Computer Science and Application 计算机科学与应用, 2022, 12(4), 1149-1156
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124117

Hans汉斯

基于视觉Transformer的面孔吸引力预测方法研究

方建安，李昶昊

东华大学，信息科学与技术学院，上海

收稿日期：2022年3月18日；录用日期：2022年4月19日；发布日期：2022年4月26日

摘　要

面孔吸引力分析预测是结合认知科学、心理学、计算机科学的一个交叉领域。是对人主观感受的客观量化——通过机器去学习面孔特征与量化的感知间的映射关系。本文提出了一种结合CNN与Transformer结构的混合模型，使用残差卷积网络提取图像的特征图，经嵌入层编码后输入到多层transformer编码器中，利用自注意力机制从全局的角度把握不同特征成分间的关系。该方法在SCUT-FBP5500数据集上取得了较好的实验效果，表明了从全局的角度将人脸图像转化为视觉词向量序列并进行属性预测是可行有效的。

关键词

面孔吸引力预测，Transformer编码器，自注意力机制，深度学习

Research on Face Attractiveness Prediction Method Based on Visual Transformer

Jianan Fang, Changehao Li

College of Information Science and Technology, Donghua University, Shanghai

Received: Mar. 18th, 2022; accepted: Apr. 19th, 2022; published: Apr. 26th, 2022

Abstract

Insert Face attractiveness analysis and prediction is a cross field combining cognitive science, psychology and computer science. It is the objective quantification of people's subjective feelings, learning the mapping relationship between face features and quantitative perception through machines. In this paper, a hybrid model combining CNN and transformer structure is proposed. The residual convolution network is used to extract the feature map of the image, which is en-

文章引用: 方建安, 李昶昊. 基于视觉 Transformer 的面孔吸引力预测方法研究[J]. 计算机科学与应用, 2022, 12(4): 1149-1156. DOI: 10.12677/csa.2022.124117

coded by the embedded layer and input into the multi-layer transformer encoder. The self attention mechanism is used to grasp the relationship between different feature components from a global perspective. This method has achieved good experimental results on scut-fbp5500 data set, which shows that it is feasible and effective to transform face image into visual word vector sequence and predict attributes from a global perspective.

Keywords

Face Attractiveness Prediction, Transformer Encoder, Self-Attention Mechanism, Deep Learning

1. 引言

面孔吸引力是人在面孔知觉上的一种认知偏好，在我们的社会交往和人际互动中有着重要作用。社会学家，心理学家通过实验研究证明了富有吸引力的面孔往往会在社会认可、职业发展、人际关系上带来更多的好处与优势。因此，对这种机制的探索和研究具有实质上的意义。大量认知心理学的实验表明[1]，这种不同个体主观的感受具有一定的共性与规律。并且是可以通过机器去学习的[2]。

面孔认知偏好分析与预测在人脸美化算法、医学美容指导、基于内容的图像检索、推荐系统等方向上具有重要的应用价值。但这个问题仍然具有挑战性，一方面它涉及到认知科学、心理学、等多学科的交叉领域，另一方面需要大量数据作为驱动，而目前这方面的数据集较少。

早期的研究方法主要是通过手工设计不同特征表述符[3] [4]，这些特征可以是基于几何、颜色、纹理，也可以是基于局部或整体尺度。随着深度学习的发展，许多学者将 CNN 结构的网络模型[5] [6] [7]应用到该领域上，采用深度网络的相比传统的机器学习可以提取到更深层、更抽象的特征。能更好地学习到贴近于人类的审美认知机制。近年来，视觉 Transformer [8]的出现打破了计算机视觉与自然语言处理的壁垒，相较于 CNN 模型，Transformer 框架更具有更大的数据容量与相对更好的性能，能捕获图像的长距离依赖关系，具有全局性。但存在一些缺点，首先 Transfomer 处理视觉任务时缺乏 CNN 网络所具有的归纳偏置，需要在大规模的数据集上作训练，并且参数量较大，计算复杂度高(与 token 的平方相关)。由此，一些混合模型兼顾了卷积的归纳偏置和多头自注意力机制捕获长距离相互作用的能力，如 BoTNet [9]、CoAtNet [10]、TranCNN [11]，表明结合卷积层和注意层可以获得更好的泛化能力和容量，在较小的数据集上也能取得不错效果。

