东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 07 期：CNN 与 ViT 两类视觉骨干有何差异？

2026 年 07 月 30 日

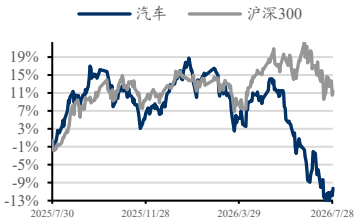
增持（维持）

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

投资要点

- **视觉骨干网络是机器视觉系统的核心模块，其作用是把原始像素转化为高层语义特征，并决定分类、检测、分割等下游任务上限。**CNN 与 ViT 是过去十余年最重要的两代视觉骨干，分别代表“将图像先验写入结构”和“将关系学习交给数据与规模”两种设计哲学；二者之争本质上是先验、数据和算力如何分工。
- **视觉架构大体经历四个阶段：**手工特征时代依赖 SIFT、HOG 等专家设计；CNN 崛起后，AlexNet、VGG、ResNet 推动机器自动学习视觉特征；ViT 通过把图像切成 patch 并当作 token 输入 Transformer，使注意力机制进入视觉；2021 年后，Swin、CoAtNet、ConvNeXt 等推动 CNN 与 ViT 相互吸收，进入融合时代。
- **CNN 的优势来自三项视觉归纳偏置：**局部连接、权值共享和平移等变性。这些结构假设使模型在数据有限时更高效，也更适合端侧和实时部署；约束在于局部感受野逐层扩大，长程依赖建模较弱。ViT 以自注意力实现第一层全局交互，弱先验、强规模化，在大数据预训练下上限更高，但对数据、算力和位置编码依赖更强。
- **核心争议在于 ViT 是否真正“强于”CNN，还是更多受益于更大数据和现代训练范式。**ConvNeXt 证明纯卷积配合现代训练技巧仍可追平部分 ViT，MLP-Mixer 也提示注意力未必是唯一变量；因此更准确的理解是，视觉架构正在从“强先验”与“弱先验”两端走向任务、数据、算力约束下的动态平衡。
- **产业落地呈现分层：**多模态大模型和研究前沿普遍采用 ViT 或类 Transformer 视觉编码器，CLIP、SigLIP、Qwen-VL 等均沿此方向推进；自动驾驶量产系统则受车端算力、时延和安全约束影响，更多采用 CNN 前端、BEV/Transformer 融合和混合骨干。研究前沿偏 ViT，车端量产偏混合，是当前现实分工。
- **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 06 期：Transformer 如何从序列任务走向基座模型？》

2026-07-30

《具身智能系列 01：本体厂商资本化加速，通用机器人产业链进入验证期》

2026-07-29

内容目录

图表目录

1. 视觉骨干网络：机器视觉的表征底座

视觉任务的核心，是将原始像素编码为可被分类、检测、分割等任务调用的高层语义特征。视觉骨干网络承担这一编码过程，是机器视觉系统的表征底座。CNN 与 ViT 是获得视觉表征的两类架构路径。CNN 将局部性、权值共享和平移等变性等视觉先验写入结构；ViT 通过自注意力将图像表示为 token 序列，并依赖数据规模学习空间关系。

视觉骨干网络是计算机视觉系统中将原始像素转化为高层语义特征的核心模块。图像在计算机中首先表现为像素矩阵，模型需要将低层数值信号转化为可表达物体、边界、位置和语义关系的特征，再交由分类、检测或分割任务头使用。因此，骨干网络的表征质量直接决定下游任务上限。

图像识别的根本难点在于语义鸿沟：低层像素变化与高层语义含义之间并不一一对应。同一对象在视角、光照、遮挡、尺度变化下像素差异可能很大，而不同类别对象在局部纹理上又可能相似。视觉模型既需要对平移、缩放、形变和光照变化保持稳定，又需要保留区分类别和实例的关键差异。

计算机视觉主流任务可分为图像分类、目标检测、语义分割与实例分割等层次，难度依次提升。它们通常共享“骨干网络+任务头”的范式：骨干网络负责提取通用视觉特征，任务头负责输出类别、位置或像素级预测。因此，CNN 与 ViT 之争本质上影响的是视觉系统的通用表征底座，而不只是单一任务模型选择。

