arXiv:2601.03392v1 [cs.CV] 6 Jan 2026

更好，但仍不够：检验视频人工神经网络能否匹配猕猴IT动态

Matteo Dunnhofer Christian Micheloni 约克大学，多伦多，加拿大 乌迪内大学，意大利 乌迪内大学，意大利
christian.micheloni@uniud.it matteo24@yorku.ca

Kohitij Kar 约克大学，多伦多，加拿大k0h1t1j@yorku.ca

摘要

在静态图像上训练的前馈人工神经网络（人工神经网络）仍是灵长类动物腹侧视觉通路的主流模型，但其本质上仅限于静态计算。灵长类动物所处的世界是动态的；猕猴腹侧视觉通路，尤其是下颞叶皮层，不仅支持物体识别，在观看自然视频时还会编码物体运动速度。*IT*的时间响应是否仅仅反映了时间展开的前馈变换、带有浅层时间池化的逐帧特征，还是体现了更丰富的动态计算？我们通过比较猕猴在观看自然视频时的*IT*反应与静态、循环和基于视频的人工神经网络模型，对此进行了检验。视频模型在神经预测性方面取得了小幅提升，尤其是在较晚的反应阶段，这引出了一个问题：它们捕捉的是何种动态？为探究这一点，我们进行了压力测试：在自然视频上训练的解码器被用于评估“无外观”变体，这些变体保留了运动，但去除了形状和纹理。*IT*群体活动能够泛化到这种操控，但所有类别的人工神经网络都未能做到。因此，当前的视频模型更擅长捕捉外观依赖的动态，而非*IT*所表现出的外观不变的时间计算，这凸显了开发能够编码生物时间统计特性与不变性的新目标的必要性。

1. 引言

过去十年来，在大型图像数据集上训练的前馈人工神经网络（人工神经网络） [8, 13] 已成为灵长类动物腹侧视觉通路的主流模型 [5]。 [10]这些网络能够近似V4 [1] 区和下颞叶皮层的反应[12, 20, 21], , 其内部表征支持线性读出物体类别和位置，其表现与识别任务中的行为成绩相符 [17, 20]。这一成功进一步强化了这样的观点：前馈图像计算或许足以解释核心物体视觉。然而，生物学上的腹侧通路并非处理静态图像——它嵌入在一个动态世界中，物体会移动、改变姿态并彼此互动。

越来越多的证据表明，猕猴下颞叶皮层支持的计算不止于静态形态识别 [11]。尤其是，IT神经元能够编码物体运动和速度，这表明腹侧皮层直接参与动态视觉 [18]。这些发现揭示了一个关键的科学空白：如果在静态图像上训练的前馈人工神经网络能够解释IT对静态物体的反应，那么是什么解释了IT处理自然视频的能力？IT的时间结构仅仅是前馈计算在“时间展开”后的延伸——即由浅层时间池化连接起来的逐帧特征——还是反映了当前模型未能捕捉到的、更丰富的动态机制？

近年来，人工神经网络建模取得了进展，引入了循环机制和视频训练方案，至少从表面上看，这些进展使模型更接近生物现实 [14, 16]。循环网络会随时间反复利用特征激活，而针对动作识别 [2, 4, 7, 15] 或目标跟踪 [3, 6, 19]等任务训练的视频人工神经网络，则能够提取跨帧的时间规律。在我们的研究中，其中一些模型在预测观看视频时的IT活动方面表现出了一定程度的提升。然而，这些模型捕捉的是何种时间信号，仍不清楚。它们主要对受外观制约的瞬态变化敏感，例如不断变化的纹理或姿态吗？还是近似实现了IT所表现出的外观不变的运动计算？这是本研究所要解决的一个范围较小但至关重要的空白。

我们直接比较了猕猴在观看短时自然视频（300毫秒；18帧）期间的IT反应与多种

人工神经网络模型，包括静态前馈、循环和基于视频的架构。我们评估了帧级预测能力和视频级解码，以检验模型是否捕捉到IT中观察到的动态时间响应。

视频人工神经网络相较于静态和循环基线模型取得了有限的提升，尤其是在较晚的响应阶段，这表明时间训练目标在一定程度上促使模型与IT动态对齐。然而，这些提升也引出了一个更深层的问题：它们反映了哪些底层计算？为探究这一问题，我们根据近期的神经生理学观察设计了一项压力测试 [9, 18]：在自然视频上训练的解码器接受了“无外观”变体的评估，这些变体保留了物体运动轨迹，却移除了纹理和形状线索。

