使用深度网络研究超低分辨率识别

王张阳，张世玉，杨应珍，刘鼎和黄Thomas

美国伊利诺伊大学厄本那-香槟分校贝克曼研究所，美国伊利诺伊州61801

{ zwang119，chang87，yyang58，dingliu2，t-huang1 } @ illinois.edu

# 抽象

*视觉识别研究通常假设感兴趣区域（ROI）的分辨率足够。这在实际情况中通常是违背的，激发我们去探索一般的超低分辨率识别（VLRR）问题。通常，VLRR问题中的ROI可以小于* 16×16*像素，即使是人类专家也难以识别。我们尝试使用深度学习方法来解决VLRR问题。我们主要利用超分辨率、领域自适应和鲁棒回归等技术，制定了一种专用的深度学习方法，并演示了这些技术是如何一步一步地结合在一起的。这就产生了一系列动机良好、功能强大的模型。分析和仿真结果都充分证明了由于引入新模型而导致的任何额外复杂性。由此产生的鲁棒的部分耦合网络同时实现了特征增强和识别，同时兼顾了应对LR-HR域失配的灵活性和对异常值的鲁棒性。最后，在三种不同的VLRR任务上评估了所提出模型的有效性，*

# 1.简介

到目前为止，尽管对象识别研究取得了可观的成就，但通常认为感兴趣区域（ROI）足够大，并且包含足够的识别信息。但是，这种假设通常在实践中不成立。一个典型的例子是来自视频监控的人脸识别[ 35 ]。由于到处都安装高清摄像机的成本高昂，因此大多数监视系统都必须依赖清晰度非常有限的摄像机。此外，通常以使观看区域最大化的方式使用广角照相机。反过来，场景中的脸部区域可能非常小且质量很差。在文本识别系统中[ 15]，廉价且用途广泛的相机使快速扫描文档成为可能，但其低清晰度也给扫描仪带来了挑战。

arXiv:1601.04153v1 [cs.CV] 16 Jan 2016

1

强大的字符分割和识别能力。在[ 6 ]中，作者解决了视觉字体识别（VFR）问题。尽管训练后的模型在合成数据上取得了良好的性能，但由于各种退化问题（包括低分辨率和压缩伪影），其实际通用性很差。尽管在野外环境下的检测碰到类似的问题很普遍，但是非常需要一种有原则的方法来解决它们。

遗憾的是，一般的*分辨率非常低识别*（VLRR）问题已在很大程度上被忽视，除了在人脸识别上某些已经有的努力之外。关于面部识别的经验研究[ 22 ]证明，独立的识别算法要求最低的面部分辨率在32×32 和64×64 之间。在传统的识别模型中，分辨率越低，识别性能越差。从HR到LR的严重信息丢失，使得不可能直接从LR图像中提取或恢复足够的可识别特征[ 35]。通常，VLRR问题中的ROI可能小于16×16 像素，甚至很难（但仍然可能）被人类观看者识别出。

在本文中，我们首次尝试使用深度学习方法来解决VLRR问题[18]。从最简单的基线模型开始，我们将逐步进行模型演化，逐步获得更复杂，功能更强大的模型。由于引入了新模型，任何额外的复杂度都可以通过分析和仿真结果来证明。最终结果命名为“ *鲁棒的部分耦合网络”*，同时实现特征增强和识别。它既具有克服跨分辨率域不匹配的灵活性，又具有对异常值的鲁棒性。所提出的模型用于解决三个不同任务上的真实VLRR问题，包括面部识别，数字识别和字体识别，所有这些均获得了非常出色的性能。

## 1.1。问题定义

在实际环境中，VLRR直接从低分辨率（LR）图像中识别视觉对象，而不存在任何相对应的高分辨率（HR）图像。但是，我们通过假设每个训练图像都有LR和HR版本，将HR图像作为“辅助变量”引入模型训练中。正如我们的以下实验所证实的那样，HR图像可以帮助发现更多易于在LR图像被忽略的可识别特征。在测试阶段，只有LR图像可用，并且应用VLRR模型而无需HR图像。

我们选择LR图像分辨率不大于16×16 ，HR不小于32×32 。在本文中，除非另有说明，否则我们将所选数据集中的原始训练图像视为HR图像，用于生成用于训练（以及测试）的LR图像。我们将原始图像下采样*s*倍，然后通过最近邻（NN）插值将它们放大回原始分辨率。放大操作旨在确保为层次卷积提供足够大的空间支持，并有助于特征传递。由于NN插值不会带来任何新信息，因此以下*将放大图像视为我们的默认LR图像*。

总之，该问题定义为：*从具有LR和HR版本的训练集图像中学习VLRR模型，并将该模型应用于LR测试集图像*。

# 2.**模型I：基本单一网络BasicSingleNetwork**

## 2.1 **动机Motivation**

深度卷积神经网络（CNN）[ 18 ]最近在诸如图像[ 29 ]，数字[ 23 ]，文本[ 16 ]和字体[[ 32 ]的识别任务中取得很好的成果而大受欢迎。但是，所有模型都假设ROI具有合适的分辨率[ 18 ]。流行的数据集，例如LFW [ 13 ]和ImageNet [ 18]，则大多具有中等到高分辨率的图像分辨率（通常每个维度约几百个像素）。对于VLRR任务，传统CNN模型的性能是否仍然合理尚待探索。因此，我们将从调查VLRR任务的最简单的基线从基本单CNN开始。

