多尺度残差通道注意机制下的人脸超分辨率网络

金炜, 陈莹*

(江南大学轻工过程先进控制教育部重点实验室　无锡　214122)
(chenying@jiangnan.edu.cn)

摘　要: 针对当前人脸超分辨率算法中存在效率不高和重建失真等问题, 提出一种基于多尺度残差通道注意机制的人脸超分辨率网络. 该网络采用多尺度递进形式的结构, 能够同时处理不同的上采样因子. 同时, 为了解决冗余和无效信息给网络造成的影响, 在网络的特征重建模块中引入了通道注意力机制, 并融合人脸解析信息提出一种残差通道注意块, 不仅提高了网络特征利用率还加强了人脸先验的约束力度. 与现有算法在 Helen, CelebA 和 LFW 数据集上进行的实验结果表明, 该算法无论是主观视觉质量, 还是峰值信噪比和结构相似性等客观评价指标, 都明显优于现有其他算法.

关键词: 人脸超分辨率; 人脸先验; 多尺度结构; 通道注意
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2020.17995

Multi-Scale Residual Channel Attention Network for Face Super-Resolution

Jin Wei and Chen Ying*

(*Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi*　214122)

Abstract: To address the problem of low efficiency and reconstruction distortion in current face super-resolution algorithms, a multi-scale residual channel attention network (MSRCAN) is proposed. The network can simultaneously process different upscale factors with a multi-scale progressive structure. Meanwhile, in order to reduce the impact of redundant and invalid features on the network, channel attention mechanism is introduced in the feature reconstruction module of the network, and a novel residual channel attention block is proposed based on face parsing maps, which not only improves the utilization rate of network features but also strengthens the constraints of facial priori. The proposed algorithm is compared with other algorithms in Helen, CelebA and LFW datasets. Extensive experiments show that the proposed algorithm is superior to other existing algorithms, both in subjective visual quality and objective evaluation index such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

Key words: face super-resolution; facial priors; multi-scale structure; channel attention

人脸超分辨率(super-resolution, SR)是一种特定的超分辨率技术, 是指通过低分辨率(low-resolution, LR)人脸图像重建出高分辨率(high-resolution, HR)人脸图像. 现有与人脸相关的任务, 如人脸识别、人脸对齐、表情识别和三维人脸重建等都是基于清晰的 HR 人脸数据集实现的, 在面对 LR 人脸图像时, 效果会有一定的折扣. 如在人脸识别中, 已有研究表明, 当人脸图像的分辨率降到 32×32 以下,

收稿日期: 2019-07-08; 修回日期: 2020-04-08. **基金项目**: 国家自然科学基金(61573168); 江苏省六大人才高峰资助项目(2015-WLW-004). **金炜**(1994—), 男, 硕士研究生, 主要研究方向为计算机视觉; **陈莹**(1976—), 女, 博士, 教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为计算机视觉、信息融合.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

识别率会严重下降[1]. 因此, 人脸 SR 技术在计算机视觉和生物识别领域显得尤其重要.

从 Baker 等[2]提出“虚幻脸”开始, 越来越多的研究者投入了人脸SR的研究. 早期传统的人脸SR算法在低等级特征上进行广泛的研究, Wang 等[3]提出的全局特征脸, 通过主成分分析尽可能从 LR 的人脸图像中提取信息特征进行特征转换来实现人脸 SR 重建; Liu 等[4]提出了一种两阶段的算法, 结合全局参数和局部纹理块来进行人脸 SR 重建. 然而这些传统算法通常只适应于单一环境, 难以处理不同的人脸姿势和图像分辨率, 以及图像模糊的情况. 自从深度学习被引入人脸 SR 中以来, 传统的人脸 SR 算法逐渐淡出人们的视野.

深度卷积神经网络强大的学习能力使其在人脸SR研究中得到了有效的应用. Cao 等[5]提出一种周期性的政策网络, 该框架可以反复发现人脸的各个部分, 并充分利用图像的全局相关性来增强面部不同部分的细节信息. URDGN[6]采用类似于SRGAN[7]的网络进行对抗性学习, 将生成对抗网络(generative adversarial networks, GAN)成功地应用到人脸SR领域研究中. Yu 等[8]提出一种基于条件 GAN 的框架, 将附加的面部属性信息融入网络的中间层, 从而实现对指定的属性执行人脸 SR 重建.

基于深度学习的人脸 SR 算法中, 将面部先验信息融入SR重建网络中, 是一种非常有效的算法. 与一般图像相比, 人脸图像具有明显的轮廓信息, 面部组件信息以及身份信息. 最近一些成功的研究已经证明, 人脸先验信息对人脸 SR 有着极大的帮助. Bulat 等[9]和 Yu 等[10]都引入了面部对齐网络, 通过端到端多任务学习来保证面部标定的一致性. Chen 等[11]不仅使用面部标定热图, 还将人脸解析图作为人脸 SR 过程的先验信息, 有效地提升了人脸 SR 重建的性能. 除了融合这种直观的面部结构信息和组件信息, Wu 等[12]和 Zhang 等[13]都以串联 SR 重建网络和识别网络的方式, 用身份信息约束 SR 人脸图像的合成.

目前, 结合人脸先验的人脸 SR 算法主要采用串联多任务训练的方式, 通过面部对齐网络或者是识别网络参与正式训练来约束人脸SR重建网络的生成结果, 然而这种方式需要更多的训练时间, 尤其是融合身份信息的识别网络需要庞大的人脸数据集训练. Zhang 等[13]提出的 SICNN(super-identity convolutional neural network)中, 仅仅识别网络就需要 150 万幅人脸图像训练, 训练 batch 为 512, 这不仅需要消耗大量训练时间, 还对实现算法的硬件有着很高的要求. 其次, 大多数结合面部标定和人脸解析信息的算法仅仅将提取的先验信息级联到 SR 重建网络的某一层, 忽略了先验信息对网络的全局约束, 极大地减弱了人脸先验的约束力度.

人脸 SR 属于 SR 研究中的一个特定分支, 目前大部分人脸SR重建网络以堆积残差块作为网络的主要部分. 为了提高模型泛化性, 通常会加深网络, 然而残差网络越深, 网络的冗余信息就越多, 网络的特征利用率就越低. 其次, 大多数人脸 SR 算法把不同的上采样因子(即HR图像和LR图像之间的分辨率比)的重建过程看做是独立的任务, 需要分别训练相对应的网络模型, 从而增加了冗余的工作量.

针对上述几个问题, 本文提出一种融合人脸先验信息的多尺度残差通道注意网络(multi-scale residual channel attention network, MSRCAN), 该网络不仅能有效地减少网络冗余, 而且在人脸先验信息的融合方式上也更具有约束力. 与其他算法相比, 本文算法在公开数据集上无论是主观评价指标还是客观评价指标都更为突出.

1 MSRCAN

本文融合了通道注意力(channel attention, CA)机制、人脸先验信息和多尺度递进训练的策略, 提出了 MSRCAN, 其整体网络框架如图 1 所示. MSRCAN 采用多尺度递进的方式, 由 4 个模块组成, 分别是人脸解析模块(face parsing module, FPM)、浅层特征提取模块(shallow feature extraction module, SFEM)、特征重建模块(feature reconstruction module, FRM)以及上采样模块(upscale module, USM). 在 FRM 中, 本文提出一种融合人脸解析图的残差通道注意模块(residual channel attention block, RCAB)结构替换了原始残差块结构, 不仅降低了网络特征的冗余度, 同时也增强了人脸先验的约束力.