猕猴IT神经元群体活动在这一操控下仍能稳健地泛化。即使去除外观信息，仍可从IT中解码出运动方向和速度。与此形成鲜明对比的是，所有类别的人工神经网络——包括视频人工神经网络——都未能通过这项压力测试。尽管这些模型能够追踪自然视频片段中的运动，但当只剩下运动信息时，其表征便会崩溃。此外，无论是按时间展开的前馈序列，还是简单的循环特征，都无法复现IT中观察到的时间响应结构，这表明当前架构缺乏必要的不变性。

我们的结果呈现出分层的图景。一方面，视频训练确实在使人工神经网络特征与下颞叶皮层晚期响应相一致方面带来了适度提升，凸显了动态建模所取得的进展。另一方面，在去除外观信息的视频上的失败揭示了一个根本性局限：当前模型能够捕捉依赖外观的时间信号，却无法捕捉下颞叶皮层所支持的外观不变的运动计算。这些发现表明，下颞叶皮层动态不能被近似为前馈计算的简单扩展。要构建与生物学相一致的视频模型，需要新的目标函数和架构约束，以编码下颞叶皮层神经元群体中观察到的时间统计特性与不变性，从而超越静态监督和浅层时间池化。

2. 结果

如上所述，核心问题在于，下颞叶皮层对自然视频的时间响应，能否归结为时间展开的前馈变换，还是反映了更丰富的动态过程。我们记录了两只猴在严格的注视控制下被动观看 300 毫秒视频时，131 个下颞叶皮层记录位点的活动，以确保获得纯粹的视觉响应。随后，我们将 IT动态与静态前馈、循环和视频训练的人工神经网络的特征进行了比较。分析分两步进行：首先，测量随时间变化的神经预测性，以检验模型是否有所提升；其次，使用“无外观”视频进行压力测试，以区分真正的动态计算与受外观限制的信号。

2.1. 动态视觉中前馈人工神经网络与视频人工神经网络的下颞叶皮层预测能力比较

我们首先评估了，在自然视频观看过程中，视频训练的人工神经网络是否比静态前馈模型更能解释下颞叶皮层反应。图1A–C总结了这一比较。前馈模型沿时间展开后，能够解释下颞叶皮层响应方差中的很大一部分，尤其是早期神经反应（~90–180毫秒）。然而，对于较晚的下颞叶皮层响应（~480–570毫秒），其预测能力明显下降。视频人工神经网络在这一晚期时间窗内表现出适度但可靠的提升，表明时间训练目标在捕捉下颞叶皮层的延展动态方面具有一定益处。循环模型的表现与前馈网络相当，这说明仅凭浅层循环并不能弥合这一差距。

至关重要的是，单神经元分析（图1B–C）显示，尽管部分下颞叶皮层神经元同样能够被静态人工神经网络和视频人工神经网络良好地解释，但有一部分神经元在其较晚的响应阶段由视频人工神经网络提供了更好的解释，而在其较早的响应阶段则由静态人工神经网络解释得更好。这一模式表明，视频训练使模型特征与IT动态的后期阶段实现了部分对齐，但并未完全解释下颞叶皮层神经元群体中时间响应的异质性。

2.2. 视频人工神经网络而非猕猴IT无法在缺乏物体外观信息的情况下区分运动方向

正如其改进的后期预测能力（第2.1节）所预期的那样，在自然视频上进行测试时，视频人工神经网络在支持运动方向解码方面比前馈网络和循环网络更加准确（图2B）。这一模式与猕猴下颞叶皮层的情况相似：在猕猴下颞叶皮层中，后期响应携带的运动信息也比早期响应更强 [18]。因此，时间训练似乎使人工神经网络与IT动态之间实现了某种程度的对齐。