## 2.2。技术方法

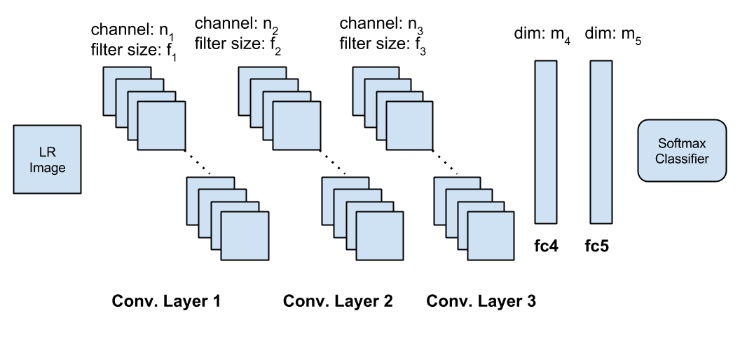


图1.传统卷积神经网络（CNN）的基本架构

图1 展示了我们的模型I（基本CNN），该模型类似于流行的ImageNet结构[ 18 ]，具有3个卷积层和2个全连接层。由于VLRR问题中的LR图像没有太多（低级和高级的）信息要分层提取，因此我们不涉及更深的体系结构。对于第I个（I = 1，2，3）卷积层，我们假设它是*Ñi*通道，卷积核尺寸*˚F i*，步长为1。注意一个LR图像（小于16×16像素）在VLRR中几乎没有可被利用自相似性，并且经常被破坏。因此，正如我们的实验所验证的那样，将大卷积核应用于提取patch（如[ 8 ]等）几乎没有什么好处。两个全连接层的维度分别为*m* 4和*m* 5。如[ 18 ]和许多其他人所建议的，网络是端到端的。

## 2.3 **仿真模拟Simulation**

我们采用流行的CIFAR-10和CIFAR-100数据集[ 17 ]作为模拟对象。CIFAR-10数据集由10类总共60,000张32 × 32彩色图像组成，每类6000张图像。每个类别有5,000张图像用于培训，而1,000张用于测试。CIFAR-100数据集与CIFAR-10一样，并且具有相同的总量，不同之处在于它具有100个类别，每个类别包含600张图像。在所有实验中，为了简化模型，我们将图像转换为灰度图。首先将原始（HR）图像按*s* = 4 比例缩小为8 × 8。然后通过NN插值将其上采样回32 ×32作为LR图像。对于每个LR图像，我们减去其均值并对其大小进行归一化，然后将其放回到恢复的HR图像中。在训练中加上默认的小的可加的高斯噪声（*σ* = 0.05）。

我们基于cuda-convnet包[ 18 ] 实现我们的模型。ReLU被用作非线性处理，并且应用了dropout。batch大小固定为128。学习率从0.1开始，当训练曲线达到平稳状态时，学习率除以10。我们还将在LR输入上训练的VLRR模型与在原始HR图像上训练的模型进行比较。最后一个完全连接层的大小始终固定为类别数，例如，对于CIFAR-10 ，*m* 5 = 10，对于CIFAR-100 ，*m* 100 = 100。*m* 4默认固定为1024。然后，我们更改其他网络配置以检验基线模型性能如何受到影响，如表1所示。容易得出以下结论：

* 与从HR图像获得的图像相比，以LR图像作为输入图像的VLRR模型的性能大大降低。
* 在VLRR实验中，较大的卷积核几乎没有好处，因为VLRR对象无法提供与HR图像一样多的空间分辨率（和细节）损失。

表1.在CIFAR-10和CIFAR-100数据集上，带有LR或HR训练图像的各种网络配置的top1的错误率（％）。

**这个表n和f画反了**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *n* 1 | *n* 2 | *n* 3 | *˚F* 1 | *f* 2 | *f* 3 | CIFAR-10 | | CIFAR-100 | |
| LR | HR | LR | HR |
| 5 | 3 | 1 | 64 | 64 | 32 | 27.47 | 14.41 | 45.03 | 38.37 |
| 7 | 3 | 1 | 64 | 64 | 32 | 28.86 | 14.24 | 49.82 | 37.07 |
| 5 | 5 | 1 | 64 | 64 | 32 | 28.82 | 14.06 | 50.01 | 36.69 |
| 5 | 3 | 3 | 64 | 64 | 32 | 29.13 | 13.78 | 49.95 | 35.02 |
| 5 | 3 | 1 | 32 | 32 | 32 | 30.03 | 17.91 | 50.81 | 42.23 |
| 5 | 3 | 1 | 64 | 32 | 32 | 27.87 | 15.45 | 48.91 | 40.27 |
| 5 | 3 | 1 | 64 | 64 | 64 | 28.08 | 14.32 | 46.67 | 37.61 |

* 在HR实验中，添加更多的卷积核通道通常会有所帮助，但这可能会损害VLRR性能。这可能是模型过度拟合的结果，因为LR主体包含很少的视觉特征。

另外，增加*m* 4的大小会稍微提高性能，同时带来显着的复杂性。我们还尝试重复第3层卷积达到更深的体系结构，最后观察到，由于LR图像没有丰富的视觉语义信息，因此深度的增加对VLRR没有明显的影响。以上发现证明了VLRR的独特特征，**即与常规的视觉识别问题相比，VLRR问题不能从更大的卷积核，更多的通道数或更深的模型中获益**。因此，我们默认采用以下配置：*n* 1 = 64，*f* 1 = 5；*n* 2 = 64，*f* 2= 3; *n* 3 = 32，*f* 3 = 1；*m* 4 = 1024。

# 3.**模型II：采用SR预训练的单一网络**

**Single Network with Super Resolution Pre-training**

## 3.1。动机

表1 提醒我们，对LR视觉对象进行直接分类是不可靠的，并且容易过拟合，因为它们的视觉特征稀缺且高度退化。另一方面，值得注意的是，尽管在实际测试中无法使用HR图像，但仍可以在训练中将其用作辅助信息以用来增强特征。