1.1 网络总体架构

为了让 MSRCAN 适应不同的上采样因子, 本文采用一种多尺度递进形式的网络结构, 网络有 3 个输入端, 可以同时处理×2, ×4 和×8 这 3 种不同大小的上采样因子, 并分别接受 16×16, 32×32 和 64×64 这 3 种不同分辨率大小的 LR 人脸图像

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

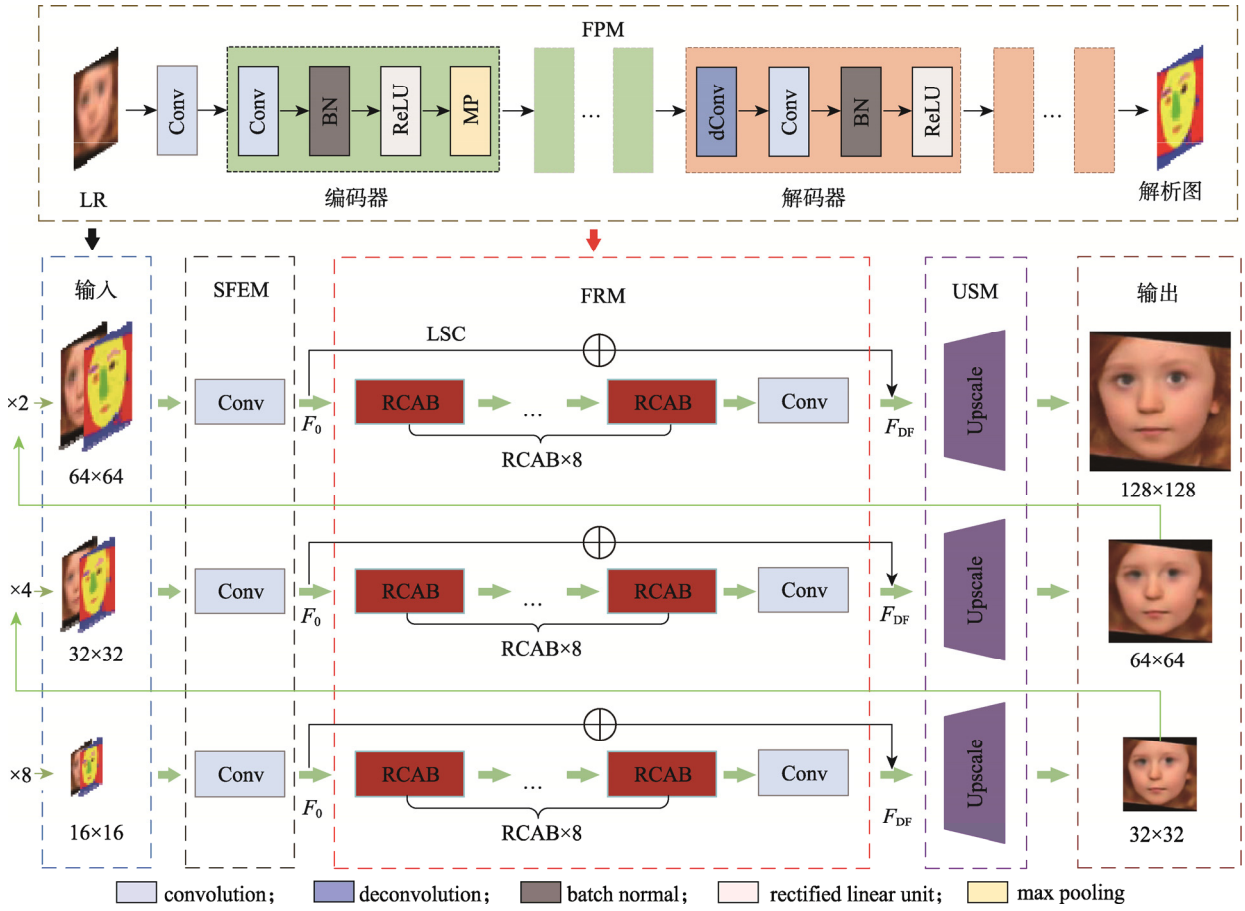


图 1　MSRCAN 结构图

$I_{\mathrm{LR}(\times 8)}$，$I_{\mathrm{LR}(\times 4)}$ 和 $I_{\mathrm{LR}(\times 2)}$. 以 ×8 的上采样因子为例，当 16×16 分辨率的 LR 图像 $I_{\mathrm{LR}(\times 8)}$ 从 ×8 的端口输入时，首先会与之相对应的人脸解析信息级联起来送入 SFEM，通过一个卷积层提取浅层特征 $F_{0(\times 8)}$，将其送入 FRM 得到重建后特征，并通过一条长跳越连接(long skip connection, LSC)与浅层特征相加得到深度特征 $F_{\mathrm{DF}(\times 8)}$；然后将 $F_{\mathrm{DF}(\times 8)}$ 送入 USM 得到 32×32 分辨率的 HR 人脸图像 $I_{\mathrm{HR}(\times 8)}$，并将 $I_{\mathrm{HR}(\times 8)}$ 当作新一轮 LR 图像 $I_{\mathrm{LR}(\times 4)}$ 送入 ×4 的端口，进行一轮相同的操作得到 64×64 分辨率的 HR 图像 $I_{\mathrm{HR}(\times 4)}$，再送入 ×2 的端口，最后完成 8 倍 SR 重建得到 128×128 分辨率的 HR 人脸图像 $I_{\mathrm{HR}(\times 2)}$. 为了同时实现 ×2，×4 和 ×8 这 3 种不同大小上采样因子的 SR 重建，网络输入会接受 3 种不同大小分辨率的 LR 人脸图像，并根据其分辨率大小分配对应的端口，所以除了 ×8 的输入口，×2 和 ×4 的输入口不仅会接受原始训练集里的 LR 图像，还会接受来自 ×8 端的输出 $I_{\mathrm{HR}(\times 8)}$ 和 ×4 端的输出 $I_{\mathrm{HR}(\times 4)}$.

1.1.1　FPM

任何真实世界的物体都有其相对应的形状和纹理分布，包括人脸. 而 LR 人脸图像即使丢失大部分面部纹理信息，面部形状信息还是被极大程度地保留下来，如果完好地提取这些形状先验信息去约束 SR 重建网络，就能保证生成的 HR 人脸图像更加贴近真实值. 相比于其他先验，人脸形状先验要更容易获取，本文采用 FPM 来提取人脸轮廓信息和不同的面部组件信息. 如图 2 所示，11 通道的人脸解析图分别代表背景、人脸轮廓、头发、左眉毛、右眉毛、左眼睛、右眼睛、鼻子、上嘴唇、口型和下嘴唇；其中人脸轮廓可以帮助恢复准确的脸型，面部组件信息可以帮助恢复出更精细的面部细节.