然而，关键的压力测试揭示了显著的分歧。当在无外观变体（类似于 [9]）上评估基于自然视频训练的解码器时，猕猴IT仍保持稳健的运动方向解码准确率（图2D，左），体现出其能够跨外观操纵进行泛化。相比之下，视频人工神经网络的表现则降至随机水平（图2D，右；IT解码准确率与模型解码准确率的t检验 p 值 < 0.001）。尽管人工神经网络表征在自然刺激上的表现有所提升，但一旦剥除外观线索，它们便无法支持运动辨别。

这一结果表明，尽管 IT动态编码了外观不变的运动，当前的人工神经网络模型仍受逐帧外观变化所限，暴露出一项关键

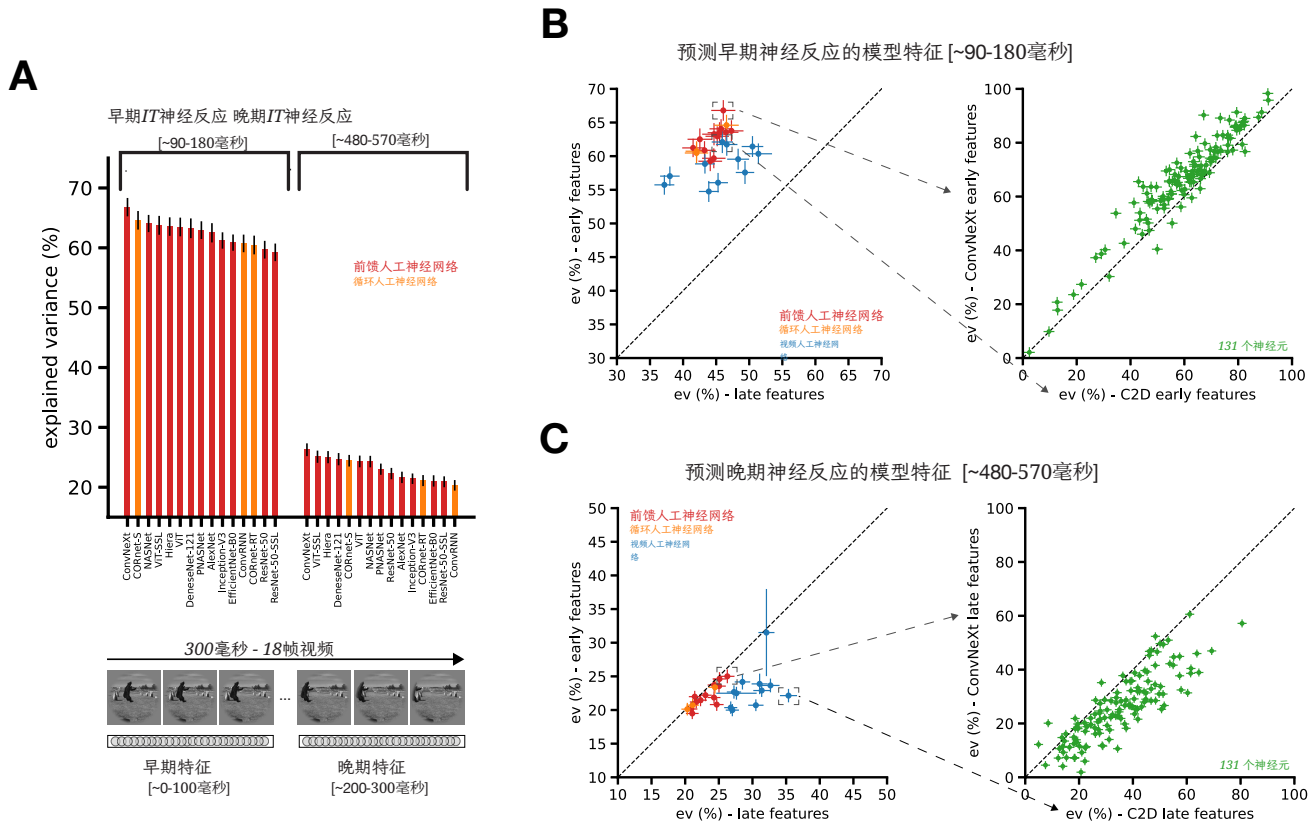


图1. **A**。静态前馈人工神经网络和循环人工神经网络从早期和晚期视频帧中提取的早期和晚期特征的神经预测性。报告跨神经元的均值和标准误。**B**。（左）早期和晚期静态与视频人工神经网络对早期神经反应的神经预测性。报告跨神经元的均值和标准误。（右）在早期反应中，最佳静态模型和视频模型对每个神经元的解释方差。报告跨重复次数的均值和标准误。**C**。（左）早期和晚期静态与视频人工神经网络对晚期神经反应的神经预测性。报告跨神经元的均值和标准误。（右）在晚期反应中，最佳静态模型和视频模型对每个神经元的解释方差。报告跨重复次数的均值和标准误。

局限：现有的视频架构和训练方式存在不足。

