FAKD: FEATURE-AFFINITY BASED KNOWLEDGE DISTILLATION FOR EFFICIENT IMAGE SUPER-RESOLUTION

Zibin He[1], $TaoDai^{1,2}$, $JianLu^{3}$, $YongJiang^{1,2}$, $Shu-TaoXia^{1,2}$

[1]$TsinghuaShenzhenInternationalGraduateSchool, TsinghuaUniversity, Shenzhen, China$

[2]$PCLResearchCenterofNetworksandCommunications, PengChengLaboratory, Shenzhen, China$

[3]$ShenzhenKeyLaboratoryofAdvancedMachineLearningandApplications, ShenzhenUniversity, China$

摘要

卷积神经网络 (Convolutional neural networks, CNNs) 在图像超分辨率 (image super-resolution, SR) 中得到了广泛的应用。大多数现有的基于 cnn 的方法都专注于通过设计更深或更宽的网络来获得更好的性能，但这样却存在大量的计算成本问题，从而阻碍了此类模型在资源有限的移动设备上的部署。为了解决这一问题，我们提出了一种新的高效的 SR 模型，即基于特征亲和力的知识蒸馏 (FAKD)，该模型将教师模型中的结构知识转移到学生模型中。为了有效地传递结构知识，FAKD 的目标是从特征图中提取二阶统计信息，并训练一个计算和内存成本低的轻量级学生网络。实验结果证明了该方法的有效性，并且在定量度量和可视化度量方面都优于其他基于知识蒸馏的方法。

关键词：图像超分辨率，知识蒸馏，模型压缩，卷积神经网络

1 引言

在海量的图像影音设备和互联网的日益普及下，人们在生产生活中对影音图像的使用热情愈加高涨，与此同时也逐渐对图像分辨率和图像处理效率提出了更高的要求。但是受到设备计算资源有限、应用终端降质或高清源文件受损等多种因素的影响，人们所能获取到的图像通常是低质量的。

图像超分辨率是指将低分辨率图像恢复为高分辨率图像的过程。初期鲜为人知的图像超分辨率方法在近几年乘着图像分类等高级计算机视觉任务在深度学习快速发展的东风逐渐走入研究者的视野，使得图像超分辨率技术在生产生活中被一一落地转化为实际应用，例如在医学影像分析中，要求医疗图像具有足够高的解析度以免遗漏微小异物；在图像成像设备中，需要在特定环境下实时将低清图像提高分辨率；在互联网影音应用上，通过在服务器端对影视图像进行压缩，于客户端再对影视图像进行超分辨率复原来降低资源维护和信息传输成本。由此可见，发展图像超分辨率技术具有极大的现实意义。

随着人工智能技术发展进入快车道，基于深度学习的技术在不同领域取得了广泛应用，推进着各行业的发展。但是神经网络依靠其中海量的网络参数共同参与计算，存在着网络层数深、结构复杂、计算复杂度高等缺点，使得现有的深度学习模型在移动端和边缘设备等的部署遇到了巨大挑战。随着深度学习网络结构和优化算法的发展与完善，众多的轻量化方法为本文研究图像超分辨率重构任务提供了极高的研究标准。

图像的超分辨率重构问题极具挑战性，而且其本质上是不适定的，也就是说任意一张低分辨率图像在不同条件环境下复原重建之后，每次生成的高分辨率图像都将大不相同。在过去的数十年中，研

究人员已经提出了各种经典的图像超分辨率方法，如基于预测的方法，基于边缘的方法，统计方法和稀疏表示方法等[1]。

随着近年来深度学习技术的快速发展，基于深度学习的图像超分辨率模型得到了积极探索，比如早期基于卷积神经网络的方法（例如 SRCNN[2-3]）到后来基于生成对抗网络[4]的方法（例 SRGAN[5]）。与以往传统方法相比，这些深度网络模型在图像重建质量和计算速度方面都得到了显著的提高。一般来说，基于深度学习的图像超分辨率模型主要在以下方面有所区别：网络架构[6-8]、损失函数[9-11]、训练策略[8,11-12]等。

目前，卷积神经网络以其强大的特征表示能力在图像超分辨率任务中取得了广泛的应用，并展示了其出色的性能[13]。在 Dong 等人率先提出了一种端到端卷积神经网络（SRCNN）之后，如 EDSR[11]、RDN[14]、RCAN[15]等网络模型都通过将残差块堆叠到数百层来构建非常深的网络，以实现最先进的结果。而后的发展趋势则转为了通过堆叠网络层数来提高测试指标分数，但人类感知到的图像重建质量却未见提升。这也导致了现有大多数基于卷积神经网络的图像超分辨率模型都存在推理速度慢的问题，因为它们包含大量的参数。这使得此类模型在实际应用中受到限制，难以在资源有限的设备（如移动设备）中部署。如果我们还无法尽快制定出足够优异且可靠的模型压缩方案，那么图像超分辨率领域的卷积神经网络方向创新发展将陷入桎梏。因此，设计轻量级图像超分辨率模型至关重要，也是近几年来图像超分辨率领域一众研究者的共同愿景。

为了获得轻量级网络模型，许多压缩网络模型的方法被提出，例如模型修剪[16-18]、轻量级网络设计[19-20]和知识蒸馏方法[21-24]。模型修剪和轻量级网络设计方法需要精心设计并且可能会导致性能下降。如果采用这两种模型压缩方法，那么从某种意义上来说，设计成本高昂的同时又难以确保方法具有普适性和鲁棒性。相比之下，在不改变网络结构的情况下，知识蒸馏方法比其他模型压缩方法更具优势。

传统的知识蒸馏方法首次被提出用于图像识别任务，它遵循师生范式，利用强大教师网络的软标签来监督小型学生网络的训练，从而在缩减模型大小和计算资源的同时，尽量使学生模型保持原教师模型的性能。到目前为止，已经有多种知识蒸馏方法被提出。例如，Romero 等人提出 FitNet[22]来提取隐藏在中间层特征图中的知识；Sergey 等人[25]提出了通过从中间层特征计算注意力图来转移注意力的方法，并鼓励学生网络生成与教师类似的注意力图；Tung 等人[26]提出了保持相似性的知识蒸馏，它使用每个输入小批量中的成对激活相似性来监督学生网络和受训练后的教师网络之间的训练过程。然而，大多数现有的知识蒸馏方法都侧重于高级任务，如图像分类[22-23,25]，而很少关注图像回归任务，如图像超分辨率。当面对图像超分辨率任务时，由于表示空间[27-28]是无限的，因此如何压缩模型仍然是一个开放的问题。

为了提高教师学习效率，我们提出了一种基于特征亲和力的知识精馏 (FAKD) 框架，从教师模型中提取结构化知识。最相关的工作[24]试图从教师模型中传播简单的一阶统计信息 (例如，通道上的平均池化)，而忽略了丰富的高阶统计信息。因此，我们将重点放在从 feature map 中提取二阶信息 (如 intrfeature correlation)，这有助于更精确地重建[29]。具体来说，FAKD 将知识从教师模型的特征关联图转移到轻量级学生模型，从而迫使轻量级学生模型模拟特征关联。实验表明，该框架有效地压缩了基于 cnn 的 SR 模型，同时通过传递强教师模型的结构性知识，提高了学生网络的性能。我们提出了一

个基于特征亲和力的知识蒸馏 (FAKD) 框架，该框架利用特征图中的相关性来监督学生网络的训练。利用空间维度上的亲和力信息改善蒸馏性能。同时实验显示了我们提出的框架在定量和可视化结果方面的优势。

2 相关工作

2.1 基于深度学习的图像超分辨率方法

2.1.1 问题建模与评估指标

图像超分辨率是指将低分辨率图像还原为高分辨率图像的过程，且如今的深度神经网络模型采用的是基于学习的方式，而非先前显式的根据统计假设选择先验的方式。首先一般情况下，低分辨率图像 I_{lr} 都是由高分辨率图像 I_{hr} 降质得到：

$$I_{lr} = \varphi(I_{hr}; \gamma) \tag{1}$$

其中是指降质的映射函数，是指降质过程的一系列参数（例如，放大比例因子或噪声）。而图像超分辨率就是将低分辨率图像 I_{lr} 恢复为真实（Ground Truth,GT）高分辨率图像 I_{hr} 的近似值 $I_{hr}^{'}$:

$$I_{hr}^{'} = F(I_{lr}; \delta) \tag{2}$$

其中 F 是超分辨率模型，δ 表示模型的参数。由于重建的高分辨率图像 $I_{hr}^{'}$ 往往和 GT 图像 I_{hr} 会有一些差别，因此在深度学习上，通常用损失函数来衡量二者之间的差异，通过最小化期望风险（empirical risk）以达到更好的重建效果：

$$\delta' = \text{argmin}_{\delta} \mathcal{L}\left(I'_{hr}, I_{hr}\right) + \lambda\Phi(\delta) \tag{3}$$

