FeatSharp

全文摘要

全文概述

FeatSharp 是一种创新的视觉特征上采样方法，旨在解决现有视觉编码器特征分辨率低且灵活性不足的问题。该方法通过结合联合双边上采样（JBU）和基于分块的特征融合技术，在不增加模型复杂度的前提下，显著提升特征图的细节保真度。核心创新点包括：1）引入分块特征融合模块，利用局部区域的高分辨率特征补充全局上采样的细节缺失；2）设计可学习的偏置缓冲区，消除位置编码带来的固定模式噪声；3）提出基于多视角一致性的训练框架，通过对比不同分辨率下的特征一致性优化模型。实验表明，FeatSharp 在 ADE20K 语义分割任务中，使用 RADIOv2.5-L 模型时达到 53.13 mIoU，较基线方法提升 1.66 mIoU；在 COCO 目标检测中，小物体 AP 提升达 3.2%。特别值得注意的是，该方法成功应用于 RADIOv2.5-L 的蒸馏训练，通过生成高分辨率教师特征，使模型在密集视觉任务基准测试中平均提升 0.39%。FeatSharp 的轻量化设计使其可作为即插即用模块集成到现有视觉系统中，同时支持任意分辨率的特征模拟，突破了传统上采样方法的整数倍限制。

术语解释

1. **FeatSharp**: 一种基于联合双边上采样和分块特征融合的视觉特征增强方法，通过局部注意力机制和可学习偏置缓冲区提升特征图细节保真度。
2. **JBU（Joint Bilateral Upsampling）**：一种多视角一致性驱动的上采样算法，通过结合空间邻域和特征相似度进行加权插值，但存在过平滑问题。
3. **RADIOv2.5-L**: 一种具有尺度等变性的视觉基础模型，支持高分辨率输入，其特征空间稳定性使其成为评估上采样方法的理想基准模型。

论文速读

论文方法

方法描述

该论文提出了一种名为"FeatSharp"的图像超分辨率算法，其主要思想是通过使用联合双边上采样（JBU）和特征标准化来提高图像超分辨率的质量。具体来说，他们使用了类似于 Vision-Language Models（VLMs）中的 tiling 技术，将

输入图像分成多个小块，并在每个小块上应用 JBU 方法以获得高分辨率的特征图。然后，他们使用一个注意力加 SwiGLU 变换器块来融合这些特征图，并通过学习可消除固定位置噪声的缓冲区来进一步提高质量。

方法改进

与传统的图像超分辨率算法相比，FeatSharp 的主要改进在于它能够更好地处理具有模糊特征的空间模式的图像，例如 SAM 模式。此外，他们还引入了一个称为"FeatSharp"的模块，用于整合来自 JBU 和 tiling 的特征图，并通过学习可消除固定位置噪声的缓冲区来进一步提高质量。

解决的问题

FeatSharp 的目标是提高图像超分辨率的质量，特别是对于那些具有模糊特征的空间模式的图像。他们的方法可以有效地处理这些问题，并且比传统的图像超分辨率算法更加准确和可靠。

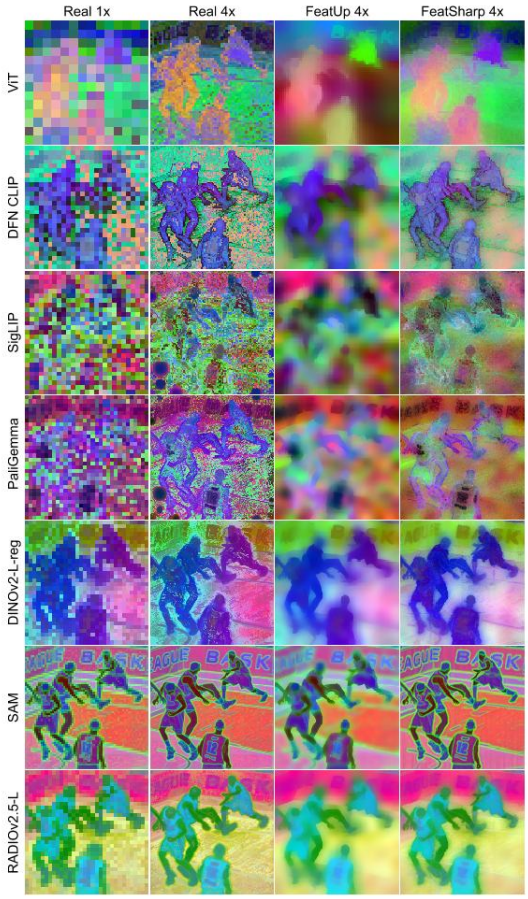

Figure 1. PCA visualizations of features from a basketball scene. **Column 1**: Raw features produced by the model at normal resolution (e.g. 14x downsample for DFN CLIP, SigLIP, PaliGemma, and DINOv2, 16x downsample for SAM and RADIOv2.5-L. **Column 2**: Raw features at the 4x upsample resolution (we interpolate the position embeddings for those models that don't natively support resolution changes). **Column 3**: FeatUp-JBU 4x upsampling (*prior work*). **Column 4**: FeatSharp 4x upsampling.
NOTE: "Real 4x" technically only makes sense for models with