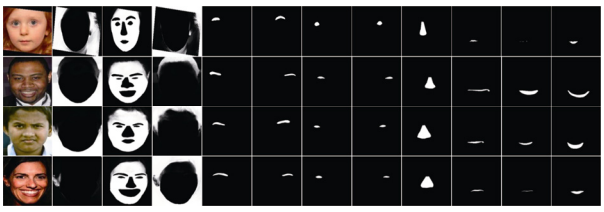
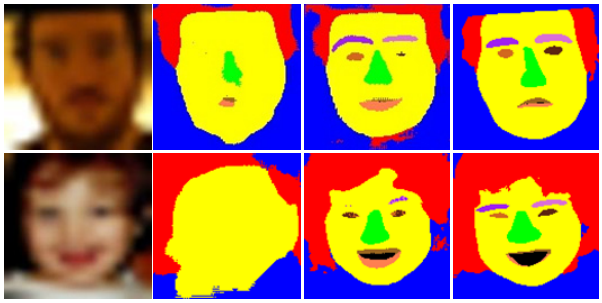


图 2　11 通道人脸解析图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

现有的人脸解析算法一般只对清晰的人脸数据集有效，并不适用于 LR 图像. 因此，在现有人脸解析网络模型[14]的基础上对其进行微调，让其能准确地捕获 LR 人脸图像的面部组件和轮廓先验；其中，网络输入是经过双三次插值(Bicubic)的 LR 人脸图像，通过类似编码解码结构的网络输出 11 通道的人脸解析图. 为了更直观地显示解析图，本文将其可视化为三通道彩色图格式.

如图 3 所示，在处理 LR 人脸图像时，经过微调后的人脸解析网络要比原始解析网络模型有着更好的性能，由清晰人脸数据集训练的原始解析模型只能捕获 LR 人脸图像的整体轮廓，微调后的模型可以精细到五官信息，其解析结果与真实解析标签相差无几.



a. Bicubic b. 原始解析图 c. 微调解析图 d. 解析标签

图 3 不同模型的人脸解析结果

1.1.2 SFEM

SFEM 由 3 个不同卷积层组成，分别对不同分辨率的 LR 人脸图像提取采样因子的浅层特征，即

$$F_{0(\times s)} = H_{\mathrm{SFEM}(\times s)}([I_{\mathrm{LR}(\times s)}, P(I_{\mathrm{LR}(\times s)})]) .$$

其中，$\times s$ 代表 ×2，×4 和 ×8 这 3 种不同的上采样因子；$H_{\mathrm{SFEM}}(\cdot)$ 表示卷积运算；[·] 表示级联操作；$I_{\mathrm{LR}(\times s)}$ 是 LR 人脸图像；$P(I_{\mathrm{LR}(\times s)})$ 是 LR 图像相对应的人脸解析图；P 表示 FPM.

1.1.3 FRM

特征重建模块是 SR 重建网络的重要组成部分，通过特征重建将低频信息重建成高频信息，而网络输出 HR 图像的质量在很大程度上取决于重建出的高频信息.

和 SFEM 类似，FRM 也有 3 个不同的尺度，分别由 8 个 RCAB 和 1 个卷积层组成. 为了加强了全局残差学习，本文在浅层特征 $F_{0(\times s)}$ 与深度特征 $F_{\mathrm{DF}(\times s)}$ 间引入一条 LSC，即

$$F_{\mathrm{DF}(\times s)} = H_{\mathrm{FRM}(\times s)}(F_{0(\times s)}) + F_{0(\times s)} .$$

其中，$F_{\mathrm{DF}(\times s)}$ 是对应上采样因子的深度特征；$H_{\mathrm{FRM}}(\cdot)$ 表示 FRM，包含 8 个 RCAB 运算和 1 个卷积运算.

1.1.4 USM

从 LR 图像到 HR 图像的过程会涉及上采样. 先前 SR 重建工作中上采样的方式是直接对 LR 图像进行 Bicubic 上采样到与 HR 图像相同的分辨率，再去学习 LR 图像与 HR 图像之前的映射函数，然而这样不仅引入了冗余的信息，还增加了计算复杂度. 近期 SR 重建工作都倾向于使用未经放大的 LR 图像作为输入来训练一个可以直接上采样到 HR 分辨率的网络，USM 通常都在网络的末端.

USM 结构如图 4 所示，主要由卷积层和亚像素卷积层又名像素洗牌(pixel shuffle，PS)[15]组成，将重建的深度特征 $F_{\mathrm{DF}(\times s)}$ 上采样成 HR 图像，即

$$I_{\mathrm{HR}(\times s)} = H_{\mathrm{USM}(\times s)}(F_{\mathrm{DF}(\times s)}) .$$

其中，$I_{\mathrm{SR}(\times s)}$ 表示不同上采样因子对应的 HR 图像；$H_{\mathrm{USM}}(\cdot)$ 表示由 Conv，PS 和 ReLU 激活函数组成的上采样运算.

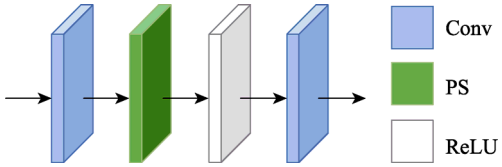


图 4 USM 结构图

1.2 RCAB

基于 CNN 的 SR 重建算法首先通过从 LR 图像中提取低频细节信息，然后通过特征重建网络恢复出其高频信息，最后生成 HR 图像. 但是泛化性能好的模型往往需要深的重建网络，从而导致网络中的冗余信息和无效信息过多，深度残差网络便是一个例子. 为了解决冗余信息和无效信息给网络造成影响，在特征重建网络的模块中引入了 CA 机制，并结合 CA 机制提出一种 RCAB 结构.

1.2.1 CA 机制

先前基于 CNN 的 SR 算法对每一层的通道特征的处理是平等的，通过可视化网络特征发现，随着网络越深，特征的冗余度就越高. 因此理论上将 CA 机制融入残差块中会降低网络特征的冗余度，从而提高网络特征利用率，CA 结构如图 5 所示.

首先经过平均池化(average pooling，AP)层将 $H \times W \times C$ 大小的输入特征 x 变为 $1 \times 1 \times C$ 大小的通道全局统计 f，即

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

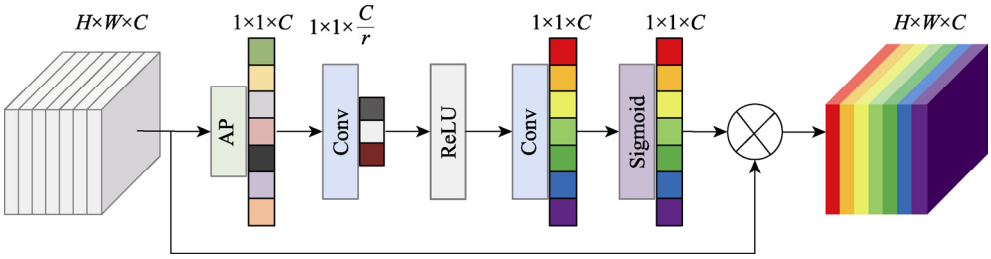


图 5　CA 机制结构图

$$f_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j).$$

其中，$x_c(i,j)$ 是 x 第 c 通道特征在 (i,j) 位置的值；f_c 是 AP 后 f 的第 c 通道的值. 除了 AP 外，还引入了 2 个 1×1 卷积来减少网络的学习参数，最后经过 Sigmoid 函数得到最终的通道统计

$$f_{\text{sig}} = S(C_{\text{up}}(\delta(C_{\text{down}}(f)))).$$

其中，$S(\cdot)$ 和 $\delta(\cdot)$ 分别表示 Sigmoid 运算和 ReLU 运算；C_{down} 是 1×1 降维卷积，其通道缩减比为 r；C_{up} 是 1×1 升维卷积，将通道还原成原始维度. 最后通过 f_{sig} 重新调节原始特征，即 $\hat{x} = f_{\text{sig}} \cdot x$. 其中，$f_{\text{sig}}$ 和 x 分别是比例因子和原始特征；$\hat{x}$ 则是最后的输出特征.

