任务驱动的文档扫描实际超分辨率

Maciej Zyrek[1], Tomasz Tarasiewicz[1], Jakub Sadel[1], Aleksandra Krzywon[2], and Michal Kawulok[1]

[1] Department of Algorithmics and Software
Silesian University of Technology, Gliwice, Poland
michal.kawulok@polsl.pl
[2] Department of Biostatistics and Bioinformatics
Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch
Gliwice, Poland

Abstract. 单图像超分辨率指从单个低分辨率观测中重建高分辨率图像。尽管最近基于深度学习的方法在模拟数据集上展示了显著的成功——其中低分辨率图像是通过降质和下采样高分辨率图像获得的——但在泛化到包括文件扫描在内的真实世界环境时，它们经常失败，因为这些环境受到复杂退化和语义变化的影响。在这项研究中，我们引入了一种任务驱动的多任务学习框架，用于训练特别针对光学字符识别任务优化的超分辨率网络。我们建议结合高层视觉任务衍生的辅助损失函数，包括使用连接主义文本提议网络的文本检测，通过卷积递归神经网络进行文本识别，使用 Key.Net 进行关键点定位，以及色调一致性。为了平衡这些多样化的目标，我们采用了一种动态权重平均机制，根据每个损失项的收敛行为自适应地调整其相对重要性。我们基于已建立的用于单图像超分辨率的 SRResNet 架构验证了我们的方法。对模拟和真实世界扫描文档数据集的实验评估显示，所提方法在保持整体图像保真度的同时，提高了用交并比测量的文本检测。这些发现强调了多目标优化在超分辨率模型中对于缩小模拟训练范围与现实场景中实际部署之间差距的重要性。

1 介绍

扫描文档的有限空间分辨率常常对光学字符识别（OCR）系统构成重大挑战，尤其是在输入图像因传感器限制、压缩伪影、运动模糊或光线条件不佳而导致的噪声和其他失真 [?] 的情况下。为了使现有 OCR 系统能够有效处理低分辨率（LR）扫描件，其质量可以通过超分辨率（SR）技术来提高，这些技术可以来自单张图像或同一场景的多次观测。最先进的单图像 SR（SISR）方法以深度学习为基础。它们包括用于 SR 的第一个卷积神经网络（CNN）（SRCNN） [?]，非常深的 SR 网络（VDSR） [?]，以及在生成对抗网络（GAN）设置下训练的残差 SR 网络 SRResNet [?]。这些技术，以及最近开发的解决方案 [?] 在模拟基准测试中展示了令人印象深刻的性能，其中低分辨率图像是从原始图像获得的，在训练和验证过程中被视为高分辨率（HR）参考。然而，正如 Cai 等人指出的那样 [?]，这些模型在退化既不均匀也不被用于下采样的核良好建模的真实数据方面往往难以推广。这促成了针对真实图像而非模拟图像进行超分辨的真实世界 SR 的发展 [?]。虽然这类技术包括利用真实世界的数据集，但它们很少在特定任务情境中得到验证 [?]。

为了解决这些局限性，我们提出了一种专为文档图像设计的任务驱动的单图像超分辨率（SISR）框架。我们的方法基于 SRResNet 架构，但将训练目标从纯粹的图像保真度——例如，以峰值信噪比（PSNR）表示——转向任务导向的性能。特别是，我们引入了多个从预训练的语义模型中导出的辅助损失。我们在训练框架中整合了四个辅助损失组件：基于连接主义文本提议网络（CTPN）[?] 的文本检测损失，促进文本存在特征的重建；源自预训练卷积递归神经网络（CRNN） [?,?] 的中间激活的损失，促进字符识别；使用 Key.Net 检测器 [?] 的关键点对齐损失，增强输出与真实值之间的结构一致性；以及基于色相饱和度-明度（HSV）色彩空间的色相组件的色彩一致性损失，帮助保持色彩的和谐一致性，而不扭曲色调关系。

这些互补的目标被整合到一个统一的多任务损失函数中。在此类多目标优化中，一个主要的挑战是确定每个任务的适当权重。静态权重可能导致不稳定的收敛或欠拟合。为了克服这个问题，我们采用动态权重平均（DWA） [?]，它在训练过程中根据各自的学习动态自动调整任务权重。这样可以鼓励平衡训练，并确保没有任何一个损失组件主导其他损失组件。

为了评估我们的方法，我们设计了一个综合实验设置，涉及真实和模拟的低分辨率—高分辨率文档对。真实世界扫描是通过受控的多分辨率采集管道获取的，而模拟对是通过标准的下采样生成的。我们展示了我们的任务驱动训练框架提高了使用交并比（IoU）指标衡量的文本检测性能，特别是在真实世界扫描中。重要的是，该模型在不同类型的文档中表现出强大的泛化能力。总体而言，我们的贡献可以总结如下：