其中 $\mathcal{L}\left(I'_{hr}, I_{hr}\right)$ 代表的是重建的高分辨率图像 I'_{hr} 和 GT 图像 I_{hr} 的损失函数，$\Phi(\delta)$ 是正则化项，λ 是权衡参数。尽管图像超分辨率任务中像素间均方误差是常用的损失函数，但强大的模型更倾向于使用多个损失函数的组合。

为了量化评估模型的性能，研究者需使用恰当的方法对生成的图像的质量进行度量。图像评估指标包括主观方法和客观方法，但这两种方法的侧重点不一致，并不能很好的统一说明图像质量优劣。在主观方法中，会由人类作为测试人员为重建图像打分，通常分数从好到差会用五分至一分进行评判，而评判标准一般是视觉感知上是否像真实自然图像或者清晰度等等。然而这种方法带有很强的主观性，即不同测试者可能会给出截然不同的评估结果，并且该方法在测试过程中会消耗大量的人力成本。

而为了更好地评估图像感知质量，同时减少人工干预，研究者试图通过在大型数据集上学习如何评估图像感知质量，提出了基于学习的感知质量方法。尽管该方法可以通过学习模拟人类视觉感知能力以表现出更好的性能，但感知质量的标准仍然有待商榷，因此具有量化指标的客观方法（如峰值信噪比、结构相似度等）仍然是当前的主流。峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）是最常被用来度量重建图像质量的评估指标之一，它所计算的是 GT 图像和重建图像之间像素级的均方误差（Mean Square Error, MSE）。假设 GT 图像为 x_i，重建图像为 y_i，则二者的均方误差为：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 \tag{4}$$

那么 PSNR 可根据图像的最大像素值 L 和均方误差来计算，如下式表示：

$$PSNR = 10\log_{10}\left(\frac{L^2}{MSE}\right) \tag{5}$$

通常情况下，我们使用的均为八位数字图像，则 L 取值为 255。根据不同的任务难度和模型性能，PSNR 的计算结果从差到优会对应于 20 至 40。

我们将采用 PSNR 作为各实验中用于评估算法性能的主要评价标准，并展示部分图像进行对比，可视化结果对算法性能评估具有更加直观的效果。

2.1.2 超分辨率的上采样框架

在图像超分辨率任务中，根据上采样操作及上采样操作在模型中的堆叠位置，当下成熟的图像超分辨率网络模型体系结构大致可归结为以下预上采样、后上采样和渐进式上采样三种类型。

(1) 预上采样

由于从低维空间到高维空间的映射难以直接学习，所以最初研究者的解决方案通常是先用传统的上采样算法将低分辨率图像上采样至目标图像尺寸，然后用深度神经网络进行非线性拟合来得到高分辨率图像。香港中文大学的 Chao Dong 等人[1-2]便是首先采用了预上采样的超分辨率框架并提出了 SRCNN，作为深度学习被引入到图像超分辨率方法中的开山之作，预上采样框架也曾被多次用于模型设计。

图 1: 预上采样框架

在预上采样框架中，由于传统的插值算法已经完成最困难的上采样任务，卷积神经网络只需要学习放大后的粗糙图像与真实高分辨率图像之间的映射，学习难度大大降低。然而，预上采样常常兼具一些诸如图像模糊和放大噪声的负面影响，而且由于大多数对特征图的操作都是在高维空间中执行的，因此会耗费大量的时间和空间成本。

(2) 后上采样

为了提高计算效率以及令上采样可学习化，研究者提出了后上采样的相关模型，该框架在低维空间中进行非线性的特征提取，将上采样层放置在一系列卷积操作之后，这大大降低了总体的计算量，取得先进性能的同时也带来了更快的训练速度和推理速度。因此该框架成为了主流的图像超分辨率框架之一[4,11,30-31]。

虽然后上采样大大降低了计算成本，但是仍然存在一些缺点。一方面，仅使用一个上采样层会使得当放大比例因子较大时难以得到优秀的学习效果；另一方面，该方法导致了每种放大比例因子都需要单独训练一个深度网络模型，无法满足多尺度超分辨率的需要。

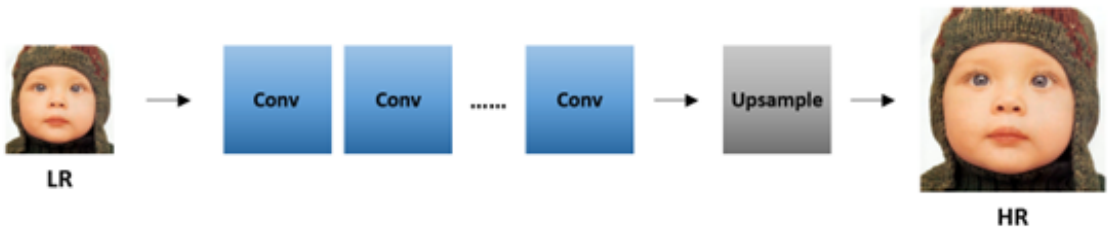

图 2: 后上采样框架

(3) 渐进式上采样

为了解决这些问题，基于渐进式上采样框架的模型被提出了，如 LapSRN[6]。该框架下的模型基于级联的卷积神经网络，并逐步重建更高分辨率的图像。在每个阶段的卷积上采样模块中，网络通过卷积操作不断学习对高分辨率需要填充的特征信息，而后再提高图像的分辨率，并不断循环调整相关权重。通过将复杂任务解耦为简单任务，该框架下的模型可以克服放大比例因子较大以及不引入过多空间和时间成本的重重困难来处理多尺度超分辨率任务。然而，该类型的模型也遇到了一些问题，由于其多阶段模型设计复杂，训练稳定性差，时至今日仍需要更多的建模规范和更先进的训练策略。

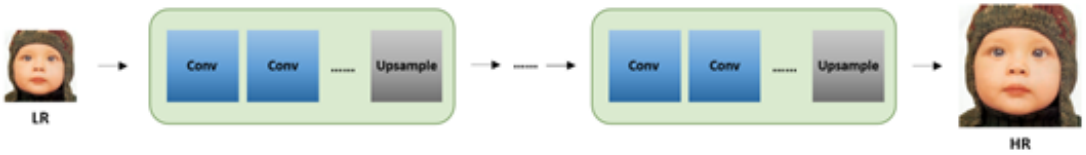

图 3: 渐进式上采样框架

2.1.3 上采样方法

除了模型中的上采样位置外，如何执行上采样也尤为重要。根据可学习权重参数的存在与否，本段落将分别介绍传统的基于图像插值的上采样方法[32-33]和基于学习的上采样方法。

(1) 基于插值的上采样方法

图像插值，又称为图像缩放，指的是调整数字图像的尺寸大小。较为常用的传统插值算法包括最近邻插值、双线性插值和双三次插值。这三种算法对于图像的重建质量随着计算复杂度的增加而增加，且由于这些算法的可解释性和易于实现，至今仍广泛应用于基于深度学习的超分辨率模型中。

最近邻插值是最经典且简单的方法。它会把待插值像素相邻区域中最近的像素值作为插值结果。该算法的计算复杂度低，但性能表现并不理想，重建的图像易出现块状效果。

双线性插值是将图像分为横轴和纵轴分别进行操作的。它先在图像的一个轴进行线性插值，再重复操作在另一个轴上。尽管每一步都是线性的采样，但是结果是对插值所在位置的 2×2 邻域的二次插值。该算法的上采样质量显然比最近邻插值算法更有竞争力，并且计算速度并未逊色几分。

