止めていない。YOLOv12は、計算効率の高さとエッジデバイスへの実装の容易さで、依然として産業界のデファクトスタンダードである。特にYOLOv12-N（Nano）のような極小モデルでは、CNNのアーキテクチャ効率がTransformerのオーバーヘッドを上回るため、圧倒的に有利である。

5. 動画生成の革命：Diffusion Transformer (DiT)

視覚分野におけるTransformerの貢献で、現在最も注目すべきは「認識（Recognition）」から「生成（Generation）」への拡張である。2024年にOpenAIが発表した動画生成AI「Sora」は、世界に衝撃を与えた。その核心技術が **Diffusion Transformer (DiT)** である。

5.1 U-NetからTransformerへ：スケーリングの壁を壊す

Stable Diffusion 1.x/2.xなどの従来の画像生成AIは、ノイズ除去のバックボーンネットワークとしてCNNベースの**U-Net**を使用していた。U-Netは、ダウンサンプリングとアップサンプリングを行う過程で画像の局所的な特徴を捉えるのには優れている。しかし、動画のような時間的・空間的に複雑な依存関係や、異なるモダリティ（テキスト、画像、動画、3D）を統合的に扱うには、CNNの固定的な構造は柔軟性を欠いていた [19]。

DiTは、このU-NetをTransformerに置き換えたものである。Saining Xieらによって提案されたこのアーキテクチャは、以下の特徴を持つ：

1. 潜在空間での処理：高解像度画像を直接ピクセル空間で扱うのではなく、VAE（Variational Autoencoder）で圧縮された潜在空間（Latent Space）で処理を行う。
2. パッチ化（**Tokenization**）：潜在表現（Latent Image）をパッチに分割し、トークン列として扱う。これにより、画像サイズやアスペクト比の変化に柔軟に対応できる。
3. スケーリング則の実践：Transformerの最大の武器である「モデルサイズとデータ量を増やせば増やすほど性能が向上する（Scaling Law）」という特性を、画像生成の世界に持ち込んだ。DiTのパラメータ数を増やすことで、生成される画像のFID（Fréchet Inception Distance）スコアが一貫して向上することが実証されている [19]。

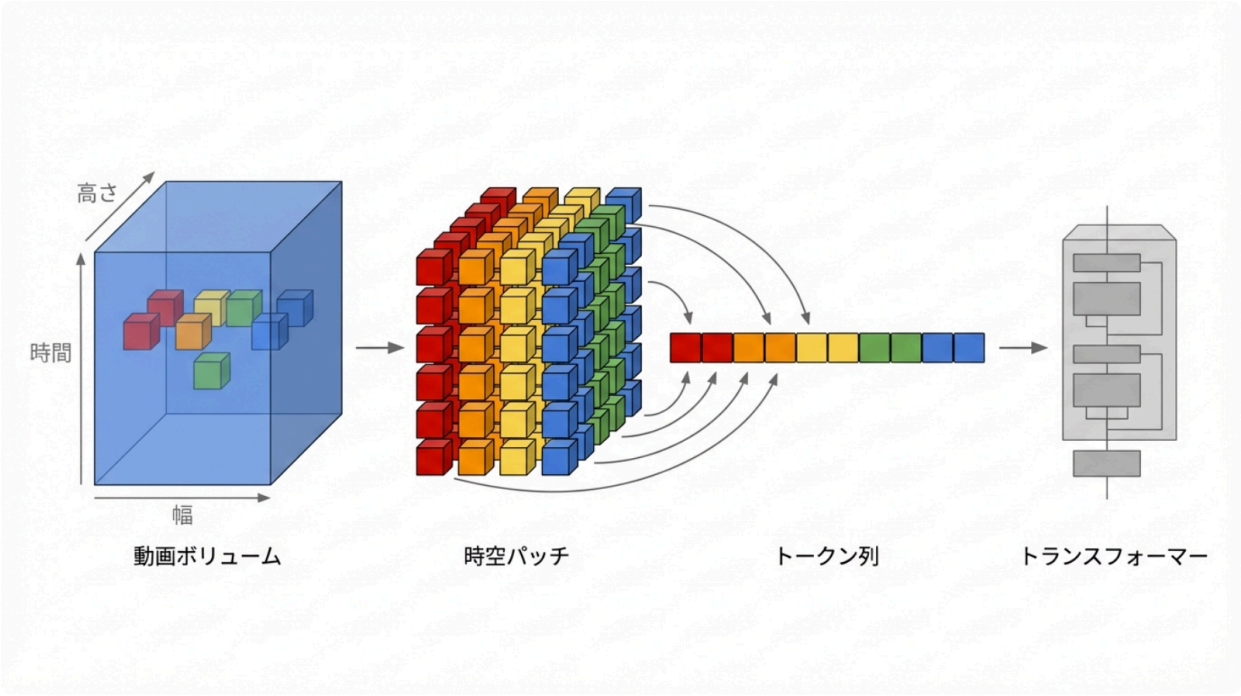
5.2「時空パッチ（Spacetime Patches）」による物理シミュレーション