1. 我们提出了一个任务驱动的单图像超分辨率（SISR）框架（见图 1），旨在通过文本检测与识别、关键点检测以及色彩一致性来实现文档图像的超分辨率。
2. 我们采用 DWA 来动态平衡损失组件，从而确保任务之间的稳定和均衡收敛。
3. 我们引入了一个包含高精度配准的现实世界数据集，该数据集具有 4 × 的放大倍率，从而允许对与光学字符识别相关任务的超分辨率进行真实评估。
4. 我们对现实世界和模拟数据集上的模型行为进行了详尽的分析，突出了任务感知 SR 的实际益处和风险。

本文的结构如下。在第 2 节中，我们概述了 SR 领域的最新技术，特别关注任务驱动的方法。我们的方法在第 **??** 节中进行了说明，实验验证的结果在第 3 节中报告。第 **??** 节总结了本文的内容。

2 相关工作

在本节中，我们展示了以深度学习为基础的 SR 技术的最新进展（第 **??** 节），我们概述了任务特定的方法，包括超分辨率文本文档（第 2.1 节），以及处理多任务优化的方法（第 **??** 节）。

卷积神经网络（CNNs）的引入革命性地改变了单图像超分辨率重建（SISR）领域。Dong 的开创性 SRCNN 模型通过优化逐像素损失来最大化自然图像上的峰值信噪比（PSNR），随后是非常深层的 VDSR 网络，该网络通过残差学习进一步提高了重建质量。Ledig 引入了 SRResNet，这是由深度残差块构建的生成器，在 Set5 和 Set14 数据集上实现了最先进的得分。SRResNet 还构成了

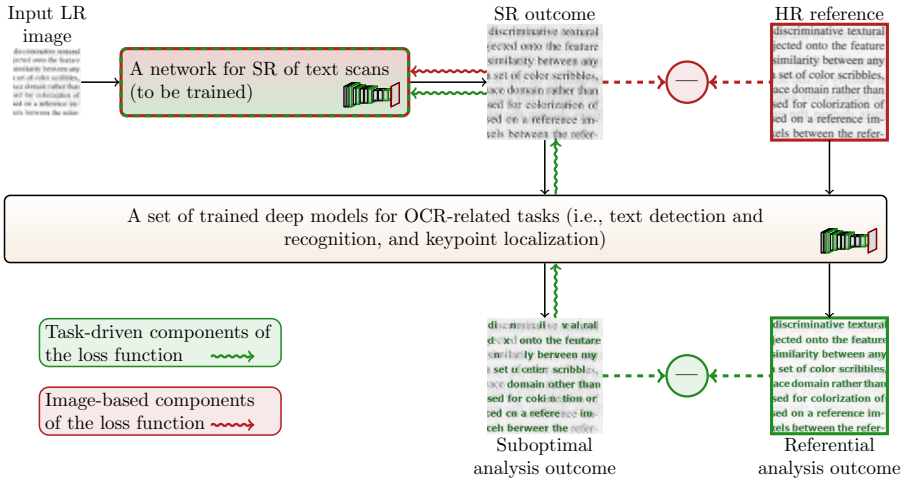

Fig. 1. 所提议的任务驱动的超分辨率网络训练大纲。执行 OCR 相关任务的 CNN 被用于处理高分辨率参考图像和超分辨率图像——它们结果之间的差异建立了损失函数的任务驱动部分，这些部分与常用的图像驱动部分相结合，以指导超分辨率网络的训练。

SRGAN 的骨干，后者通过基于使用 VGG 网络提取的特征的对抗性感知损失扩展了生成器。后来的架构如增强型深度 SR 网络（EDSR）和残差密集网络（RDN）改进了残差和注意力机制，以进一步提升视觉保真度。然而，这些方法主要集中在具有合成下采样的自然图像基准测试上，并且在训练期间不集成更高层次的语义目标。

最近在超分辨率（SR）架构的发展中，人们越来越强调在不影响重建质量的情况下缩小模型规模 [?] 。特别值得注意的是，近期的工作表明，单帧图像超分辨率（SISR）可以从视觉变换器（transformers）中受益 [?] ，后者通过动态调整特征图的大小来降低模型复杂度。此方法仅需 $6.5 \cdot 10^5$ 个参数，据报道其性能优于许多显著更大的最先进模型。快速无参数注意力网络 [?] 引入了一种新颖的无参数注意力机制，在图像质量和推理速度之间取得了平衡，使其适合实时时间应用。Xie 等人 使用核蒸馏来简化模型结构并增强注意力模块 [?] ，在减少计算成本的同时提升了性能。

2.1 特定任务的超分辨率

尽管大多数努力关注于与特定目的无关的增强，但在特定计算机视觉任务背景下验证新兴的超级分辨率技术正受到越来越多的研究关注。这也包括聚焦于文本检测和识别的研究。Wang 使用基于 GAN 的超级分辨率来提升场景文本识别的性能，而 Honda 利用多任务转换器进行场景文本的超级分辨率。这些方法通常报告了识别准确性的提升。在文档领域，ICDAR 稳健阅读挑战推动了对扫描收据和书页的超级分辨率的发展。受这些发现的启发，我们的方法明确地包含了与 OCR 相关的特征损失，以引导超级分辨率朝向文本清晰的重建，同时监控并减轻潜在的幻觉。

