第 59 卷 第 4 期/2022 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

融合自注意力机制的人物姿态迁移生成模型

赵宁，刘立波*

宁夏大学信息工程学院，宁夏 银川 750021

摘要 针对人物姿态迁移生成图像存在纹理细节丢失、姿态转移不合理等问题，提出一种融合自注意力机制的人物姿态迁移生成模型。首先，在两阶段姿态迁移生成模型的基础上，通过把改进的自注意力模块嵌入到生成对抗网络中，降低相似特征间的相互影响，强化对纹理细节的学习能力及丰富信息的捕获能力，增强姿态特征的显著性建模。然后，使用马尔可夫判别模型，进一步增强对生成图像细节的鉴别能力。最后，采用优化的内容损失函数，约束整个模型的图像特征信息损失计算，促进生成图像与真实图像语义内容一致性，加强姿态转移的合理性。实验验证，本模型在Deepfashion数据集上比PG^{2}方法的IS值与SSIM值分别提升了0.388和0.032，在Market-1501数据集上比PG^{2}方法的IS值与SSIM值分别提升了0.036和0.065，改善了图像生成质量。

关键词 图像处理；深度学习；生成对抗网络；图像生成；自注意力机制

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202259.0410014

Generation Model of Character Posture TransferBased on Self-attention Mechanism

Zhao Ning, Liu Libo*

School of Information Engineering, Ningxia University, Yinchuan, Ningxia 750021, *China*

Abstract This paper proposes a character pose transfer generative model fused with the self-attention mechanism to address the issues of loss of texture details and unreasonable pose transfer in images generated by character pose transfer. First, based on the two-stage pose transfer generative model, the improved self-attention module is introduced into the generative adversarial network to reduce the interaction between similar features, which improves the ability to learn texture details and capture information, enhances saliency modeling of posture features. Then, the Markov discriminative model is used to enhance the ability to discriminate the details of the generated image. Finally, the optimized content loss function is used to constrain the image feature information loss of the entire model, promote semantic consistency between the generated and the real images, and strengthen the rationality of pose transfer. The experimental results demonstrate that, compared with the PG^2 method on the DeepFashion and Market-1501 datasets, the IS and SSIM values of our model has increased in 0.388 and 0.032, 0.036 and 0.065, respectively.

Key words image processing; deep learning; generative adversarial net; image generation; self attention mechanism

1 引 言

人物姿态迁移是指将输入的人物图像从当前姿势转换到目标姿势的图像生成任务，该任务可为行人重识别、视频生成等领域提供基础数据，具有较高的应用价值，成为图像生成领域的研究热点之一[1]。

主流的人物姿态迁移方法主要基于深度生成模型，其中生成对抗网络(GANs)[2]是最常用的方法，

收稿日期：2021-03-10；**修回日期**：2021-03-26；**录用日期**：2021-04-02

基金项目：国家自然科学基金(61862050)、宁夏自然科学基金(2020AAC03031)

通信作者：*liulib@163.com

目前，已有大量工作通过与GANs结合来实现人物姿态迁移。早期Isola等[3]提出一种基于U-net结构的图像翻译框架，该方法是生成对抗网络的变形，可以应用在人体姿态迁移中将输入图像转换为目标图像，但其假设输入和输出图像信息对齐，当目标图像前景空间相对输入图像产生较大变形时，将很难处理不一致姿态区域的迁移。Ma等[4]提出利用两阶段生成方法（PG^2）缓解这一问题，在第一阶段生成网络中训练U-net生成器，产生具有目标姿态的中间图像，在第二阶段通过生成对抗训练，输出令中间图像更接近目标图像的外观差异图，该方法可以合成任意姿态人物图像，但生成图像的服装细节纹理容易丢失。Huang等[5]在PG^{2}模型基础上引入特征信息反馈机制，增强生成网络对特征的学习能力。Zhu等[1]提出基于条件GAN的模型，利用叠加卷积层来形成特征的注意力图，通过注意力图获取图像外观特征，模型包含一系列姿态转移块，每个块关注特定区域，逐步生成所需图像，该方法利用了注意力机制思想，能更好捕捉原始输入图像小区域的细节特征。文献[6]利用两个引入注意力机制的分支交叉传递信息，引导形状和外观生成，分支互相促进，通过融合逐步生成图像。当前的人物姿态迁移生成方法研究尽管有了一定的进展，但是仍存在生成图像纹理细节丢失和姿态转移不合理的问题。

综上所述，本文在原有两阶段生成框架PG^{2}模型的基础上，提出一种融合自注意力机制的人物姿态迁移生成模型，主要方法如下：1）在外观细化阶段生成对抗网络中，引入改进后的自注意力机制，增强对像素间关系的捕获能力及相似像素的学习能力，改善多层卷积难以捕捉图像全局特征的问题，加强图像像素远距离建模能力，进而学习到更加细粒度级别的纹理细节信息；2）将判别网络替换成马尔可夫判别模型，通过对图像感受野进行真假判断，提高对高频信息的学习，强化图像纹理风格训练；3）为增强模型对图像内容合理性和真实性的学习能力，引入VGG-19预训练模型，在特征层面进行信息损失计算，建立内容损失函数，增强生成图像内容的语义一致性，进一步提升姿态转移准确性。

在Market-1501数据集和DeepFashion数据集上进行定量与定性分析，并与当前主流方法进行对比实验，验证了模型生成图像符合人类视觉感知，生成质量比主流方法有较高的提升。

2　模型分析

采用两阶段姿态迁移生成模型PG^{2}作为基本框架，通过改进自注意力机制，引入自注意力机制的生成对抗网络和优化损失函数，提出一种融合自注意力机制的人物姿态迁移生成算法。

2.1　两阶段姿态迁移生成模型

PG^{2}首先提出采用姿态转移阶段（G1）和纹理细化阶段（G2）两个阶段生成模型来解决姿态迁移问题[4]。在姿态转移阶段，生成结果是一张学习到目标姿势的粗糙图像，该中间图像包含大部分目标图像的姿态特征和少量的纹理信息。在纹理细化阶段通过生成对抗训练细化图像，使模型在上一步生成结果基础上，着重学习图像缺失的外观细节，PG^{2}具体流程如图1所示。

