东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 17 期：视觉智能为何引入 Transformer？

2026 年 08 月 06 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn

投资要点

- **视觉 Transformer 化的本质，不是 ViT 对 CNN 的一次性替代，而是图像被 token 化、视觉归纳偏置回归、视觉基础模型形成的渐进过程。**CNN 之所以在 AI1.0 时代成为视觉核心架构，源于局部连接、权重共享和平移等变三项视觉归纳偏置与图像数据高度契合；但同样的强先验也带来局部感受野、任务专用、封闭标签和全局建模不足等约束，构成 Transformer 进入视觉任务的前提。

- **Transformer 进入视觉的关键，是 ViT 确立“图像 patch 即 token”的表示方式，使二维图像可被改写为一维序列，并接入自注意力和大规模预训练范式。**相较 CNN，Transformer 具备全局建模与长距离依赖更优、接口统一利于跨任务与跨模态、可扩展性强等优势，且归纳偏置较弱，在大规模预训练下更能释放数据规模价值；但视觉数据并非天然离散序列，且高分辨率、多尺度和空间先验需求更强，因此视觉演化路径与语言不同，更多表现为融合与再平衡，而非彻底替换。

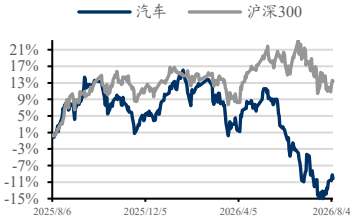
- **本报告将视觉 Transformer 演进归纳为三大阶段、九个子阶段。**第一阶段是图像被 token 化，Non-local 等注意力模块先补足 CNN 全局建模，ViT 和 DETR 进一步改写图像输入与检测输出范式；第二阶段是可扩展通用骨干，Swin、PVT、CoAtNet 等重新吸收局部性、层级结构和多尺度先验，ConvNeXt 则证明 ViT 训练经验可反哺卷积网络；第三阶段是视觉基础模型，DINOv2/DINOv3、CLIP、SAM/SAM2/SAM3 分别推动自监督通用特征、开放词表图文对齐和可提示分割。

- **从落地看，视觉智能仍主要是“感知性”中间能力。**图像分类、检测、分割、视觉特征等结果通常需要嵌入既有业务流程才能兑现价值，尚未独立形成语言模型式消费级交互入口；视觉的“互动性”更多通过 CLIP、GPT-4V、Florence-2 等视觉语言模型接入语言中枢获得。因此，以语言智能的四次跃迁（架构、规模、对齐、推理）为参照，视觉已完成架构与规模两次跃迁，对齐仅以图文对齐形式部分实现且依附于语言，推理跃迁在视觉侧尚未展现。

- **我们认为，数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**

- **风险提示：技术迭代不及预期的风险；大模型厂商 ARR 增速放缓的风险；云服务商资本开支不及预期的风险；智能驾驶及机器人商业化落地不及预期的风险。**

行业走势



相关研究

《AI 系列深度 16 期：语言智能为何从 RNN 转向 Transformer？》

2026-08-06

《重卡全球化专题 02：基建物流双轮驱动东南亚扩容，中国重卡出海进入收获期》

2026-08-06

内容目录

图表目录

1. 复盘 CNN：视觉归纳偏置如何定义视觉任务上限

1.1. 基本原理：三项视觉归纳偏置与层级特征抽象

卷积神经网络（CNN）是一类专门处理图像等网格结构数据的神经网络。与全连接网络为每个像素单独赋权不同，CNN 通过卷积核在图像上滑动，对局部区域做加权求和，从中提取边缘、纹理等基础特征；再通过池化对特征图下采样，扩大观察范围，并增强对轻微位移的鲁棒性。可以把卷积核理解为一组会移动的“滤镜”：浅层滤镜先识别边缘和纹理，深层网络再把这些局部线索组合成部件和语义，最终由全连接层或检测/分割头输出结果。

CNN 之所以高度契合图像，源于其架构内置的三项视觉归纳偏置。

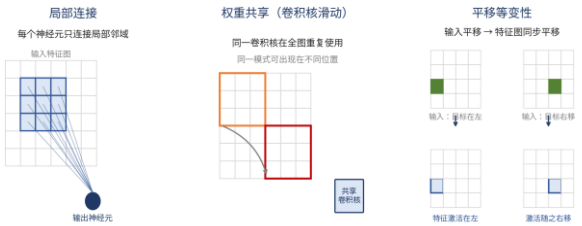
其一，**局部连接**：每个神经元只看输入的局部邻域，对应“基础视觉模式通常出现在局部”的先验。