Soraや、TencentのHunyuan Video [21] などの最新動画生成モデルでは、動画を「時間の経過する画像（2D + Time）」としてではなく、「3次元の時空ボリューム（Spacetime Volume）」として扱うアプローチが採用されている。

- **Spacetime Patches:** 動画データをフレームごとに処理するのではなく、時間軸を含めた3次元データとして捉え、それを小さな立方体（Cube）のパッチに分割する。これを「時空パッチ」と呼ぶ。
- 物理法則の学習**:** Transformerは、これらの時空パッチ間の相互関係を学習する。これは単に「次のフレームの画素を予測する」こと以上の意味を持つ。ボールが壁に当たって跳ね返る、水が流れる、光が反射するといった現象は、時空パッチ間の因果関係としてモデル化される。Soraが「世界シミュレータ（World Simulator）」と呼ばれる所以はここにある。Transformerは、

大量の動画データを学習することで、世界の動的な遷移ルール——すなわち物理法則の近似——を、自己教師あり学習によって獲得している可能性が示唆されている[23]。

時空パッチ化：動画を3次元ボリュームとして処理する仕組み



動画（左）は一連のフレームとしてではなく、時間軸（T）を含む3次元の直方体として扱われる。これを「時空パッチ」という小さな立方体に分割し、フラット化してTransformer（右）に入力する。これにより、空間的な位置関係と時間的な変化を同時に学習できる。

6. 新たな挑戦者：Mambaと線形計算量の可能性

2024年から2026年にかけて、Transformerの「王座」を脅かす可能性のある新たなアーキテクチャが急速に注目を集めている。**State Space Models (SSM)**、特に**Mamba**である[25]。

6.1 二次関数的計算量の壁とMambaの解

TransformerのSelf-Attentionは、入力トークン数 N に対して $O(N^2)$ の計算量とメモリを必要とする。224x224の画像（パッチ数196）なら問題ないが、4K画像や長尺の動画、高解像度の3D医療画像（CT/MRI）では、トークン数が数万～数十万に達し、計算量が爆発する。

Mambaは、古典的な状態空間モデルに、入力依存の選択メカニズム（Selection Mechanism）を導入したモデルである。リカレントニューラルネットワーク（RNN）のようにデータを逐次的に処理するた

め、推論時のメモリ消費は一定であり、学習時の計算量はトークン数に対して線形（$O(N)$）である。それでいて、並列学習が可能であるというTransformerの利点も併せ持つ。

6.2 Vision Mamba (Vim) の革新と効率性

2024年初頭に発表された**Vision Mamba (Vim)**は、画像をパッチ列として扱い、それを双方向（Forward & Backward）のSSMで処理するアーキテクチャである。画像にはテキストのような明確な「順序」がないため、双方向処理によって空間的な位置関係を捉える工夫がなされている [26]。

- **効率:** Vimは、DeiT（標準的なViT）と比較して、高解像度画像（1248x1248）において2.8倍高速であり、GPUメモリ消費を86.8%削減できることが示されている。これは、エッジデバイスや高解像度タスクにおいて革命的な効率化である [26]。