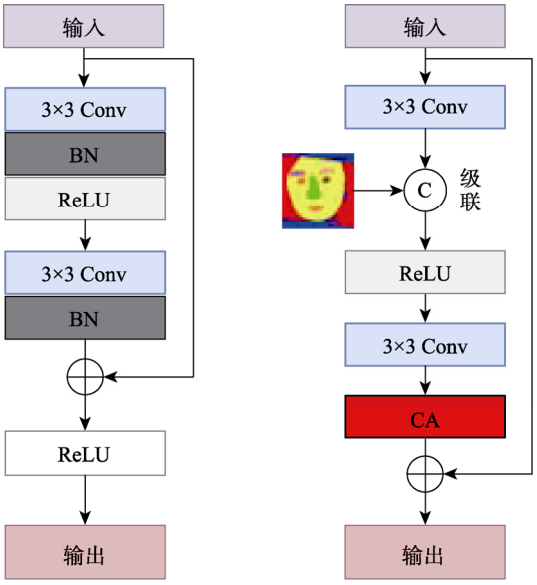
1.2.2　融合人脸先验的 RCAB

不同于原始残差块，RCAB 去除了 BN(batch normal)层. 对于 SR 算法而言，BN 层不仅会带来庞大的计算复杂度，还会降低 SR 任务的指标. 同时为了加强人脸先验的全局约束，本文将人脸解析图融入改进的残差块中，如图 6 所示，RCAB 的第 1 个卷积层输出和先验信息级联起来，作为 ReLU 激活层的输入.

本文分别将 SRResNet 和 MSRCAN 的 64 通道的深度重建特征进行可视化，并以热图形式展示在图 7 中. 从图 7 可以看出，SRResNet 深度重建特征十分稀疏(大多值都接近 0，相当于热图中的蓝色部分)，而 MSRCAN 深度重建特征的冗余无效特征就相对较少，并且在视觉上要显得更加清晰.

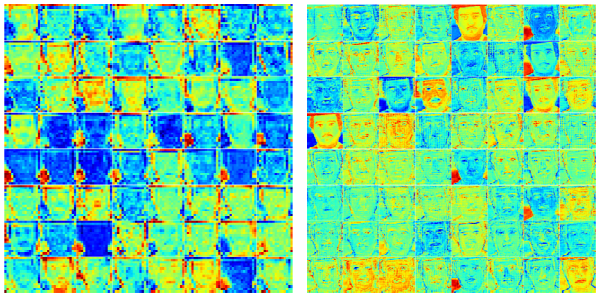
1.3　损失函数

MSRCAN 采用一种多尺度递进训练的方式，必然会产生不同尺度的 HR 图像输出，因此本文提出一种多尺度内容损失作为损失函数之一. 同时为了提升生成的 HR 人脸图像逼真度和感知相似性，对抗损失和感知损失也被引入到算法中. 整体损失组合为 $L = L_{\text{con}} + \alpha L_{\text{adv}} + \eta L_{\text{per}}$. 其中，$L_{\text{con}}$，$L_{\text{adv}}$ 和 L_{per} 分别是内容损失、对抗损失和感知损失；



a. 原始残差块　　b. RCAB

图 6　原始残差块和改进的 RCAB



a. SRResNet[7]　　b. MSRCAN

图 7　不同网络的深度特征可视化

α 和 η 是损失权重.

为了与其他算法进行公平比较，本文采取 2 种不同的损失组合：第 1 种只用内容损失训练网络模型，称为 MSRCAN；第 2 种加入了对抗损失和感知损失，称为 MSRGAN 模型.

1.3.1　多尺度内容损失

本文算法将原始真实值(ground truth, GT)下采样到 64×64 和 32×32 的分辨率，分别作为 ×4 和 ×8 端的 GT，多尺度内容损失采用 L_1 损失的形式，即

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$L_{\text{con}}=\frac{1}{m}\sum_{s=2,4,8}\sum_{i=1}^{m}\left\|I_{\text{HR}(\times s,i)}-I_{\text{GT}(\times s,i)}\right\|_1.$$

其中，m 表示批次的大小；$I_{\text{HR}(\times s,i)}$ 和 $I_{\text{GT}(\times s,i)}$ 分别表示不同尺度的第 i 幅输出 HR 图像和 GT 图像.

1.3.2 对抗损失

最近几年，很多研究者成功地证明对抗网络能够让生成的图像更加逼真. 因此，本文还以所提出的 SR 重建网络为生成网络，并增加了一个判别网络进行对抗学习，网络结构如表 1 所示. 对抗损失定义为

$$L_{\text{adv}}=\underset{I_{\text{GT}}\sim p_{\text{GT}}(I_{\text{GT}})}{E}[\ln D(I_{\text{GT}})]+\underset{I_{\text{LR}}\sim p_{\text{LR}}(I_{\text{LR}})}{E}[1-\ln D(G(I_{\text{LR}}))].$$

其中，E 表示概率分布的期望；G 是本文 SR 网络；D 是判别网络. 当输入是原始 GT 时，D 会最大化该项损失(只更新的 D 的参数)；当输入是由 G 生成的 HR 图像 $G(I_{\text{LR}})$ 时，G 会最小化该项损失(只更新 G 的参数)，从而形成对抗学习.

表 1 判别网络具体参数配置

网络层	输入大小	卷积核	步长	输出
卷积层 1	128×128×3	3×3	1	128×128×32
卷积层 2	128×128×32	3×3	2	64×64×64
卷积层 3	64×64×64	3×3	1	64×64×64
卷积层 4	64×64×64	3×3	2	32×32×128
卷积层 5	32×32×128	3×3	1	32×32×128
卷积层 6	32×32×128	3×3	2	16×16×256
卷积层 7	16×16×256	3×3	1	16×16×256
卷积层 8	16×16×256	3×3	2	8×8×512
卷积层 9	8×8×512	3×3	1	8×8×512
池化层	8×8×512			1×1×512
全连接层	1×1×512			1

1.3.3 感知损失

L_1 损失和 L_2 损失是 SR 工作中常见的损失，这种像素级别的损失虽然能够带来极高的峰值信噪比(peak signal-to-noise ratio, PSNR)，却很难恢复丢失的高频内容，使生成的图像过于平滑. 因此，本文采用了感知损失[16]，它在高维特征空间约束生成图像与真实图像的差异，有助于高频信息的恢复，使生成的图像有着较高的质量. 感知损失定义为 $L_{\text{per}}=\sum_l\left\|\phi_l(I_{\text{HR}})-\phi_l(I_{\text{GT}})\right\|_2^2$. 其中，$\phi_l(\cdot)$ 是损失网络 ϕ 的第 l 层的特征. 本文算法选取 VGG-FACE[17] 网络的 Pool2 和 Pool5 层计算感知损失.

