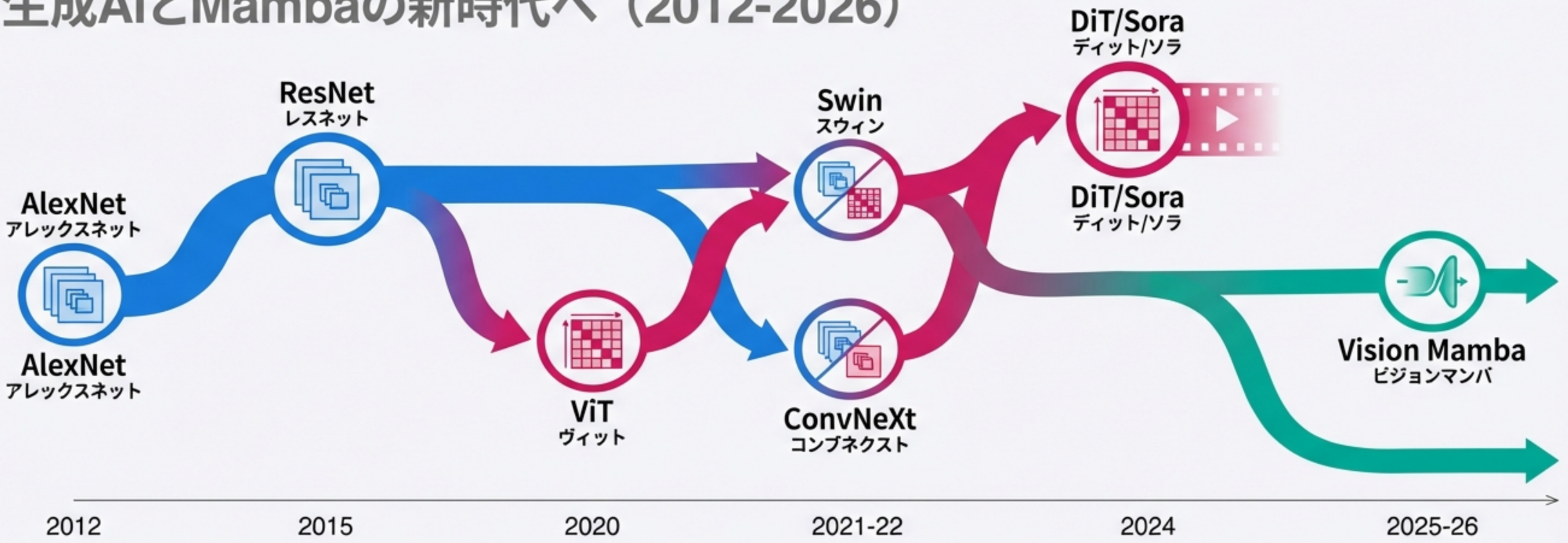
視覚革命の全貌

CNNの覇権からTransformer、そして生成AIとMambaの新時代へ（2012-2026）

Introduction: 2010年代の「絶対王者」CNNの時代から、2020年のTransformerによる破壊的イノベーション、そして2026年の「統一された知覚」に至るまでの技術史を紐解く。

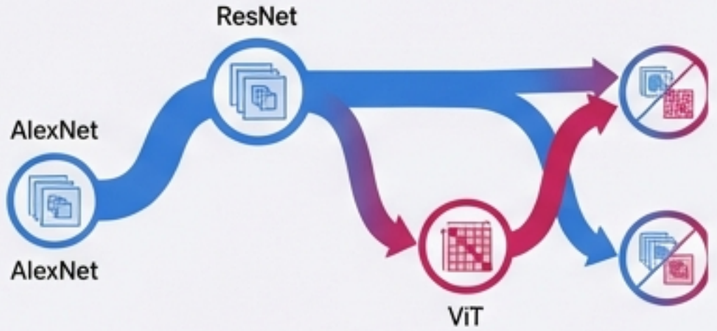
Key Themes: これは単なるアルゴリズムの競争ではない。「帰納的バイアス（Inductive Bias）」の設計と放棄、そして再導入を巡る、人類とAIの壮大な実験の記録である。



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。

パラダイムシフトの構造：帰納的バイアスの死と再生

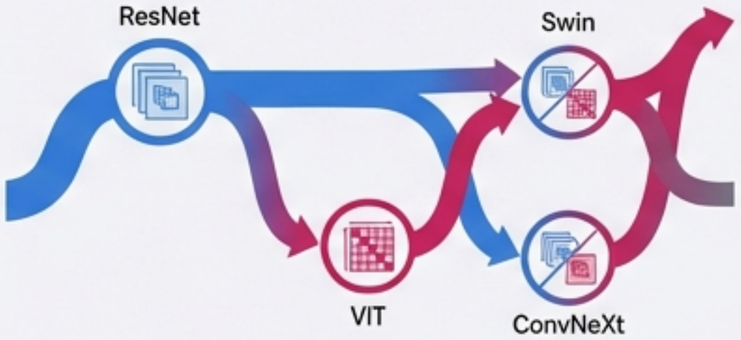
Thesis (2012-2020)
CNNの覇権



Concept: 局所性と移動不変性。人間が設計した「視覚のルール」が勝利した時代。

Representative: ResNet.

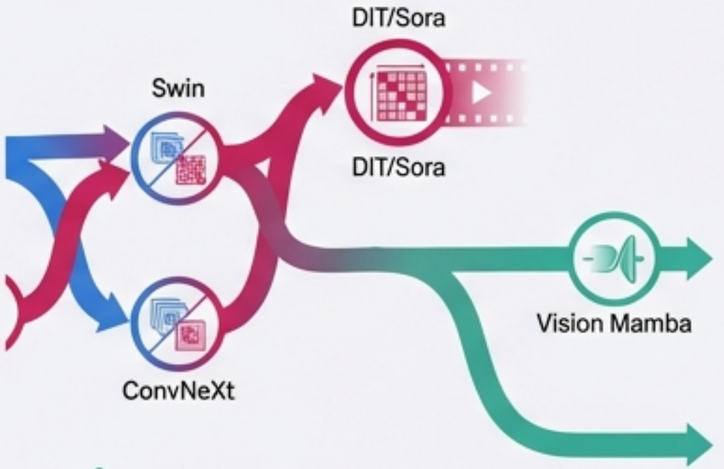
Antithesis (2020-2023)
ViTの破壊



Concept: ルールの撤廃。大域的受容野とSelf-Attentionによる「全体性の獲得」。スケーリング則がバイアスを凌駕する。

Representative: Vision Transformer (ViT).