3. 讨论

我们的结果表明，当前的视频人工神经网络虽然在自然视频观看过程中略微提升了对下颞叶皮层的预测能力，但仍未能捕捉猕猴下颞叶皮层中观察到的全部时间计算。视频训练目标带来的晚期阶段（~480-570 毫秒）增益表明，模型对时间统计特性具有一定敏感性；但在无外观条件下的灾难性失效（图2D，右图）揭示，这些模型仍依赖逐帧外观线索，而未能编码外观不变的运动动态。

除了这一总体结论之外，我们认为，未来的分析应旨在识别IT反应的多种动态属性，以便将其作为建模的定量约束。首先，应测量神经元随时间改变反应的速度，从而揭示IT神经元群体在时间稳定性与波动性方面的异质性。

其次，量化活动主要受物体身份调节的神经元，与同时编码物体身份及其动态特征的神经元各自所占的比例。第三，刻画时间反应曲线的多样性，这或许能揭示：视频展开时，一些IT神经元维持稳定信号，而另一些则发生系统性变化。对于基于机器学习的计算机视觉，这可以提供一份路线图。仅在大型动作识别数据集上训练视频人工神经网络，虽能使其趋近IT，但无法充分复现其表征结构。将实测的时间属性纳入模型设计或优化——例如神经元可靠性的分布、物体与运动的混合选择性普遍程度，以及时间响应的群体层面几何结构——或可推动新一代生物学对齐架构的发展。这些约束可以通过辅助目标、架构偏置或动态正则化项引入，促使模型遵循IT中体现的时间不变性。