经典算法[ 35 ]依赖于第一次首先将超分辨率（SR）算法应用于LR图像，然后将其SR结果分类。恢复的HR图像不仅遭受不可避免的过度平滑和细节损失，还遭受重构过程[ 2 ] 引入的各种伪影，进一步破坏了随后的识别性能。[ 35 ]中的作者结合了判别性约束，以学习用于识别的HR和LR面部图像之间的关系。在[ 12 ]中，特定类别的面部特征作为先验已包括在SR中。这种“闭环”方法的性能始终优于传统的两阶段（SR再识别）pipelines。

## **3.2。技术方法 Technical Approach**

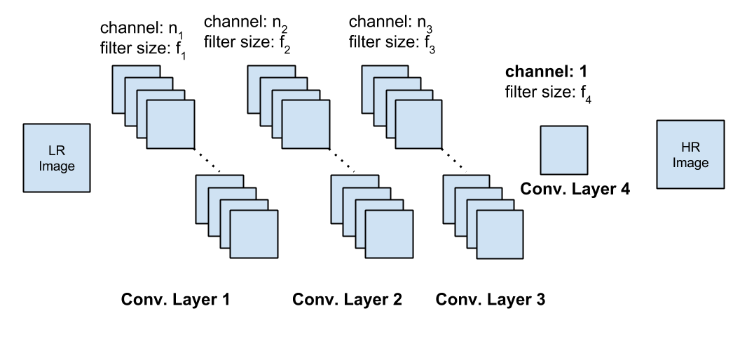


图2. 作为模型II的预训练部分的超分辨率（SR）子网络。注意它不包含完全连接的识别子网，与图1中的相同（fc4，fc5和softmax）。新的卷积层4层只有一个通道，经过预训练后将被丢弃。

模型II是模型I的“分解”版本，分为两个子网：用于无监督预训练的超分辨率（SR）子网络，以及用于监督微调的识别子网络。如图2 所示，SR子网由四个卷积层组成，将LR图像作为输入，将HR图像作为输出。它以无监督方式进行训练，以预测LR和HR图像之间的非线性映射。与图1中的卷积部分相比，很容易注意到新添加的Conv.4层，只有一个通道，将其输入特征图卷积以产生最终的重构。SR子网经过训练后，Conv。4层已弃用，并且两个完全连接的层以及softmax分类器添加在其顶部。现在，整个网络都具有与Model I完全相同的拓扑，并以受监督的方式联合优化。

较早的多层特征学习方法，例如自动编码器和降噪自动编码器，仅使用未标记的数据进行重构来进行权重的预训练，然后进行有监督的微调（[ 4 ]）。先前的工作[ 9 ]表明，无监督的预训练似乎主要起着正则化作用。模型II具有类似的想法，并且针对SR的无监督部分，而不是进行重建。最终的监督微调将那些未受监督的特征与类标签相关联。最终的网络旨在同时实现分辨率增强和识别。

## **3.3。模拟Simulation**

我们遵循默认的网络配置和与2.2节相同的实验设置。LR和HR图像首先用于训练SR子网，其放大倍数为4。CIFAR-10的分类错误率下降了2.32％（最终下降为25.15％），而在CIFAR-100上，错误率降低到46.50％，降低了3.45％。

备注：由于SR是一对多的问题，因此“很多”中的任何一个都是可行的解决方案，并且无法确保所添加的细节对于原始HR图像是真实的。相反，物体识别试图从“许多”中识别出一个。那么，人们可能会对SR预训练可能会改善视觉识别性能的原因有些担忧？我们的解释是，尽管不能保证一定是真实的，但SR预训练中产生的幻觉细节有助于发现有助于识别的细微但有区分度的特征。否则，它们很容易从LR图像中被忽略。

# 4.模型III：具有预训练和LR-HR特征转移的单一网络

**:SingleNetworkwithPre-trainingandLRHRFeatureTransfer**

## 4.1。动机（不太明白）

尽管通过引入SR预训练获得的改进有限，但在数据集CIFAR-10 / 100上仍然存在约10％的巨大性能差距。为什么会有这样的性能差距？

一个假设是，SR进程在VLRR问题中非常不适定，因为LR图像是缺少大量的细节[ 21 ]。当我们查看卷积核的可视化效果时，该假设得到了进一步的支持。虽然大多数HR滤波器学习清晰可见的边缘和曲线，但LR滤波器学习的可识别的结构特征却很少。后者也遭受更多“死”过滤器的困扰。因此，我们推测SR预训练过程没有为识别目的添加足够的判别能力。

通过从HR图像中导出更具区分性的细节，可以很容易地想到增强LR图像的细节。**这相当于学习从HR到LR的更多可转移细节，这将我们的情况与基于深度学习的域适应相关联。**HR图像可以视为源域，而LR图像构成目标域。Bengio等。人。在[ 3 ]中假设中间表示可以在输入分布的变体之间共享。作为跨域利用“共享”特征的具体示例，[ 10]使用堆叠式自动编码器（SAE）以无监督的方式从所有可用域中学习了跨域特征提取。使用SAE对线性SVM进行源域转换后的​​标记数据的训练，并在目标域上进行测试。

## 4.2。技术方法

超分辨率（SR）模型III的子网络，如图所示3 ，与域相关的适应技术[3 ，10 ，31 ]有关。**在域适应的情况下，我们将LR和HR图像视为来自两个相关但不重叠的域。**SR子网可以看作是两个完全耦合的通道。

