Computer Science and Application 计算机科学与应用, 2021, 11(4), 876-884
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114090

Hans汉斯

基于嵌套残差注意力模块的人物姿态转换方法

钟晓静，谭台哲

广东工业大学计算机学院，广东 广州
Email: 15757301765@163.com, 969313709@qq.com

收稿日期：2021年3月15日；录用日期：2021年4月10日；发布日期：2021年4月21日

摘　要

近几年，人们围绕人物图像合成技术展开了多项研究，姿态转换就是其中一个。作为条件输入的姿态信息的引导有局限性，视角变换时生成模型难以处理复杂的人物外观特征。注意力机制可以有效提取图像中的重要部分，通过将提取特征用的残差块嵌入到残差注意力模块中，通过短跳跃连接来逐步学习姿态相关性，自适应地选择空间像素，充分利用姿态转换过程中的全局空间信息，提高生成网络的表征能力，生成具有目标姿态的高质量人物图像。在多类别大型服装数据集DeepFashion上进行测试，验证了所提出算法的有效性。

关键词

深度学习，注意力机制，姿态转换

Nested Residual Attention Module-Based for Human Pose Transfer

Xiaojing Zhong, Taizhe Tan

School of Computers, Guangdong University of Technology, Guangzhou Guangdong
Email: 15757301765@163.com, 969313709@qq.com

Received: Mar. 15th, 2021; accepted: Apr. 10th, 2021; published: Apr. 21st, 2021

Abstract

In recent years, several researches have been carried out around the techniques of human image synthesis, and pose transfer is one of them. The guidance of pose information as conditional input

文章引用: 钟晓静, 谭台哲. 基于嵌套残差注意力模块的人物姿态转换方法[J]. 计算机科学与应用, 2021, 11(4): 876-884. DOI: 10.12677/csa.2021.114090

has limitations, and it is difficult to generate models to handle complex character appearance features during perspective transformation. The attention mechanism can effectively extract the important parts of the image, learn the pose correlation step by step by through the short skip connections embedding the residual blocks for feature extraction into the residual attention module, select spatial pixels adaptively making full use of the global spatial information in the pose transformation process, improve the representational capability of the generative network and generate high-quality person images with target pose. The effectiveness of the proposed algorithm is verified by testing on a multi-category large clothing dataset DeepFashion.

Keywords

Deep Learning, Attention Mechanism, Pose Transfer

1. 引言

近几年，随着深度学习中生成对抗网络(Generative Adversarial Network, GAN) [1]的发展，人物图像合成技术受到越来越多的关注，已经成为一个研究的热点问题，在图像编辑、电影制作、行人重识别和虚拟换衣等方向上有广泛的应用[2] [3] [4]。通过姿态节点引导的人物图像合成是目前最流行的方法之一，生成具有目标姿态的人物图像实现姿态转换技术的实现，可以方便用户在线上服装类购物平台了解更多的信息，也可以促进虚拟试衣间的完善。然而，由于人体是非刚性的，在姿态变换的过程中可能会引起合成图像中人物的变形和伪影的产生，同时，在不同视角下的不同姿态的人物外观有很大的不同，这使得生成网络必须具有捕捉图像分布中大范围变化的能力和推理被遮挡区域像素的能力。另一方面，服装的纹理和人物外观的细节在生成过程中很容易丢失。因此，目前的人物图像合成技术还存在着许多挑战。

文献[5]通过对人体的 3D 模型进行建模，再将 3D 人物渲染到 2D 图像中，这种方法通常比较耗时，并且计算量很大。文献[6]首先提出用两阶段的模型来实现人物图像的合成，但这种 coarse-to-fine 策略的计算步骤很复杂。随后文献[7]在之前的基础上将图像的前景、背景和人物的姿态分解和编码成不同的特征，通过一定的组合方式再把它们结合在一起生成目标姿态人物图像，虽然提升了对生成过程的编码控制，但是最后生成的图像质量并不高。文献[8]通过提出一种神经网络中跳跃连接的变体来解决因姿态不同造成的人物外观的像素错位，这种方法不适用于形变较大的姿态转换。

受人类感知过程的启发，注意力机制在计算机视觉领域中常用来提取重要区域的特征[9] [10]。同时残差网络 ResNet [11]的提出有效解决了网络退化和网络难训练的问题，被广泛应用于深层网络中。