双三次插值，与双线性插值法类似。它是执行三次二维插值，是目前传统方法中主流的上采样算法，被广泛应用在预采样中。该算法依然是在横轴和纵轴上分别对图像进行操作，只是由线性插值更换为三次的多项式插值，对插值所在位置的 4×4 邻域被纳入该位置像素值的计算中。该算法的上采样质量有所提升，结果更平滑且伪影更少，但计算速度相对较慢。

基于插值的上采样方法仅根据图像本身的像素信息来放大分辨率，像素间的特征及相关性没有被挖掘出来，这使其本质上无法生成像素间额外的细节信息，因此该类方法具备较高计算复杂度这一沉

重负担的同时，也往往会放大原有噪声，反而让采样结果变得模糊。

(2) 基于学习的上采样方法

为了解决上述传统方法引发的问题，研究者引入了转置卷积层和亚像素层两种方式来让模型学习如何上采样。转置卷积层，又称为反卷积层。它进行了一个与一般卷积运算相反的变换，即根据卷积输出大小的特征映射来预测可能的输入。为了实现这一过程，它通过对待插值像素位置插入零值并执行指定尺寸大小的卷积操作以扩展图像，从而达到提高图像分辨率的目的。转置卷积被广泛应用于多种图像超分辨率网络模型中，但是由于该方法在进行卷积运算的时候，卷积掩模对输入图像或特征的每个像素的计算是非均匀且重叠的，这导致重建的图像极可能会伴随有“棋盘状”的伪影。

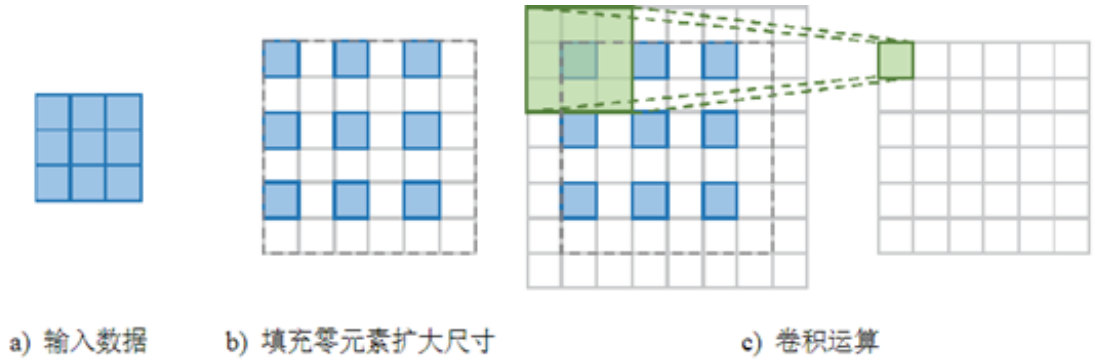

图 4: 转置卷积层示意图

亚像素变换层先通过卷积运算增大输入图像特征的通道数，然后对通道上的元素进行重新排列。与反卷积层相比较，亚像素层具有更大的感受野，能提供更丰富的上下文信息帮助生成更精确的细节特征，但是亚像素层感受野的分布也是非均匀的，块状区域之间几近相同的感受野会导致不同的块边界出现伪影。

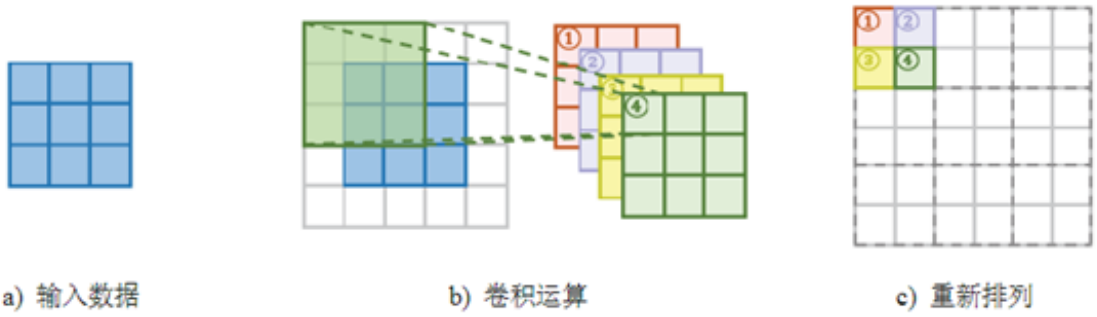

图 5: 亚像素变换层示意图

基于学习的上采样方法相比于传统基于插值的上采样方法计算效率更高，性能表现也往往较好。通常可模块化地添加在图像超分辨率模型的尾部以实现相应深度网络模型的训练。

2.2 知识蒸馏

在模型压缩和知识迁移方面，知识蒸馏是一项成功应用于人工智能诸多场景的可靠技术，它使用强大且带有大量参数的教师模型来指导轻量化的学生模型训练，逐渐改善学生模型的精度以无限逼近教师模型的性能。目前，根据知识信息蒸馏关系可以将知识信息大体分为响应的、特征的和关系的知识三类，而知识蒸馏技术按照蒸馏策略主要可以分为多教师蒸馏、多学生蒸馏和自蒸馏三类。

2.2.1 知识类型

基于响应的知识通常是指教师模型最后输出层的神经响应。主要思想是直接模仿教师模型的最终预测。基于响应的知识蒸馏简单但有效地进行了模型压缩，已被广泛用于不同的任务和应用中。最流行的基于响应的图像分类知识被称为软目标。基于响应的知识可用于不同类型的模型预测。例如，对象检测任务中的响应可能包含标签以及边界框的偏移量。在语义地标定位任务中，例如人体姿态估计，教师模型的响应可能包括每个地标的热图。最近，基于响应的知识得到了进一步的探索，以解决将真实标签信息作为条件目标的问题。从另一个角度看，软目标的有效性类似于标签平滑或正则化器。但是，基于响应的知识通常依赖于最后一层的输出（例如，软目标），因此无法解决教师模型在监督，这对于使用非常深层神经网络的表示学习非常重要。由于软标签实际上是类概率分布，因此基于响应的知识蒸馏也仅限于监督学习。

基于特征的知识通常是指教师模型与学生模型各中间层的特征信息。深度神经网络擅长通过增加抽象来学习多个级别的特征表示。因此，最后一层的输出和中间层的输出，即特征图，都可以用作监督学生模型训练的知识。具体来说，来自中间层的基于特征的知识是基于响应的知识的良好扩展，尤其是对于更薄和更深的网络的训练而言。中间表示法首先在 Fitnet 中引入，通过提供 hints，以改善学生模型的训练。主要思想是直接匹配老师和学生的特征激活。有趣的是，教师模型中间层的参数共享以及基于响应的知识也可以被用作教师知识。尽管基于特征的知识转移为学生模型的学习提供了有利的信息，但是如何有效地从教师模型中选择提示层和从学生模型中选择引导层仍然有待进一步研究。由于提示（hint）层和被指导（guided）层的大小之间存在显着差异，因此还需要探索如何正确匹配教师和学生的特征表示。

基于响应的知识和基于特征的知识都使用教师模型中特定层的输出。基于关系的知识进一步探索了不同层或数据样本之间的关系。为了探索不同特征图之间的关系，Yim 等人[23]提出了一种解决方案流程（FSP），该流程由两层之间的 Gram 矩阵定义。FSP 矩阵总结了特征图对之间的关系。它是使用两层要素之间的内积来计算的。Tung 与 Mori[26]提出了一种保留相似性的知识提炼方法，即将教师网络中输入对的相似激活所产生的保持相似性的知识转移到学生网络中，并保持成对相似性。Peng 等人[34]提出了一种基于相关同余的知识蒸馏框架，提取的知识中既包含实例级信息，又包含实例之间的相关性。使用关联一致性进行蒸馏，学生网络可以了解实例之间的关联。尽管基于关系的知识往往有更好的蒸馏表现，但是如何根据特征图或数据样本对关系信息进行建模仍然值得进一步研究。

2.2.2 蒸馏策略

现有的大多数知识蒸馏方法是建立在单教师网络和单学生网络的基础上，也有很多工作研究多个教师网络、多个学生网络和单个学生网络的蒸馏策略。

首先是基于多级教师网络的模型蒸馏方法，Mirzadeh 等人[35]发现当教师网络和学生网络的容量差距过大时，不利于学生网络的学习，提出了一种基于助教网络的知识蒸馏方法来层级指导学生网络的学习，让学生学习一个知识容量差距较小的助教，而不是直接学习原始的教师网络。实验证明，多级教师蒸馏在模型层面软化了输出分布，提高了输出信息熵，从而为学生网络的学习提供了更多的知识。