Synthesis (2024-2026)
融合と生成



Concept: 効率化と物理シミュレーション。バイアスの再導入（Swin/ConvNeXt）と、動画生成（DiT）への応用。そして線形計算量（Mamba）への挑戦。

Representative: Sora, ConvNeXt, Vision Mamba.

絶対王者CNN：なぜ「畳み込み」は最強だったのか

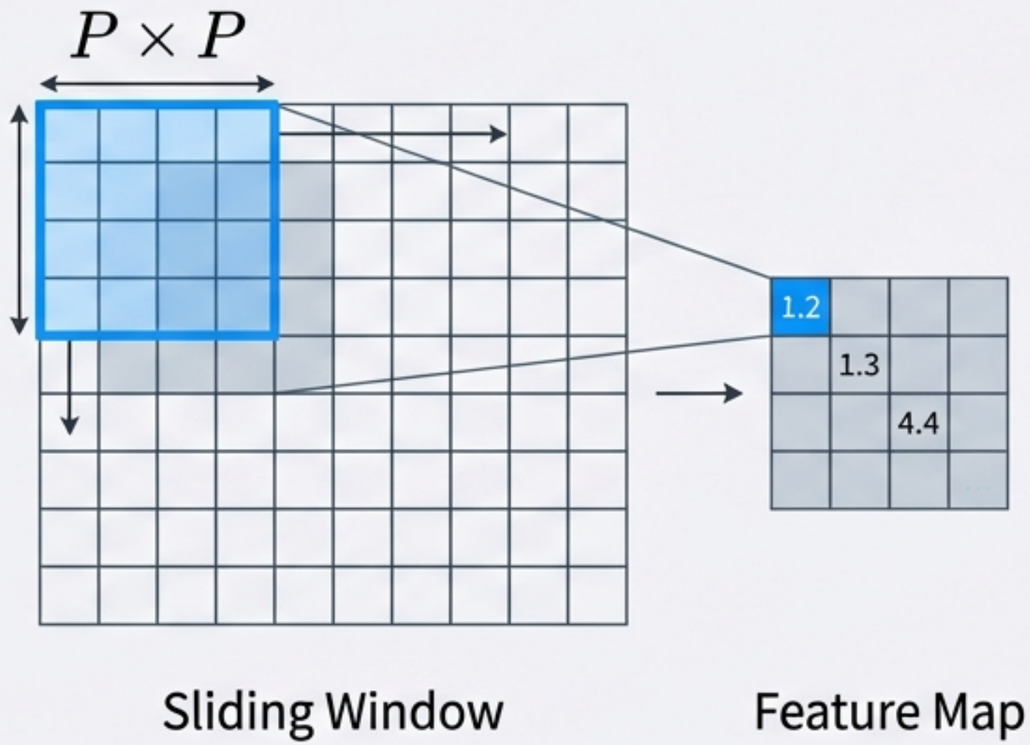
画像データの本質を捉えた強力な「帰納的バイアス」

Definition: **帰納的バイアス (Inductive Bias)**とは、モデルに先験的に組み込まれた「仮定」や「制約」。CNNは以下の強力なバイアスを持っていた：

1. **局所性** (Locality): 画像のピクセルは、遠くよりも近隣同士に強い相関がある。3x3などのフィルタで局所特徴（エッジ、テクスチャ）を効率的に抽出。
2. **移動不変性** (Translation Invariance): 猫が画像のどこにいても「猫」である。重み共有 (Weight Sharing) により、位置に依存しない認識能力を獲得。

Impact: これにより、少ないデータ量でも高い汎化性能 (Sample Efficiency) を発揮し、ディープラーニングの黄金時代を築いた。

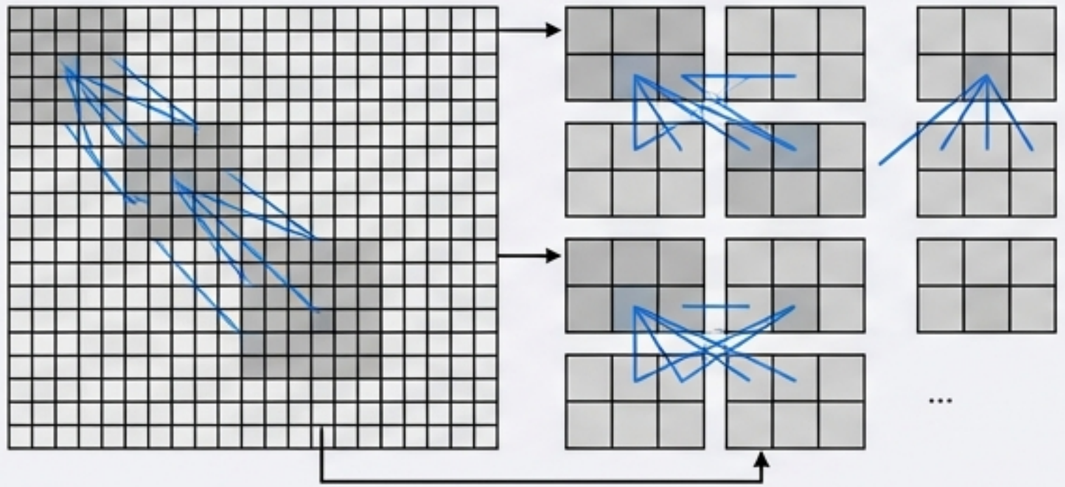
$$x \in \mathbb{R}^{H \times W \times C}$$



