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。

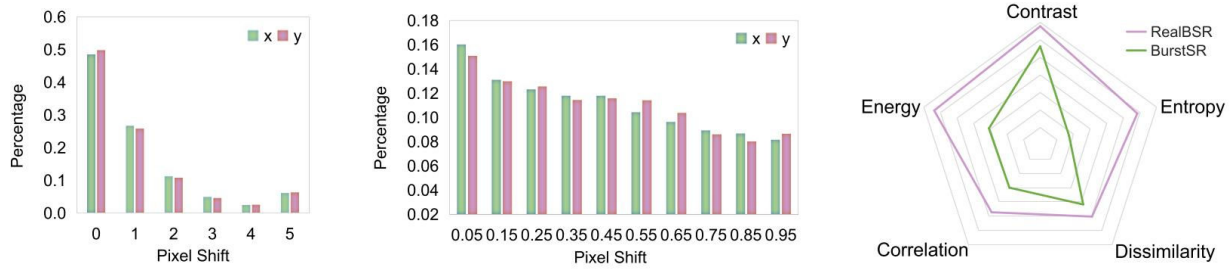
3）高阶特征学习（TNN[9])：针对传统网络难以捕捉高频细节的问题，我们提出了泰勒神经网络(TNN)。基于泰勒级数近似理论，TNN 通过引入泰勒跳跃连接（TSC），在不同网络层级显式地构建和聚合图像的高阶导数信息，从而显著增强了网络对微细纹理恢复能力。

