主办 CCF计算机视觉专业委员会

COMPUTER VISION NEWSLETTER

CCF 计算机视觉专委会简报

01 2026

总第 47 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2026 年第 01 期

总第 47 期

CCF 计算机视觉专委会

主　办　CCF 计算机视觉专业委员会

编委会

CONTENTS
简报目录

CCF 计算机视觉专委会
CCFCV.CCF.ORG.CN
CCFCVN@GMail.com

CCF-CV 走进高校系列报告会

第 149 期 华东师范大学

2025 年 12 月 20 日下午，由中国计算机学会计算机视觉专委会（CCF-CV）主办，华东师范大学计算机科学与技术学院承办的第 149 期 CCF-CV 走进高校系列报告会——“视觉大模型前沿进展与挑战”学术论坛在华东师范大学计算机科学与技术学院理科大楼 B112 报告厅成功举行。本次报告会邀请了来自厦门大学**纪荣嵘**教授、中国科学院大学**蒋树强**教授、清华大学**郭雨晨**副研究员、南京理工大学**潘金山**教授、中国科学技术大学**张天柱**教授及南京理工大学**张姗姗**教授多位顶尖专家学者，共同探讨多模态大模型、视觉导航及医学模型等领域的前沿进展与未来发展。本次活动由华东师范大学**林绍辉**研究员担任执行主席，报告由华东师范大学**何高奇**教授、**林绍辉**研究员，**张志忠**副教授共同主持。

活动由**纪荣嵘**教授、**蒋树强**教授、**潘金山**教授、**郭雨晨**副研究员、**张天柱**教授、**张姗姗**教授五位专家针对多模态感知、具身智能与视觉理解等前沿方向的创新成果分别做主题报告，为师生搭建了与顶尖学者对话的高水平交流平台。

第 150 期 河南科技学院

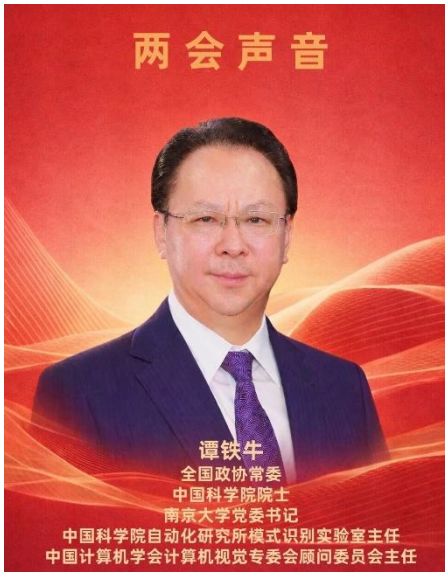
2025 年 12 月 19 日，由中国计算机学会主办，中国计算机学会计算机视觉专委会（CCF-CV）、河南科技学院软件学院联合承办的第 150 期 CCF-CV 走进高校系列报告会——“计算机视觉前沿技术及应用”在河南科技学院行政楼 206 报告厅成功举行。本期活动执行主席是河南科技学院软件学院院长**古乐声**教授和河南科技学院软件学院副院长**马玉琨**副教授。邀请了北京大学**查红彬**教授、北京工业大学**毋立芳**教授、清华大学**高跃**教授等 3 名业内专家作特邀报告。河南科技学院师生聆听了三位专家的精彩报告。报告由**马玉琨**主持。

活动由**查红彬**教授、**毋立芳**教授和**高跃**教授针对具身视觉、多模态感知及超图计算的前沿进展与未来趋势分别做主题报告。随后，三位专家与师生积极互动，展开了热烈的讨论，针对师生们提出的一系列问题，专家们给予了详尽的解答，分享了大量宝贵的学术观点与见解，为师生搭建了一个难得的学习与交流平台。

责任编委 毋立芳、朱安娜

CCF-CV 委员声音

2026 年全国两会胜利召开，时代的春风激荡神州，各界代表委员齐聚一堂，为国家发展注入澎湃动力。在这场关乎国计民生、民族未来的盛会中，中国计算机学会计算机视觉专委会（CCF-CV）的委员们带着对科技与教育的深刻洞察、对未来的热切期许，走进两会会场，发出专业强音，让我们一同聆听他们的智慧之声，感受科技赋能时代的磅礴力量。

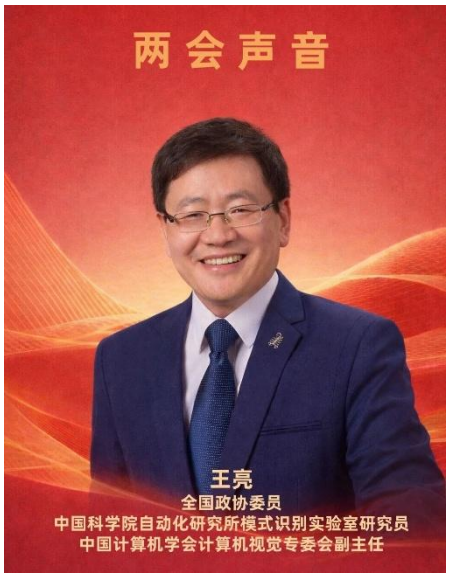


“十五五”规划建议提出，“建设具有全球影响力的教育中心、科学中心、人才中心”。**谭铁牛**院士强调，这赋予了高等教育特别是高水平研究型大学新的时代使命。高水平研究型大学需以更加开放、包容、自信的姿态，在服务国家战略大局和促进文明交流互鉴中彰显更大作为。

谭铁牛院士表示深化拓展高等教育对外开放可以从四方面入手：一是进一步优化全球伙伴关系网络。既要深化与世界一流大学的机制化、长效化合作，也要推动与非洲、东南亚、南美等地区的教育交流。二是进一步提升国际化人才培养质量。扎根中国大地，融通中外优质教育资源，着力培养兼具家国情怀与国际视野、能够参与全球治理的拔尖创新人才；完善留学生招生、培养、管理与服务，加强国际化课程和国际化师资队伍建设，推进“留学中国”品牌建设。三是进一步做强高水平开放创新平台。打造与国际接轨的学术科研体系，依托高能级科研创新载体，开展高水平国际联合攻关，实施更加积极、开放、灵活、有效的人才引进政策，打造全球创新要素集聚的高地。四是进一步深化国际人文交流。推动文明交流互鉴，以更富感染力的方式讲述中国故事、展示中国形象，为构建人类命运共同体贡献高等教育力量。

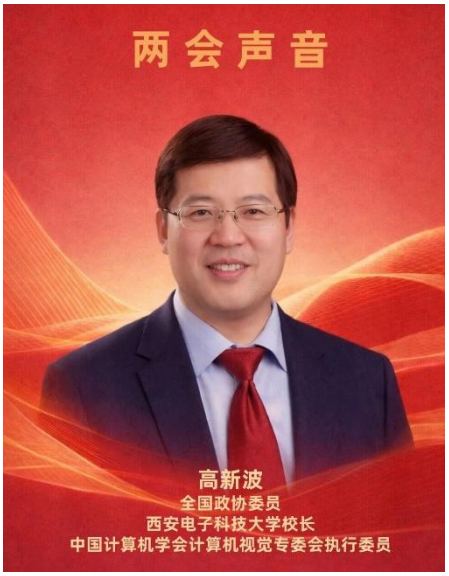
“十五五”时期经济社会发展，必须以推动高质量发展为主题。谭铁牛院士强调，发展新质生产力是推动高质量发展的内在要求和重要着力点，高水平研究型大学是助推新质生产力发展的重要力量，必须充分发挥自身在基础研究、学科交叉、人才培养等方面的综合优势，

切实担负起以发展新质生产力推动高质量发展的重任。一是要发挥基础研究优势，筑牢新质生产力发展之基。二是要发挥学科交叉优势，催生新质生产力发展之果。三是要发挥人才培养优势，造就新质生产力发展之才。



王亮委员在参加《焦点访谈》推出两会特别节目“跟着总书记上两会”时说：“总书记当时特别强调了应当重视基础研究和应用基础研究的发展。在‘十四五’期间，我们国家人工智能发展特别迅速，我们是自然而然延续以前的工作基础，做到多模态大模型和具身智能相关的研究工作，我们希望让机器人能够自动感知、认知、决策、规划到执行，更好地去推动科技创新的发展过程。”

王亮委员在参加两会期间表示，未来机器人是我们生活的重要伙伴。目前，我国具身智能产业已经进入全球“第一梯队”，未来拥有广阔的应用和发展前景。他认为，这得益于技术上的突破、硬件系统的改进、性能的提升、成本的下降、国家政策资金的支持以及实际的社会需求。随之而来的问题是人才需求量增大，王亮委员建议应加强相关领域人才培养，尤其是工程实践型的人才，以满足具身智能产业发展需求。



高新波委员在两会期间接受记者采访时提出：突破具身智能“卡脖子”环节，应优先推动具身智能在封闭场景闭环应用，通过“深水区”场景倒逼技术升级。

高新波委员提出，“具身智能产业承载着重大战略价值与深远意义。”从经济维度看，其不仅能赋能传统制造业智能化转型升级，更有望催生万亿级市场空间；从社会维度看，可有效替代危险或繁重劳动，显著提升公共服务效能，为民众生活带来更多便利；从技术维度看，将促进人工智能、物联网等前沿技术深度融合，辐射带动新材料、新能源等关联产业发展，助力我国在新一轮科技革命和产业变革中抢占国际竞争制高点。

高新波委员提出，随着人工智能、大数据等新兴技术在教育领域的深度应用，教育数字化转型已成为推动教育公平、提升教育质量的重要途径。然而，受多项因素制约，城乡学校在数字化转型进程中呈现出明显的“智能鸿沟”，这种“数字鸿沟”极有可能演变为“教育鸿沟”，进一步加剧区域教育差距。高新波委员认为，目前，农村学校数字化转型迫切需要外部支持。“为促进教育公平，为乡村振兴战略的实施提供有力支撑，建议尽快探索城乡学校‘数字结对帮扶’新模式，防止‘智能鸿沟’加剧区域差距。”

责任编委 潘金山

2026 年度 CCF-CV 秘书处工作会议召开

2026 年 3 月 1 日，中国计算机学会计算机视觉专委会 (CCF-CV) 秘书处年度工作规划会议于北京召开。会议由秘书长**王瑞平**研究员主持，秘书处成员**马伟**、**潘金山**、**任传贤**、**毋立芳**、**魏秀参**、**朱安娜**参加了会议。本次会议主要围绕专委会年度活动规划、活动优化方案及推进过程中的难点问题等议题进行了深入研讨。

首先，秘书长简要回顾并总结了过去一年秘书处的工作开展情况及取得的各项成效。随后，各活动负责人分别就相关工作的计划方案及进展进行了详细汇报，内容涵盖 RACV 及讲习班的程序组织、“视界无限”“走进高校”“走进企业”等品牌活动推进情况、科普工作规划、专委简报及公众号等宣传平台状况、PRCV 活动组织与财务收支情况，以及委员服务工作的落实情况等，同时明确了当前工作中存在的问题及拟采取的解决方案。秘书处全体成员围绕专委会活动组织及宣传推广工作中的痛难点问题，重点就提升委员参与积极性、优化品牌活动质量、扩大宣传平台覆盖面等内容，展开了深入研讨，最终形成了切实可行、易于落地的解决办法。

在新的一年，秘书处全体成员将持续主动推进各项工作落地，着力提升专委会的组织活力，确保各类活动高质量实施，以务实作风用心服务各位委员，以及计算机视觉相关领域的专家学者，助力领域发展。

责任编委 马伟

专题综述

真实图像超分辨率探索：数据，模型，联合优化

中山大学 魏朋旭 赵清 李冠彬 林倞

如何突破“理想退化”局限，让超分辨率技术从合成退化实验环境走向真实复杂世界，在现实场景中清晰还原且辨识图像目标？核心看点：构建真实数据基准，设计鲁棒模型架构，突破多帧与带噪超分，实现底层复原与高层感知任务的统一联合优化。

本文总结了中山大学HCP实验室在真实图像超分辨率领域的一系列工作成果，涵盖了系列数据基准的构建、模型设计和联合优化三个方面，从多维度拓展真实超分辨率任务研究：从单帧到多帧真实超分、从单一真实超分到真实带噪图像超分，从底层复原任务到与下游高层感知模型联合学习，探讨了面向真实图像超分辨率的异质退化、跨设备迁移、真实带噪联合优化、底层与高层统一联合优化学习问题（发表在NeurIPS2025）。

图像超分辨率（Super-Resolution, SR）旨在从低分辨率（Low-Resolution, LR）观测图像中重建出高分辨率（High-Resolution, HR）图像。尽管SR技术在重建质量上取得了显著进展，但是主流的研究大多依赖于合成退化模型（如双三次下采样）构建训练数据，无法反映真实世界复杂异质退化，模型泛化性差。真实世界超分辨率（Real-World Super-Resolution, RWSR）因此应运而生，旨在解决实际应用场景中的图像重建问题。RWSR直接处理由物理成像系统捕获的真实退化图像，这对数据集的构建质量和算法对复杂退化的建模能力提出了极高的要求。RWSR面临着诸多严峻挑战：一是真实场景下的成对数据（LR-HR）获取极其困难；二是不同成像设备（相机、传感器）导致的退化机制具有高度多样性和异质性；三是传统网络架构难以有效捕捉图像的高阶细节或解耦复杂的退化特征。本文综合探讨了我们在真实世界超分辨率领域的系列研究成果，主要从数据基准构建、模型设计和联合优化三个维度，阐述现有工作的局限性以及我们的相关工作。

一、研究背景

1.1 真实世界超分辨率数据构建

数据是驱动深度学习 SR 模型发展的基石。在真实世界 SR 领域，构建高质量、多场景、对齐精准的配对数据集是核心挑战。早期的 SR 基准数据集（如 Set5[1], Set14[2], DIV2K[3]等）主要由高分辨率图像经人工合成降质得到，无法反映真实成像系统的退化。随后的 RealSR[4]和 City100[5]等数据集尝试利用数码相机变焦拍摄真实 LR-HR 图像对，但仍存在局限性：City100 仅包含室内印刷场景，RealSR 虽然包含自然场景但采集设备较少。此外，针对多帧连拍（Burst）SR 的 BurstSR 数据集存在 LR 与 HR 图像由不同设备（手机与单反）拍摄导致的严重空间错位和色彩差异，且评估方式不公。而在含噪 SR 领域，现有数据集通常倾向于采用低 ISO 设置以避免噪声，忽略了真实世界中高噪声对超分辨率重建的干扰。

针对上述问题，我们构建了三个具有代表性的大规模真实世界基准数据集，填补了现有数据的空白：

1) DRealSR[6] (大规模多样化真实 SR 数据集)：我们构建了 DRealSR 数据集，这是目前规模较大、极具挑战性的真实 SR 基准。该数据集利用五款不同型号的数码单反（DSLR）相机采集，涵盖室内外丰富场景，不仅弥补了以往数据集场景单一、设备少的缺陷，还通过 SIFT 匹配与迭代对齐策略确保了像素级配准精度，为研

究不同设备间的异质性退化提供了数据支撑。

2）RealBSR[7] (真实多帧 SR 数据集)： 单帧图像超分时模型从单张图像中获取到的有效信息非常有限，限制了重建质量。为此，针对多帧连拍超分辨率，我们构建了 RealBSR 数据集。与 BurstSR 不同，RealBSR 采用光学变焦策略，使用同一相机在连拍模式下捕获 LR 序列和 HR 图像，有效避免了跨设备带来的色彩风格差异和严重畸变。该数据集保留了真实的亚像素位移和 RAW/RGB 数据特性，为多帧细节重建提供了真实基准。

3）RealNSR (真实世界带噪 SR 数据集)：真实场景应用下，图像通常是包含真实噪声的，因此更真实的超分研究应该是面向真实带噪场景下的超分。为此，我们建立了 RealNSR 基准。通过调整 ISO 设置(如 ISO 1600 和 3200)，我们采集了不同噪声强度的 LR 图像及其对应的低噪 HR 图像。这是首个专门针对真实世界“含噪”图像超分辨率构建的数据集，旨在解决真实噪声在 SR 过程中被放大且难以去除的难题。

1.2 真实世界超分辨率模型设计

在真实世界场景中，图像退化具有高度的复杂性、异质性和空间变异性。现有的 SR 方法往往难以有效应对这些挑战。传统的基于 L1/L2 损失的 SR 模型倾向于生成平滑结果，且对图像不同区域 (平坦、边缘、纹理) 采用均一化处理，导致难以恢复复杂纹理。在处理跨设备 SR 时，现有无监督域自适应（UDA）方法通常假设源域为合成数据，未能解决不同真实相机间各异的退化模式。此外，现有方法在多帧融合时过度依赖参考帧，或未能有效解耦噪声与图像内容，导致细节丢失或伪影产生。针对上述挑战，我们提出了一系列创新方法，从不同维度提升了真实世界 SR 的性能：

1) 成分分治策略（CDC[6]）：针对不同图像区域重建难度不均的问题，我们提出了 CDC 模型。该模型将图像解耦为平坦、边缘和角点三类成分，通过三个并行的“成分注意力模块”（CAB）分别进行处理，并设计了梯度加权损失（GW Loss）以自适应地强调高难度纹理区域的重建，实现了“分而治之”。

2) 跨设备域自适应（DADA[8]）：针对不同相机设备间的退化差异，我们提出了 DADA 模型。该模型采用双重对抗自适应策略，通过域间（InterAA）和域内（IntraAA）对抗学习，首次实现了从“真实源域”到“真实目标域”的无监督迁移，利用域不变注意力模块有效解决了跨设备 SR 的难题。

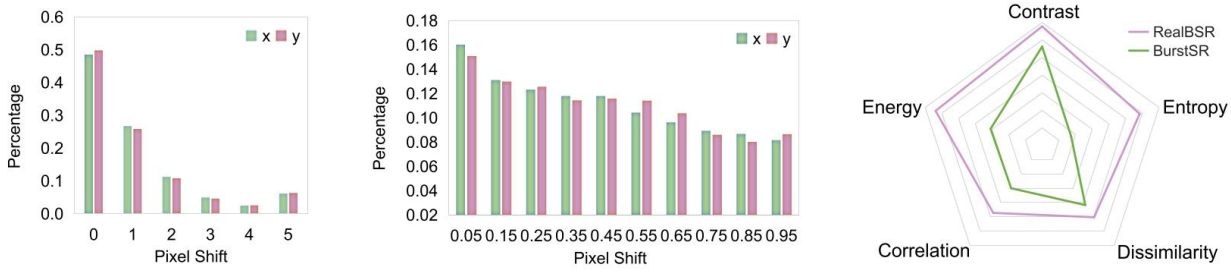
3）高阶特征学习（TNN[9]）：针对传统网络难以捕捉高频细节的问题，我们提出了泰勒神经网络(TNN)。基于泰勒级数近似理论，TNN 通过引入泰勒跳跃连接（TSC），在不同网络层级显式地构建和聚合图像的高阶导数信息，从而显著增强了网络对微细纹理恢复能力。

4）多帧信息融合（FBAnet[7]）：针对多帧连拍图像中的像素位移和融合难题，我们提出了 FBAnet。该网络摒弃了复杂的逐像素对齐，采用稳健的单应性对齐，并引入联邦亲和融合（FAF）策略。FAF 不仅关注与参考帧相似的内容，还通过亲和差异图聚合帧间的互补信息，从而更有效地利用多帧数据重建细节。

1.3 真实底层视觉任务的联合优化

真实图像超分的研究往往存在一个前提假设：不考虑真实噪声条件下处理真实超分任务。但是，真实噪声对于真实图像来说是无法避免的。为此，需要考虑真实去噪与真实超分之间的联合优化学习问题。另外，现有的超分任务往往只关注图像质量的提升，很少考虑其对下游高层任务的影响，或者只是简单地利用底层图像修复算法得到的图像输入给高层感知模型，但是该策略往往是不考虑底层图像修复任务和高层感知任务之间的差异性问题，前者是回归任务，后者是判别任务，两者之间的任务差异性无法保证下游高层任务在恶劣成像条件下的感知精度。在采用级联策略实现两类任务优化时，两类任务的不一致很可能会引入不稳定性，即在图像恢复阶段中引入微小噪声，在目标检测等高层感知过程中也会被放大，导致感知结果不可靠。此外，这两项任务在建模特性上的根本差异尚未得到充分探索，这在一定程度上限制两者更有效的集成与整体鲁棒性提升。

1）噪声分离范式（NSSR）：真实图像往往包含真实噪声，因此真实超分模型需要考虑真实去噪与真实超分的联合优化学习问题。针对含噪图像 SR 中噪声被放大



(a) Pixel shift computed by SIFT method according to keypoint shift

(b) Pixel shift within a range of one pixel

(c) Image Diversity

图 1　我们 RealBSR 数据集的特征分析

的问题，我们提出了 NSSR 网络。该网络采用全新的“噪声分离”范式，先从含噪特征中显式解耦出噪声特征和纯净图像特征，再将分离出的噪声作为条件，指导图像特征的高频细节重建，有效解决了真实噪声对 SR 干扰。

2）Lipschitz 正则化协同(LROD[10])：针对真实世界中图像恢复与以目标检测为例的高层感知任务级联时的不稳定性问题，我们提出了 LROD 框架。我们从 Lipschitz 连续性的视角分析发现，图像恢复网络通常是平滑的（低 Lipschitz 常数），而目标检测网络则具有不连续的决策边界（高 Lipschitz 常数），这种功能不匹配导致微小扰动被放大。为此，LROD 将图像恢复任务直接集成到检测器的特征学习中，通过低-Lipschitz 恢复正则化在输入空间约束检测特征的敏感性，并通过参数空间平滑正则化稳定梯度流动。该方法实现了恢复与检测任务在 Lipschitz 连续性上的协同，显著提升了复杂退化条件下的系统鲁棒性与检测精度。

二、真实世界超分辨率数据构建

2.1 DRealSR：大规模真实世界超分辨率数据集

为更深入探索真实世界中复杂的超分辨率退化机制，我们构建了一个大规模、多样化的超分辨率基准数据集——DRealSR（Diverse Real-world Super-Resolution），通过调整数码单反（DSLR）相机的光学变焦，分别采集真实世界的低分辨率（LR）与高分辨率（HR）图像对。DRealSR 覆盖 4 种尺度因子（即×1 至×4，实际用于训练与评估的为×2、×3、×4）。

数据采集。图像采集自五款主流 DSLR 相机（Canon、Sony、Nikon、Olympus 和 Panasonic），涵盖丰富多样的自然场景，包括室内与室外环境，并刻意避开动态物体（如行人、车辆），典型内容包括广告海报、植物、办公室、建筑物等。针对每种尺度因子，我们采用 SIFT 特征匹配方法[11]对 HR 与 LR 图像进行配准，并裁剪 LR 图像以精确对齐其对应的 HR 内容。为提升配准精度，我们将图像划分为局部块，并结合迭代配准算法与亮度对齐策略进行精细化校正。

由于在真实场景中同步采集精确对齐的 HR-LR 图像对极为困难，大量超分辨率方法仍依赖于模拟退化（如双三次下采样）构建的数据集进行验证。相比之下，真实世界 SR 呈现以下新挑战：

退化机制远比双三次下采样复杂。双三次下采样仅通过固定核对 HR 图像进行简单下采样以生成 LR 图像。而在真实成像过程中，下采样通常发生在各向异性模糊之后，并伴随信号相关噪声的引入。因此，LR 图像的获取同时受到模糊、下采样、噪声以及相机内部图像信号处理（ISP）流水线的联合影响。这种非均匀、非线性的退化特性可通过图 1 中不同图像区域（或成分）的重建难度差异得到验证。通常，在双三次退化数据上训练的 SR 模型在处理真实退化图像时表现显著下降。

不同设备间的退化过程高度多样化。在实际中，不同相机的镜头与传感器特性决定了其成像模式的差异，这是导致真实世界 SR 退化机制多样性的根本原因。因此，在某一真实 LR 数据上训练的 SR 模型，往往难以泛化至其他设备捕获的图像，甚至在同设备不同设置下也表现不稳定，这对 SR 技术实际部署构成了严峻挑战。

尽管如此，我们观察到图像中不同类型的区域对退化具有不同的鲁棒性。例如，平坦区域受退化多样性的

图 2　我们 RealBSR 数据集的实例

影响较小；而边缘与角点区域对退化设置的变化极为敏感，其重建结果随设备或成像条件变化呈现显著差异。这一现象启发我们将图像解析为平坦区域、边缘与角点三类低层成分，通过成分感知的建模策略缓解真实世界 SR 的训练难度，提升模型的适应性与泛化能力。

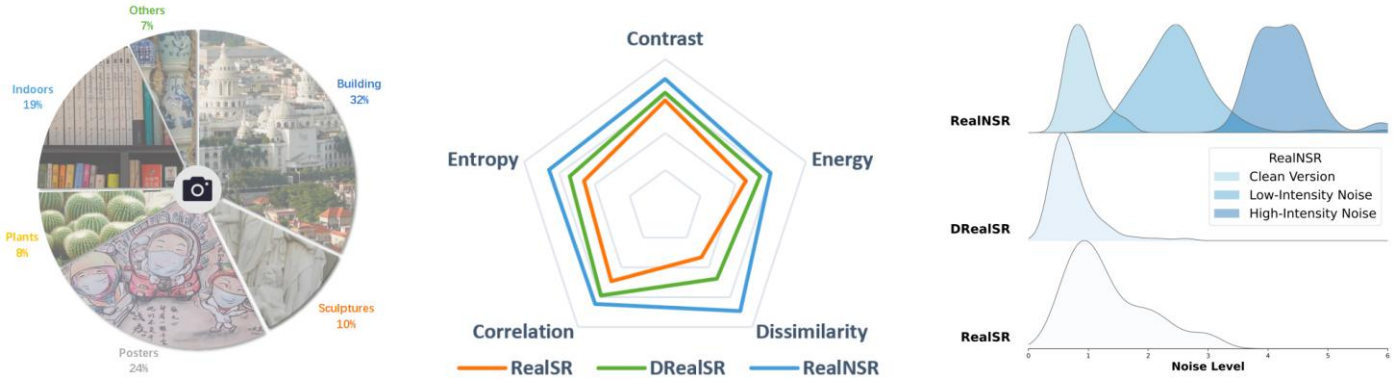
2.2 RealBSR：真实世界多帧超分辨率数据集

BurstSR[12]是唯一现有的真实世界多帧图像超分辨率数据集，存在三个典型问题。一是数据错位。L 图像与 HR 对应图像之间的畸变较为明显，可能导致 LR 和 HR 图像对之间出现严重错位。这种严重的不匹配会导致超分辨率结果不尽如人意，从多帧 LR 图像中重建的细节很少。二是跨设备分布。由于 LR 序列和 HR 对应图像分别由智能手机和单反相机拍摄，不同成像设备之间的差异不可避免地导致它们之间存在跨设备差异。因此，必须将 BurstSR 数据集上的任务视为多帧图像超分辨率和增强的组合。三是评估不足。BurstSR 的评估流程是将生成的最终 SR 图像以 GT HR 为参考进行扭曲，然后使用该扭曲后的 SR 图像与 GT HR 计算评估指标。这种评估方式存在较大问题，甚至无法公平地评估模型性能，借助 GT 进行评估。此外，计算出的指标值（例如峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)）无法很好地反映视觉质量，这意味着在 BurstSR 数据集上追求更高 PSNR 并不一定与更好的重建质量相关。这种评估策略主要归因于数据错位和跨设备分布，给真实世界多帧超分辨率带来了巨大挑战。