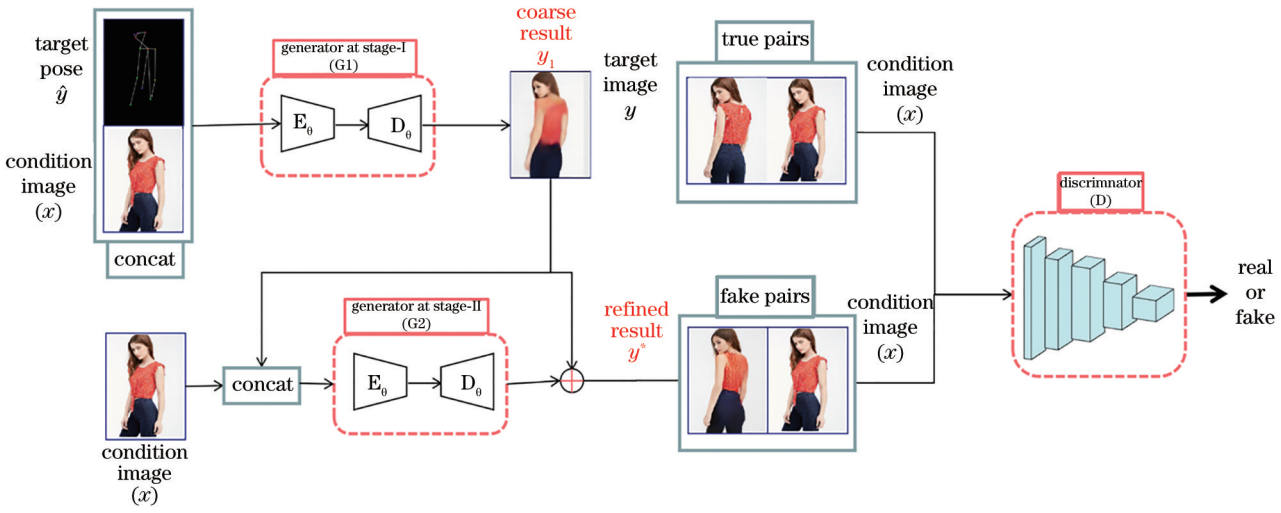

图1　两阶段人物姿态迁移生成模型

Fig. 1　Two-stage character pose transfer image generation model

姿态转移阶段将成对条件图像 $\boldsymbol{x}$ 和目标姿势 $\hat{\boldsymbol{y}}$ 作为输入，生成该姿态下的粗糙图像 $\boldsymbol{y}_1$，$\hat{\boldsymbol{y}}$ 由目标图像 $\boldsymbol{y}$ 通过现有先进的姿态提取模型[7]提取关键点后得到。$\boldsymbol{y}_1$ 和条件图像 $\boldsymbol{x}$ 共同构成输入送入外观细化网络，先经过生成网络 G2 得到细节更丰富的生成图像 $\boldsymbol{y}^*$，再利用判别网络(D)鉴别生成图像的真伪，由于判别器输入为成对图像而非噪声，在判别器中生成结果 $\boldsymbol{y}^*$ 与条件图像 $\boldsymbol{x}$ 组成假图像对，目标图像 $\boldsymbol{y}$ 和条件图像 $\boldsymbol{x}$ 构成真图像对，以保证生成图像在 $\boldsymbol{y}_1$ 结果上进行细化，而非从零生成一张新的图像。

2.2　自注意力机制的改进

原始的 GAN 的感受野取决于卷积核大小，在学习和对图像建模时难以捕获图像全局内容，会丢失内部信息以及远距离像素间关联，而自注意力机制可以较好地解决这一问题，SAGAN[8]首先将自注意力机制引入生成对抗网络，并在生成器和判别器中分别添加自注意力模块，极大地提升了模型生成质量。该自注意力机制基于 non-local 思想，引入图像领域时，对像素内部关联与像素边界信息的学习在训练中相互干扰，阻碍对纹理细节的建模[9]。针对此问题，采用改进后的自注意力机制降低相似像素的冗余影响[10]，增加像素的显著性建模，进一步强化对特征丰富性与显著性的学习，提升注意力特征图对整体图像建模能力，具体改进如下：

1) 为了学习到同类区域内的像素关联，降低相似像素间的冗余影响，使特征在局部区域内关联性更高，生成的图像细节更加充实，进一步增强姿态与纹理细节的贴合度，特征图增加白化操作，对捕捉到的像素进行白化计算，可表示为

$$\alpha_{j,i}=\sigma\left\{\left[\boldsymbol{f}\left(\boldsymbol{u}_i\right)-\boldsymbol{\mu}_{\mathrm{f}}\right]^{\mathrm{T}}\left[\boldsymbol{g}\left(\boldsymbol{u}_j\right)-\boldsymbol{\mu}_{\mathrm{g}}\right]\right\},\tag{1}$$

式中，$\boldsymbol{f}\left(\boldsymbol{u}_i\right)$ 是 i 像素的特征值，$\boldsymbol{\mu}_{\mathrm{f}}$ 是普遍像素特征的均值，$\boldsymbol{g}\left(\boldsymbol{u}_j\right)$ 是 j 像素特征值，$\boldsymbol{\mu}_{\mathrm{g}}$ 是全局关联特征的均值，经过 σ 将像素的协方差矩阵归一化，得到的 $\alpha_{j,i}$ 是一个白化点乘项，代表像素 i，j 间成对关系。