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。

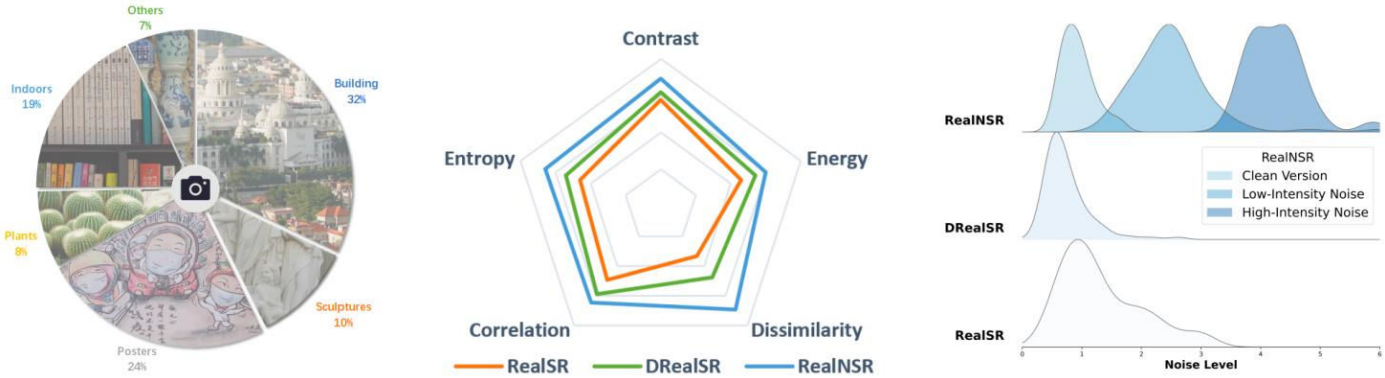
2.2 RealBSR：真实世界多帧超分辨率数据集

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

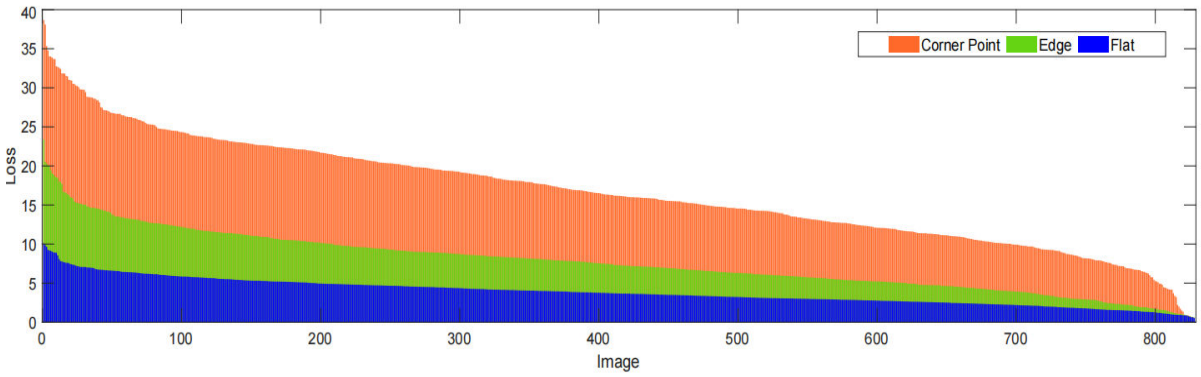


图 4　真实超分辨率中的成分分析。为探究其挑战性，我们分析了 EDSR 在 L1 损失下三类图像成分（平坦区域、边缘和角点）的占比，并通过平均逐像素损失评估各成分对超分辨率重建的影响。结果表明，三类成分的恢复难度存在显著差异：平坦区域与边缘的损失值较低，而角点的损失值明显更高。

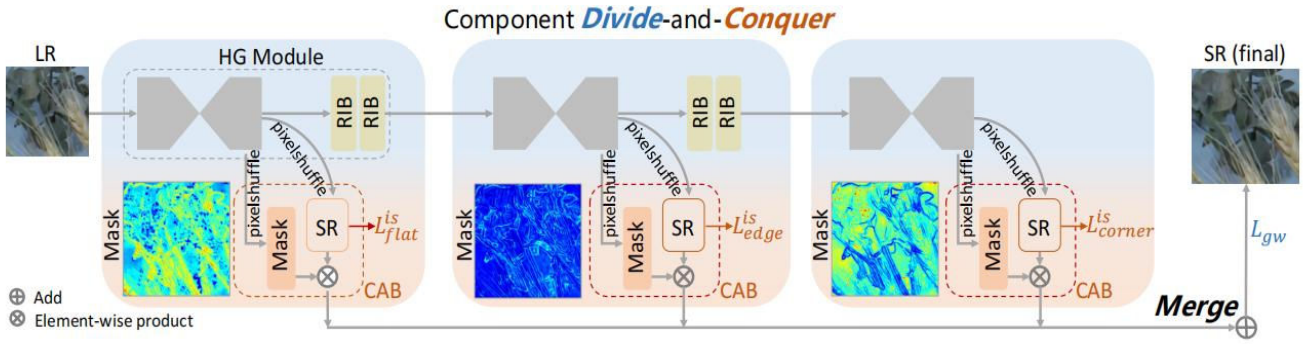


图 5　CDC 框架示意图。堆叠式架构使 CDC 能够在三个成分注意力模块（CABs）中分别建模平坦区域、边缘与角点。每个 CAB 分支独立生成一个注意力掩码和一个中间超分辨率（SR）结果。该掩码用于对中间 SR 进行正则化，使其与其他分支协同工作，并无缝融合各分支信息，最终生成自然、细节丰富的超分辨率图像。

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。

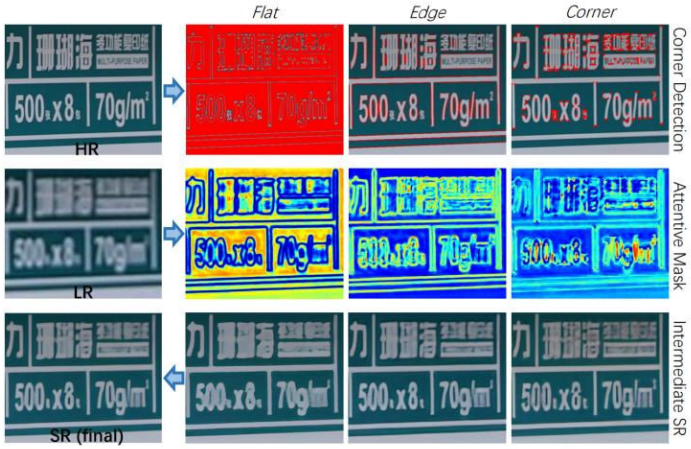


图 6　Harris 角点检测结果、各 CAB 学习到的成分注意力掩码以及对应的中间超分辨率图像。每个 CAB 生成的成分注意力掩码分别与平坦区域、边缘和角点高度相似，表明模型成功解耦并聚焦于三类低层图像成分。