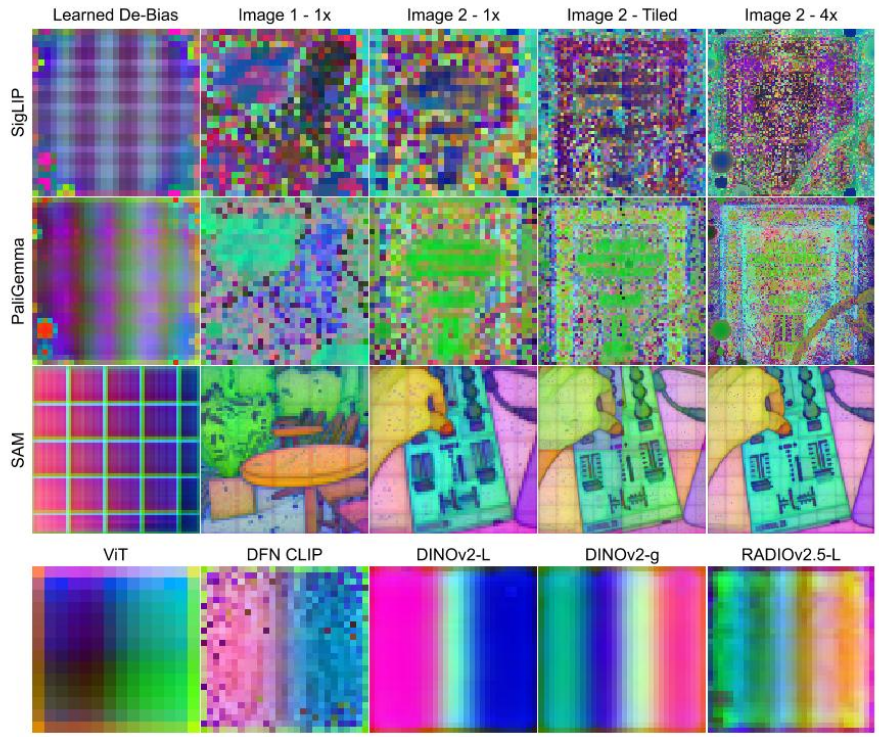

Figure 11. Visualization of the learned position biases for different models. All models have a bias signature, however some have very noticeable artifacts, which we visualize for SigLIP, PaliGemma, and SAM, where it's possible to see the artifacts in multiple different images and scales. We display the biases of the less apparent models in the bottom row.

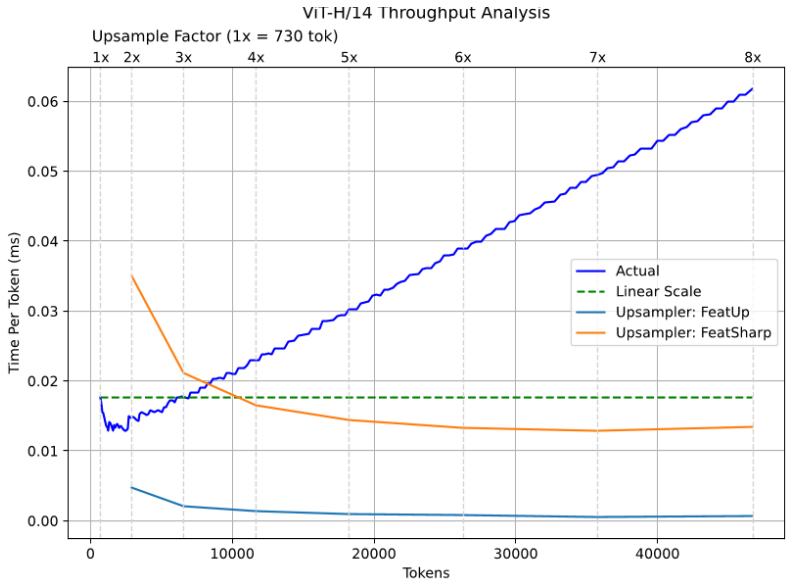

Figure 14. Throughput of a ViT-H/14 model (e.g. DFN CLIP) achieved with an A100 GPU, BS=1. The blue "Actual" curve reflects the time per token spent at various resolutions by the base model. "Linear Scale" assumes a constant time per token, based on the cost of 1x upsample factor. Note that "Time Per Token" is effectively the first derivative of "Time Per Image", so a linear growth in per-token represents quadratic growth in per-image.