2) 为了增强像素的个性捕获能力，获取更丰富的特征细节信息，增加一个像素显著性建模操作，通过 softmax 归一化和 expand 操作，以及多分支特征图互相补充增强跨维度的细节信息融合，强化加权融合后特征图的丰富性以及显著性，最终达到注意力特征图的整体建模操作。该相似度计算可表示为[10]

$$\chi_{j,i}=\sigma\left[\boldsymbol{\mu}_{\mathrm{f}}^{\mathrm{T}}\boldsymbol{g}\left(\boldsymbol{u}_j\right)\right]=\sigma\left[\boldsymbol{\mu}_{\mathrm{f}}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{u}_j\right]\rightarrow\boldsymbol{m}_j=\boldsymbol{W}_m\boldsymbol{u}_j,\tag{2}$$

式中，$\boldsymbol{\mu}_{\mathrm{f}}^{\mathrm{T}}\boldsymbol{g}\left(\boldsymbol{u}_j\right)$ 为 non-local 的注意力计算公式解耦为 $\alpha_{j,i}$ 后余下的单一项，像素 j 对其他所有像素 i 具有相同的影响，为了能将(2)式应用于图像领域，与 SAGAN 引入方式类似，以 $\boldsymbol{W}_g\boldsymbol{u}_j$ 计算 $\boldsymbol{g}\left(\boldsymbol{u}_j\right)$ 得到像素 j 的特征值[8]，同时，为了学习该路特征更多信息，因此用独立线性变换 $\boldsymbol{W}_m$ 替代 $\boldsymbol{W}_g$，建立新的分支 $\boldsymbol{m}(u)$，将 $\boldsymbol{\mu}_f^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{u}_j$ 近似于 $\boldsymbol{m}_j$，代表提取的像素 j 的显著特征信息。

3) 通过以上白化操作以及像素显著性建模两大优化，构建成一个新的自注意力模块，经过这两步改进，降低相似像素的冗余影响和增加像素的显著性建模使得模型能够生成更具突出边界的特征图像，以便生成图像保持更细节的图像纹理，具体模块原理如图 2 所示。

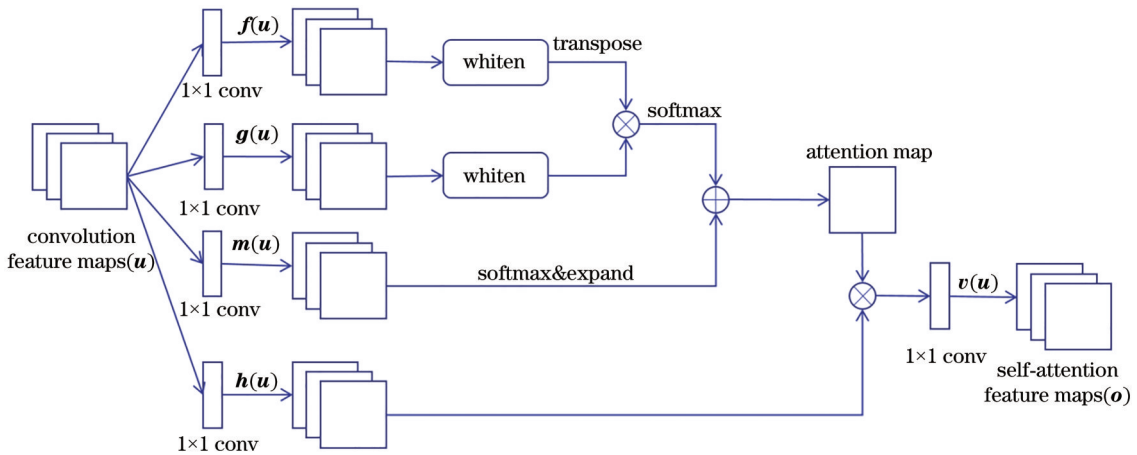

图 2　改进后的自注意力机制

Fig. 2　Improved self-attention mechanism

输入是前一层卷积提取的图像特征图 $\boldsymbol{u}\in\mathbb{R}^{C\times N}$，特征图尺寸是输入图像的1/8。自注意力模块分为四个分支，其中 $\boldsymbol{f}(\boldsymbol{u})=\boldsymbol{W}_f\boldsymbol{u}$ 用于提取像素普遍特征，$\boldsymbol{g}(\boldsymbol{u})=\boldsymbol{W}_g\boldsymbol{u}$ 用于提取像素全局特征[10]，$\boldsymbol{m}(\boldsymbol{u})=\boldsymbol{W}_m\boldsymbol{u}$ 用于提取像素显著信息。

通过对 $\boldsymbol{f}(\boldsymbol{u})$，$\boldsymbol{g}(\boldsymbol{u})$ 和 $\boldsymbol{m}(\boldsymbol{u})$ 变换来计算注意力图 $\beta_{j,i}$，可表示为

$$\beta_{j,i}=\alpha_{j,i}+\chi_{j,i}=\sigma\left\{\left[\boldsymbol{f}\left(\boldsymbol{u}_i\right)-\boldsymbol{\mu}_{\mathrm{f}}\right]^{\mathrm{T}}\left[\boldsymbol{g}\left(\boldsymbol{u}_j\right)-\boldsymbol{\mu}_{\mathrm{g}}\right]\right\}+\sigma\left(\boldsymbol{m}_j\right),\tag{3}$$

式中，$\beta_{j,i}$ 表示在合成第 j 个区域时模型对第 i 个位置的关注度，$\sigma\left(\boldsymbol{m}_j\right)$ 表示对像素显著信息的关注。

注意力层的输出表示为 $\boldsymbol{o}=\left(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_i,\cdots,\boldsymbol{o}_N\right)\in\mathbb{R}^{C\times N}$

$$\boldsymbol{o}_j=\boldsymbol{v}\left[\sum_{i=1}^{N}\boldsymbol{\beta}_{j,i}\boldsymbol{h}\left(\boldsymbol{u}_i\right)\right],\boldsymbol{h}(\boldsymbol{u}_i)=\boldsymbol{W}_h\boldsymbol{u}_i,\boldsymbol{v}(\boldsymbol{u}_i)=\boldsymbol{W}_v\boldsymbol{u}_i,\tag{4}$$