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

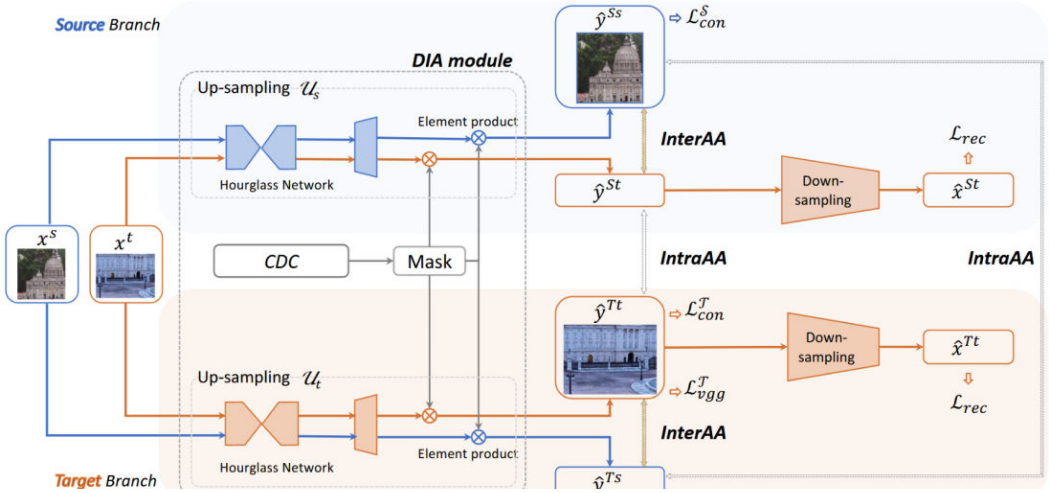


图 7　提出的 DADA 框架，包括一个源分支和一个目标分支，两者都采用 LR-HR-LR 重建结构。在其上采样网络中，一个域不变注意力（DIA）模块通过一个预训练的 CDC 模型提供组件注意力编码。对于 InterAA，来自不同域的两张 LR 图像被输入到上采样网络中，以便在一个分支内进行对抗自适应。对于 IntraAA，一张 LR 图像分别被输入到源分支和目标分支的上采样网络中，进行对抗自适应。在测试阶段，目标分支中经过训练的上采样网路被用于推理。