曾有过一些尝试，通过使损失函数专注于这些特定任务，以任务导向的方式对网络进行超分辨率（SR）训练，从而相应地引导训练过程。Haris 等人 应用了一个物体检测损失来训练单图像超分辨率（SISR）网络 [?] 。尽管基于任务的训练在峰值信噪比（PSNR）得分上比依赖于 L1 损失的得分更低，但已经证明超分辨率图像对于物体检测而言是更有价值的来源。类似的任务导向损失函数也被定义用于语义图像分割 [?] 和文本识别 [?,?] 。语义分割还被用于学习 [?] 中的 SR 网络。重要的是，所有这些技术仅应用于模拟的低分辨率（LR）图像，并未在真实场景中进行测试。这也涉及到我们最近关于任务导向的 SR [?] 的研究中，我们以任务导向的方式训练了几种 SISR 架构。在这篇报告中，我们将该技术调整适用于真实图像，并结合多种在训练过程中动态平衡的语义损失。

在多任务学习中，平衡异构损失项是至关重要的。固定权重常导致一个目标占据主导地位，从而导致次优收敛。文献中提出了几种动态策略，包括用于平衡跨任务梯度大小的 梯度归一化 、使用任务不确定性作为自适应系数的 不确定性加权 、动态任务优先级和根据最近损失变化率更新每个权重的 DWA。DWA 通过给予损失下降较慢的任务更大关注，促进平衡收敛。

基于在第 2 节中回顾的先前工作，在本节中我们提出了一种方法，旨在增强针对真实世界扫描文档的超分辨率重建（SISR）。尽管许多研究已经在模拟环境中展示了令人期待的性能，但在真实场景中实现稳健且与任务相关的重建仍然是一个重大挑战。受到传统 SISR 技术在保持文本级语义方面存在局限性的启发，我们提出了一种任务驱动的多损失训练框架，并在 SRResNet 架构的基础上进行了验证。我们的方法将多重目标整合到训练流程中，不仅在像素级别上实现重建保真度，还在文档分析至关重要的语义和结构域中实现保真。

本节的其余部分详细介绍了我们提出的解决方案。在第 2.2 节中，我们描述了选择用于研究的 SRResNet 架构以及所采用的具体损失函数。第 2.3 节详述了我们使用 DWA 结合这些目标的策略。

2.2 网络架构和损失函数

为了应对超分辨率处理现实世界扫描文档以用于 OCR 相关任务的挑战，我们提出了一种以 SRResNet 骨干为中心的模块化架构 [?] ，并通过辅助组件增强语义监督。我们根据之前的研究 [?] 选择了该网络，其中我们考虑了几种不同的单图像超分辨率架构——SRResNet 在重建精度和训练时间方面提供了最佳平衡。我们的设计遵循任务驱动的范式，其中重建质量不仅仅通过像素相似度来评估，还通过其保留光学字符识别所需任务相关语义的能力来评估。

我们框架的关键元素是将预训练的 CTPN、CRNN 和 Key.Net 模型集成到 SR 网络训练所利用的损失函数中。重要的是，这些预训练模型的权重在训练过程中保持不变，以保持其最初的训练行为。用于文字检测（CTPN）和识别（CRNN）的模型被集成到一个单一的管道中 [3] ，从而允许以端到端的方式解决 OCR 任务 [?] 。在训练过程中，这些网络充当代理监督者：我们不需要手动标注标签，而是从它们的最深层中提取高级特征激活，并将其与超分辨图像 $\hat{I}$ 和真实的高分辨率图像 I_{HR} 之间的 L1 距离进行比较。通过这种方式，每个冻结网络基于学习的表示隐式生成自己的“标签”，产生一种半自我监督的训练机制，该机制在没有显式标注的情况下强制一致的语义。

[3] CTPN 和 CRNN 模型可在 `https://github.com/courao/ocr.pytorch` 获取

此外，我们引入了一个预训练的 Key.Net 模型（与 OCR 无关）来提取图像关键点，假设文本区域表现出独特的关键点模式。尽管 Key.Net 最初是为通用关键点检测设计的，我们测试了其通过鼓励突出特征的结构对齐来补充 OCR 驱动损失的能力。这些辅助损失函数引导超分辨率模型保留文本和结构线索。

在接下来的小节中，我们概述了核心的 SRResNet 模型，并描述了 CTPN、CRNN 和 Key.Net 模块的架构及其作用，这些模块被集成到训练流程中以产生任务对齐的超级分辨率输出。我们特别关注 SRResNet 作为生成器的主干，因为在我们之前发表的 FedCSIS 论文 [?] 中，该模型展示了良好的结果。其在重建质量、模型复杂性和多损失训练兼容性之间的平衡使其成为进一步任务驱动调整的一个引人入胜的选择。

SRResNet 架构概述 SRResNet 模型最初由 Ledig 等人提出 [?]，作为我们方法的架构骨干。该架构由一个初始特征提取层组成，该层使用宽 9×9 卷积将三通道低分辨率输入图像 $I_{\mathrm{LR}} \in \mathcal{R}^{3\times H\times W}$ 转换为高维表示。随后是一个深度堆叠的 N 残差块，每个残差块通过跳跃连接在保持稳定训练动态的同时优化提取的特征。形式上，如果 F_{i-1} 是第 i 个残差块的输入，则

$$F_i = F_{i-1} + \mathcal{R}(F_{i-1}), \quad i = 1, \ldots, N, \tag{1}$$

，其中 $\mathcal{R}(\cdot)$ 表示带有参数化修正线性单元（PReLU）激活的两层卷积变换。这些残差连接使网络能够专注于学习高分辨率和低分辨率表示之间的高频差异，而不是整个映射，从而提高了收敛速度和最终性能。