6.3「MambaOut」論争と2026年の結論

しかし、2025年に発表された論文「MambaOut: Do We Really Need Mamba for Vision?」は、このブームに冷静な視点を提供した [29]。ImageNetのような標準的な画像分類タスクでは、MambaはViTや最新のCNNに対して、必ずしも優位性を示さなかったのである。 同論文の分析によれば、ImageNet分類のようなタスクでは「長期依存性（Long-range dependency）」があまり重要ではなく、むしろCNN的な局所性が依然として有効である。そのため、Mambaのような長距離相関に強いモデルの恩恵が薄いとされる。

2026年の現時点での結論としては、MambaはViTを「完全に置き換える」ものではなく、「補完する」ものとして位置付けられている。

- **ViTの領域:** 一般的な画像認識、中解像度の処理、生成モデル（Soraなど）。
- **Mambaの領域:** 超高解像度画像（衛星写真、ギガピクセル病理画像）、長尺動画解析、3D医療画像、時系列データとのマルチモーダル処理など、シーケンス長が極めて長く、Transformerでは扱いきれないタスク [28]。
- **InceptionMamba:** 2026年には、CNN（Inceptionモジュール）の局所特徴抽出能力と、Mambaの大域的モデリング能力を組み合わせたハイブリッドモデルも登場し、医療画像分類でSOTAを達成している [31]。

7. ドメイン特化型の革命：医療・科学・エッジ

汎用モデルの進化と並行して、特定のドメインにおいては、Transformerの特性（大域的な文脈理解）が決定的なブレイクスルーをもたらしている。

7.1 医療画像とAlphaFold 3

医療分野、特にデジタルパソロジー（病理画像診断）では、WSI（Whole Slide Imaging）と呼ばれるスライド画像が扱われる。これらは1枚で数ギガピクセル（数万×数万ピクセル）に達する巨大な画像であり、CNNの小さな受容野だけでは、「組織全体の構造」や「腫瘍の広がり」、「微小環境（Microenvironment）」を把握することが困難であった。ここで、ViTの大域的コンテキスト理解能力が

不可欠となる [32]。

さらに、Google DeepMindの**AlphaFold 3**（2024年）は、タンパク質構造予測において「Atom Transformer」と「Diffusion Module」を組み合わせることで、従来のAlphaFold 2を超え、タンパク質だけでなく、低分子リガンド、DNA、RNA、修飾基を含むあらゆる生体分子の複合体構造を予測可能にした。AlphaFold 3は、アミノ酸残基だけでなく原子（Atom）レベルでの相互作用をAttention機構で捉え、生成モデル（Diffusion）で構造を組み立てる。これは、ViTの技術が単なる画像認識を超え、科学的発見（Scientific Discovery）の基盤ツール（AI for Science）となったことを象徴している [34]。

7.2 自動運転とオクルージョン：見えないものを見る

自動運転の認識システムにおいて、最大の課題の一つが「オクルージョン（遮蔽）」である。駐車車両の影から飛び出す歩行者や、他の車に半分隠れた車両を認識しなければならない。CNN単体では、見えている部分の特徴しか捉えられないため、遮蔽された物体の認識は困難であった。しかし、Transformerベースのモデル（例：BEVFormerやFAOD）は、時間方向と空間方向のAttention（Spatio-temporal Attention）を用いることで、過去のフレームからの情報を現在に統合したり、他のセンサー（LiDARやRadar）からの情報を融合（Sensor Fusion）したりすることで、見えていない部分を「補完」する能力に優れている。CNNでは困難だった「文脈に基づく推論」が可能になり、安全性が飛躍的に向上した [37]。

7.3 エッジAIへの実装：JetsonとRaspberry Piの戦い

Transformerは計算コストが高いため、かつてはクラウド上のGPUでしか動かないとされていた。しかし、2025-2026年にはエッジデバイスでの実行が現実的になりつつある。