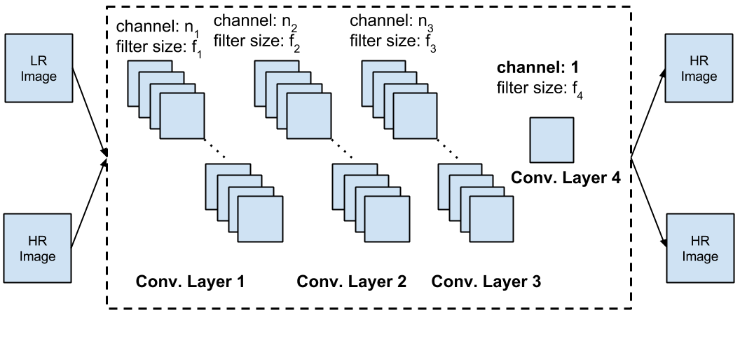


图3.作为模型III的预训练部分，通过LRHR特征传递增强的超分辨率（SR）子网。注意，它不包含完全连接的识别子网，该子网与图1中的相同（fc4，fc5和softmax）。

LR-HR通道从LR域样本重构HR域样本，这基本上等效于Model II的SR子网。另一条HR-HR通道从自身重构HR域样本，与前者相比，HR应当学习更多判别特征。通过强制两个渠道使用完全共享的隐藏层，该模型旨在学习跨域，可转移的特征。通过HR-HR通道的共享滤波器，LR-HR通道也有望获得用于SR预训练任务的增强特征。

## 4.3。模拟

我们首先对图3中通过特征转移增强的无监督SR模型进行预训练，然后进行有监督的调整。所有实验设置都遵循先前的设置，除了我们发现在训练这种“混合”模型时，选择较小的学习率有助于更快，更稳定地收敛。实际上，我们从0.01的学习率开始，并且在训练过程中不对其进行退火。在CIFAR-10上，错误率降低到21.72％，而在CIFAR-100上，错误率降低到43.03％。

# 5.模型IV：部分耦合的网络

# 5.1。动机

从模型I到模型III，我们在CIFAR-10 / 100模拟实验中获得了高达10％的精度提升。还有进一步的改进空间吗？

我们退后一步，重新思考模型III中的关键前提：我们利用LR和HR数据来学习跨域的完全共享的特征表示。由于每个LR-HR对表示同一场景或对象，可以合理的假设存在一个隐藏空间被LR和HR特征近似的共享[ 20 ，34 ]。然而，假设LR和HR数据在耦合子空间中具有严格相等的特征表示，这一假设似乎过于强烈。

如许多先前的工作[ 30 ] 所 认识的，LR和HR空间之间的映射可以是复杂的，空间变换的和非线性的。以人脸识别为例，实验研究[ 22 ]证明了在非常低分辨率的图像空间中的人脸相似性度并不能反映HR图像的实际人脸相似性，因此在HR Gallery中进行probe图像的直接匹配性能很差。这构成了LR和HR域之间的不匹配。常见的图像退化以及低分辨率导致的歧义进一步扩大了这种差异。

模型III中的简单假设导致其完全耦合的网络结构。相反，我们将大胆的假设，即使经过高度复杂的转换，LR和HR域特征也不会完全重叠。由于LR和HR域的不匹配，我们试图稍微relax一下模型III，以便有一定的灵活性来应对特征差异。

## 5.2。技术方法

图4 描述了*部分耦合的SR网络*（PCSRN），作为模型IV的预训练部分。PCSRN在图2中具有两个非监督通道。PCSRN和模型III（图3 ）中SR子网之间的主要区别在于，前者通过仅共享每一层中的*k i个*卷积滤波器（*i* = 1、2、3）而“松散地”耦合。他们倾向于捕获LR和HR域的共性特征模式[[1]](#footnote-0)[1]。基于上述假设，PCSRN还允许每个单个通道通过*m i* - *k i个*非共享卷积核学习特定于域的特征（[[2]](#footnote-1)*i* = 1、2、3），可以补充共享特征并处理不匹配问题。图3 也可以被看作是图4的特殊情况当*mi* = *ķ i*。

在对PCSRN进行预训练之后，我们遵循相同的程序来替换Conv.4，用两个全连接层替换，然后在每个后面独立地带有softmax分类器。每个通道（包括共享卷积核）都具有与Model II或III完全相同的拓扑。整个两通道的Model IV以有监督的方式进行微调。任一通道都会对其自己的输入进行分类，而这两个部分仍然可以利用共享卷积核进行交互。在测试中，我们“解耦”模型，并仅使用图4中的左LR-HR通道，包括所有共享卷积核。

“部分耦合”架构证明了我们的重要假设：VLRR问题中的域不匹配只会减少，即使经过学习转换也不会消失。它带来了重要的额外灵活性，可以跨域传输知识。尽管看起来像一个鲁莽的主意，但其好处将在接下来的工作中得到验证。

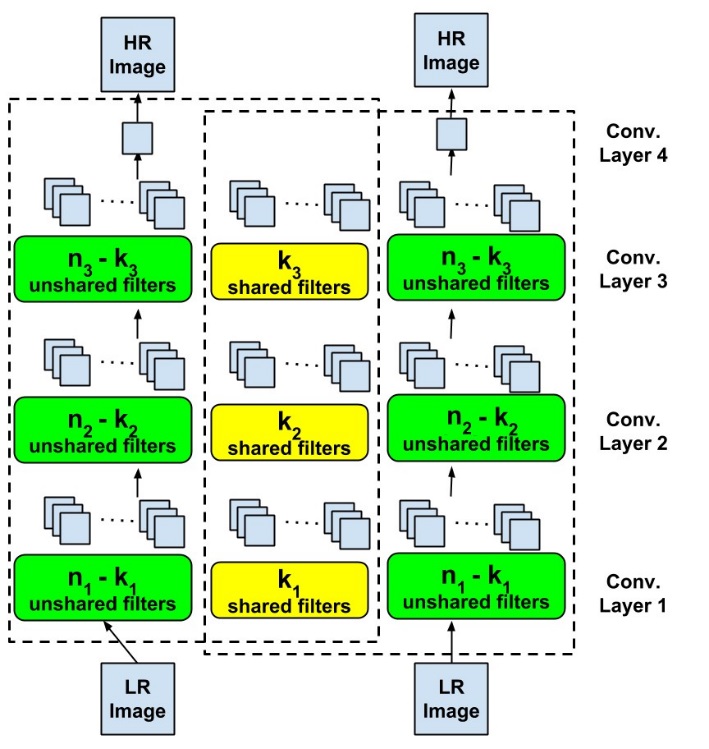


图4.部分耦合的SR网络（PCSRN），作为模型IV的预训练部分。请注意，它不包含完全连接的识别子网。对于每个通道，其识别子网与图1中的相同（fc4，fc5和softmax）。两个通道在第*i*层共享*k i*卷积滤波器。