在残差块之后，另一个卷积层将学到的特征合并为一个单一的张量，随后通过连续的子像素卷积（像素洗牌）层进行上采样。每个子像素块将空间分辨率提高 2 倍 $\times$（总的缩放因子为 $4\times$），将特征通道信息重新排列到更细粒度的像素网格中。最后，一个重建层（一个宽 9×9 卷积后接 Tanh 激活）将上采样的特征图映射回一个三通道 RGB 图像：

$$\hat{I} = G(I_{\mathrm{LR}}; \theta_G) \in \mathcal{R}^{3\times(4H)\times(4W)}. \tag{2}$$

在我们的工作中，θ_G 表示 SRResNet 的所有可训练参数。因为 SRResNet 在之前的工作中已展示出强大的文本中心 SISR 基线性能 [?]，我们通过在 MS COCO 数据集上使用纯 MSE 损失的预训练权重进行初始化，之后再用我们的多任务目标进行微调。

用于文本区域检测的 CTPN 架构 为了提取与文本内容相关的空间特征，我们引入了 CTPN 网络 [?]，该网络最初是为检测场景图像中的水平文本行而引入的。CTPN 在识别可能包含文本的局部区域方面特别有效，使其成为我们任务驱动的 SR 框架中一个合适的辅助监督模块。

CTPN 模型的架构基于截断的 VGG16 骨干网络，并在 ImageNet 数据集上进行了预训练 [?]。给定一张超分辨率图像 $\hat{I}$ 或高分辨率的参考图像 I_{HR}，卷积编码器输出一个大小为 $(C \times H' \times W')$ 的特征图。后续的轻量级 3×3 卷积层将通道维数减少到 512，从而得到一个张量 $X \in \mathcal{R}^{512\times H'\times W'}$。该张量被重塑为长度为 512 的向量序列，并由双向门控循环单元（Bi-GRU）按行处理，捕获图像宽度上的水平上下文依赖性——这对于精确的文本区域检测是必

不可少的。输出序列由 Bi-GRU 的 256 维特征组成，并被重新塑造成一个空间图 $X_{\text{seq}} \in \mathcal{R}^{256\times H'\times W'}$ 。

这个中间表示然后通过一个 1×1 卷积来恢复到 512 通道的维度。得到的张量由两个并行的 1×1 卷积分支处理。第一个分支是分类头，$\psi_{\text{CTPN-clss}}$ ，用于预测每个锚点（即输入的固定宽度垂直切片）的包含文本与背景的概率。这个分支的训练使用屏蔽交叉熵损失进行，其中中性锚点不参与监督。第二个分支是回归头，$\psi_{\text{CTPN-reg}}$ ，用于估计对应每个锚点的边界框的垂直坐标偏移。这一输出仅对正样本应用平滑 L1 损失进行训练。通过这种设计，CTPN 模型捕捉到文本内容的存在及其精确位置，使我们的 SR 框架能够聚焦于输入图像中具有语义意义的区域。

具体来说，如果 $F_{\text{clss}}(\cdot)$ 和 $F_{\text{reg}}(\cdot)$ 表示分类和回归卷积的输出，那么：

$$\mathcal{L}_{\text{CTPN-clss}} = \text{CrossEntropy}\big(F_{\text{clss}}(\hat{I}),\, F_{\text{clss}}(I_{\text{HR}})\big), \tag{3}$$

$$\mathcal{L}_{\text{CTPN-reg}} = \text{SmoothL1}\big(F_{\text{reg}}(\hat{I}),\, F_{\text{reg}}(I_{\text{HR}})\big). \tag{4}$$

然而，当用作网络监督的代理时，我们并不直接比较最终预测框。相反，我们从每个头部的最后一个预激活层中提取深度特征——记为 $\psi_{\text{CTPN-deep}}(\cdot) \in \mathcal{R}^D$ ——并测量它们在 $\hat{I}$ 和 I_{HR} 之间的 L1 距离：

$$\mathcal{L}_{\text{CTPN-deep}} = \big\|\psi_{\text{CTPN-deep}}(\hat{I}) - \psi_{\text{CTPN-deep}}(I_{\text{HR}})\big\|_1. \tag{5}$$

这些冻结特征损失促使 SR 网络在关键区域保持文本提案的语义。

用于文本识别的 CRNN 架构 为了提供与文本识别对齐的高层语义监督，我们在训练框架中集成了一个冻结的 CRNN 模块 [**?**] ，该模块在大量扫描文档语料库上训练。它在 SR 训练过程中保持冻结，并作为文本识别特征的代理。