Vision Transformer (ViT) の衝撃：ルールからの脱却

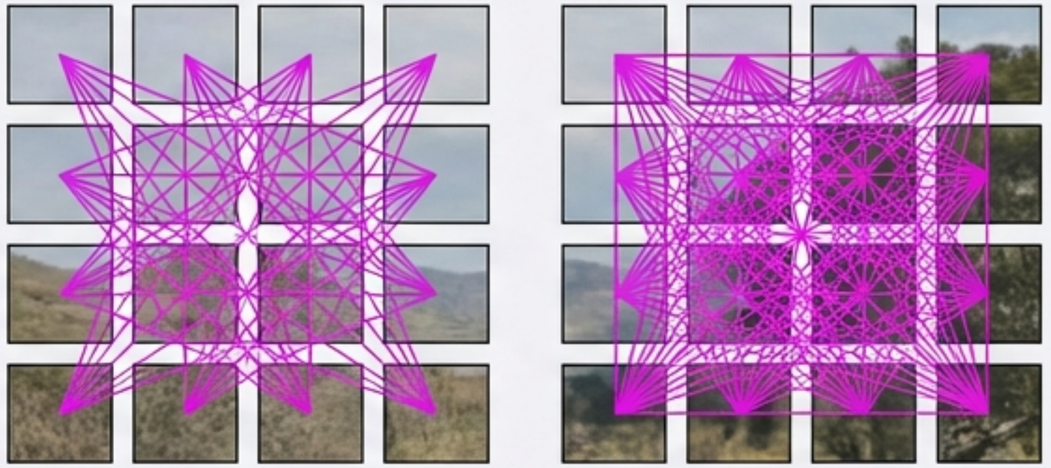
「An Image is Worth 16x16 Words」 — 画像を言語のように読む

CNN (Convolutional Neural Network)



局所的接続

ViT (Vision Transformer)



大域的相互作用
全結合アテンション

The Shift: 2020年、GoogleがCNNの常識を否定。畳み込み層を廃止し、画像をパッチ（16x16ピクセル）に分割して単語列として処理。
Mechanism: **Self-Attention（自己注意機構）**の導入。
Key Concept: **大域的受容野** (Global Receptive Field):

- **CNN**: 層を深く積み上げないと全体が見えない（Bottom-up）。
- **ViT**: 第1層目から、画像の左上と右下の関係性を直接捉える（Global-first）。

Implication: オクルージョン（遮蔽）や複雑な背景があっても、文脈から物体を認識可能に。

High-Fidelity Scientific Editorial

スケーリング則：「量」が「質」に転化する瞬間

ハードコードされたバイアスが「足かせ」になる時

- **The Paradox:** 小規模データ（ImageNet-1k）では、ViTはResNetに劣る。局所性や平行移動の概念をゼロから学ぶ必要があるため「データ貪欲（Data Hungry）」である。
- **The Breakthrough:** Google独自の超巨大データセット **JFT-300M（3億枚）**での事前学習。
- **The Conclusion:** データ量が閾値を超えた時、CNNの固定された仮定は制約となり、データから柔軟にパターンを学ぶ**ViT**が圧倒する。
- **Scaling Laws:**「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、手動設計よりも優れた表現を獲得できる」ことが実証された。

Swin Transformer：階層構造と局所性の再導入

二次関数的な計算量の壁を突破する

The Problem

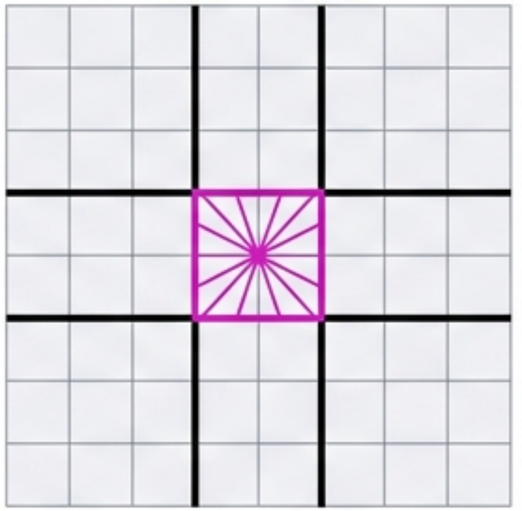
$$O(N^2)$$

$$H \times W$$

The Problem: 初期のViTは、画像を解像度が上がると計算量が爆発する（$O(N^2)$）。高解像度タスクは不可能だった。

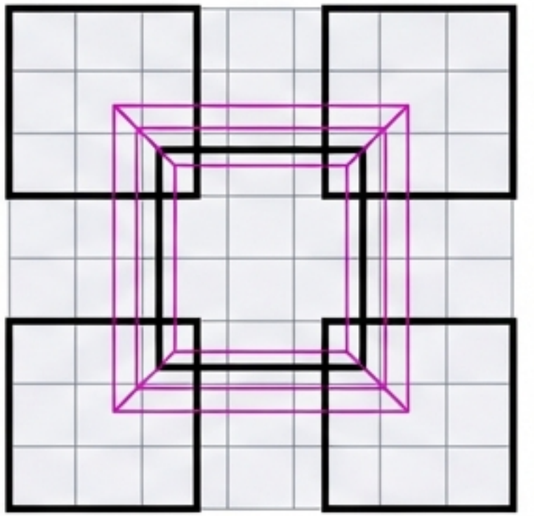
The Solution (2021 ICCV Best Paper): **Swin Transformer** (Microsoft)