2 实验结果与分析

2.1 数据集与训练设置

本文分别在 Helen[18]，CelebA[19]和 LFW[20]这 3 个公开人脸数据集上进行了实验. Helen 数据集是一个相对较小的人脸数据集，它由 2330 幅人脸图像组成，并且有相对应的 11 通道的解析标签. 本文遵循 FSRNet[11]算法，选取前 2280 幅人脸图像作为训练集，剩下 50 幅用来测试. CelebA 数据集是大规模人脸数据集，它由 10177 个不同身份的 202599 幅人脸图像组成. 本文选取前 18000 幅图像作为训练集，接下来的 100 幅作为测试集. LFW 数据集由 5749 个不同身份的 13233 幅人脸图像组成，所有人脸图像均采集于野外非受控环境. 本文遵循 LFW 数据集的官方协议，选取 9526 幅人脸图像作为训练集，3707 幅人脸图像作为测试集.

本文用 Helen 数据集中的 2000 幅训练集图像对人脸解析模块进行微调，将经过下采样的 LR 图像作为网络的输入，相对应的 11 通道的解析图作为标签. 学习率设为 10^{-6}，总共训练 30 个 epoch.

在训练 SR 重建网络时，本文采用 Bicubic 来下采样原始 128×128 人脸图像，生成 16×16，32×32 和 64×64 这 3 种不同分辨率大小的 LR 人脸图像作为网络的 3 种输入. 为了防止过拟合，本文对训练数据随机翻转和旋转 90°, 180° 和 270° 作为数据增强. 学习率在前 150 个 epoch 设置为 10^{-4}，后 150 个 epoch 学习率会线性的衰减到 0. 网络模型训练 batch 为 16；并根据经验设置 $\alpha=10^{-3}$，$\eta=10^{-5}$. 本次实验中使用 TensorFlow[21]深度学习工具，并采用 Adam 优化器[22]训练网络模型. 本文算法在 Helen 数据集上训练一个 MSRACN 基础模型需要在一块 GTX 1080 Ti GPU 上消耗 2.5 h.

2.2 评价指标

目前绝大多数的图像 SR 算法采用 PSNR 值和结构相似性(structural similarity index, SSIM)值作为评价指标.

PSNR 是基于像素误差敏感的图像质量评价，它非常简单地通过均方误差(mean square error, MSE)

$$\text{MSE}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(\boldsymbol{X}(i,j)-\boldsymbol{Y}(i,j))^2$$

来定义，即 $\text{PSNR}=10\lg\frac{(2^n-1)^2}{\text{MSE}}$. 其中，MSE 表示当前图像 $\boldsymbol{X}$ 和参考图像 $\boldsymbol{Y}$ 的均方误差；H 和 W

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分别为图像的高度和宽度；n 为每像素的比特数，一般的灰度图像为 n=8，彩色图像 n=24. PSNR 的数值越大，表示图像失真越小.

SSIM 模拟了人眼视觉系统对图像结构信息的敏感性，是一种衡量 2 幅图像相似度的指标，其定义为

$$\mathrm{SSIM}=\frac{(2\mu_f\mu_{\hat{f}}+C_1)(2\sigma_{f,\hat{f}}+C_2)}{(\mu_f^2+\mu_{\hat{f}}^2+C_1)(\sigma_f^2+\sigma_{\hat{f}}^2+C_2)}.$$

其中，μ_f 和 $\mu_{\hat{f}}$ 分别表示真值图像和待评价图像的灰度平均值；σ_f 和 $\sigma_{\hat{f}}$ 分别表示真值图像和待评价图像的方差；$\sigma_{f,\hat{f}}$ 表示真值图像和待评价图像之间的协方差；C_1 和 C_2 是一个比较小的数，为了防止分母为 0. SSIM 取值范围为[0,1]，其值越大，表示图像失真越小.

为了验证生成的 SR 人脸的轮廓与面部组件位置的准确性，本文同时也引入了 F-score，即

$$F=\frac{(1+\beta^2)PR}{\beta^2P+R}.$$

其中，β 是平衡精确率 P 和召回率 R 的非负参数，本文设置 $\beta^2=2$. F-score 反映了不同面部组件位置的预估准确性，值越大代表 SR 人脸的面部组件位置与真实值越接近.

为了体现人脸 SR 能够有效地提升 LR 人脸图像识别的性能，本文还引入了特征相似距离(feature similarity distance, FSD)作为评价指标. 本文选择现有公开人脸识别模型 FaceNet[23]计算 SR 结果与 GT 间的平均 FSD，距离值越低，则代表越容易识别成功，与 GT 越接近.

2.3　消融研究

为了证明本文算法中多尺度递进网络和人脸先验及其融合方式的有效性，本文用 5 个的不同的模型分别在 Helen 和 CelebA 人脸数据集测试了其 PSNR 值和 SSIM 值，结果如表 2 所示.

表 2　不同模型在 Helen 和 CelebA 数据集上的 PSNR/SSIM 指标

数据集	尺度	SRResNet[7]		RCAN		MSRCAN−−		MSRCAN−		MSRCAN+	
		PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
Helen	×2	37.26	0.9753	37.65	0.9775	37.60	0.9740	37.85	0.9771	**38.06**	**0.9780**
	×4	30.41	0.8945	30.71	0.9021	31.05	0.9092	31.30	0.9161	**31.93**	**0.9273**
	×8	25.30	0.7297	25.54	0.7359	25.96	0.7577	26.19	0.7648	**26.83**	**0.7872**
CelebA	×2	37.62	0.9726	37.75	0.9730	37.80	0.9729	37.92	0.9732	**38.24**	**0.9776**
	×4	31.04	0.9014	31.22	0.9026	31.51	0.9154	32.02	0.9180	**32.66**	**0.9274**
	×8	25.82	0.7369	25.94	0.7451	26.21	0.7526	26.54	0.7597	**27.12**	**0.7787**

在表 2 中，第 1 个模型是采用原始残差块的 SRResNet[7]；第 2 个 RCAN 模型是在 SRResNet 基础上将原始残差块替换成未结合人脸先验的 RCAB；第 3 个 MSRCAN−−模型则是基于原始 RCAB 的多尺度递进模型，并没有结合人脸先验信息；第 4 个 MSRCAN−模型是在多尺度递进模型的基础上只在输入端融合了人脸先验信息；最后一个模型是 MSRCAN+，与 MSRCAN−不同之处是其将人脸先验信息融入每个 RCAB 中.