CRNN 包含一个卷积编码器，它将 RGB 输入转换为一个紧凑的特征图，该特征图的高度缩减为 1，从而有效地将二维图像沿宽度轴转换为一维特征向量序列。此卷积编码器由多个堆叠的卷积层构成，具有批归一化（BatchNorm）和线性修正单元（ReLU）、池化层，这些层逐步将高度维度减半，直至等于一。此阶段的输出是 $H \in \mathcal{R}^{C\times 1\times W'}$ ，其被重塑为序列 $S \in \mathcal{R}^{W'\times C}$ 。

接下来，S 被送入两个堆叠的双向长短期记忆（Bi-LSTM）层。每个 Bi-LSTM 层在前向和后向两个方向处理序列，捕获邻近字符区域的上下文。最终的循环输出通过线性嵌入投射到字符类加上一个空白符号的分布上。在我们的多损失设置中，我们从最后的 Bi-LSTM 层提取中间循环特征，记为 $\psi_{\text{CRNN}}(\cdot) \in \mathcal{R}^{W'\times d}$ ，并在 $\hat{I}$ 和 I_{HR} 之间进行比较：

$$\mathcal{L}_{\text{CRNN}} = \big\|\psi_{\text{CRNN}}(\hat{I}) - \psi_{\text{CRNN}}(I_{\text{HR}})\big\|_1. \tag{6}$$

。这鼓励 SR 模型产生与 HR 参考相匹配的特征，从而保持字符级的可辨别性。

结构一致性的 Key.Net 虽然 CTPN 和 CRNN 专注于基于 OCR 的监督，我们也引入了 Key.Net [**?**] ，这是一种用于自然图像中通用关键点检测的预训练网络。我们的假设是，文本区域会产生特征性的局部关键点（例如，笔划连接点），因此在 $\hat{I}$ 和 I_{HR} 之间对齐这些关键点可能会进一步增强结构的忠实度。Key.Net 取自其原始实现并保持冻结状态。

给定一幅图像 X ，设 $\boldsymbol{p}_k(X) \in \mathcal{R}^2$ 表示检测到的第 k 个关键点的二维坐标（或热图响应）。然后我们定义关键点对齐损失为：对 HR 图像中检测到的所有关键点进行求和。在实践中，我们从 Key.Net 中抽取固定数量的得分最高的关键点，并确保它们的空间排列在超分辨率后被保留。

多损失方法 在本研究中，我们考虑了八个损失组件来捕捉像素级的保真度、结构一致性和高级语义对齐。这些是：（ i ） 逐像素均方误差 ，（ ii ）一致性损失组件，（ iii ）CTPN 特征空间中的距离（三个组件），（ iv ）CRNN 特征空间中的距离，（ v ）Key.Net 损失，以及 (vi) 色调差异。

逐像素 MSE 损失确保超分辨率输出在 L2 意义上与高分辨率真实值匹配。形式上：

$$\mathcal{L}_{\text{MSE}} \;=\; \big\|\hat{I} - I_{\text{HR}}\big\|_2^2. \tag{7}$$

最小化 $\mathcal{L}_{\text{MSE}}$ 驱使网络减少每个像素的差异，这通常与较高的 PSNR 值相关。一致性损失的目标是确保当超分辨率图像降采样回原始尺寸时，它与输入图像相似。通过对 $\hat{I}$ 进行双三次降采样回到低分辨率域，并与原始 I_{LR} 进行比较，我们施加循环一致性：

$$\mathcal{L}_{\text{cons}} \;=\; \big\|D(\hat{I}) - I_{\text{LR}}\big\|_2^2, \tag{8}$$

其中 $D(\cdot)$ 表示以 $4\times$ 降采样双三次。这一项惩罚无法由原始低分辨率图像合理解释的不真实高频伪影。

为了引导网络保留与文本检测相关的空间特征，我们结合了一组来源于冻结 CTPN 模型的辅助损失函数。具体来说，我们从超分辨率图像 $\hat{I}$ 及其高分辨率对应项 I_{HR} 中提取并比较三种中间表示：（i）深度卷积特征，（ii）分类 logits，以及（iii）回归输出。每个特征空间产生一个独立的 L1 损失，其公式为：

$$\mathcal{L}_{\text{CTPN-deep}} = \big\|\,\psi_{\text{CTPN-deep}}(\hat{I}) - \psi_{\text{CTPN-deep}}(I_{\text{HR}})\big\|_1, \tag{9}$$

$$\mathcal{L}_{\text{CTPN-clss}} = \big\|\,\psi_{\text{CTPN-clss}}(\hat{I}) - \psi_{\text{CTPN-clss}}(I_{\text{HR}})\big\|_1, \tag{10}$$

$$\mathcal{L}_{\text{CTPN-reg}} = \big\|\,\psi_{\text{CTPN-reg}}(\hat{I}) - \psi_{\text{CTPN-reg}}(I_{\text{HR}})\big\|_1. \tag{11}$$

。其中，$\psi_{\text{CTPN-deep}}(\cdot)$ 表示最深的卷积激活图，$\psi_{\text{CTPN-clss}}(\cdot)$ 指的是文本提议置信度的分类输出 logits，而 $\psi_{\text{CTPN-reg}}(\cdot)$ 捕获预测的垂直边界框回归。