Window-based Attention



画像全体ではなく、小さなウィンドウ内（例: 7x7）でのみAttentionを計算し、計算量を線形に抑える。

Shifted Window

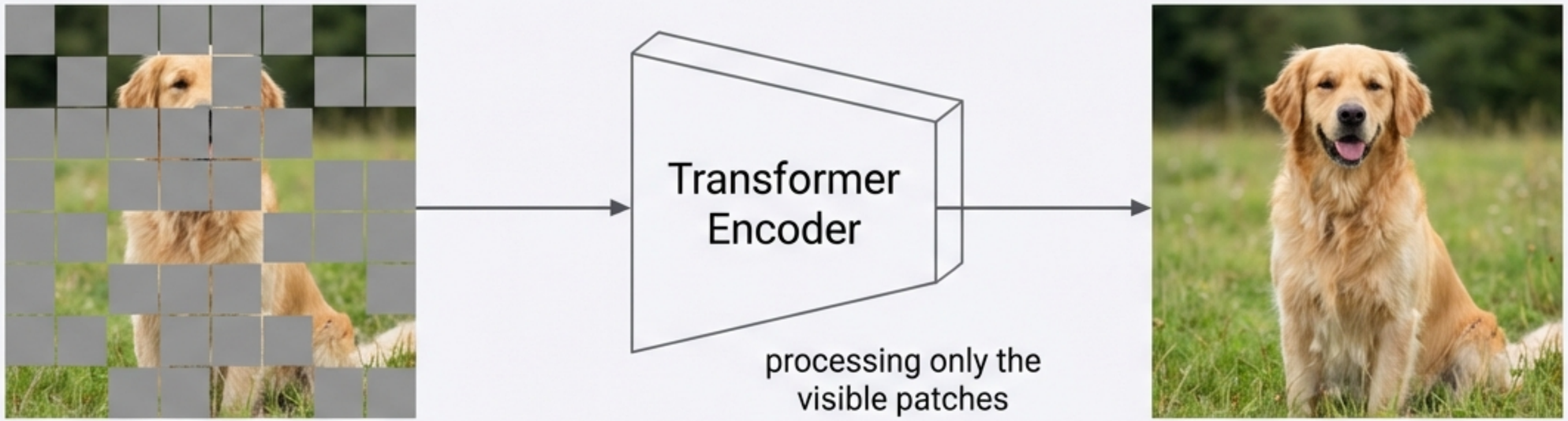


ウィンドウをずらすことで、隣接エリア間の情報伝達を確保。

Result: CNNのような「階層構造」を取り戻し、物体検出やセグメンテーションでもSOTAを記録。

Masked Autoencoders (MAE)：見えないものを視る力

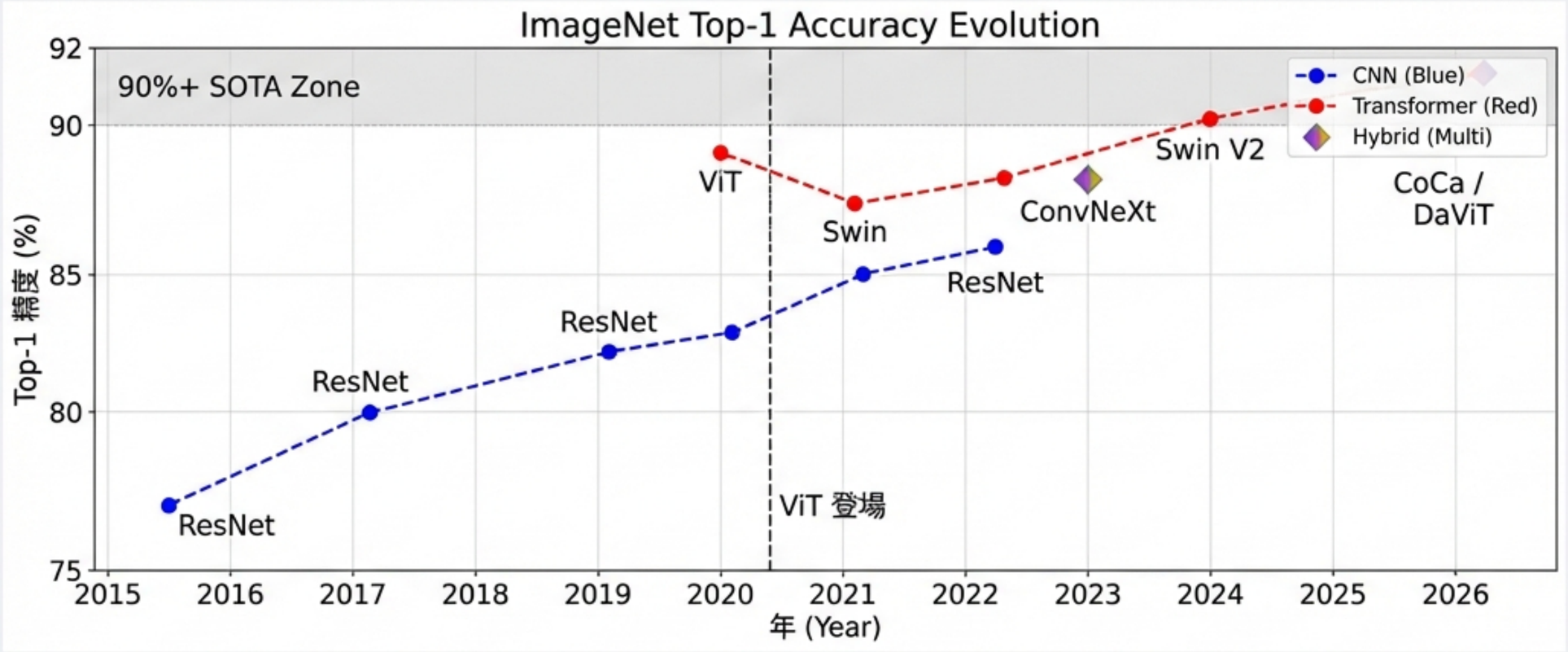
75%の欠落から全体を復元する自己教師あり学習



- **Concept:** 入力画像の大部分（75%）をマスクし、残りの25%から元の画像を復元させるタスク。
- **Why 75%?:** 言語（BERT）と異なり、画像は冗長性が高い。極端に隠すことで、モデルは「物体の形状」「テクスチャの連続性」「文脈」を深く理解せざるを得なくなる。
- **Efficiency:** マスクされた部分はエンコーダーに入力されないため、学習が高速。
- **Impact:** ラベルなしデータだけで強力な表現を獲得（Self-Supervised Learning）。ViTの「データ貪欲性」を克服し、実用化を加速させた。