式中，$\boldsymbol{h}(\boldsymbol{u}_i)$ 为特征图经过卷积后得到的特征分支图，$\boldsymbol{v}(\boldsymbol{u}_i)$ 将 $\boldsymbol{h}(\boldsymbol{u}_i)$ 与注意力图相乘后的结果进行卷积处理。

进一步将注意力层的输出乘以比例参数并添加回输入特征图，最终输出为

$$\boldsymbol{z}_i=\gamma\boldsymbol{o}_i+\boldsymbol{u}_i,\tag{5}$$

式中，γ 为预先定义的系数，$\boldsymbol{z}_i$ 表示最终的输出。

最后通过把输入的特征图与生成的特征图进行加权求和，得到特征图任意两个位置的全局依赖关系。

2.3　引入自注意力机制的生成对抗网络

2.3.1　生成网络G2

为加强网络对特征的捕获能力，提升对远距离像素关联捕获能力和对特征内容的建模能力，丰富纹理细节，将2.2中的自注意力机制引入纹理细化阶段生成对抗网络中，改进后G2如图3所示。

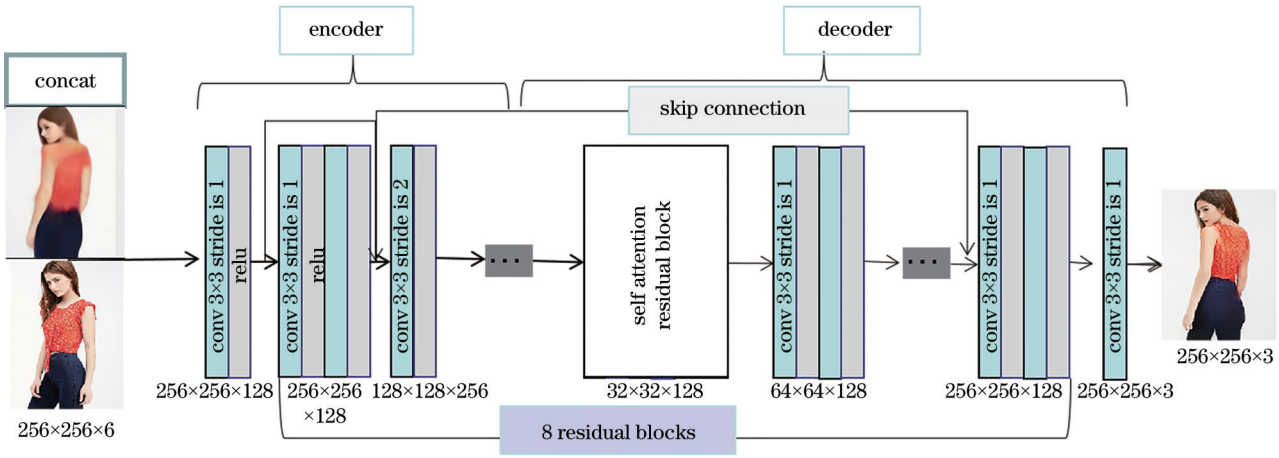

图3　纹理细化生成网络

Fig. 3　Texture refinement generation network

G1生成图像 $\boldsymbol{y}_1$ 与条件图像 $\boldsymbol{x}$ 一同输入G2，G2基于U-net架构，下采样层和上采样层间对称增加跳跃连接，并为了保存输入的更多细节，去掉了U-net的全连接层，以防全连接层压缩输入的大量信息，编码器和解码器模块采用残差结构加强对深层特征提取能力，减弱梯度消失[11-12]。当输入是256×256像素图像时，生成网络中的编码器和解码器分别由四个残差块组成，每个残差块仅由两个stride值为1的卷积层和一个stride值为2的子采样卷积层组成，卷积层由3×3滤波器组成，滤波器数量随每块线性增加，除全连接层和输出卷积层外，每层采用线性整流单元(ReLU)作为激活函数。解码器第二层残差块引入自注意力机制后结构如图4所示。

在自注意力层前增添了归一化层，通过自注意力层进行self-attention权重矩阵计算得到特征图，图像全局结构及其相关性，计算参数在训练中通过反向传播算法自适应调整。

2.3.2　判别网络D

为增强图像纹理细节约束能力，判别网络D采用马尔可夫判别器(PatchGAN)[3]替代基础的判别器结构，在卷积层后引入2.2中的自注意力机制模块，并通过谱归一化(spectral normalization)[13]约束每一层网络权重矩阵的谱范数进而约束lipschitz-1常数，增强模型训练稳定性，减弱梯度消失。

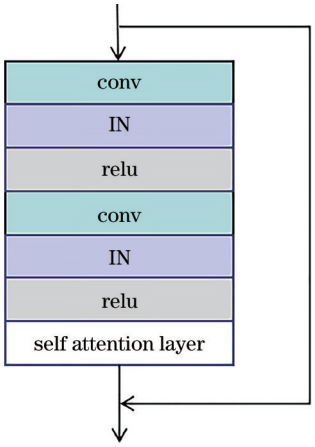

图 4　引入自注意力机制的残差块

Fig. 4　Residual block with self-attention mechanism

由于G1已经生成含有部分低频信息的图像，G2进一步重建了生成图像的低频特征，因此D应更多约束高频细节的生成。普通的判别网络输出是一个代表真假的矢量，对于高分辨率、高细节生成图像的判别效果不足，而PatchGAN是一个全卷积构成的网络，只对图像中每个大小为$N\times N$的patch判断真假，输出是一个$N\times N$矩阵，最后取矩阵均值作为判别网络真假的输出，矩阵中每一个值代表原图一个感受野[3]。PatchGAN没有将整张图像输入判别器判断，而是对相对独立的patch进行鉴别，因此其更多关注GAN对高频信息构建是否正确，对抗学习中补充了生成图像的纹理细节。判别器由五个卷积块和一个全连接层构成，卷积层使用5×5的滤波器，stride值为2时，每层卷积之间使用LeakeyRelu激活函数，并在第2，3，4层卷积后进行谱归一化处理，在第1层卷积后加入self-attention模块，增强判别器对图像远距离关联可信度的判断能力，进而约束图像生成。