其次是基于深度互学习的模型蒸馏方法，Ying Zhang 等人[36]从蒸馏的学习策略入手，提出了无教师网络监督多个学生网络相互学习的学习策略，让学生网络学习多模态的局部最优解，提高所有学生网络的鲁棒性，其性能逼近甚至高于 Hinton 所提的知识蒸馏方法。

基于自蒸馏的模型蒸馏方法是一种提升模型泛化能力的方法，使用与学生网络完全相同的网络作为教师来自我指导。Furlanello 等人和 Bagherinezhad 等人[37]将教师网络的 softmax 输出迁移到学生网络，并进行传代训练，得到泛化能力更强的学生网络。

3 本文方法

3.1 本文方法概述

基于特征亲和力的知识蒸馏 (FAKD) 框架如图 7所示。退化图像 LR 通过老师 T 和学生 S 网络传播。教师模型是一个功能强大的笨重网络，而学生模型是一个轻量级网络。在 FAKD 框架中，它们共享相同的架构，但有不同的超参数 (例如网络深度)。如图 6所示，它们分别由 m 和 n 个残差块 (m > n) 组成。为了有效地将知识从教师模型转移到学生模型，强制学生网络的中间特征图模仿教师模型的特征亲和矩阵。此外，教师输出图像和真实图像还分别通过教师监督 (TS) 和数据监督 (DS) 对学生网络进行监督。

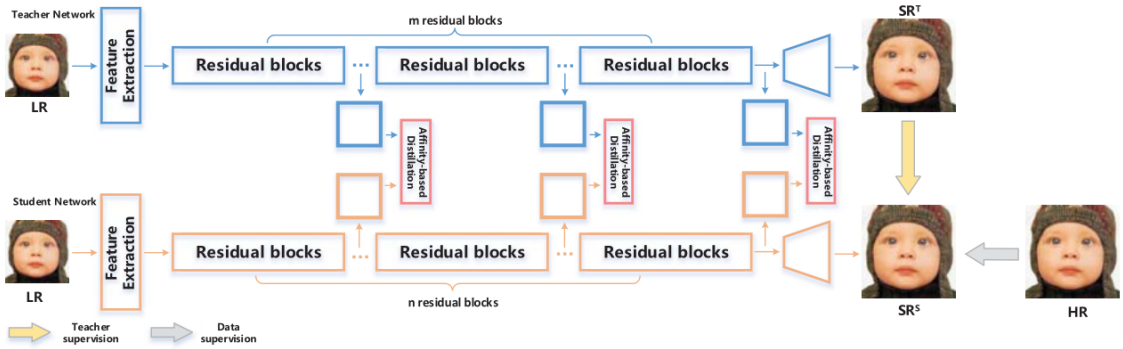

图 6: 基于特征相关性的知识蒸馏框架流程图

3.2 基于特征亲和的蒸馏 (FAKD)

知识蒸馏的关键是设计一个合适的模仿损失函数，能够成功传播有价值的信息，指导学生模型的训练过程。

以往的研究 zhang2019fast,saputra2019distilling 表明回归问题的特征表示空间是无界的。因此，现有的针对分类任务设计的蒸馏方法[22-23,25]可能并不适合图像 SR，因为存在巨大的解空间。为了有效地对图像 SR 进行知识蒸馏，必须对解空间进行限制。为此，我们设计了一个通用的基于特征亲和的知识蒸馏框架，以实现高效的知识蒸馏。

给定一批特征图 $F \in R^{b\times C\times W\times H}$，我们首先将其重塑为三维张量 $F \in R^{b\times C\times W\times H}$，分别为实例维、通道维和空间维。为了利用特征图中的一致性，我们计算亲和力矩阵 A。它是使用低层次、中层和高层的特征图生成的，以表示不同的相关性级别。鼓励学生网络产生与教师网络相似的亲和力矩阵，基于特征亲和力的蒸馏损失可以表述为:

$$L_{AD} = \frac{1}{|A|}\sum_{l=1}^{l'} \left\| A_l^S - A_l^T \right\|_1, \tag{6}$$

其中，A_l^T 和 A_l^S 为从第 l 层特征图中提取的师生网络亲和力矩阵；l' 是我们选择提取的层数。$|A|$

表示亲和矩阵中元素的个数。

为了保持像素之间的空间连续性，我们从空间角度考虑亲和矩阵，旨在探索像素之间的关系。该通道如图 7所示，其中每个像素都被视为一个 C 维向量 (红色列)，并且对每个列进行归一化。归一化后，每一列都是单位长度，因此两个像素点之间的余弦相似度可以简单地通过内积得到，经验上效果很好。空间亲和矩阵公式为:

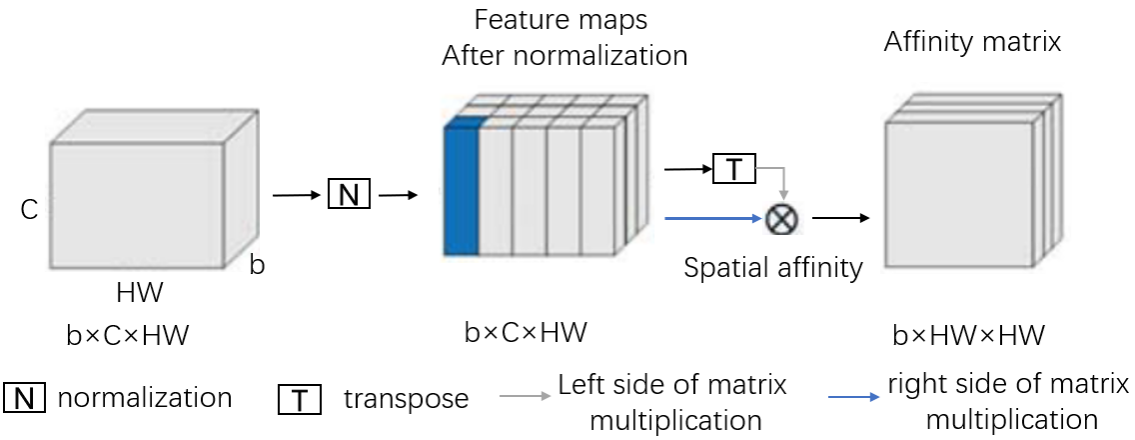

图 7: 空间亲和力计算，C 表示通道数，H,W 表示空间大小，b 表示批处理大小，蓝色像素是亲和的维度，将被归一化。

$$\widetilde{F}_{[i,:,j]} = \frac{F_{[i,:,j]}}{\left\|F_{[i,:,j]}\right\|_2^2}, \tag{7}$$

$$A = \widetilde{F}^T \cdot \widetilde{F}, \tag{8}$$

其中 $\widetilde{F}$ 为归一化特征映射。生成的空间亲和矩阵大小为 $b{\times}HW{\times}HW$。空间亲和矩阵中的每个元素表示两个像素之间的空间相关性。

3.3 总损失函数

基于特征亲和力的蒸馏，我们实证发现教师监督 (TS) 和数据监督 (DS) 也有助于提高蒸馏性能，如图 6所示。TS 和 DS 分别将学生的输出与教师的和地面的真实图像进行比较，如式 9和式 10所示。因此，学生网络既可以接收到教师分布的监控信号，也可以接收到真实数据分布的监控信号。总损失函数如式 11所示。

$$L_{TS} = \sum_{l=1}^{N} \left\| I_{SR}^{S(i)} - I_{SR}^{T(i)} \right\|_1, \tag{9}$$

$$L_{DS} = \sum_{l=1}^{N} \left\| I_{SR}^{S(i)} - I_{HR}^{T(i)} \right\|_1, \tag{10}$$

$$L(\theta) = \alpha L_{DS} + \beta L_{TS} + \gamma L_{AD}, \tag{11}$$

其中，I_{SR}^{S}、I_{SR}^{T} 和 I_{HR} 分别是来自学生输出、教师输出和真实图像数据集的图像。α，β 和 γ 是权重系数，使用该损失函数，可以使学生模型在基于特征相关性的蒸馏框架中学习到更多教师模型的有效知识。