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

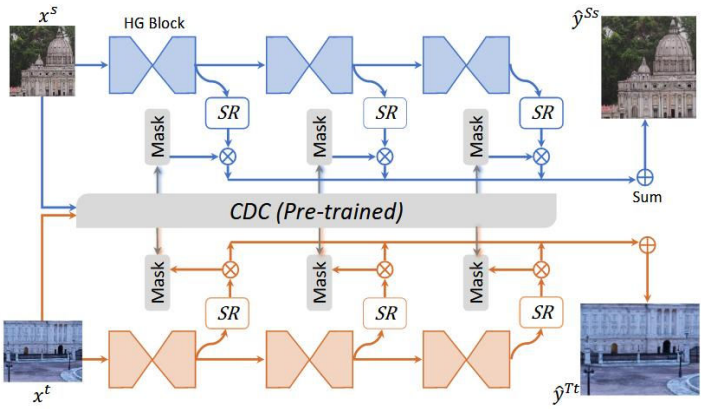


图 8　域不变注意力模块。为了简洁起见，我们通过将两张 LR 图像分别作为源分支和目标分支的输入，来展示其详细的网络结构。

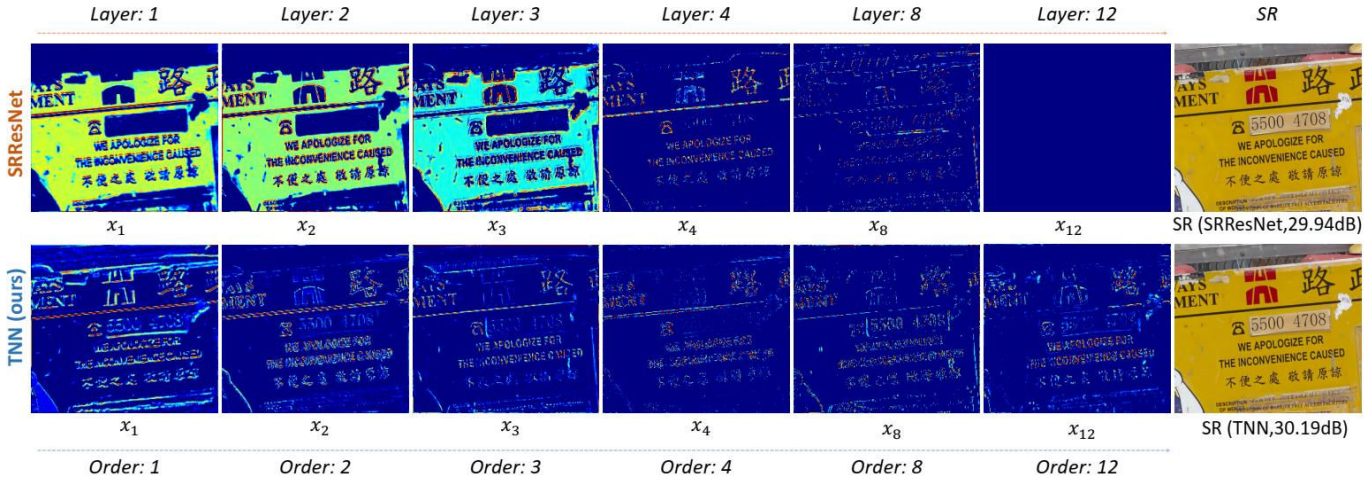


图 9　不同网络层级所学特征图的可视化对比。在各层中，所提出的泰勒神经网络（TNN）生成的泰勒特征图，相较于 SRResNet 能激活更多高响应值的像素点。SRResNet 的特征图易趋于平坦化区域（这类区域比纹理更易重建），尤其随着网络层级加深，其激活的像素数量显著减少。而本方法采用泰勒架构的 TNN 则能够学习到更丰富的图像细节特征。

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。

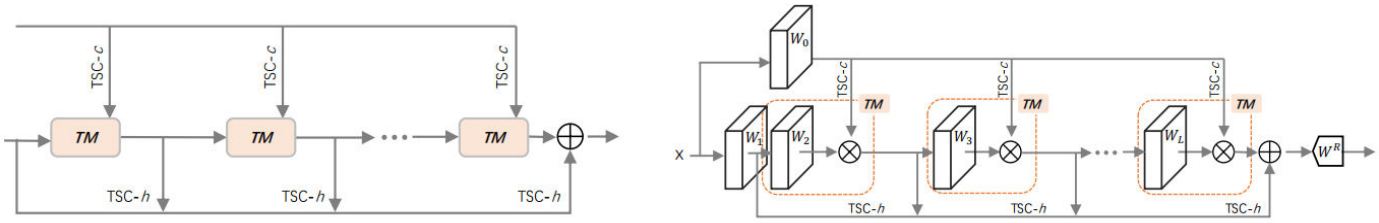
同理，在目标分支中，源域 LR 图像和目标域 LR 图像都被$\mathcal{U}_t$进行上采样，生成两个 SR 结果$\hat{y}^{Ts}$和$\hat{y}^{Tt}$，然后它们由判别器$\mathcal{D}_t^{inter}$进行判别。这里与源分支有两个不同点：为了避免源域对目标分支中的$\mathcal{U}_t$产生过于强大和主导的监督，不对源域 SR 结果$\hat{y}_j^{Ts}$施加任何监督。由于目标分支中不使用源域监督，为了稳定$\mathcal{U}_t$的训练，我们对目标域的 SR 图像施加了监督。由于缺乏目标域的 HR 图像，我们对目标域 SR 结果$\hat{y}_j^{Tt}$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