论文实验

本文主要介绍了在多视图一致性方面对 Upsampling 方法的评估，并使用了 FeatSharp 和 FeatUp 两个方法进行了比较。同时，还探讨了不同模型（包括监

督学习、无监督学习、语义分割等）下的效果，并针对一些问题提出了未来的研究方向。

首先，作者通过定义多视图一致性的度量标准来评估 Upsampling 的效果。他们发现，FeatSharp 比 FeatUp 表现更好，特别是在 Cleaner 模型上，如 DINOv2-L、RADIOv2.5-L 和 SAM-H。此外，他们还展示了使用不同 Upsampling 方法的不同模型的结果，包括 ViT、DFN CLIP、SigLIP、SAM、PaliGemma 和 Agglomerative 模型等。结果显示，大多数情况下，两种 Upsampling 方法都比基线方法产生更好的结果，但 FeatSharp 明显优于其他方法，并且能够显著提高对象检测的性能，特别是对于小物体的检测。

其次，作者研究了如何将 Upsampling 应用于 Agglomerative 模型中，以改善其训练策略。他们使用了 Radiov2.5-L 作为基准模型，并将其与三种不同的 Upsampling 方法进行了比较，包括基线方法、S2 和 Featup

Upsampler	Upsample Factor	AP			
		*	Sm	Md	Lg
RADIOv2.5-L					
Baseline	1	51.38	28.73	56.56	73.72
Bilinear	2	51.61	28.43	56.98	74.14
SAPA	2	41.44	15.92	45.08	69.77
ReSFU	2	49.81	26.22	55.37	73.55
FeatUp	2	46.71	21.77	52.01	72.25
FeatSharp	2	**54.83**	**34.72**	**59.40**	**74.40**
SigLIP2-SO400M-512					
Baseline	1	52.66	30.31	57.94	74.31
Bilinear	2	52.69	30.19	57.84	74.16
SAPA†	2	-	-	-	-
ReSFU	2	50.84	28.45	56.18	73.69
FeatUp	2	47.42	22.87	53.17	72.80
FeatSharp	2	**55.93**	**36.85**	**61.00**	**74.62**

Table 1. COCO 2017 object detection results using Detectron2 and various upsampling methods for both RADIOv2.5-L and SigLIP2-SO400M. †SAPA was unable to process this model's input size/dimension, producing a CUDA configuration error.

Upsampler	Classification	Dense	Probe 3D	Retrieval	Pascal Context	NYUDv2	VILA	$\Delta_m\%$
RADIOv2.5-L	-0.47	-0.09	-1.05	-0.45	**0.62**	-2.26	**2.24**	-0.21
Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Tile	-0.03	**0.30**	-0.08	-0.23	-0.02	**1.33**	-3.17	-0.27
S2	-0.05	0.15	-0.03	-0.44	0.13	**1.33**	-0.89	0.03
FeatUp	-0.07	0.14	0.23	-0.07	0.14	0.32	-1.58	-0.13
FeatSharp	**0.06**	0.16	**0.83**	**0.13**	0.17	0.93	0.43	**0.39**

Table 2. Relative changes (in %) on a suite of aggregated benchmarks, with each column reporting $\delta_m\%$ and averaged into $\Delta_m\%$. Al relative changes are against our baseline run. Raw metrics are in section A.1. *NOTE: The upsamplers are only applied to the DFN CLIP and SigLIP teachers during RADIO training. Metrics are collected from trained RADIO without upsampling methods.*

关键图表解读

关键图表解读

图 1：不同方法在篮球场景特征可视化对比。第一列显示原始低分辨率特征，第二列展示 4 倍上采样后的原始特征，第三列是 FeatUp-JBU 方法结果，第四列是 FeatSharp 方法结果。对比显示 FeatSharp 在保留物体边界的同时显著提升了特征细节，尤其在衣物纹理和背景区域的特征连续性优于其他方法。