在本工作中，我们构建了一个名为 RealBSR 的真实世界多帧超分辨率数据集。它包含 579 组(RAW 版本)和 639 组（RGB 版本）的图像，放大比例为 4。每组包含 14 张多帧 LR 图像和一张 GT HR 图像，如图 2 所示。

特征分析。计算基准帧（第一张 LR 帧）与其他 13 帧之间的像素位移。如图 1（a）所示，50%的帧间偏移量小于 1 个像素，但 25 的偏移量在（1,2）范围内，还有 25%的偏移量大于 2 个像素，这表明模型仍需要一个对齐模块来消除较大的偏移量。与移动物体和不一致的颜色等其他外部因素不同，RealBSR 中的较大像素位移是由剧烈的手抖引起的。如图 1（b）所示，亚像素位移在（0,1）范围内均匀分布，为超分辨率提供了丰富的信息。

图像多样性。我们采用广泛用于测量图像纹理的灰度共生矩阵（Grey-Level Co-occurrence Matrix, GLCM）来分析图像多样性[13]。通过 GLCM，我们从所有训练图像中导出五个二阶统计特征，即 Haralick GLCM 特征[13]，包括图像对比度、熵、不相似性、相关性和能量。如图 1（c）所示，对比度衡量相邻像素之间的剧烈变化，不相似性与对比度类似，但呈线性增加。能量衡量纹理均匀性，熵衡量图像的无序程度，与能量呈负相关。相关性衡量线性依赖性。



(a) Scenario Distribution (b) Image Diversity (c) Noise Level Distribution

图 3　我们 RealNSR 数据集的统计信息：(a)展示了场景多样性统计，而(b)和(c)则对比了现有的 RealSR 数据集（即 RealSR 和 DRealSR 与我们的 RealNSR 数据集在图像多样性和噪声水平分布上的差异（x 轴为噪声水平，y 轴为密度)。得益于包含多样的现实场景和合格的噪声强度，RealNSR 比现有的 RealSR 更具实用性。

2.3 RealNSR：真实世界含噪超分辨率数据集

为了促进超分辨率（SR）方法在真实世界噪声图像上的学习和评估，我们构建了一个新的数据集，称为真实世界噪声图像超分辨率（RealNSR）数据集。我们利用全画幅单反相机捕捉了涵盖不同现实场景、光照条件和成像设置的真实世界噪声图像。通过调整 ISO 级别（低、中、高）来改变噪声强度，同时通过更换相机镜头获得多种分辨率（真值 HR 和 LR），从而生成逼真的噪声 LR 和 HR 图像对。RealNSR 数据集包含 580 个场景，具有不同的噪声强度，缩放因子为 4。每个场景包含一张无噪 LR 图像和两张具有不同噪声水平的噪声 LR 图像，以及它们对应的无噪 HR 图像。总共，RealNSR 包含 1,740 对 HR 和 LR 图像，其中 LR 图像分为三类：干净、低强度噪声和高强度噪声。

图像多样性。我们采用广泛用于测量图像纹理的灰度共生矩阵，来分析图像的多样性。利用 GLCM，我们从 RealNSR 数据集的所有训练图像中得出了五个二阶统计特征，包括图像对比度(Contrast)、熵(Entropy)、差异性（Dissimilarity）、相关性（Correlation）和能量（Energy），如图 3(b)所示。

噪声水平。我们计算了现 RealSR 数据集（即 RealSR 和 DRealSR）和我们 RealNSR 数据集中 LR 图像的噪声水平，其分布如图 3(c)所示。现有的 RealSR 数据集包含在小 ISO 设置下拍摄的无噪/低噪水平的 LR

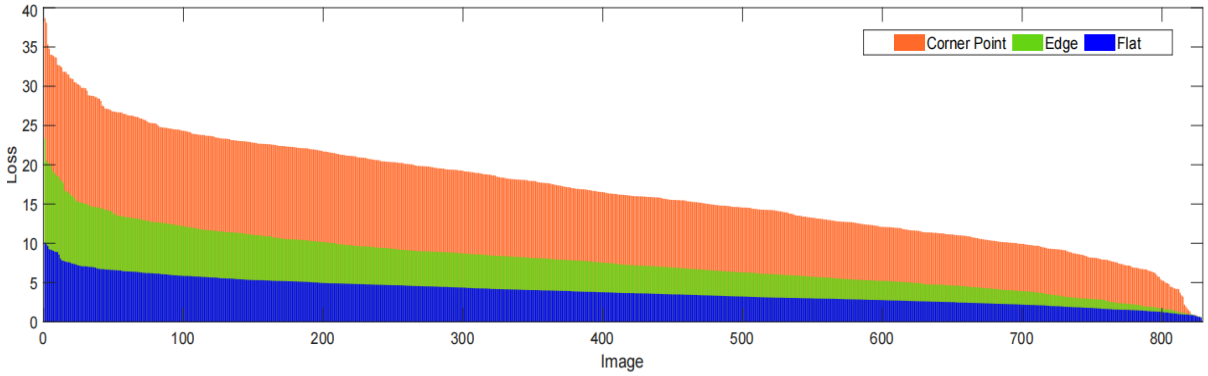


图 4　真实超分辨率中的成分分析。为探究其挑战性，我们分析了 EDSR 在 L1 损失下三类图像成分（平坦区域、边缘和角点）的占比，并通过平均逐像素损失评估各成分对超分辨率重建的影响。结果表明，三类成分的恢复难度存在显著差异：平坦区域与边缘的损失值较低，而角点的损失值明显更高。

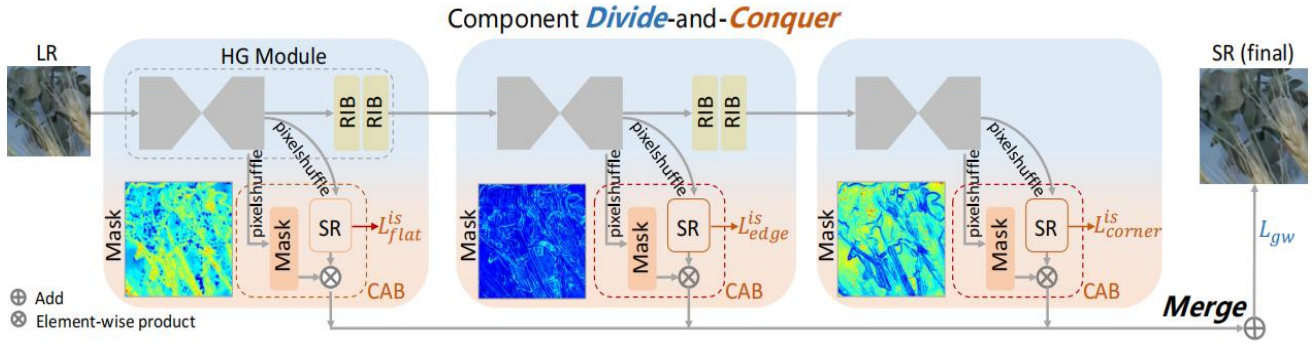


图 5　CDC 框架示意图。堆叠式架构使 CDC 能够在三个成分注意力模块（CABs）中分别建模平坦区域、边缘与角点。每个 CAB 分支独立生成一个注意力掩码和一个中间超分辨率（SR）结果。该掩码用于对中间 SR 进行正则化，使其与其他分支协同工作，并无缝融合各分支信息，最终生成自然、细节丰富的超分辨率图像。

图像，而我们的 RealNSR 数据集包含具有不同噪声水平的 LR 图像，分为三类：干净版本（低 ISO）、低强度噪声（中等 ISO）和高强度噪声（高 ISO）。由于具备合格的真实世界噪声强度，RealNSR 数据集比现有的 RealSR 数据集更具实用性。

三、真实世界超分辨率模型设计

3.1 CDC：成分分治策略

为应对真实世界中多样化的图像退化问题，一种直观的思路是学习各向异性模糊核。然而，自然图像内容高度复杂，难以设计普适的各向异性核估计方法。受 Harris 角点检测机制的启发，我们将图像内容依据其梯度变化划分为三类基本视觉成分：平坦区域（flat）、边缘（edges）与角点（corner points），显式地建模这三类成分，因其能有效表征图像内容的复杂性，从而反映其在超分辨率任务中的重建难度，如图 4 所示。例如，角点蕴含关键的方向性线索，对边缘形态与纹理外观具有决定性作用，因而在图像重建中具有潜在价值。基于此，我们探索利用这三类成分指导 SR 模型训练，以摆脱对多样化退化过程建模的依赖。

如图 5 所示，我们提出一种基于堆叠结构的 Hourglass 超分辨率网络（HGSR），并在此基础上构建成分分治模型（Component Divide-and-Conquer, CDC），结合梯度加权损失（Gradient-Weighted Loss, GW Loss）以解决真实世界 SR 问题。其中：

1）分（Divide）：按照从易到难的顺序（平坦→边缘→角点）引导网络分阶段学习特征，逐步提升对复杂结构的建模能力；分（Divide）考虑到平坦、边缘与角点区域在内容复杂度上的差异，我们引导三个 Hourglass（HG）模块分别从低分辨率（LR）图像中学习对应的成分注意力掩码（component-attentive masks），其监督信号来自高分辨率（HR）图像的成分解析结果。值得注意的是，我们并未直接在 LR 图像上使用现成检测器（如 Harris 算法）提取三类成分。主要原因有二：其一，LR 图像质量较低，导致角点等精细结构检测不准确；其二，该策略避免了在测试阶段对每张输入图像执行额外的角点检测，提升推理效率。

2）治（Conquer）：为每一类成分分别生成中间超分辨率结果；治（Conquer）三个成分注意力模块（Component-Attentive Blocks, CABs）分别生成针对平坦、边缘与角点的注意力图及对应的中间 SR 预测。所学注意力图清晰呈现出三类成分的语义特性，且中间 SR 结果亦与各自区域的结构特征高度一致。

3）合（Merge）：将三路中间结果依据其对应的成分注意力图加权融合，生成最终 SR 预测。合（Merge）为生成最终 SR 结果，我们以各 CAB 输出的中间 SR 图像为基，通过其对应的成分注意力图进行加权融合。特别地，我们提出梯度加权损失（GW Loss），驱动模型根据各区域的重建难度自适应调整学习目标。

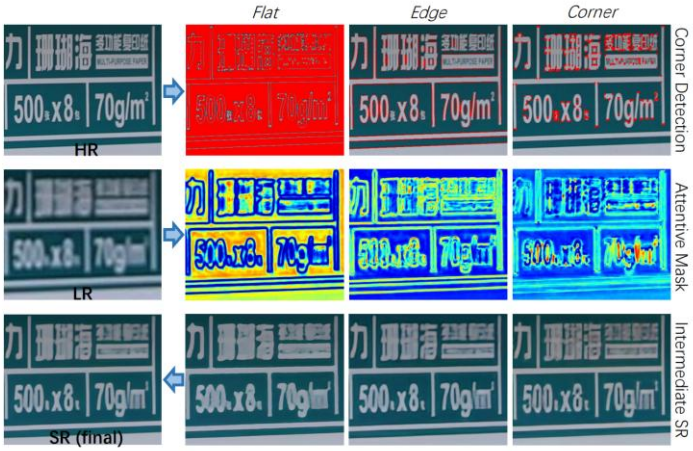


图 6　Harris 角点检测结果、各 CAB 学习到的成分注意力掩码以及对应的中间超分辨率图像。每个 CAB 生成的成分注意力掩码分别与平坦区域、边缘和角点高度相似，表明模型成功解耦并聚焦于三类低层图像成分。

Hourglass 超分辨率网络（HGSR）。我们提出 HGSR 作为基础模型，其采用堆叠沙漏（Stacked Hourglass）架构，后接 PixelShuffle 层。沙漏结构旨在捕获多尺度信息，在关键点检测任务中表现优异。每个沙漏模块可视为带跳跃连接的编码器-解码器：输入首先经卷积层处理，随后通过最大池化逐级下采样至最低分辨率；在自底向上阶段，通过最近邻插值逐级上采样，并利用跳跃连接融合多尺度特征，最终恢复原始分辨率。

成分分治模型（CDC）。CDC 以 HGSR 为骨干网络，遵循“分而治之”原则进行建模。其核心在于聚焦三类图像成分（平坦、边缘、角点），而非仅关注边缘或纹理。该策略使不适定的真实 SR 问题更易求解。三类成分通过 Harris 角点检测算法从 HR 图像中显式提取，并在各 CAB 中分别处理；最终通过最小化 GW 损失实现无缝融合，生成自然的 SR 结果。尽管成分指导信号来自 HR 图像，但 CDC 在测试阶段无需任何显式检测，即可推断成分概率图。如图 6 所示，每个 CAB 生成的成分注意力掩码分别与平坦区域、边缘和角点高度相似，表明模型成功解耦并聚焦于三类低层图像成分。

3.2 DADA：跨设备域自适应

由于复杂的成像过程，不同相机捕获的统一场景可能表现出不同的成像模式，这导致在不同设备图像上训练的超分辨率模型表现出不同的性能。我们考虑跨设备无监督域自适应的真实世界超分辨率问题，其中源域提供了来自一台相机的真实 LR-HR 图像对，而目标域只能获取来自另一台相机的真实LR图像。值得注意的是，这里提到的设备域差距源于两台不同型号相机对同一场景的成像差异。利用源域的 LR-HR 真实图像对和目标域的真实 LR 图像，我们的目标是训练一个 UDA 模型，用于在目标域中实现真实的 SR。在本节中，我们将详细阐述所提出的 DADA 模型，以探索跨设备真实世

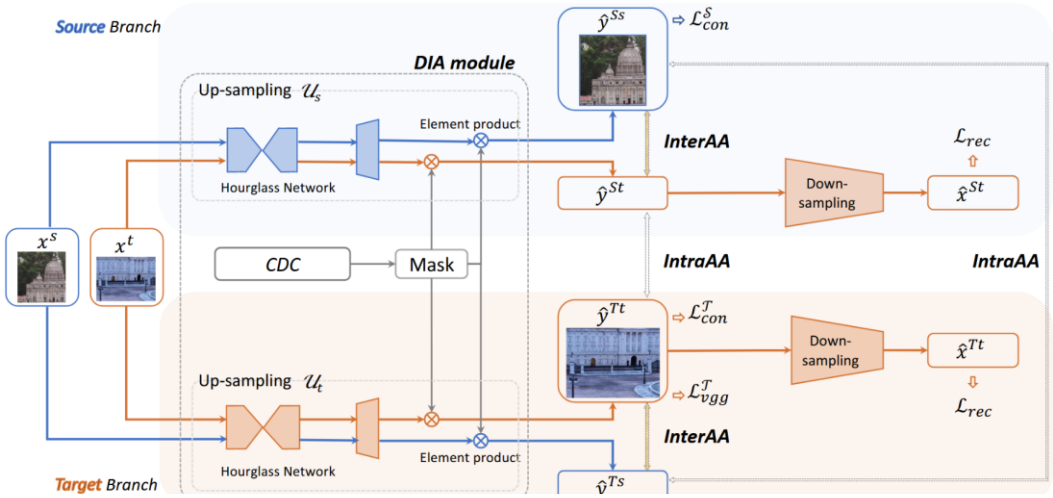


图 7　提出的 DADA 框架，包括一个源分支和一个目标分支，两者都采用 LR-HR-LR 重建结构。在其上采样网络中，一个域不变注意力（DIA）模块通过一个预训练的 CDC 模型提供组件注意力编码。对于 InterAA，来自不同域的两张 LR 图像被输入到上采样网络中，以便在一个分支内进行对抗自适应。对于 IntraAA，一张 LR 图像分别被输入到源分支和目标分支的上采样网络中，进行对抗自适应。在测试阶段，目标分支中经过训练的上采样网路被用于推理。

界图像超分辨率的无监督域自适应问题。

概述。高层视觉中的 UDA 有一个潜在的前提，即源域和目标域共享相同的语义类别。这有利于学习域不变的语义特征，将其作为模型自适应的基础。然而，对于真实 SR 中的 UDA 而言，要找出模型自适应的基础是很困难的。为了解决这个问题，一方面，我们的 DADA 模型利用从低层图像像素中提取中层图像组件的稳定性，来构建一个域不变注意力模块，而不是直接学习域不变特征。另一方面，所提出的 DADA 模型继承了针对目标域的循环一致性重建结构（LR→HR→LR）。考虑到目标域中 HR 图像的不可获取性，DADA 采用了一种在双重架构下进行域间和域内对抗自适应的训练策略。如图 7 所示，我们的 DADA 由两个对称的分支构成。每个分支都是一个 LR→HR→LR 的重建网络，包括一个上采样模块和一个下采样模块。其中一个分支主要由带有成对 LR-HR 监督信息的源数据主导，命名为源分支；另一个分支主要负责目标域，命名为目标分支。我们分别将源域 LR 图像x^s和目标域 LR 图像x^t作为输入，它们被送入这两个分支中的上采样模块，从而得到四个 SR 输出。我们使用域间对抗自适应（InterAA）方案来处理同一分支中来自不同输入的 SR 输出。相反，域内对抗自适应（IntraAA）方案则处理不同分支中来自相同输入的 SR 输出。正如 CDC 中所述，图像组件（即平面、边缘和角点）是内容依赖的，因此它们是域不变的，且受不同相机的影响较小。这促使我们基于一个预训练的 CDC 模型来建立一个域不变注意力模块。这两个分支通过域间和域内对抗自适应协同训练。

双重对抗自适应模型。对于源域，可以获取成对数据$\{x_i^s, y_i^s\}_{i=1,\dots,M}$，其中$x_i^s$是 LR 图像，而$y_i^s$是相应的 HR 图像。对于目标域，只能获取 LR 图像$\{x_j^t\}_{j=1,\dots,N}$，而他们对应的 HR 图像$\{y_j^t\}$是未知的。DADA 接收一个源域 LR 图像x_i^s和一个目标域 LR 图像x_j^t作为输入。这两张 LR 图像分别被同时送入源分支和目标分支。在一个分支中，除了 LR→HR→LR 重建过程之外，x_i^s和x_j^t通过同一个上采样模块（即生成器）被重新解析成 SR 图像，这些 SR 图像随后被送入一个判别器进行源域/目标域的判别。我们将这个过程命名为域间对抗适应(InterAA)。在两个分支之间，一张目标域（或源域）的 LR 图像会经历每个分支中的 LR→HR→LR 重建，而由两个分支的上采样模块生成的 SR 图像，将被进一步送入一个分支判别器来区分它们来自哪个分支。我们将这个过程命名为域内对抗适应(IntraAA)。

域不变注意力模块。我们分别用$\mathcal{U}_s$表示源分支的上采样模块，用$\mathcal{U}_t$表示目标分支的上采样模块，网络结构如图 8。它们遵循与 CDC 相同的神经网络架构。CDC 以沙漏网络为骨干，构建了三个组件注意力块，以解决模型在重建过程中对简单图像区域或内容的过拟合问题。这三个图像组件分别是平面、边缘和角点。考虑到这些组件相对而言对相机硬件是不变的，因此可以认为它们对于跨设备的 SR 任务是域不变的。这是实现真实到真实自适应的根本基础。因此，DADA 利用一个在源域数据对上预训练的 CDC 模型，为 LR 输入图像提供域不变注意力掩码（即 CDC 中的组件掩码）。具体来说，如图 7 所示，每个分支中的两个上采样模块共享相同的参数权重来生成组件掩码。与 CDC 类似，这些掩码分别对三个中间 SR 结果进行加权，然后将加权结果求和以生成最终 SR 结果。

域间对抗自适应。在每个分支中，来自源域和目标域的 LR 图像都由同一个上采样模块处理。由于源域的输入拥有 GT HR 图像，因此我们可以对源域的 SR 图像施加内容监督。InterAA 过程旨在将目标域与源域对齐。

在源分支中，上采样模块$\mathcal{U}_s$分别前向传输一个源域 LR 图像x_i^s和一个目标域 LR 图像x_j^t，并生成它们相应的

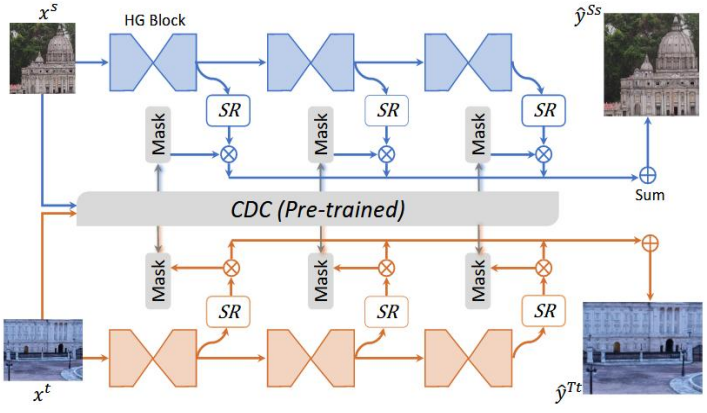


图 8　域不变注意力模块。为了简洁起见，我们通过将两张 LR 图像分别作为源分支和目标分支的输入，来展示其详细的网络结构。

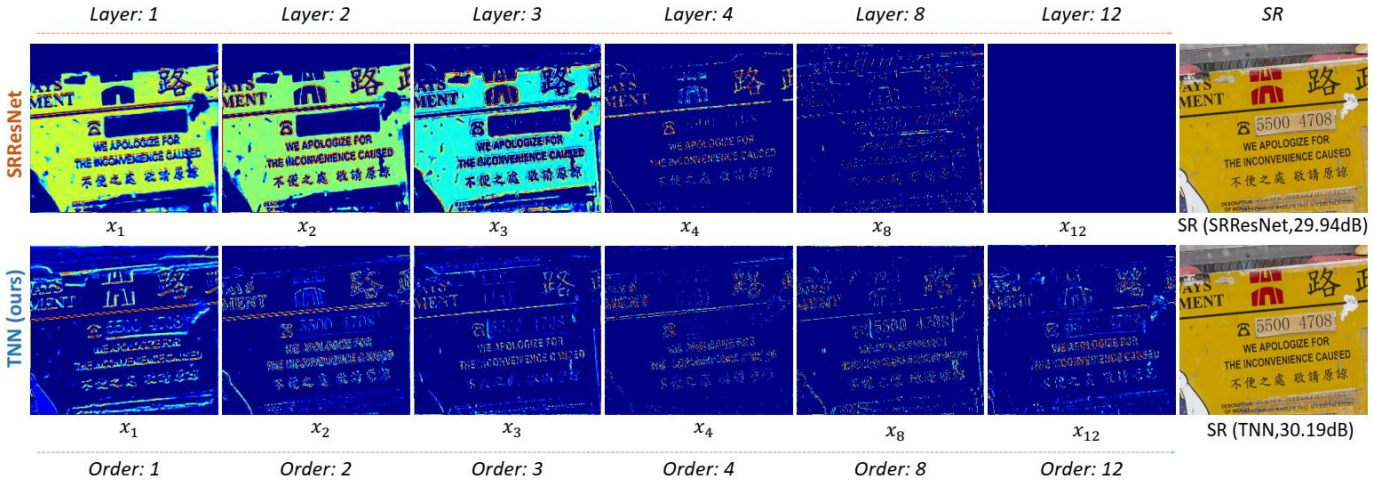


图 9　不同网络层级所学特征图的可视化对比。在各层中，所提出的泰勒神经网络（TNN）生成的泰勒特征图，相较于 SRResNet 能激活更多高响应值的像素点。SRResNet 的特征图易趋于平坦化区域（这类区域比纹理更易重建），尤其随着网络层级加深，其激活的像素数量显著减少。而本方法采用泰勒架构的 TNN 则能够学习到更丰富的图像细节特征。

超分辨率结果$\hat{y}_j^{Ss}$和$\hat{y}_j^{St}$。对于这两个 SR 图像，即源分支中的源域 SR 结果和目标域 SR 结果，我们采用判别器$\mathcal{D}_s^{inter}$来区分它们来自哪个域。通过在$\hat{y}_j^{Ss}$和$\hat{y}_j^{St}$之间施加对抗损失，我们强迫网络对目标输入生成接近源域的结果。以这种方式，通过对抗性的方式实现了域对齐，从而提高了模型处理目标数据的能力。值得注意的是，由于来自两个域的 LR 输入，例如，内容和颜色差异很大，这增加了它们之间对抗训练的难度，因此我们让判别器在 Y 通道上区分它们的 SR 图像。这样做可以避免对图像内容或风格产生偏差。

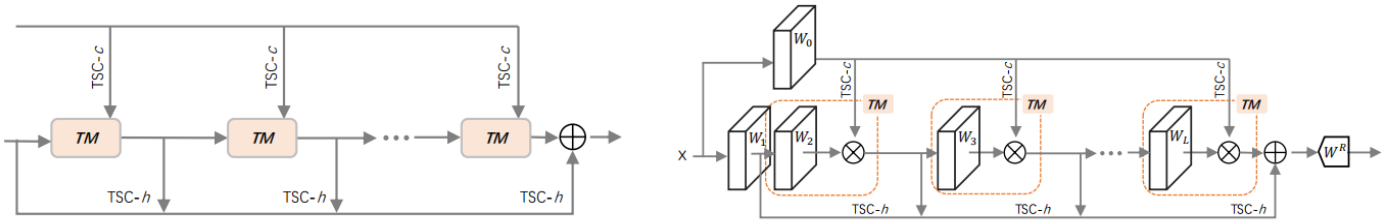
同理，在目标分支中，源域 LR 图像和目标域 LR 图像都被$\mathcal{U}_t$进行上采样，生成两个 SR 结果$\hat{y}^{Ts}$和$\hat{y}^{Tt}$，然后它们由判别器$\mathcal{D}_t^{inter}$进行判别。这里与源分支有两个不同点：为了避免源域对目标分支中的$\mathcal{U}_t$产生过于强大和主导的监督，不对源域 SR 结果$\hat{y}_j^{Ts}$施加任何监督。由于目标分支中不使用源域监督，为了稳定$\mathcal{U}_t$的训练，我们对目标域的 SR 图像施加了监督。由于缺乏目标域的 HR 图像，我们对目标域 SR 结果$\hat{y}_j^{Tt}$使用了伪标签。具体来说，我们使用目标输入在源分支中生成的 SR 结果$\hat{y}_j^{St}$作为$\hat{y}_j^{Tt}$的标签。