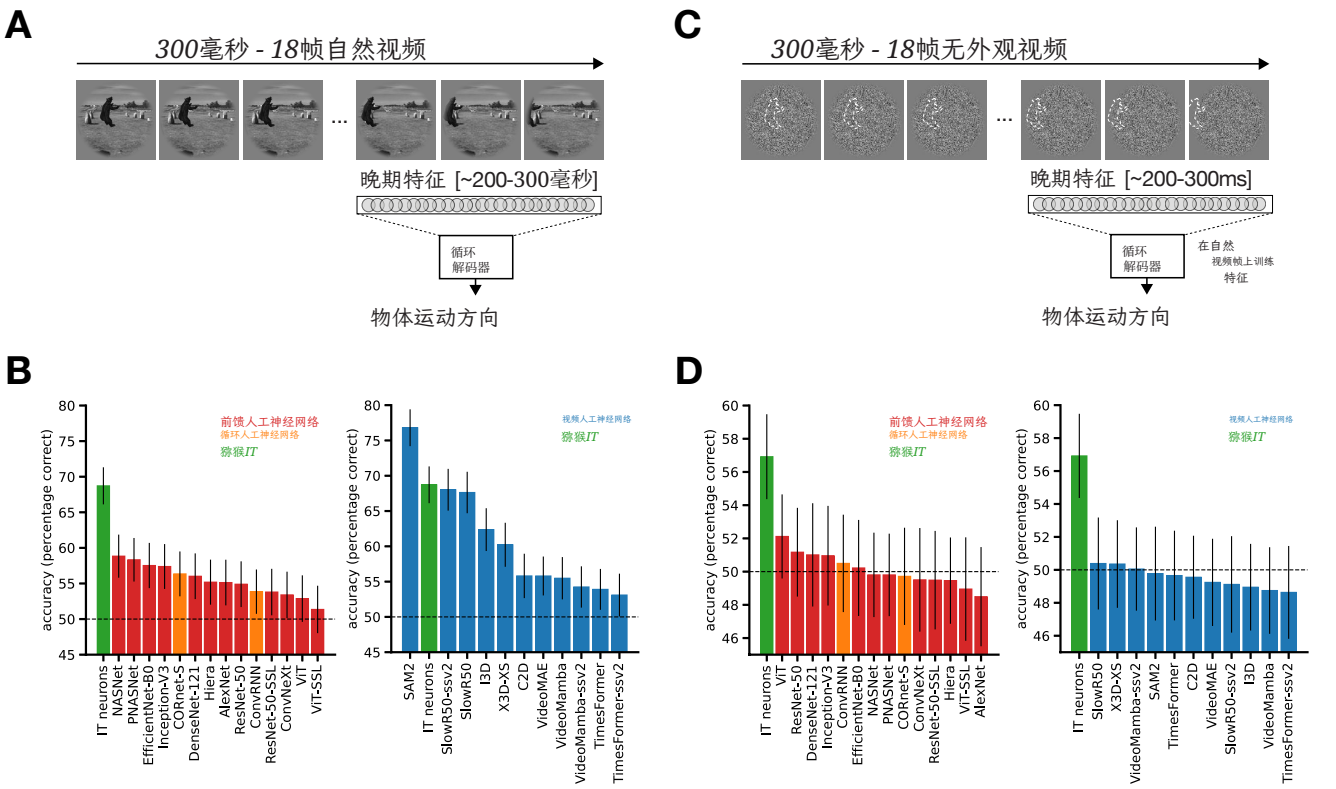


图 2。 **A**。基于长短期记忆网络的解码机制，用于评估模型在自然视频中表征动态物体属性的能力。 **B**。对于在自然视频上运行的静态与视频人工神经网络，从模型晚期特征和下颞叶皮层晚期响应（83 个神经元）中解码物体运动方向的准确率。我们报告跨视频的均值和标准误。 **C**。基于长短期记忆网络的解码机制，用于评估模型在无外观视频中表征动态物体属性的能力。无外观视频由帧序列组成，其中外观被随机像素值取代，这些随机像素值按照原始物体运动进行移动 [9]。 **D**。对于在无外观视频上运行的静态与视频人工神经网络，从模型晚期特征和下颞叶皮层晚期响应（83 个神经元）中解码物体运动方向的准确率。我们报告跨视频的均值和标准误。

总之，我们发现，动态视觉的人工神经网络模型虽已接近目标，却仍不足够：它们部分捕捉了IT较晚阶段的视频反应动态，但在面对不含外观信息的运动时便难以应对。与此同时，我们的分析为一套更丰富的经验约束奠定了基础，可为新架构的开发提供指引。通过以行为基准和IT反应的动态统计为未来模型奠定基础，我们可以进一步理解腹侧通路如何支持动态世界中的视觉。

致谢。 MD 获得了欧盟地平线欧洲计划研究与创新项目的资助，该项目依据玛丽·斯克沃多夫斯卡-居里资助协议编号 101151834（PRINNEVOT CUP G23C24000910006）开展。KK 获得了加拿大研究讲席计划、Google Research、加拿大卓越研究基金（VISTA计划）以及加拿大自然科学与工程研究委员会的资助，（NSERC）。

参考文献

[1] Pouya Bashivan、Kohitij Kar 和 James J DiCarlo。通过深度图像合成实现神经元群体控制。《科学》，364（6439）：eaav9436，2019年。1[2] Gedas Bertasius、Heng Wang 和 Lorenzo Torresani。视频理解是否只需时空注意力？载于国际机器学习会议，2021年。1[3] Luca Bertinetto、Jack Valmadre、Joao F Henriques、Andrea Vedaldi 和 Philip HS Torr。用于目标跟踪的全卷积孪生网络。载于欧洲计算机视觉会议，页码 850 865 2016 1 –。施普林格，。[4] Joao Carreira 和 Andrew Zisserman。动作识别何去何从？一种新模型与 Kinetics 数据集。载于*IEEE* 计算机视觉与模式识别会议，页码 6299–6308，2017年。1[5] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于*IEEE* 计算机视觉与模式识别会议，页码 248–255。IEEE，2009年。1