2. 视觉架构演进：从手工特征到融合骨干

2.1. 手工特征时代：依赖专家设计视觉描述子

在深度学习兴起前，图像特征主要依赖人工设计的算子。SIFT 通过关键点描述子提升对尺度和旋转变化的稳定性，HOG 通过方向梯度直方图刻画局部形状并常用于行人检测，再配合 SVM 等浅层分类器完成识别。2010—2011 年 ImageNet 大规模视觉识别挑战赛获胜方案多属于这一类。

手工特征的主要约束在于依赖专家经验、跨任务迁移能力弱、难以随数据规模扩展。在特定场景中调优后的特征，换到新类别、新光照或新背景下往往需要重新设计。随着 ImageNet 等大规模数据集出现，人工特征的表达上限迅速暴露，深度网络自动学习特征成为视觉架构演进的核心方向。

2.2. CNN 崛起：深度网络自动学习视觉特征

AlexNet 是 CNN 成为视觉主流架构的关键转折。2012 年，Krizhevsky、Sutskever 与 Hinton 提出的 AlexNet 在 ImageNet 上将 Top-5 错误率显著降低，核心变化在于特征不再由人工设计，而由深度卷积网络从海量图像中自动学习。此后，CNN 成为分类、检测、

分割等任务的主流视觉骨干。

2014—2020 年，CNN 沿深度、宽度和结构效率持续演进。VGG 通过连续 3×3 小卷积堆叠提升深度，GoogLeNet/Inception 通过多尺度并行模块提升效率，ResNet 通过残差连接缓解深层网络训练困难，将网络深度推向百层以上。该阶段 CNN 在自动驾驶、安防、医疗影像等视觉应用中广泛落地，形成视觉骨干的产业基础。

2.3. ViT 登场：注意力机制进入视觉

ViT 将 Transformer 引入视觉任务。2020 年，Dosovitskiy 等提出 Vision Transformer，将图像切分为固定大小 patch，并将 patch 投影为 token 序列，再输入标准 Transformer 编码器。在谷歌内部 JFT-300M 大规模数据集上预训练后，ViT 分类精度超过当时强 CNN 模型，验证了弱视觉先验、强数据规模路径的可行性。

ViT 成功的关键前提是数据规模。若仅在 ImageNet-1k 等中等规模数据集上从头训练，ViT 通常不优于同级别 CNN；在 ImageNet-21k、JFT-300M 等更大规模数据上预训练后，其优势才逐步显现。由此可见，ViT 的核心特征是在大数据和大模型预训练下具备更高可扩展性。

表1：视觉架构演进的四个时代

时代	标志性工作	表征方式	主要局限
手工特征时代，—2012	SIFT、HOG + SVM	人工设计的特征算子	依赖专家经验、难以规模化
CNN 崛起，2012—2020	AlexNet、VGG、ResNet	卷积自动学习局部特征	感受野受限、长程依赖偏弱
ViT 登场，2020—2022	ViT、DeiT	自注意力建模全局关系	数据饥渴、算力开销大
融合时代，2022—今	Swin、ConvNeXt、CoAtNet	局部性与全局注意力融合	设计空间复杂、缺统一范式

数据来源：东吴证券研究所整理

3. CNN 机制：卷积为何适合图像

3.1. 核心机制：卷积、池化与任务头

卷积是 CNN 的核心运算。卷积核在图像局部区域滑动并进行加权求和，用于检测边缘、颜色、纹理等局部模式；浅层卷积通常捕捉低级视觉特征，深层卷积进一步组合成部件、物体和场景语义。多组卷积核形成多张特征图，记录不同视觉模式在空间位置上的响应强度。

池化负责降采样和增强局部稳定性。通过在局部区域内保留最大值或平均值，池化

降低特征图分辨率和计算量，同时提高对微小平移的鲁棒性。尽管现代网络中池化形式有所变化，但降低空间维度、扩大感受野和控制计算开销仍是视觉骨干的重要设计目标。

在骨干网络末端，全连接层或任务头将高层视觉特征转化为具体输出，例如分类概率、检测框或像素级标签。现代检测和分割网络中，全连接层常被卷积结构或专门任务头替代，但“骨干提取特征、任务头完成预测”的主线保持不变。