其二，**权重共享**：同一个卷积核在整张图像上重复使用，对应“同一种局部模式可出现在不同位置”的先验，并大幅减少参数量。

其三，**平移等变性**：在理想卷积条件下，输入图像发生平移，输出特征图也随之平移。

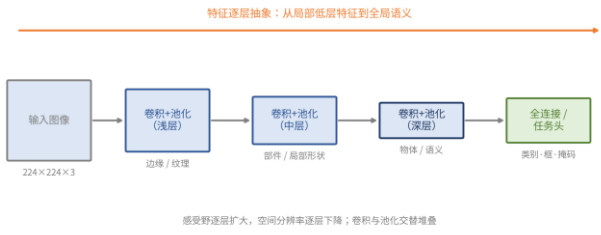
我们认为，正是这些由架构直接注入、而非完全依靠数据学习得到的空间先验，使 CNN 在数据规模有限时仍能高效学习，并奠定其在 AI 第一次爆发中的核心地位。

图1：卷积神经网络的三项视觉归纳偏置



数据来源：《Gradient-Based Learning Applied to Document Recognition》，东吴证券研究所绘制

图2：卷积神经网络的典型层级结构与特征抽象



数据来源：《Gradient-Based Learning Applied to Document Recognition》，《ImageNet Classification with Deep Convolutional Neural Networks》，东吴证券研究所绘制

1.2. 从 LeNet 到 ResNet 的演进

卷积思想可追溯至 Fukushima（1980）提出的神经认知机；LeCun 等（1998）将反

向传播与卷积结构结合，提出 LeNet-5 并成功用于手写数字识别，确立了现代 CNN 的基本范式。其后十余年，受限于数据与算力，相关进展相对平缓。2012 年，Krizhevsky 等提出的 AlexNet 在 ImageNet 大规模图像识别竞赛中大幅领先，标志 AI 第一次爆发的开端。此后，VGG（2014）用小卷积核加深网络，GoogLeNet（2014）引入 Inception 多分支结构，ResNet（2015）以残差连接缓解深层网络退化问题，将网络深度推进到上百层，并把 ImageNet 图像分类的 top-5 错误率降至约 3.5%，已低于人类标注者约 5%的 top-5 错误率水平。

1.3. 结构性局限

CNN 在图像分类等任务上取得了显著工程成功，但其能力也受到若干结构性因素约束。这些约束构成 Transformer 进入视觉任务的前提。

其一，感受野偏局部，长距离依赖建模效率较低。卷积是局部操作，单层只能覆盖邻近区域；如果要理解图像中相距较远的两个区域之间的关系，需要堆叠很多层或使用大卷积核，效率较低。这一局限直接催生了在 CNN 中引入注意力、补足长距离关系建模的早期工作。

其二，强空间先验是一把双刃剑。局部性和平移等变在数据有限时是优势，相当于提前告诉模型“图像应当这样看”；但它也限定了模型的假设空间。当训练数据规模足够大时，人工注入的强归纳偏置反而可能成为上限约束。

其三，任务专用，缺乏统一可迁移底座。CNN 时代的视觉模型大多按任务单独设计：分类、检测（如 Faster R-CNN）、分割（如 FCN、Mask R-CNN）各有专门网络结构与训练流程。跨任务复用主要依赖在 ImageNet 上有监督预训练的骨干权重，迁移能力有限。

其四，标签体系相对封闭。分类模型只能识别固定类别，检测与分割模型通常绑定特定数据集和标签集，难以根据自然语言或提示完成开放词表与零样本（zero-shot）任务。

我们认为，CNN 的三项归纳偏置赋予其高效视觉感知能力，但也通过局部感受野、强空间先验、任务专用和封闭标签四种形式限定能力边界。理解这四点，是理解 Transformer 为何进入视觉、又为何无法对 CNN 形成彻底替换的前提。

2. Transformer 爆发之前，卷积网络的应用上限

2.1. 卷积网络时代视觉任务的主要成果

在 Transformer 进入视觉之前，CNN 已将主流视觉任务推进到较高工程水平，代表性进展覆盖图像分类、目标检测、语义/实例分割等方向。在图像分类上，从 AlexNet（2012）