综上所述，本文提出了基于嵌套残差注意力模块的人物姿态转换方法。一方面，保留了传统提取特征用的残差模块用于优化深层网络的优势，同时为了减少静默神经元的出现本文将里面的激活函数替换成 LeakyReLu [12]。另一方面，将残差模块嵌入在注意力模块里，可以在深层网络传播过程中根据注意力机制的动态变化学习全局特征的局部相关性，高权重聚集重要信息，通过多方位残差学习融合特征信息，使得人物图像可以递进式转换姿态，相比之前的方法不会丢失更多细节，平缓地完成姿态转换的过程，同时添加 Style Loss [13]损失函数来引入 Gram 矩阵提升姿态转换前后图像风格的匹配程度，使生成图像的风格和原图像风格保持一致。

2. 基础理论

2.1. 人体姿态估计

基于深度学习的人物图像合成算法往往通过 OpenPose [14]做姿态估计。OpenPose 用热力图 $S=\left(S_1,S_2,\cdots,S_j\right)$来表示要检测的关节点数，用向量图 $L=\left(L_1,L_2,\cdots,L_c\right)$来表示要检测的关节对数。通过 2D 高斯分布建模，第 k 个人的第 j 个关节点在 p 点的热力图 $S_{j,k}^{*}\left(p\right)$可表示为：

$$S_{j,k}^{*}\left(p\right)=\exp\left(-\frac{\left\|p-x_{j,k}\right\|_2^2}{\sigma^2}\right) \tag{1}$$

其中，$x_{j,k}$表示坐标位置，σ用于控制峰值的范围。

通过关节联通区域(Part Affinity Fields，PAFs)建模骨骼区域，对骨骼区域内的每一个像素点使用 2D 向量表征位置和方向信息。第 k 个人的第 c 对关节对在 p 点的向量图 $L_{c,k}^{*}\left(p\right)$可表示为：

$$L_{c,k}^{*}\left(p\right)=\begin{cases}\dfrac{x_{j2,k}-x_{j1,k}}{\left\|x_{j2,k}-x_{j1,k}\right\|_2}, & \text{if } p \text{ on } limbc,k\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

得到各个关节点的坐标热力图和关节对的向量图后，将各个关节连接成一段骨骼，快速对应到不同人物个体中。图 1 为 OpenPose 做双向匹配的示意图。

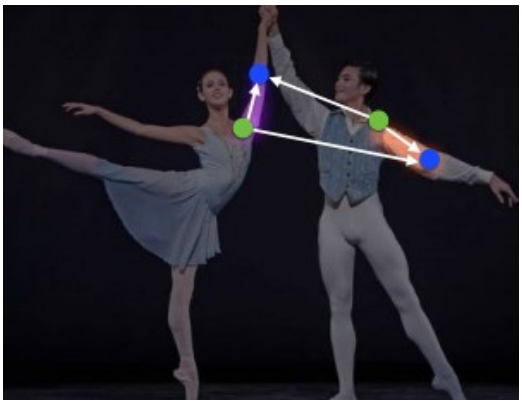

Figure 1. OpenPose bipartite matching
图 1. OpenPose 双向匹配

2.2. 注意力机制

注意力是人类大脑中一项不可或缺的复杂认知功能，人们会无意识地对外界输入的大量信息在大脑中初步筛选出重点部分，并暂时性地忽略其它部分，这种能力就叫做注意力。深度学习中把注意力机制分为软性注意力机制和硬性注意力机制，考虑到硬性注意力机制不方便通过反向传播方法进行训练，一般常用的是软性注意力机制，计算所有输入信息的加权平均，再输入到神经网络中。能够以高权重去聚焦重要信息，低权重去忽略不相关的信息，并且还可以不断调整权重，有很高的可拓展性。其基本网络框架如图 2 所示。在由卷积层、池化层等常见操作层构建出来的卷积神经网络中，每一层对于一副图像的所有空间特征是一视同仁的，为了让网络更加关注重点区域，通常会在网络中添加一层或几层注意力模块[15]来获得想要的信息。