2. 任务流程及定义

2.1. 任务流程

面孔吸引力分析预测本质是一个多范式计算问题，可视作回归、分类、排序问题处理。其关键在于如何去学习人脸特征到评价统计量的映射关系。如图 1 所示，首先将原始的图像数据集引入评价系统，量化每位评价者的主观感受，对这种主观感受的量化指标做数据处理后生成具有代表性的统计量。统计量的不同代表处理这个问题的不同范式，如取样本均值可作为回归问题，取众数作为分类问题，取在不同评级上的分布率可作为一个标签分布问题[12]。将这个统计量作为该图像对应的标签。另一方面是将数据集做预处理，比如剪切、旋转、对齐、归一化等方式转为便于机器处理的数据。一部分作为训练集和

验证集输入到模型中进行训练和调优。传统的机器学习方法中，分为特征工程和输出算法模块(如高斯回归、支持向量机分类等)，而深度学习模型则是一个由输入到输出端到端的整体架构。另一部分数据作为测试集测试检验模型的学习效果，对新的输入图像进行评估预测。另外，可以迁移到新的数据集上对其进行训练微调。

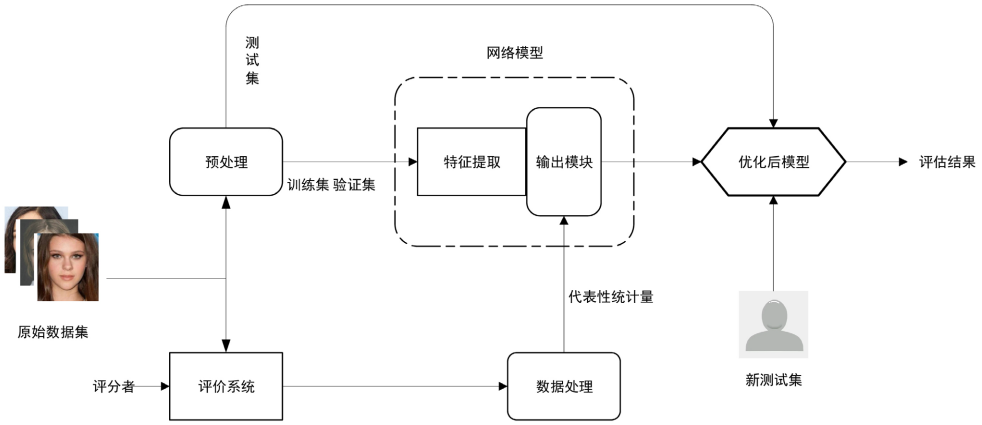

Figure 1. Task flow-process diagram
图 1. 任务流程图

2.2. 任务定义

对给定的人脸图像集：$F=\{F_1,F_2,\cdots,F_n\}$，n 为数据集容量。其中对给定的人脸 F_i，由 k 个评价者给出的指标集：$P^i=\{p_1^i,p_2^i,\cdots,p_k^i\}$，在 P^i 上抽取出统计量 S_i。训练一个预测模型，对特定的人脸 F_i，抽取特征 $\{f_1^i,f_2^i,f_3^i,\cdots,f_\xi^i\}$，并学习抽取的人脸特征到统计特征 S_i 间的映射：$\{f_1^i,f_2^i,f_3^i,\cdots,f_\xi^i\}\to S_i$。