ConvNeXt：現代化されたCNNの逆襲

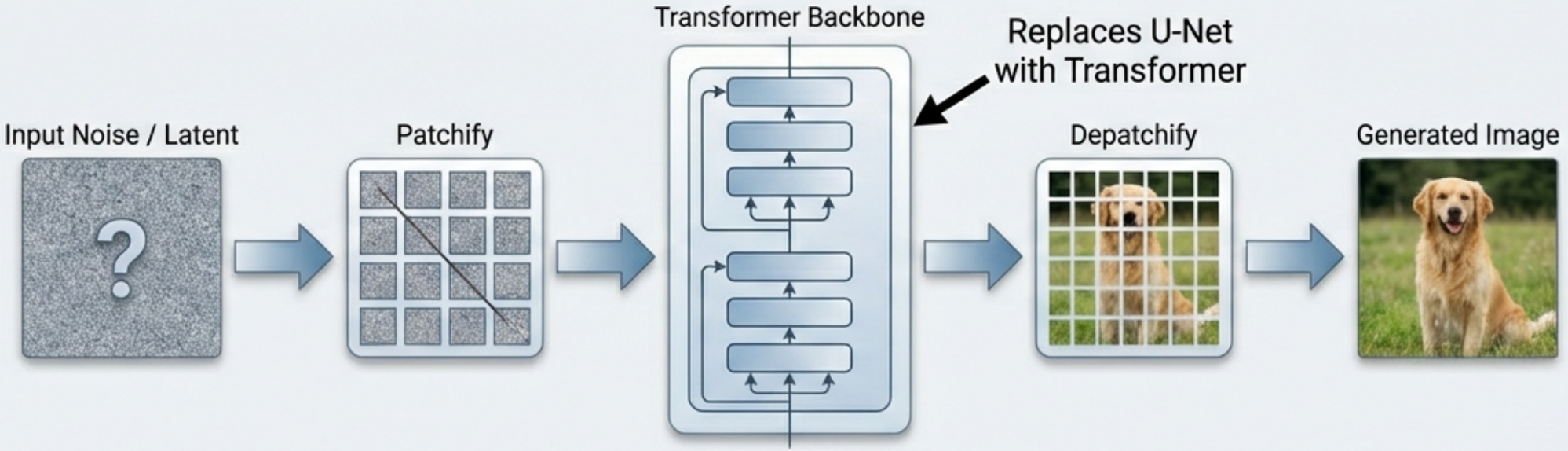
「重要なのはAttentionではなく、学習レシピだった」



- **The Insight:** ResNetをベースに、ViTの設計思想（パッチ化、GELU、LayerNorm、AdamWなど）を徹底的に模倣。
- **Result:** Self-Attentionを一切使わない純粋なCNNでありながら、Swin Transformerと同等の精度を達成。
- **Key Takeaway:** 2026年現在、CNNは死んでいない。**CoAtNet**のようなハイブリッドモデルや、**ConvNeXt V2** (FCMAE採用）として、トップティアの精度（ImageNet Top-1 88-91%）を維持している。

Diffusion Transformer (DiT)：認識から生成へ

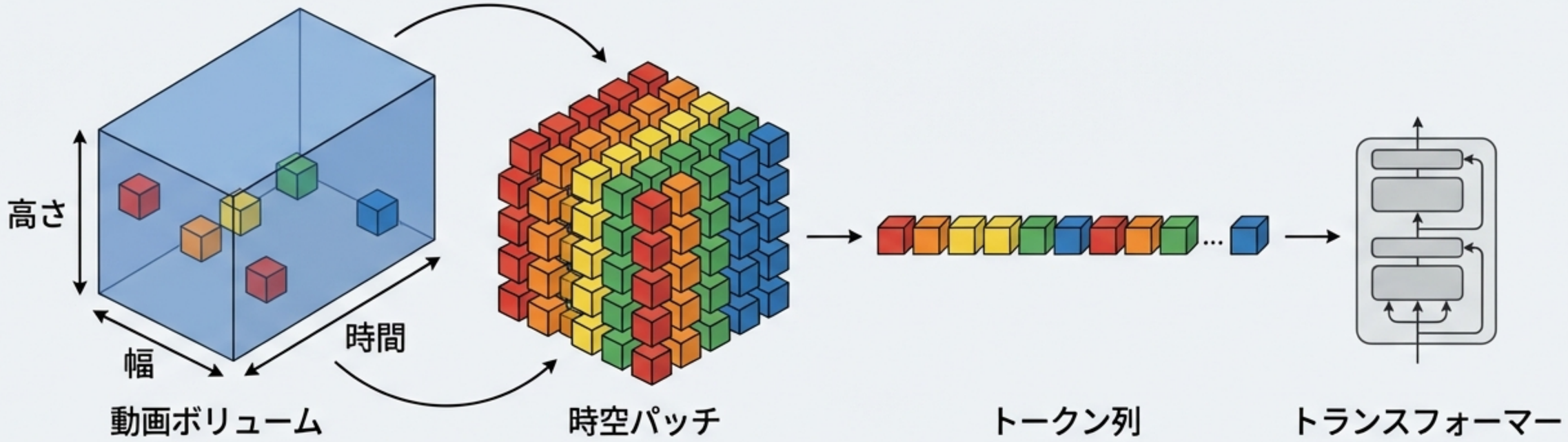
U-Netを置き換え、スケーリング則を生成モデルに持ち込む



- **Evolution:** 従来の画像生成（Stable Diffusion 1.x/2.x）はCNNベースのU-Netを使用していた。
- **Innovation:** U-NetをTransformerバックボーンに置換。
- **Mechanism:**
  1. Latent Space: 高解像度画像を圧縮した潜在空間で処理。
  2. Patchify: 潜在表現をパッチ（トークン）化し、Transformerに入力。
- **Impact:** モデルサイズとデータを増やせば増やすほど生成品質（FIDスコア）が向上するスケーリング則が、画像生成でも成立することを実証。