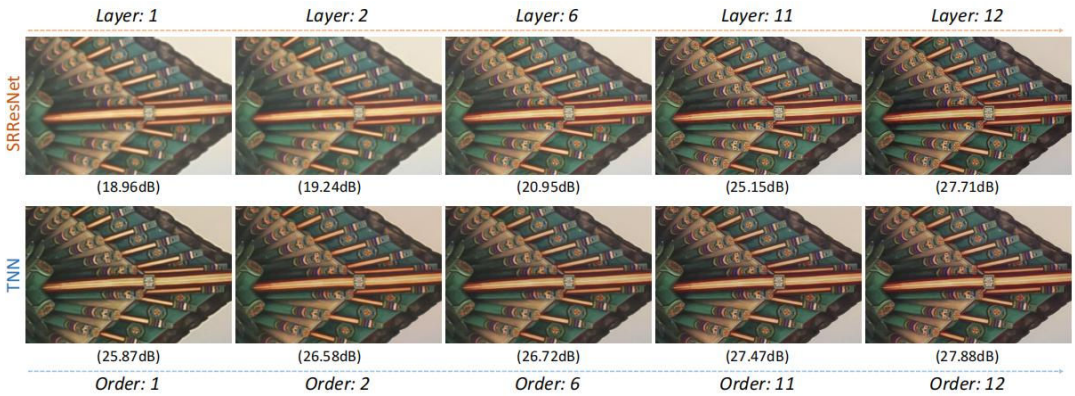


图 11　不同网络层级的超分辨率结果可视化对比

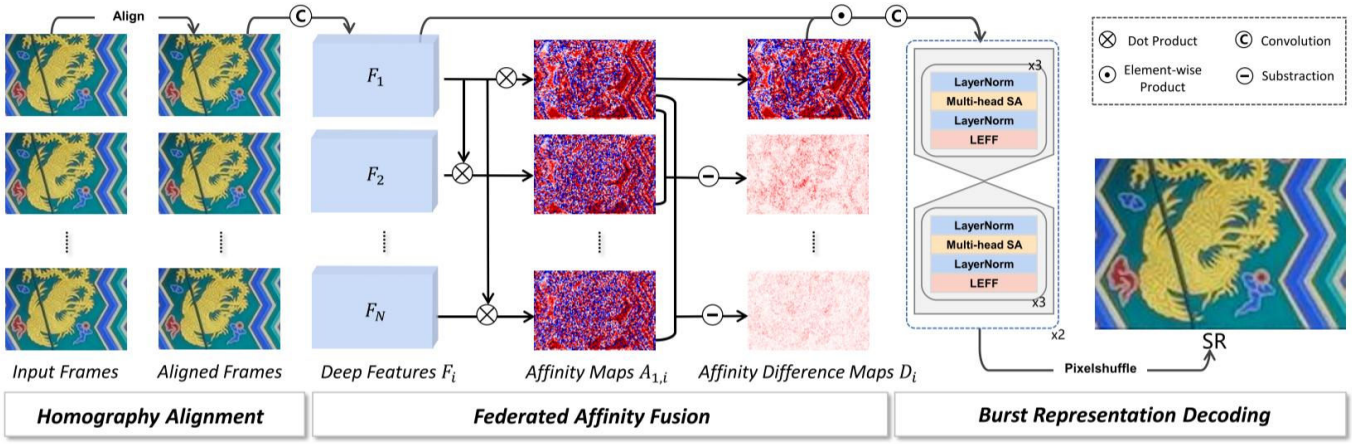


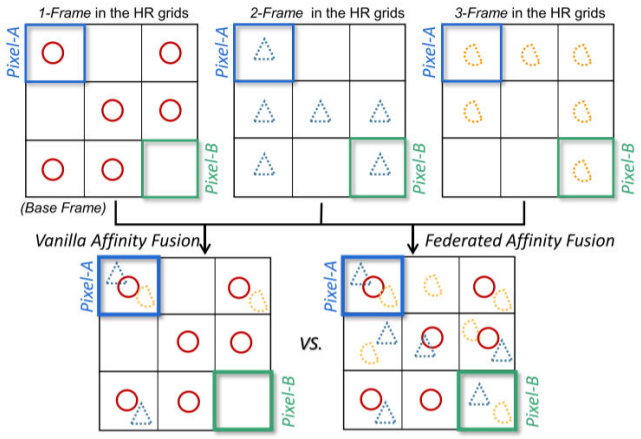
图 12　所提 FBAnet 的工作流程示意图，包含三个主要组件，包括单应性对齐、联邦亲和力融合以及多帧表示解码