域内对抗自适应。InterAA 通过将来自两个域的不同 LR 图像作为输入，在源分支或目标分支内部对模型进行自适应。相反，DADA 通过将相同的 LR 图像作为输入，利用源分支和目标分支之间进行域内对抗自适应。具体来说，以源域 LR 图像x_j^s作为两个分支的输入，即两个分支分别使用$\mathcal{U}_s$和$\mathcal{U}_t$将其超分辨率为 SR 图像$\hat{y}_j^{Ss}$和$\hat{y}_j^{Ts}$。判别器$\mathcal{D}_s^{intra}$用于识别$\hat{y}_j^{Ss}$和$\hat{y}_j^{Ts}$是由哪个分支生成的。类似地，目标域 LR 图像也会被发送到两个分支的上采样模块中，以获得两个对应 SR 图像，然后由判别器$\mathcal{D}_t^{intra}$进行判别。也就是说，两个上采样模块$\mathcal{U}_s$和$\mathcal{U}_t$协同欺骗$\mathcal{D}_t^{intra}$。通过采用 IntraAA，我们迫使两个分支中的上采样模块产生彼此接近的结果，即使它们处于不同的监督之下，即源域的 GT HR 监督和目标域的伪 SR 监督。通过这种方式，我们对目标分支施加了间接的、对抗性的监督。

3.3 TNN：高阶特征学习

现有网络通过堆叠层级对输入图像进行序列化处理，其架构本质上难以显式学习图像的高阶信息。这将成为学习高频图像内容 (对超分辨率至关重要) 的障碍。如图 9 所示，SRResNet 在浅层主要关注比纹理更易重建的平坦区域，且随着层级加深，被激活的像素（尤其是纹理区域像素）持续减少，这将难以保证重建出更丰富的图像细节。

针对该问题，提出一种基于泰勒级数近似推导的通用神经网络架构——泰勒神经网络（TNN），示意图如图 10 所示。首先从泰勒级数近似的视角重新审视超分



(a) Taylor architecture　　(b) Taylor Neural Network

图 10　泰勒架构示意图：(a) 为泰勒架构原理图，(b) 为其具体实现。TNN 基于公式(5)的泰勒公式化表述推导得出，通过带有泰勒跳跃连接（TSC_s）的序列化泰勒模块（TM_s）构建。其中泰勒-c 跳跃连接（TSC_c）用于引入输入以推导高阶项，而泰勒-h 跳跃连接（TSC_h）则负责聚合来自不同层级的差异化阶数信息。

辨率任务中的传统卷积神经网络，揭示其与泰勒多项式近似的等效性。继而详细阐述泰勒神经网络及其与泰勒展开的理论关联。

重审超分辨率任务中的卷积神经网络。给定输入图像$\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$，一个包含$L$层的深度神经网络由参数$\boldsymbol{\theta} = \{(\mathcal{W}_l, b_l), l = 1, \ldots, L\}$，参数化，其本质是学习一个映射函数$\mathcal{F}: \boldsymbol{x} \mapsto \boldsymbol{\mathcal{F}_\theta}(\boldsymbol{x})$。在第$l$层的输出通过如下序列层的递归方式计算：

$$F_l = \sigma(W_l \ldots \sigma(W_1 \cdot x + b_1 + \cdots) + b_l) \qquad (1)$$

其中$\sigma(\cdot)$为线性整流单元（ReLU）激活函数。对于卷积滤波器$\omega_{ijc}^l \in \mathcal{W}_l$（其中$(i,j)$为滤波器在通道$c$上的二维索引），其在图像坐标位置$(s,t)$处生成的卷积响应$F_l(s,t)$可简化为如下表达式：

$$F_l(s,t) = \sigma\left(\sum_c \sum_{i,j} \omega_{ijc}^l F_{l-1}(s+i, t+j, c) + b_{l-1}\right) \qquad (2)$$

用于高分辨率图像重建的卷积神经网络。如上述公式所示，特征图F_l的每个元素均由图像感受野局部确定。由于卷积运算的线性特性与 ReLU 激活函数的分段线性特性，由上述公式可知：当激活函数σ采用 ReLU 时，神经网络所学习的映射函数与输入$\boldsymbol{x}$呈一阶线性关系。

基于这一观察，我们进一步分析神经网络结构，发现这种一阶网络架构难以高质量重建具有复杂纹理或细节的图像——这类图像内容在重建问题中通常以高频成分为主。另一方面，公式表明重建信号（图像）可被简单视为输入的分段线性组合。在超分辨率任务中，这种建模方式使重建图像实际上源于低分辨率输入图像的插值结果。然而，该策略难以有效应对现实场景中复杂异构的图像退化问题。

卷积神经网络中的高阶信息在图像超分辨率中的应用。从本质上讲，图像超分辨率过程可视为将低质量图像映射为高质量图像的过程。然而，神经网络中使用的 ReLU 激活函数具有分段线性特性，无法提供输入图像的高阶信息。ReLU 网络具有分段线性特性，这意味着输出图像的每个像素都是对应输入图像像素的线性组合。换言之，ReLU 网络本质上是一种针对不同像素值及邻域的自适应插值方法。特别需要指出的是，ReLU 网络的二阶导数处处为零，因此无法建模自然信号高阶导数中所包含的信息。鉴于 ReLU 网络的这一局限性，有必要引入图像的高阶信息来进一步提升深度神经网络的学习能力。

与泰勒展开的关联性分析。一方面，基于卷积运算的线性特性与 ReLU 激活函数的分段线性特性，映射函数$\mathcal{F}$可简化为如下表达式：

$$\mathcal{F}(\boldsymbol{x}) = b + \sum_{i,j\in\mathbb{R}} \alpha_{ij}\, \boldsymbol{x}_{ij} \qquad (3)$$

为简化表述，其中$\boldsymbol{x}_{(i,j)} \in \boldsymbol{x}$代表输入图像中与卷积核处于对应位置的像素，$\mathcal{R}$表示感受野，$\alpha$由对应相同感受野的不同层网络参数决定。另一方面，函数可通过以下泰勒展开式近似表示：

$$\mathcal{F}(\boldsymbol{x}) = b + \sum_{(i,j)} \nabla\mathcal{F}\boldsymbol{x}_{(i,j)} + \sum_{(i,j),(e,k)} \Delta\mathcal{F}\boldsymbol{x}_{(i,j)}\boldsymbol{x}_{(e,k)} + \cdots \qquad (4)$$

其中$\Delta\mathcal{F}$表示一阶偏导数$\frac{\partial\mathcal{F}}{\partial \mathrm{x(I,j)}}$，$\Delta\mathcal{F}$表示二阶偏导数$\frac{\partial^2\mathcal{F}}{\partial \mathrm{x(i,j)}\,\partial \mathrm{x(e,k)}}$。由公式(3)和(4)可知，从公式(1)推导出的映射

函数可视为与泰勒展开式的前两项完全一致，即等价于一阶网络。因此，我们可以通过类似方式构建高阶神经网络——随着网络层数的增加，其训练过程等价于实现 n 阶泰勒展开近似。

泰勒神经网络。基于上述分析，我们旨在将泰勒多项式近似思想融入现有深度神经网络，以改进其学习高阶信息能力不足的问题，从而提升高质量图像重建效果。具体而言，我们提出一种包含泰勒模块的泰勒架构，该架构无需复杂修饰即可实现特征学习，且兼容多种主流深度神经网络（如卷积神经网络、残差网络等）。基于泰勒架构衍生的网络被命名为泰勒神经网络（TNN）。本节以普通卷积网络为基准架构进行说明；若采用残差网络，只需将泰勒架构中的卷积层替换为残差块即可。TNN 由多个分布于不同层级的泰勒模块构成。每个泰勒模块通过附加分支（即泰勒跳跃连接 TSC）与各层级相连，用于构建含高阶信息的泰勒项。

泰勒跳跃连接。在每个层级中，泰勒模块（TM）会输出一个泰勒特征图F_i，其特性主要由输入信号直接主导——这得益于我们提出的泰勒跳跃连接（TSC）。该设计与带有跳跃连接的其他架构（如密集连接网络）存在本质区别：我们的架构通过不同层级聚合图像多阶内容的信息流。具体而言，TNN 引入两种泰勒跳跃连接形式：泰勒-c 跳跃连接(TSC_c)与泰勒-h 跳跃连接(TSC_h)。在第i层中，TM 首先通过卷积运算（卷积层可替换为残差块）或残差中的残差块等）处理前一层输出F_{i-1}，得到响应映射$\mathcal{W}_i\boldsymbol{x}$；随后，$\mathrm{TSC}_c$通过逐元素乘积运算将原始输入$\boldsymbol{x}$引入 TM，生成泰勒特征图；而$TSC_h$则通过加法运算聚合当前层与所有前置层的泰勒特征图信息。

1) 单层网络架构分析：我们首先考察主干网络仅含单层的泰勒神经网络，其参数集为$\{\mathcal{W}_1, b\}$。该主干网络的输出特征图（泰勒特征图）可表示为$\mathcal{F} = b + \mathcal{W}_1 * \boldsymbol{x}$。因此，该特征图亦可表述为 $\mathcal{F} = b + \sum \alpha_{ij}\boldsymbol{x}_{(i,j)}$。

2) 双层网络架构分析：对于主干网络包含两层的泰勒神经网络（参数集为$\{b, \mathcal{W}_0, \mathcal{W}_1, \mathcal{W}_2\}$），其主干网络输出的泰勒特征图可表示为$\mathcal{F} = b + \mathcal{W}_1 * \boldsymbol{x} + \mathcal{W}_2(\mathcal{W}_1 * \boldsymbol{x}) \circ (\mathcal{W}_0 * \boldsymbol{x})$基于卷积运算的线性特性，我们可以通过以下简化形式理解TNN的数学本质：

$$\begin{aligned}\mathcal{F} &= b + \sum_{(i,j)\in\mathcal{R}_1} \alpha_{ij}\boldsymbol{x}_{(i,j)} + \left(\sum_{(i,j)\in\mathcal{R}_{12}} \omega_{ij}\boldsymbol{x}_{(i,j)}\right)\left(\sum_{(e,k)\in\mathcal{R}_0} \omega_{ek}\boldsymbol{x}_{(e,k)}\right) \\ &= b + \sum \alpha_{ij}\boldsymbol{x}_{(i,j)} + \sum \beta_{ijek}\boldsymbol{x}_{(i,j)}\boldsymbol{x}_{(e,k)}\end{aligned} \tag{5}$$

其中，α_{ij}和β_{ijek}为网络在训练过程中学习得到的参数。$\mathcal{R}_0$、$\mathcal{R}_1$和$\mathcal{R}_2$分别表示由参数集$\{\mathcal{W}_0, \mathcal{W}_1, \mathcal{W}_2\}$所定义的单个或双层卷积操作的感受野范围。

显然，我们提出的网络最终输出特征与上述公式具有相似性，且各层参数恰好对应各阶偏导数。这表明通过样本训练，网络学习的参数可类比于图像超分辨率任务中的各阶偏导数。对于包含更多层的泰勒神经网络（TNN），这一结论同样成立。此外，网络层数可自由设定，其数值直接对应泰勒展开的阶数。

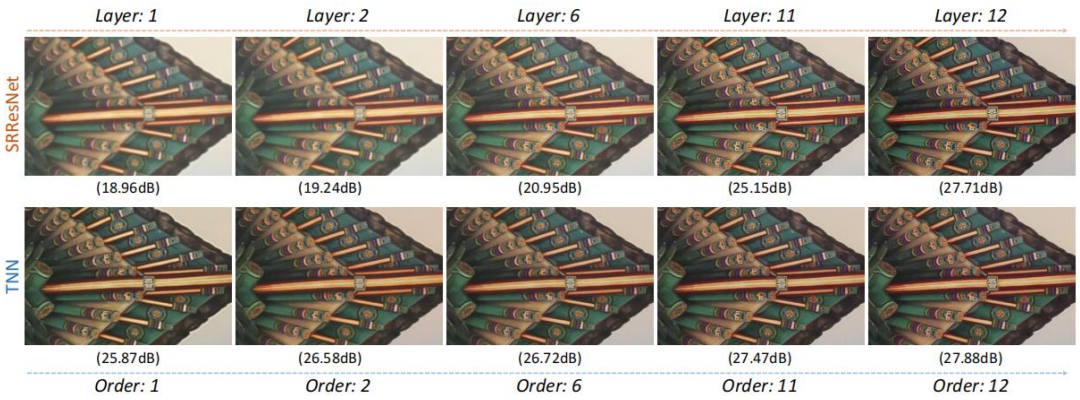


图 11　不同网络层级的超分辨率结果可视化对比

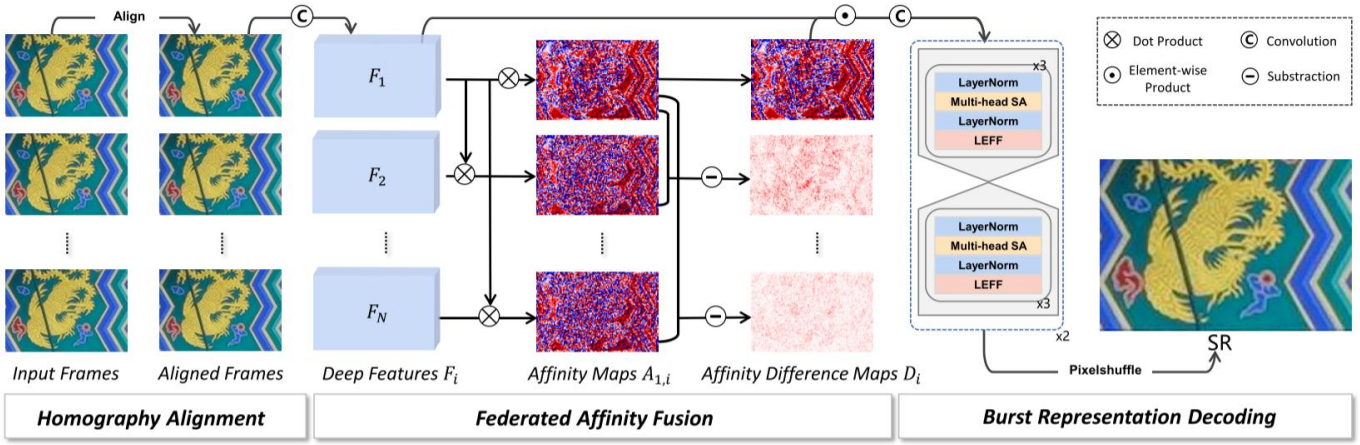


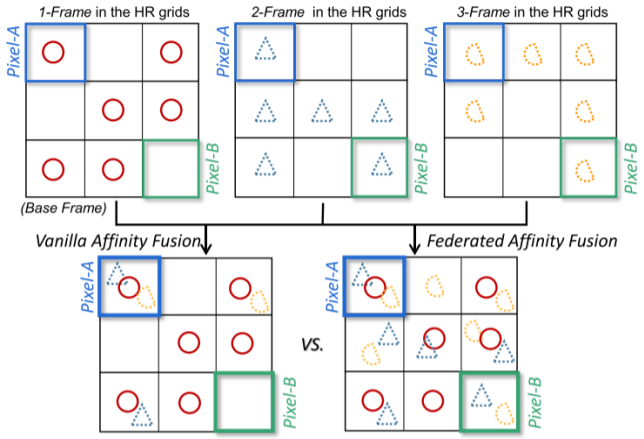
图 12　所提 FBAnet 的工作流程示意图，包含三个主要组件，包括单应性对齐、联邦亲和力融合以及多帧表示解码

泰勒特征图分析。我们在图9中进一步分析了不同网络层级学习到的泰勒特征图，并同步展示了如图11所示的SRResNet生成特征图作为对比。可以观察到，本方法生成的泰勒特征图在不同层级均对图像细节呈现强激活响应；而SRResNet的特征图则倾向于激活更易重建的平坦区域，且随着层级加深，激活的细节特征逐渐减少。这一现象可归因于SRResNet仅学习图像残差的简化策略——该策略会驱使模型优先拟合容易学习的视觉成分，从而导致其难以有效学习纹理等复杂视觉成分的细节特征。

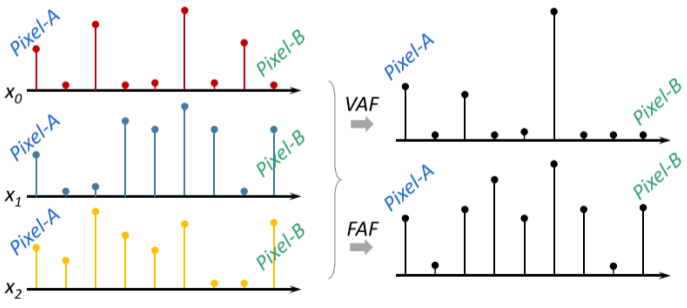
3.4 FBANet：多帧信息融合

与SISR相比，MFSR追求有利的像素级位移，以促进真实细节的重建。由于很难准确确定不同多帧LR图像之间的位移关联，因此如何融合多帧图像仍然是一个棘手的问题。更糟糕的是，由于成像过程中的相机抖动，会产生更多意外且不均匀的像素位移。为解决这一问题，我们提出联邦多帧亲和网络（FBAnet），通过有效整合多帧图像中的信息亚像素细节，迈向真实世界的多帧超分辨率。

FBAnet遵循传统的对齐-融合范式，如图12所示。具体来说，给定一个包含 N 张多帧连拍观测图像的LR图像序列 $\{x_i\}_{i=1}^{N}$ 作为输入，我们的模型将输出一个放大比例为 s 的高分辨率图像预测值 $\hat{y}_i$，其对应的真值HR图像表示为 y_i。由于不同多帧连拍图像之间存在随机的像素级位移，融合后的特征可能不足以直接支持图像细节的重建。不失一般性，将第一帧 x_1 视为参考帧，



(a) Fusion from pixels in different burst LRs. (*Circle*, *Triangle* and *Semicircle* indicate pixels in 3 burst frames, shown in the corresponding positions of HR pixel grids.)



(b) Fusion results on each of 9 HR pixels. (Height: signal intensity.)

图 13　FAF 的直观说明。FAF 除了考虑与基础帧相似的细节（如（a）中的 Pixel-A）外，还会考虑来自其他帧的更多互补细节（如（a）中的 Pixel-B）。(b) FAF 能够校正融合信息，避免来自具有极高相似度的易重建区域的负面影响，并鼓励亚像素进行融合。

通过单应性矩阵 H 对序列中的其他图像进行对齐。然后，FBAnet采用联邦亲和融合策略聚合多帧图像，并利用两个沙漏形Transformer块接管融合后的特征，以进行最终的高分辨率图像预测阶段。

单应性对齐。快速连续拍摄的多帧图像之间的像素级位移主要源于相机运动和场景变化，通常被认为是重建更多细节的补充信息。在融合图像之前，我们首先对它们进行对齐，避免信息混淆或差异，从而导致超分辨率预测模糊甚至出现不愉快的伪影。我们采用简单的单应性矩阵进行从全局和结构角度出发的对齐。具体而言，第 i 帧 x_i 与基准帧 x_1 之间的3×3单应性矩阵 H_t 表示基于相关系数最大化（Enhanced Correlation Coefficient, ECC）标准变换。每一帧都通过单应性矩阵变换，以基准帧为参考进行对齐。

联邦亲和融合。为了充分利用潜在信息，对齐后的图像会经过融合模块。由于融合后的输出不仅应与基准帧一致，还应整合来自其他帧的额外信号，因此我们提出联邦亲和融合（FAF）模块来聚合帧间和帧内的信息，如图13所示。我们的FAF通过为每一帧分配像素级融合权重来确定最终结果的有用性，这构成了整个多帧范式的核心。值得注意的是，一阶亲和图，即两个亲和图之间的差异，被用来确定权重，而不是亲和图或注意力图。具体来说，我们通过两个卷积层从对齐后的图像中提取深度特征 F_i。两帧之间的亲和图A是它们特征的点积，即 $A_{\{i,j\}} = F_i \cdot F_j$。融合的特征图可以表示为 $M = \sum_{i=1}^{N} A_{1,i} \circ F_i$，其中∘表示逐元素乘积。如图13a直观所示，VAF 关注与基准帧中像素一致的其他帧中的像素。因此，与基准帧相似的信息会被保留（例如图13a中的像素A），而仅在其他帧中出现的重要细节则被忽略（例如图13b中的像素B）。

FAF。尽管 $x_i(\forall i \neq 1)$ 对 x_1 的亲和度越高，表明其与基准帧的相似度越高，但也存在两个不利影响，尤其是对于真实世界的多帧图像：（a）那些易于重建的区域（例如平坦区域）也会对 $x_i(\forall i \neq 1)$ 产生较大的亲和度值，这将意外地促使模型更多地关注这些区域，从而导致过拟合。（b）由于不完美的对齐和像素位移，即使在细节丰富的区域中的关键像素也可能没有较大的亲和度值以突出显示用于融合。

多帧表征解码。为了聚合全局信息以重建更精细的高频细节，我们采用自注意力机制来模拟像素之间的长程依赖关系。具体来说，FBAnet 采用多帧表示解码模块，明确地模拟通道之间的相互依赖性。该模块包含两个级联块，如图 12 所示。每个块包含一个编码器和一个解码器，它们都级联了三个局部增强窗口（Locally-enhanced Window, LeWin）Transformer 块。每个块包含一个 LayerNorm、一个多头自注意力模块、一个 LayerNorm 和一个局部增强前馈（LeFF）层。该模块后面跟着像素重排（Pixelshuffle），生成最终 HR 预测。

图 14　给定一张真实世界的含噪 LR 图像，我们的 NSSR 与三类解决方案的预测结果对比，包括单任务方法（SR 和 DN）、单任务组合（SR→DN 和 DN→SR）以及联合训练解决方案（SwinIR-BSRGAN、SwinIR 和我们的方法）。注意，所有方法均使用我们的真实世界含噪数据进行训练，SwinIR 和 Restormer 分别作为 SR 和 DN 方法的示例。

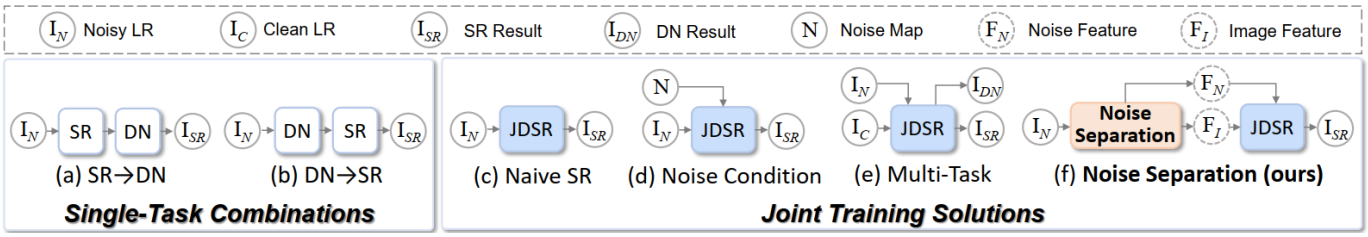


图 15　解决真实世界含噪图像 SR 问题的现有范式：单任务组合（SR→DN 和 DN→SR）、联合训练解决方案（即真实世界盲 SR、朴素 SR、噪声条件和多任务）以及我们的噪声分离（Noise Separation）。我们的噪声分离并非直接处理含噪 LR，而是首先解耦噪声特征和相对干净的图像特征，减轻了真实世界噪声对后续图像细节重建的干扰。

我们的训练目标包括用于SR图像重建的平均绝对误差（Mean Absolute Error, MAE）损失。此外，在RAW版本数据集上，为了减轻RAW版本数据集轻微错位带来的负面影响，我们还引入了CoBi损失，以简化训练并提高最终结果的视觉质量。而在RGB版本数据集上，我们采用梯度加权（Gradient Weighted, GW）损失，用于高频细节重建。

四、真实底层视觉任务的联合优化

4.1 NSSR：噪声分离优化范式

现有的真实世界SR研究局限于以干净的LR为输入，忽略了真实世界噪声的潜在影响。本质上，由于相机传感器、成像条件等原因，真实世界噪声在更现实的场景中是不可避免的。现有方法分为两类策略，如图14和15所示。(1) 单任务组合，即SR→DN和DN→SR。在SR→DN中，结果中残留可见噪声，而DN→SR难以重建更精细的图像细节；(2)联合DN和SR（JDSR）训练，包括真实世界盲SR、朴素SR、噪声条件和多任务等任务。具体而言，真实世界盲SR方法使用更复杂的退化流程来生成用于训练的含噪LR图像。然而，这些用于下采样和噪声的退化是合成的，导致与真实世界领域存在差距 。噪声条件在实际噪声场景中被证明是不切实际的，因为难以提前获得精确的噪声信息 。使用含噪-LR和GT对进行JDSR 的朴素SR训练难以获得与原始 SR 相当的结果 。一些现有的工作提出了多任务学习，在训练时同时在一个模型中学习 DN 和SR 。然而，当模型联合推理时，这些方法仍然遵循SR→DN和DN→SR，产生的结果与单任务组合中的SR→DN和DN→SR相似。

图像中真实世界噪声的存在给真实世界含噪图像 SR 带来了巨大挑战，导致真实世界噪声与 SR 之间的矛盾 。受语音分离的启发，我们尝试提取不含真实世界噪声的相对干净的图像特征，然后在这些相对干净的图像特征上执行重建模块，而不是在浅层含噪特征上 。我们提出了图16所示的噪声分离超分辨率(NSSR)网络 。我们采用噪声分离(NS)来估计两个掩码 M_N 和 M_I ，分别指示浅层含噪特征中噪声特征和相对干净图像特征的比例 。随后，利用噪声条件重建 (NCR) 在分离出的噪声特征条件下恢复相对干净图像特征的高频细节。