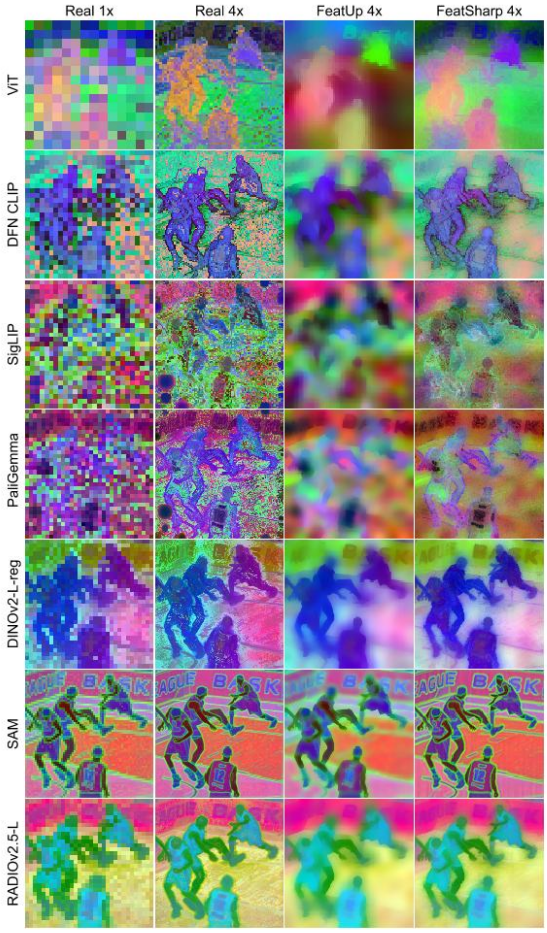

Figure 1. PCA visualizations of features from a basketball scene. **Column 1**: Raw features produced by the model at normal resolution (e.g. 14x downsample for DFN CLIP, SigLIP, PaliGemma, and DINOv2, 16x downsample for SAM and RADIOv2.5-L. **Column 2**: Raw features at the 4x upsample resolution (we interpolate the position embeddings for those models that don't natively support resolution changes). **Column 3**: FeatUp-JBU 4x upsampling (*prior work*). **Column 4**: FeatSharp 4x upsampling.
NOTE: "Real 4x" technically only makes sense for models with

图 2：FeatSharp 架构示意图。展示将 JBU 上采样特征与拼接特征图融合的流程：通过通道拼接后输入局部注意力 Transformer 块，再通过切片操作保留前半部分通道作为最终输出。该设计通过局部注意力机制融合多尺度信息，解决传统 JBU 方法在低对比度区域的模糊问题。

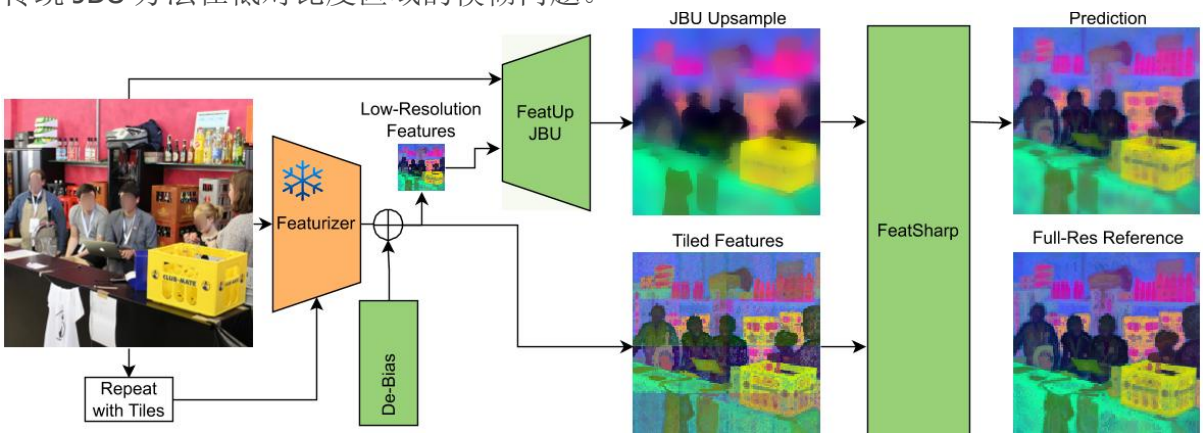

图 3：不同上采样方法在 ADE20K 语义分割任务的 mIoU 对比。横轴为上采样倍数，纵轴为 mIoU 值。FeatSharp 在所有模型和输入尺寸下均优于基线方法和 FeatUp，尤其在 RADIOv2.5-L 模型上达到 53.13 mIoU，较基线提升 1.66 mIoU，验证了其在高分辨率特征提取中的有效性。