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。

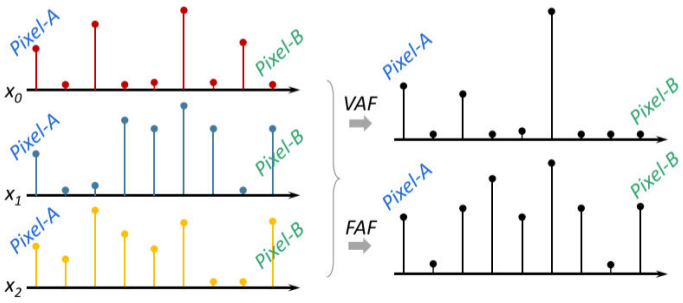
3.4 FBANet：多帧信息融合

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。

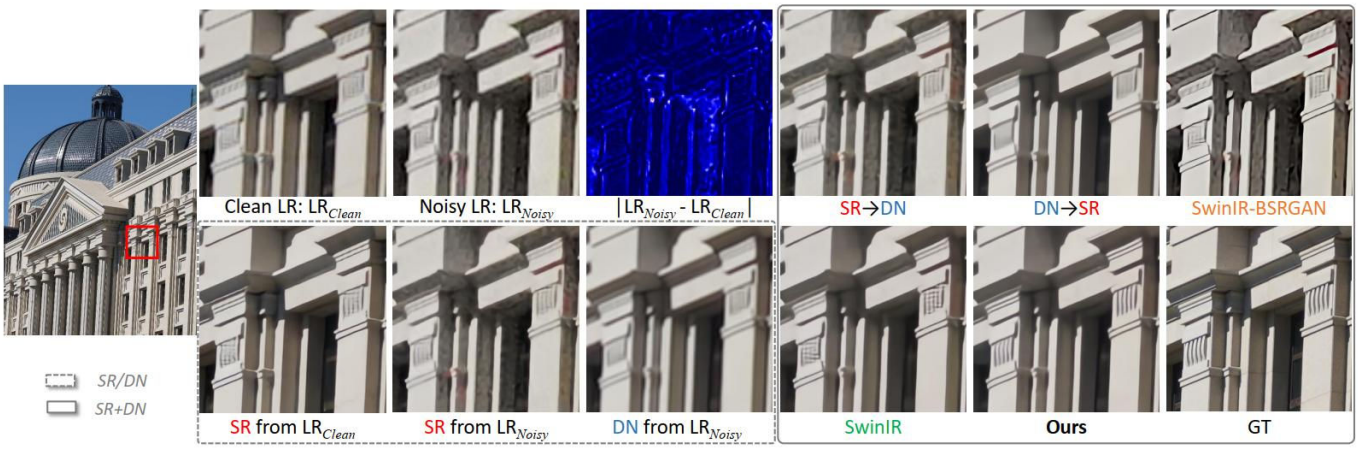


图 14　给定一张真实世界的含噪 LR 图像，我们的 NSSR 与三类解决方案的预测结果对比，包括单任务方法（SR 和 DN）、单任务组合（SR→DN 和 DN→SR）以及联合训练解决方案（SwinIR-BSRGAN、SwinIR 和我们的方法）。注意，所有方法均使用我们的真实世界含噪数据进行训练，SwinIR 和 Restormer 分别作为 SR 和 DN 方法的示例。

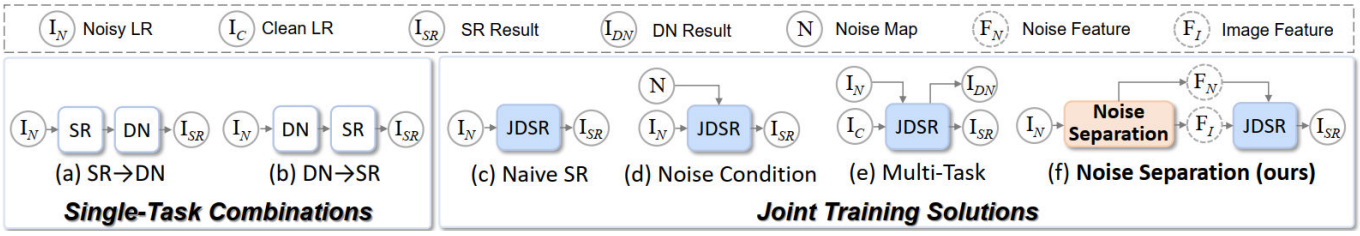


图 15　解决真实世界含噪图像 SR 问题的现有范式：单任务组合（SR→DN 和 DN→SR）、联合训练解决方案（即真实世界盲 SR、朴素 SR、噪声条件和多任务）以及我们的噪声分离（Noise Separation）。我们的噪声分离并非直接处理含噪 LR，而是首先解耦噪声特征和相对干净的图像特征，减轻了真实世界噪声对后续图像细节重建的干扰。