整体流程。 给定一张含噪 LR 图像I_{LR}，一个带有

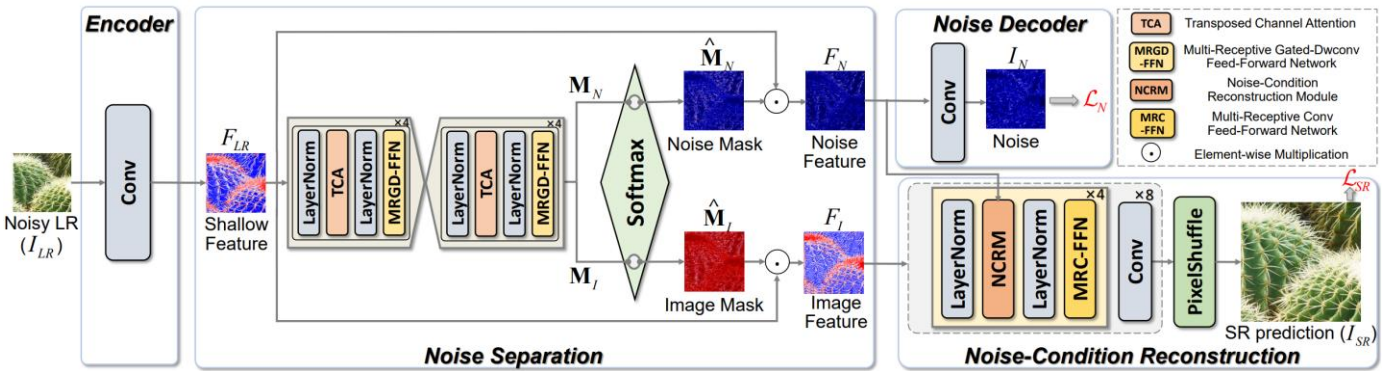


图 16　我们的噪声分离超分辨率（NSSR）网络框架

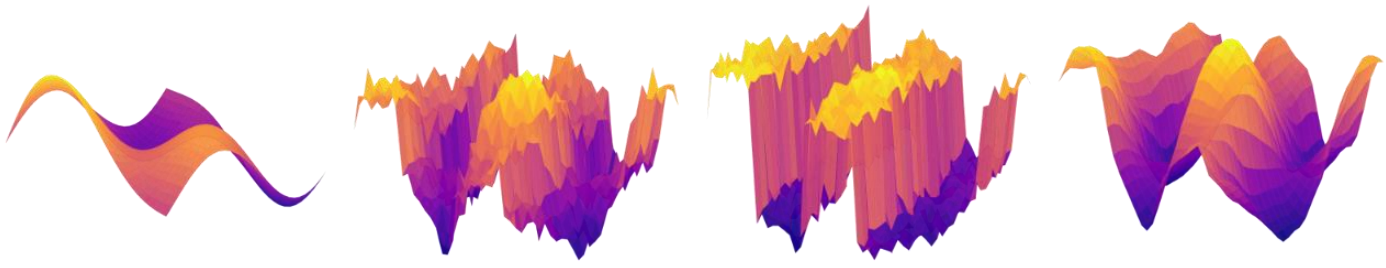
3×3 卷积的编码器（H_E）提取浅层含噪特征F_{LR}。H_E将含噪 LR 从低维空间映射到高维空间 。然后，NS（H_{NS}）从F_{LR}中解耦噪声特征F_N和相对干净的图像特征F_I。F_N和F_I在两个分支中处理：（1）带有卷积层的噪声解码器（D_N）将F_N重新映射为估计的噪声输出I_N，使用参考噪声$\widehat{I_N}$进行训练，$\widehat{I_N}$通过计算 GT 图像I_{GT}的双三次下采样与含噪LR图像I_{LR}之间的绝对差值计算得出；（2）NCR（H_{NCR}）在噪声特征F_N的条件下恢复相对干净图像特征F_I的高频细节，利用 PixelShuffle 获得 SR 结果I_{SR}，其中s是缩放因子。

4.2 LROD：基于 Lipschitz 连续性视角的图像恢复与目标检测联合优化

为提高在恶劣条件（如雾和低光环境）下的目标检测鲁棒性，图像恢复通常作为一种预处理操作来提升图像质量，从而为检测器提供更高质量的输入。然而，图像恢复网络和目标检测网络之间的功能不匹配会引入不稳定性，并阻碍两者的有效集成。本文从Lipschitz连续性的角度重新审视这一局限性，分析了恢复和检测网络在输入空间和参数空间中的功能差异。我们的分析表明，图像恢复网络执行的是平滑且连续的变换，而目标检测器则具有不连续的决策边界，因此对微小扰动极其敏感。这种特性差异导致传统级联框架中出现显著的不稳定性，即在图像恢复过程中引入的微小噪声，在后续的目标检测阶段也可能被放大，进而扰乱梯度回传并阻碍优化过程。针对上述问题，我们提出了Lipschitz正则化目标检测框架（Lipschitz-Regularized Object Detection, LROD），直接将图像恢复任务集成到检测器的特征学习过程中，在训练阶段协调两者的Lipschitz连续性。我们将该方法实现为Lipschitz-Regularized YOLO（LR-YOLO），可以无缝地扩展到现有的YOLO检测器。

从输入空间的角度来看，我们借助Lipschitz连续性的概念来刻画模型输出对输入扰动的敏感程度。具有较低Lipschitz值的网络通常表现出更平滑、可预测的行为，而较高值则意味着更强的输入敏感性和不稳定性。通过计算在不同雾霾密度变化下的雅可比范数（Jacobian norm），发现目标检测网络的Lipschitz常数相较于图像恢复网络高出近一个数量级，这突显了两者在平滑性上的显著差异。具体而言，图像恢复网络展现出平滑且连续的映射特性，其中微小的输入扰动只会导致恢复图像在像素层面的渐进式调整。这种平滑性源于其逐像素处理机制，该机制能够一致地增强局部区域，并将变化在整个图像中平稳传播。相比之下，目标检测网络本质上具有不连续性，表现为在分类和边界框回归任务中存在突变的决策边界。即使是微小的像素级扰动，也可能引发类别预测或边界框坐标的剧烈变化，从而在输出中产生非平滑、阶跃式的跳变。这种行为差异在两个网络级联时会引入显著的不稳定性。为更直观地说明这一现象，我们在图17（a）和（b）中可视化了这两类网络的功能行为，其中图像恢复网络呈现出平滑过渡的响应曲线，而目标检测网络则表现出明显的跳跃性变化，二者形成鲜明对比。这种不一致性导致两个级联网络时产生不稳定现象：在复原阶段引入的微小噪声会在检测阶段被放大，最终造成整个级联框架出现非平滑的运行状态，如图17（c）所示。

为更深入地理解传统“图像恢复→目标检测”级



（a）图像复原　（b）目标检测　（c）级联　（d）我们的方法

图 17　面向底层复原与高层感知任务的网络函数行为的可视化

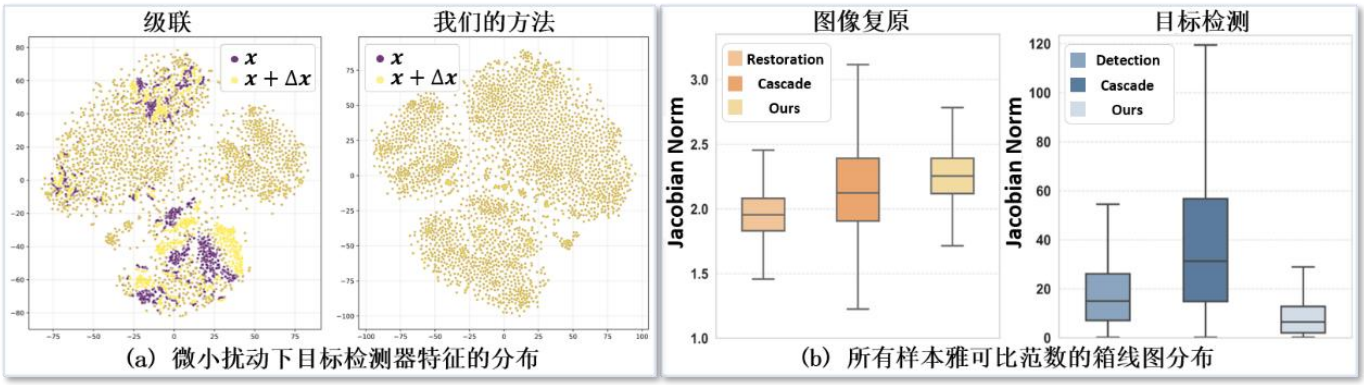


图 18　不同退化变化对特征稳定性和 Lipschitz 连续性的影响

联框架中所表现出的不稳定性，我们将分析进一步拓展至网络的参数空间。在此空间中，Lipschitz连续性被用来刻画模型输出对其参数变化的敏感程度。我们的研究表明，图像恢复网络在训练过程中通常具有相对较低的Lipschitz常数，这使其展现出平滑且稳定的优化轨迹。与之形成鲜明对比的是，目标检测网络往往具有显著更高的Lipschitz常数，导致其在优化过程中出现剧烈的梯度变化和不稳定的收敛路径。这种在参数空间中的行为差异造成了两者之间的不平衡，不仅扰乱了梯度流动，还引入了任务间的相互干扰，进一步加剧了联合优化过程的不稳定性。最终，这种机制成为传统级联框架中不稳定性的又一重要来源。

在雾霾或低光等恶劣成像条件下，目标检测网络对退化强度的微小变化（如雾霾密度、亮度波动）表现出高度敏感性。传统的“图像恢复→目标检测”级联框架难以有效应对这些变化，从而导致检测结果不稳定。如图18 (a) 所示，即使通过图像恢复可以缓解图像退化，但输入中的细微扰动仍可能引发目标检测网络特征的显著偏移，暴露出该框架在鲁棒性方面的缺陷。为深入理解这一不稳定性，我们借助Lipschitz连续性的理论工具，从输入扰动对模型输出的影响角度出发进行建模分析。实验结果表明，目标检测网络的Lipschitz常数比图像恢复网络高出近一个数量级，这使其在面对噪声干扰时更容易放大扰动，从而破坏整体系统的稳定性。

为定量比较图像恢复网络与目标检测网络在输入空间上的Lipschitz特性，我们在Pascal VOC数据集上针对不同雾霾密度变化，计算了每个样本的雅可比范数。如图18（b）所示，图像恢复网络的雅可比范数通常在1到3.5之间，而目标检测网络的雅可比范数则高出近一个数量级。这一结果表明，目标检测网络具有更高的Lipschitz常数，因此其对输入扰动更为敏感。当这两个任务被直接级联时，这种连续性上的显著差异将导致系统不稳定：图像恢复网络本身就会放大输入中的微小变化（因其样本雅可比范数已大于1），而高Lipschitz常数的目标检测网络将进一步加剧这种扰动效应，最终影响检测性能。

因此，图像复原网络通常呈现平滑且连续的映射关系，而从Lipschitz连续性的角度来看，目标检测网络则表现出更强的非平滑性。当这两个任务直接以级联方式组合时，由于它们在Lipschitz连续性方面存在显著差异，会进一步加剧整体系统的非平滑性，从而在面对不同退化强度变化时，引发显著的不稳定性问题。

图像恢复网络与目标检测网络在Lipschitz连续性上的差异不仅体现在输入空间，也可延伸至参数空间，对模型训练的稳定性产生直接影响。为了深入理解这种影响，我们从参数空间的角度出发，分析梯度更新如何在优化过程中影响模型的稳定行为。实验分析表明，图像恢复网络具有较低的Lipschitz常数，因此在整个训练过程中表现出平滑且稳定的优化轨迹。相比之下，目标检测网络的Lipschitz常数显著更高，导致其在优化过程中出现剧烈的梯度跳变，收敛路径不稳定。这种不平衡会扰乱梯度流动，加剧任务之间的相互干扰，最终导致基于级联的设计在联合训练过程中出现不稳定现象，优化效率下降。为直观展示参数扰动对优化平滑性的影响，我们可视化模型的损失景观（Loss Landscape）。如图19 (a) 所示，图像恢复网络的损失景观呈现出光滑、连

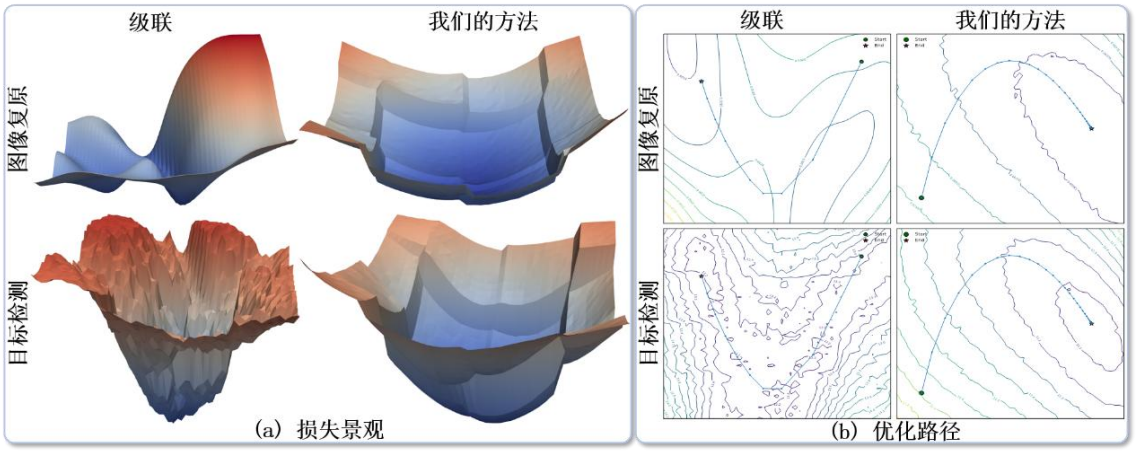


图 19　级联框架和我们的 LROD 框架之间的参数空间平滑度和优化稳定性比较

续的形态，而目标检测网络的损失景观则显得崎岖不平，反映出两者在参数空间中Lipschitz常数的巨大差异。图19（b）展示了对应的优化轨迹：图像恢复任务沿稳定路径收敛，而目标检测任务则频繁发生偏移，显示出明显的不稳定性。这种参数空间中的行为差异会在联合训练过程中进一步放大，破坏梯度流动，导致收敛困难并降低整体优化效率。

因此，具有较低Lipschitz常数的图像复原网络通常具有平滑的优化轨迹，而Lipschitz常数较高的目标检测网络则容易出现梯度剧烈波动与不稳定收敛。这种在参数空间中的不平衡性会导致二者在级联设计中的训练过程不稳定、优化效率下降，严重制约了图像复原与目标检测任务的高效协同。

上述分析中，我们发现图像恢复网络与目标检测网络在输入空间与参数空间中的Lipschitz连续性存在显著差异。基于这一关键发现，我们提出了一种简洁且有效的目标检测框架——Lipschitz-Regularized Object Detection (LROD)，通过有针对性的Lipschitz正则化策略来协调这两项任务，提升整体系统的稳定性与一致性。该方法的核心思想包含两个关键组成部分：（1）低-Lipschitz 恢复正则化（Low-Lipschitz Restoration Regularization），利用图像恢复任务本身所具有的低Lipschitz特性，将其集成到检测器的特征学习过程中，从而在输入空间中隐式地约束目标检测网络的Lipschitz常数，提升其对输入扰动的鲁棒性；（2）参数空间平滑正则化（Parameter-Space Smoothing Regularization），在参数空间中引入梯度范数约束，直接限制模型输出对参数变化的敏感程度，稳定训练过程中的梯度流动，增强优化的平滑性与收敛稳定性。

1）低-Lipschitz 恢复正则化：输入空间中的Lipschitz 连续性分析表明，图像恢复网络表现出平滑、连续的映射特性，而目标检测网络则更加非平滑。通过利用图像恢复任务的低 Lipschitz 特性，我们将图像恢复过程直接集成到目标检测骨干网络的特征学习中，从而约束目标检测任务在输入空间中的 Lipschitz 常数。$f_{\theta_b,\theta_d} = f_{\theta_d} \circ f_{\theta_b}$表示目标检测模型，其中$f_{\theta_b}(\cdot;\theta_b)$是由参数$\theta_b$定义的骨干网络，$f_{\theta_d}(\cdot;\theta_d)$是由参数$\theta_d$定义的检测头。类似地，令$g_{\theta_b,\theta_r} = f_{\theta_r} \circ f_{\theta_b}$表示图像恢复模型，其中$f_{\theta_r}(\cdot;\theta_r)$是由参数$\theta_r$定义的恢复头，共享相同的骨干网络$f_{\theta_b}$。给定检测损失和恢复损失的加权组合：

$$\mathcal{L}(\theta_b,\theta_d,\theta_r) = \mathcal{L}_{det}(f_{\theta_b,\theta_d}) + \lambda \cdot \mathcal{L}_{res}(g_{\theta_b,\theta_r}), \lambda > 0 \quad (6)$$

令$\mathrm{Lip}(f_{\theta_b}) := \sup_x |J_{f_{\theta_b}}(x)|$表示由雅可比范数定义的$f_{\theta_b}$的 Lipschitz 常数。如果满足以下条件：

1)$\mathcal{L}_{res}$是 Lipschitz 连续的，并且对于较小的$G < ||\nabla_{\theta_b}\mathcal{L}_{det}(f_{\theta_b,\theta_d})||$，有$||\nabla_{\theta_b\mathcal{L}_{res}}(g_{\theta_b,\theta_r})|| \le G$；

2) 存在一个训练样本x^*和$\gamma > 0$，使得$\left\langle \nabla_{\theta_b}|J_{f_{\theta_b}}(x^*)|, \nabla_{\theta_b}\mathcal{L}_{res}(g_{\theta_b,\theta_r}) \right\rangle \ge \gamma$；

那么在连续时间梯度下降过程$\theta_b^{(t+1)} \leftarrow \theta_b^{(t)} - \mu \cdot \nabla_{\theta_b}\mathcal{L}(\theta_b,\theta_d,\theta_r)$（其中$\mu$表示学习率），Lipschitz 常数的演化满足$\frac{d}{dt}[\mathrm{Lip}(f_{\theta_b})] \le -\lambda \cdot \gamma + \xi(t)$，其中$\xi(t) := \left\langle \nabla_{\theta_b}|J_{f_{\theta_b}}(x^*)|, \nabla_{\theta_b}\mathcal{L}_{det}(f_{\theta_b,\theta_d}) \right\rangle$是检测损失引起的无约束变化，而$\gamma$是通过图像恢复任务引入的正则化项。

这表明，通过共享检测器的骨干网络将图像恢复任务直接集成到检测器的特征学习中，有助于抑制模型对输入扰动的敏感性，从而有效地起到 Lipschitz 正则化的作用。

具体而言，我们提取目标检测骨干网络的前三层低级特征，这些特征保留了图像恢复所需的空间和纹理信息。然后，这些特征通过特定的恢复头以获得恢复后的图像。通过利用图像恢复任务固有的更平滑的 Lipschitz 连续性，恢复损失隐式地正则化了训练期间用于目标检测的特征表示，从而约束了检测网络在输入空间中的 Lipschitz 常数。如图 18（a）和（b）所示，我们的 Lipschitz 正则化目标检测方法相较于原始目标检测和级联框架，在不同退化强度下展现出更平滑的 Lipschitz 连续性，具有更低的 Lipschitz 常数和更稳定检测特征。

参数空间中的 Lipschitz 连续性分析表明，低 Lipschitz 的图像恢复网络能够维持平滑的优化轨迹，而高 Lipschitz 的目标检测网络则会经历剧烈的梯度变化和不稳定的收敛过程。为协调这两项任务并确保训练稳定性，我们通过约束目标检测网络在参数空间中的 Lipschitz 常数来促进更平稳的优化动态。

2) 参数空间平滑正则化：设$\theta = \theta_b \cup \theta_d$是检测模型的完整参数集合。参数空间正则化项定义为模型参数的梯度范数，记为$|\nabla_\theta f_\theta(x)|$。Lipschitz 正则化的目标检测框架 (LROD) 结合了上述两种正则化方法，以确保稳定和高效的训练。

总损失函数定义为：

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda \cdot \mathcal{L}_{res} + \lambda_p \cdot ||\nabla_\theta f_\theta(x)|| \quad (7)$$

其中，$\mathcal{L}_{det}$是检测损失，$\mathcal{L}_{res}$是恢复损失（计算恢复图像与真实干净图像之间的 Charbonnier 损失），而$|\nabla_\theta f_\theta(x)|$是正则化项。权重$\lambda$和$\lambda_p$分别用于平衡恢复项和正则化项。我们将该框架实现为 Lipschitz 正则化的 YOLO（LR-YOLO），基于 YOLO 检测器构建。如图 19 所示，ConvIR 和 YOLOv8 分别作为代表性的图像恢复和目标检测方法被采用。与传统级联框架相比，LR-YOLO 平滑了目标检测的损失景观，使其更好地与训练期间的图像恢复对齐。这带来了更平稳的梯度流动、更高的稳定性以及更高效的优化过程。

五、结论与未来展望

本文系统地探讨了面向真实世界场景的图像超分辨率重建问题，旨在解决传统基于合成退化假设的方法在实际应用中泛化能力差、细节恢复不真实等核心痛点。我们从数据基准构建、模型设计和联合优化三个维度出发，通过一系列创新性工作，显著提升了RWSR的性能与实用性。

尽管我们在真实世界超分辨率领域取得了一定进展，但面对极其复杂的现实成像环境和多元化的应用需求，仍存在诸多挑战与开放性问题，未来的研究可重点关注以下方向。1) 面向机器感知的复原优化：目前的 SR 研究多以提升人眼视觉质量(如 PSNR、LPIPS)为目标。然而，LROD 的工作表明，视觉上“好看”的图像并不一定利于机器识别。未来的研究应进一步探索“人眼视觉”与“机器视觉”在特征需求上的异同，设计能够同时兼顾视觉美感与下游任务（检测、分割、自动驾驶感知等）精度的联合优化模型；2）向动态视频场景的拓展： 目前我们的 FBAnet 主要针对静态场景的多帧图像。然而，在具有大幅度运动或复杂遮挡的视频超分辨率（VSR）任务中，简单的单应性变换可能失效。未来的工作需要探索更鲁棒的对齐机制（如光流引导的可变形卷积或基于 Transformer 的轨迹建模），将现有的空间自适应与噪声分离技术推广至视频领域，解决时序一致性与动态场景下的细节重建问题。

责任编委 王金甲

参考文献

[1] Bevilacqua M, Roumy A, Guillemot C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[J]. 2012.

[2] Zeyde R, Elad M, Protter M. On single image scale-up using sparse-representations[C]//International conference on curves and surfaces. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 711-730.

[3] Timofte R, Agustsson E, Van Gool L, et al. Ntire 2017 challenge on single image super-resolution: Methods and results[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 114-125.

[4] Cai J, Zeng H, Yong H, et al. Toward real-world single image super-resolution: A new benchmark and a new model[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3086-3095.

[5] Chen C, Xiong Z, Tian X, et al. Camera lens super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1652-1660.

[6] Wei P, Xie Z, Lu H, et al. Component divide-and-conquer for real-world image super-resolution[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 101-117.

[7] Wei P, Sun Y, Guo X, et al. Towards real-world burst image super-resolution: Benchmark and method[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13233-13242.

[8] Xu X, Wei P, Chen W, et al. Dual adversarial adaptation for cross-device real-world image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5667-5676.

[9] Wei P, Xie Z, Li G, et al. Taylor neural network for real-world image super-resolution[J]. IEEE Transactions on Image Processing, 2023, 32: 1942-1951.

[10] Zhao Q, Deng W, Wei P, et al. Delving into Cascaded Instability: A Lipschitz Continuity View on Image Restoration and Object Detection Synergy[J]. Advances in Neural Information Processing Systems, 2025.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[12] Bhat G, Danelljan M, Van Gool L, et al. Deep burst super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9209-9218.

[13] Haralick R M, Shanmugam K, Dinstein I H. Textural features for image classification[J]. IEEE Transactions on systems, man, and cybernetics, 2007 (6): 610-621.

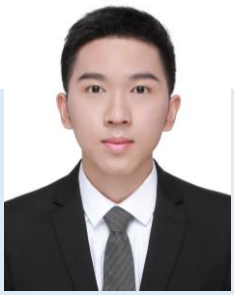
[14] Xiong Y, Li Z, Chen Y, et al. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 5652-5661.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

魏朋旭

中山大学计算机学院副教授。博士毕业于中国科学院大学。已发表 30 余篇高水平学术论文，两次举办 IEEE Seasonal School，并在领域顶会 ECCV 上举办真实图像超分辨率学术竞赛。获广东省科技进步一等奖（4/13），主持/参与国自然面上、青年、国家重点研发计划项目、国自然联合基金项目。现在主要研究兴趣：底层视觉模型预训练，真实图像超分辨率数据基准构建及方法研究，图像和视频内容生成，多模态内容鉴伪。

Email: weipx3@mail.sysu.edu.cn



赵 清

中山大学博士生。2023 年本科毕业于中山大学。在 NeurIPS 国际会议发表一作论文 1 篇，主要研究方向为真实世界超分辨率、目标检测。
Email: zhaoq78@mail2.sysu.edu.cn

李冠彬

中山大学计算机学院教授，博士生导师，国家优秀青年基金获得者。主要研究领域为图像视频内容理解与生成。 迄今为止累计发表 CCF A 类/中科院一区论文 200 余篇，谷歌学术引用超过 20000 次。曾获得中国图象图形学学会青年科学家奖、吴文俊人工智能优秀青年奖、ACM 中国新星提名奖、中国图象图形学学会科学技术一等奖、ICCV2019 最佳论文提名奖、CVPR2024 最佳论文候选等荣誉。主持了包括国家自然科学基金优青、面上、青年、重点研发课题、广东省杰青、 CCF-腾讯犀牛鸟科研基金、CCF-快手科研基金等 20 多项科研项目。担任广东省大数据分析与处理重点实验室副主任、广东省图象图形学会计算机视觉专委会主任等职务。担任人工智能领域顶级会议 CVPR、ICCV 等顶会领域主席，获得 10 余项人工智能领域国际顶级会议竞赛冠军，研究成果应用于智能交通分析、智慧医疗诊断、数字人驱动的智慧教育等。
Email: liguanbin@mail.sysu.edu.cn

林 倞

中山大学计算机学院教授，二级教授，IEEE Fellow，国家杰出青年基金获得者。长期从事多模态人工智能、机器学习等领域的应用基础研究，作为首席科学家/项目负责人，承担国家 2030 科技创新重大项目，入选国家万人计划青拔人才。在国际顶级学术期刊和会议发表论文 300 余篇，谷歌学术引用 4.4 万次，多次入选全球高被引学者榜单；获权威期刊 Pattern Recognition 年度最佳论文奖，多媒体计算旗舰会议 ICME 最佳论文钻石奖，计算机视觉旗舰会议 ICCV 最佳论文奖提名；获广东省科技进步一等奖（1/13）、中国图像图形学会科学技术一等奖、吴文俊人工智能自然科学奖，省级自然科学一等奖；指导博士生获得 CCF 优秀博士论文奖、ACM China 优秀博士论文奖及 CAAI 优秀博士论文奖。（曾）担任知名期刊 IEEE Trans. Human-Machine Systems, IEEE Trans. Multimedia, IEEE Trans. Neural Networks and Learning Systems 编委 (Associate Editor)，十余次担任 IEEE CVPR、ICCV、NeurIPS、KDD、ACM Multimedia 等国际会议的领域主席。
Email: linliang@ieee.org

热点追踪

多模态视频-文本定位

中山大学 梁天铭 谭超镭 李谨轩 胡建芳

如何突破高成本标注与静态理解的局限，让机器从“看懂一张图”走向“看懂一段视频”，并按语言描述准确找到目标内容？核心看点：构建低标注的学习方式，增强时序运动的理解能力，拓展图像文本基础模型的视频定位能力，推动视频理解走向应用场景。

本文主要介绍中山大学计算机学院胡建芳副教授课题组在多模态视频-文本定位领域的系列研究成果，具体包括发表在CVPR 2024的SiamGTR, ICCV 2025的ReferDINO和AAAI 2026的TubeRMC，下面将围绕以上工作的技术细节和创新点展开讨论。

一、背景

多模态视频-文本定位是连接视频内容与自然语言描述的核心技术，在视频检索和视频剪辑等应用起着至关重要的作用。然而，该领域面临着依赖昂贵的细粒度标注和模型时序理解不足等关键挑战。为了克服这些难题，本文聚焦于弱监督学习和基础模型继承两个方面展开具体研究：针对弱监督视频段落定位（WSVPG），提出了SiamGTR[1] 框架，通过新颖的孪生学习机制和伪监督数据生成，在无需时间戳标签的情况下实现高效的单阶段定位；针对指代视频目标分割(RVOS)任务，提出了端到端框架ReferDINO[2]，该模型继承了预训练视觉定位基础模型的强大能力，并引入对象一致性时序增强器和定位引导的可变形掩码解码器，有效解决了时序一致性和像素级分割难题。此外，针对弱监督时空视频定位（WSTVG），设计了TubeRMC [3] 框架，其核心在于时空条件重建学习和互约束学习机制，通过强制模型利用生成的时空管（spatio-temporal tube）来重建查询语句中的关键语义成分，从而在没有细粒度标注的情况下有效捕获管-文对应关系。这些研究旨在突破现有方法的局限，助力多模态视频语言理解技术在更复杂和更具挑战性的实际场景中落地应用。