4 复现细节

4.1 创新点

我们主要工作在于提出了一种新颖的基于空间特征特征相关性的知识蒸馏方法，该方法普遍适用于各种基于卷积神经网络的图像超分辨率模型。我们在基于特征相关性的知识蒸馏框架下对师生模拟损失函数和特征矩阵进行改善，在其提取特征图中二阶信息的基础上，将空间相关性信息与通道相关性信息相结合，其高维度的复杂相关性有助于指导学生模型模拟教师模型强大的图像重建能力。

我们在四个标准的开源数据集上对我们所提出的蒸馏特征矩阵和损失函数分别进行了详细的实验，包括与最新方法的对比实验和消融实验。实验结果表明，我们提出的蒸馏方法进一步有效压缩了基于卷积神经网络的图像超分辨率模型，将特征知识从强大的教师模型中转移出来，提高了学生网络的性能。

基于知识蒸馏的模型，一个较为重要的创新点在于如何表达教师模型和学生模型各个层次的特征关系，即如何设计蒸馏矩阵。以往论文的创新点同样主要在于此，如 FitNet 是直接将教师模型和学生模型的特征图对齐进行对比；转移注意力 (AT)[25]是对比师生模型各自在通道维度上聚合特征图生成的注意力图；解决方案流程 (FSP) 是指对比所有中间层中每相邻两个特征图之间所计算的 Gram 矩阵；通道亲和力 (CA) 是指把每个通道视为 HW 维向量，在每个通道中归一化后用内积的方式获得两个通道间的亲和力，得到的通道亲和矩阵是一个 $b\times C\times C$ 的矩阵，将该矩阵作为蒸馏特征矩阵；实例亲和力 (IA)[34]是指把每个批次（实例）视为 CHW 维向量，在每个实例中归一化后用内积的方式获得两个实例间的亲和力，得到的实例亲和矩阵是一个 b×b 的矩阵；空间亲和力 (SA)[13]与 CA 相似，是指把每个通道视为 HW 维向量，在每个通道中归一化后用内积的方式获得空间维度的亲和力，得到的空间亲和矩阵是一个 $b\times HW\times HW$ 的矩阵；相关一致性 (CC)[34]在每个实例中归一化后用内积的方式获得两个实例间的亲和力，再通过高斯 RBF 核函数计算得到实例内在相关性的一致性，得到的相关一致矩阵依然是一个 b×b 的矩阵。

CA 和 SA 仅仅以通道或者空间关系作为特征的表达方式，并且通过实验证明其有效性，我们考虑到，单一方面的特征是可能并不能完整的表示蒸馏的知识，能否结合空间和通道两种关系设计蒸馏矩阵，达到更好效果。因此，我们在实验中设计了**基于空间通道相关性矩阵的知识蒸馏机制**，用于提高图像超分辨率模型的性能。

4.1.1 空间通道特征的图像超分辨率模型的知识蒸馏框架

于是，为了实现高效的图像超分辨率方法，我们通过激励学生模型学习教师模型中空间相关性和通道相关性的知识，提出了一个新颖的基于空间通道特征相关性的知识蒸馏方法，如图 8所示是我们的知识蒸馏框架流程图，我们首先从退化的低分辨率图像提取特征，然后通过教师和学生的网络分别进行传播，传播过程中我们会对教师模型和学生模型在中间层产生的特征图进行处理，每次都会将得到的特征图用来计算各自的空间通道相关性矩阵，然后求取两个矩阵之间的特征相关性蒸馏损失，以此来监督学生模拟教师模型的图像重建能力。该框架的教师模型和学生模型分别由 t 个和 s 个残差块组成，它们以不同的超参数共享相同的体系结构，最后我们还将教师输出图像和真实图像分别作为教

师监督信号和数据监督信号来辅助学生模型的知识学习。

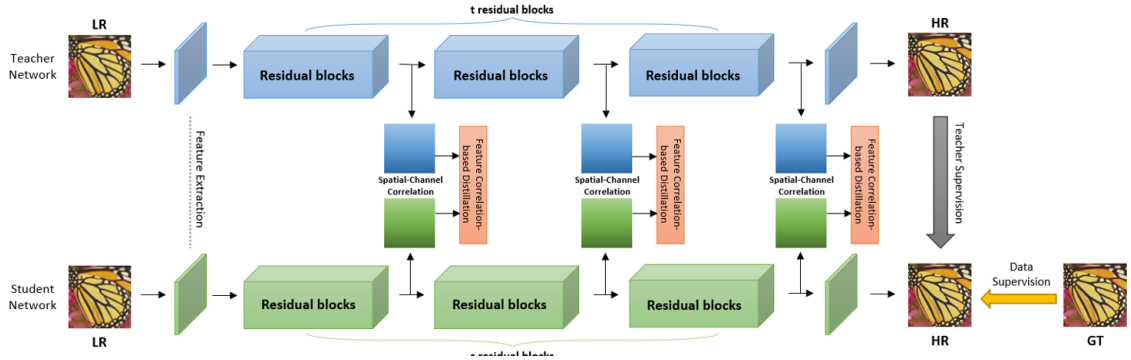

图 8: 基于空间通道相关性的知识蒸馏框架流程图

4.1.2 空间通道相关性矩阵

在我们提出的框架中，通过空间通道相关性来计算蒸馏损失是我们主要的创新点，而该损失是由教师模型和学生模型的空间通道相关性矩阵计算得到。计算过程如图 9所示，给定单一批次的特征图 $F \in R^{b\times C\times W\times H}$, 我们首先将其分别重塑为三维张量 $M_{cw} \in R^{b\times CW\times H}$ 和 $M_{ch} \in R^{b\times CH\times W}$，其中 b 代表实例，C 代表通道，H 和 W 分别代表空间维度的高和宽。为了保持各通道像素间的邻接关系，我们从不同空间维度的角度来考虑特征矩阵，以探索像素之间的关系。两个三维张量可以被视为二维空间中不同维度视角下与通道维度的信息结合，其中每个像素可被视为一个 H 维向量或 W 维向量，并在所在列上进行归一化。归一化后，每一列都是单位长度，因此两个像素之间的余弦相似性可以通过内积简单得到，这在经验上效果不错。空间通道相关性矩阵的公式如下：

$$\widetilde{M}_{[i,j,:]} = \frac{M_{[i,j,:]}}{\left\| M_{[i,j,:]} \right\|_2^2} \tag{12}$$

$$F = \widetilde{M}_{cw}^{T} \cdot M_{cw} + \widetilde{M}_{ch}^{T} \cdot M_{ch} \tag{13}$$

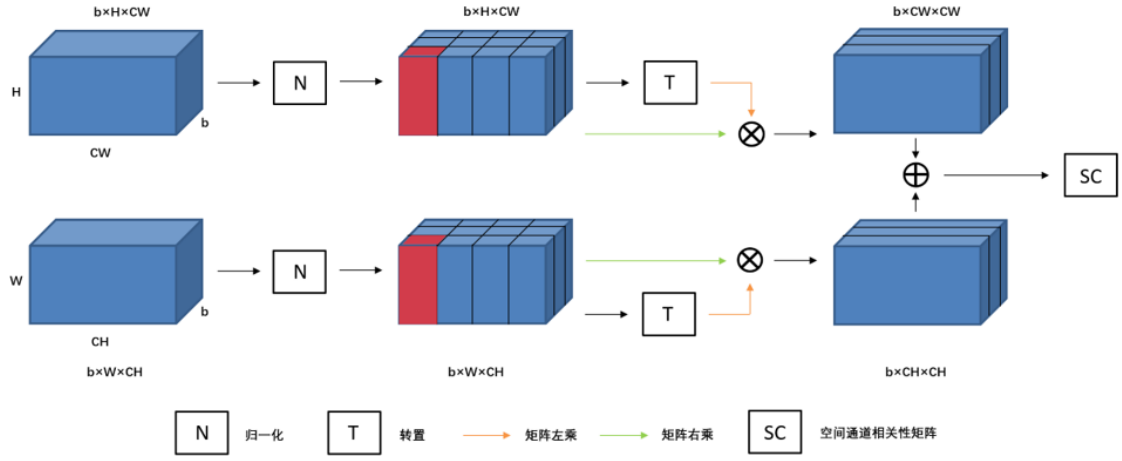

图 9: 空间通道相关性矩阵计算过程

其中 $\widetilde{M}$ 是归一化的特征图，两个三维张量内积得到的矩阵大小分别为 $b\times CW\times CW$ 和 $b\times CH\times CH$。由于预处理图像时通过平铺使得 W 与 H 相等，所以最终可以得到空间通道相关性矩阵 F。该矩阵中的每个元素表示在两个空间维度下各通道像素间的相关性。