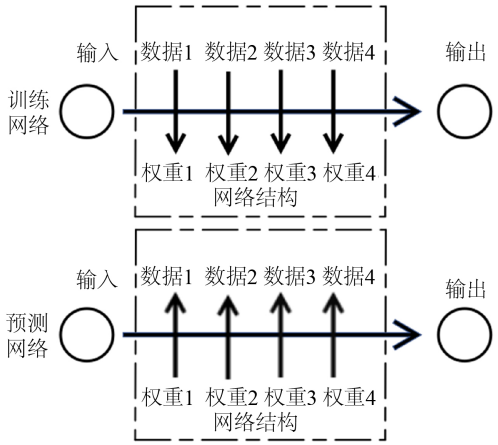

Figure 2. Attention mechanism
图 2. 注意力机制

3. 本文算法

3.1. 嵌套残差注意力模块

针对于通过姿态节点引导的人物姿态转换任务，将人体姿态估计得到的姿态节点和人物外观信息结合在一起做卷积可能会造成语义信息的缠结和混乱，因此本文提出将姿态信息和外观信息分成两个分支来处理。从姿态信息分支中提取特征注意力，对全局姿态信息进行编码，再通过嵌套残差注意力模块将软注意力机制下提取出的掩膜自适应选择外观信息分支中的人物特征信息，加强全局特征和局部特征的依赖性。不同于以往大多数双支路网络在提取特征过程中互不干扰的做法，本文的两个分支相互作用，不仅外观信息分支会受到引导姿态的影响，姿态信息分支也会随着人物外观特征的更新而更新。

虽然网络层数越深能提取的特征更加丰富，但也存在着难训练、优化效果差等问题。为了解决上面提到的问题，发挥重要作用的就是残差学习。ResNet 通过短路机制引入残差单元，在恒等映射的基础上避免了由深度网络引起的网络退化问题。不仅可以通过残差学习来提取深层特征，同时还可以将短跳跃连接和注意力机制相结合，提取出深层的高级语义信息来引导人物外观特征图的形成，使得人物外观在姿态转换的过程中不会产生巨大的形变。以下有两个把残差块嵌入在注意力模块里的好处：(1) 通过在外观信息分支做短跳跃连接，使得注意力因子的影响范围得到可控，在生成网络面对形变大的姿态转换场景时可以更好地保留特征。(2) 通过把外观信息经过注意力提取后的通道特征叠加到外观信息上，对引导的姿态提供一个反馈信息，起到一个修正作用。图 3 为本文提出的一个嵌套残差注意力模块的结构图。其中，conv 为若干层卷积操作，concat 为特征图在通道层进行叠加操作。

此外，传统的残差模块将 ReLu 函数[16]作为激活函数，当它的输入值为负的时候，输出始终为 0，其一阶导数也始终为 0，这样会导致神经元不能更新参数，这种现象叫做“Dead Neuron”。为了避免“Dead Neuron”的发生，本文将内部的残差块激活函数替换为泄漏型修正线性单元 Leaky ReLU 函数。Leaky ReLu 函数的表达式如下：