二、视频段落定位

2.1 背景介绍

视频段落定位（VPG）旨在从未剪辑视频中定位具有语义关系和时间顺序的多个句子，是视频语言理解领域的新兴任务。然而，现有的VPG方法主要依赖于大量的时间边界标注，这些标注的获取过程极其繁琐且耗时费力，严重限制了模型在大规模数据上的应用。虽然弱监督方法试图消除对时间标注的需求，但现有的工作主要集中在单句定位上，且多基于多示例学习或重建学习的“提议-排序”范式，不仅计算复杂度高，而且在缺乏时间敏感监督的情况下难以准确衡量提议质量。为此，本文引入了一种新颖的孪生学习框架（SiamGTR），旨在通过联合学习跨模态特征对齐和时间坐标回归，在无需时间戳标签的情况下实现简洁高效的单阶段弱监督视频段落定位。

2.2 SiamGTR

本文提出了一个新颖的孪生学习框架SiamGTR，专门用于弱监督视频段落定位（WSVPG）任务，旨在无需时间戳标注的情况下，实现高效简洁的单阶段定位。如图1，该框架通过增广分支（Augmentation Branch）和推理分支（Inference Branch）的协作，联合学习跨

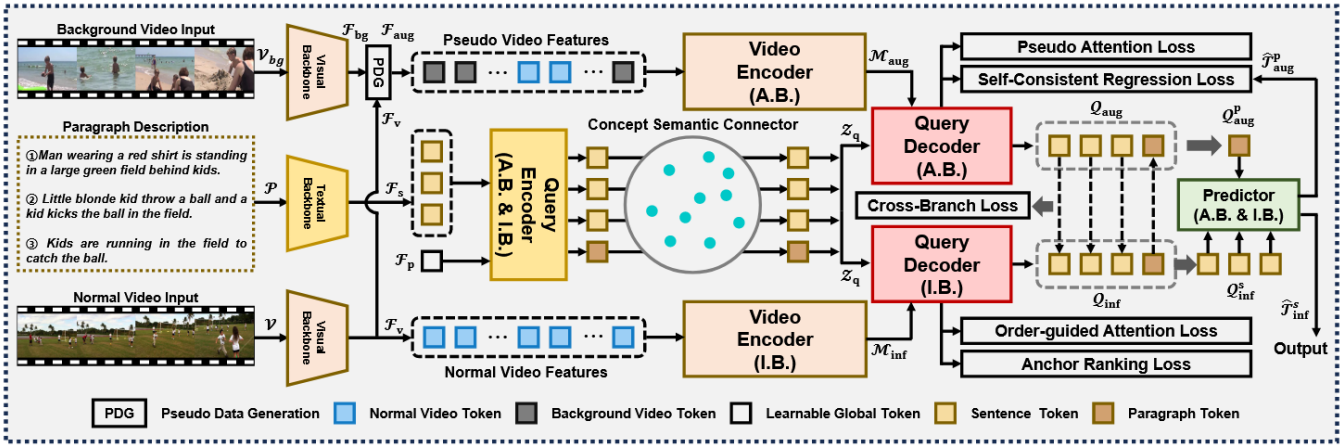


图 1　SiamGTR 架构示意图。增强分支（A.B.）接收通过将查询相关视频特征随机插入无关视频特征生成的伪视频特征，该分支学习以段落为查询条件，从伪视频中回归出目标时间区间。推理分支（I.B.）接收正常视频特征，用于学习视频中多句子的跨模态特征对齐。

模态特征对齐和时间坐标回归能力。增广分支通过合成具有已知时间边界的伪视频，利用自洽(self-consistent)回归损失来训练模型掌握精准的时间坐标回归能力。推理分支则用于处理真实的无标签数据，通过引入顺序导注意力损失和锚框排序损失，确保模型能够学习到文本叙述顺序与视频事件时序相符的正确时序对齐。两个分支共享一个带有动态锚框机制的查询解码器和一个概念语义连接器，以确保定位预测的精确性和语义理解的鲁棒性，从而在弱监督环境下实现高性能的段落定位。

增广分支与伪监督数据生成

增广分支（Augmentation Branch, A.B.）是该论文实现弱监督学习的核心部件。在弱监督视频段落定位（WSVPG)任务中，最大的挑战在于视频数据缺乏时间边界标注。传统的弱监督方法通常依赖多示例学习(MIL)或重建学习生成大量的提议（proposal），这些方法往往受限于低质量的提议或对时间不敏感的重建损失。该论文的创新之处在于，通过增广分支构造“伪视频”，从而在没有人工标注的情况下创造出精准的“伪标签”用于训练模型回归能力。该模块的核心思想利用了段落描述的特性：一段完整的段落描述通常对应视频中的一系列连贯事件。如果将一段主要视频插入到另一段完全无关的背景视频中，那么这段主要视频在合成视频中的位置就是天然的Ground Truth。

对于输入视频V和其对应的段落 P，系统首先随机采样一个背景视频V_{bg}。然后，通过特定的特征拼接策略生成伪视频特征F_{aug}。公式如下：

$$F_{aug} = NRS(Concat(F_{bg}^{1:I}, RRS(F_v), F_{bg}^{I+1:}))$$

其中，I是随机插入位置，$RRS(\cdot)$是随机重采样操作，$NRS(\cdot)$ 是归一化重采样操作，用于固定特征长度。与之对应的伪时间边界$(\tau_{aug}^{st}, \tau_{aug}^{ed})$可以通过插入位置和缩放比例精确计算得出：

$$\tau_{aug}^{st} = \frac{I + \Delta I^{st}}{rL + L_{bg}}, \tau_{aug}^{ed} = \frac{I + rL - \Delta I^{ed}}{rL + L_{bg}}$$

其中L和L_{bg}分别表示视频和背景视频的特征长度，r表示$RRS(\cdot)$的随机缩放因子。这里引入了ΔI^{st}和ΔI^{ed}作为随机边界偏移（Random Boundary Shifting, RBS），目的是模拟真实数据的边界不确定性，减轻合成伪影，提高模型的鲁棒性。

增广分支通过这种方式获得了强监督信号，并计算“自洽回归损失”(Self-Consistent Regression Loss)。该损失函数结合了L1损失和GIoU损失，但仅在注意力权重高度集中于伪真值区间时才进行优化。这种设计确保了模型仅在特征对齐较为可靠时才学习回归，从而避免了噪声干扰。通过该分支，模型在无需人工标注的情况下，学会了“如何从视频中回归出与文本对应的片段”。

推理分支与顺序引导注意力

推理分支（Inference Branch, I.B.）是模型处理真

实视频数据的核心部分，也是模型最终用于测试的通路。在孪生网络架构中，推理分支与增广分支共享所有权重（视频编码器、查询编码器、解码器）。由于真实视频没有时间标签，推理分支无法直接进行有监督的回归训练，因此它依赖于增广分支传递过来的回归能力，并侧重于解决“跨模态特征对齐”的问题。

该模块利用了视频段落定位任务中至关重要的先验知识：段落中句子的叙述顺序通常与视频中事件发生的时序一致。在推理分支中，输入是未经篡改的正常视频特征M_{inf}和文本查询特征Z_q。为了在没有标签的情况下让模型学会正确地“看”视频，我们设计了“顺序引导注意力损失”。该损失函数强制要求解码器中的交叉注意力图遵循时间顺序。具体而言，对于段落中的第j个句子和第$j+1$个句子，它们在视频上的注意力重心应当保持先后关系。公式定义如下：

$$L_{oga} = \max\left(0, \Delta mT + \sum_{t=1}^{T} t\alpha_s^j(t) - \sum_{t=1}^{T} t\alpha_s^{j+1}(t)\right)$$

其中$\alpha_s^j(t)$表示第j个句子在时间步t上的注意力权重，$\sum t\alpha_s^j(t)$即计算该句子的注意力时间重心。Δm是设定的最小间隔距离。

通过这一约束，模型被迫将前面的句子与视频较早的片段对齐，将后面的句子与较晚的片段对齐。这种机制有效地将增广分支学到的“定位能力”应用到了正确的“语境位置”上。如果没有这一模块，模型虽然具备回归能力，但面对包含多个事件的长视频时，无法区分哪个事件对应哪个句子。此外，该分支还引入了锚框排序损失（Anchor Ranking Loss），进一步强化了预测结果的时序一致性 。通过这种孪生互补的学习方式（增广分支教回归，推理分支教对齐），SiamGTR实现了高效的弱监督学习。

动态锚框查询解码器

该模块是连接特征编码与最终预测的关键桥梁，位于增广分支和推理分支的末端。本文并没有使用标准的Transformer解码器，而是受DETR[4]类目标检测模型的启发，设计了带有动态锚框（Dynamic Anchor Boxes）的查询解码器。此外，为了弥补句子特征与段落特征之间的语义鸿沟，该模块还辅以概念语义连接器（Conceptual Semantic Connector, CSC）。

在解码过程中，每个查询不仅包含内容信息，还关联一个动态更新的锚框。初始锚框设为零，通过MLP基于编码器内存特征M_{aug}和查询特征Z_q的交互来预测一组种子锚框。在每一层解码器中，锚框的坐标被转换为高维正弦嵌入（Sinusoidal Embeddings），作为位置编码参与自注意力计算和交叉注意力计算。

每一层输出的查询特征不仅用于下一层，还通过一个MLP预测相对偏移量ΔA，用于逐层精细更新锚框坐标：

$$A^{(i+1)} \leftarrow A^i + \Delta A^i$$

这种设计使得模型能够显式地建模查询的位置信息，而非仅仅依赖隐式的注意力机制，对于精确定位至关重要。

由于模型同时处理短句子和长段落，两者在语义空间上存在差异。CSC模块通过引入一个包含高频动词和名词的外部概念字典，计算查询特征与概念字典的匹配度，并利用多热编码标签（Multi-hot labels）进行监督 。损失函数如下：

$$L_{csc} = BCE(y_{cept}^p, \hat{y}_{cept}^p) + BCE(y_{cept}^s, \hat{y}_{cept}^s)$$

这确保了无论是句子还是段落查询，都能准确捕捉到关键的语义概念，从而增强了特征的判别力。

2.3 实验结果及分析

详细的实验设置、定量或定性实验可参考相关方法原论文，下面仅对关键实验结果进行展示。

本研究将SiamGTR与主流的弱监督及全监督方法在Charades-CD-OOD[5]、ActivityNet Captions[6]和TACoS[7]三大数据集上进行了对比。结果显示，在仅使用视频-段落对级别标签的弱监督设置下，SiamGTR在定位精度上展现出显著优势。例如，在Charades- CD-OOD数据集上， SiamGTR在R@1, IoU=0.5指标上取得了 52.66% 的成绩，相比最佳的WSVPG方法提升了超过4个百分点。这证明了本文提出的孪生学习框架（结合伪监督回归与顺序引导对齐）在

图 2　不同模型的预测结果可视化

单阶段、弱监督环境中实现高质量定位的优越性。更值得注意的是，SiamGTR的性能表现甚至能够逼近或在某些指标上超越部分全监督方法，显著缩小了弱监督与全监督方法之间的性能差距。这种高效性源于其简洁的单阶段定位设计，避免了传统弱监督方法中复杂且易出错的“提议生成-排序”两阶段流程。

在图2中，我们直观展示了WSTPN与弱监督/半监督设置下本模型预测的时间戳。总体而言，WSTPN预测的时间戳虽能粗略反映句子顺序关系，但边界存在偏差；相比之下，我们的弱监督模型取得了更精准的结果。值得注意的是，本研究的半监督模型生成了更精细的边界，这印证了孪生框架在联合利用少量弱标注数据实现高效学习方面的优势。

三、视频目标分割

3.1 研究背景

指代视频物体分割（RVOS）旨在根据文本描述在整个视频序列中定位并分割出目标物体，作为一种新兴的跨模态任务，其在交互式视频应用中极具潜力，近年来备受计算机视觉社区关注。尽管已有显著进展，现有的RVOS模型受限于视频-语言理解能力的不足，在面对涉及复杂物体属性和空间位置的描述时往往难以有效应对。目前的主流方法主要面临两方面挑战：传统的

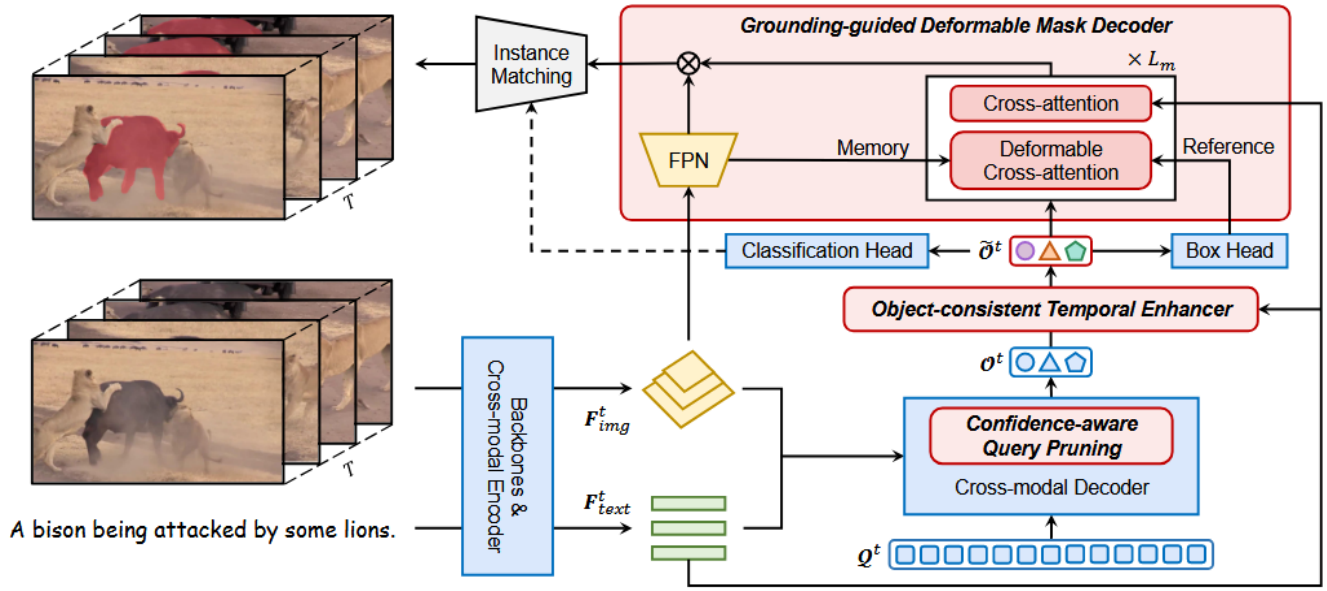


图 3　ReferDINO 的整体架构。蓝色标注模块继承自 GroundingDINO，红色标注模块为本论文新增组件。基于逐帧对象特征$\{O^t\}_{t=1}^T$，我们提出的对象一致性时序增强器利用跨模态文本特征实现帧间对象交互。随后，基于位置预测、跨模态文本特征及高分辨率特征图，接地引导的可变形掩码解码器为候选对象生成掩码。为提升视频处理效率，我们在跨模态解码器中引入了置信度感知的查询剪枝策略。

RVOS模型通常受限于现有数据的规模与多样性，泛化能力有限；而近期尝试将视觉定位基础模型（如GroundingDINO[8]）与分割模型（如SAM2[9]）进行简单组合的方法，虽然引入了开放世界知识，但这种非端到端的集成方式效率低下且不可微，同时缺乏对视频时序动态的捕捉能力，性能严重依赖于单帧检测质量。因此，本文提出了ReferDINO，这是一种端到端的RVOS框架，旨在直接继承预训练视觉定位基础模型强大的视觉-语言理解能力，并通过引入针对性的时序增强与分割模块，填补其在时序一致性和像素级分割上的空白，从而实现高效且精准的视频物体分割。

3.2 ReferDINO

ReferDINO 是一种用于指代视频物体分割（RVOS）的端到端框架，其核心思想是利用预训练的视觉定位基础模型 (如 GroundingDINO) 的强大视觉-语言理解能力，并针对视频任务引入三大创新模块：对象一致性时序增强器，利用文本相关性实现鲁棒的跨帧物体跟踪和时序特征聚合；定位引导的可变形掩码解码器，将定位框先验融入像素级分割过程，高效生成精确掩码；以及置信度感知的 Query 剪枝策略，在每一层迭代地移除低置信度的 Query，显著提高训练和推理效率。整体架构图如图 3。

目标一致性时序增强器

ReferDINO建立在GroundingDINO基础之上，但GroundingDINO是针对静态图像设计的，缺乏对视频时序动态的捕捉能力。在处理如“猫摆动尾巴”这类涉及动作描述的文本或面对视频中常见的运动模糊和受限视角时，仅靠单帧检测往往会导致物体识别不稳定甚至丢失。为了解决这一问题，本文设计了“目标一致性时序增强器”，旨在利用跨模态的文本表征来促进帧间的物体交互从而增强时序理解能力和物体的一致性。

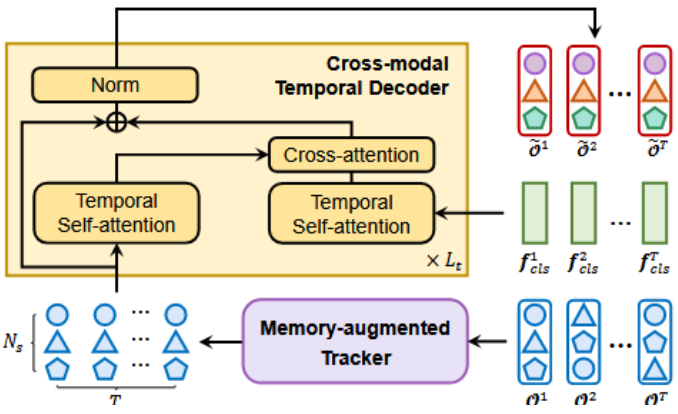


图 4　目标一致性时序增强器的示意图

如图4所示，该模块主要由记忆增强跟踪器（Memory-augmented Tracker）和跨模态时序解码器（Cross-modal Temporal Decoder）两部分组成。记忆增强跟踪器：传统的RVOS方法通常只考虑相邻帧之间的对齐，而ReferDINO引入了记忆机制以实现稳定的长期跟踪。首先，利用匈牙利算法基于余弦相似度将当前帧的对象O^t与上一帧的记忆M^{t-1}进行对齐：

$$\hat{O}^t = Hungarian(M^{t-1}, O^t)$$

其次，在更新记忆时，本文通过引入文本相关性来防止物体不可见的帧干扰长期记忆，公式如下：

$$M^t = (1 - \alpha c)M^{t-1} + \alpha c\hat{O}^t$$

其中α是动量系数，c是对象与其对应的句子嵌入之间的余弦相似度。这一设计的精妙之处在于，如果当前检测到的物体与文本描述的相关性较低（例如物体被遮挡或检测错误），c值变小，模型会自动保留更多的历史记忆M^{t-1}，从而保证跟踪的鲁棒性。

跨模态时序解码器：该部分负责提取视频级的物体信息。它将对齐后的物体嵌入$\{\hat{O}^t\}$和句子嵌入$\{f_{cls}^t\}$作为输入。通过在时序维度上应用自注意力机制，模型能够实现帧间的物体交互。随后，利用交叉注意力机制，以句子嵌入为Query，物体嵌入为Key和Value，提取出包含丰富时序信息的视频级物体表征 O_v^t。最后，通过以下公式增强逐帧的物体嵌入：

$$\tilde{O}^t = LayerNorm(\hat{O}^t + O_v^t)$$

这种设计使得单帧物体特征能够感知到整个视频上下文从而极大提升了对动作和时序变化的理解能力。

定位引导的可变形掩码解码器

视觉定位基础模型（如GroundingDINO）通常输出的是边界框，而非像素级的分割掩码。将检测框转换

为高质量的分割掩码是RVOS的核心难点。现有的方法（如Grounded-SAM2[10]）通过简单的模型堆叠来实现这一步，但这导致了流程不可微且效率低下。为了实现端到端的优化并充分利用预训练模型的空间定位知识，论文提出了“定位引导的可变形掩码解码器”。

该解码器旨在融合文本条件和定位预测来生成精确的掩码。其工作流程主要依赖于可变形交叉注意力（Deformable Cross-attention）和普通交叉注意力（Vanilla Cross-attention）的协同工作。

可变形交叉注意力：这是该模块利用预训练知识的关键。它以物体Query为查询对象，以FPN输出的高分辨率特征为记忆，并直接利用预测框的中心点作为参考点。通过这种方式，注意力机制不再需要全局搜索，而是聚焦于预测框附近的区域。这不仅提高了计算效率，更重要的是它显式地利用了GroundingDINO强大的定位能力来引导特征聚合，从而实现对物体边缘的精细化。

普通交叉注意力：为了确保分割结果符合文本描述，该层以物体Query为Query，以文本特征为Key和Value。这确保了掩码生成过程始终受到语言描述的约束，避免分割出与描述不符的背景或干扰物。

经过L_m层解码块的迭代优化后，得到的掩码嵌入与高分辨率特征图进行点积运算，最终生成实例掩码。

置信度感知的Query剪枝

基础视觉模型通常使用大量的Query（例如GroundingDINO使用900个Query）来存储丰富的物体信息，以覆盖开放世界中的各种目标。然而，在视频处理任务中，如果对每一帧都保留如此庞大的Query数量进行迭代解码，计算成本将呈指数级增长，严重限制了模型的推理速度和训练效率。直接减少Query数量又可能破坏预训练模型原有的知识结构。为此，本文提出了一种“置信度感知的Query剪枝”策略，在不牺牲性能的前提下显著降低计算负载。

如图5所示，该策略的核心思想是在跨模态解码器的每一层迭代地剔除低置信度的Query，只保留对当前任务最关键的Query子集。置信度评分s_j的计算复用了

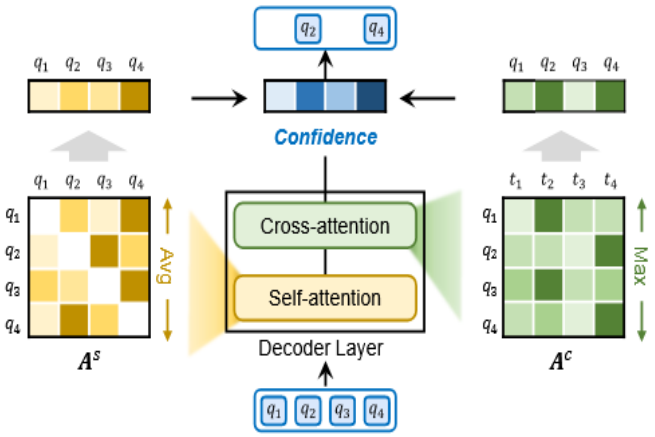


图 5　置信度感知的 Query 剪枝示意图

注意力权重，无需额外的参数，公式如下：

$$S_j = \frac{1}{N_l - 1} \sum_{i=1, i \neq j}^{N_l} \boldsymbol{A}_{ij}^{s} + \max_k \boldsymbol{A}_{kj}^{c}$$

该公式由两部分组成，分别衡量了Query的重要性和相关性。第一部分代表第j个Query从其他Query处获得的平均注意力权重。如果一个Query被其他Query频繁关注，说明它承载了不可替代的全局上下文信息（即“不可替代性”）。第二个部分则代表第j个Query与文本Token之间的最大匹配度。这一项直接衡量了该物体是否是文本所指代的目标（即“文本相关性”）。结合这两项指标，模型能够有效地识别出既在视觉上重要又符合文本描述的Query。在每一层解码后，模型会根据得分s_j过滤掉$p\%$（例如 50%）的低分Query，最终只保留极少量的Query进行后续计算。

3.3 实验结果与分析

详细的实验设置、定量或定性实验可参考相关方法原论文，下面仅对关键实验结果进行展示。

ReferDINO在与SOTA的比较中，显著地超越了所有五个RVOS数据集[11-14]上的SOTA性能。具体而言，在使用Swin-T骨干网络时，ReferDINO在高难度数据集MeViS上取得了48.0%的J&F评分，性能超过现有SOTA 1.6%。在Ref-YouTube-VOS数据集上，ReferDINO的J&F达到66.4%，将SOTA性能提升了2.8%。当采用更大规模的Swin-B骨干网络时，ReferDINO 在Ref-YouTube-VOS上的J&F进一步提

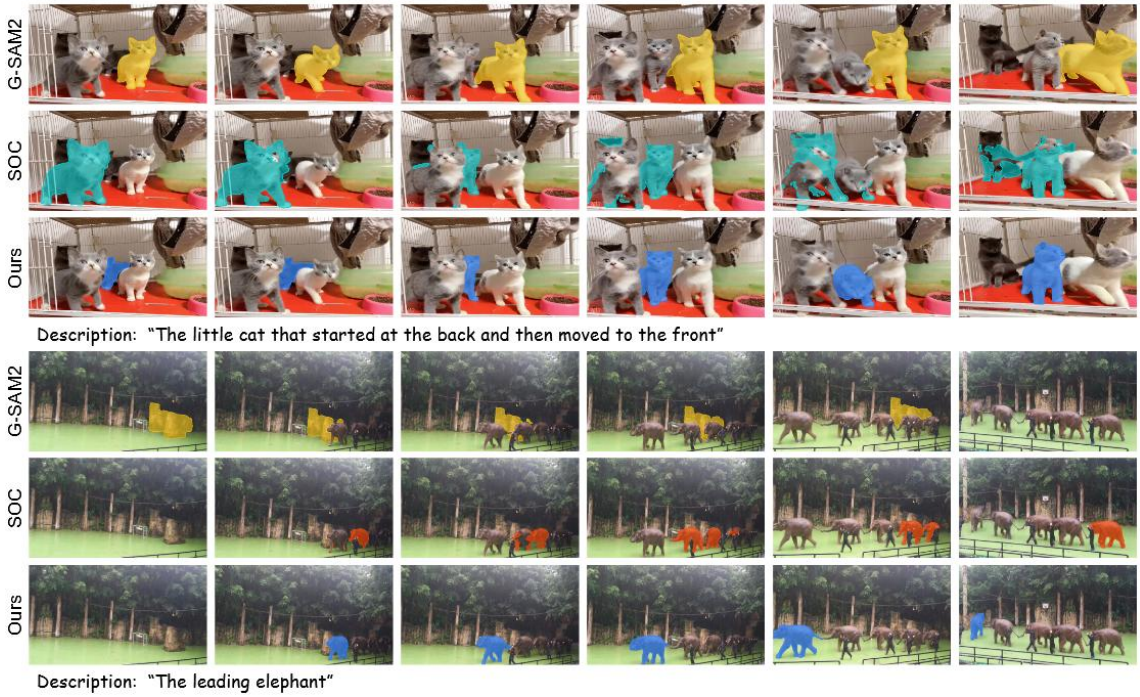


图 6　ReferDINO 和 SOTA 方法的定性比较

升至69.3%，领先SOTA超过2.2%。这种跨数据集的性能提升的一致性，有力地证明了 ReferDINO 模型的优越性。

为论证ReferDINO模块设计相对于 GroundingDINO的有效性，本研究将其与近期的基线模型Grounded-SAM2进行了对比。Grounded-SAM2 采用简单的模型集成方式，即利用 GroundingDINO 进行单帧物体识别，再将边界框输出输入SAM2进行视频全程分割。Grounded-SAM2的性能高度依赖 GroundingDINO的检测质量，例如，将 GroundingDINO从Swin-B切换到Swin-T导致Ref-YouTube-VOS上的J&F骤降10.4%。相比之下，ReferDINO在不同数据集和骨干网络上始终以显著幅度超越Grounded-SAM2。这表明，简单的模型集成无法有效解决RVOS任务，并且有力证明了ReferDINO在将GroundingDINO适配至RVOS任务时所引入方法（如时序增强和定位引导解码器）的有效性。