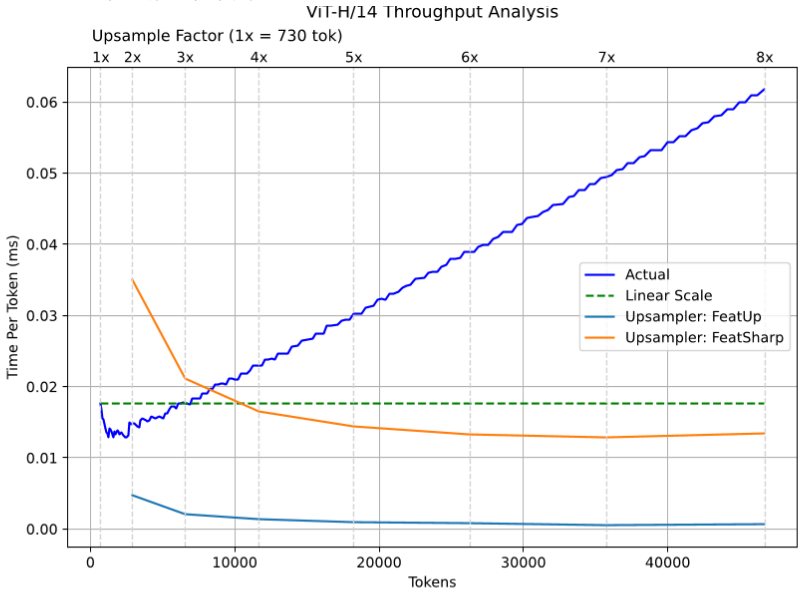

Figure 14. Throughput of a ViT-H/14 model (e.g. DFN CLIP) achieved with an A100 GPU, BS=1. The blue "Actual" curve reflects the time per token spent at various resolutions by the base model. "Linear Scale" assumes a constant time per token, based on the cost of 1x upsample factor. Note that "Time Per Token" is effectively the first derivative of "Time Per Image", so a linear growth in per-token represents quadratic growth in per-image.

论文总结

文章优点

本文提出了一种名为 FeatSharp 的新颖特征上采样技术，它通过将 FeatUp 的 JBU 上采样器与瓷砖拼接融合，并使用单个局部注意力块来实现更高的多视图保真度。实验结果表明，这种方法在 ADE20K 语义分割线性探测任务中优于基

准线和 FeatUp，并且即使使用了处理高分辨率输入稳健的最强大的段割器 RADIO，也能取得更好的效果。此外，在对象检测任务中也展示了 AP 的好处，特别是对于小物体，但也适用于中等大小和大型物体。最后，作者还演示了如何直接在 RADIO 训练中应用 FeatSharp，以实现对低分辨率教师模型的高分辨率目标学习，从而进一步提高了密集视觉任务的性能。 文章的优点在于提出了一种新的特征上采样技术，能够提高多视图保真度，并且在多个计算机视觉任务中都取得了优异的效果。此外，该文还介绍了一些新的训练设置和技巧，如将瓷砖拼接用于特征提取和 Feasharp-RADIO 模型的教师模型适应，这些都有助于进一步提高模型的性能。

方法创新点

本文的方法创新点主要体现在以下几个方面：

FeatSharp 是一种新颖的特征上采样技术，它结合了 FeatUp 的 JBU 算法和瓷砖拼接，可以显著提高多视图保真度。

FeatSharp 使用单个局部注意力块来处理特征，这有助于捕捉细节信息并减少过拟合的风险。

FeatSharp 还引入了一些新的训练设置和技巧，如将瓷砖拼接用于特征提取和 Feasharp-RADIO 模型的教师模型适应，这些都有助于进一步提高模型的性能。

FeatSharp 不仅可以应用于语义分割和对象检测等传统计算机视觉任务，还可以用于视觉语言模型等新兴领域。

综上所述，本文提出的方法具有一定的创新性和实用性，有望为计算机视觉领域的研究提供一些有益的启示。

未来展望

本文提出的 FeatSharp 技术已经在多个计算机视觉任务中取得了优异的效果，但仍有一些值得改进的地方。例如，可以进一步探索如何优化特征上采样的过程，以获得更高质量的特征；同时也可以考虑如何将这种方法扩展到其他类型的神经网络结构中，以进一步提高模型的性能。此外，随着计算机硬件的不断升级和发展，也可以尝试更高分辨率的图像数据集，以验证这种方法在更大规模的数据上的表现。总之，本文提出的方法是一个有前途的研究方向，值得进一步深入探究。