3. 网络结构

3.1. 结构总览

如图 2 所示，本文提出的模型可以大致分为三个阶段：特征图提取、视觉词向量嵌入、自注意力计算、表示特征回归

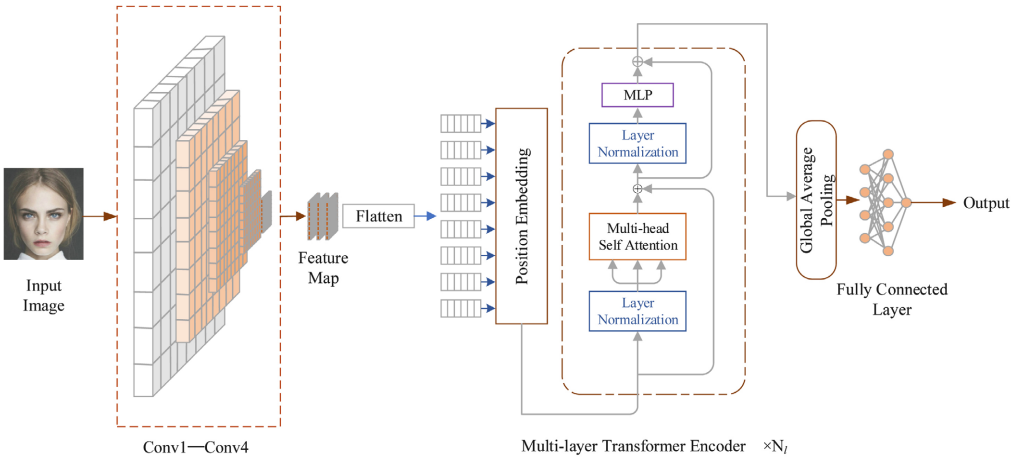

Figure 2. Overall structure of our model
图 2. 模型框架图

1) Backbone 部分，由预训练过的 Resnet-18 的前四个阶段组成，Conv_1 到 Conv_4 作为前置网络提取特征图。

2) 将提取的特征图在空间维度上展平并加上可学习的位置编码，再通过线性层映射到特定维度的嵌入空间，作为具有表征性的视觉词向量。

3) 利用多层的 Transformer 编码器模块，计算嵌入的词向量中各元素的自注意力。把握中不同视觉特征成分在全局上的相对关系。面部特征词嵌入中各元素间的复杂联系。

4) 最后的部分做一个全局平均池化后使用一个简单的全连接层计算回归输出。

3.2. Backbone 模块

本文采用在 MS-Celeb-1M 人脸检测数据集上预训练过的 Resnet-18 [13]的部分网络层作为 Backbone 来提取图片特征，对给定的输入为 224 × 224 × 3 的 RGB 人脸图像，经过如表 1 所示四个的阶段后，抽取出人脸特征图 $X_f \in \mathbb{R}^{H_f \times W_f \times C_f}$，其输出形状为[14, 14, 256]。

Table 1. Backbone structure of our model
表 1. 特征提取网络结构

网络层数	输出维度 $H \times W \times C$	网路结构
Conv_1	112 × 112 × 64	7 × 7，64，stride2 3 × 3 maxpooling，stride2
Conv_2	56 × 56 × 64	$\begin{pmatrix} 3\times3 & 64 \\ 3\times3 & 64 \end{pmatrix} \times 2$
Conv_3	28 × 28 × 128	$\begin{pmatrix} 3\times3 & 128 \\ 3\times3 & 128 \end{pmatrix} \times 2$
Conv_4	14 × 14 × 256	$\begin{pmatrix} 3\times3 & 256 \\ 3\times3 & 256 \end{pmatrix} \times 2$

3.3. Transformer 模块

Transformer [14]是 2017 年提出的一个自然语言处理框架，包括词嵌入、位置编码、编码器与解码器四个模块。一般地视觉任务中会将解码器模块去掉。

在本文中由上一级的残差网络抽取的二维特征图组 $X_f \in \mathbb{R}^{H_f \times W_f \times C_f}$，需要将其转化为一维的视觉词向量序列，这里将其展平为 $\overline{X}_f \in \mathbb{R}^{H_f W_f \times C_f}$ 后经由一个线性层 E 投射到 $X_p \in \mathbb{R}^{H_f W_f \times C_p}$，在这里 $C_f = 256$，$C_p = 384$。并加上一个一维的可学习位置嵌入向量 E_{pos}，由此得到视觉词嵌入向量序列：

$$Z_0 = \left[X_p^1; X_p^2; \cdots, X_p^{H_f W_f} \right] + E_{pos} \tag{3.1}$$

位置向量 E_{pos} 学习嵌入向量在每一个位置上的信息，最后生成的 Z_0 则代表了对位置敏感的特征序列，为了把握面部特征词嵌入中各元素间的复杂联系，将 Z_0 作为多层标准的 Transformer 编码器结构的输入。Transformer 编码器通过多头注意力模块(MSHA)来计算嵌入向量的权重，包含可学习的三个矩阵：查询矩阵 Q、键矩阵 K 和值矩阵 V，其中每一个注意力头的注意力计算表达式如式(0.1)：

$$\begin{aligned} head_i &= \text{Attention}\left(Q_i, K_i, V_i\right) \\ &= \text{Softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d}}\right) V_i \\ &= \text{Softmax}\left(\frac{Z_j W_i^Q \left(Z_j W_i^K\right)^{\mathrm{T}}}{\sqrt{d}}\right) Z_j W_i^V \end{aligned} \tag{3.2}$$