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

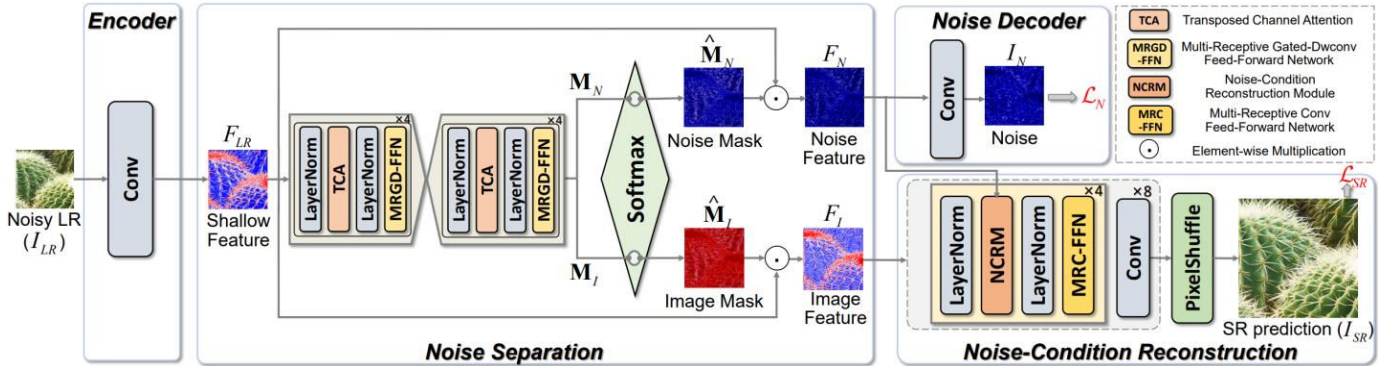


图 16　我们的噪声分离超分辨率（NSSR）网络框架

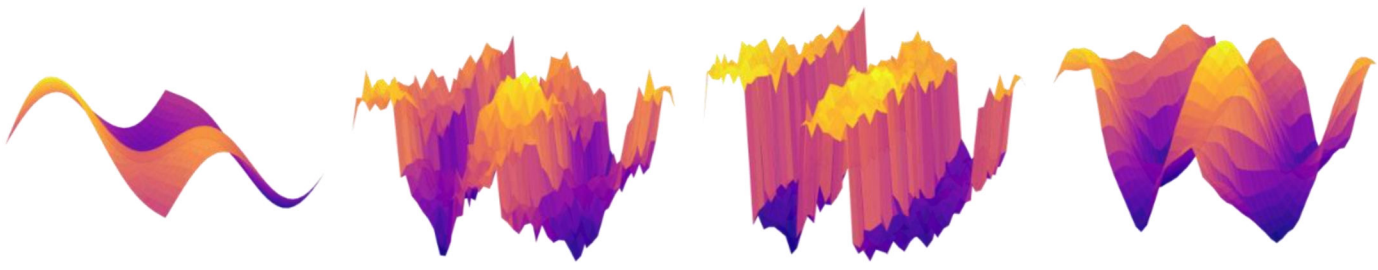
3×3 卷积的编码器（H_E）提取浅层含噪特征F_{LR}。H_E将含噪 LR 从低维空间映射到高维空间 。然后，NS（H_{NS}）从F_{LR}中解耦噪声特征F_N和相对干净的图像特征F_I。F_N和F_I在两个分支中处理：（1）带有卷积层的噪声解码器（D_N）将F_N重新映射为估计的噪声输出I_N，使用参考噪声$\hat{I}_N$进行训练，$\hat{I}_N$通过计算 GT 图像I_{GT}的双三次下采样与含噪LR图像I_{LR}之间的绝对差值计算得出；（2）NCR（H_{NCR}）在噪声特征F_N的条件下恢复相对干净图像特征F_I的高频细节，利用 PixelShuffle 获得 SR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