Soraと時空パッチ：物理世界をシミュレートする

動画を「フレームの連続」ではなく「3次元ボリューム」として捉える

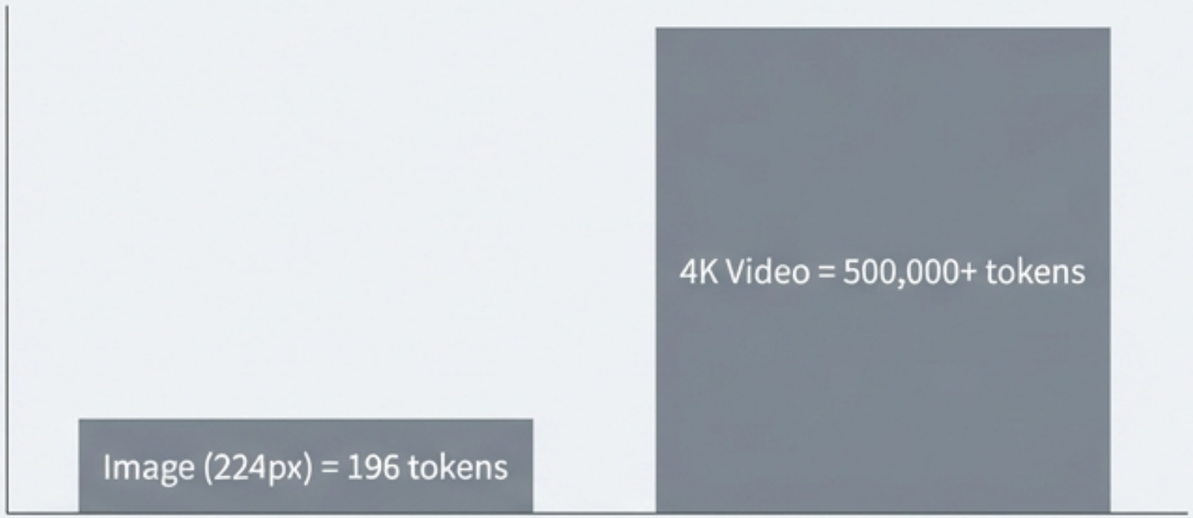


- Core Technology: **時空パッチ (Spacetime Patches)**。動画を時間軸(T)を含めた3次元の直方体として切り出し、フラット化して処理。
- **Physics Simulation**: Transformerはパッチ間の因果関係を学習する。ボールの落下、水の反射、衝突などの物理現象を、画素の遷移則として近似。
- **World Simulator**: OpenAIはこれを「**世界シミュレータ**」と呼ぶ。大規模学習により、AIは世界の物理法則を（暗黙的に）獲得しつつある。

計算量の壁：Transformerの限界

シーケンス長が増大する未来（4K動画、DNA解析）への課題

$$O(N^2)$$



The Bottleneck: Self-Attentionの計算量はトークン数の二乗（$O(N^2)$）で増える。

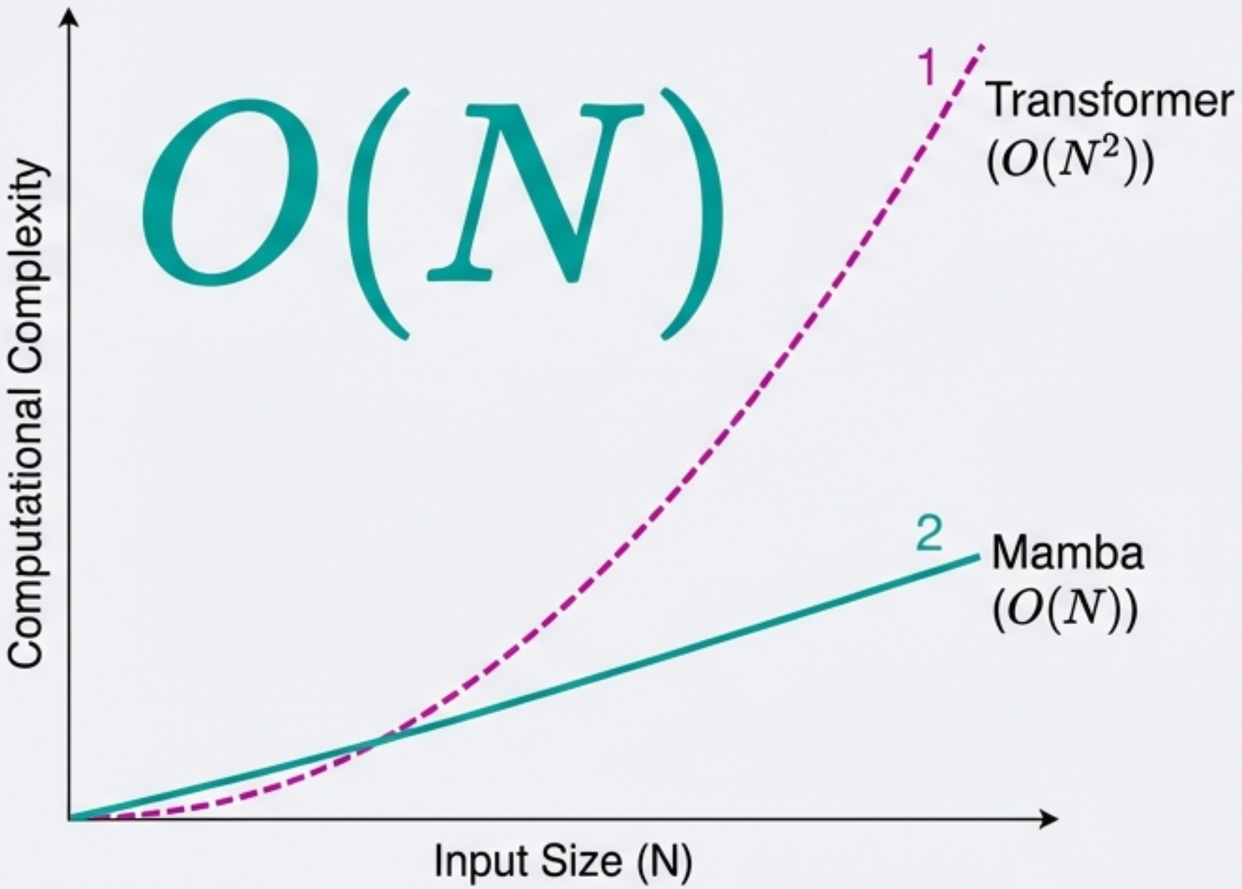
Impact:

- 画像（224x224）= 196トークン → 問題なし。
- 4K動画や3D医療画像 = 数万〜数十万トークン → 計算爆発。

Need: Transformerの精度を維持しつつ、線形計算量（$O(N)$）で動く次世代アーキテクチャが必要とされた。

Vision Mamba (SSM)：線形計算量の革命

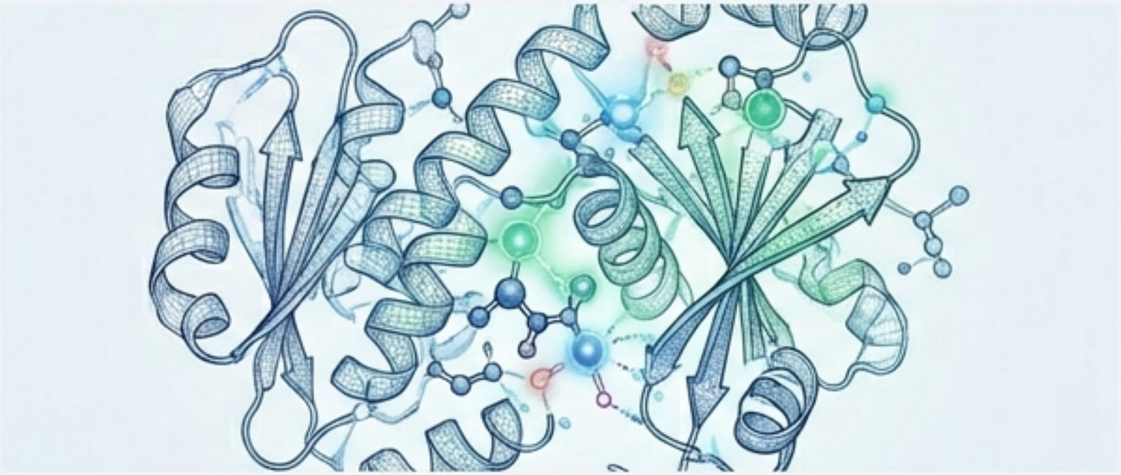
Transformerレベルの性能を、推論メモリ一定で実現

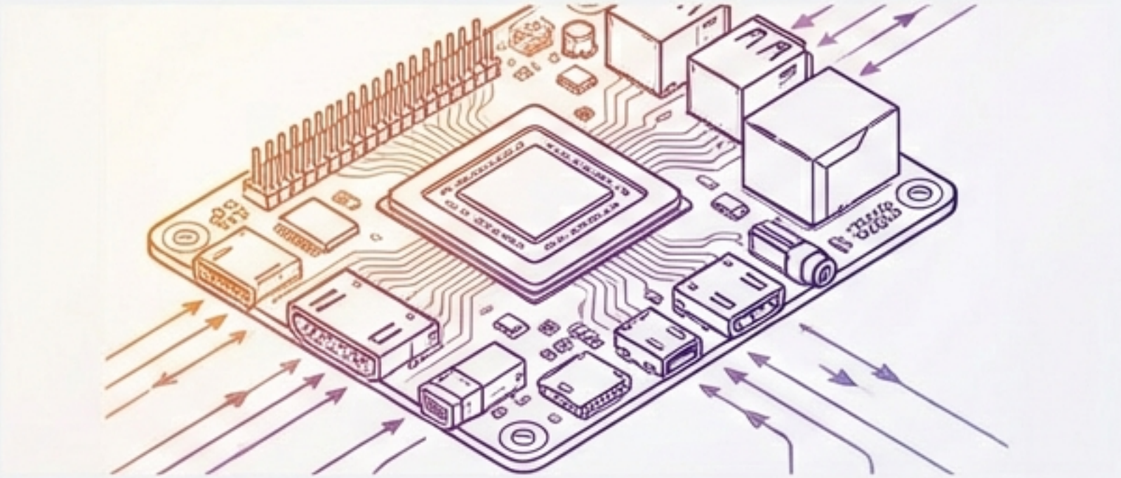


- Architecture: **状態空間モデル (State Space Models / Mamba)**。入力依存の選択メカニズムを持つRNN的な挙動。
- Advantage:
  - 学習時は並列化可能（Transformerの利点）。
  - 推論時はメモリ消費一定、計算量は線形（$O(N)$）。
- Performance: 高解像度画像（1248x1248）でDeiTより2.8倍高速、メモリ消費86.8%削減。
- Current Verdict (2026): 「MambaOut」論争。ImageNetのような短期依存タスクではViTと差がないが、超高解像度や長尺動画ではMambaが不可欠な「補完関係」にある。

科学とエッジの最前線

AlphaFold 3からRaspberry Piまで





Zone 1: Science (AlphaFold 3).
Google DeepMind (2024). **Atom Transformer**とDiffusionの融合。アミノ酸だけでできなく原子レベルの相互作用を捉え、DNA/RNAを含む全生体分子の構造を予測。AI for Scienceの基盤。

Edge AI (Jetson & Pi)

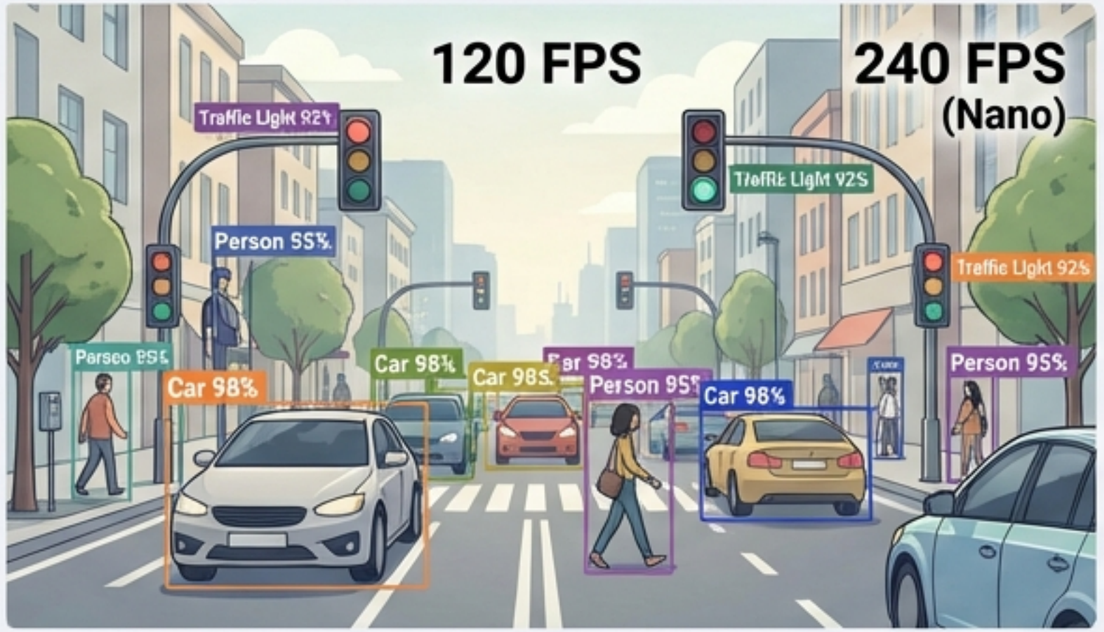
- **NVIDIA Jetson Orin**: 量子化ViTの実用化。
- **Raspberry Pi 5**: 基本は軽量CNN (MobileNet) や **YOLOv12-Nano** が主流だが、Hailo-8などのNPU追加でTransformerも動作可能に。