相较于单个注意力头的情况，多头注意力计算的实现采用多个查询矩阵、键矩阵、值矩阵将 Z_0 投影到 N_h 个不同的表示子空间：

$$\text{MHSA}\left(Z_j\right) = \text{Concat}\left(head_1, head_2, \cdots, head_{N_h}\right) W^o \tag{3.3}$$

其中的 $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{H_f W_f \times d}$，$i \in \{1,2,\cdots,N_h\}$，$d = \frac{C_p}{N_h}$，$N_h$ 表示不同注意力头的个数，每个头的 d 等于信道数除以头数 Z_j。最后通过拼接操作(Concat)将不同注意力头的输出整合成一个矩阵，整个注意力计算过程如图 3 所示。每个 Transformer 编码器由多个多头自注意力模块堆叠而成，这里 $j \in \{1,2,\cdots,N_l\}$，超参数 N_l 表示堆叠的个数。该结构输出如式(0.2)、(0.3)：

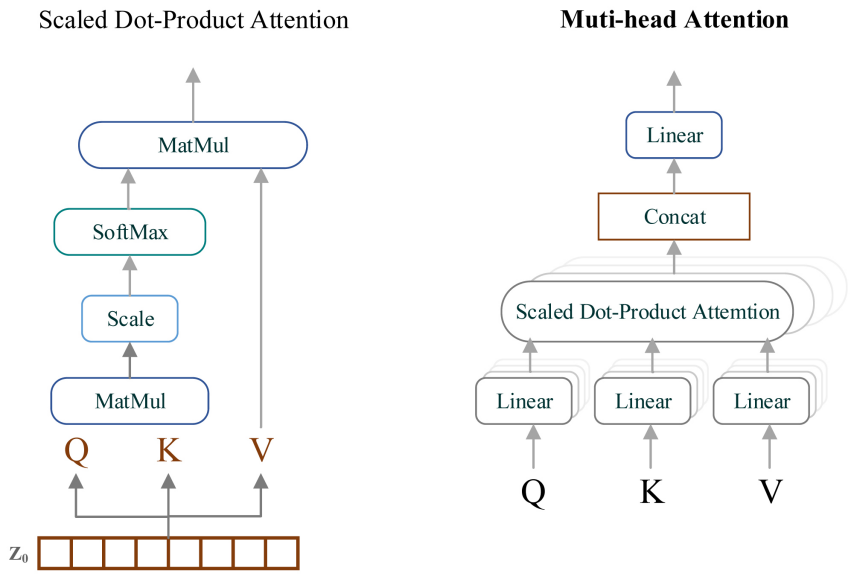

Figure 3. Diagram of the practical teaching system of automation major
图 3. 自注意力计算模块

$$\tilde{Z}_j = \text{MHSA}\left(\text{LN}\left(Z_{j-1}\right)\right) + Z_{j-1} \tag{3.4}$$

$$Z_j = \text{MLP}\left(\text{LN}\left(\tilde{Z}_j\right)\right) + \tilde{Z}_j \tag{3.5}$$

其中，$\tilde{Z}_j$，Z_j 表示在第 j 层中的中间输出与最终输出。LN 代表层正则化(Layer Normalization)。每个模块最后 MLP 层由两个前馈层和一个 GELU 非线性激活函数。输入层与输出层大小相同均为 384，隐藏层是输入层的四倍，这里设置为 1536。

对最后的输出 Z_{Nl} 进行一个全局池化操作后，再输入到一个隐藏层为 5 输出层为 1 的全连接层，其设置目的是为了更好地归纳 5 个类级的权重，最终得到输出结果。

4. 实验与分析

4.1. 数据集及预处理

本文的实验数据集来源于华南理工大学 2018 年制作的 SCUT-FBP5500 数据集[15]，这是一个多属性多范式的数据集，包含有 5500 张中性表情的正脸图像，共分为四个子集：亚洲男性、亚洲女性图像各 2000 张，高加索男性、高加索女性图像各 750 张。每张图片都对应有 60 个志愿者的评分，评分范围为{1, 2, 3, 4, 5}。其中 5 表示吸引力最高的指标，其余依次递减。图 4 为数据集上的标签分布情况。