$$F(x)=\begin{cases}x, x\geq 0\\ 0.01x, x<0\end{cases} \tag{3}$$

该函数的输出对负值输入有很小的坡度，由于导数总是不为零，这能减少静默神经元的出现，允许基于梯度的学习，解决了 ReLu 函数进入负区间后神经元不能学习的问题。

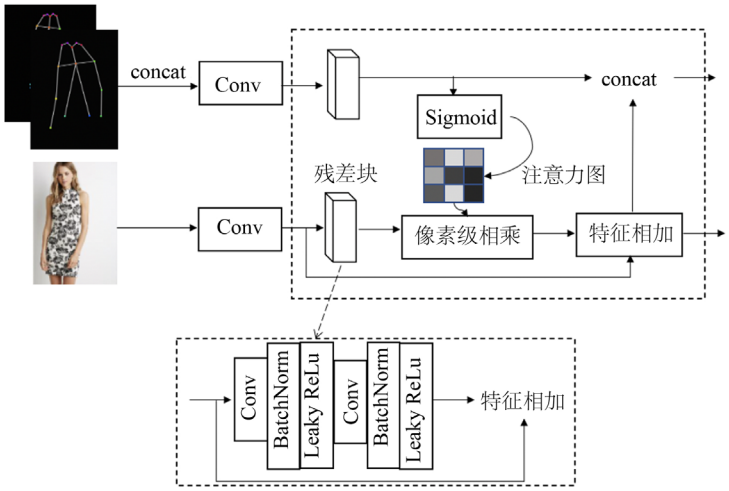

Figure 3. Nested residual attention module
图 3. 嵌套残差注意力模块

目前的姿态转换方法中，大多通过第二次训练做一个加强模块来增强人物图像中的衣服纹理，将姿态和人物外观看作两个单独的问题，是忽略了它们之间的相关性，在这样的基础上人物外观信息的生成和姿态的正确转换很难达到一个优势的平衡点。而在本文提出的人物图像姿态转换网络的生成过程中，把这两个分支通过注意力机制联系在一起，随着每一个嵌套残差注意力模块的更新，经过 Sigmoid 函数生成大小为 $H\times W\times 1$ 的注意力图，$H\times W$ 为特征图空间大小，得到的权重值范围在[0, 1]，使得像素注意力权重在目标区域越接近 1，在非目标区域越接近 0。网络在局部上可以把源人物图像中的像素块转移到目标姿态中对应的位置上去，在全局上可以学会人体各个部位的连接关系，从而推理出因视角局限性而未观察到的外观信息，被分配到权重越大的像素在生成过程中得以保留，权重越小的像素则会被逐步清除。

经过多个嵌套残差注意力模块后，再通过连接 2 个反卷积层恢复图像的分辨率，使其和输入图像保持同样的大小。

3.2. 目标函数

L1 最小化绝对误差损失虽然能以清晰的边缘重建为代价函数去除斑点和伪影，但是基于像素级的损失函数无法判断生成图像和真值样本的相似性，比如两张一样的图像，只不过在位置上有一点点偏移，这两张图像得到的损失值可能会非常大。为了让生成图像与网络的输入条件更匹配，根据条件生成对抗网络(conditional Generative Adversarial Network, cGAN) [17]通过改变标签得到条件相关的生成图像的原理，把对抗损失里的概率表达转换成条件概率来训练判别器使得生成器生成的图像越接近目标条件。

生成网络的目的是将目标姿态的人体解析图迁移成人物图像的风格。Gram 矩阵是特征图之间的偏心协方差矩阵，通过计算各个维度的内积来度量它们的互相关程度。对于特征图来说，浅层网络提取出的是局部的纹理细节特征，深层网络提出的是更抽象的轮廓、大小等信息，由这些向量计算出来的 Gram 矩阵可以把图像特征之间隐藏的联系提取出来，也就是各个特征之间的相关性高低。因此，本文引入了 Style Loss 来比较每一层特征图的 Gram 矩阵判断生成图像和真值图像的风格的相似性。Gram 矩阵通过以下公式来计算：

$$G_j^{\phi}(x)_{c,c'}=\frac{1}{C_jH_jW_j}\sum_{h=1}^{H_j}\sum_{w=1}^{W_j}\phi_j(x)_{h,w,c}\,\phi_j(x)_{h,w,c'} \tag{4}$$

其中，$C_jH_jW_j$ 分别指的是相对应的通道值和空间值，$\phi_j(x)$ 指的是图像经过 VGG-16 [18]特征提取网络得到的第 j 层中间层的特征图。通过比较每一层特征图的 Gram 矩阵，再计算它们之间的欧式距离，最后将不同层得到的损失相加得到最后的 Style Loss。风格损失通过下面公式计算得到：

$$\ell_{\text{style}}^{\phi,j}(\hat{y},y)=\left\|G_j^{\phi}(\hat{y})-G_j^{\phi}(y)\right\|_2^2 \tag{5}$$

4. 实验及其分析

4.1. 数据集

本文实验使用的数据集为香港中文大学多媒体实验室提供的多类别大型服装数据集 DeepFashion 中的子集‘In-shop Clothes’，其中包含 52712 张人物图像，7982 种衣服商品，本文将 48674 张图像划分为训练集，剩余的 4038 张图像作为测试集。训练时，batch_size 设置为 8，Epoch 设置为 800，生成器和判别器均使用 Adam 优化器。实验部分从定性和定量两个方面评估所提出的嵌套残差注意力模块方法的有效性。

4.2. 评价标准

由于本文做的是生成任务研究，通过肉眼观察可以评判出生成图像的质量好坏。另一方面，本文从数值定量方面比较了之前提出的 2 种比较主流的人物图像姿态转换算法，包含基于编码控制的 PG2 算法[6]，和基于跳转连接变体的 Deform 算法[8]。