到 ResNet，ImageNet 基准错误率快速下降并接近饱和；在目标检测上，形成了以区域提议为代表的两阶段路线与以单阶段为代表的实时路线；在分割上，全卷积网络确立端到端语义分割范式，U-Net 在医学影像中广泛应用，Mask R-CNN 统一检测与实例分割。此外，在人脸识别、姿态估计、图像超分辨率等方向，CNN 也曾是主流方法。

表1：卷积网络在主要视觉任务上的代表方法

任务方向	代表方法	主要时期
图像分类	AlexNet、VGG、GoogLeNet、ResNet	2012—2016
目标检测（两阶段）	R-CNN、Fast/Faster R-CNN	2014—2016
目标检测（单阶段/实时）	YOLO、SSD	2016
语义/实例分割	FCN、U-Net、DeepLab、Mask R-CNN	2015—2017
图像超分辨率/低层视觉	SRCNN 及其后续	2014—2017
人脸识别与度量学习	DeepFace、FaceNet	2014—2015

数据来源：《Deep Residual Learning for Image Recognition》，《Mask R-CNN》等其他文献，东吴证券研究所

2.2. 应用上限：封闭标签、任务专用与全局建模约束

我们认为，CNN 把视觉任务推进到“卷积范式所能达到的工程高峰”，但其应用上限也由前述结构性局限决定，集中体现为四点。

第一，在有监督、单任务条件下，CNN 可在分类、检测、分割等基准上达到当时最优水平，ImageNet 分类更接近性能饱和；但能力高度依赖人工标注的封闭数据集。

第二，不同任务缺乏统一可复用底座，分类、检测、分割大多需要单独设计网络结构，迁移与复用主要依靠 ImageNet 有监督预训练骨干，难以零样本迁移到新任务和新类别。

第三，封闭标签体系限制开放性，模型只能识别训练时定义的类别，无法依据自然语言描述识别任意目标。

第四，长距离与全局关系建模能力不足，这一点至关重要：在 Transformer 进入视觉之前，研究者已开始在 CNN 中引入注意力机制以补足全局建模，说明纯卷积在全局依赖建模上已显不足，而真正的全局关联能力来自注意力。

换言之，CNN 已把视觉任务推到卷积范式的工程上限；但任务统一化、开放化和全局建模三条道路，分别受到任务专用性、封闭标签和局部感受野限制。下一代架构需要回答的问题由此明确：能否以一种统一、可大规模预训练、可建模全局关系的架构，承接并超越卷积网络。

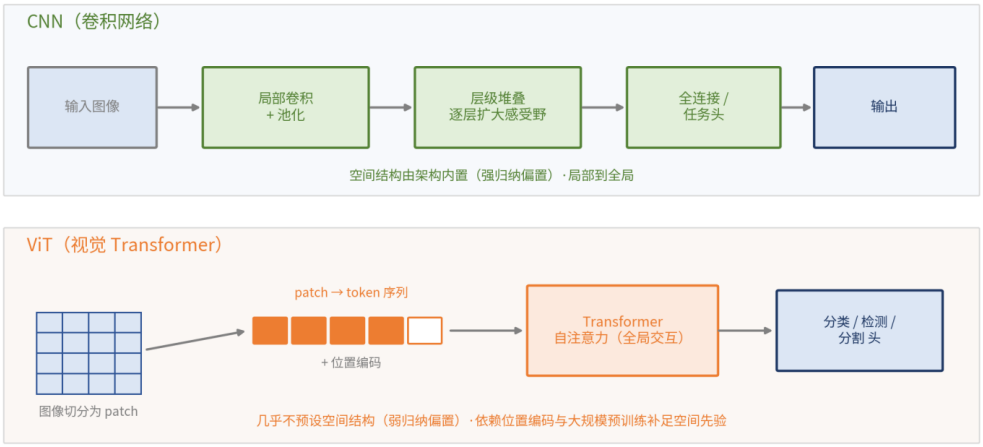
3. Transformer 与卷积网络的结构对比及其优点

3.1. 两种架构的结构对比

Transformer（2017）最初为序列建模而提出，将其应用于图像的关键，是 ViT（2020）确立的“图像 patch 即 token”范式：先把图像切分为固定大小的图块（如 16×16），再把每个图块线性映射为一个 token，并加入位置编码显式注入空间位置信息，从而把二维图像转换为一维 token 序列，输入标准 Transformer。