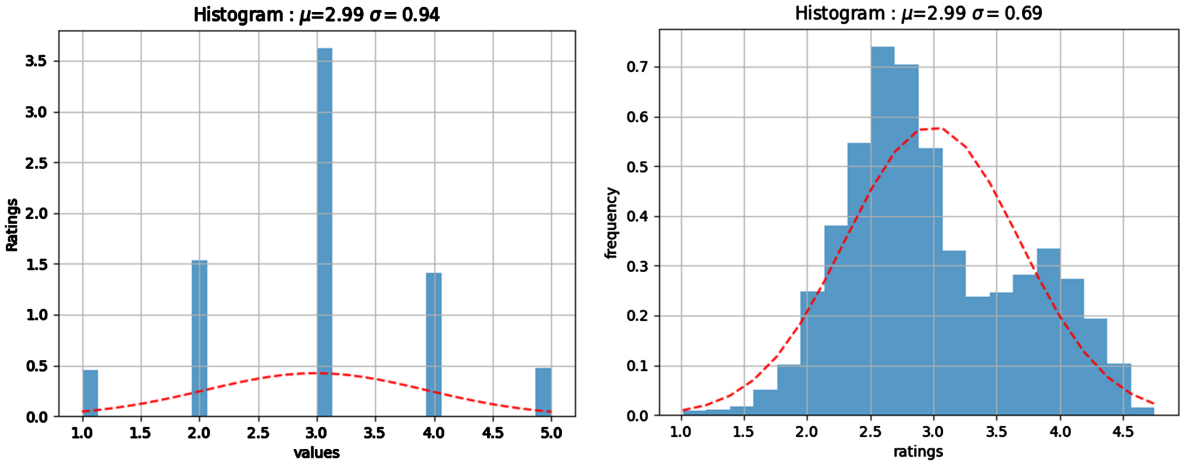

Figure 4. Label distribution on SCUT-FBP5500
图 4. 数据集上的标签分布

本文以每张图像的所有评分的均值作为其对应标签。其分布情况如图 4。基于 SCUT-FBP5500 数据库中包含有每张图像上标注的 86 个面部特征点数据，这里采用第 44 号，第 52 号(左右眼中点)和第 75 号点(上唇中点)的坐标作对齐处理，归一化至 224 × 224 大小。将训练集、验证集、测试集按 60%、20%、20%的比例做随机划分。由于数据集样本量较小，为了避免的过拟合，采用左右翻转的方式将训练集、测试集、验证集扩充一倍。

4.2. 评价指标

本文将面孔吸引力评估预测视作回归问题，采用三个回归模型评价指标：皮尔逊相关系数(PC)，平均绝对值误差(MAE)，均方根误差(RSME)。皮尔逊相关系数是用以刻画真实值与预测值的相关程度的指标，其取值范围为[−1, 1]，越接近于 1 表示相关性越强。而平均绝对值误差，均方根误差表预测值与真实值的接近程度，用以刻画模型的拟合质量，它们的值越接近于零效果越好。其计算公式如下：

$$\mathrm{PC}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\left(Y_i-\bar{Y}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}} \tag{4.1}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|X_i-Y_i\right| \tag{4.2}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2} \tag{4.3}$$

4.3. 实验设置

本次实验所用环境为谷歌实验室 Colar，显卡为 Nvidia TeslaK80，显存 11 G，深度学习框架为 Pytorch1.5 + cuda10.1，优化器选用 Adam，学习率初始时设置为 5×10^{-4}，batch_size 设置为 50，通过在验证集上的实验筛选出合适的超参数.前 10 轮训练中冻结特征提取网络的参数，后面轮次的训练对整个网络的参数进行微调。

4.4. 实验结果与分析

经过实验验证，最终选定超参数 $N_l = 3$，$N_h = 8$，N_l 会显著地提升参数量，过高会导致过拟合，N_h 较小则表现较差，因为没有足够的子空间来学习潜在的特征信息。为了验证本文模型的有效性，与该数据集上的其它模型作对比，本文深度学习模型在皮尔逊相关系数、平均绝对值误差、均方根误差三个指标上均取得到了提升。如表 2 所示。其中 Psychological inspired CNN [6]是一种采用级联微调的方法优化，将输入的图像提取色彩、纹理、几何成分融合输出的模型，ResNeXt-50 based R^3CNN [7]是一种通过排序信息引导的 CNN 网络。