本文使用的评价指标为衡量生成图像质量中常用的结构相似性(Structural Similarity Index Measure, SSIM) [19]和起始分数(Inception Score, IS) [20]。SSIM 指标比较了生成图像和目标图像的亮度、对比度以及结构，SSIM 值越大和目标图像结构越相似。SSIM 值可通过下面公式计算：

$$\text{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)} \tag{6}$$

其中，μ_x 和 μ_y 分别代表图像 x 和图像 y 的均值，σ_x 和 σ_y 分别代表它们的方差，σ_{xy} 代表图像 x 和图像 y 的协方差，C_1 和 C_2 为两个取决于图像像素值的随机常数，避免除零。

IS 指标比较了生成图像和目标图像的清晰度和多样性，IS 值越大，意味着生成的图像质量越高。IS 值可以通过下面公式计算：

$$\text{IS}(G)=\exp\left(E_{x\sim p_g}D_{\text{KL}}\left(p(y\mid x)\,\|\,p(y)\right)\right) \tag{7}$$

其中，x 为生成图像，y 为生成图像输入到分类网络中得到的向量，D_{KL} 为对 $p(y\mid x)$ 和 $p(y)$ 求 KL 散度。

4.3. 定性分析

图 4 所示是本文提出的嵌套残差注意力模块数量从 1 到 9 分别得到的注意力热力图。可以看出从 AM1 到 AM3 的时候由于网络比较浅，能学习到的语义信息还不能描述出具体的人体轮廓。从 AM4 到 AM6 的时候，能看出靠近目标姿态的区域权重逐步增大，网络可以通过残差注意力模块正确学习到姿态信息，并和人物信息相融合。从 AM6 到 AM9 的时候，能看出区域轮廓从点线到块的转变，此时通过掩膜提取得到的是有连接性的语义信息，这意味着姿态转变过程中，覆盖区域的像素可以直接迁移过来，而不属于目标姿态中的区域则会为赋予低权重。说明嵌套的残差注意力模块在深度网络的训练效果下可以取得有效的注意力提取。

从图 5 可以看出，不管是从正面到侧面、侧面到正面、反面到正面，以及姿态转换的视角远近不同，本文提出的人物图像姿态转换方法都可以较好地保留人物的外观细节特征(脸部和头发)，还有衣服的完整

形状和人体的正确轮廓，同时可以恢复目标姿态中在源图像被遮挡或者覆盖的像素，接近真值样本。

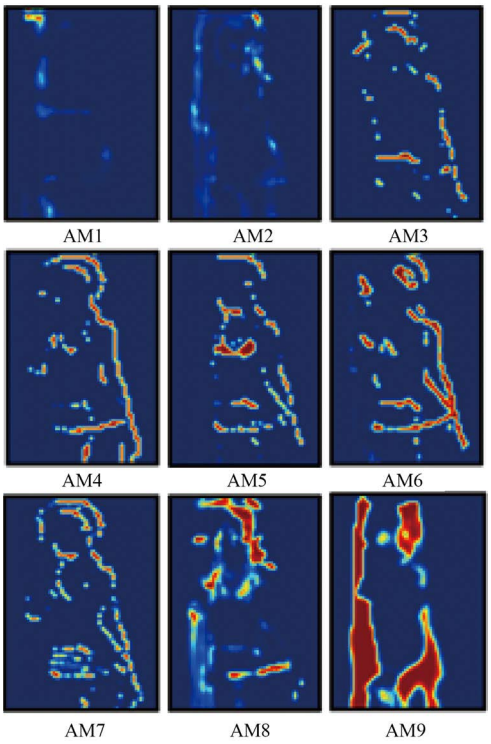

Figure 4. Attention heatmap
图 4. 注意力热力图

Figure 5. Qualitative comparison with other methods
图 5.与其它方法的定性比较

4.4. 定量分析

表 1 为各人物图像姿态转换算法的 SSIM 值和 IS 值的对比，可以看出本文算法相较之前的方法均取得了最高的 SSIM 值和 IS 值，SSIM 值比之前最好的分数提高了 3.84%，IS 值比之前最好的分数提高了 4.02%。由此可以看出，本文提出的方法生成的图像不仅清晰度高，不容易失真，并且和目标图像具有更相似的结构。

Table 1. Comparison of SSIM and IS values with other methods
表 1. 与其它方法的 SSIM 和 IS 值的比较