- **NVIDIA Jetson Orin Nano / Super:** 67 TOPSのAI性能を持ち、量子化（INT8）やTensorRTによる最適化を施したViTモデルをリアルタイムで動作させることができる。ロボティクスや高度な監視カメラなど、GPUパワーが利用可能な環境では、CNNからTransformerへの移行が進んでいる [39]。
- **Raspberry Pi 5:** CPU推論が主であり、GPU性能は限定的（VideoCore VII）であるため、依然としてMobileNetやEfficientNetといった軽量CNN、あるいはYOLOv8/v11/v12のNanoモデルが主流である。しかし、Hailo-8などのAIアクセラレータ（NPU）をアドオンすることで、Raspberry Piでも軽量なViTや物体検出Transformerを実行する試みが進んでおり、その差は縮まりつつある [42]。

8. 結論：統一された知覚（Unified Perception）へ

「CNNが王座を譲る日」という問いに対する2026年の答えは、「王座の意味が変わった」ということである。

1. **SOTA**の覇者としての**Transformer:** 精度（Accuracy）、スケーラビリティ、生成能力において、Transformerは疑いようもなく現在の王者である。ImageNetのトップスコア、Soraによる動画生成、AlphaFoldによる科学的発見は、すべてTransformer（およびその派生であるDiT）によってもたらされた。
2. 実務のワークホースとしての**CNN:** 一方で、CNNは死んでいない。YOLOv12やConvNeXtのよう

に、Transformerから学んだ教訓を取り入れ、エッジデバイスやリアルタイム性が求められる現場、あるいは学習データが限られるタスクにおいて、その効率性と安定性でしぶとく生き残っている。
3. 新たなフロンティアとしての**SSM:** さらに、Mambaのような線形計算モデルが、Transformerの弱点である計算量を克服し、長尺動画や3D解析という新たな領域を開拓しつつある。

我々は今、AIが単に画像を「分類（Classify）」する段階を終え、テキスト、画像、動画、3D、そして科学データを統一的に扱い、世界を「理解（Understand）」し「シミュレート（Simulate）」する段階へと突入した。この「統一された知覚（Unified Perception）」の時代において、CNN、Transformer、Mambaは対立するものではなく、それぞれの特性（局所性、大域性、線形性）を活かして適材適所で組み合わされる構成要素（Building Blocks）となったのである。歴史的転換点は通過した。我々はその先にある、より広大で統合的なAIの景色を見ている。

参考文献・データソース: 本レポートの分析は、提供されたスニペット資料 [1] - [44] に基づく。特にViTの基礎理論については [1]、最新のImageNet動向は [14]、DiTに関しては [19]、Mambaに関しては [26] を主要な根拠としている。

引用文献

1. A Survey on Visual Transformer - arXiv, 2月 3, 2026にアクセス、https://arxiv.org/pdf/2012.12556
2. A Survey of Visual Transformers - arXiv, 2月 3, 2026にアクセス、https://arxiv.org/pdf/2111.06091
3. Model-agnostic Measure of Generalization Difficulty, 2月 3, 2026にアクセス、https://proceedings.mlr.press/v202/boopathy23a/boopathy23a.pdf
4. Are Transformers replacing CNNs in Object Detection? - Picsellia, 2月 3, 2026にアクセス、https://www.picsellia.com/post/are-transformers-replacing-cnns-in-object-detection
5. Understanding and Incorporating Mathematical Inductive Biases in ..., 2月 3, 2026にアクセス、https://cs.nyu.edu/media/publications/Marc_Finzi_Thesis__12_.pdf
6. The Rise of Hybrid CNN-Transformer Architectures - Medium, 2月 3, 2026にアクセス、https://medium.com/@savindufernando/the-rise-of-hybrid-cnn-transformer-architectures-5e101986f51d
7. Object Detection using Vision Transformer and Deep Learning for ..., 2月 3, 2026にアクセス、https://www.researchgate.net/publication/393769299_Object_Detection_using_Vision_Transformer_and_Deep_Learning_for_Computer_Vision_Applications
8. A THEORETICAL UNDERSTANDING OF SHALLOW ... - OpenReview, 2月 3, 2026にアクセス、https://openreview.net/pdf?id=jClGv3Qjhb
9. Vision Transformers for Image Classification: A Comparative Survey, 2月 3, 2026にアクセス、https://www.mdpi.com/2227-7080/13/1/32
10. Masked autoencoders are effective solution to transformer data-hungry, 2月 3,