2.4　损失函数的优化

2.4.1　特征层面信息损失计算

由于直接使用L_1范数计算损失只是从像素层面衡量图像间差距，不能直接模拟图像的感知质量，为了保证生成图像的全局内容与目标图像接近，提出基于VGG-19[14]预训练模型的内容损失函数，利用预训练好的VGG-19网络特定层提取生成图像和目标图像特征值，比较其语义上的差异并计算损失，以增强生成图像语义一致性。将生成图像和目标图像成对放入VGG-19预训练网络中，分别选取VGG-19中conv1_2，conv2_2，conv4_2，conv5_2四层提取图像特征，并针对每一层计算特征差，用α_k控制每一层特征差所占权重，得到内容损失函数。姿态转移阶段G1与纹理细化阶段G2的生成图像均送入该预训练模型强化生成结果。

G1阶段内容损失函数为

$$\mathcal{L}_{\text{per}}(\boldsymbol{x},\hat{\boldsymbol{y}})=\sum_k\alpha_k\left\|\boldsymbol{\theta}_k\left[\boldsymbol{G}_1(\boldsymbol{x},\hat{\boldsymbol{y}})\right]-\boldsymbol{\theta}_k(\boldsymbol{y})\right\|_1,\tag{6}$$

式中，$\boldsymbol{\theta}_k$是计算感知相似度的网络（即VGG-19模型），α_k控制不同层k在θ总损失中的权重，$\boldsymbol{x}$是条件图像，$\hat{\boldsymbol{y}}$是目标姿势图，$\boldsymbol{G}_1(\boldsymbol{x},\hat{\boldsymbol{y}})$是生成图像，$\boldsymbol{y}$是目标图像。

G2阶段内容损失函数为

$$\mathcal{L}_{\text{per}}(\boldsymbol{x},\boldsymbol{y}_1)=\sum_k\alpha_k\left\|\boldsymbol{\theta}_k\left[\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)\right]-\boldsymbol{\theta}_k(\boldsymbol{y})\right\|_1,\tag{7}$$

式中，$\boldsymbol{y}_1$是G1生成图像，$\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)$是第二阶段生成图像，$\boldsymbol{y}$是目标图像。

通过特征层面信息损失计算，改进生成图像的姿态合理性，并进一步加强纹理细节约束。

2.4.2　优化后的模型损失

细化阶段生成对抗损失可表示为

$$\mathcal{L}_{\text{adv}}^{D}=\mathcal{L}_{\text{bce}}\left[D(\boldsymbol{x},\boldsymbol{y}),1\right]+\mathcal{L}_{\text{bce}}\left\{D\left[\boldsymbol{x},\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)\right],0\right\},\tag{8}$$

$$\mathcal{L}_{\text{adv}}^{G}=\mathcal{L}_{\text{bce}}\left\{D\left[\boldsymbol{x},\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)\right],1\right\},\tag{9}$$

式中，$D(\boldsymbol{x},\boldsymbol{y})$代表给定条件图像与目标图像真伪判断，$D\left[\boldsymbol{x},\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)\right]$代表给定条件图像和第二阶段生成图像真伪判断，$\mathcal{L}_{\text{bce}}$是二元交叉熵损失。

姿态转移网络G1中，原始的图像重建损失为$\mathcal{L}_1$损失，可表示为

$$\mathcal{L}_1=\left\|\boldsymbol{G}_1(\boldsymbol{x},\hat{\boldsymbol{y}})-\boldsymbol{y}\right\|_1,\tag{10}$$

式中，$\boldsymbol{x}$代表条件图像，$\boldsymbol{y}$代表目标图像，$\hat{\boldsymbol{y}}$是$\boldsymbol{y}$提取后的姿态图像，$\boldsymbol{G}_1(\boldsymbol{x},\hat{\boldsymbol{y}})$是生成图像。

为了将人物姿态信息与外观信息更好地集成，减少图像背景的影响，输入图像时，对人物姿态关键点提取后，对周围像素进行膨胀，并建立添加姿态掩模的L_1损失衡量各阶段图像重建质量[4]，其损失为

$$\mathcal{L}_m=\left\|\left[\boldsymbol{G}_1(\boldsymbol{x},\hat{\boldsymbol{y}})-\boldsymbol{y}\right]\odot(1+M_y)\right\|_1,\tag{11}$$

式中，M_y是目标图像的姿态掩模，通过形态学操作得到，前景掩模设为1，后景掩模设为0。

在掩模损失基础上，加入利用VGG-19预训练模型计算的感知损失函数，则G1阶段损失函数为

$$\mathcal{L}_{\text{G1}}=\mathcal{L}_1+\mathcal{L}_m+\omega_1\mathcal{L}_{\text{per}},\tag{12}$$

式中，ω_1是$\mathcal{L}_{\text{per}}$占比权重系数。

外观细化网络 G2 中，由于对抗损失和最小 L_p 距离损失混合可以使得图像生成过程规范化，因此损失函数由生成对抗损失、重建损失和内容损失加权构成为

$$\mathcal{L}_{G2}=\mathcal{L}_{adv}^{G}+\omega_2\left\|\left[\boldsymbol{G}_2(\boldsymbol{x},\boldsymbol{y}_1)-\boldsymbol{y}\right]\odot(1+M_y)\right\|_1+\omega_3\mathcal{L}_{per}(\boldsymbol{x},\boldsymbol{y}_1),\tag{13}$$