两种架构最根本的差异，在于空间信息如何交互、先验从哪里来。CNN 依靠局部卷积核滑动并逐层扩大感受野，空间结构由架构直接注入；Transformer 则通过自注意力机制，让每个 token 直接与所有其他 token 交互。模型为每个位置生成查询（Query）、键（Key）、值（Value），根据查询与各位置键的相似度计算注意力权重，再对值加权求和，从而在单层内建立全局关联。换言之，Transformer 几乎不预设空间结构，更多依靠数据学习图像内部关系。

图3：卷积网络（CNN）与视觉 Transformer（ViT）的结构对比



数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Attention Is All You Need》，东吴证券研究所绘制

3.2. Transformer 相对卷积网络的主要优点

基于上述结构差异，我们将 Transformer 相对 CNN 的优点归纳为四点。

第一，全局建模与长距离依赖更优。在自注意力中，任意两个位置可以在单层内直接建立联系，关联路径不随空间距离拉长；而 CNN 需要通过多层卷积逐步扩大感受野，才能覆盖全局。

第二，归纳偏置较弱，有利于释放数据规模价值。CNN 预设较强空间结构假设，Transformer 不预设强结构，主要依靠数据学习关系。在大规模预训练条件下，这更有利于发挥数据规模的价值。ViT 实验表明，在足够大的数据（如数亿量级图像）上预训练后，纯 Transformer 的图像分类性能可超过同等条件下的 CNN。

第三，接口统一，利于跨任务与跨模态。将图像表示为 token 序列后，视觉任务的输入形式与语言任务趋于一致，便于复用预训练—迁移范式，也为图像与文本在同一框架下对齐、构建多模态模型提供结构基础。

第四，可扩展性强。Transformer 结构均一、便于堆叠加深与加宽，配合并行训练，天然适配“扩大模型与数据规模”的路线，为视觉领域的规模化提供了结构载体。

但同时，上述优势也伴随明显代价，而这些代价决定了视觉的 Transformer 化无法成为一次彻底替换。

其一，缺乏视觉归纳偏置：标准 ViT 不内置局部性、二维空间邻接和平移等变，在中小数据规模下未必优于 CNN，需依赖大规模预训练。

其二，计算开销随分辨率快速上升：全局自注意力的计算量会随 token 数量增加而快速放大，高分辨率图像的 token 数量较多，直接使用全局注意力成本较高。

其三，缺乏顺序与空间先验，必须依赖位置编码补足。

表2：Transformer 与卷积网络的多维对比

对比维度	卷积网络（CNN）	视觉 Transformer（以 ViT 为代表）
空间信息交互	局部卷积，逐层扩大感受野	自注意力，全局直接交互
长距离依赖	需堆叠多层，路径随距离增长	O(1)常数级路径
归纳偏置	强（局部性、权重共享、平移等变）	弱（依赖数据学习，需位置编码）
计算复杂度	随像素近似线性	全局注意力随 token 数平方（$O(n^2)$）
空间先验来源	架构内置	依赖位置编码与大规模预训练
小数据表现	较好	较弱，依赖大规模预训练
跨任务/跨模态	任务专用，迁移有限	接口统一，利于迁移与多模态对齐
可扩展性	受结构约束	利于规模扩展

数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Attention Is All You Need》等其他文献，东吴证券研究所

我们认为，Transformer 之于 CNN 的意义，不在于提供一个“更聪明的局部滤镜”，而在于用统一、可扩展、可建模全局关系的序列框架，为视觉任务接入大规模预训练与多模态对齐打开通道。但由于 Transformer 缺少 CNN 的视觉归纳偏置，这一通道也伴随对数据规模、计算成本和空间先验的新要求，决定了视觉演化路径不是替换，而是融合与再平衡。