从表 2 可以发现，对于×2，×4 和×8 的 SR，与 SRResNet 相比，融入 CA 机制的 RCAN 模型指标有一定的提升. 而采用多尺度递进形式的 MSRCAN−−模型的优越性更加体现在×4 和×8 的 SR 任务上，虽然在×2 的 SR 上指标与 RCAN 差别不大，但这种多尺度递进形式的网络决定了×2 的 SR 只占用整体网络的三分之一. 与 MSRCAN−−模型相比，结合了人脸先验信息的 MSRCAN−模型在 3 种上采样因子的 SR 任务上的性能都有着明显的提升. 而 MSRCAN+加强了人脸解析的全局约束，在 Helen 和 CelebA 数据集上，其 PSNR 和 SSIM 指标均达到最高.

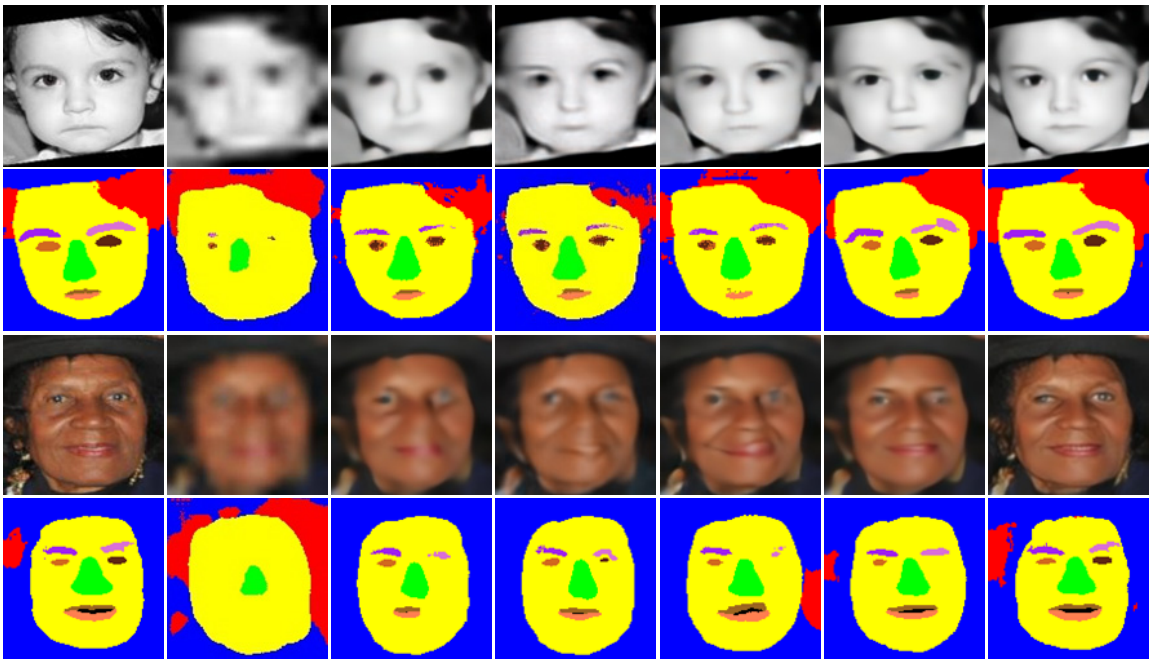
不同上采样因子的 SR 模型学习的难易程度是不同的，因此在 Helen 数据集上增加以一组额外的实验，用不同的残差块数目的 SRResNet 分别对×2，×4 和×8 的 SR 进行测试，结果如表 3 所示. 对于×8 的 SR 来说，一定程度上加深网络会显著地提升性能；而对于×2 和×4 的 SR，8 个残差块和 16 个残差块的指标变化就相对细微. 本文的 MSRCAN+这种多尺度递进的网络在×2，×4 和×8 的 SR 上的深度逐渐递增，最大化地节约了网络参数量.

表 3　不同数目残差块的 SRResNet 的 PSNR 指标 dB

尺度	残差块数目		
	4	8	16
×2	37.14	37.27	37.26
×4	30.19	30.35	30.41
×8	24.85	25.01	25.30

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

为了更直观地体现人脸先验的作用，本文分别测试不同模型在Helen数据集×8 SR的输出结果，并展示其解析图的可视化结果，如图8所示．同时，用原始人脸解析网络评估了 MSRCAN−−，MSRCAN−和 MSRCAN+模型的每个面部组件的 F 值，如表4所示．



a. GT b. Bicubic c. SRResNet[7] d. RCAN e. MSRCAN−− f. MSRCAN− g. MSRCAN+

图 8 不同模型在 Helen 数据集上的×8 SR 及人脸解析结果

表 4 不同模型在 10 个面部组件上的 F 值

面部组件	MSRCAN−−	MSRCAN−	MSRCAN+
面部轮廓	0.828	0.831	0.827
左眉毛	0.429	0.453	0.538
右眉毛	0.391	0.417	0.478
左眼睛	0.555	0.573	0.573
右眼睛	0.528	0.569	0.572
鼻子	0.893	0.898	0.895
上嘴唇	0.576	0.619	0.638
下嘴唇	0.389	0.434	0.465
口型	0.693	0.706	0.719
头发	0.637	0.643	0.647
平均值	0.592	0.614	0.635

从图8中可以看出，加入人脸先验前，MSRCAN−−模型的人脸解析可能会生成错误的形状甚至丢失组件(如眼睛和上嘴唇)；结合人脸先验 MSRCAN−模型的解析结果基本上不会发生丢失面部组件的情况，但面部组件的形状与真实值有部分差异；而 MSRCAN+模型在 RCAB 中融合了人脸先验，加强了人脸先验的约束力度，解析结果与真实解析标签几乎一致．

表4则从 F-score 指标上进一步证明了上述结论，MSRCA−融入了人脸先验，平均 F-score 比 MSRCAN−−高 0.022；MSRCAN+加强了人脸先验的全局约束，平均 F-score 达到最高．上下文出现的 MSRCAN 皆指 MSRCAN+.

2.4 与其他先进算法的比较

为了让本文算法更具有说服力，选取6种近年来所提出的 SR 算法进行比较，包括 SRResNet[7]，RDN[24]，GLN[25]，URDGN[6]，AAFH[5]和 FSRNet[11]算法．其中，SRResNet 和 RDN 属于普通 SR 研究中代表性算法．SRResNet 加深了重建网络的深度，并通过残差学习提升了高频特征．RDN 在残差块中融合了密集连接，使网络每一层的梯度都能直接传到最低层，从而加强了梯度传播，进一步提升了 SR 的性能．GLN，URDGN，AAFH 和 FSRNet 属于人脸 SR 重建算法．GLN 先通过全局网络来约束整体轮廓，再用局部网络来约束面部细节，从而得到最终 SR 结果．URDGN 采用对抗学习的方式，利用判别网络的反馈使上采样的人脸图像更接近真实的人脸图像．AAFH 利用周期性的政策网络不断发现人脸组件，并用局部增强网络提升人脸组件块．FSRNet 将人脸先验融入网络的中间层，采用多任务的方式训练网络．上述算法都不能用统一的模