指标	PG^2	Deform	本文算法
SSIM	0.614	0.756	0.785
IS	3.228	3.362	3.497

表 2 对比了各人物图像姿态转换算法的推理时间，可以看出本文算法的推理时间最短。本文提出的算法是端到端训练，结构简单，具有轻量级嵌入的优势，不同于之前的一些方法需要经过两阶段 coarse-to-fine 的合成，能很好地满足有实时性要求的应用场景。

Table 2. Comparison of inferencing time with other methods
表 2. 与其它方法的推理时间的比较

指标	PG^2/FPS	Deform/FPS	本文算法/FPS
推理时间	17.367	9.336	8.912

综上所述，定量分析的结果与定性分析结果一致，充分表明本文算法的优越性,在姿态转变较大的场景中能较好地应对遮挡和外观形变等挑战。

5. 结束语

本文针对现有的人物图像姿态转换算法生成的人物图像清晰度不高，人物外观容易变形和产生伪影等问题，提出了一种嵌套残差注意力模块，集合了残差模块和注意力模块原本的优势，通过姿态信息分支的引导提取注意力掩膜改变外观信息分支中空间特征的权重，使生成网络具有捕捉图像分布中大范围变化的能力，实现平缓的姿态转换。在公开数据集上测试表明，从定性和定量两方面本文算法都很好地解决了上述提到的问题，本文提出的姿态转换算法相比其它主流算法能够生成更清晰、更精确以及保留更多细节信息的人物图像，生成的图像结构分布也与真实分布更为接近。

参考文献

[1] 孙义博, 张文靖, 王蓉, 李冲, 张琪. 基于通道注意力机制的行人重识别方法[J/OL]. 北京航空航天大学学报, 2021: 1-10. https://doi.org/10.13700/j.bh.1001-5965.2020.0684, 2021-03-07.

[2] 刘若雯, 杨建喜, 赵海博. 基于对偶学习的图像翻译技术研究[J]. 北京电子科技学院学报, 2020, 28(2): 12-18.

[3] 禹立. 基于纹理修复的虚拟试衣网络[D]: [硕士学位论文]. 上海: 东华大学, 2020.

[4] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al.* (2014) Generative Adversarial Networks. arXiv preprint arXiv: 1406.2661.

[5] Lassner, C., Pons-Moll, G. and Gehler, P.V. (2017) A Generative Model of People in Clothing. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 853-862. https://doi.org/10.1109/ICCV.2017.98

[6] Ma, L., Sun, Q., Georgoulis, S., *et al.* (2018) Disentangled Person Image Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 99-108. https://doi.org/10.1109/CVPR.2018.00018

[7] Ma, L., Jia, X., Sun, Q., *et al.* (2017) Pose Guided Person Image Generation. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., Eds., *Advances in Neural Information Processing Systems*, 406-416.

[8] Siarohin, A., Sangineto, E., Lathuiliere, S. and Sebe, N. (2018) Deformable GANs for Pose-Based Human Image Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3408-3416. https://doi.org/10.1109/CVPR.2018.00359

[9] Mnih, V., Heess, N. and Graves, A. (2014) Recur Rent Models of Visual Attention. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, **2**, 2204-2212.

[10] Xiao, T.J., Xu, Y.C., Yang, K.Y., *et al.* (2015) The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-Grained Image Classification. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 842-850.

[11] He, K., Zhang, X., Ren, S., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[12] Maas, A.L., Hannun, A.Y. and Ng, A.Y. (2013) Rectifier Nonlinearities Improve Neural Network Acoustic Models. *International Conference on Machine Learning* (*ICML*), **30**, 3.

[13] Johnson, J., Alahi, A. and Li, F.F. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *European Conference on Computer Vision*, Springer, Cham, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[14] Cao, Z., Hidalgo, G., Simon, T., *et al.* (2019) OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186. https://doi.org/10.1109/TPAMI.2019.2929257

[15] Hu, J., Shen, L. and Sun, G. (2020) Squeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023. https://doi.org/10.1109/TPAMI.2019.2913372

[16] Yarotsky, D. (2017) Error Bounds for Approximations with Deep ReLU Networks. *Neural Networks*, **94**, 103-114. https://doi.org/10.1016/j.neunet.2017.07.002

[17] Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784.

[18] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

[19] Wang, Z., Bovik, A.C., Sheikh, H.R., *et al.* (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/TIP.2003.819861

[20] Salimans, T., Goodfellow, I., Zaremba, W., *et al.* (2016) Improved Techniques for Training Gans. arXiv preprint arXiv:1606.03498.