式中，ω_2 和 ω_3 分别是图像重建损失和内容损失的所占权重，M_y 是 $\boldsymbol{y}$ 的姿态掩模。

通过以上损失函数的优化，改进判别器，并在生成对抗网络中引入改进后的自注意力机制，可以有效加强姿态转移生成的合理性，生成纹理细节更丰富的图像。

3　实验结果与分析

实验环境为 Ubuntu 16.04LTS 操作系统，深度学习框架选用 Pytorch，软件环境为 cuda9.0 和 python3.6.10，GPU 采用 NVIDIA quadro p5000。实验在 DeepFashion[15] 和 Market-1501[16] 两个数据集上进行，采用 adam 优化器，$\beta_1=0.5$，$\beta_2=0.999$，初始学习率设为 $2\exp(-5)$，在 DeepFashion 数据集上共运行 10 个 epoch，每个 epoch 迭代 30000 次，在 Market-1501 数据集运行 16 个 epoch，每个 epoch 迭代 22000 次，由于输入图像大小不一样，本模型对不同数据集进行模块微调，对于 DeepFashion 数据集，G1 和 G2 的编码器、解码器分别有 6 个、4 个残差块，对于 Market-1501 数据集，G1 和 G2 的编码器、解码器分别有 5 个、3 个残差块。

3.1　数据集

为了证明方法的有效性，分别在含有复杂背景信息的 Market-1501 数据集[16] 和不含背景信息的 DeepFashion 数据集[15] 上进行验证。在数据预处理阶段，将 Market-1501 和 DeepFashion 数据集中图像大小分别调整为 128×64 和 256×256，检测数据集是否包含人物图像，将不包含人物图像的图片去除。接着为了得到人物图像的骨骼姿势，利用文献[7]提取人体关节，为每张图像生成对应的包含 18 个关节位置的姿态热图。其中 Market-1501 数据集是背景多样的 1501 个不同人物的 32668 张图像，训练集包含 452420 对图像，测试集随机选择 12800 对图像；DeepFashion 数据集背景干净，包含 52712 张衣服图像，训练集包含 89262 对图像，测试集包含 12000 对图像，训练集与测试集均不存在重复人物。

3.2　评价指标

采用 Inception Score(IS)[17] 和结构相似性指标(SSIM)[18] 对图像生成质量进行定量评估。对于 Market-1501 数据集，由于条件图像和目标图像可能拥有不同的背景，若输入图像缺少目标图像背景信息，则生成网络很难模拟真实图像的背景构成，因此为了减少生成图像背景在定量评估时的影响，增加对应掩模版本评价指标 mask Structural SIMilarity[4](mask-SSIM) 和 mask Inception Score[4](mask-IS)，即在计算性能指标前，条件图像和目标图像添加姿态掩模，评价指标专注于评估人物整体生成质量。

IS 是衡量生成模型性能的指标，利用 Inception V3 判断图像类别[17]，定义为

$$i_{IS}(G)=\exp\left\{E_{x\sim p_g}D_{KL}\left[p(\boldsymbol{y}|\boldsymbol{x})\,\|\,p(\boldsymbol{y})\right]\right\},\tag{14}$$

式中，$p(\boldsymbol{y}|\boldsymbol{x})$ 用于衡量图像清晰度，$p(\boldsymbol{y})$ 用于衡量图像多样性，D_{KL} 计算两个分布之间的 KL 散度，$E_{x\sim p_g}$ 是分布函数期望。

SSIM 模拟人类视觉感知，从成对样本的亮度、对比度和结构三个角度进行衡量，由于局部计算结构相似度效果更好，因此利用 $N\times N$ 滑动窗口不断在图像上取值计算，将所有计算结果取平均值得到图像全局的 SSIM，其值始终小于等于 1，当 SSIM 值为 1 时表示成对图像完全一致[18]。

3.3　消融实验

为了验证和评估本文改进模块的有效性，通过以下四个不同的模型进行消融实验：1)基线模型(baseline：复现 PG^2)；2)改进判别器(baseline+D)；3)引入 VGG-19 内容损失(baseline+D+L_{per})；4)引入自注意力层的完整模型(baseline+D+L_{per}+SA)。为验证模型多情况下的泛化性，在简单背景数据集 DeepFashion 和复杂背景数据集 Market-1501 上分别选取 4 种不同姿态下的人物图像进行实验，结果如图 5 和图 6 所示。其中，不同行代表了不同姿态，对于每种姿态，第一列表示原始条件图像，第二列表示转换时的目标姿势，第三列表示原始目标图像，第四列表示基线模型生成的图像，第五列表示改进判别器的模型生成的图像，第六列表示引入 VGG-19 内容损失的模型生成的图像，第七列表示引入自注意力层的完整模型生成的图像。

从图 5 可以看出，baseline 模型可以生成较完整的人物图像，但仍存在细节丢失、姿态转移失败等问题。改进判别器(baseline+D)后，a5，d5 人物服

图 5　在 DeepFashion 数据集进行消融实验。(a)样本 1,(b)样本 2,(c)样本 3,(d)样本 4

Fig. 5　Perform ablation experiments on the DeepFashion dataset. (a) Sample 1; (b) sample 2; (c) sample 3; (d) sample 4

装轮廓明显变清晰,并纠正少数颜色生成错误区域,加强生成图像高频细节约束。c5 中人物的服装仍有部分融合,胳膊依然存在扭曲,引入内容损失函数(baseline+D+L_{per})后减弱这一现象。对比 d6 和 d7,d7 背部服装结构更加完整,较远区域间像素仍具备合理性关联,并且图像纹理细节更为真实,证明了引入自注意力机制(baseline+D+L_{per}+SA)可以提高模型细节生成能力。

在图 6 中,对比第五列和第四列,明显看出 PatchGAN 判别器对局部区域判别能力更强,第五列人物外观更自然,c5 没有 c4 中噪点区域,学习到目标条件图像纹理;对比第六列与第五列,由于在该数据集上验证时,L_{per} 损失权重赋值较高,因此第六列图像在具有更清晰边界同时,a6、c6 都产生了一定的锐化现象,证明 L_{per} 对图像内容结构具有一定的约束能力;加入自注意力机制后,第七列图像明显结构清晰,纹理丰富,像素边界更分明。