## 5.3。模拟

图4 概述了总体思路，但仍不清楚从更复杂的体系结构中能否以及获得多少性能提升。另外，应该耦合多少个卷积核？为了回答这些问题，我们对CIFAR-10 / 100案例进行了定量研究。我们固定卷积核的数量*n* 1，*n* 2和*n* 3，并改变共享卷积核的数量*k* 1，*k* 2和*k* 3。我们为第*i*层定义*“耦合比”*：*c i* = *k i* / *ni*（ *i* = 1，2，3）。

为了确定最佳配置，我们对每个*c i*（*i* = 1、2、3）进行course grid search，从0到1，步长为0.25。即使这样，枚举搜索仍会导致训练模型至少具有5 3 = 125种配置。为了避免这种复杂的工作，我们通过参考领域适应的共同belief[ 10 ] 来制定一种策略，即随着层越深，两个域之间的差异有望逐渐减小[ 31 ]。因此，我们始终优先考虑增加较深的层的*c i*（从层3开始，然后是2和1）。一旦错误率趋于上升，我们将撤销最后的更改，转而去增加前一层的耦合率。如表2所示，结果强烈支持我们之前的“不重叠”假设。

表2.在CIFAR-10和CIFAR-100数据集上，使用不同配置的耦合卷积核的前1名分类错误率（％）。（*n* 1 = 64，*n* 2 = 64，*n* 3 = 32）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *ķ* 1 | *k* 2 | *k* 3 | *ç* 1 | *c ^* 2 | *c* 3 | CIFAR-10 | CIFAR-100 |
| 0 | 0 | 0 | 0 | 0 | 0 | 25.15 | 46.50 |
| 0 | 0 | 16 | 0 | 0 | 0.50 | 21.81 | 43.48 |
| 0 | 0 | 32 | 0 | 0 | 1.00 | 22.85 | 44.03 |
| 0 | 32 | 16 | 0 | 0.50 | 0.50 | 21.43 | 43.27 |
| 0 | 64 | 16 | 0 | 1.00 | 0.50 | 21.87 | 44.09 |
| 32 | 32 | 16 | 0.50 | 0.50 | 0.50 | 21.04 | 43.21 |
| 32 | 32 | 24 | 0.50 | 0.50 | 0.75 | 19.91 | 41.52 |
| 32 | 32 | 32 | 0.50 | 0.50 | 1.00 | 20.72 | 41.90 |
| **32** | **48** | **24** | **0.50** | **0.75** | **0.75** | **18.77** | 41.64 |
| 32 | 64 | 24 | 0.50 | 1.00 | 0.75 | 19.52 | 42.99 |
| 48 | 48 | 24 | 0.75 | 0.75 | 0.75 | 19.06 | 42.15 |
| 64 | 64 | 32 | 1.00 | 1.00 | 1.00 | 21.72 | 43.03 |

事实上，我们在表2 (0 <ci <1)中尝试的所有部分耦合模型都获得了比模型II (ci = 0)更好的性能，而其中大多数模型也优于模型III (ci = 1)。而且，观察到较高的层比较低的层对过度耦合更敏感。例如，当*c* 1和*c* 2都固定为05 时，只需将*c* 3从0.75增加到1.00，即可将CIFAT-10的错误率从19.91％提高到20.72％。结果，我们选择*c* 1 = 0.50，*c* 2 = 0.75，*c* 3 = 0.75，作为下文中的默认配置。生成的模型随着层数的增加逐渐缩小了两个域之间的差距，但仍允许每个通道具有特定于域的特征的一定灵活性。

除VLRR之外，“部分耦合”架构还可以潜在地应用于许多其他跨域识别问题[ 27 ]。我们已经证明，与完全耦合的模型相比，部分耦合的体系结构可以带来额外的收益。但是，如何针对特定情况自适应地选择最佳*c i*而不是专门的去尝试，任然未被解决。一种潜在的解决方案是学习一种映射函数，该函数可以规范两个特定域的表示形式之间的固有关系，例如[ 30 ]。我们把它留作将来的工作。

6.模型V：稳健的部分耦合网络

## 6.1。动机

到目前为止，借助SR预训练以及灵活的“部分耦合”域自适应，我们在CIFAR-10 / 100 LR数据上获得了令人满意的结果，这些数据是从清晰的HR图片人工下采样的。但是，来自低分辨率源的典型真实数据（例如视频监控）通常会伴随传感器噪声和脉冲异常值。即使在高分辨率情况下，也存在可能严重破坏结构的异常因素。在面部识别中可以找到一个这样的例子：由于面部既不是完美的凸面也不是传感器[ 22]，在定向照明下拍摄的面部图像通常会遭受自遮蔽成影，镜面反射或亮度饱和的困扰。姿势和表情的更改还引入了更多的空间局部伪影，从而使一部分数据严重损坏。