在图6中，我们将ReferDINO与Grounded-SAM2和SOC的定性结果进行对比。基于SAM2的Grounded-SAM2能够生成高质量的掩码，但缺乏对时序动态和运动的理解能力，经常导致所指对象的识别错误。SOC在对象识别方面表现更好，但其分割质量和时序一致性仍有不足。相比之下，我们的ReferDINO有效解决了这两方面挑战，实现了准确的所指对象识别、稳定的目标跟踪和高质量的对象分割。

四、视频定位

4.1 研究背景

时空视频定位 (STVG) 旨在根据自然语言查询从未剪辑视频中定位出目标对象的时空管道，是一项涉及复杂跨模态理解的挑战性任务。由于全监督学习依赖昂贵的细粒度边界框和时间标注，弱监督设置 (WSTVG) 近年来备受关注。现有的弱监督方法大多采用简单的后期融合策略，即生成独立于文本描述的候选管道后再进行匹配，这种方式忽略了视频区域与句法成分的对应关系，往往导致目标识别错误和跟踪不一致。虽然引入预训练视觉定位模型可提供空间线索，但简单的帧级连接难以捕捉连贯的时空语义。因此，本文提出了 TubeRMC 框架，通过引入基于管道条件的重构机制以及时空互约束策略，在无需细粒度标注的情况下有效捕获管-文对应关系，显著提升了定位性能。

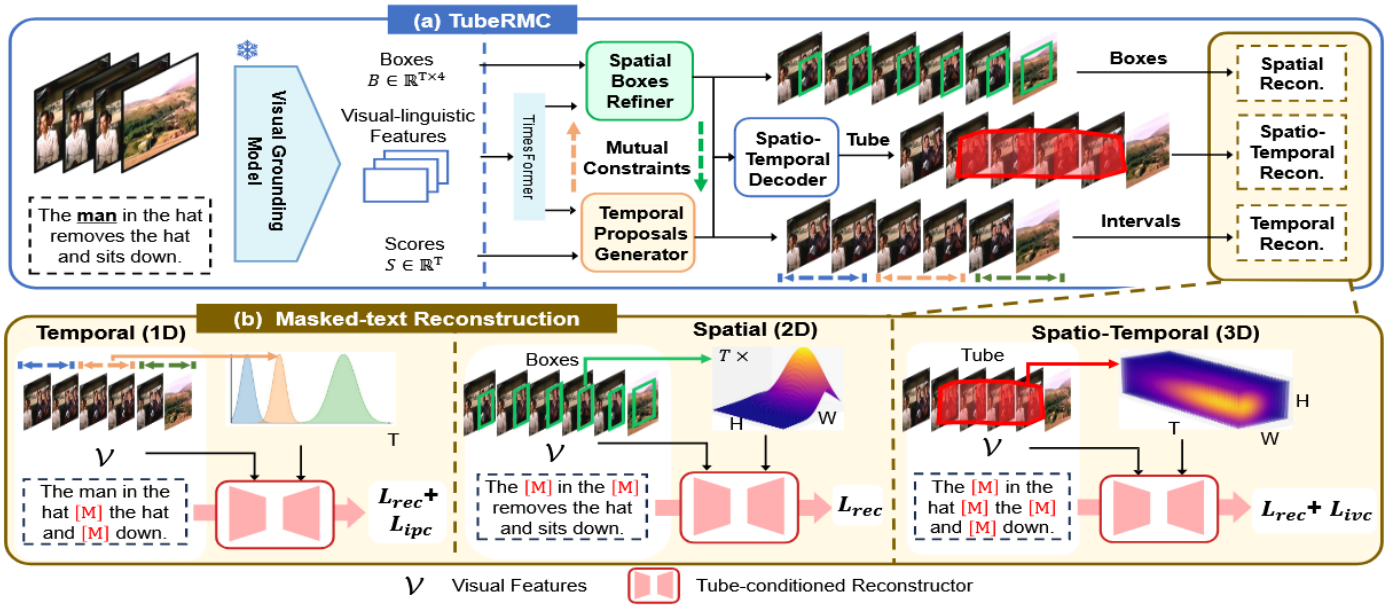


图 7 （a）TubeRMC 整体架构。该模型在逐帧中为主语词选择最相关的边界框，同时提取图像级视觉-语言特征（左图）；并生成时间区间（1D）、空间边界框（2D）与时空管道提案（3D）以进行重构（右图）。（b）管道条件重构学习通过掩码句子中包含动态、静态或整体事件信息的短语，随后基于 1D、2D 与 3D 提案重构被掩码内容。该方法进一步在 1D 与 2D 提案间建立相互约束，通过时间-空间与空间-时间的双向约束机制，增强提案生成过程中的时空一致性。

4.2 TubeRMC

TubeRMC 框架（如图 7 所示）旨在解决弱监督时空视频定位（WSTVG）中的管（Tube）-文本对齐问题，其核心是一个端到端的结构，由三个相互协作的模块构成：特征编码模块、提议生成模块和管条件重建学习模块。特征编码模块负责从视频和查询语句中提取视觉和语言特征。提议生成模块在弱监督下，通过互约束学习（Mutual Constraints Learning）机制——包括时空约束（Time-to-Space）和空时约束（Space-to-Time）——来不断细化和生成高质量的时空管提议。最关键的管条件重建学习模块，则将生成的时空管作为条件，利用管条件重构器（TR）在遮蔽语言建模（MLM）的框架下，通过时间、空间和时空三种重建策略，强制模型学习管提议与查询语句中关键语义成分（如动作和对象）的精细对应关系，最终通过最小化重建损失和互约束损失，实现高性能的弱监督时空视频定位。

管条件重建学习

管条件重建学习是TubeRMC框架的核心创新点，旨在无需细粒度标注的情况下，学习视频管道（tube）与文本描述之间精细的语义对齐关系。其基本思想是将管（视觉线索）作为条件，去重建查询语句（文本线索）中被遮蔽（masked）的关键线索，以此来增强管与文本的对应关系学习和事件理解。只有与目标事件高度匹配

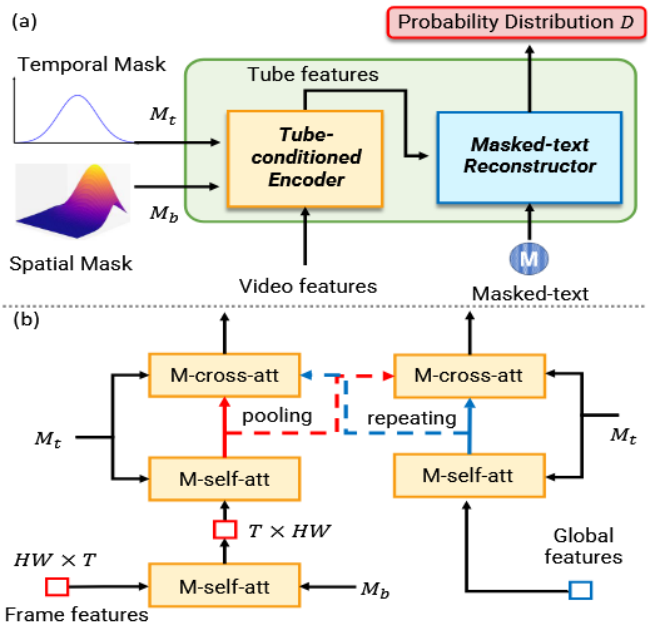


图 8 管条件重构器的架构：（a）总览；（b）管条件编码层的设计。

的管，才能成功重建出查询语句中被遮蔽的关键短语。

该机制的核心组件是管条件重构器（Tube-conditioned Reconstructor, TR）。如图8所示，TR能够同时处理时间和空间注意力掩码，从而捕获丰富的时空视觉-语言对应关系。为了实现梯度反向传播，模型将时空管、时间区间和空间边界框等提议转化为高斯分布来表示时空掩码。具体而言，时间提议转换为1D高斯函数生成时间注意力掩码$M_t \in \mathbb{R}^T$，空间边界框转换为2D高斯函数 $M_b \in \mathbb{R}^{HW}$，而时空管则转换为3D高斯函数来生成$T \times HW$大小的掩码。TR（如图 8(a)所示）通过一个局部分支和一个全局分支的编码器，使模型注意力集中在与高斯掩码对应的视觉区域，从而生成时空管特征F_T。最后，这些F_T经过遮蔽文本重构器（Masked-text Reconstructor）来预测被遮蔽的词汇分布D，并使用交叉熵损失$L_c(\cdot)$计算重建损失L_{rec}，其中L_{rec}综合了空间、时间、时空三种重建策略的损失。

三种重建策略

为了全面捕获时空管与文本之间多方面的对应关系，TubeRMC框架提出了三种从1D、2D到3D维度的重建策略，而非仅仅使用传统的时间重建。

时间重建：该重建主要关注运动感知的时间信息，选择性地遮蔽句子中谓语及其直接相关的名词。正向时间提议被转换为1D高斯时间注意力掩码M_t，用于条件重建。为实现无需时间标注的时空视觉-语言理解，模型引入了负样本提议并设计了提议间对比损失（Inter-Proposal Contrastive Loss, L_{ipc}）。该损失确保正样本的重建损失低于负样本的重建损失，从而增强模型对时间边界的识别能力。L_{ipc}的公式为：

$$L_{ipc} = \sum_{i=1, i\neq k^*}^{K} \max\left(L_c\left(D_t^{(k^*)}\right) - L_c\left(D_t^i\right) + \beta_1, 0\right) + \sum_{i=1}^{K} \max\left(L_c\left(D_t^{(k^*)}\right) - L_c\left(D_{tn}^i\right) + \beta_2, 0\right)$$

其中k^*是损失最小的正样本索引，β_1和β_2是边界超参数。

空间重建：该重建侧重于对象的外观信息和静态属性。通过遮蔽语句中的名词和形容词，模型被迫依赖于提议管中2D空间高斯掩码M_b所突出的视觉特征。

时空重建：该重建进一步捕获时空耦合的、动态的对应关系，是三种策略中最全面的。它使用3D高斯掩码作用于管条件重构器，以更高的概率随机包含动词、名词和形容词等一系列关键语义成分。

互约束学习

图 9　时空预测的可视化结果，其中粉色、黄色和绿色分别代表我们的方法、MDETR-Zero 和 ground truth

为了进一步提升用于重建任务的管提议（包括边界框和时间区间）的质量，TubeRMC框架引入了互约束学习机制。该机制包含时到空约束和空到时约束，它们共同提升了对象的一致性和时间预测的准确性。

空到时约束：该约束旨在利用空间定位的置信度分数来指导更精确的时间提议生成。框架首先利用空间边界框细化器（Spatial Boxes Refiner）输出的置信度分数。选择得分最高的K帧作为时间中点，并结合预设宽度生成初始时间提议。然后时间提议生成器（Temporal Proposals Generator）预测偏移量来得到最终的正向时间提议。同时，使用得分最低的 K 帧构建负向时间提议。空时互约束损失被定义为最小化不同正向提议之间、以及正向提议与负向提议之间的重叠，这促使时间边界与高置信度的帧对齐，提升了时间提议的质量。

时到空互约束：该约束确保同一场景相邻帧中目标保持空间上的连续性，对于实现一致的目标跟踪至关重要。在每个时间提议内，应用一个惩罚损失，对相邻帧中预测边界框的交并比低于某一阈值的情况进行惩罚。

这两种约束相互作用，时到空约束提高了预测框的帧间平滑度和一致性，而空到时约束则利用细化后的空间定位线索来优化时间边界的定位，从而显著提高了整体的时空管提议质量。

4.3 实验结果与分析

TubeRMC 在与现有 SOTA 方法的比较中取得了显著优势。在 HCSTVG-v1 数据集[15]上，TubeRMC 在核心指标 m_vIoU 上，相较于先前最佳的两阶段和单阶段方法，分别实现了 12.50%和 4.74%的性能提升。此外，它在 VidSTG[16] Declarative 和 Interrogative 子集上的表现也全面超越了所有对比方法，充分验证了 TubeRMC 设计的有效性。论文引入了两个新的基线：MDETR-Zero 和 MDETR+CPL[17]。虽然这两个基线已能达到或超越部分现有 SOTA，但由于它们未能有效捕获时空对应关系，性能仍存在局限。相比之下，TubeRMC 优势明显，例如在 HC-STVG-v2[15]上，它在 m_vIoU 指标上分别超越 MDETR-Zero 和 MDETR+CPL 8.43%和 5.55%。特别是在与预训练模型语料差异较大的 VidSTG Interrogative 子集上，TubeRMC 成功弥补了性能鸿沟，取得了优异成果，有力证明了其管条件重建和互约束设计的卓越性能和泛化能力。

我们展示了由 MDETR-Zero 基线模型与 TubeRMC 预测的多组可视化样本。如图 9 所示，MDETR 模型存在时序预测不可靠（第 1 行）与空间定位不稳定（第 2 行）的问题。相比之下，TubeRMC 能够有效捕捉管道-文本对应关系，提供精确的时空预测，这凸显了管道条件重构在弱监督时序视觉定位学习中的优势。第 3 行展示了一个挑战性案例：目标“戴墨镜的男子”出现剧烈的视角变化和物体遮挡，MDETR 错误地将边界框分配给另一名执行相似动作的男子。尽管受到 MDETR 错误结果干扰，TubeRMC 仍能给出相对准确的时空预测。

五、总结

本文针对多模态视频-文本定位领域面临的关键挑战，聚焦于解决视频-文本标注数据缺失和模型时序理解不足问题，提出了三个创新框架以突破现有方法的局限：针对弱监督视频段落定位，提出了 SiamGTR 框架，通过孪生学习机制和伪监督数据生成，在无需时间戳标签的情况下实现高效定位。针对指代视频目标分割任务，提出了端到端框架 ReferDINO，该模型继承了预训练视觉定位基础模型的强大能力，并引入对象一致性时序增强器和定位引导的可变形掩码解码器，有效解决了时序一致性和像素级分割难题。针对弱监督时空视频定位，设计了 TubeRMC 框架，其核心在于管条件重建学习和互约束学习机制，通过强制模型利用生成的时空管来重建查询语句中的关键语义成分，在没有细粒度标注的情况下有效捕获管-文对应关系。大量实验证明上述所提系列方法的有效性和优越性，助力多模态视频语言理解技术在更复杂和更具挑战性的实际场景中落地应用。

责任编委 张 青

参考文献

[1] Tan C, Lai J, Zheng W S, et al. Siamese learning with joint alignment and regression for weakly-supervised video paragraph grounding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13569-13580.

[2] Liang T, Lin K Y, Tan C, et al. Referdino: Referring video object segmentation with visual grounding foundations[J]. arXiv preprint arXiv:2501.14607, 2025.

[3] Li J, Zhang Y, Hu J F, et al. TubeRMC: Tube-conditioned Reconstruction with Mutual Constraints for Weakly-supervised Spatio-Temporal Video Grounding[J]. arXiv preprint arXiv:2511.10241, 2025.

[4] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[5] Yuan Y, Lan X, Wang X, et al. A closer look at temporal sentence grounding in videos: Dataset and metric[C]//Proceedings of the 2nd international workshop on human-centric multimedia analysis. 2021: 13-21.

[6] Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos[C]//Proceedings of the IEEE international conference on computer vision. 2017: 706-715.

[7] Regneri M, Rohrbach M, Wetzel D, et al. Grounding action descriptions in videos[J]. Transactions of the Association for Computational Linguistics, 2013, 1: 25-36.

[8] Liu S, Zeng Z, Ren T, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 38-55.

[9] Ravi N, Gabeur V, Hu Y T, et al. Sam 2: Segment anything in images and videos[J]. arXiv preprint arXiv:2408.00714, 2024.

[10] Grounded-SAM-2. https://github.com/IDEAResearch/Grounded-SAM-2.

[11] Ding H, Liu C, He S, et al. MeViS: A large-scale benchmark for video segmentation with motion expressions[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 2694-2703.

[12] Seo S, Lee J Y, Han B. Urvos: Unified referring video object segmentation network with a large-scale benchmark[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 208-223.

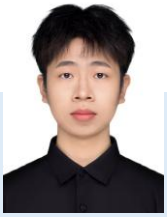
[13] Khoreva A, Rohrbach A, Schiele B. Video object segmentation with language referring expressions[C]//Asian conference on computer vision. Cham: Springer International Publishing, 2018: 123-141.

[14] Gavrilyuk K, Ghodrati A, Li Z, et al. Actor and action video segmentation from a sentence[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5958-5966.

[15] Tang Z, Liao Y, Liu S, et al. Human-centric spatio-temporal video grounding with visual transformers[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(12): 8238-8249.

[16] Zhang Z, Zhao Z, Zhao Y, et al. Where does it exist: Spatio-temporal video grounding for multi-form sentences[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10668-10677.

[17] Kamath A, Singh M, LeCun Y, et al. Mdetr-modulated detection for end-to-end multi-modal understanding[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1780-1790.



梁天铭

中山大学计算机学院博士研究生，导师为郑伟诗教授和胡建芳副教授，主要研究方向为视频-语言理解和生成。

Email:liangtm@mail2.sysu.edu.cn

谭超镭

中山大学计算机学院硕士研究生，导师为胡建芳副教授，主要研究方向为视频-语言理解，尤其是语言引导的视频定位问题。

Email: ctanak@cse.ust.hk

李谨轩

中山大学计算机学院硕士研究生，导师为胡建芳副教授，主要研究方向视频理解和视频-语言学习。

Email:lijx267@mail2.sysu.edu.cn

胡建芳

中山大学计算机学院副教授，博士生导师。主要研究方向为视频内容理解与生成，发表包括 TPAMI、CVPR、NeurIPS 在内的论文 70 余篇。先后主持国家自然科学基金面上项目、青年项目和广东省杰青等项目，获中国图像图形学学会优秀博士论文奖。

Email: hujf5@mail.sysu.edu.cn

顶会观察

NeurIPS 2025

南京大学人工智能学院副教授 叶翰嘉

神经信息处理系统大会（Conference on Neural Information Processing Systems）作为机器学习领域的顶级盛会，持续保持 CCF-A 类会议的权威地位，在全球学术界与产业界拥有深远影响力。第 39 届 NeurIPS 大会创新采用“双会场+全虚拟并行”的举办模式，分别于 2025 年 11 月 30 日至 12 月 5 日在墨西哥城希尔顿 reforma 酒店、2025 年 12 月 2 日至 7 日在美国圣地亚哥会议中心设立线下会场，并同步开放全球虚拟参会通道。会议日程涵盖多元学术环节，包括专题 Tutorial、行业博览会、核心学术会议、海报展示、前沿研讨会等，其中墨西哥城会场设置 6 场线下 Tutorial 与 8 个主题 Workshop，圣地亚哥主会场则围绕强化学习、AI 安全、大模型优化等核心方向开展 3 天正式会议及 2 天深度 Workshop，通过直播联动实现学术内容互通，为全球约 2.6 万余名注册参会者打造沉浸式、跨地域的学术交流体验。

一、会议亮点

首创“双会场+全虚拟”并行模式：本届 NeurIPS 会议突破往届单一举办地限制，首次在墨西哥城与圣地亚哥设立双线下会场，搭配全球虚拟参会通道，在满足不同地区参会者的线下交流需求的同时，又通过直播联动技术实现双会场内容同步互通，兼顾学术交流的沉浸感与覆盖广度。

评审机制革新：针对投稿量激增 37.7%，会议延续并强化“Responsible Reviewing”相关举措：对未履行评审义务的作者可采取限制 rebuttal 期间获取评审意见等措施；对严重失职（如低质/敷衍）评审者，其合著投稿可能被 desk-reject；明确 LLM 使用规范，要求作者披露相关使用情况。评审团队扩容至 2.2 万余人，通过优化校准流程维持 24.5%录用率，同时强化保密机制，保障评审公正安全。

学术赛道拓展：新增立场论文赛道，聚焦 AI 伦理等宏观议题；引入期刊整合赛道，精选 34 篇顶级期刊成果展示，打通资源壁垒。数据集与基准测试赛道投稿量达 1995 篇，成为独立核心赛道，推动 AI 研究向数据质量与基准公平性深化。

二、录用情况

投稿与录用：大会共收到了 **21,575 篇**有效投稿，相较于 2024 年增长了约 **37.7%**，十年间投稿量更是从 2015 年的 1,838 篇增长超十倍，创下历史最高纪录，充分彰显了该领域全球研究热度的爆发式提升。经过多轮严格评审，最终接收 **5,290 篇**论文，录用率约为 **24.52%**，延续了近年“投稿量激增但录用率稳定在 24%-26% 区间”的趋势。在录用的论文中，仅有 **77 篇**被选为 Oral Presentations，**Oral 率仅为 0.36%**，不仅显著低于往年水平，也远低于同领域其他顶会，成为本届会议极具含金量的学术荣誉。此外，另有 **688 篇**论文被选为 Spotlight（重点海报），约占投稿总数的 **3.19%**，其余 4,525 篇录用论文以普通 Poster 形式展示，整体录用结构凸显了会议对学术成果质量的追求。

评审团队：为了处理海量投稿，会议组织了庞大的评审团，包括 21921 名技术审稿人、641 名伦理审稿人、1985 名领域主席和 240 名高级主席，同时优化评审校准流程，在投稿量翻倍的情况下仍维持稳定录用率，

确保质量门槛不降低。

中国力量：据第三方统计，NeurIPS 2025 接收的 5290 篇主赛道论文中，**中国及港澳台区域机构参与的论文占比接近 50%**，成为全球贡献量最高的区域。这一比例较 2021 年（约 35%-40%）显著提升，反映出中国在机器学习与 AI 领域的科研生产力已实现从**重要参与者**到**核心引领者**的转变。中国内地高校与企业是 NeurIPS 2025 中国力量的核心参与主体。高校方面，清华、北大等顶尖院校稳居机构贡献榜前列，其中清华大学以 138 篇录用论文位列**全球第二、高校第一**；港澳台高校持续稳定输出，合计贡献约 80 篇录用论文。企业方面，阿里巴巴以 146 篇录用论文位居**全球企业第三**，华为、腾讯、字节跳动等合计超 200 篇，覆盖大模型、图神经网络、数据集基准测试等多个关键方向，且产出多篇高质量论文与口头报告成果。

三、热门研究方向

根据投稿和录用分布，NeurIPS 2025 呈现出以下技术趋势：

架构革新聚焦效率优化：彻底告别盲目堆参的发展模式，转向“高效智能”的核心诉求。线性复杂度架构（如 YAAD、MONETA）表现亮眼，在多项基准测试中反超传统 Transformer，打破其长期垄断地位。同时，动态计算分配、200 万 token 超长上下文窗口等技术突破，既解决了超长序列处理难题，又实现了计算效率的线性增长，让模型性能与部署效率同步优化。

智能体与具身智能全面崛起：智能体相关研究成为主流，“长上下文、效率优化、智能体/具身、安全与评测”等成为高频主题。核心探索聚焦大规模智能体协作（解决数千智能体的通信过载与目标对齐问题）与零样本机器人控制(如通过 YouTube 视频训练“世界模型”，实现机器人无专门数据即可模仿人类动作)，推动 AI 从“对话工具”向“可行动实体”演进，而多模态融合技术则为这一演进提供关键支撑。

研究导向转向数据与可信：AI 研究从“模型中心”逐渐转向“数据中心”，数据集与基准测试赛道成为核心领域，投稿量达 1995 篇，聚焦数据质量提升、基准测试公平性完善与研究可复现性。同时，AI 安全（鲁棒性、公平性、隐私保护）与 LLM 推理可控性成为刚需，通过测试时优化、门控注意力架构等技术，让模型输出更稳定、可追溯，筑牢学术诚信与技术落地的安全防线。

四、热点论文

本届 NeurIPS 会议共选出 **4 篇最佳论文**（含 1 篇数据集与基准赛道专属获奖论文）。

Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)[1]：大语言模型在生成创意内容时，常因趋向同质化而难以呈现人类风格的多样性，而当前可规模化评估这种多样性的方法仍较为匮乏。为填补这一研究空白，论文作者构建了大规模数据集 Infinity-Chat，该数据集包含 26K 条多样且源自真实世界的开放式用户查询，成为首个支持系统研究真实场景下开放式查询的核心资源。这项研究聚焦现代语言模型的多样性表现、价值多元性及社会影响，通过扎实的数据集构建与系统性分析，为理解和优化语言模型的生成同质化问题提供了关键支撑，具备重要且及时的学术与实践意义。

Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free[2]：随着大语言模型向更大规模、更长上下文方向演进，训练稳定性与注意力行为可控性已成为核心瓶颈。门控机制的有效性虽已得到广泛验证，但在注意力机制中的应用价值及规模化扩展能力尚未被充分探讨。通义千问团队通过一系列控制实验，系统性证实门控机制对大语言模型的优化作用，其核心优势源于增强注意力机制的非线性特征与提供输入相关的稀疏性。此外，该机制还能有效消除注意力池（Attention Sink）和巨量激活（Massive Activation）等问题，显著提升模型训练稳定性，大幅减少训练过程中的损失波动（loss spike），同时在长度外推任务上较基线模型实现明显突破。研究团队在不同尺寸、架构及训练数据规模的模型上验证了方法的通用性，并成功应用于 Qwen3-Next 模型。在当下 LLM 领域科学成果公开共享逐渐减少的环境中，作者公开分享大量工业级计算资源支持下的研究结果，将有力推动社区对大语言模型注意力机制的深入理解，