物体検出の現在地：YOLO vs DETR

リアルタイム性とEnd-to-End学習の競争

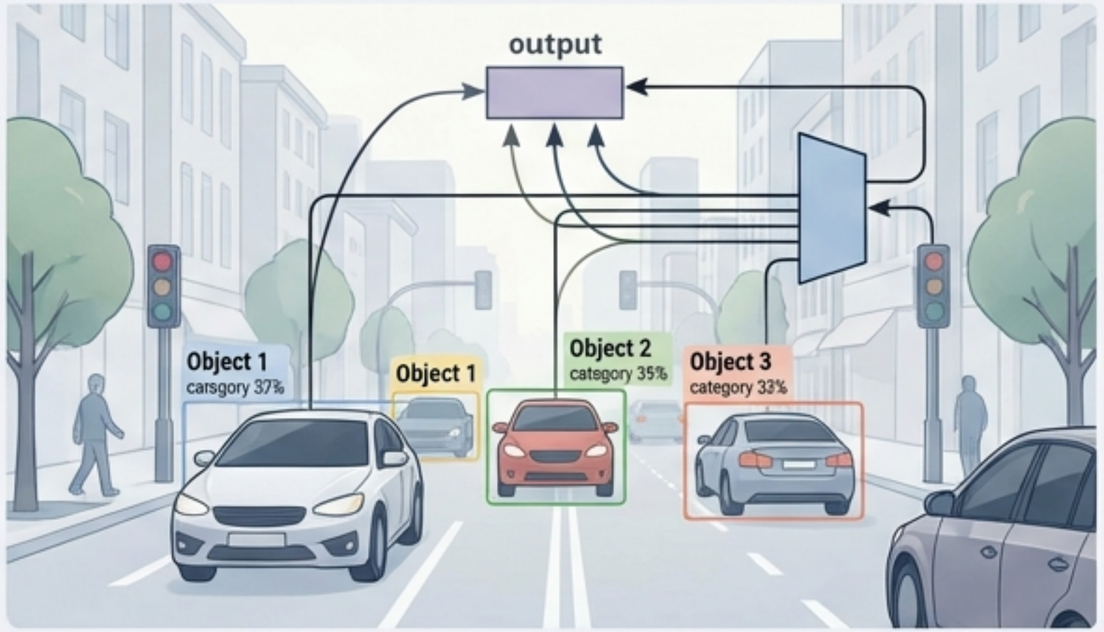
The Incumbent: YOLOv12

依然として産業界のデファクト。CNNベースの圧倒的な効率性。特にエッジ（Nanoモデル）ではTransformerのオーバーヘッドがない分、有利。



The Challenger: RF-DETR

DEtection TRansformer。物体検出を「セット予測問題」として解く。NMS不要のEnd-to-End学習。2025年、YOLOに匹敵する速度と精度（60.6 mAP）を達成。



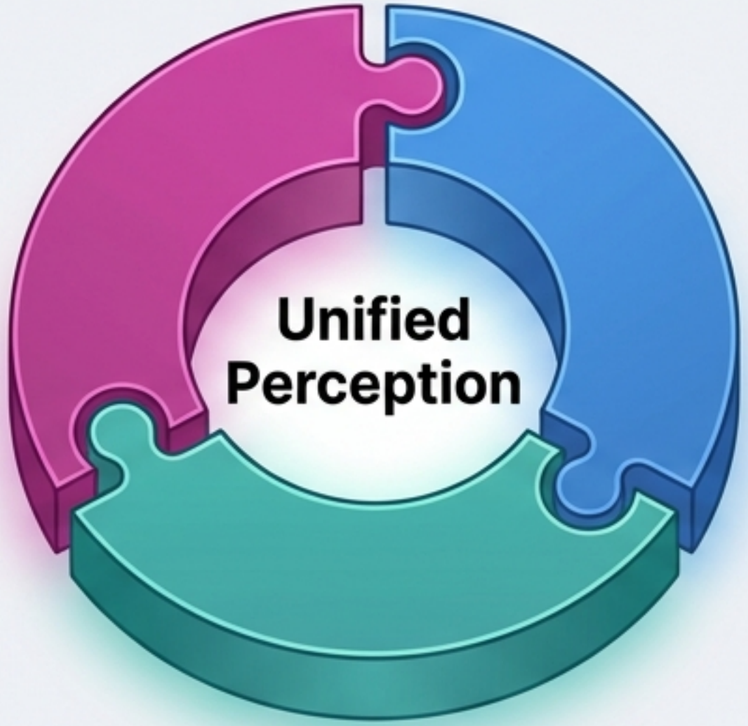
Legacy: かつてはNMS（重複除去）などの手動ヒューリスティクスが必須だったが、Transformerがそれを不要にした。

結論：統一された知覚（Unified Perception）へ

適材適所のアーキテクチャで世界を「理解」し「シミュレート」する

Transformer

High-level Understanding & Generation (Sora, AlphaFold)



CNN

Industrial Efficiency & Edge (YOLO, ConvNeXt)

Mamba

Long-Sequence Processing (3D Medical, DNA)

Final Thought: 我々は「分類（Classify）」の時代を終え、マルチモーダルな「統一された知覚」の時代に突入した。CNN、Transformer、Mambaは対立するものではなく、目的に応じて選択される「構成要素（Building Blocks）」となったのである。

The King didn’t die; the kingdom just got much, much bigger.