在 Market-1501 和 DeepFashion 数据集上分别对消融实验进行定量评估,其中 IS 与 mask-IS 值越大,生成模型效果越好,SSIM 与 mask-SSIM 值越大,生成图像与目标图像的相似性越高,结果如表 1 所示。

由表 1 可以看出,各阶段改进后模型均有一定程度性能,指标均高于 baseline,证明模型能够增强图像生成质量。由于该模型内容损失函数权重分配较高,引入自注意力机制后,Market-1501 数据集上图像的 SSIM、Mask-SSIM 值分别比基线模型提升了 0.044 和 0.111,但是 IS 和 Mask-IS 值尚未超越(Baseline+D+L_{per}),根据图 6 分析可知,(Baseline+D+L_{per})模型存在局部锐化严重问题,主观感知远不如(Baseline+D+L_{per}+SA)方法,并且 Baseline+D+L_{per}+SA 相对于基线方法,IS、Mask-IS 分别也提升了 0.024、0.209,在 DeepFashion 数据集上,SSIM 和 IS 指标均取得消融

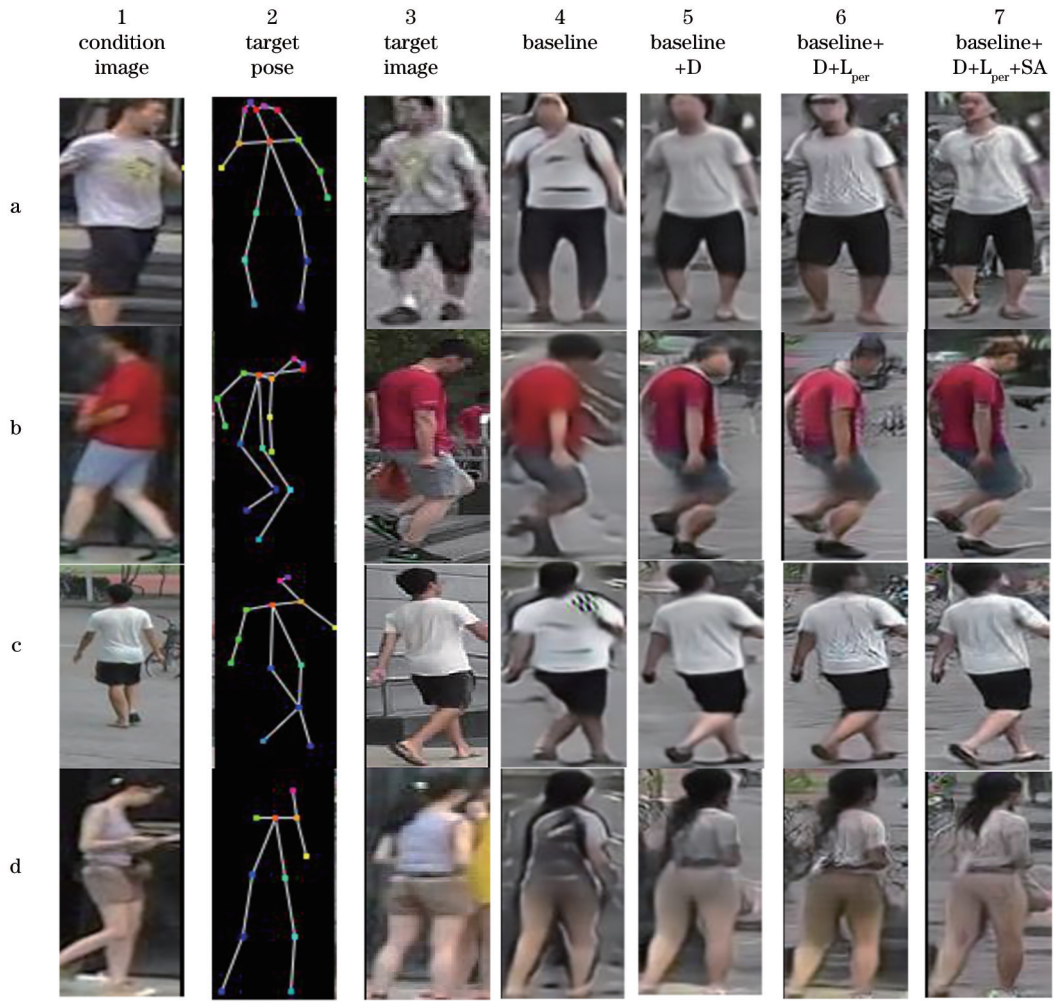

图 6　在 Market-1501 数据集进行消融实验。

Fig. 6　Perform ablation experiments on the Market-1501 dataset. (a) Sample 1; (b) sample 2; (c) sample 3; (d) sample 4

表 1　消融实验

Table 1　Ablation experiment

Model	Market-1501				Deepfashion	
	SSIM	IS	Mask-SSIM	Mask-IS	SSIM	IS
Baseline	0.274	3.472	0.712	3.456	0.614	3.112
Baseline+D	0.315	3.470	0.789	3.449	0.761	3.108
Baseline+D+L_{per}	0.311	**3.501**	0.767	**3.672**	0.759	3.301
Baseline+D+L_{per}+SA	**0.318**	3.496	**0.823**	3.665	**0.794**	**3.478**
Real data	1.000	3.890	1.000	3.706	1.000	3.898

对照组中最优的结果，因此证明 Baseline+D+L_{per}+SA 模型有效，能生成更真实的图像。

3.4　对比实验

为了进一步证明模型有效性，选取姿态迁移人物图像生成领域中近年来主流方法 PG²[4]，DPIG[19]，Def-GAN[20]，UPIS[21]、PATN[1]和 XingGAN[6]与本方法进行对比。定量评价结果如表 2 所示。

从表 2 结果可知，本研究的模型在 Market-1501 数据集上，SSIM 和 Mask-SSIM 上取得了比其他方法更优的结果达到了 0.318 和 0.823，在 DeepFashion 数据集上 SSIM、IS 值达到了 0.794 和 3.478，比其他方法效果更好，相对同样是两阶段生成模型 PG²，在 Market-1501 数据集上 SSIM、IS、Mask-SSIM 和 Mask-IS 指标分别提升了 0.065、0.036、0.031 和 0.230，在 DeepFashion 数据集上 SSIM、IS 分别提升了 0.032 和 0.388，证明生成图像