4.1.3 空间通道特征相关性蒸馏损失

凭借特征映射中的一致性，不同中间层之间的空间通道相关性矩阵 F 可被视为由低层、中层和高层的特征映射生成的，以表示不同的关联等级。而学生模型会被激励生成与教师模型相似的空间通道相关性矩阵，因此空间通道特征相关性蒸馏损失可以表示为：

$$L_{SC} = \frac{1}{|F|} \sum_{l=1}^{n} \left\| F_l^S - F_l^T \right\|_1, \tag{14}$$

其中 F_l^S 与 F_l^T 是从第 l 层的特征图中提取的学生模型和教师模型的空间通道相关性矩阵，n 是所有选择提取的中间层数量，$|F|$ 是空间通道相关性矩阵中的元素数量。除了基于特征相关性的提取，由经验发现教师监督（TS）和数据监督（DS）同样有助于提高蒸馏性能。教师监督和数据监督指将学生模型的输出图像分别与教师模型的输出图像和真实图像进行比较。由此，学生模型可以接受来自教师模型分布和真实数据分别的共同监督信号，则总损失函数如下式所示：

$$L_{TS} = \sum_{l=1}^{N} \left\| I_i^S - I_i^T \right\|_1, \tag{15}$$

$$L_{DS} = \sum_{l=1}^{N} \left\| I_i^S - I_i^{GT} \right\|_1, \tag{16}$$

$$L(\theta) = \alpha L_{DS} + \beta L_{TS} + \gamma L_{SC}, \tag{17}$$

其中 I_i^S，I_i^T 和 I_i^{GT} 分别是指学生模型的输出图像、教师模型的输出图像和真实数据集的高分辨率图像，α，β 和 γ 则分别是对应的权重系数。使用该损失函数，可以使学生模型在基于特征相关性的蒸馏框架中学习到更多教师模型的有效知识。

4.2 与已有开源代码对比

基于知识蒸馏的的图像超分辨率技术已经发展了一段时间了，现阶段的改进都是围绕着蒸馏机制进行的。在本次实验中，我们同样在这方面进行改进，我们参考了 FAKD 作者的代码，但作为蒸馏机制最关键的蒸馏关系矩阵和蒸馏总损失函数，我们采用自己提出的基于空间通道特征的蒸馏机制。部分关键代码如图所示。

```
#空间通道特征相关性
def testdis(fm):
    if (fm.size(1) > 64):
        fm = F.avg_pool3d(fm, (4, 1, 1))
    hc = fm.view(fm.size(0), -1, fm.size(2))
    wc = fm.transpose(1, 2)
    wc = wc.reshape(wc.size(0), wc.size(1), -1)
    norm_hc = hc / (torch.sqrt(torch.sum(torch.pow(hc, 2), 1)).unsqueeze(1).expand(hc.shape) + 0.0000001)
    norm_wc = wc / (torch.sqrt(torch.sum(torch.pow(wc, 2), 1)).unsqueeze(1).expand(wc.shape) + 0.0000001)
    hchc = norm_hc.bmm(norm_hc.transpose(1, 2))
    wcwc = norm_wc.transpose(1, 2).bmm(norm_wc)
    ans = hchc.unsqueeze(1) + wcwc.unsqueeze(1)
    gamma, P_order = 0.4, 2
    corr_mat = torch.zeros_like(ans)
    for p in range(P_order + 1):
        corr_mat += math.exp(-2 * gamma) * (2 * gamma) ** p / math.factorial(p) * torch.pow(ans, p)
    return corr_mat
# 通道相关性
def channel_similarity(fm):
```

图 10: 空间通道相关性计算的部分代码

```
    def forward(self, student_sr, teacher_sr, hr, student_fms, teacher_fms):
        # DS Loss
        DS_loss = self.loss[0]['function'](student_sr, hr) * self.loss[0]['weight']
        self.log[-1, 0] += DS_loss.item()

        # TS Loss
        TS_loss = self.loss[1]['function'](student_sr, teacher_sr) * self.loss[1]['weight']
        self.log[-1, 1] += TS_loss.item()

        loss_sum = DS_loss + TS_loss


        if self.feature_loss_used == 0:
            pass
        elif self.feature_loss_used == 1:
            assert(len(student_fms) == len(teacher_fms))
            # assert(len(student_fms) == len(self.feature_loss_module))

            for i in range(len(self.feature_loss_module)):
                feature_loss = self.feature_loss_module[i](student_fms[i], teacher_fms[i])
                self.log[-1, 2 + i] += feature_loss.item()
                loss_sum += feature_loss

        self.log[-1, -1] += loss_sum.item()

        return loss_sum
```

图 11: 总体损失函数部分代码

其中，我们已经在创新点部分介绍过了基于空间通道特征的蒸馏机制，并以此设计了一套基于蒸馏该机制的图像超分辨率算法框架作为实验的创新内容。在之后的实验设计和结果部分，可以看到我们提出的蒸馏机制相对于作者和以往的各种方法性能都有了显著提升。

蒸馏机制和损失函数是算法框架最为关键的一部分，我们通过自己编写代码实现。我们的任务与以往的任务都是相同的，所以，在网络的架构、图像的处理等方面我们大体上是参考了 FAKD 作者的代码进行编写，但是在实验设计上，我们采用了不同的超参数，并且在网络框架中设计了不同的残差组合残差快。EDSR[11]和 RCAN[15]两个经典框架我们直接使用了原作者提供的模型。在与其它结果比较的过程中，FAKD[13]作者已经实现的用于比较的其它蒸馏方法，我们在额外实现了用于比较的蒸馏机制相关一致性 (CC) 的蒸馏方法，不过为保持相同的蒸馏框架, 没有计算实例一致性，只保留了相关一致性，即把每个批次（实例）视为 CHW 维向量，在每个实例中归一化后用内积的方式获得两个实例间的亲和力，再通过高斯 RBF 核函数计算得到实例内在相关性的一致性，得到的相关一致矩阵依然是一个 $b \times b$ 的矩阵。

4.3 实验设计

4.3.1 实验数据

在接下来基于知识蒸馏的图像超分辨率方法研究实验中，我们将收集的 DIV2K 数据集中的 800 份训练集作为蒸馏框架中师生模型的训练数据，DIV2K 是图像超分辨领域经典的训练图片。除此之外，我们还引入经典的开源基准测试集 Set5、Set14、B100 和 Urban100 作为本次实验的测试图像集，其中 Set5、Set14、B100 数据集中包含的是一系列的自然图像，Urban100 则是关于城市建筑的图像。

4.3.2 实验参数

我们使用的知识蒸馏框架对网络结构和优化算法毫无影响，以下实验都严格遵守控制变量原则。首先我们对数据集进行双三次插值以得到对应尺寸的低分辨率图像，并以 PSNR 作为评估指标。其次，我们使用 EDSR 和 RCAN 均分别作为教师模型和学生模型，以验证我们的通用蒸馏方法的有效性。其中教师模型和学生模型我们通过堆叠包含大量残差块的残差组来构建，不过学生模型的残差块数量有所减少，以减少其总参数数量至教师模型的 30% 左右，即两个网络的教师模型和学生模型均堆叠 10 个残差组，单个残差组中教师模型的残差块数量为 20 个，而学生模型的残差块数量为 6 个。

学生模型使用 Adam 优化器进行训练，初始学习率设置为 $1\times10^{(}-4)$，训练一共迭代 4×10^{5} 次。

5 实验结果分析

5.1 蒸馏有效性实验

我们以上述实验数据和实验参数设置进行了后续几个实验，首先是消融实验，消融实验是为了验证损失函数中不同模块的影响，我们使用各种损失函数来训练 EDSR。整体结果如表 1所示，其中数据是代表损失函数中不同模块对于超分性能的影响。在第一行中，仅采用数据监督（DS），即对学生模型不进行蒸馏，仅使用高分辨率真实图像作为监督。其他三行的学生模型均使用不同的蒸馏策略进行训练。当分别添加教师监督（TS）和特征相关性蒸馏损失（SC）时，性能表现均有所提升。通过进一步研究整体蒸馏框架的影响，我们同时加入了数据监督、教师监督和特征相关性蒸馏损失时，该策略得到了最佳的结果，如表 1中最后一行所示。与其他三种蒸馏策略相比，我们的蒸馏策略可以在不同的数据集上均获得理想的性能增益，这证明了我们的蒸馏方法的有效性。在之后的实验中，我们将数据监督、教师监督和特征相关性蒸馏损失的组合策略作为默认实验设置。