2026にアクセス、https://www.semanticscholar.org/paper/f4991092d16960ed4629bd31029af93243cb8777
11. Breaking Down Swin Transformer: Understanding Relative Position ..., 2月 3, 2026にアクセス、https://medium.com/@ovularslan/breaking-down-swin-transformer-understanding-relative-position-bias-and-masked-self-attention-437d692ab7cf
12. Local pattern aware 3D video swin transformer with masked ..., 2月 3, 2026にアクセス、https://pubmed.ncbi.nlm.nih.gov/40594635/
13. Swin MAE: Masked Autoencoders for Small Datasets - arXiv, 2月 3, 2026にアクセス、https://arxiv.org/pdf/2212.13805
14. Image Classification: State-of-the-Art Models in 2025 - HiringNet, 2月 3, 2026にアクセス、https://hiringnet.com/image-classification-state-of-the-art-models-in-2025
15. A Hybrid Network of CNN and Transformer for Lightweight Image ..., 2月 3, 2026にアクセス、https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/papers/Fang_A_Hybrid_Network_of_CNN_and_Transformer_for_Lightweight_Image_CVPRW_2022_paper.pdf
16. A Survey of Dense Object Detection Methods Based on Deep ..., 2月 3, 2026にアクセス、https://ieeexplore.ieee.org/iel8/6287639/10380310/10770219.pdf
17. Best Object Detection Models 2025: RF-DETR, YOLOv12 & Beyond, 2月 3, 2026にアクセス、https://blog.roboflow.com/best-object-detection-models/
18. Object Detection: State-of-the-Art Models in 2025 - HiringNet, 2月 3, 2026にアクセス、https://hiringnet.com/object-detection-state-of-the-art-models-in-2025
19. Diffusion Transformers Explained: The Beginner's Guide - Lightly, 2月 3, 2026にアクセス、https://www.lightly.ai/blog/diffusion-transformers-dit
20. Diffusion Transformer (DiT) Models: A Beginner's Guide - Encord, 2月 3, 2026にアクセス、https://encord.com/blog/diffusion-models-with-transformers/
21. Video Generation: Evolution from VDM to Veo2 and SORA, 2月 3, 2026にアクセス、https://learnopencv.com/video-generation-models/
22. Hunyuan Video | Best AI for Video | Find AI Tools & Apps - ai-search.io, 2月 3, 2026にアクセス、https://ai-search.io/tool/hunyuan-video
23. Diffusion-Based Video Transformer - Emergent Mind, 2月 3, 2026にアクセス、https://www.emergentmind.com/topics/diffusion-based-video-transformer
24. Autoregressive Video Generation with Reward Feedback, 2月 3, 2026にアクセス、https://www.researchgate.net/publication/400072021_Reward-Forcing_Autoregressive_Video_Generation_with_Reward_Feedback
25. TRANSFORMER VS. MAMBA AS SKIN CANCER CLASSIFIER, 2月 3, 2026にアクセス、https://scinews.kpi.ua/article/view/301028
26. Vision Mamba: Efficient Visual Representation Learning with ..., 2月 3, 2026にアクセス、https://raw.githubusercontent.com/mlresearch/v235/main/assets/zhu24f/zhu24f.pdf
27. Vision Mamba: Like a Vision Transformer but Better, 2月 3, 2026にアクセス、