4. Transformer 应用于视觉任务的学术成果梳理及发展阶段总结

我们以论文与代表性学术成果为基础，梳理 2017 年以来 Transformer 在视觉任务中的学术演进，并对其发展阶段进行归纳。

4.1. 发展阶段框架

基于相关论文与代表成果复盘，我们将 Transformer 在视觉任务中的演化归纳为三个大阶段、九个子阶段。三个大阶段分别为：

1）图像被 token 化，即注意力先作为 CNN 增强模块，继而图像在输入与输出两端被改写为 token/query 序列；

2）视觉架构转向可扩展通用骨干，即 Transformer 重新吸收 CNN 的局部性、层级和多尺度，成为可适配密集任务的通用骨干，并与 CNN 融合再平衡；

3）从封闭标签分类走向视觉基础模型，即自监督预训练、开放词表对齐与可提示分割三条路线，使视觉模型走向可预训练、可迁移、可提示、可与语言对齐。

表3：Transformer 视觉任务发展三大阶段—子阶段框架

大阶段	子阶段	时间	核心变化	代表成果
一、图像被 token 化	1.1 注意力视觉化前史	2017—2019	注意力先作为 CNN 的增强模块，建模长距离依赖	Non-local 、 Image Transformer 、 Stand-Alone Self-Attention
一、图像被 token 化	1.2 图像输入 token 化	2020—2021	图像被切成 patch token，像文本一样输入	ViT、DeiT
一、图像被 token 化	1.3 输出 token/query 化	2020—2021	检测、分割从手工流程转为 token/query 输出	DETR、SETR
二、可扩展通用骨干	2.1 视觉归纳偏置回归	2021—2022	重新吸收局部性、层级结构与多尺度特征	PVT、Swin、MViT、CoAtNet、MaxViT
二、可扩展通用骨干	2.2 通用视觉骨干形成	2021—2023	成为分类、检测、分割、视频的通用骨干	Swin 、 PVT 、 ViT-Adapter
二、可扩展通用骨干	2.3CNN 与 Transformer 融合再平衡	2022—2024	CNN 借鉴 Transformer 训练范式，二者结构互吸收	ConvNeXt 、 InternImage 、 Hybrid ViT
三、走向视觉基础模型	3.1 自监督视觉预训练	2021—2026	从有监督训练转向大规模无标签预训练	DINO、BEiT、MAE、EVA 、 DINOv2 、 DINOv3
三、走向视觉基础模型	3.2 开放词表视觉模型	2021—2026	从固定类别转向自然语言监督与零样本识别	CLIP 、 ALIGN 、 Florence、SigLIP
三、走向视觉基础模型	3.3 可提示视觉基础模型	2023—2026	转向点、框、文本、mask 提示驱动	SAM、SAM2、SAM3、Grounding DINO

数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Swin Transformer: Hierarchical Vision Transformer using Shifted Windows》等其他文献，东吴证券研究所

4.2. 里程碑学术成果与论文

在该阶段框架下，我们将以下成果列为里程碑：DETR（2020，将检测改写为集合预测）、ViT（2020，确立“图像 patch 即 token”范式）、Swin Transformer（2021，以窗口注意力适配高分辨率密集任务）、MAE（2021，确立可扩展的视觉掩码自监督）、CLIP（2021，图文对齐打开开放词表与零样本）、DINOv2（2023，免微调可迁移的通用视觉特征）、SAM（2023，可提示分割与 SA-1B 数据集）。这些成果分别标志了输出范式改写、输入范式确立、骨干工程化、自监督可扩展、开放词表对齐、通用特征迁移与可提示基础模型等关键转折。

表4：Transformer 视觉任务重要成果分类表（里程碑红色加粗）

大阶段	子阶段	代表论文/成果	核心贡献
一、图像被 token 化	前史	Non-local 、 Stand-Alone Self-Attention	在 CNN 中引入自注意力，建模长距离依赖
一、图像被 token 化	输入 token 化	ViT、DeiT	确立“图像 patch 即 token”，纯 Transformer 胜任分类
一、图像被 token 化	输出 token/query 化	DETR、SETR	将检测、分割改写为 token/query 输出，端到端化
二、可扩展通用骨干	归纳偏置回归	PVT 、 Swin Transformer、MViT	引入金字塔、窗口与层级结构，适配高分辨率密集任务
二、可扩展通用骨干	融合再平衡	ConvNeXt 、 InternImage	ViT 经验反哺 CNN，二者双向借鉴
三、走向视觉基础模型	自监督预训练	DINO、BEiT、MAE、EVA 、 DINOv2 、 DINOv3	摆脱人工标签，学习可迁移的通用视觉表征
三、走向视觉基础模型	开放词表	CLIP 、 ALIGN 、 Florence、SigLIP	图文对齐，实现零样本分类、检索与开放词表识别
三、走向视觉基础模型	可提示	SAM 、 SAM2 、 SAM3 、 Grounding DINO	转向点、框、文本提示驱动的开放视觉任务