我们采用均方误差（MSE）作为SR预训练中的损失函数，该函数可能对异常值非常敏感。有人可能会争辩说，MSE在大多数无监督的深度学习模型中都能很好地发挥作用，例如Burger等。[ 5 ]表明，普通的多层感知器可以产生不错的结果并处理不同类型的噪声。实际上，已经观察到，这些异常值会削弱深度系统的学习能力，这些异常值在现实世界的数据中普遍存在[ 7 ]。已经提出了一些健壮的学习方法来免疫由异常值引起的有害影响。提出了一种对抗异常值的相关损失函数[ 24 ]。[ 33] 的作者]将另一种特定的降噪CNN连接起来，以处理复杂且依赖信号的离群值，但由于合并而导致参数量增加。

表3.在清晰和损坏的CIFAR-10 / 100数据集上，SR预训练中使用MSE或huber损失函数的top 1分类错误率（％）。

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | CIFAR-10 | CIFAR-100 |
| 0％（干净） | MSE（模型IV） | 18.77 | 41.64 |
| Huber（Model V） | 18.81 | 41.70 |
| 5％损坏 | MSE（模型IV） | 19.99 | 43.43 |
| Huber（Model V） | 19.04 | 42.18 |
| 10％损坏 | MSE（模型IV） | 23.03 | 46.20 |
| Huber（Model V） | 20.28 | 43.79 |
| 15％损坏 | MSE（模型IV） | 25.05 | 49.49 |
| Huber（Model V） | 22.24 | 46.18 |

## 6.2。技术方法

在SR预训练中，我们用凸和连续的Huber损失函数[ 14 ] 代替了IV模型的MSE损失。我们将其命名为“ *稳健的部分耦合网络”*，即模型V。由于huber损失对异常值的敏感性较低，在鲁棒回归中得到了广泛的应用。它包括两个部分对应l1和l2损失：

C:\Users\pc\AppData\Roaming\Tencent\Users\904759374\QQ\WinTemp\RichOle\DT}`1UQ7QWTXBM$UG(G@7MA.png

刀沿参数 *c* 直观地指定了误差的最大值，之后必须应用*l1误差函数*。我们知道可以设计出更稳健的损耗，在这里选择Huber损耗，因为它易于实现和计算。

在VLRR中，我们确实发现异常值具有更强的负面影响（我们将在仿真中进行验证）

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 表4.所有模型的审查和比较。（∗ ：异常值存在15％的异常值。）   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | 改善  最后的模型 | 加  复杂？ | 减少错误率（％） | | | CIFAR-10 | CIFAR-100 | | 模型二 | 利用人力资源培训数据并进行SR预培训 | 没有 | 2.32 | 3.45 | | 模型三 | 从HR数据中学习可转移的功能，以增强SR预训练 | 没有 | 3.43 | 3.47 | | 模型四 | 引入了额外的灵活性，并避免了部分耦合结构的过度规范化 | 是 | 2.95 | 4.86 | | 型号V | 在预训练中采用Huber损失以抵抗异常值 | 没有 | 2.81 \* | 3.31 \* | |

与常规识别相比，由于LR图像可能无法提供足够代表性的特征分布来绘制正确的数据流形。使用huber损失极大地缓解了该问题。

## 6.3。模拟

我们采用“干净”的CIFAR-10 / 100数据，同时通过在下采样之前在5％，10％和20％的随机选择像素上添脉冲噪声来创建“损坏”版本。然后，我们在两个数据集上测试模型VI和V，它们的唯一区别在于在SR预训练中选择MSE还是Huber损失。Dropout应用于两个模型。我们根据[ 14 ]和其他许多建议，令经验值*c* = 1.345固定在（1 ）中。在预训练期间，我们观察到Huber损失似乎在一定程度上加速了模型收敛。如表3，在没有异常值的情况下，huber损失的性能几乎与MSE相同（略差），但在存在异常值的情况下变得更好，

在具有15％异常值的情况下，最终比MSE的收益高出3％。

## 6.4。审查与比较

我们将表4中的所有五个模型进行比较，并评估它们是否带来额外的复杂性[[3]](#footnote-2)[2]。设计和发展模型的一般方法是专注于使模型既适应于特定任务（SR预训练等），又适应于特定数据域（LR-HR特征转移，部分耦合，huber损失等）。 ）[[4]](#footnote-3)

# 7.解决实际的VLRR问题

我们将我们提出的模型用于解决三种VLRR应用。除非另有说明，否则所有模型配置，参数和培训详细信息均与CIFAR-10 / 100仿真案例相似。

## 7.1。VLRR人脸识别

**数据集：**尽管已经开发了许多人脸图像数据集，例如LFW [ 13 ]，SCface [ 11 ]和CMU-PIE [ 28 ]，但它们大多数是在受控的室内环境中收集的，变化有限。最近在[ 25 ]中提出了一个新的UCCS数据集，作为无约束人脸识别的更大且更具挑战性的基准。[ 25 ] 的作者报告了通过提取73个面部属性[ 19 ]并使用SVM进行分类的原始分辨率图像的180个对象子集的基准面部识别结果，准确性约为78％top-1 。我们将裁剪后的面部区域归一化为80 × 80作为 HR，并将其下采样*s=5，*生成16 × 16的LR图像。我们对180个对象的子集进行评估，其中每个对象有25张或更多的图像。我们使用4,500张图像进行训练，其余935张用于测试。



图6.来自UCCS数据集的裁剪面部示例[ 25 ]。

**性能**：考虑到训练数据不足（对于不受约束的面部数据集，这是一个普遍的缺陷），我们对所有模型[[5]](#footnote-4)[3]进行逐层贪婪无监督训练[ 4 [[6]](#footnote-5)]。如图5[[7]](#footnote-6)（a），随着我们模型的发展，性能不断提高，最终达到40.97%@top-1和22.35%@top-5，与最简单的基线模型相比，可观的边际利润约为10％@ top-1。模型V在前1个结果中正确分类了935个测试样本中的552个，在前5个结果中正确分类了726个。模型II比模型I改善了4.60％，而模型III进一步提高了2.57％。通过SR预训练和LR-HR传递，可以确定幻觉特征在此任务中的关键作用。此外，由于有了更多的免费模型参数，部分耦合的体系结构使性能提高了1.71%@top-1。最后，胡贝尔损耗的效用带来了1.06%@top-1的额外收益。