Module			PSNR			
DS	TS	SC	Set5	Set14	B100	Urban100
✓			31.823	28.361	27.398	25.546
✓	✓		31.845	28.375	27.411	25.578
✓		✓	31.899	28.392	27.434	25.592
✓	✓	✓	**31.903**	**28.406**	**27.445**	**25.607**

表 1: 损失函数中不同模块的影响，此处实验模型为 EDSR，放大比例为 4

5.2 特征知识蒸馏对比实验

我们用其他七种先进的知识蒸馏的特征提取模块或者变体来替换我们的特征相关性蒸馏损失，以验证我们的蒸馏方案的优越性。参照对象包括 FitNet、注意力转移（AT）、解决方案流程（FSP）、通道亲和力（CA）、实例亲和力（IA）、空间亲和力（SA）和相关一致性（CC）。

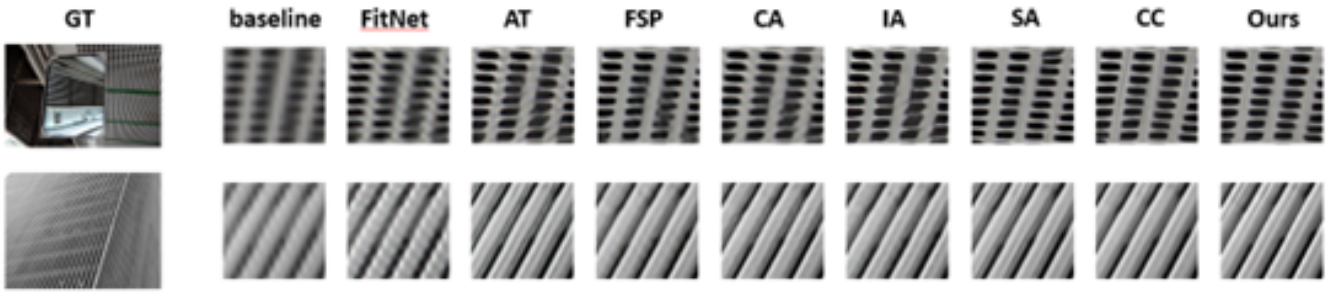

图 12: 各种特征模块蒸馏策略下的复原图像可视化对比

Method	PSNR			
	Set5	Set14	B100	Urban100
Baseline	31.823	28.361	27.398	25.546
FitNet	31.875	28.391	27.427	25.588
AT	31.893	28.387	27.426	25.579
FSP	31.892	28.382	27.427	25.573
CA	31.900	28.378	27.415	25.566
IA	31.868	28.381	27.423	25.561
SA	31.842	28.382	27.436	25.572
CC	31.905	28.393	27.439	25.591
Ours	**31.907**	**28.406**	**27.445**	**25.607**

表 2: EDSR(x4) 使用我们的特征知识蒸馏方法与其他方法的对比实验结果

从表 2可以看出，我们的基于空间通道相关性矩阵的特征相关性蒸馏方法优于其他蒸馏策略。与 FitNet 相比，我们的优势可以归因于特征转换后的有界表示空间；与 AT 与 FSP 相比，我们的方法将教师模型中的知识转化为一个信息量更大的压缩空间；与其余基于亲和力的方法相比，我们的方法能够结合空间和通道的角度捕获二者的相关性信息，比单独在实例域、通道域和空间域提取的信息要丰富的多；而与相关一致性的方法相比，在非海量实例（即非 $b \gg (CW + CH)$) 的情况下，从特征矩阵空间上我们的方法蕴含更多的知识信息，其次相比于在中间层特征图上探究各实例内在相关的一致性，特征图中通道域和空间域的信息更有助于学生模型学习如何重建图像。

5.3 评估结果

EDSR 和 RCAN 两种网络模型在同师生蒸馏环境下的定量评估结果（PSNR）如表 3和表 4所示。我们以批大小为 4 于 100 个时期（epoch）的实验条件训练各自的学生模型。在三个尺度 ×2、×3 和 ×4 下的所有数据集，我们的基于特征相关性的知识蒸馏方法均取得了最佳表现，这成功证明了我们方法的有效性和优越性。EDSR 和 RCAN 的平均性能增益约为 0.048dB，RCAN 的峰值信噪比增益大于 EDSR，这主要归功于 RCAN 模型的性能更强大，可以提供更多的特征知识。因此 RCAN 的学生模型可以从教师模型身上捕获更多有效的监督信号，从而提高性能表现。除此之外，我们还进行了模型尺寸分析，结果表明，在性能合理可接受的情况下，模型参数量大幅下降。我们的蒸馏方法在不引入额外参数的同时，可以稳定提高性能，并且大幅减小模型尺寸。

我们对 RCAN 在 ×2 尺度下的评估结果做了可视化对比，如图 13所示。对比起未使用我们的知识蒸馏策略的输出图像，我们蒸馏策略下的学生模型重建后的图像更加清晰，也恢复了更多的线条细节等等，这同样证明了我们的知识蒸馏机制可以有效地从教师模型那里转移修复图像的相关知识。

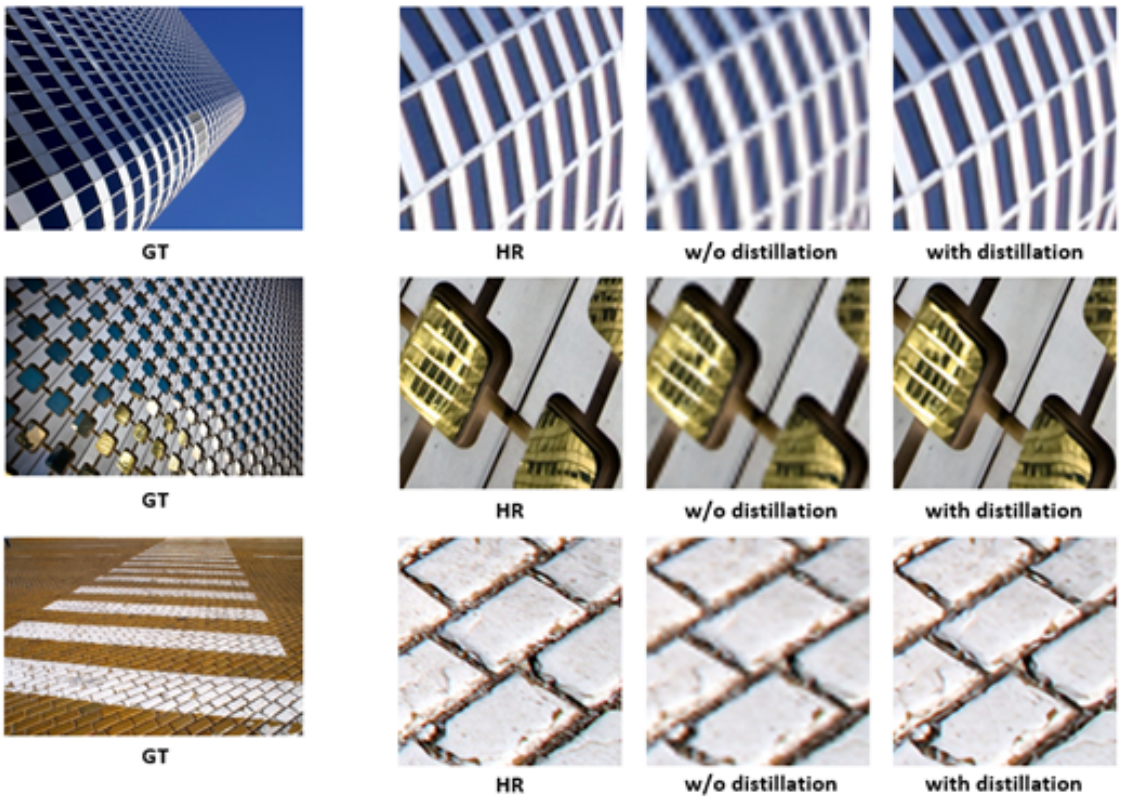

图 13: 该框架下 RCAN(x2) 的复原图像可视化对比

Datasets	Scale	EDSR		
		TN	SN w/o KD(Ours)	SN with KD(Ours)
Set5	x2	37.824	37.752	**37.793**
	x3	34.299	34.114	**34.165**
	x4	32.001	31.842	**31.907**
Set14	x2	33.323	33.228	**33.311**
	x3	30.189	30.108	**30.163**
	x4	28.423	28.356	**28.406**
B100	x2	32.035	31.985	**32.021**
	x3	28.988	28.907	**28.969**
	x4	27.457	27.411	**27.445**
Urban100	x2	31.336	31.279	**31.319**
	x3	27.679	27.619	**27.666**
	x4	25.621	25.570	**25.607**
Params		4.71M	1.56M	1.56M