https://towardsdatascience.com/vision-mamba-like-a-vision-transformer-but-better-3b2660c35848/
28. Vision Mamba for efficient Tuberculosis Detection based on Chest X ..., 2月 3, 2026にアクセス、https://www.researchgate.net/publication/394944876_Vision_Mamba_for_efficient_Tuberculosis_Detection_based_on_Chest_X-Rays_A_comparative_study_with_CNN_and_Vision_transformers
29. MambaOut: Do We Really Need Mamba for Vision?, 2月 3, 2026にアクセス、https://openaccess.thecvf.com/content/CVPR2025/papers/Yu_MambaOut_Do_We_Really_Need_Mamba_for_Vision_CVPR_2025_paper.pdf
30. An Efficient Vision Mamba–Transformer Hybrid Architecture ... - MDPI, 2月 3, 2026にアクセス、https://www.mdpi.com/1424-8220/25/21/6785
31. A Lightweight and Effective Model for Medical Image Classification ..., 2月 3, 2026にアクセス、https://www.researchgate.net/publication/399534959_InceptionMamba_A_Lightweight_and_Effective_Model_for_Medical_Image_Classification_Revealing_Mamba's_Low-Frequency_Bias
32. Spatially Aware Transformer Networks for Contextual Prediction of ..., 2月 3, 2026にアクセス、https://www.medrxiv.org/content/10.1101/2023.02.20.23286044v1.full-text
33. (PDF) Comparative analysis of convolutional neural networks and ..., 2月 3, 2026にアクセス、https://www.researchgate.net/publication/393185463_Comparative_analysis_of_convolutional_neural_networks_and_transformer_architectures_for_breast_cancer_histopathological_image_classification
34. An Overview of Alphafold's Breakthrough - Frontiers, 2月 3, 2026にアクセス、https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.875587/full
35. How does AlphaFold 3 work? - EMBL-EBI, 2月 3, 2026にアクセス、https://www.ebi.ac.uk/training/online/courses/alphafold/alphafold-3-and-alphafold-server/introducing-alphafold-3/how-does-alphafold-3-work/
36. Atom Transformer in AlphaFold3 - by Dr. Zhongqiang DING - Medium, 2月 3, 2026にアクセス、https://medium.com/@ding.zhongqiang/atom-transformer-in-alphafold3-5dfc57be5878
37. RSO-YOLO: A Real-Time Detector for Small and Occluded Objects ..., 2月 3, 2026にアクセス、https://pmc.ncbi.nlm.nih.gov/articles/PMC12610148/
38. Robust Occluded Object Detection in Multimodal Autonomous Driving, 2月 3, 2026にアクセス、https://www.mdpi.com/2079-9292/15/1/245
39. Raspberry Pi vs NVIDIA Jetson: the ultimate 2025 comparison, 2月 3, 2026にアクセス、https://monraspberry.com/en/raspberry-pi-vs-nvidia-jetson-the-ultimate-2025-comparison/
40. NVIDIA Jetson Orin Nano vs Raspberry Pi 5: The Ultimate Edge ..., 2月 3, 2026にアクセス、

https://thinkrobotics.com/blogs/learn/nvidia-jetson-orin-nano-vs-raspberry-pi-5-the-ultimate-edge-computing-showdown
41. Jetson Nano Super vs Raspberry Pi 5: AI vs Affordability, 2月 3, 2026にアクセス、https://www.thinclientdirect.com/nvidia-jetson-nano-super-vs-raspberry-pi-5-comparison/
42. (jetson nano) vs (Raspberry pi5 8g + hailo 26 TOP) - Ultralytics, 2月 3, 2026にアクセス、https://community.ultralytics.com/t/jetson-nano-vs-raspberry-pi5-8g-hailo-26-top/532
43. Jetson Nano vs Raspberry Pi 5 for AI: The Ultimate Performance and ..., 2月 3, 2026にアクセス、https://thinkrobotics.com/blogs/learn/jetson-nano-vs-raspberry-pi-5-for-ai-the-ultimate-performance-and-value-comparison
44. AlphaFold 3 as a Differentiable Framework for Structural Biology, 2月 3, 2026にアクセス、https://arxiv.org/html/2508.18446v1
45. Training a State-of-the-Art ImageNet-1K Visual Transformer Model ..., 2月 3, 2026にアクセス、https://developer.nvidia.com/blog/training-a-state-of-the-art-imagenet-1k-visual-transformer-model-using-nvidia-dgx-superpod/