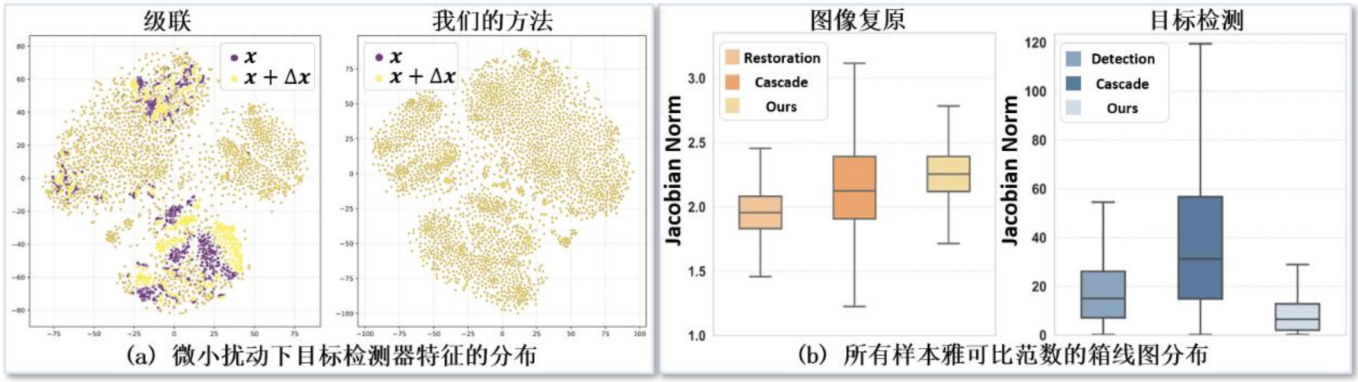


图 18　不同退化变化对特征稳定性和 Lipschitz 连续性的影响

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

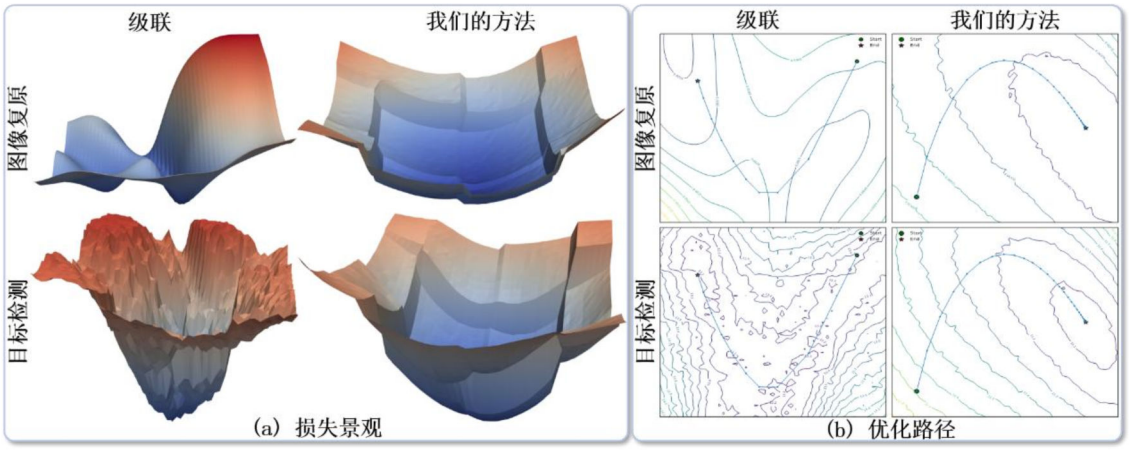


图 19　级联框架和我们的 LROD 框架之间的参数空间平滑度和优化稳定性比较

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$|\nabla_\theta f_\theta(\boldsymbol{x})|$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.

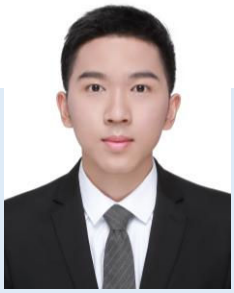
[14] Xiong Y, Li Z, Chen Y, et al. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 5652-5661.

[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

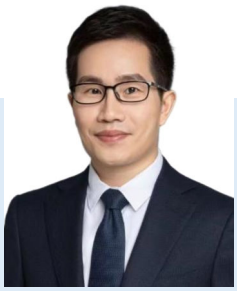
中山大学计算机学院副教授。博士毕业于中国科学院大学。已发表 30 余篇高水平学术论文，两次举办 IEEE Seasonal School，并在领域顶会 ECCV 上举办真实图像超分辨率学术竞赛。获广东省科技进步一等奖（4/13），主持/参与国自然面上、青年、国家重点研发计划项目、国自然联合基金项目。现在主要研究兴趣：底层视觉模型预训练，真实图像超分辨率数据基准构建及方法研究，图像和视频内容生成，多模态内容鉴伪。

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