表 2 不同方法的 SSIM、IS 分数对比

Table 2 Comparision of SSIM and IS by different methods

Model	Market-1501				DeepFashion	
	SSIM	IS	Mask-SSIM	Mask-IS	SSIM	IS
PG^{2}[4]	0.253	3.460	0.792	3.435	0.762	3.090
DPIG[19]	0.099	3.483	0.614	3.491	0.614	3.228
Def-GAN[20]	0.290	3.185	0.805	3.502	0.756	3.439
UPIS[21]	0.151	3.431	0.742	3.485	0.747	2.971
PATN[1]	0.311	3.323	0.811	3.773	0.773	3.209
XingGAN[6]	0.313	**3.506**	0.816	**3.872**	0.778	3.476
Ours	**0.318**	3.496	**0.823**	3.665	**0.794**	**3.478**
Real data	1.000	3.890	1.000	3.706	1.000	3.898

更符合人物视觉感知，与目标图像相似性更高。由于 IS 指标反映图像清晰度与多样性，其值受神经网络参数影响较大，而 SSIM 指标反映了结构相似度，对于姿态迁移生成工作而言，结构正确性远比图像多样性重要，因此，尽管本方法在复杂背景数据集 Market-1501 上 IS 和 Mask-IS 的值未超过 XingGAN，但是 SSIM 值达到以上方法最优，可以证明本方法具有有效性，生成质量更高。

4 结 论

在两阶段姿态迁移生成模型 PG^{2}的基础上，提出一种融合自注意力机制的人物姿态迁移生成模型。在纹理细化阶段生成对抗网络中引入改进后的自注意力机制，加强相似像素区域的学习，增强特征的捕获，进而提升生成图像的细节丰富度。设计判别网络对成对图像局部区域进行真假判断，增强对高频细节的捕捉，保持图像风格。两阶段均利用已训练好的深度卷积神经网络对图像语义一致性进行约束，增强姿态转移合理性。实验从定性角度分别在 Market-1501 和 DeepFashion 数据集上验证了本方法的有效性，并利用初始分数 IS 和结构相似性 SSIM 对图像进行定量分析，获得模型的生成图像细节更丰富，结构更清晰，生成质量有一定的提升，在 DeepFashion 数据集和 Market-1501 数据集上，本方法比 PG^{2}方法 IS 值与 SSIM 值分别提升了 0.388、0.036 和 0.032、0.065。

参 考 文 献

[1] Zhu Z, Huang T T, Shi B G, et al. Progressive pose attention transfer for person image generation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 2342-2351.

[2] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. [S.l.: s.n.], 2014: 2672-2680.

[3] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, Honolulu, HI, USA. New York: IEEE Press, 2017: 5967-5976.

[4] Ma L Q, Jia X, Sun Q R, et al. Pose guided person image generation[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. [S.l.: s.n.], 2017: 406-416.

[5] Huang Y W, Zhao P, You Y D. Pose-guided human image synthesis based on fusion feature feedback mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141011.
黄友文, 赵朋, 游亚东. 融合反馈机制的姿态引导人物图像生成[J]. 激光与光电子学进展, 2020, 57(14): 141011.

[6] Tang H, Bai S, Zhang L, et al. XingGAN for person image generation[M]// Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12370: 717-734.

[7] Cao Z, Simon T, Wei S, et al. Realtime multi-person 2D pose estimation using part affinity fields [EB/OL]. (2016-11-24) [2021-02-10]. https://arxiv.org/abs/1611.08050.

[8] Zhang H, Goodfellow I J, Metaxas D, et al. Self-attention generative adversarial networks[C]// Proceedings of the 36th International Conference on Machine Learning, ICML 2019, June 9-15, 2019, Long Beach, California, USA. Cambridge: PMLP, 2019: 7354-7363.

[9] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[10] Yin M H, Yao Z L, Cao Y, et al. Disentangled non-local neural networks[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12360: 191-207.

[11] Yan B, Zhang L, Zhang J L, et al. Image generation method for adversarial network based on residual structure[J]. Laser & Optoelectronics Progress, 2020, 57(18): 181504.
颜贝, 张礼, 张建林, 等. 基于残差结构的对抗式网络图像生成方法[J]. 激光与光电子学进展, 2020, 57(18): 181504.

[12] Chen Q J, Qu M. Low-light image enhancement based on cascaded residual generative adversarial network[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141024.
陈清江, 屈梅. 基于级联残差生成对抗网络的低照度图像增强[J]. 激光与光电子学进展, 2020, 57(14): 141024.

[13] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks [EB/OL]. (2018-02-16) [2021-02-10]. https://arxiv.org/abs/1802.05957.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2021-02-10]. https://arxiv. org/abs/1409.1556.

[15] Liu Z W, Luo P, Qiu S, et al. DeepFashion: powering robust clothes recognition and retrieval with rich annotations[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1096-1104.

[16] Zheng L, Shen L Y, Tian L, et al. Scalable person Re-identification: a benchmark[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1116-1124.

[17] Salimans T, Goodfellow I J, Zaremba W, et al. Improved techniques for training GANs[C]//Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. [S.l.: s.n.], 2016: 2234-2242.

[18] Yao J C, Liu G Z. Improved SSIM IQA of contrast distortion based on the contrast sensitivity characteristics of HVS[J]. IET Image Processing, 2018, 12(6): 872-879.

[19] Ma L Q, Sun Q R, Georgoulis S, et al. Disentangled person image generation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 99-108.

[20] Siarohin A, Sangineto E, Lathuilière S, et al. Deformable GANs for pose-based human image generation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3408-3416.

[21] Pumarola A, Agudo A, Sanfeliu A, et al. Unsupervised person image synthesis in arbitrary poses[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8620-8628.