通过鼓励从 $\hat{I}$ 和 I_{HR} 提取的这些中间表示之间的对齐，我们激励网络以保留其原始空间结构和显著性的方式重构文本区域。即使没有明确的边界框标签，该公式也为与文本相关的几何结构提供了间接监督。

为了进一步确保超分辨率输出保留字符级的可识别性，我们结合了一个基于预训练 CRNN 模型的特征激活的辅助损失 [**?**] 。设 $\psi_{\text{CRNN}}(\cdot)$ 表示 CRNN 模型的特征提取函数，它包含卷积和递归表示。我们计算从超分辨图像 $\hat{I}$ 及其高分辨率参考 I_{HR} 中提取的特征图之间的 L1 距离：

$$\mathcal{L}_{\text{CRNN}} \;=\; \big\|\,\psi_{\text{CRNN}}(\hat{I}) - \psi_{\text{CRNN}}(I_{\text{HR}})\big\|_1. \tag{12}$$

这种公式鼓励网络重建图像，以不仅保留低级纹理还保留准确文本转录所需的语义结构。通过在输入对之间对齐 CRNN 派生的特征，即使在没有显式文本标签的情况下，模型也被隐含地引导保持光学字符识别（OCR）相关的细节。

为了在超分辨率输出中保持结构模式和空间一致性，我们引入了源自 Key.Net 架构的多尺度索引提案（MSIP）损失。与以前比较中间特征图的损失不同，MSIP 直接作用于从多个尺度的关键点区域提取的局部描述符。给定高分辨率参考图像 I_{HR} 和超分辨率图像 $\hat{I}$ ，Key.Net 识别出可重复的关键点并计算基于局部梯度的描述符。MSIP 损失比较多个尺度下对应关键点周围的响应，在重建的几何结构中强制稳定性和鲁棒性。形式上，此损失定义为：

$$\mathcal{L}_{\text{Key.Net}} = \sum_{s\in\mathcal{S}} \sum_{k} \left\| \phi_{s,k}(\hat{I}) - \phi_{s,k}(I_{\text{HR}}) \right\|_2^2, \tag{13}$$

，其中 $\phi_{s,k}(\cdot)$ 表示关键点 k 和尺度 s 处的局部描述符，而 $\mathcal{S}$ 是使用的尺度集合。此损失惩罚几何结构的差异，确保超分辨率图像在多个分辨率上保持稳定的关键点表示。

通过利用 MSIP 损失，我们不仅促进了纹理或颜色上的对齐，还促进了内在结构线索的对齐，这对于包含精细排版细节或图形注释的文档至关重要。

整体目标函数被定义为前面描述的所有单个损失组件的加权组合。这种聚合允许网络同时优化低级保真度、结构对齐、语义一致性和色彩协调性。为了确保没有单一任务在训练过程中占主导地位，每个权重 λ_i 在优化过程中使用 DWA 策略 [?] 动态调整。具体而言，对那些损失随时间减少较慢的任务给予更高的重视，从而促进异构目标之间的平衡收敛。总损失计算如下：

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \lambda_{\text{MSE}}\,\mathcal{L}_{\text{MSE}} + \lambda_{\text{cons}}\,\mathcal{L}_{\text{cons}} + \lambda_{\text{CTPN-deep}}\,\mathcal{L}_{\text{CTPN-deep}} \\ & + \lambda_{\text{CTPN-clss}}\,\mathcal{L}_{\text{CTPN-clss}} + \lambda_{\text{CTPN-reg}}\,\mathcal{L}_{\text{CTPN-reg}} \\ & + \lambda_{\text{CRNN}}\,\mathcal{L}_{\text{CRNN}} + \lambda_{\text{Key.Net}}\,\mathcal{L}_{\text{Key.Net}} + \lambda_{\text{Hue}}\,\mathcal{L}_{\text{Hue}}.\end{aligned} \tag{14}$$

这种公式确保超分辨率输出不仅最小化像素级重建误差，还保留与下游文档分析任务（如文本检测和识别）相关的结构和语义完整性。DWA 机制在每个训练周期结束时重新计算，基于相对损失下降率，从而使模型能够在整个训练过程中调整其关注点。权重 $\{\lambda_i(t)\}$ 在每个训练周期 t 时更新。设

$$r_i(t) = \frac{\mathcal{L}_i(t-1)}{\mathcal{L}_i(t-2)} \tag{15}$$

表示两个较早周期之间的相对损失改进。那么：

$$\lambda_i(t) = \frac{N \exp\big(r_i(t-1)/T\big)}{\sum_{j=1}^{N} \exp\big(r_j(t-1)/T\big)}, \tag{16}$$

，其中 $N = 8$ 是损失项的数量，$T = 2$ 控制权重的柔和度。这鼓励平衡收敛，防止任何单一任务在训练中占主导地位。

2.3 多任务损失聚合和优化策略

上述损失函数被组合成一个统一的多任务训练目标，其中每个组件根据其动态更新的权重对整体优化作出贡献。这种复合策略使模型在训练过程中能够同时满足多个重建目标——从像素级精确度到高层次语义保留。