## 7.2。VLRR数字识别

|  |
| --- |
| （a） （b） （c）  图5.在三个实际的VLRR应用程序上，所有五个模型的top-1和top-5错误率。 |

数据集SVHN [ 23 ]是具有高度复杂场景的真实世界大量数字图像数据集。请注意，所有数字高度的中位数为28像素，而超过10,000张图像的原始高度小于10像素[ 23 ]。这证实了SVHN是VLRR的适当研究对象。我们利用604,388张（包括额外的图片）图像进行训练，并使用26,032张图像进行测试，涵盖10个课程。将所有图像调整为HR时的大小为32 × 32，并通过*s* = 4（8 × 8）的下采样获得LR图像。为了进行训练，我们同时提供了HR和LR图像，而测试集仅查看了LR图像。请注意，SVHN的预处理会在感兴趣的某些数字的侧面引入“分散数字” [23 ]。考虑到超低分辨率图像的高度噪声敏感性，这些强烈的离群值可能会极大地混淆模型[ 35 ]。

性能我们计算所有五个模型的top-1和top-5错误率。如图5 （b）所示，模型仍然实现了最低的错误率，记录为43.02%@to-1和29.18%@top-5。请注意，对于高度小于25像素的原始SVHN图像，人类top-1精度仅为82.0％[ 23 ]。另外，值得我们特别注意的是，模型V比模型IV获得4.05%@top-1的余量，显示出强大的Huber损失的巨大好处，特别是在存在异常值（分散数字）的情况下。II型和III型之间还有4.03%@top-1的性能差距，这证明LR-HR传输增强了特征学习。此外，全耦合和部分耦合的体系结构还展示了

2.67%@top-1差异，揭示了潜力

## 7.3。VLRR字体识别



图7. VFR测试集[ 6 ]中的示例（标准化）。

在[用于数据集的数据集VFR 6 ，32 ]包括2383字体类。对于每个班级，将渲染1,000张合成文本图像进行训练，标准化高度为105。

收集了一个测试集，其中包含325个真实世界的93个类别的图像，并且具有各种ROI（文本）尺寸。进一步的统计数据表明，大约有1/6的测试图像（52）属于VLRR情况（例如，字符高度小于16像素），即使使用最新型号，大多数“ VLRR图像”也无法识别[ 32 ]，由于归一化之后非常大量的细节损失和压缩伪像。我们的目标是通过在训练过程中明确考虑提高分辨率，提出VLRR问题来克服这一问题。通过[5，15]之间的随机因子*s*对所有训练图像进行降采样，以获得高度在[7，21]像素之间的LR图像，目的是使所学习的模型在测试中能够在很低的分辨率范围内保持鲁棒性。

表现通过相同的培训和测试集，我们可以完成相同的任务，重点是提高VLRR主题的识别能力。我们遵循[ 32 ]中的数据扩充来训练所有模型。尽管SR预训练，LR-HR传输和部分耦合的网络仍然可以带来明显的收益，但有趣的是，注意到模型V几乎无法从集线器损耗中受益。这可能部分是由于样本量太小所致；但是更重要的是，训练集包含所有渲染的合成数据，而没有任何异常值，这使得在无监督训练中不需要鲁棒的损失。该模型的错误率达到36.31%@top-1和16.31%@top-5。胜过之前VFR上报告的最佳准确度数字：38.15%@top-1和20.62%@top-5 [ 32]。仔细检查发现，根据前5个结果，我们的模型正确地将52个“ VLRR图像”中的33个进行了分类。

# 8.结论

在本文中，我们使用深度网络工具研究具有挑战性的VLRR问题。从最简单的CNN基线开始，我们逐步发展我们的模型，每一步都有充分的动机和理由。除了强大的学习能力外，最终模型还受益于SR预训练，具有部分灵活性的领域自适应和鲁棒的损失。在三种不同的VLRR任务上评估了所提出模型的有效性，并获得了出色的性能。