3.2. 三项设计先验：局部连接、权值共享、平移等变性

局部连接体现了图像的空间局部性假设。CNN 让单个神经元只连接局部区域，而不是直接连接全图，从而显著降低参数量，并使模型优先学习边缘、纹理、角点等局部结构。该假设符合图像中局部像素相关性较强的特征，也是 CNN 数据效率较高的重要原因。

权值共享意味着同一个卷积核在整张图上复用同一组参数。一个局部模式的检测器可在不同空间位置生效，既降低参数规模，也使模型具备跨位置泛化能力。相比全连接网络，权值共享是 CNN 在视觉任务中更高效的关键结构设计。

平移等变性来自卷积核在空间上的滑动复用。当目标在图像中发生位置平移时，对应特征响应也随之平移，使模型更容易识别位置变化下的同一视觉模式。局部连接、权值共享和平移等变性共同构成 CNN 的归纳偏置，即在训练前写入模型结构的视觉先验。这也是 CNN 在中小数据规模下仍具备较好表现的重要原因。

4. ViT 机制：将图像 patch 化为 token 序列

4.1. 前置概念：Self-Attention 机制

自注意力是 Transformer 的核心机制。每个输入元素生成 Query、Key、Value 三个向量，通过 Query 与所有 Key 计算相似度获得注意力权重，再按权重汇总 Value 信息。由此，任意两个元素即使相距较远，也可以在一层中建立直接关系。这一机制使模型能够动态选择当前输入中最相关的信息。

自注意力与卷积的核心差别在于交互范围和权重生成方式。卷积主要处理局部窗口，卷积核参数训练后固定；自注意力从第一层即可进行全局交互，且注意力权重随输入内容动态变化。因此 Transformer 更擅长捕捉长距离关系，但标准自注意力计算量随 token 数量呈平方增长，分辨率提升会显著增加计算开销。

4.2. 核心设计：patch 切分、token 化与 Transformer 编码

ViT 将图像切分为固定大小 patch，并将每个 patch 拉平、投影为向量，作为视觉 token 输入 Transformer。一张 224 × 224 图像若按 16 × 16 切分，将形成 196 个 patch token，并通常加入用于汇总全图信息的分类 token。通过这一转换，图像被表示为 token 序列，

从而可直接复用 Transformer 编码器。

由于自注意力本身不包含空间顺序信息，ViT 需要为每个 token 加入位置编码，以保留 patch 在原图中的位置信息。随后，token 序列进入标准 Transformer 编码器，通过多层自注意力和前馈网络形成全局视觉表征。与 CNN 相比，ViT 对图像结构的专门假设更少，更依赖数据规模学习局部性和空间关系。

ViT 的代价与红利均来自弱视觉先验。由于模型几乎不内置局部性和平移等变性，小数据条件下 ViT 需要自行学习这些规律，表现往往不及 CNN；但在大规模预训练下，弱先验减少了结构约束，使模型能够从数据中学习更灵活的关系。ViT-H/14 在 JFT-300M 预训练后取得较高 ImageNet 精度，体现其规模化潜力。

5. CNN 与 ViT：先验、数据与部署的四组分歧

CNN 与 ViT 的差异主要体现在感受野、归纳偏置、数据效率和部署约束四个维度。核心问题在于，视觉规律在多大程度上由结构先验提供，又在多大程度上由数据和训练过程学习。

5.1. 感受野：逐层扩大与首层全局交互

CNN 的感受野自下而上逐层扩大。浅层神经元主要处理局部区域，随着层数加深和空间降采样，模型逐步获得更大的全局视野。该路径层级清晰、数据效率较高，但建模远距离关系通常需要更深网络或额外结构。

ViT 从第一层自注意力开始即可实现全局交互，每个 patch 可直接与所有其他 patch 建立关系。Raghu 等在 2021 年研究中发现，ViT 在较浅层即可同时利用局部与全局信息，而 CNN 浅层更多集中于局部信息。这是两类架构最本质的结构差异之一。

5.2. 归纳偏置：强先验与数据驱动

CNN 具有较强归纳偏置，局部连接、权值共享和平移等变性在训练前已写入结构。这些先验提升了样本效率和训练稳定性，使 CNN 在中小数据集、端侧视觉任务和安全攸关场景中仍具竞争力。