所有的损失值被汇总为总损失：

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{N} \lambda_i(t) \cdot \mathcal{L}_i, \tag{17}$$

，并相应地传播梯度以使用 Adam 优化器更新 SRResNet 参数 θ 。

3 实验验证

在本节中，我们报告了为评估所提出的任务驱动 SR 框架而进行的实验。我们描述了数据集的构建和训练设置（第 3.1 节），接着是在多个测试场景中的定量和定性比较（第 3.2 节），以及最后的讨论（第 **??** 节）。

3.1 实验设置

为了评估我们方法的有效性，我们进行了系列对照实验，比较了在各种监督策略下训练的模型。每个模型都在各种测试数据集上进行了评估，其中包括真实和模拟的质量下降，以评估其泛化能力和对实际世界失真的鲁棒性。接下来，我们将描述数据集准备过程、训练设置和研究中使用的评估协议。

数据集 为了评估我们任务驱动的单图像超分辨率（SISR）框架的有效性，我们构建了一个高质量的数据集，该数据集由在不同条件下获得的真实文档扫描组成 [4] 。数据集设计受到了先前研究的指导，并继续沿着我们早期工作的研究方向，重点研究 ×4 的放大因子。这个上采样因子提供了一个实用的平衡：虽然 2× 的放大可能不足以使模型具备重建精细细节的挑战性，但更高的倍率常常会导致不稳定的优化和虚幻的内容。因此，4× 设置为恢复质量提供了一个有意义且真实的基准。该数据集由几个部分组成，即：(i) 大学公告（大学公告的扫描件），(ii) 科学文章（印刷科学文章的扫描件），(iii) 新冠测试传单（一个医学传单的扫描件），以及 (iv) 一个公开可用的古籍数据集 [**?**] 。此外，我们还利用了包含自然图像的 MS COCO 数据集 [**?**] 。

大学公告和科学文章扫描是使用三星 SCX-3400 平板扫描仪获取的——一种广泛可用的消费级设备。每页按照五个分辨率级别依次扫描：75、150、200、300 和 600 点每英寸（DPI）。通过一个自定义脚本自动化扫描，以保持一致的获取条件。在所有 DPI 设置下扫描后，每页被稍微平移一点并重新扫描，每页获得九个空间偏移。这总共产生了 $32 \times 5 \times 9 = 1440$ 张扫描图像。所得的多分辨率、多偏移数据集支持单图像和多帧超分辨率场景。在我们的研究中，我们专注于具有正好 4× 缩放比的分辨率对（75-300 DPI 和 150-600 DPI），确保实际保真度并与我们的实验目标一致。

为了确保低分辨率（LR）和高分辨率（HR）图像对之间的空间对齐，我们应用了基于刚性二维平移的全局配准。首先，每幅 LR 图像通过三次样条插值放大，以匹配其 HR 对应图像的分辨率。为了评估对齐质量，我们计算了绝对差分图，然后在中央裁剪区域内（不包括 32 像素的边缘）计算均方误差（MSE）。然后，执行随机网格搜索以识别最小化 MSE 的整数像素平移向量。为了确保估计

[4] 我们将在论文被接收后发布数据集

变换的可靠性，我们在 20 对随机选择的图像对中验证了所选择的平移。最终确定对 75–300 DPI 对的平移为 $[-5, -1]$ ，对 150–600 DPI 对的平移为 $[-6, -7]$ 。随后使用仿射变换将这些位移应用于 LR 图像。对齐的图像对被裁剪成 LR 为 256×256 像素和 HR 为 1024×1024 像素的图像块对，保持 $4\times$ 的比例。步幅与 LR 图像块大小相匹配，块在图像边界处被夹住以避免越界采样。对于每个页面，零偏移扫描用于所有 DPI 级别的验证示例。其余八个偏移扫描用作训练集，每个分辨率产生的比例为 8:1。

使用配备有接触式图像传感器的消费级佳能 imageFORMULA P-208II 线性扫描仪扫描了 COVID 测试说明。这些扫描由于扫描过程中的传输不稳定导致产生显著的挑战，结果是在几何上产生非线性失真，如局部拉伸和扭曲。因此，这个数据集为在现实和不受控采集条件下评估 SR 模型的空间弹性提供了一个强大的基准。每个扫描被分割成相互重叠的 512×512 像素的图块。使用相位相关进行局部对齐，注册后的图块被重新组装成完整图像。由于失真是非刚性的，所有的注册结果都经过手动验证。这些扫描被纳入测试集。

旧书 数据集 [?] 可公开获取，包含历史印刷文档的高分辨率扫描。它具有多种多样的印刷风格、由于纸张老化导致的物理退化以及不均匀的照明。重要的是，该数据集包括每个扫描的文本转录和二值化的真实版本，这使其特别适合评估超分辨输出上的 OCR 相关性能。

为了检验在超越以文件为中心的图像方面的泛化能力，我们采用了模拟的 MS COCO 数据集子集 [?] 。从这个语料库中，我们提取了包含文本内容的图像区域，并通过双三次下采样生成了对应的低分辨率图像。这种受控设置使我们能够在自然场景输入上对比我们的以文件为导向的超分辨率模型，从而提供关于其跨域适应性的洞察。