数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Segment Anything》等其他文献，东吴证券研究所

4.3. 逐阶段梳理

第一大阶段：图像被 token 化（2017—2021）。这一阶段的关键在于，Transformer 进入视觉的第一步并非直接替代 CNN，而是先把图像改造成 Transformer 能处理的 token 序列。

前史阶段（2017—2019），该阶段的工作还未真正摆脱 CNN，更多是在卷积网络中引入注意力，补足卷积难以直接建模的长距离依赖。真正的范式转折来自 ViT（2020）：它将图像切分为固定大小 patch 并映射为 token 序列，使 Transformer 第一次能以接近自然语言处理的方式处理图像，并证明在大规模预训练下纯 Transformer 可胜任图像分类；DeiT（2021）随后通过蒸馏证明，仅用 ImageNet 也能训练出有竞争力的视觉 Transformer，缓解了对超大数据和算力的依赖。

token 化不只发生在输入端：DETR（2020）将目标检测改写为基于 object query 的集合预测问题，去除非极大值抑制等手工环节；SETR 将语义分割改写为序列到序列预测。这意味着 Transformer 不仅改变图像输入形式，也开始改变视觉任务的输出范式。

第二大阶段：视觉架构转向可扩展通用骨干（2021—2024）。ViT 证明图像可以被 token 化，但原始 ViT 存在三个问题：高分辨率图像 token 数量多，全局自注意力计算量过大；视觉任务天然需要多尺度特征，而原始 ViT 结构较平坦；CNN 的局部性、层级结构和平移等变虽限制通用性，但在视觉任务中非常有效。

因此，2021 年之后，视觉 Transformer 主线不是继续追求“纯 Transformer”，而是重新吸收 CNN 中已被验证有效的视觉归纳偏置。PVT（2021）引入金字塔层级结构，使 Transformer 能像 CNN 骨干一样输出多尺度特征；Swin Transformer（2021）通过窗口注意力和移位窗口机制，在高分辨率图像上实现高效注意力，成为可服务检测、分割等密集任务的通用骨干；MViT、CoAtNet、MaxViT 等也在不同程度上融合层级结构、局部窗口和卷积思想。

与此同时，ConvNeXt（2022）把 ViT 训练范式与设计经验反哺纯卷积网络，使 CNN 重新达到与 Swin 相当的水平，说明视觉架构进入 CNN 与 Transformer 双向借鉴阶段。

我们认为，这一大阶段是视觉与语言演化路径分野的关键：语言中 Transformer 基本替换 RNN，视觉中 Transformer 则与 CNN 融合。

第三大阶段：从封闭标签分类走向视觉基础模型（2021—2026）。这一阶段指明视觉 Transformer 的最终方向：不是做一个更强分类器，而是成为可迁移、可对齐语言、可接受提示、可服务多种任务的基础模型，主要沿三条路线推进。

第一条是自监督视觉预训练：DINO（2021）发现自监督 ViT 会涌现较强物体区域感知能力，BEiT、MAE（2021）以掩码图像建模使 ViT 自监督预训练高效可扩展，EVA、DINOv2（2023）进一步以大规模无监督数据训练通用视觉特征，强调免微调即可迁移到分割、深度估计、医学影像、遥感等任务；DINOv3（2025）将自监督扩展到约 17 亿无标签图像与约 70 亿参数规模，并通过 Gram Anchoring 抑制长训练下稠密特征退化，使单一冻结骨干在检测、分割、视频跟踪等任务上首次超过专用方案。

第二条是开放词表视觉模型：CLIP（2021）以约 4 亿图文对训练图文对齐模型，把图像与自然语言映射到同一语义空间，使分类、检索、开放词表识别进入零样本与提示

范式，ALIGN、Florence、SigLIP 等沿同一方向推进。

第三条是可提示视觉基础模型：SAM（2023）提出可提示分割模型与 SA-1B 数据集（约 10 亿掩码），将分割从固定类别改造为点、框、mask 提示驱动的开放任务；SAM2（2024）扩展到视频并引入流式记忆支持实时分割；SAM3（2025）进一步引入概念提示，可依据名词短语一次性返回全部匹配实例，Florence-2、Grounding DINO 则推动文本提示、视觉定位与检测分割统一。