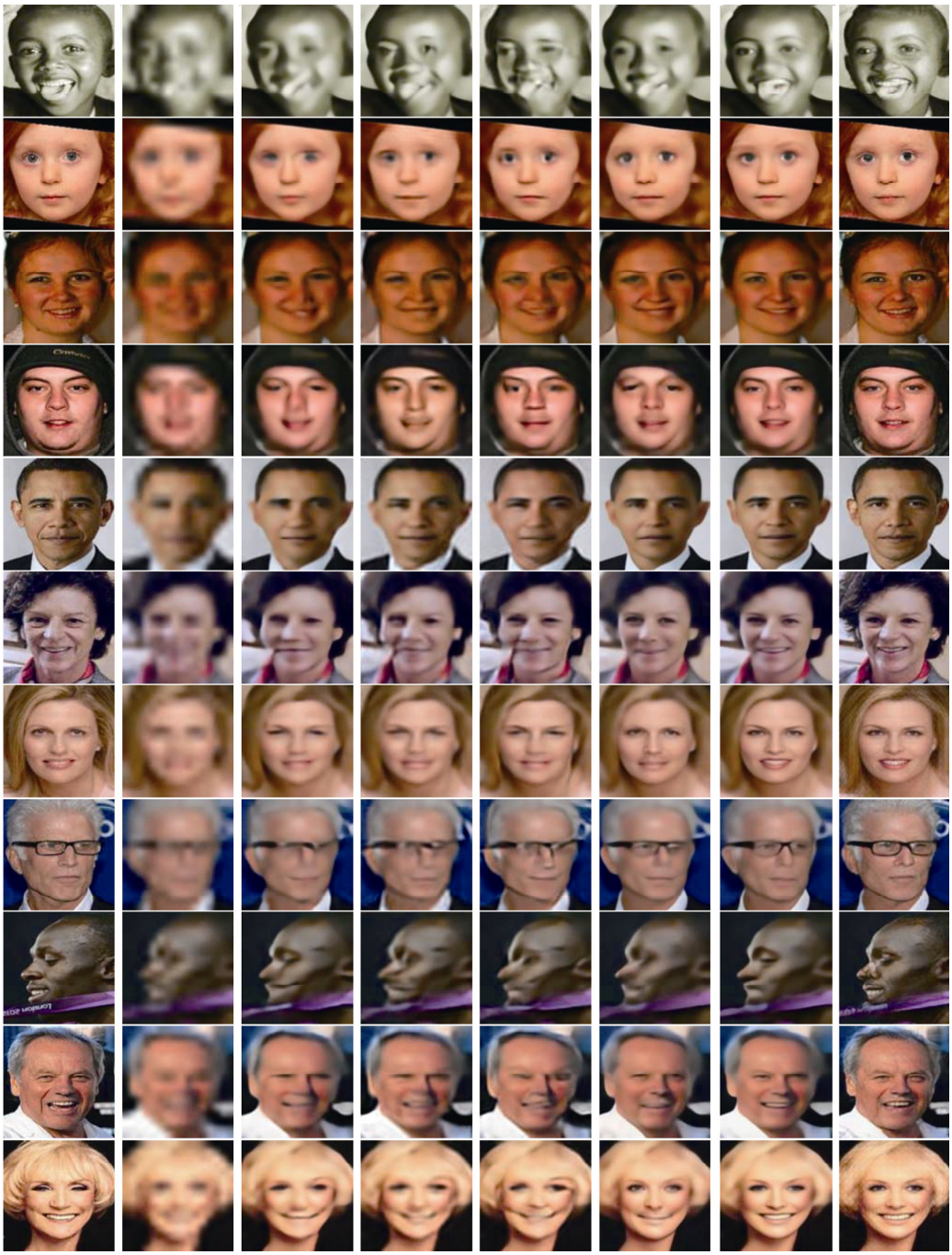
(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

型实现不同上采样因子的 SR.

为了公平比较，对于公开的 SR 算法，使用上述模型发布的代码，用相同的训练集训练所有模型；对于非公开的 SR 算法，本文算法则遵循其论文里介绍训练测试设置进行比较. 所有对比算法的结果均由原文献作者公开网页提供或者源代码生成.

2.4.1　主观视觉质量评估

为了更直观地体现出不同算法的差异，本文在图 9 中展示了不同算法 ×8 SR 的视觉对比结果.



a. GT　b. Bicubic　c. SRResNet[7]　d. RDN[24]　e. URDGN[6]　f. FSRNet[11]　g. MSRCAN　h. MSRGAN

图 9　不同算法在 Helen 和 CelebA 数据集上 ×8 SR 的视觉效果对比

与其他算法相比，MSRCAN 加强了人脸先验的全局约束，产生了相对清晰的边缘和形状，一些五官细节信息如眼睛、鼻子和嘴巴等都能够很好地重建出来，视觉结果都更加接近 GT.

从图 9 可以看出，像素级别的 L_1 或者 L_2 损失过多地关注低频信息间的差异，往往会造成过于

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

平滑的 SR 结果. 而 MSRGAN 在 MSRCAN 的基础上加入了对抗损失和感知损失, 能够更好地恢复出人脸的高频细节, 让生成的 SR 人脸结果更加逼真. 例如, 人物的头发的纹理和面部的细节都与 GT 都更加接近, 更加符合人类的视觉感官.

2.4.2 客观质量评估

表 5 显示了本文算法在 Helen, CelebA 和 LFW 数据集上的性能, 并与其他先进的算法进行了比较. 本文算法在×8 SR 上的 PSNR, SSIM 和 FSD 指标明显优于对比算法. 以 PSNR 指标为例, MSRCAN 在 3 个数据集上分别超过了第 2 算法 0.62 dB, 0.52 dB 和 0.97 dB; 在 LFW 数据集×4 SR 的 PSNR 指标上, MSRCAN 超过第 2 算法 AAFH 0.75 dB; MSRCAN 在 Helen 和 CelebA 数据集×2 SR 的指标上与 RDN 算法相比提升不大, PSNR 指标分别提升 0.14 dB 和 0.07 dB, 但在 LFW 数据集提升比较明显.

绝大多数人脸识别模型往往都无法处理 LR 人脸图像. 为了进一步证明本文人脸 SR 重建算法的有效性, 用现有公开人脸识别模型 FaceNet[23]分别计算不同模型的×2 ,×4 和×8 SR 的结果与 GT 间的平均 FSD. 从表 5 可以看出, 在 3 种不同的上采样因子上, MSRCAN 均超过其他算法, 以 CelebA 数据集为例, 平均 FSD 分别是 0.119, 0.376 和 0.656; 而结合对抗损失和感知损失的 MSRGAN 更注重恢复人脸图像的高频细节, 在 3 个数据集上取得最优的 FSD, 但同时也造成了相对较低的 PSNR 和 SSIM 指标.