对于 COVID 公告和旧书数据集，我们准备了通过物理扫描获得的真实低分辨率输入和通过下采样相应高分辨率图像生成的合成低分辨率-高分辨率对。这种双模式评估有助于在不同退化机制之间进行公平和一致的比较，帮助分离采集噪声与纯分辨率相关损失的影响。相比之下，由于没有真实的低分辨率对应物，MS COCO 子集仅被纳入模拟数据集中。

训练策略 SRResNet 模型使用从 MS COCO 数据集上以像素级 MSE 损失预训练的变体权重进行初始化，遵循了在 [?] 中描述的过程。为了研究退化真实性对 SR 性能的影响，我们采用了一个两阶段微调策略，将真实和模拟数据条件的影响分开。

在第一个阶段，称为真实数据微调阶段，我们使用实际扫描过程获取的图像对来训练模型，具体分辨率比例为 $4\times$ （例如，75–300 DPI 和 150–600 DPI）。这些图像对固有地包含现实中的失真，包括特定扫描仪的模糊、噪声、压缩伪影以及印刷介质的物理缺陷。这个设置使模型能够学习反映现实世界扫描退化情况的重建模式。

在第二阶段，称为合成数据微调，我们复用了相同的 HR 图像，并采用双三次下采样来生成它们的 LR 对应图像。这使得在匹配的语义内容下实现了可控的退化建模，隔离了仅由下采样过程造成的学习动态。这个双步协议旨在解开物理采集伪影和基于插值的退化过程的贡献，从而促进在真实和合成条件下训练的模型之间的公平比较。

为了平衡任务驱动框架的多重目标，我们采用了 DWA 策略进行自适应损失加权。在实践中，我们观察到 DWA 使模型能够根据输入特性动态调整重点：例

如，在处理稀疏或低对比度文本的区域时，CRNN 为基础的监督贡献增加，而在纹理丰富的区域，则更加强调像素级损失和基于 Key.Net 的结构对齐。这种自适应机制促进了稳定优化，并提高了模型在测试数据集上有效泛化的能力。

研究的变体和评价指标 在表格 1 中，我们列出了本研究中调查的变体。正如之前的研究（针对模拟数据集）所展示的那样，CTPN 损失对于优化用于 OCR 相关任务的 SR 网络至关重要，因此我们在此着重研究与 CRNN、Key.Net 和色调特征有关的组件的影响。除此之外，我们还报告了使用双三次插值（Int.）以及从模拟 MS COCO 数据集训练的基线 SRResNet 模型（$\mathcal{M}_{\mathrm{B}}$）获得的结果。

Table 1. 我们实验研究中考虑的训练设置。

Variant name	Training dataset	$\mathcal{L}_{\mathrm{CRNN}}$	$\mathcal{L}_{\mathrm{Key.Net}}$	$\mathcal{L}_{\mathrm{Hue}}$
$\mathcal{M}^{R}_{+++}$	Real-world	✓	✓	✓
$\mathcal{M}^{R}_{++-}$	Real-world	✓	✓	
$\mathcal{M}^{R}_{+-+}$	Real-world	✓		✓
$\mathcal{M}^{R}_{+--}$	Real-world	✓		
$\mathcal{M}^{R}_{-++}$	Real-world		✓	✓
$\mathcal{M}^{R}_{-+-}$	Real-world		✓	
$\mathcal{M}^{R}_{--+}$	Real-world			✓
$\mathcal{M}^{R}_{---}$	Real-world			
$\mathcal{M}^{S}_{+++}$	Simulated	✓	✓	✓
$\mathcal{M}^{S}_{++-}$	Simulated	✓	✓	
$\mathcal{M}^{S}_{+-+}$	Simulated	✓		✓
$\mathcal{M}^{S}_{+--}$	Simulated	✓		
$\mathcal{M}^{S}_{-++}$	Simulated		✓	✓
$\mathcal{M}^{S}_{-+-}$	Simulated		✓	
$\mathcal{M}^{S}_{--+}$	Simulated			✓
$\mathcal{M}^{S}_{---}$	Simulated			

对于每个变体，我们报告了从超分辨率和高分辨率（HR）参考图像中提取的文本定位之间的 PSNR、结构相似性指数 (SSIM)、学习的感知图像块相似性 (LPIPS) 和交并比 (IoU)。为了验证所调查的变体之间的差异是否具有统计显著性，我们使用了统计测试。组间比较通过 Kruskal-Wallis H 检验和事后 Dunn 检验，并采用 Benjamini-Hochberg p 值校正进行。双侧 p 值 <0.05 被认为具有统计显著性。所有计算分析均在用于统计计算的 R 环境中进行（版本 4.4.3）。

3.2 结果

表格 **??** 报告了使用所研究的变体在测试数据集上获得的重建质量得分。可以看出，所有训练的超分辨率（SR）模型在图像保真度得分（PSNR、SSIM 和 LPIPS）上都比双三次插值差，但由于依赖于从真实世界数据集中以任务驱动方式训练的模型，它们的 IoU 得分更好（无论对于模拟还是真实世界的数据集）。