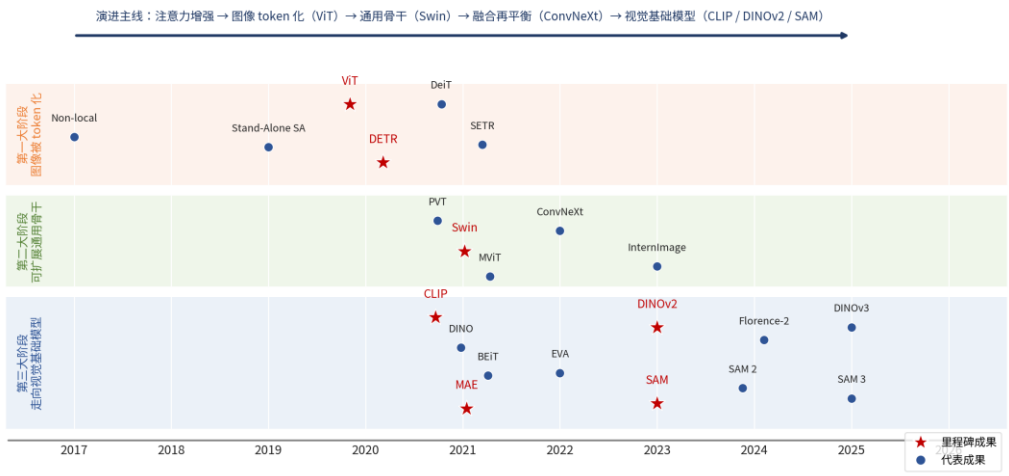
由此，视觉智能从 CNN 时代的任务专用网络，逐步演化为可预训练、可迁移、可提示、可与语言对齐的视觉基础模型。

4.4. 技术演进时间轴

从发展主线看，注意力作为 CNN 增强（2017—2019）→ViT 确立图像 token 化（2020）→DETR/SETR 改写输出范式（2020—2021）→Swin/PVT 形成通用骨干（2021）→ConvNeXt 推动融合再平衡（2022）→DINOv2/CLIP/SAM 走向视觉基础模型（2021—2023）→DINOv3/SAM3 持续扩展（2025），构成逐层递进的升级链。

从技术路线看，自监督预训练、开放词表与可提示三条线，分别从“无标签表征”“图文对齐”“提示驱动”三个方向收敛到“视觉基础模型”这一共同终点。我们认为，这条演进链的突出特征，是 Transformer 与 CNN 由对立走向融合：每一阶段都在前一阶段基础上吸收、再平衡，而非推倒重来。

图4：Transformer 在视觉任务中的技术演进时间轴（2017—2026）



数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Segment Anything》等其他文献，东吴证券研究所

5. Transformer 之后：视觉从封闭感知走向基础模型

5.1. Transformer 缺少 CNN 的视觉归纳偏置

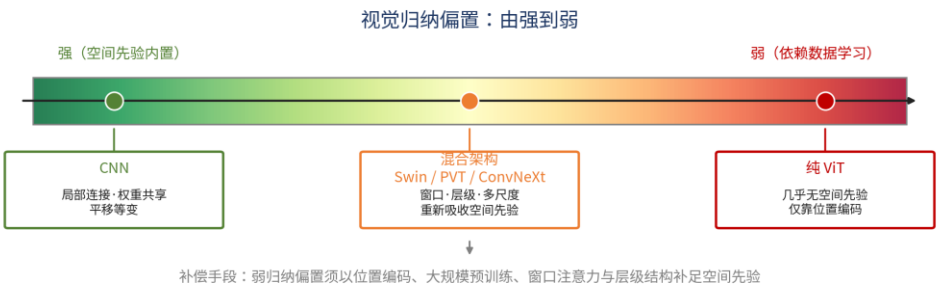
要理解 Transformer 出现之后视觉领域的变化，须先理解二者在归纳偏置上的根本差异。

CNN 的视觉归纳偏置主要来自架构设计，而非完全从数据中学得：局部连接假设基础视觉模式存在于局部邻域；权重共享假设同一局部模式可出现在不同空间位置；在理想卷积条件下进一步带来平移等变性。

Transformer 最初为文本等序列建模任务提出，处理的是一维 token 序列，而非二维图像网格。其自注意力机制善于建模 token 之间的全局关系，但不天然内置 CNN 那样的局部感受野、二维空间邻接和平移等变。