表 5 不同算法在 Helen 和 CelebA 和 LFW 数据集上的 PSNR/SSIM 指标

算法	尺度	Helen			CelebA			LFW		
		PSNR/dB	SSIM	FSD	PSNR/dB	SSIM	FSD	PSNR/dB	SSIM	FSD
Bicubic		33.96	0.9476	0.223	33.58	0.9436	0.246	34.02	0.9593	0.134
SRResNET[7]		37.26	0.9753	0.118	37.62	0.9726	0.131	38.00	0.9729	0.079
RDN[24]	×2	37.92	**0.9790**	0.114	38.33	0.9763	0.123	38.59	0.9781	0.070
MSRCAN		**38.06**	0.9780	0.109	**38.40**	**0.9776**	0.119	**38.98**	**0.9821**	0.067
MSRGAN		37.45	0.9524	**0.093**	38.01	0.9617	**0.106**	38.04	0.9674	**0.055**
Bicubic		26.87	0.8191	0.675	27.40	0.8092	0.743	26.79	0.8469	0.531
SRResNET[7]		30.41	0.8945	0.393	31.04	0.9014	0.428	31.45	0.8960	0.336
RDN[24]		30.84	0.9064	0.382	31.42	0.9018	0.406	32.33	0.9084	0.316
GLN[25]	×4							30.34	0.8922	0.353
AAFH[5]								32.93	0.9104	0.312
MSRCAN		**31.93**	**0.9273**	0.351	**32.66**	**0.9274**	0.376	**33.68**	**0.9298**	0.277
MSRGAN		30.27	0.8835	**0.344**	30.66	0.8763	**0.352**	31.41	0.8917	**0.249**
Bicubic		22.37	0.6281	1.113	22.88	0.6162	1.191	21.92	0.6712	1.043
SRResNET[7]		25.30	0.7297	0.867	25.82	0.7369	0.975	25.19	0.7412	0.748
RDN[24]		25.49	0.7436	0.757	25.96	0.7482	0.824	25.50	0.7553	0.718
GLN[25]		24.11	0.6922	0.891	24.55	0.6867	0.947	24.51	0.7031	0.832
URDGN[6]	×8	24.22	0.6909	0.875	24.63	0.6851	0.955			
FSRNet[11]		26.21	0.7720	0.604	26.60	0.7628	0.745			
AAFH[5]								26.17	0.7604	0.709
MSRCAN		**26.83**	**0.7872**	0.554	**27.12**	**0.7787**	0.656	**27.04**	**0.7830**	0.591
MSRGAN		25.04	0.7286	**0.519**	25.50	0.7215	**0.582**	25.34	0.7411	**0.508**

注. 粗体为最优值.

2.5 速度比较

表 6 所示为不同算法在一幅 16×16 LR 人脸图像上进行×8 SR 所花费的时间比较. AAFH 和 FSRNet 算法所消耗时间是由原文献所记录在 Titan X GPU 上测试的时间, 其他 SR 模型均在 GTX 1080 Ti GPU 上进行测试. 从表 6 中可以看出, URDGN 算法的速度虽然最快, 但同时在 CelebA 数据集上的×8 SR 的 PSNR 指标却是最低, 比本文算法低 2.94 dB. 与

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

FSRNet 算法相比，本文算法引入 CA 机制，每幅图像的处理速度相对要慢 6 ms，但在 CelebA 数据集上 ×8 SR 的 PSNR 指标却高于 FSRNet 算法 0.52 dB.

表 6　不同算法在不同数据集耗时与 PSNR 对比

算法	耗时/ms	PSNR/dB	
		CelebA	LFW
SRResNet[7]	24	25.82	25.19
RDN[24]	38	25.96	25.50
URDGN[6]	**9**	24.63	
AAFH[5]	81		26.17
FSRNet[11]	12	26.60	
MSRCAN	18	**27.12**	**27.04**

注. 粗体为最优值.

3　结　语

本文提出一种结合人脸先验信息的 MSRCAN. 该网络通过多尺度递进的训练方式能同时处理 3 种不同上采样因子的超分辨率任务，最大化节省了网络参数及训练时间，并将人脸先验信息与 CA 机制融入残差块，不仅提高了网络特征利用率，还加强了人脸先验的全局约束. 与当前 SR 算法相比，本文算法能产生更清晰的边缘和形状，并在 ×2 ,×4 和 ×8 的 SR 指标上均超过当前人脸 SR 算法. 在速度上，虽然本文算法不能取得最优，但综合考虑指标因素，其速度还是相对较快的. 然而本文算法只能处理偶数倍上采样因子，未来可以进一步改进模型，让其可以同时处理奇数倍上采样因子.

参考文献(References):

[1] Torralba A, Fergus R, Freeman W T. 80 million tiny images: a large data set for nonparametric object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958-1970

[2] Baker S, Kanade T. Hallucinating faces[C] //Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2000: 83-88

[3] Wang X G, Tang X O. Hallucinating face by eigen transformation[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2005, 35(3): 425-434

[4] Liu C, Shum H Y, Zhang C S. A two-step approach to hallucinating faces: global parametric model and local nonparametric model[C] //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2001: 192-198

[5] Cao Q X, Lin L, Shi Y K, *et al*. Attention-aware face hallucination via deep reinforcement learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 690-698

[6] Yu X, Porikli F. Ultra-resolving face images by discriminative generative networks[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 318-333

[7] 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. Los Alamitos: IEEE Computer Society Press, 2017: 105-114

[8] Yu X, Fernando B, Hartley R, *et al*. Super-resolving very low-resolution face images with supplementary attributes[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 908-917

[9] 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/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 109-117

[10] Yu X, Fernando B, Ghanem B, *et al*. Face super-resolution guided by facial component heatmaps[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 217-233

[11] Chen Y, Tai Y, Liu X M, *et al*. FSRNet: End-to-end learning face super-resolution with facial priors[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2492-2501

[12] Wu J Y, Ding S Y, Xu W, *et al*. Deep joint face hallucination and recognition[OL]. [2019-07-08]. https://arxiv.org/abs/1611.08091.pdf

[13] Zhang K P, Zhang Z P, Cheng C W, *et al*. Super-identity convolutional neural network for face hallucination[C] // Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 183-198

[14] Liu S F, Yang J M, Huang C, *et al*. Multi-objective convolutional learning for face labeling[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3451-3459

[15] Shi W Z, Caballero J, Huszár F, *et al*. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1874-1883

[16] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 694-711

[17] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C] //Proceedings of the British Machine Vision Conference. Guildford: BMVA Press, 2015: 41.1-41.12

[18] Le V, Brandt J, Lin Z, *et al*. Interactive facial feature localization[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2012: 679-692

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[19] Liu Z W, Luo P, Wang X G, *et al*. Deep learning face attributes in the wild[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 3730-3738

[20] Huang G B, Ramesh M, Berg T, *et al*. Labeled faces in the wild: a database forstudying face recognition in unconstrained environments[R]. Amherst: University of Massachusetts. Technical Report 07-49, 2007

[21] Abadi M, Barham P, Chen J M, *et al*. TensorFlow: a system for large-scale machine learning[C] //Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Berkeley: USENIX Association, 2016: 265-283

[22] Kingma D P, Ba J. Adam: a method for stochastic optimization[OL]. [2019-07-08]. https://arxiv.org/abs/1412.6980.pdf

[23] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 815-823

[24] Zhang Y L, Tian Y P, Kong Y, *et al*. Residual dense network for image super-resolution[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2472-2481

[25] Tuzel O, Taguchi Y, Hershey J R. Global-local face upsampling network[OL]. [2019-07-08]. https://arxiv.org/abs/1603.07235.pdf

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net