表 3: 评估结果，以 EDSR 作为教师网络，最接近教师模型性能的最优结果已通过加粗标示

6 总结与展望

6.1 工作总结

随着注意力机制、重参数化、知识蒸馏等方法在图像分类等高级任务的模型压缩问题中取得不错表现，研究者们也开始着手于将这些方法策略应用于图像超分辨率任务中，这也使得超分辨率网络模型开始朝着轻量化的方向发展。

我们从知识蒸馏在模型压缩方案中突出独有的特点出发，旨在提出一种普遍适用于各种现有的图像超分辨率卷积神经网络，且能够高效压缩模型且尽可能保留原有优秀性能的训练策略。具体来说，我们的主要工作包括以下几点：

Datasets	Scale	RCAN		
		TN	SN w/o KD(Ours)	SN with KD(Ours)
Set5	x2	37.942	37.857	**37.922**
	x3	34.397	34.301	**34.383**
	x4	32.208	32.144	**32.191**
Set14	x2	33.501	33.436	**33.489**
	x3	30.334	30.273	**30.303**
	x4	28.559	28.515	**28.551**
B100	x2	32.141	32.066	**32.108**
	x3	29.111	29.049	**29.098**
	x4	27.555	27.523	**27.537**
Urban100	x2	31.787	31.712	**31.758**
	x3	28.052	28.008	**28.033**
	x4	25.933	25.891	**25.925**
Params		15.59M	5.17M	5.17M

表 4: 评估结果，以 RCAN 作为教师网络，最接近教师模型性能的最优结果已通过加粗标示

我们介绍了图像超分辨率方法和知识蒸馏技术的研究背景和意义，同时总结了知识蒸馏技术在图像超分辨率领域中的研究现状，然后从问题模型、常用的上采样框架、上采样方法这几个方面展开了对研究现状的具体分析，其次也对目前主流用于评估模型性能的峰值信噪比指标进行阐述。除此之外，我们还对知识蒸馏的知识类型和蒸馏策略进行介绍。

通过以上调研，我们发现了现有的知识蒸馏方法对于图像超分辨率卷积神经网络的模型压缩问题上仍存在巨大的提升空间。于是我们致力于将改善基于特征的知识蒸馏的模拟损失函数作为主要创新工作，并提出了一种新颖的基于特征相关性的知识蒸馏方法。该方法通过提取教师模型在通道域和空间域上重建图像过程中的特征相关性信息，去鼓励学生模型以少量参数进行模仿学习。

我们提出的方法可适用于时下各种前沿的图像超分辨率卷积神经网络，通过仿照原模型改变师生模型残差组与残差块的堆叠数量及堆叠方式，则可适应不同尺寸和通道数的特征图谱之间的知识迁移。

我们在四个标准的数据集上进行了详细的实验，包括蒸馏有效性实验、与最新方法的对比实验和性能定量评估实验。实验结果显示，我们的方法均优于其他知识蒸馏方法，量化结果和可视化结果均证明了该方法的有效性和优越性。

6.2 不足与展望

目前我们的研究工作是围绕利用知识蒸馏方法对图像超分辨率卷积神经网络模型的轻量化问题展开的，适逢如今单图像超分辨率领域的卷积神经网络模型均包含残差学习的模块，而对于其他网络设计框架下的模型无法达到较优的蒸馏效果，因此在普适性上仍存在一定限制。其次我们提出的基于特征相关性的知识蒸馏方法对于空间域特征和通道域特征及其之间的内在相关性探索程度较低，即我们简单地将分解空间域信息和通道域信息做内积求取余弦相关性作为特征矩阵，而其中各域之间的权重和解耦方式未有得到充足的实验探究。最后我们的方法中的特征相关性蒸馏损失中也未能寻求到与实例域信息的相关性关系，与相关一致性方法的实验结果中也可看出，实例域信息和内在相关一致性仍存在丰富知识值得探索，其中核函数对于我们方法的增益策略也是一个值得研究的问题。以上问题将在未来的工作中进一步深入探索。

如果以上问题能够得到有效的解决，那么我们所提出的基于特征相关性的知识蒸馏方法将有望成

为图像超分辨率卷积神经网络模型压缩过程的一种标准方法，从而在此基础上开拓出更为新颖的蒸馏机制，具有极大的现实意义。

参考文献

[1] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution [C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. 2014: 184-199.

[2] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[3] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[4] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[5] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1646-1654.

[6] LAI W S, HUANG J B, AHUJA N, et al. Deep laplacian pyramid networks for fast and accurate super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 624-632.

[7] AHN N, KANG B, SOHN K A. Fast, accurate, and lightweight super-resolution with cascading residual network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 252-268.

[8] SAJJADI M S, SCHOLKOPF B, HIRSCH M. Enhancenet: Single image super-resolution through automated texture synthesis[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4491-4500.

[9] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution [C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. 2016: 694-711.

[10] BULAT A, TZIMIROPOULOS G. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 109-117.

[11] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.

[12] WANG Y, PERAZZI F, MCWILLIAMS B, et al. A fully progressive approach to single-image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 864-873.

[13] HE Z, DAI T, LU J, et al. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution[C]//2020 IEEE International Conference on Image Processing (ICIP). 2020: 518-522.

[14] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.

[15] ZHANG Y, LI K, LI K, et al. Image super-resolution using very deep residual channel attention networks [C]//Proceedings of the European conference on computer vision (ECCV). 2018: 286-301.

[16] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convnets[J]. arXiv preprint arXiv:1608.08710, 2016.

[17] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2736-2744.

[18] HE Y, ZHANG X, SUN J. Channel pruning for accelerating very deep neural networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1389-1397.

[19] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[20] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.

[21] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[22] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.

[23] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4133-4141.

[24] GAO Q, ZHAO Y, LI G, et al. Image super-resolution using knowledge distillation[C]//Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II. 2019: 527-541.

[25] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint arXiv:1612.03928, 2016.

[26] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1365-1374.

[27] ZHANG F, ZHU X, YE M. Fast human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3517-3526.

[28] SAPUTRA M R U, DE GUSMAO P P, ALMALIOGLU Y, et al. Distilling knowledge from a deep pose regressor network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 263-272.

[29] DAI T, CAI J, ZHANG Y, et al. Second-order attention network for single image super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 11065-11074.

[30] TONG T, LI G, LIU X, et al. Image super-resolution using dense skip connections[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4799-4807.

[31] HAN W, CHANG S, LIU D, et al. Image super-resolution via dual-state recurrent networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1654-1663.

[32] TIMOFTE R, DE SMET V, VAN GOOL L. A+: Adjusted anchored neighborhood regression for fast super-resolution[C]//Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12. 2015: 111-126.

[33] SCHULTER S, LEISTNER C, BISCHOF H. Fast and accurate image upscaling with super-resolution forests[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3791-3799.

[34] PENG B, JIN X, LIU J, et al. Correlation congruence for knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5007-5016.

[35] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant [C]//Proceedings of the AAAI conference on artificial intelligence: vol. 34: 04. 2020: 5191-5198.

[36] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4320-4328.

[37] FURLANELLO T, LIPTON Z, TSCHANNEN M, et al. Born again neural networks[C]//International Conference on Machine Learning. 2018: 1607-1616.