Table 2. Experiment results comparison
表 2. 实验结果对比

在该数据集上的模型	PC	MAE	RMSE
Geometric features + Gassian Regression	0.7472	0.3554	0.4599
ResNest-18	0.8513	0.2818	0.3703
ResNeXt-50	0.8777	0.2518	0.3325
Psychological inspired CNN	0.8978	0.2267	0.3016
ResNeXt-50 based R^3CNN	0.9142	0.2120	0.2800
Ours Model	**0.9273**	**0.2014**	**0.2740**

5. 总结

本文将多头自注意力模块应用在面部属性感知分析的面孔吸引力预测方向，提出了一种结合了残差网络结构与 Transformer 编码器结构的混合模型。该模型先使用残差网络作为 Backbone 提取特征图，再将特征图转化为视觉词嵌入向量，通过多头自注意力机制模块计算各元素间位置信息的长依赖关系，从全局的建模方式来把握各层特征的相对联系。最后通过一个池化层与全连接层来获得回归输出。为了验证了该混合模型的可行性与有效性，在 SCUT-FBP5500 数据集上进行实验，取得了较好的效果。

参考文献

[1] Sala, E., Terraneo, M., Lucchini, M. and Knies, G. (2013) Exploring the Impact of Male and Female Facial Attractiveness on Occupational Prestige. *Research in Social Stratification and Mobility*, **31**, 69-81. https://doi.org/10.1016/j.rssm.2012.10.003

[2] Eisenthal, Y., Dror, G. and Ruppin, E. (2006) Facial Attractiveness: Beauty and the Machine. *Neural Computation*, **18**, 119-142. https://doi.org/10.1162/089976606774841602

[3] Kagian, A., Dror, G., Leyvand, T., Cohen-Or, D. and Ruppin, E. (2006) A Humanlike Predictor of Facial Attractiveness. *Advances in Neural Information Processing Systems* 19: *Proceedings of the* 2006 *Conference*, 649-656.

[4] Schmid, K., Marx, D. and Samal, A. (2008) Computation of Face Attractiveness Index Based on Neoclassic Canons, Symmetry and Golden Ratio. *Pattern Recognition*, **41**, 2710-2717. https://doi.org/10.1016/j.patcog.2007.11.022

[5] Xu, J., Jin, L., Liang, L., Feng, Z. and Xie, D. (2015) A New Humanlike Facial Attractiveness Predictor with Cascaded Fine-Tuning Deep Learning Model. arXivpreprint arXiv:1511.02465.

[6] Xu, J., Jin, L., Liang, L., Feng, Z., Xie, D. and Mao, H. (2017) Facial Attractiveness Prediction Using Psychologically Inspired Convolutional Neural Network (PI-CNN). 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orlean, 5-9 March 2017, 1657-1661. https://doi.org/10.1109/ICASSP.2017.7952438

[7] Lin, L.J., Liang, L.Y. and Jin, L.W. (2019) Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction. *IEEE Transactions on Affective Computing*.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

[9] Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P. and Vaswani, A. (2021) Bottleneck Transformers for Visual Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 16519-16529. https://doi.org/10.1109/CVPR46437.2021.01625

[10] Dai, Z., Liu, H., Le, Q. and Tan, M. (2021) Coatnet: Marrying Convolution and Attention for All Data Sizes. *Advances in Neural Information Processing Systems*, **34**, 3965-3977.

[11] Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A. and Van Gool, L. (2021) Transformer in Convolutional Neural Networks. arXiv preprint arXiv:2106.03180.

[12] Fan, Y.Y., Liu, S., Li, B., Guo, Z., Samal, A., Wan, J., *et al*. (2017) Label Distribution-Based Facial Attractiveness Computation by Deep Residual Learning. *IEEE Transactions on Multimedia*, **20**, 2196-2208. https://doi.org/10.1109/TMM.2017.2780762

[13] Wu, Y., Li, J., Kong, Y. and Fu, Y. (2016) Deep Convolutional Neural Network with Independent Softmax for Large Scale Face Recognition. *Proceedings of the* 24*th ACM international conference on Multimedia*, Amsterdam, 15-19 October 2016, 1063-1067. https://doi.org/10.1145/2964284.2984060

[14] Vaswani, A., Shazeer, N., Parmar, N., (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5998-6008.

[15] Liang, L., Lin, L., Jin, L., Xie, D. and Li, M. (2018) SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. 2018 24*th International Conference on Pattern Recognition* (*ICPR*), Beijing, 20-24 August 2018, 1598-1603. https://doi.org/10.1109/ICPR.2018.8546038