值得高度称赞。

1000 Layer Networks for Self-Supervised RL: Scaling Depth Can Enable New Goal-Reaching Capabilities[3]：该论文聚焦自监督强化学习的可扩展性问题，深入探索关键积木模块，指出网络深度是影响性能的核心因素。近年来多数强化学习（RL）研究依赖 2-5 层的较浅网络结构，而本文通过实验证实，将网络深度提升至 1024 层可显著增强模型性能。研究在无监督的目标条件设定下开展：不提供示范数据或奖励信号，要求智能体从零开始探索，自主学习最大化达成指令目标的概率。在模拟运动控制（locomotion）与操作（manipulation）任务的评估中，该方法在自监督对比式 RL 算法上实现 2×-50×的性能提升，且表现优于其他目标条件基线方法。增加模型深度不仅提高了任务成功率，还在质量上改变了智能体学到的行为模式。这项研究挑战了“强化学习提供的信息不足以引导深度神经网络中大量参数，大型 AI 系统应主要依赖自监督训练、RL 仅用于微调”的传统观点，提出一种新颖且易于实现的 RL 范式，使极深神经网络的有效训练成为可能，为强化学习的深度扩展提供了全新思路。

Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training[4]：本文探讨了扩散模型为何能避免对训练数据的死记硬背并具备良好泛化能力，这一核心问题尚未得到充分解答。该论文聚焦训练动力学在模型从泛化走向记忆转变过程中的作用，通过大量实验与理论分析，识别出两个关键时间尺度：较早的时间点t_g与更晚的时间点t_m。模型在t_g之后开始能够生成高质量样本，而超过t_m后则会出现记忆化（memorization）现象。研究围绕扩散模型的隐式正则化动力学展开基础性探索，将经验观察与形式化理论有机统一，得出强有力的研究结论，为理解扩散模型的泛化机制、优化训练过程提供了重要的理论支撑，对扩散模型的进一步发展具有关键指导意义。

五、大会获奖和竞赛奖

时间检验奖：时间检验奖是 NeurIPS 的重要荣誉，旨在表彰 **10 年前发表于 NeurIPS**、经长期实践验证对人工智能领域产生深远且持续性影响的奠基性研究成果。该奖项无固定首次设立年份，核心目的是回溯并认可那些突破当时技术局限、塑造后续研究方向，且在产业落地或学术探索中持续发挥价值的经典工作，为领域树立“经得住时间考验”的创新标杆。本年度获得该奖的是：

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[5]. 作者：任少卿、何恺明、Ross Girshick、孙剑。该论文创新性提出“区域提议网络（RPN）”，首次将目标检测的“区域提议”与“特征提取、分类回归”整合为端到端深度学习网络，解决了传统目标检测“速度与精度难以兼顾”的核心痛点——将检测速度从秒级提升至毫秒级，首次实现实时目标检测，为计算机视觉技术从实验室走向产业应用奠定基础。

Sejnowski-Hinton 奖：Sejnowski-Hinton 奖**首次设立于 2025 年**，其起源于人工智能领域两位先驱——Geoffrey Hinton 与 Terry Sejnowski 的长期合作与学术慷慨。奖项核心目的是表彰**过去 10 年内发表**、在“基于人工智能洞见的大脑计算理论”领域做出重大贡献，且对 NeurIPS 社区产生显著影响的研究，以此纪念二人的合作精神，并推动“脑启发人工智能”与“计算神经科学”的交叉发展。本年度获得该奖的是:

Random Synaptic Feedback Weights Support Error Backpropagation for Deep Learning[6]. 作者是 Timothy Lillicrap、Daniel Cownden、Douglas Tweed、Colin Akerman。论文聚焦“大脑神经回路如何仅通过局部信息高效学习”这一经典问题，提出“反馈对齐（Feedback Alignment）”机制——证明多层人工神经网络无需像反向传播那样依赖“精确的反向权重对称性”，仅通过固定的随机反馈权重即可有效学习。训练过程中，网络的前向权重会自然与随机反馈信号对齐，生成有用的损失梯度估计，首次为“权重传输问题”（真实神经元如何在无跨区域信息传递的情况下遵循损失梯度）提供了符合生物学合理性的具体解决方案。

六、总结展望

NeurIPS 2025 继续扮演全球 AI 研究的“风向标”。从本届日程与教程主题可见，研究重心正从单纯的规模扩张转向“在效率与可靠性约束下提升智能”：算力与能耗被当作一等设计指标，对齐与安全、可解释/可审计、隐私与合规等问题被更前置地讨论。

展望未来，架构创新将更紧密地与系统工程、数据与评测方法协同推进：一方面缓解长上下文、训练/推理成本与真实场景泛化的瓶颈，另一方面把可控性、可追溯性与人类价值约束内生到模型与流水线中，推动 AI 从“能跑分”走向“更稳、更省、更可用”规模化落地。

责任编委　魏秀参

参考文献

[1] Liwei Jiang, et al. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). NeurIPS 2025.

[2] Zihan Qiu, et al. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. NeurIPS 2025.

[3] Kevin Wang, et al. 1000 Layer Networks for Self-Supervised RL: Scaling Depth Can Enable New Goal-Reaching Capabilities. NeurIPS 2025.

[4] Tony Bonnaire, et al. Why Diffusion Models Don’t Memorize: The Role of Implicit Dynamical Regularization in Training. NeurIPS 2025.

[5] Shaoqing Ren, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NeurIPS 2015.

[6] Timothy P. Lillicrap, et al. Random Synaptic Feedback Weights Support Error Backpropagation for Deep Learning. Nature Communications 2016.

叶翰嘉

现任南京大学人工智能学院副教授、博导，在南京大学机器学习与数据挖掘研究所（LAMDA）从事学术研究工作；主持国家自然科学基金青年科学基金 B 类项目、国家重点研发计划专项项目，获中国人工智能学会吴文俊人工智能青年科技奖。

Email: yehj@nju.edu.cn

北京大学袁粒助理教授访谈

2026 年 2 月 2 日，《CCF-CV 专委简报》在线采访了北京大学博士生导师袁粒助理教授。下面是采访实录。

袁老师，您好！首先，请您分享一下您的个人学习和研究经历。

大家好!我于 2013 年进入中国科学技术大学 9 系，就读精密仪器与自动化专业。虽然本科并非计算机科班出身，但得益于系内在人工智能与视觉领域深厚的学术沉淀和优秀系友的影响，我从本科阶段就开始持续关注视觉与人工智能领域的前沿发展。2016 年我有幸赴澳大利亚交流学习，期间接触到中科大很多在海外 CV 届的前辈的研究工作，这进一步坚定了我将研究方向聚焦于视觉与人工智能的决心。

2017 年本科毕业后，我选择通过本科直博的方式，前往新加坡国立大学攻读博士学位。在博士期间，我有幸加入冯佳时教授与颜水成教授的团队，这段科研经历对我后续的研究取向和学术成长产生了深远影响。此外，我还曾赴哈佛大学，在 Livingstone Vision Lab 进行交流学习，拓宽了学术视野。

博士毕业后，我在工业界与学术界的发展机会之间进行了权衡，基于对“大算力驱动人工智能研究”这一趋势的判断——我认为算力与平台支撑在视觉乃至更广泛的人工智能研究中正变得愈发关键，最终于 2022 年加入北京大学深圳研究生院，并同时在鹏城实验室双聘开展科研工作。

您曾在新加坡国立大学攻读博士，又作为访问博士前往哈佛大学，这样的跨文化学术经历对您的研究视角和方法论带来了哪些独特的影响？

我觉得这段经历的影响，首先不完全是“跨文化”层面的，更准确地说是一次非常重要的“跨学科”经历。哈佛大学是一所综合性大学，而我当时所在的是脑科学相关实验室，它与人工智能，尤其是计算机视觉，在问题意识与研究方法上其实有着高度的共通性。

从更长期的研究视角来看，在脑科学实验室的学习，让我逐渐形成一种习惯：在进行人工智能研究时，会更自觉地从“大脑启发”与“机制解释”的角度，反过来审视模型的设计与能力边界。例如，我学习到的“雪貂实验”就给我留下很深印象——当视觉相关通路受损、听觉神经元接管后，动物经过一段时间适应，又能恢复某种功能表现。这让我意识到，很多时候底层神经网络的结构可能高度相似，真正拉开差异的是数据、任务与功能化方式，这些认知也持续影响着我后续的研究选择与方法论偏好。

您的研究涵盖了从计算机视觉、神经模型架构到多模态学习等多个领域，您是如何确定并不断扩展自己的研究主线的？

我的研究主线其实是沿着“以视觉为起点、围绕表征与架构展开、再走向以视觉为中心的多模态”这一脉络逐步扩展的。最早选择计算机视觉，一方面与我博士

阶段的经历密切相关——读博时我主要在冯佳时老师、颜水成老师团队中开展研究，而团队的核心方向正是视觉，因此以视觉作为切入点是非常自然、也最能积累方法论与成果的选择；另一方面，我也受到视觉领域前辈与学术氛围的持续影响，所以在研究起步阶段，我更倾向于先把视觉这条线做深做透。

在视觉研究的基础上，我进一步延伸到神经网络架构研究，本质上也是为视觉表征服务。因为我长期关注的是视觉神经网络的表征学习，而表征学习必然会牵引到模型结构。随后在博士期间恰逢 Transformer 快速兴起，我也较早开始尝试设计视觉 Transformer 相关结构，这段探索帮助我建立了“架构演进与表征能力之间强相关”的研究直觉。

至于为什么我的关注点会进一步转向以视觉为中心的多模态，我觉得主要有两类原因：一是研究发展上的客观推动。多模态不仅涉及视觉表征本身，也需要考虑与其他模态的融合方式、跨模态对齐，以及更高层的多模态推理等关键问题；这些问题本身就很有研究价值，也代表了领域的前沿方向。二是个人路径上的主动选择。在视觉领域里，我的导师与前辈都非常出色，因此我也希望在保持视觉优势的同时，拓展新的增长点，探索更广阔的研究边界。

您作为 ChatExcel 和 Open-Sora Plan 等项目的负责人或发起人，是如何想到将这些前沿技术开源，并推动它们被广泛使用的？

SORA 刚发布时，很多人会觉得它“很神秘、很遥远”。但我们当时的判断是，它的核心技术并非不可理解，我们发起 Open-Sora Plan 这个项目就是想打破这种认知。ChatExcel 更像是一个意外之喜，最初来自我一位博士生的朴素想法——希望通过大模型来降低处理各类表格的劳动量。虽说它与我的研究主方向并不完全一致，但我认为它足够有价值，具备很强的实用性。

所以，真正让我们决定发起开源的原因，一方面是普惠，希望把思路、框架、经验沉淀成可复用的资源，实现让技术真正帮到人的目标；另一方面也很现实，学术界往往缺资源、缺算力，所以我们希望主动推动和产业界合作，由企业提供算力与数据，我们提供技术与工程化能力，大家一起把项目做起来。开源在这里既是目标，也是手段，通过提高影响力来吸引更多伙伴加入，从而形成持续迭代的正循环。

您有多项专利实现成果转化，例如 13 项专利评估 4800 万入股企业，您如何看待学术研究与产业落地之间的关系？在推动技术转化过程中有哪些心得？

在成果转化过程中，我们始终致力于将团队的研究成果真正落地到企业的产品和业务场景中。很多时候，开源项目与产业合作并非割裂的，相反，开源往往能带动更紧密的产学研协作，进一步推动技术在真实系统里的应用与落地。

从更宏观的角度来看，尤其是在人工智能，特别是多模态方向，学术研究与产业转化之间的距离其实非常近。我也常跟组里同学强调两点：第一，业界的大厂在很多方向上已经跑得比学术界更快，我们不能再慢，必须和他们保持紧密合作。这种合作往往是双赢的，企业能提供算力和数据等关键资源，而我们的技术也能在企业中真正形成落地应用；第二，人工智能研究不应只停留在论文或概念层面，如果完全没有产业落地的验证，很多时候就很难证明它在当下环境中的有效性与价值，因此需要更深度地与产业界协同。

成果转化的前提是做“有用”的科研，这里的“有用”既可以是纯粹的学术价值，也可以是面向场景的工程落地，两者都很重要。但就人工智能而言，我更倾向于把研究往更可落地、更能解决实际问题的方向推进。

您负责多项国家级和校企合作重大项目，总经费超过千万。在领导这些大型科研项目时，您是如何协调团队、规划研究方向并确保项目高效推进的？

我认为，开展国家级和校企合作重大项目，首先离不开一个好的平台机制。如今，北大以及国内不少高校都呈现出一个积极的趋势——愿意让年轻人挑大梁，我非常感谢平台和团队给予的机会与支持。

至于项目推进中的协调，我的核心做法是“把学生放在正确的位置”，同时兼顾学生的发展节奏。大型项目往往周期长、工程量大，确实可能在某一阶段影响学生的论文发表。项目不仅要跑通工程链路，还要持续产出高质量的研究成果。以 Open-Sora Plan 为例，我们在推进项目的同时，也形成了一系列系统性工作，累计产出了接近二十篇顶会论文。具体到学生层面，我会让每个人负责清晰的模块，确保他们既能对项目做出实质性贡献，也能在自己的研究方向上实现创新、形成学术论文发表。

最后，我想特别强调，我们非常重视开源。开源能最大化科研影响力，让学生的成果在业界被真正看见。影响力越大，被认可、被支持、被合作的空间也就越大。

作为博士生导师和两门研究生课程的授课教师，您在培养学生方面有哪些独特的理念或方法？您认为优秀的青年研究者应具备哪些素质？

要说“独特”，其实谈不上，我更多是将自己求学与科研路上的心得体会传递给学生。在授课中，我始终遵循“不把简单问题复杂化，而是要把复杂问题简单化“的思路，将复杂问题拆解为层层递进的逻辑结构，让同学们能够循序渐进地理解和掌握。

在学生培养方面，我认为优秀的青年研究者至少要具备三个能力：第一是**提出问题**，这是科研的起点，也是大家普遍具备的意识；第二，也是最核心的能力，是**定义问题**，即把模糊的想法清晰化、具体化；第三才是**解决问题**。很多时候，只要前两步走得扎实，解决问题的过程便会水到渠成。我之所以特别强调 “定义问题”，是因为这一步直接决定了研究的核心方向，以及我们将用何种尺度去衡量和解决问题。

您担任多个顶级会议和期刊的编辑、领域主席和审稿人，这些学术服务工作对您个人和研究团队有哪些积极影响？您如何看待学术服务对学科发展的作用？

我始终觉得，学术服务是每一位学者都应主动承担的责任。从学科发展的角度来看，负责任的审稿与审核工作，是推动整个领域稳健、健康发展的关键。当审稿足够严谨、判断足够客观公正时，反馈给投稿人的意见才会更精准、更具建设性，从而真正帮助作者发现不足、完善工作，共同守护学术研究的质量。

对我个人及团队而言，这份工作也带来了两方面的直接增益：其一，在处理大量稿件的过程中，我会特别留意优秀论文在结构组织、叙事逻辑和表达方式上的长处，并将这些好的方法总结出来，传递给团队的学生，助力他们提升学术写作能力；其二，这让我能够近距离洞察领域动态，清楚把握一段时间内学界的研究热点、方法演进方向以及具有潜力的创新思路，从而更早捕捉到前沿趋势，并反过来精准校准我们团队的研究布局。

您曾入选福布斯亚洲 30U30 名单、获得国家优秀留学生奖等多项荣誉，这些成就中哪一项对您来说意义最特别？能否分享一些小故事？

在众多荣誉中，我反而认为**广东省人工智能科技进步一等奖**的意义更为特别。它更有意义的原因有两点：其一，它代表了产业界与区域创新体系对我们研究成果的肯定，是学术价值与应用价值结合的直接体现；其二，这份成果大多积累于我入职北大、成为独立 PI 之后，更像是对我这几年独立开展科研工作的阶段性总结。

此外更让我感到骄傲的是学生培养方面，我申请教职时并没有博士后研究经历，在北大校级评审会上，也有老师曾提出疑问“没做过博后就当博导，能不能带好博士生？”正是因为有过这样被质疑的时刻，当我们最终斩获这项奖，以及今年我第一届博士生毕业拿到各方面的机会时，让我格外感慨，它仿佛是对当年那份质疑

的一份答卷，证明了我在入职后的这几年，交出了一份还算合格的成绩。同时，我也希望这段经历能为未来同样没有博士后经历、却想回到高校从事科研工作的人，提供一个可参考的先例。

您目前在多模态大模型、神经架构设计等领域已取得显著成果，未来 3-5 年，您和团队计划在哪些方向进行重点突破呢？

未来 3-5 年，我和团队的核心突破方向，毫无疑问是多模态统一。从工业界的技术路线选择中，我们已能清晰看到趋势：真正的原生多模态系统，正朝着更统一的方向演进，但目前远未实现完全统一，仍有大量核心问题亟待解决。

在我看来，多模态统一绝非简单地将多种输入融合，更关键的是要回答三个核心难题：一是梳理理解和生成各自存在的关键瓶颈，二是探索理解与生成如何在同一套体系中实现深度融合起来，三是验证生成侧编码器与理解侧编码器能否真正实现统一。

现在很多所谓的“世界模型”系统，本质上是语言、视觉、视觉理解、规划、动作生成等多个模型的串行拼接，流程长、推理慢，而且每串行衔接还会造成信息损失，最终导致决策与理想状态存在明显差距。而“原生统一”的核心价值，就在于让复杂任务尽可能在同一个模型中完成端到端的处理，而非依赖多个模型反复接力。

因此，面向未来，我们将围绕“三层结构”推进多模态统一的突破。首先，实现多模态理解编码器与生成编码器的统一；其次，在 backbone 架构设计层面，研发全新的神经网络架构以及计算路由架构，缓解不同模态表征在同一骨干网络中的摩擦冲突；最后，聚焦面向不同任务的解码器设计，实现架构与任务的精准适配。

如果吐露研究工作者的心声，您最想说的是什么？

科研既可以“有用”，也可以“快乐”，这条路注定伴随着反复的试错与过程的艰辛，但当研究成果落地真实世界、创造出实际价值时，那份成就感与喜悦，会变得格外踏实而厚重。

责任编委 余烨 赵振兵

袁粒

北京大学深圳研究生院助理教授/研究员，博士生导师，主要围绕视觉为中心的多模态学习开展研究。作为多模态大模型应用与开源计划推动者，发起万星开源项目 Open-Sora Plan，并建设多模态开源生态（课题组开源十余项多模态算法和模型 https://github.com/PKU-YuanGroup，两年累计 Star 超三万，Huggingface 累计下载量超两千万）。学术方面以一作及通讯发表 Nature 子刊/TPAMI/IJCV/CVPR 等 50 余篇，谷歌学术引用 19000+。主持 2030 新一代人工智能重大项目课题、海外优青及华为昇腾科研创新使能重大计划等。获广东省卓越人工智能科技进步一等奖（2/11）、福布斯亚洲 30U30、国家优秀留学生奖等荣誉。多次任 ICLR/ICML/ACL 等领域主席，并受邀为 Nature、IEEE TPAMI 等期刊审稿。授权/申请发明专利多项，并完成成果转化：多模态视觉模型 13 项专利评估四千余万转移入股企业，ChatExcel 技术转移并获数百万技术捐赠。

委员好消息

- 2026 年 1 月 6 日，2025 年中国人工智能学会-联想蓝天科研基金入选名单揭晓，CCF-CV 专委会执行委员、南京大学**叶翰嘉**等的成果《基于模型表示的端云多模态大模型路由技术研究》入选。

- 2026 年 1 月 26 日，2025 中国电子学会科学技术奖公布、CCF-CV 专委会 10 位执行委员获奖。北京航空航天大学**刘祥龙**、国防科技大学**刘丽**入选青年科学家计划。清华大学**鲁继文**等的成果《视觉信息端侧智能分析关键技术及应用》入选技术发明一等奖。中国科学院计算技术研究所**王瑞平**等的成果《数实空间融合的社会安全风险感知防控关键技术与应用》；北京科技大学**殷绪成**等的成果《面向开放环境的低质图像表征与复杂图文识别技术及应用》；北京交通大学**白慧慧**等的成果《工业视觉智能关键技术及应用》获科技进步一等奖。重庆邮电大学**高新波**等的成果《边缘智能驱动的通信计算协同理论与方法》；华中科技大学**高常鑫**等的成果《复杂动态场景精细化视觉理解理论与方法》获自然科学二等奖。北京大学**施柏鑫**等的成果《超大型多场景跨域网络差异化体验运营关键技术及产业化应用》；电子科技大学**张乐**等的成果《高效协作的智能视联网关键技术及规模应用》获科技进步三等奖。

- 2026 年 1 月 27 日，中国电子学会 2025 年第四季度高级会员评定结果公布，CCF-CV 专委会北京航空航天大学**刘祥龙**、清华大学**鲁继文**、重庆大学**张磊**、北京科技大学**徐婧林** 4 位执行委员入选。

- 2026 年 2 月 2 日，中国电子学会-腾讯 Robotics X 犀牛鸟专项研究计划（2025）入选项目发布，CCF-CV 专委会 3 位执行委员入选。中国科学院自动化研究所**申抒含**的项目《面向具身数据增强的高斯世界模型研究》；上海交通大学**卢策吾**的项目《面向具身智能的机器人操作数据增广与人类数据迁移方法研究》；北京科技大学**徐婧林**的项目《面向意图理解的语言-肢体动作跨模态语义解析及人机协同研究》入选。

- 2026 年 2 月 25 日，2025 中国自动化学会科学技术奖评审结果公布，CCF-CV 专委会 3 位执行委员获奖。大连理工大学**刘日升**获 2025 中国自动化学会青年科技奖。山东财经大学**蹇木伟**等的成果《视觉显著性目标检测与自然场景分割方法研究》；上海大学**王日英**等的成果《多源非理想信息下随机跳变系统控制》获自然科学奖二等奖。

- 2026 年 2 月 27 日，2025 年度中国人工智能学会激励计划入选名单揭晓，CCF-CV 专委会 8 位执行委员入选。山东大学**张伟**等的成果《“四维自主，多元协同”：拔尖创新人才培养模式的系统性改革与实践》；北京理工大学**付莹**等的成果《价值引领-前沿融入-产教融通的计算机视觉教学创新探索与实践》；北京航空航天大学**徐迈**等的成果《国际视野与工程实践融通的电子信息-人工智能复合型人才培养模式》入选 CAAI 教学成果激励计划一类成果。燕山大学**吴培良**等的成果《地方高校人工智能新工科人才培养模式探索与援疆迁移》；福州大学**赵铁松**等的成果《AI 驱动的“四育协同、五维贯通”电子信息人才培养模式创新与实践》入选 CAAI 教学成果激励计划二类成果。清华大学**鲁继文**所指导的《面向视觉内容理解的高效深度神经网络模型研究——依托于大数据芯片》；北京理工大学**付莹**所指导的《基于领域先验的图像分割方法研究》；四川大

学**彭玺**所指导《面向复杂数据的深度聚类方法研究》；华中科技大学**王兴刚**所指导的《基于知识迁移的高效神经架构搜索与设计》入选 CAAI 博士学位论文激励计划。

✪ 2026 年 2 月 27 日，2025 年度吴文俊人工智能科学技术奖奖励公布，CCF-CV 专委会 21 位执行委员获奖。清华大学**鲁继文**、上海科技大学**高盛华**等的成果《视觉信息相似性协同度量理论与方法》；大连理工大学**刘日升**、大连理工大学**樊鑫**等的成果《异模态信息融合的数据对齐与任务协同理论方法》；东南大学**耿新**、南京大学**吴建鑫**等的成果《标记分布学习理论与方法》；重庆邮电大学**高肖斌**、西北工业大学**韩军伟**、重庆邮电大学**高新波**等的成果《多模态图感知、表征与安全》；浙江大学**杨易**等的成果《跨媒体模型协同的多重知识表达理论与方法》获自然科学奖一等奖。中国科学院大学**黄庆明**等的成果《视频层次化表示与智能编码》；深圳大学**雷柏英**等的成果《面向智慧医疗的多模态计算理论与方法》；武汉大学**叶茫**、中山大学**郑伟诗**等的成果《增广驱动的跨时空目标感知理论与方法》获自然科学奖二等奖。山东大学**张伟**等的成果《感算控一体化机器人通用控制器关键技术及应用》获技术发明奖一等奖。东北大学**杨金柱**等的成果《面向重大疾病的多模态影像智能诊断关键技术及应用》；上海交通大学**马超**等的成果《星载遥感影像智能处理关键技术及应用》获科技进步奖一等奖。中山大学**操晓春**等的成果《面向海上安防的通感算一体化大数据智能处理关键技术及产业化》；同济大学**赵才荣**、华中科技大学**白翔**、同济大学**何良华**等的成果《电力作业安全风险智能监测与防控关键技术及应用》；北京邮电大学**何召锋**等的成果《知识增强的可信多模态交互关键技术及应用》获科技进步奖二等奖。上海大学**曾丹**等的成果《面向高端制造的智能化检修机器人关键技术与应用》获得科技进步三等奖。

✪ 2026 年 3 月 16 日，2025 年度中国电子学会博士/硕士学位论文激励计划结果公布，CCF-CV 专委会 2 位执行委员指导的工作获选。中山大学**操晓春**所指导的《基于解耦知识蒸馏的目标检测增量学习》；南京大学**王利民**所指导的《面向过程式长视频的规划与理解方法研究》入选 2025 年度中国电子学会硕士学位论文激励计划。

责任编委 徐天阳

基于扩散模型的风格迁移方法及开源代码

香港理工大学 付陈平 大连理工大学 樊鑫

扩散模型（Diffusion Models），例如去噪扩散概率模型，近年来在图像生成、图像编辑以及文本条件生成等任务中取得了显著成功。这些研究表明，与变分自编码器、自回归模型、流模型以及生成对抗网络等其他生成模型相比，扩散模型能够生成更高质量的结果。随着扩散模型的发展使其逐渐适用于风格迁移任务。一些方法利用基于文本的算法，通过对预训练扩散模型进行提示控制，在无需微调模型的情况下实现图像风格化。这类方法通过文本描述引导生成过程，并通过提示工程和潜空间操作来实现风格迁移。下面，本文将简单介绍 2 篇基于扩散模型的风格迁移方法及其代码。

1、StyleDiffusion 方法

介绍：给定一幅参考风格图像，风格迁移的目标是将其中的艺术风格（如颜色分布和笔触纹理）迁移到任意内容图像上。为了实现这一目标，首先需要有效地将风格信息与内容信息分离，随后再将分离得到的风格信息迁移到新的内容图像中。由此带来了两个基本挑战：

(1) 如何实现内容与风格的解耦；

(2) 如何将提取的风格有效地迁移到另一幅内容图像上。

针对上述问题，文章提出一种“内容–风格”解耦风格迁移框架，其核心思想是显式提取内容信息，同时隐式学习其互补的风格信息，从而实现具有良好可解释性和可控性的内容–风格解耦与风格迁移。具体而言，文章在 CLIP 图像空间中引入一种基于 CLIP 的风格解耦损失，并结合风格重建先验，以协同实现内容与风格的解耦。进一步地，通过利用扩散模型在风格去除与图像生成方面的强大能力，使框架不仅能够取得优于现有最先进方法的结果，还能够实现更加灵活的“内容–风格“解耦以及二者之间的权衡控制。 方法流程见图 1。