因此，Transformer 直接用于视觉任务时，相比 CNN 具有更弱的视觉归纳偏置，往往需要依赖位置编码、大规模数据预训练，或通过窗口注意力、层级结构等改造来补充空间结构先验。

图5：视觉归纳偏置的强弱谱与补偿手段



数据来源：《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》，《Swin Transformer: Hierarchical Vision Transformer using Shifted Windows》等其他文献，东吴证券研究所绘制

我们认为，这一差异是理解视觉与语言两条主线分野的关键。语言数据天然是离散一维序列，与 Transformer 输入形式高度契合，因此 Transformer 可对 RNN 形成较为彻底的替换；图像数据是二维网格，且 CNN 的归纳偏置确有工程价值，因此 Transformer 进入视觉必须以“补足空间先验”为前提，演化路径表现为补充、回归与融合，而非简单替代。

5.2. 是否所有视觉任务都已改用 Transformer

就主流与前沿水平而言，图像分类、目标检测、语义/实例分割、视频理解以及多模态等任务的最优结果，目前多由基于 Transformer 或含 Transformer 的混合模型取得。但这并不意味着所有视觉任务在所有场景下都已改用 Transformer。

其一，在低层视觉任务（如超分辨率、去噪、去模糊）中，卷积结构仍被广泛使用，即便 SwinIR、Restormer 等引入 Transformer 的方法，也多与卷积混合。

其二，在实时、边缘与移动端场景，以 YOLO 系为代表的卷积检测器仍是工程主力，ConvNeXt（2022）也证明纯卷积网络经现代化训练即可达到与视觉 Transformer 相当的精度。

其三，在数据规模有限的场景，缺乏强归纳偏置的纯 ViT 未必优于 CNN。因此，“Transformer 化”在能力前沿较充分，但在落地场景中仍是有选择的。

5.3. Transformer 是作为核心本体还是作为组件

在语言领域，Transformer 多以“核心本体”的方式被直接运用，GPT、BERT 等模型本身就是端到端承载任务的主干，输入直接是离散词元序列，无需额外特征提取前端。

视觉领域则不同：其一，图像须先切分为 patch 并线性映射为序列后才能输入，并非天然离散序列；**其二**，视觉任务输出形态多样，包括类别、检测框、分割掩码、像素等，通常需要在骨干之上叠加专门任务头；**其三**，高分辨率与多尺度需求使视觉 Transformer 多采用层级结构，并常与卷积混合。因此在视觉中，Transformer 更多扮演骨干网络（backbone）角色，被嵌入检测、分割与多模态等更大处理流程，是“组成部分之一”而非本体。

表5：Transformer 在语言与视觉任务中的角色对比

对比维度	语言任务	视觉任务
输入形式	天然离散词元序列，直接输入	图像须切分 patch 并序列化（ViT）
归纳偏置处理	弱偏置即可，无需补足空间先验	需以位置编码、窗口、层级补足空间先验
典型角色	本体/主干，端到端直接承载（GPT、BERT）	多作骨干网络组件，常与 CNN 混合
与原架构关系	基本替换循环网络	与 CNN 融合再平衡，CNN 未被淘汰
多模态中的定位	常作统一接口与中枢	作为视觉编码器被语言中枢接入
本体化程度	高（最为彻底）	中（以组件化、混合为主）

数据来源：《Attention Is All You Need》，《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》等其他文献，东吴证券研究所

6. Transformer 在语言、图像任务发展的对比

若以语言智能中 Transformer 的四次跃迁（架构、规模、对齐、推理）观察，视觉主线呈现出与语言不同的完成度。

架构跃迁对应 ViT 确立图像 token 化（2020），视觉获得统一序列建模接口；**规模跃迁**对应大规模预训练（ViT 的数亿级图像、CLIP 约 4 亿图文对、DINOv2/DINOv3 的大规模数据），在视觉领域印证规模化价值；**对齐跃迁**在视觉中主要以“图文对齐”的形式部分实现，且依附于语言；**推理跃迁**在视觉领域则尚未展现。

7. 风险提示

技术迭代不及预期的风险：数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

大模型厂商 ARR 增速放缓的风险：若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

云服务商资本开支不及预期的风险：若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

智能驾驶及机器人商业化落地不及预期的风险：若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn