Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network

汇报人：王李悦

- 研究简介

- 3D人脸重建和人脸对齐是计算机视觉中两个基本且高度相关的主题。在过去的几十年中，这两个领域的研究相互受益。

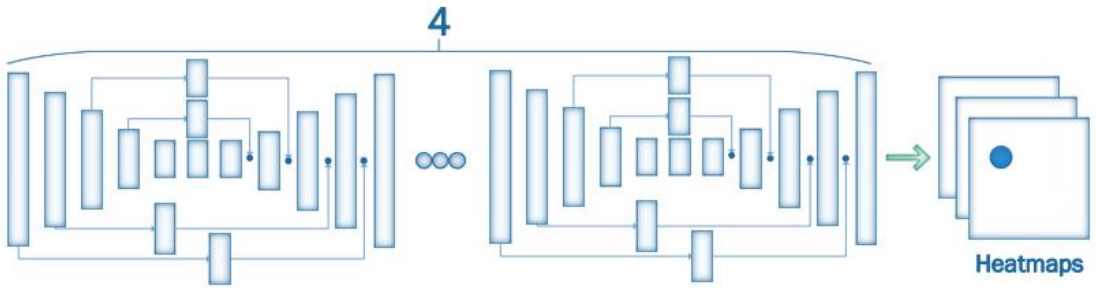
- 运用CNN来依据单张图片估计3DMM模型系数受限于空间基础模型的限制

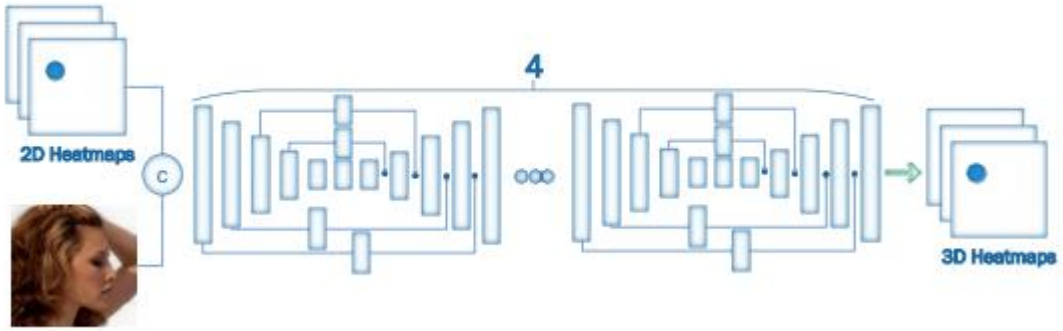
研究简介

- How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)
- ICCV 2017

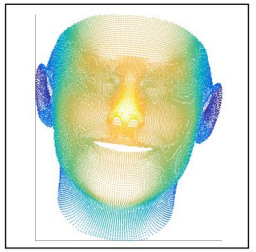
- 根据单个图像回归68个2D面部坐标
- 根据单个图像与2D面部坐标回归3D面部坐标
- 提供数据库

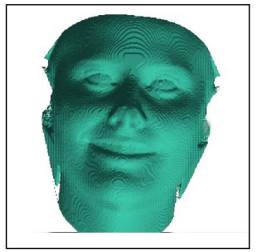
- 回归网络复杂（Hourglass）
- 无密集对齐
- 需要额外的深度值估计网络

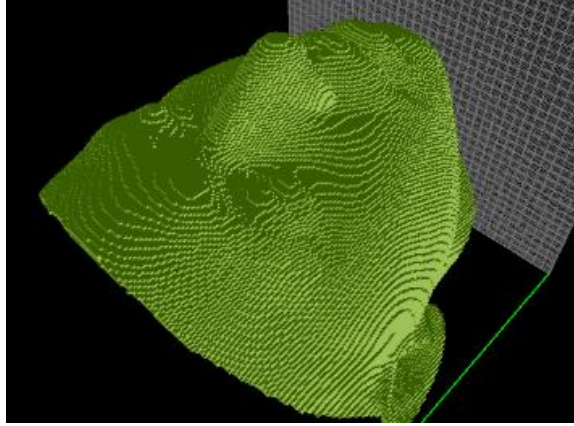




- 研究简介



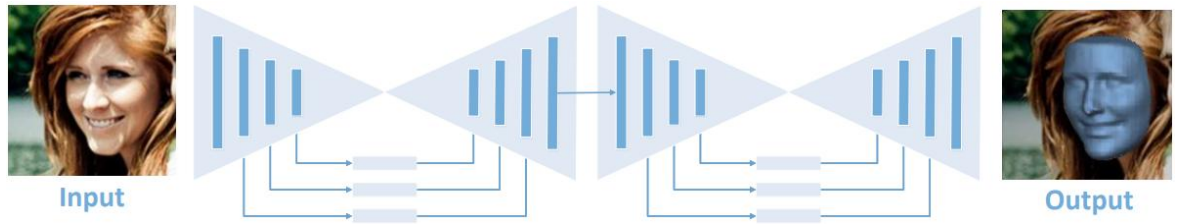


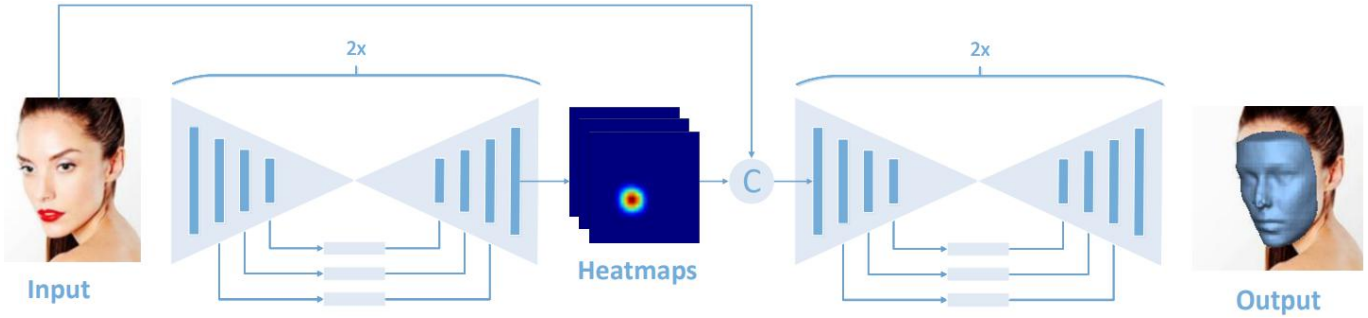


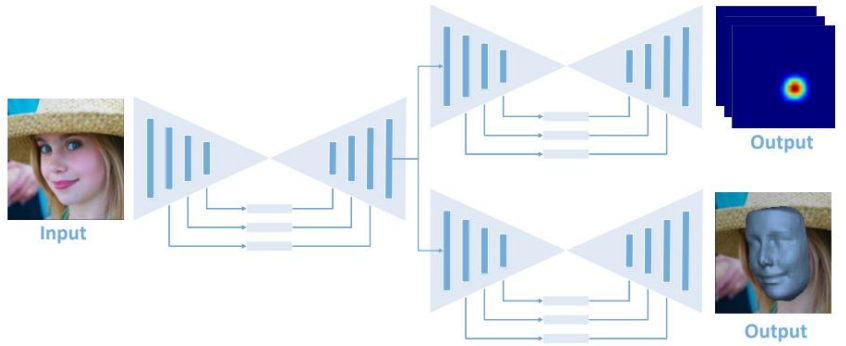
- Large pose 3d face reconstruction from a single image via direct volumetric

- 将mesh转换为voxel

- 丢弃了2D坐标点的语义信息
- 重构结果的分辨率受限
- 回归网络复杂







研究简介

- 本文提出了一种端到端的网络：Position map Regression Network (PRN)
- 3D人脸重建方面，本文方法框架模型自由，采用轻量级模式
- 人脸对齐方面，本文方法通过位置图，直接完成整张人脸与3D模板之间的稠密关系回归而无需3DMM系数和TPS Warping参数，无需复杂透视变换与TPS变换
- 将密度对齐与3D面部重构同时实现
- 超越了所有其他同时期的先进方法

3D人脸重构

- 我们任务目的是从单2D人脸图像回归出人脸3D几何结构信息以及它们之间的稠密关系。所有我们需要适合的表示方法，并用一个网络来直接预测
- 简单且普遍的方法是用一个向量来表示，即将3D点信息用一个向量来表示，然后用网络预测，但这种方法丢失了空间信息
- 相关研究中也提到了类似于3DMM等模型表示，但这些方法太过依赖3DMM模型
- 最近的VRN用Volumetric来表示，成功摆脱了上述问题，但是其网络需要输出192x192x200的一个Volume，计算量相当大，重建分辨率将会受到限制。

- 3DMM模型

- 3DMM模型表示

$$S_{model} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i S_i$$

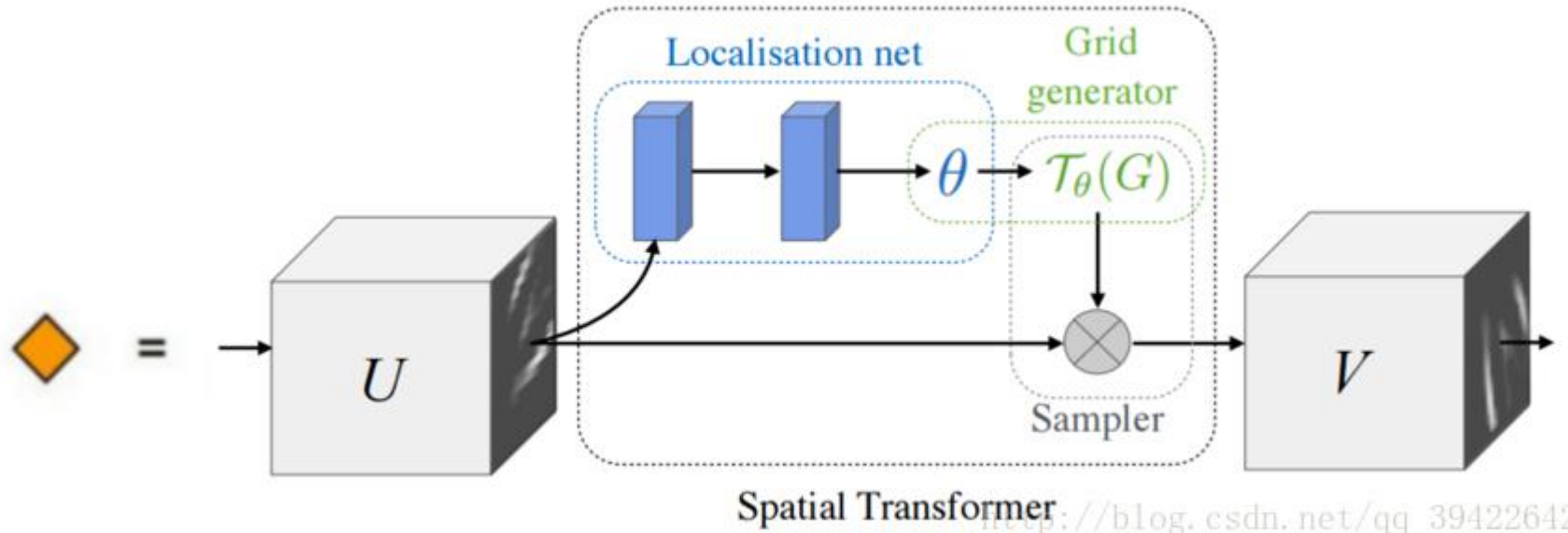
$$T_{model} = T^2 + \sum_{i=1}^{m-1} \beta_i T_i$$

$$S_{newModel} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i s_i + \sum_{i=1}^{n-1} \beta_i e_i$$

- 3D人脸重构

- 本文的训练数据来源是开源数据集，如300W-LP，但数据集的ground-truth是没有用UV表示的。因此训练用到的UV标签是基于3DMM的，最终仍然与BFM建立联系来建立。
- 当然作者也提出，可以采用其他的数据集来训练，那样的话可能就不需要3dmm模型等相关的处理技巧，反正无论如何，流程都是：得到3D点云-> 3DMMasSTN插值形成uv位置图->训练网络这个步骤。只是得到3d点云的方式不同而已。

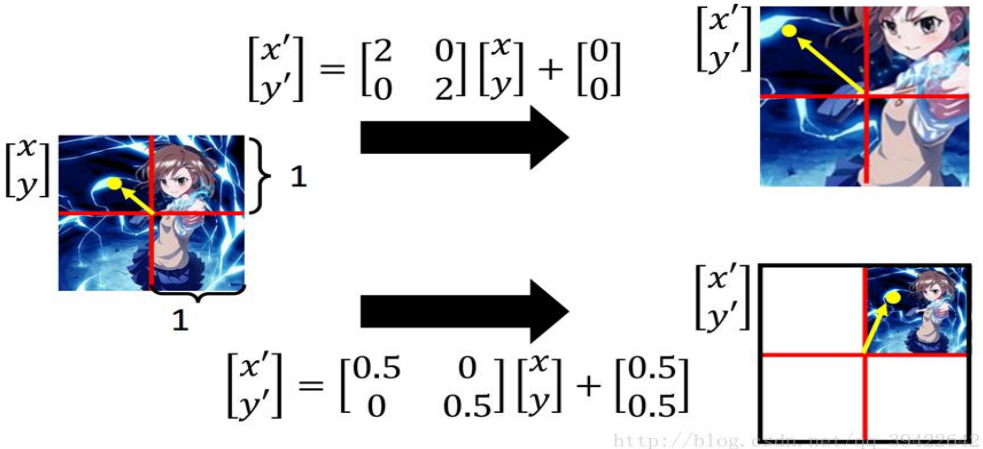
- STM网络架构

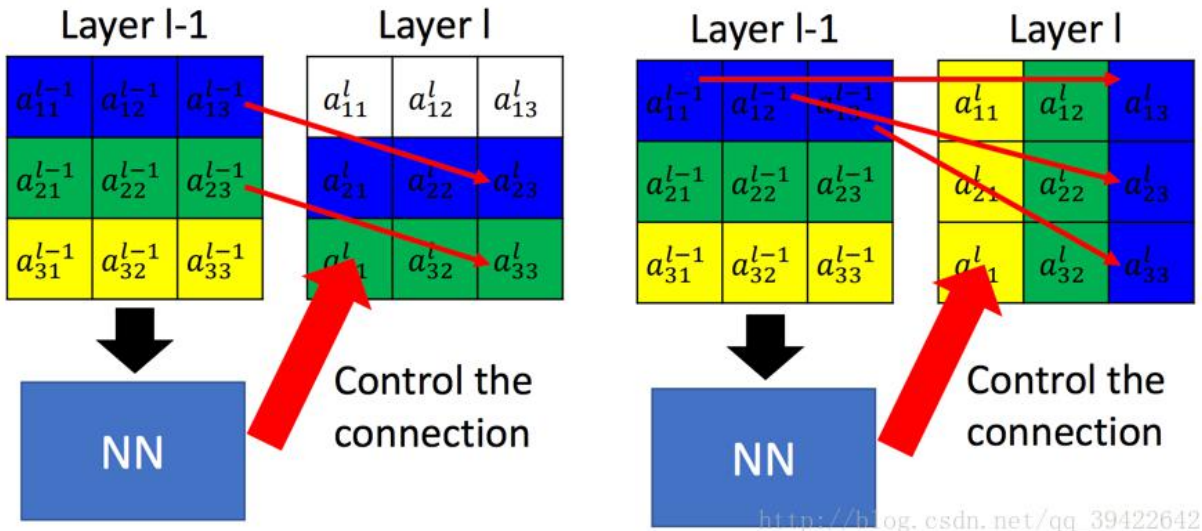


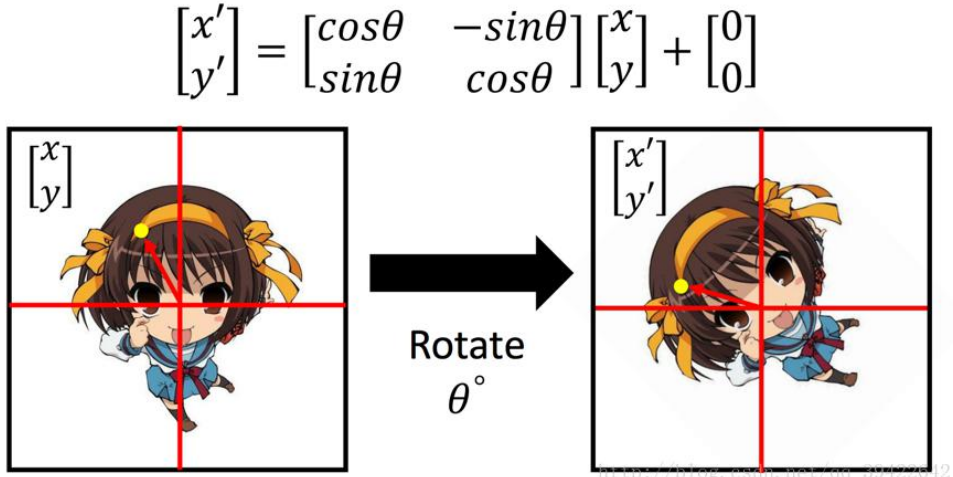
- U为原始输入图片，V为输出图片
- 该网络的主要部分有三个：1.参数预测：Localisation net　2.坐标映射：Grid generator
  3.像素的采集：Sampler

- STM-Localisation net

- 通过卷积或全连接等过程回归出参数θ用于下一步计算



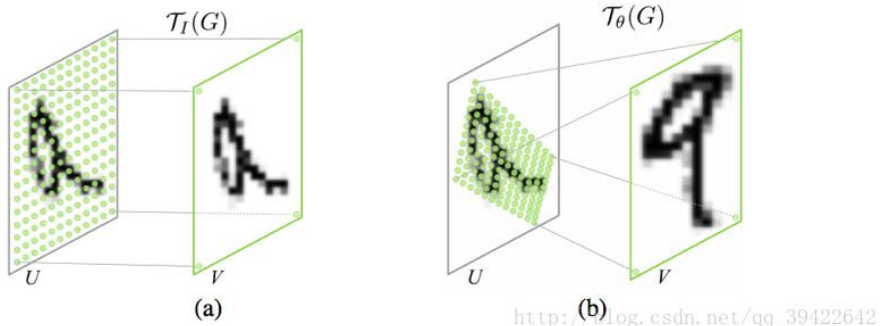




- STM-Gred generator

- 进行目标图片->原始图片的映射

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$



- (xti,yti) 是输出的目标图片的坐标
- (xsi,ysi)是原图片的坐标
- Aθ表示仿射关系

- 3DMM-STM

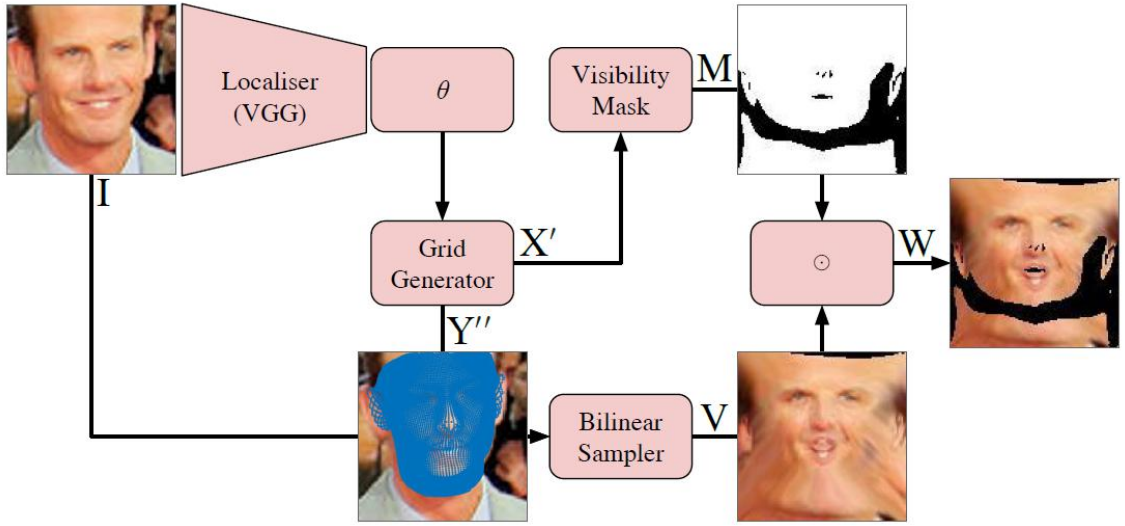
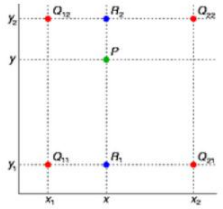


Figure 1. Overview of the 3DMM-STN.

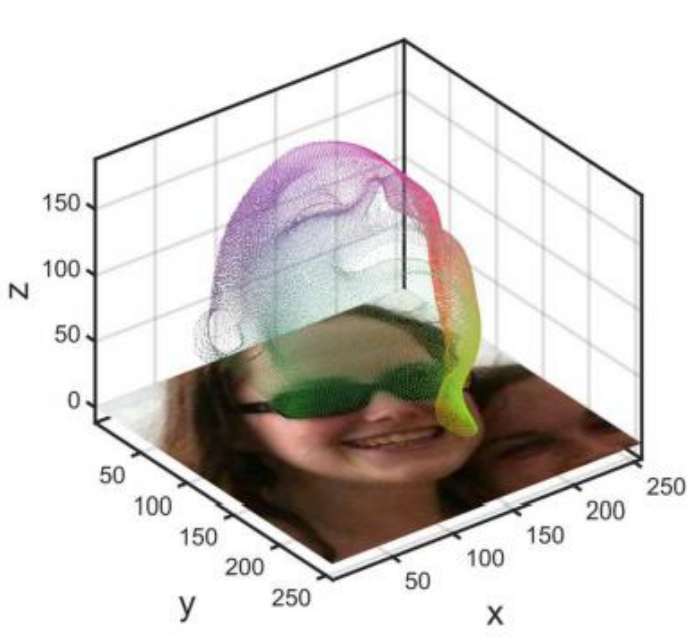


假如我们想得到未知函数 f 在点 P = (x, y) 的值，假设我们已知函数 f 在 Q11 = (x1, y1)、Q12 = (x1, y2), Q21 = (x2, y1) 以及 Q22 = (x2, y2) 四个点的值。**最常见的情况，f就是一个像素点的像素值**。首先在 x 方向进行线性插值，得到

$$f(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \quad \text{where} \quad R_1 = (x, y_1),$$

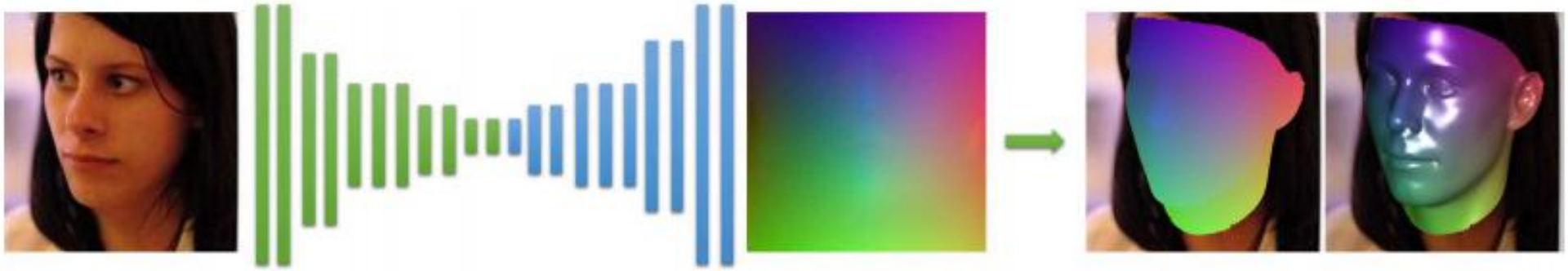
$$f(R_2) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \quad \text{where} \quad R_2 = (x, y_2).$$

- UV position map　　a 2D image recording 3D positions of all points in UV space.



- 第一排第一张是原始图片，第二张是将人脸RGB通道映射到UV空间的对应位置的UV空间纹理图，第一排第三张是UV空间位置图(UV position map)，每个通道表示对应纹理的xyz。
- 第二排是 UV空间的位置图三个通道的展开。

- 网络结构



本文encoder-decoder网络基于tensorflow实现，共包含1层卷积层，10层resblock残差块（encoder）和17层转置卷积(decoder)，其中中间各层卷积、转置卷积后跟着batchnorm和relu，最后一层转置卷积的激活采用Sigmoid。encoder中各层的kernel均为4，out channel从16到512，输出尺寸从256到8(每两个resblock尺寸减半)，decoder则反之。输入和输出均为256x256 RGB图像，其中输入是人脸图像，输出是3D点云坐标（共65536个）。最后分别从中提起相应点的2d、3d坐标来进行3D重建或人脸关键点。

- 损失函数与 weight mask

均方误差是这类研究常用的损失函数，然而每个点在其中的权重是相等的，该函数平等对待每一个点，这对于位置图的学习是不恰当的。鉴于面部中心位置的点有更大的学习意义，该研究对不同位置的点进行了不同权重的判定。

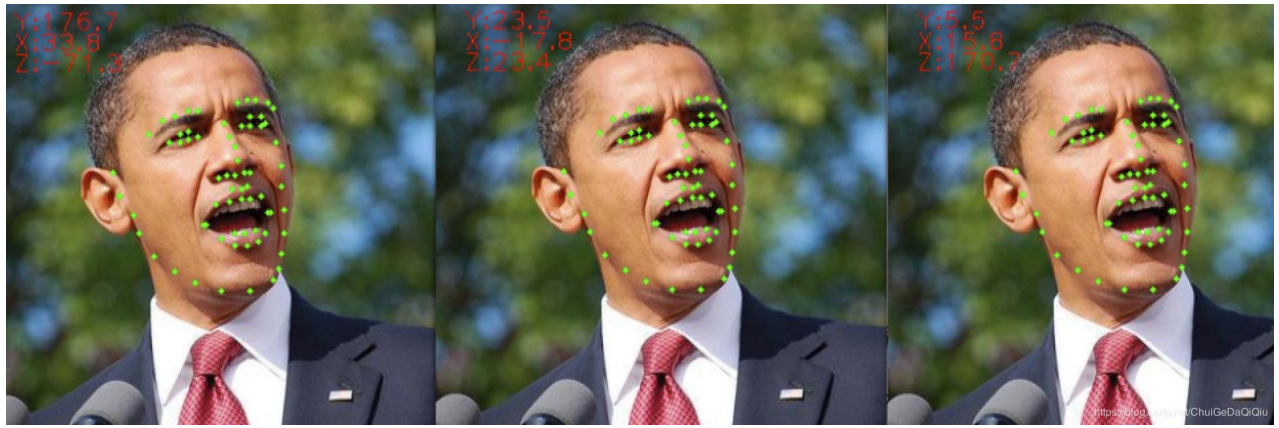
其将所有点分为4大类：

1.68个关键点

2.眼睛鼻子嘴巴

3.面部的其他区域

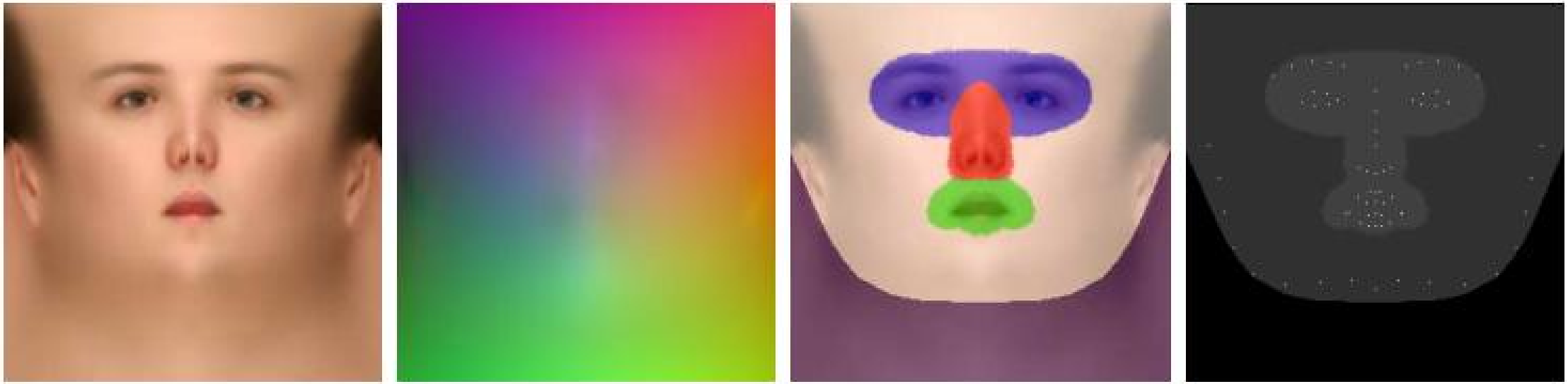
4.脖子



其中将68个面部关键点设置为第一序列是为了确保网络能够准确到这些关键点，而脖子在不同图片上可能被头发或衣服等遮挡，因此设置为第四等级。

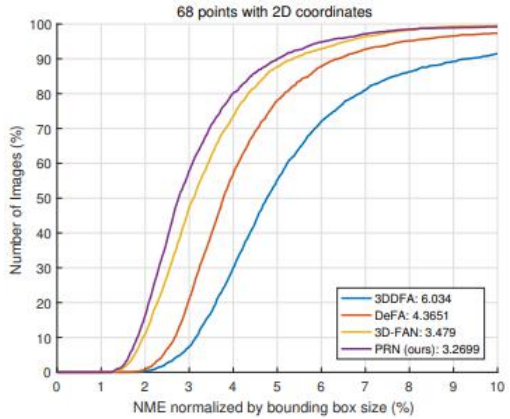
- 损失函数与 weight mask

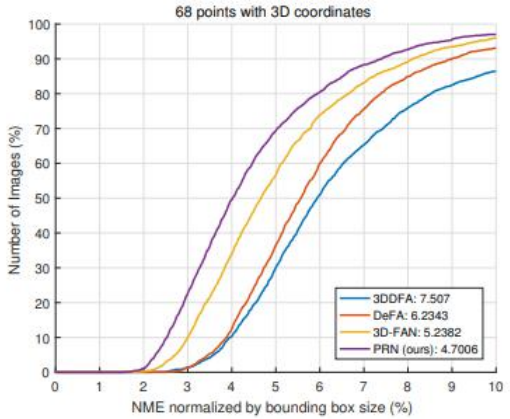
研究提出的weight mask如下图所示：

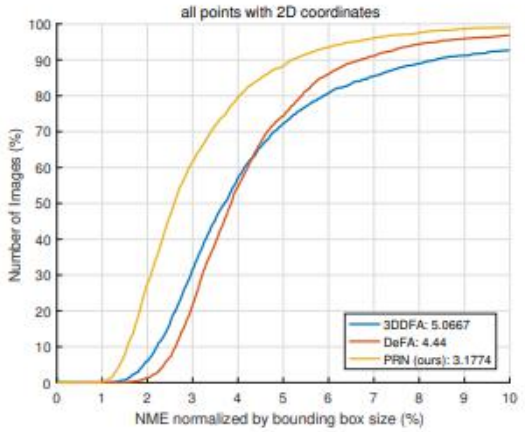


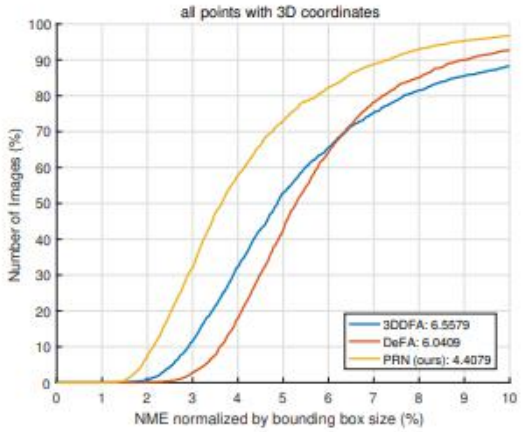
从左到右依次是UV纹理图、UV位置图，根据分级标准上色的纹理图，最终的weight mask灰度图。

- 密集对齐实验结果


68 points with 2D coordinates
Number of Images (%)
NME normalized by bounding box size (%)
3DDFA: 6.034
DeFA: 4.3651
3D-FAN: 3.479
PRN (ours): 3.2699


68 points with 3D coordinates
Number of Images (%)
NME normalized by bounding box size (%)
3DDFA: 7.507
DeFA: 6.2343
3D-FAN: 5.2382
PRN (ours): 4.7006

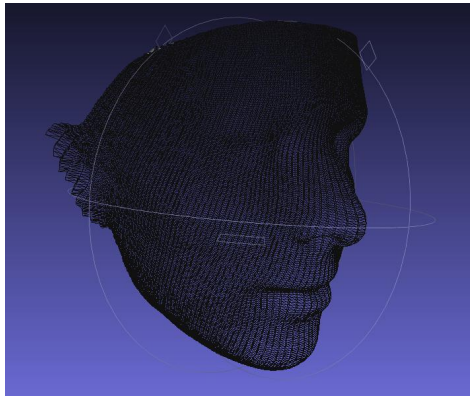

all points with 2D coordinates
Number of Images (%)
NME normalized by bounding box size (%)
3DDFA: 5.0667
DeFA: 4.44
PRN (ours): 3.1774

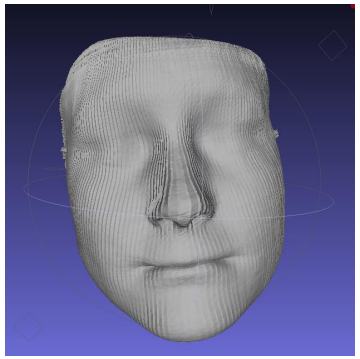

all points with 3D coordinates
Number of Images (%)
NME normalized by bounding box size (%)
3DDFA: 6.5579
DeFA: 6.0409
PRN (ours): 4.4079

- 3D人脸重构实验结果

Fig. 9: Left: CED curves on Florence dataset with different yaw angles. Right: the qualitative comparison with VRN-Guided. The first column is the input images from Florence dataset and the Internet, the second column is the reconstructed face from our method, the third column is the results from VRN.

- demo





THANKS