论文地址： https://arxiv.org/pdf/2308.07863.pdf

代码地址：https://github.com/rafaelheid-it/StyleDiffusion

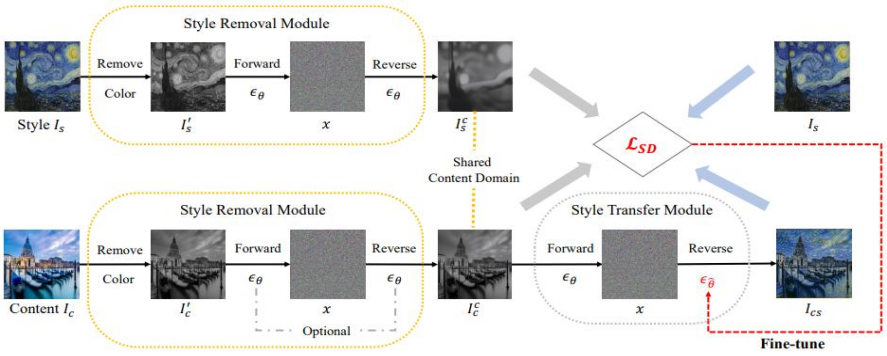


图 1 StyleDiffusion 方法框架图

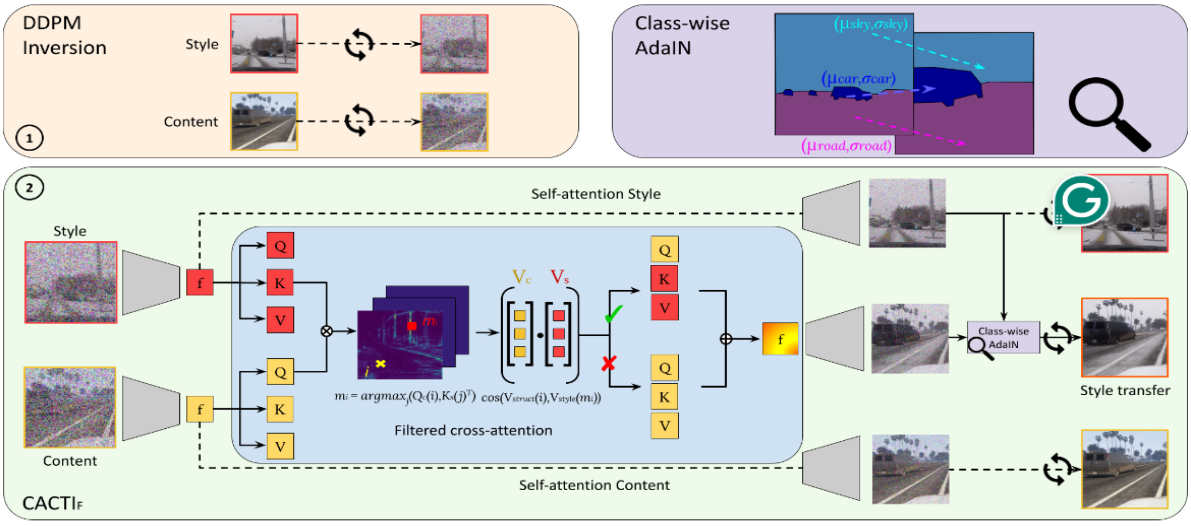


图 2 CACTI 方法框架图

2、Swin-UMamba 方法

介绍：近年来，扩散模型的快速发展显著提升了图像生成能力，在内容生成的质量与灵活性方面展现出前所未有的表现。这类模型通过迭代去噪过程逐步生成图像，为风格迁移提供了新的强大机制，相较于以往方法，可能更有效地缓解合成到真实域差异问题。然而，将基于扩散模型的风格迁移专门应用于域适应仍面临一些独特挑战，这些问题目前仍缺乏系统研究。

本文研究了基于扩散模型的风格迁移在语义分割中的合成到真实领域自适应问题，特别关注仅具有少量目标域数据的少样本场景。现有方法主要存在两个局限：

(1) 传统风格迁移方法通常采用全局变换，忽略语义类别边界，导致同一目标类别内部出现不真实的外观变化；

(2) 直接应用跨注意力机制在不同域之间结构对应关系较弱时，容易产生伪影。

针对上述问题，文章提出了两种新的技术方法：首先， CACTI 根据语义类别选择性地应用统计归一化，从而在语义边界内合理地迁移风格特征。其次，CACTI 在此基础上进一步扩展，通过基于特征相似度选择性地应用跨注意力机制，避免在结构不一致区域进行直接风格迁移，从而减少由此产生的伪影。实验结果表明，所提出的方法不仅能够生成视觉上连贯且真实的图像，而且在用于域适应时能够显著提升语义分割性能。方法流程见图 2。

论文地址： https://arxiv.org/pdf/2505.16360

代码地址： https://github.com/echigot/cactif

责任编委 贾同 王田

付陈平

博士后，香港理工大学，研究方向为计算机视觉，目标检测，图像增强。

樊 鑫

博士生导师，大连理工大学国际信息与软件学院从事教学与科研工作，担任软件学院院长。研究方向为计算机视觉与图像处理、医学影像分析。

个人主页：http://faculty.dlut.edu.cn/Xin_Fan/zh_CN/index.htm

具身智能操作数据集

北京航空航天大学 彭程浩 蔡锦涵 王田

具身智能强调智能体通过与环境的物理交互来感知、理解和行动。与传统计算机视觉以“识别”为代表的任务不同，具身智能不仅要“看懂”，更要“动手”——在真实或仿真的物理世界中触摸、抓取、推动、装配，通过试错与反馈不断习得灵活的操作技能。正如人类婴儿通过抓握玩具、堆叠积木逐渐理解重力、摩擦与因果关系，具身智能体也必须在动态交互中建立对物理世界的认知。构建高质量的机器人操作数据集远比收集静态图像或剪辑视频复杂得多，每一次操作示范都需要精细记录多模态信息：视觉、力触觉、关节状态，同时还要标注动作序列、物体姿态变化、任务意图乃至失败尝试。本简报将介绍几个具有代表性的具身智能操作数据集，以期为相关研究提供参考。

1、真实世界机器人操作数据集 RH20T

图 1 RH20T 提供的多视角机器人操作图

介绍： RH20T 是一个面向真实世界的机器人操作数据集，包含超过 20,000 条操作示范，涵盖抓取、放置、推开、折叠、倒水等多种日常任务。数据采集自多个真实机器人平台，配备多视角视觉、力觉、触觉等多模态传感信息，图 1 展示了某一任务下采集的视觉图像。该数据集可用于模仿学习、多任务学习与跨场景泛化研究。

RH20T 通过远程操作方式采集人类示范，采集环境涵盖家庭、办公室、厨房等多种场景。每个操作序列都经过精细化标注，包括物体姿态、操作意图、动作标签等。数据集还提供了跨场景、跨物体的泛化评估任务。图 2 展示了该数据集中的一个任务与采集的图像。

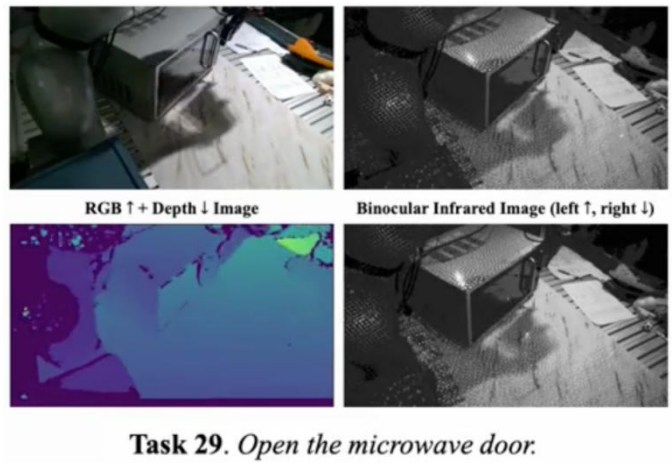


图 2 RH20T 任务示例

为支持多模态学习，RH20T 同步采集了 RGB-D 视频、触觉信号、关节角度、力反馈等信息，并确保了多模态数据的对齐。其包含内容如**错误!未找到引用源。**所示。

Modal	Size	Frequency
RGB image	1280 x 720 x 3	10 Hz
Depth image	1280 x 720	10 Hz
Binocular IR images	2 x 1280 x 720	10 Hz
Robot joint angle	6 / 7	10 Hz
Robot joint torque	6 / 7	10 Hz
Gripper Cartesian pose	6 / 7	100 Hz
Gripper width	1	10 Hz
6-DoF Force/Torque	6	100 Hz
Audio	N/A	30 Hz
Fingertip tactile	2 x 16 x 3	200 Hz

图 3 RH20T 数据集包含内容

项目主页（包含数据集）： https://rh20t.github.io/

论文链接： https://arxiv.org/abs/2307.00595

代码链接： https://github.com/rh20t/rh20t_api

2、大规模分布式数据集 DROID

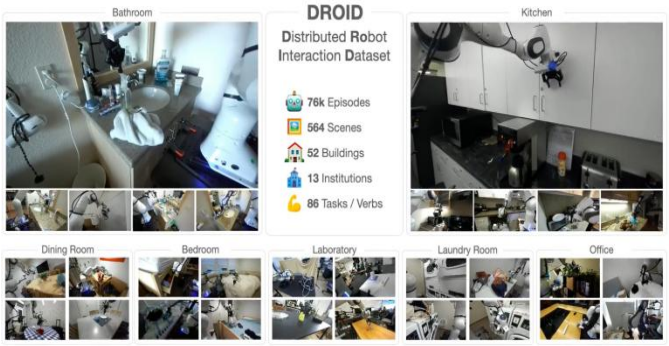


图 4 DROID 数据集示意图

介绍： 如图 4 所示，DROID（Distributed Robot Interaction Dataset）是一个由多所研究机构联合采集的大规模真实世界机器人操作数据集，包含超过 20,000 条示范轨迹，涵盖 86 种任务，涉及家庭、办公、工业等多种场景。数据采集采用分布式架构，在多个实验室同步进行，确保场景和物体布局的多样性。配备了多视角 RGB-D 相机、腕部力觉传感器和关节状态记录。DROID 采集自全球 12 个实验室，使用了多种机器人平台，如图 5 所示。数据集强调“分布外泛化”测试，设计了跨场景、跨物体、跨光照等泛化任务，为评估机器人操作模型在真实世界中的鲁棒性提供了重要基准。

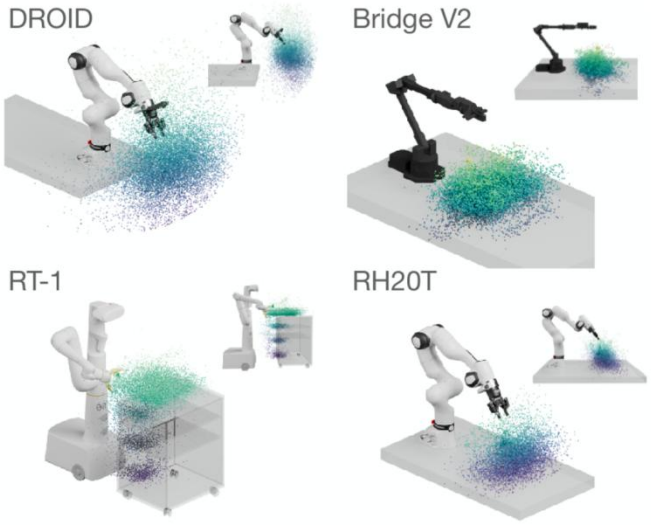


图 5 DROID 的多个机器人平台的点云交互示意图

DROID 的另一个亮点是其高精度的标注与一致性控制。所有示范数据都经过时间对齐和动作分割，并提供了详细的物体姿态、任务标签、场景描述等元信息。数据集还配套了标准化的训练及评估脚本和基线模型，支持研究人员快速复现和比较。

项目主页： https://droid-dataset.github.io

论文链接： https://arxiv.org/abs/2403.12945

代码链接（硬件代码）： https://github.com/droid-dataset/droid

代码链接（策略学习代码）： https://github.com/droid-dataset/droid_policy_learning

数据集使用方法： https://droid-dataset.github.io/droid/the-droid-dataset

介绍： 如图 6 所示，Open X-Embodiment 是迄今为止规模最大、最具多样性的跨机器人操作数据集，汇集

3、跨平台数据集 Open X-Embodiment

了来自全球 21 个机构的 60 余个数据集的示范数据，总计超过 100 万条轨迹，涵盖了抓取、放置、推拉、折叠、倒水、工具使用等超过 100 种任务和 Franka、UR、

Panda、Kuka、Fetch、Aloha 等 20 种不同的机器人平台。其核心目标是推动“通用机器人策略”的研究，使模型能够跨越机器人硬件、场景和任务差异进行泛化。

图 6 Open X-Embodiment 数据集

数据集统一了不同来源的数据格式，提供了标准化的观察空间（如 RGB-D 图像、关节位置、力觉）和动作空间（如末端执行器位姿、关节速度），并设计了跨平台的模型训练与评估协议。

Open X-Embodiment 不仅提供了海量数据，还配套发布了基于 Transformer 的通用策略模型 RT-1-X 和 RT-2-X，其结构如图 7 所示。

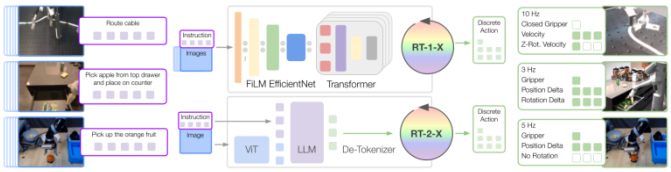


图 7 RT 模型结构示意图

这些模型在多个下游任务中表现出优异的零样本泛化能力，验证了大规模异构数据对具身智能的重要性。

项目主页： https://robotics-transformer-x.github.io

论文链接： https://arxiv.org/abs/2310.08864

代码链接： https://github.com/google-deepmind/open_x_embodiment

数据集链接：
https://docs.google.com/spreadsheets/d/1rPBD77tk60AEIGZrGSODwyyzs5FgCU9Uz3h-3_t2A9g/edit?gid=0#gid=0

责任编委 李策 樊鑫

彭程浩

北京航空航天大学人工智能学院，硕士研究生，研究方向为边缘计算、具身智能等。

蔡锦涵

北京航空航天大学人工智能学院，硕士研究生，研究方向为多模态大模型、视频异常检测等。

王　田

教授，博士生导师，北京航空航天大学人工智能学院系主任，研究方向为模式识别与智能系统、异常检测、类脑计算等。

好 文 推 荐

北京航空航天大学的最新研究成果“Beyond Heat Dissipation: Optimizing Diffusion Models in Frequency Domain”发表于 IEEE TPAMI 2026。

论文：Wang Q, Zhao Y, Li J. Beyond Heat Dissipation: Optimizing Diffusion Models in Frequency Domain, IEEE TPAMI, 2026.

当前扩散模型虽然在图像生成任务中取得了显著成功，但大多数标准扩散模型采用的是逐像素、各向同性的退化方式，这种建模方式没有充分利用图像本身在频域上的多尺度结构特性。为弥补这一不足，研究者提出了基于半正定退化(Positive Semi-Definite Degradations, PSD)的广义扩散模型，如 heat dissipation 和 blur diffusion，这类方法能够在频域中体现从低频到高频、由粗到细的生成过程。然而，已有 PSD-based 方法大多停留在退化构造层面，对其训练过程背后的最大似然估计和变分下界优化机理缺乏系统解释，也难以根据不同数据分布和训练阶段自适应调整频率归纳偏置。因此，如何从理论上揭示 PSD-based 扩散模型的频域优化本质，并进一步提升其频域建模能力，是这篇论文试图解决的核心问题。

针对这一问题，论文从频域角度系统分析了基于 PSD 的扩散生成模型的优化机制，证明了其在空间域和频域中的优化目标本质上是等价的，但在频域展开后会显式表现为对不同频率分量的非均匀加权，从而解释了这类模型 coarse-to-fine 归纳偏置的来源。在这一理论基础上，作者提出了频率归纳偏置自举优化(Frequency Inductive Biases Bootstrapping Optimization, FIBBO)方法，将原本固定的前向退化核参数化为可学习形式，并通过 bootstrapping optimization 交替优化前向退化参数与反向去噪模型，使模型能够根据训练状态进行动态调整不同频率分量的退化—生成轨迹，逐步学习更适合当前数据分布的频域归纳偏置。实验结果表明，FIBBO 在 CIFAR-10、CIFAR-100 和 MNIST 等数据集上均优于标准 Denoising Diffusion 、Heat Dissipation 和 BDM 等基线方法，例如在 CIFAR-10 上可达到 FID 2.83–3.19 的结果，体现出较好的生成质量与训练适应性。该工作的重要价值不仅提出了新方法，更从理论上阐明了 PSD-based 扩散模型为何有效，并将原本固定的频域偏置转化为可学习、可优化的建模机制。

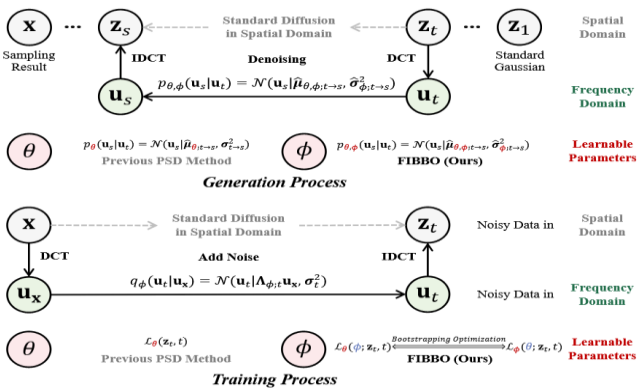


图 1 所提方法模型

好 文 推 荐

清华大学的最新研究成果“FlowTurbo: Accelerating Flow-based Image Generation Models via Multi-stage Refinement”发表于 IEEE TPAMI 2026。

论文：Zhao W, Shi M, Yu X, et al. FlowTurbo: Accelerating Flow-based Image Generation Models via Multi-stage Refinement, IEEE TPAMI, 2026.

近年来，基于 flow matching 的生成模型由于采样轨迹更平滑、生成效率更高，逐渐成为继扩散模型之后重要的生成建模方向。但与扩散模型已经拥有较丰富的快速采样策略与蒸馏研究不同，flow-based 生成模型在实际推理中仍较多依赖于 Euler 和 Heun 等传统数值求解器，导致其速度优势尚未被充分发挥。同时，单步蒸馏类方法虽然能显著加速生成，但往往会破坏原有多步采样结构，从而削弱模型在图像编辑、修复等任务上的灵活性。因此，如何在保留多步采样能力的前提下实现流模型的高效推理，是该领域一个具有代表性的关键问题。

针对上述问题，论文提出了 FlowTurbo 框架，旨在加速 flow-based 生成模型生成过程。其核心思想是利用 flow-based 模型在采样过程中速度场预测更稳定的这一特性，引入一个参数量仅占原始速度预测器约 5% 参数量的轻量级速度细化器，用于估计相邻采样步之间的速度偏移，可在部分采样步骤中用该细化器替代原始模型，从而降低推理成本。在此基础上，作者又设计了伪校正器，通过复用上一步速度预测减少 Heun 采样中的额外模型评估；并进一步提出样本感知编译，突破仅对网络模型编译的传统方式，将模型前向传递、无分类器指导与采样更新合并为独立的样本块并编译为静态计算图，以实现模型加速。针对于流模型在消费级 GPU 上因显存限制导致的推理瓶颈，论文提出了空间多级细化和阶段感知部署的拓展方案，通过低分辨率粗生成与高分辨率细化两阶段流程提升模型的生成效率。该论文提出的 FlowTurbo 及其拓展方案，在提升生成质量的同时显著加速了流基图像生成模型的采样过程，其中分类条件生成加速 53.1%~58.3%，文本到图像生成加速 29.8%~38.5%。该方法不仅实现了 ImageNet 上的实时生成并刷新了 SOTA，还支持在消费级 GPU 上高效部署大参数量的流模型，并保留了多步采样范式对下游视觉生成任务的良好适配性。

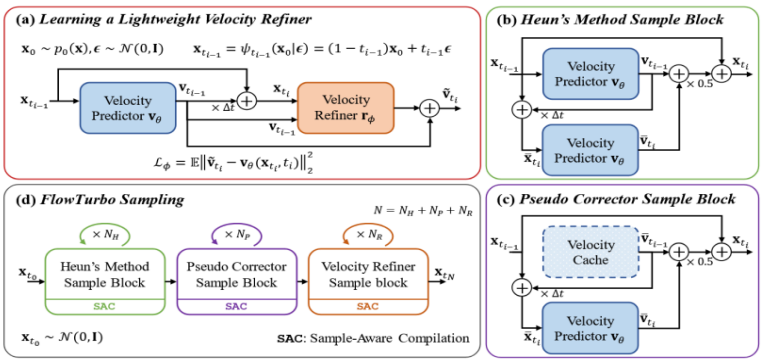


图 1 所提方法模型

责任编委 贾同 李策

好 文 推 荐

伊利诺伊大学厄巴纳-香槟分校、斯坦福大学、亚马逊和哥伦比亚大学团队的最新成果“A Real-to-Sim-to-Real Approach to Robotic Manipulation with VLM-Generated Iterative Keypoint Rewards”发表在 ICRA 2025。

论文：Patel S, Yin X, Huang W, et al. A real-to-sim-to-real approach to robotic manipulation with vlm-generated iterative keypoint rewards[C]. 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025: 8258-8266.

机器人在开放世界环境中的操作任务极具挑战性，因为这需要灵活且自适应的任务目标。这些目标既要能够与人类意图保持一致，又要能够通过迭代反馈不断演化和调整。现有基于视觉语言模型（VLM）的机器人技术普遍存在两方面局限：一方面，缺乏在三维空间中对目标位置进行精确指定的能力；另一方面，难以在任务执行过程中根据环境变化进行动态适应，因而在多步操作和复杂动态场景中的表现仍然受限。

为解决上述问题，文章提出包含四大核心组件的完整框架，如图 1 所示。首先，在给定 RGB-D 观测和自然语言指令的条件下，在场景中采样关键点，并利用 VLM 为多步操作任务自动生成并迭代优化奖励函数（IKER）。该函数通过建模关键点之间的空间关系来定义奖励，同时引入关于目标行为的常识先验，从而实现对 SE(3)空间中的精确控制。进一步地，采用 BundleSDF 方法生成物体的三维网格模型，来在仿真环境中重建真实场景，并利用生成的奖励函数训练强化学习（RL）策略。随后将训练得到的策略直接部署到真实世界机器人中，形成一个真实到仿真，再由仿真到真实的闭环。此外，框架还创新引入难负样本优化逻辑，进一步提升模型鲁棒性。

文章在 XArm7 机械臂平台上开展了大量实验，并使用 4 个固定安装的 RealSense 相机进行环境感知。具体而言，文章在 4 种典型场景中对所提出的方法进行了系统评估，包括鞋子放置、鞋子推动、书本推动以及书本重新定向任务，实验对象涵盖 5 双鞋子、2 个鞋架和 9 本不同类型的书籍。实验结果表明，IKER 框架在单步任务中表现出良好的性能，在真实环境中的鞋子放置任务成功率达到 0.7；在多步任务中，该框架能够通过迭代重规划，成功完成“推鞋盒-放置左鞋-放置右鞋”的连续操作。此外，领域随机化技术的引入使真实环境中推鞋任务的成功率从 0.2 提升至 0.7，充分验证了框架的有效性和鲁棒性。

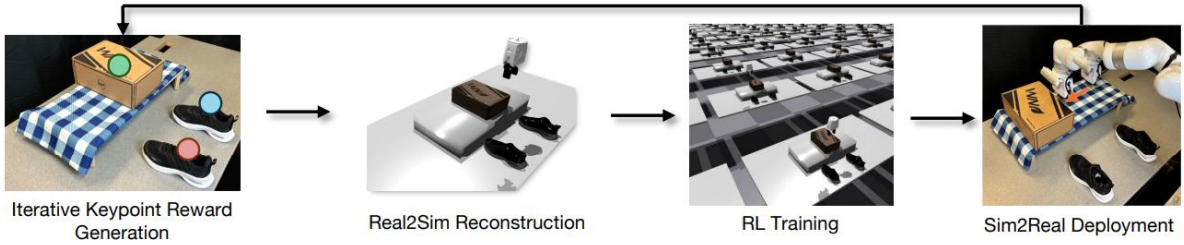


图 1 所提出的网络框架图

责任编委 李策 王田

征文通知

1 会议征文

计算机视觉领域相关国内外会议的征文通知如表 1 所示。同时，可继续关注每个会议举办的 workshop 或 special session。

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 Image and Vision Computing，Ocean Engineering, Pattern Recognition 和 Displays。

3 会议简介

中国模式识别与计算机视觉学术会议 PRCV (Chinese Conference on Pattern Recognition and Computer Vision)，由中国图象图形学学会（CSIG）、中国人工智能学会（CAAI）、中国计算机学会（CCF）和中国自动化学会（CAA）联合主办。旨在汇聚全球学者展示前沿研究成果，推动技术创新与应用发展，已进入 CCF 分区（CCF-C）。会议通常包括主旨报告、专题论坛及产业交流环节，覆盖学术研究、智能驾驶、智能制造等热点方向。

第九届 PRCV 将于 2026 年由哈尔滨工程大学承办。本届会议将秉持团结模式识别与计算机视觉领域科技工作者的宗旨，进一步推动开放合作，广泛吸引学术界和工业界的人才，提升会议的国际化水平，力求打造一个高品质的学术交流平台。大会的举办将为学术界与工业界提供更多产学研合作机会，推动模式识别与计算机视觉领域的协同创新和可持续发展。

责任编辑：刘帅奇

表 1 计算机视觉领域相关国内外会议

会议名称	会议时间	会议地点	截稿日期	会议网站
MM 2026	2026.11.10-14	Rio, Brazil	2026.04.02	https://2026.acmmm.org/
NeurIPS 2026	2026.12.06-12	Sydney, Australia	2026.05.07	https://neurips.cc/Conferences/2026
Siggraph Asia 2026	2026.12.01-04	Kuala Lumpur, Malaysia	2026.05.13	https://asia.siggraph.org/2026/

表 2 计算机视觉领域相关国内外期刊专刊

期刊名称	专刊题目	投稿网址	截稿日期
IMAVIS	Security-AI: Attacks on AI Systems in Computer Vision	https://www.sciencedirect.com/special-issue/321405/security-ai-attacks-on-ai-systems-in-computer-vision	2026.04.15
OCEAN ENG	Multimodal 3D Perception for Underwater Engineering: Acoustics, Optics and Integrated Solutions	https://www.sciencedirect.com/special-issue/325118/multimodal-3d-perception-for-underwater-engineering-acoustics-optics-and-integrated-solutions	2026.04.01
PR	Special Section on Pattern Recognition Letters with the Best Papers from SIBGRAPI 2025	https://www.sciencedirect.com/special-issue/326659/special-section-on-pattern-recognition-letters-with-the-best-papers-from-sibgrapi-2025	2026.04.30
Displays	Vision-Based Intelligent Evaluation Technology	https://www.sciencedirect.com/special-issue/326609/vision-based-intelligent-evaluation-technology	2026.06.30

COMPUTER VISION NEWSLETTER

01 2026

总第 47 期

计算机视觉专委会简报

CCF 计算机视觉专委会