ViT 视觉先验较弱，主要依靠 patch 切分和位置编码，并将图像规律交由数据学习。弱先验意味着小数据下样本效率较低，但在大规模数据和算力支撑下更具扩展空间。由此形成“先验是资产还是约束”的核心争议。

5.3. 数据效率与规模化特性

在中小数据规模下，CNN 通常更具优势。ImageNet-1k 从头训练时，CNN 与 ViT 精度接近，部分 CNN 仍可略优，原因在于 CNN 结构先验降低了数据需求。数据越有限，

局部性、权值共享和平移等变性带来的样本效率越重要。

在大规模预训练条件下，ViT 的扩展性更强。随着预训练数据从 ImageNet-1k 扩大到 ImageNet-21k、JFT-300M，ViT 性能持续提升，而部分 CNN 更早进入平台期。ViT 能够吸收更多数据、参数和算力，与大模型时代的 Scaling Law 更为一致，这也是多模态大模型视觉编码器偏向 ViT 的重要原因。

5.4. 计算复杂度与部署场景

标准自注意力计算量随 token 数量呈平方增长，处理高分辨率图像时开销较大；卷积计算量随分辨率近似线性增长，算子规则、硬件支持成熟。Swin Transformer 等改进通过局部窗口注意力将复杂度降回近似线性，正是为了解决高分辨率视觉任务中的算力约束。

部署层面，CNN 在 GPU、车载 SoC、手机 NPU 等端侧硬件上的算子支持、量化和加速工具链更成熟，延迟和功耗更可控；纯 ViT 部署生态仍在追赶。现实分工通常表现为：云端训练、研究前沿和多模态大模型偏向 ViT 或类 Transformer 视觉编码器，端侧量产和实时系统更多采用 CNN 或 CNN 与 Transformer 的混合架构。

表2：CNN 与 ViT 四组路线分歧对照

对比维度	CNN	ViT
感受野	逐层扩大，深层才全局	第一层即全局
归纳偏置	自带先验假设：侧重局部特征、权重共享、平移不变性	弱先验，数据驱动
数据效率	小数据集表现更好，训练效果容易提前到达瓶颈	对数据量要求高，数据规模越大，性能提升空间越明显
计算复杂度	随分辨率线性，硬件友好	自注意力随分辨率平方
长程依赖	需堆深度，相对困难	天然擅长
部署成熟度	端侧工具链成熟	部署配套工具还在持续完善，生态不及 CNN

数据来源：东吴证券研究所整理

6. 核心争议：架构优势与约束边界

6.1. ViT 更强，还是数据和训练范式贡献更大

关于 ViT 是否真正强于 CNN，核心质疑在于早期优势是否来自更大数据和更现代训练策略。ViT 首次超越 CNN 时依赖 JFT-300M 等私有超大数据集；ConvNeXt 进一步证明，在纯卷积架构中引入现代训练技巧、数据增强、正则化和结构调整后，CNN 仍可追平部分 ViT 表现。因此，早期 ViT 优势不能完全归因于注意力结构本身，数据规模和

训练范式同样是关键变量。

支持 ViT 的一方则强调，自注意力的全局建模能力和可扩展性具有结构性价值，越是在超大规模、多模态场景中越能体现。中等规模数据集上 CNN 与 ViT 表现接近，并不意味着二者上限一致。更审慎的判断是：在同等数据与训练条件下，CNN 与 ViT 精度可能相近，架构选择取决于数据规模、算力条件、部署约束和任务类型。

6.2. 归纳偏置：资产还是约束

数据稀缺时，归纳偏置是资产。工业质检、医疗影像、车载长尾等场景标注数据有限，CNN 内置的局部性、权值共享和平移等变性可提升样本效率与稳健性。数据越有限，结构先验带来的训练稳定性和可部署性价值越高。

数据和算力充足时，强先验可能成为上限约束。过强的结构假设可能限制模型从数据中学习更灵活的关系。Swin、CoAtNet、ConvNeXt 等架构的共同方向，是在保留必要视觉先验与释放数据学习空间之间寻找平衡。由此看，视觉骨干并不存在统一最优解，关键在于任务、数据、算力和部署约束的匹配。

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。

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