[6]Matteo Dunnhofer、Antonino Furnari、Giovanni Maria Farinella 和 Christian Micheloni。第一人称视觉中的视觉目标跟踪。《国际计算机视觉杂志》，2023年。1[7] Christoph Feichtenhofer、Haoqi Fan、Jitendra Malik 和何恺明。用于视频识别的 SlowFast 网络。载于 IEEE/CVF 国际计算机视觉会议，页码 6202–6211，2019年。1[8] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于 IEEE 计算机视觉与模式识别会议，页码 770–778，2016年。1[9] Filip Ilic、Thomas Pock 和 Richard P Wildes。无外观信息能否实现动作识别？载于欧洲计算机视觉会议，页码 156–173。施普林格，2022年。2, 4[10] Kohitij Kar 和 James J DiCarlo。探寻一套整合的神经机制，以解释灵长类动物的物体识别。《视觉科学年度评论》，第10卷，2024年。1[11] Kohitij Kar、Jonas Kubilius、Kailyn Schmidt、Elias B Issa 和 James J DiCarlo。有证据表明，循环回路对腹侧通路执行核心物体识别行为至关重要。《自然-神经科学》，22(6)：974–983，2019年。1[12] Seyed-Mahdi Khaligh-Razavi 和 Nikolaus Kriegeskorte。深度监督模型（而非无监督模型）或可解释其皮层表征。《公共科学图书馆·计算生物学》，10（11）：e1003915，2014年。1[13] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。使用深度卷积神经网络进行 ImageNet 分类。神经信息处理系统进展，25，2012年。1[14] Jonas Kubilius、Martin Schrimpf、Kohitij Kar、Rishi Rajalingham、Ha Hong、Najib Majaj、Elias Issa、Pouya Bashivan、Jonathan Prescott-Roy、Kailyn Schmidt 等。利用高性能浅层循环人工神经网络实现类脑目标识别。神经信息处理系统进展，32，2019年。1[15] Kunchang Li、Xinhao Li、Yi Wang、Yinan He、Yali Wang、Limin Wang 和 Yu Qiao。VideoMamba：用于高效视频理解的状态空间模型。载于欧洲计算机视觉会议，页码 237–255。施普林格，2024年。1[16] Aran Nayebi、Javier Sagastuy-Brena、Daniel M Bear、Kohitij Kar、Jonas Kubilius、Surya Ganguli、David Sussillo、James J DiCarlo 和 Daniel LK Yamins。灵长类腹侧视觉通路中的循环连接介导了任务表现与网络规模之间的权衡，这一权衡发生于核心物体识别过程中。《神经计算》，34(8)：1652–1675，2022年。1[17] Rishi Rajalingham、Elias B Issa、Pouya Bashivan、Kohitij Kar、Kailyn Schmidt 和 James J DiCarlo。对人类、猴和最先进的深度人工神经网络的核心视觉物体识别行为进行大规模高分辨率比较。《神经科学杂志》，38(33)：7255–7269，2018年。1[18] Hamidreza Ramezanpour、Filip Ilic、Richard P Wildes 和 Kohitij Kar。猕猴腹侧通路中的物体运动表征——通往理解大脑直觉物理引擎的入口。BioRxiv，页码 2024–02，2024年。1, 2[19] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman R¨adle、Chloe Rolland、Laura Gustafson 等。Sam 2：分割任何物体的图像和视频。*arXiv*预印本 *arXiv:2408.00714*，2024年。1[20] Martin Schrimpf、Jonas Kubilius、Ha Hong、Najib J Ma- jaj、Rishi Rajalingham、Elias B Issa、Kohitij Kar、Pouya Bashivan、Jonathan Prescott-Roy、Franziska Geiger等。Brain-score：哪种人工神经网络在物体识别方面最像大脑？*BioRxiv*，第407007页，2018年。1[21] Daniel LK Yamins、Ha Hong、Charles F Cadieu、Ethan A Solomon、Darren Seibert和James J DiCarlo。性能优化的分层模型能够预测高级视觉皮层中的神经反应。美国国家科学院院刊，111(23):8619–8624，2014年。1