# 参考

1. F. Bach，R。Jenatton，J。Mairal，G。Obozinski等。具有稀疏诱导规范的凸优化。*《机器学习的优化》*，第19–53页，2011年。6
2. S. Baker和T. Kanade。超分辨率的限制以及如何打破它们。*TPAMI*，24（9）：1167–1183，2002. 3
3. Y. Bengio。为AI学习深度架构。*机器学习的基础和趋势* R ，2（1）：1-127，2009. 4
4. Y. Bengio，P。Lamblin，D。Popovici和H. Larochelle。深入网络的贪婪分层训练。*NIPS*，19：153，2007。 3 ，7
5. HC Burger，CJ Schuler和S. Harmeling。图像降噪：普通神经网络可以与bm3d竞争吗？在*2012年IEEE会议的计算机视觉和模式识别（CVPR）中*，第2392–2399页。IEEE，2012年6
6. G. Chen，J。Yang，H。Jin，J。Brandt，E。Shechtman，A。Agarwala和TX Han。大规模视觉字体识别。在*CVPR中*，第3598–3605页。IEEE，2014年1 ，8
7. L. Chen，H。Qu，J。Zhao，B。Chen和JC Principe。具有肾上腺素引起的丢失功能的高效而强大的深度学习。*《神经计算与应用》*，第1-13页，2015年。6
8. C. Dong，CC Loy，K。He和X. Tang。学习深度卷积网络以获得图像超分辨率。在*ECCV中*，第184-199页。2014年2 ，3
9. D. Erhan，P.-A.。Manzagol，Y。Bengio，S。Bengio和P. Vincent。训练深层架构的难度以及无人监督的预训练的效果。在*国际人工智能与统计会议上*，第153-160页，2009年。3
10. X. Glorot，A。Bordes和Y. Bengio。针对大规模情感分类的领域适应：一种深度学习方法。在*ICML*，页513-520，2011年4 ， 5
11. M. Grgic，K。Delac和S. Grgic。Scface监控摄像机的人脸数据库。*多媒体工具与应用*，51（3）：863–879，2011. 7
12. Hennings-Yeomans，S。Baker和BV Kumar。同时进行超分辨率和特征提取，以识别低分辨率人脸。在*CVPR中*。IEEE，2008年1 ，3
13. GB Huang，M。Ramesh，T。Berg和E.Learned-Miller。带有标签的野外面孔：用于在不受限制的环境中研究面部识别的数据库。2 ，7
14. PJ Huber等人。位置参数的可靠估计。*数理统计年鉴*，35（1）：73-101，1964年6 ，7
15. C. Jacobs，PY Simard，P。Viola和J. Rinker。低分辨率文档图像的文本识别。在*ICDAR中*，第695-699页。IEEE，2005年。1
16. M. Jaderberg，A。Vedaldi和A. Zisserman。文字识别的深层功能。在*ECCV中*，第512-528页。2014年2
17. 克里热夫斯基。从微小图像中学习多层功能，2009年。2
18. A. Krizhevsky，I。Sutskever和GE Hinton。深度卷积神经网络的图像网络分类。在

*NIPS*，1097-1105页，2012年1 ，2

1. N. Kumar，AC Berg，PN Belhumeur和SK Nayar。

用于面部验证的属性和明喻分类器。在*ICCV中*，第365-372页。IEEE，2009年7

1. D. Lin和X. Tang。图像样式转换的耦合空间学习。在*ICCV中*，第1699-1706页。IEEE，2005年。4
2. Z. Lin和H.-Y.。岑。本地翻译下基于重构的超分辨率算法的基本限制。*模式分析和机器智能，IEEE*期刊，26（1）：83–97，2004年 。4
3. 吕YM，D。Bolme，BA Draper，JR Beveridge，G。Givens和PJ Phillips。人脸识别协变量的荟萃分析。在*IEEE BTAS*，第1-8页，2009年1 ， 5 ，6
4. Y. Netzer，T。Wang，A。Coates，A。Bissacco，B。Wu和AY Ng。通过无监督的特征学习来读取自然图像中的数字。在*深学习和无监督功能学习NIPS车间*，第5页，2011年2 ，7 ， 8
5. Qi。Y. Wang，Y。Wang，J。Zhang，J。Zhu，和X. Zheng。强大的深层网络，具有最大的癫痫发作检测标准。*生物医学研究的国际*，2014年，2014年6
6. A. Sapkota和T. Boult。大型无约束开放式面孔数据库。在*IEEE BTAS中*，第1–8页，2013年。7
7. S. Shekhar，VM Patel和R. Chellappa。基于合成的鲁棒低分辨率人脸识别。*IEEE TIP*，2014年1
8. X. Shu，G.-J.。Qi，T。Tang和J.Wang。异构领域知识传播的弱共享深度传输网络。在*第23届ACM多媒体会议年会记录中*，第35–44页。ACM，2015年。6
9. T. Sim，S。Baker和M. Bsat。cmu姿势，照明和表情（饼）数据库。在*FG中*。IEEE，2002年。7
10. 孙Y，王X，唐X。预测10,000个班级的深度学习人脸表示。在*CVPR中*，第1891至1898页。IEEE，2014年2
11. S. Wang，L。Zhang，LY和PanQ。半耦合字典学习及其在图像超分辨率和照片素描合成中的应用。在*CVPR中*。IEEE，2012。4 ，6
12. W. Wang，Cui Zui，H。Chang，S。Shan和X. Chen。用于交叉视图分类的深层耦合自动编码器网络。*的arXiv预印本的arXiv：1402.2031*，2014年4 ， 5
13. Z. Wang，J。Yang，H。Jin，E。Shechtman，A。Agarwala，J。Brandt和TS Huang。使用深度网络和域自适应的现实世界字体识别。*的arXiv预印本的arXiv：1504.00028*，2015年2 ，8
14. 徐升，任建勋，刘长和贾建。用于图像反卷积的深度卷积神经网络。《*神经信息处理系统的发展》*，第1790-1798页，2014年。6
15. J. Yang，J。Wright，TS TS Huang和Y. Ma。通过稀疏表示的图像超分辨率。*IEEE TIP*，19（11）：2861-2873，2010. 4
16. 邹WW和PC Yuen。分辨率极低的人脸识别问题。*TIP*，21（1）：327-340，2012。1 ，3 ，8

原文

提供更好的翻译建议

1. 这些共享特征通常是基本的对应关系，大多数情况下会在分辨率变化下持续存在，例如像素强度直方图，强边缘和特定的结构布局[22]，这构成了交叉分辨率识别的基础 [↑](#footnote-ref-0)
2. [↑](#footnote-ref-1)
3. 这里我们指的是参数复杂度。例如，训练数据从模型II到模型III翻倍，但自由模型参数的数量保持不变。因此，参数复杂度是相同的，但是训练时间肯定会增加。 [↑](#footnote-ref-2)
4. [↑](#footnote-ref-3)
5. 对于模型I，预训练仅用于重建LR图像本身。对于II型到V型，预训练仍然是SR型，但是是逐层执行的，而不是端到端的。 [↑](#footnote-ref-4)
6. [↑](#footnote-ref-5)
7. [↑](#footnote-ref-6)