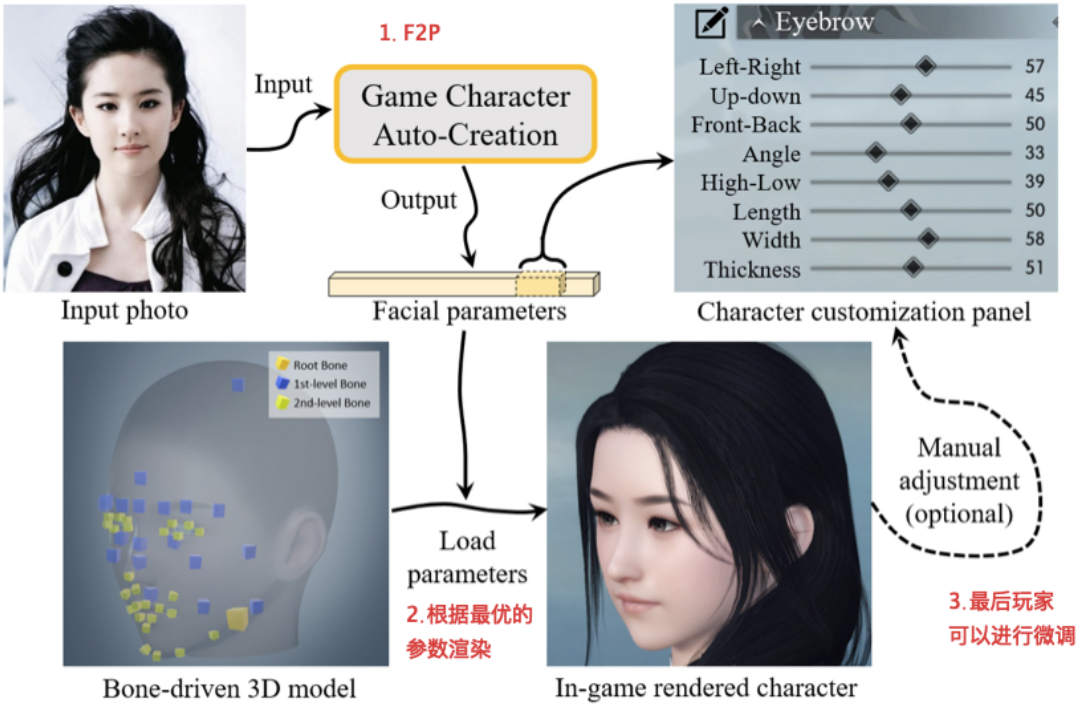
Face-to-Parameter Translation for Game Character Auto-Creation

本篇论文主要是解决 RPG 游戏中人物角色的自动创建，即根据玩家自定义的一张 2 维真实脸部图片生成游戏人物的 3 维面部模型。减少 role-playing games (RPGs) 游戏中角色头像个性化时参数设置的繁琐步骤 （脸型，发型，肤色等）。

一、基本流程

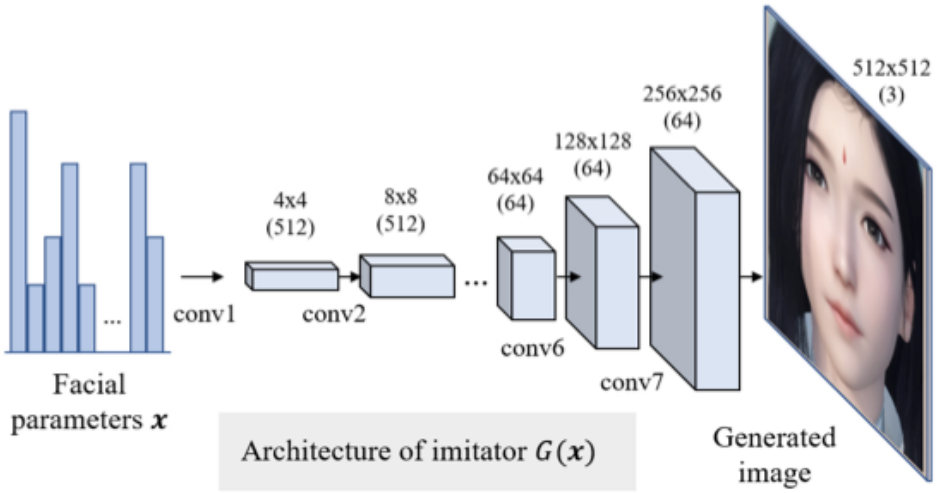
因为游戏引擎可以生成最终的游戏人物面部模型， 所以只能对其输入的参数进行优化，为了使其生成的面部模型最接近用户输入的真实面部图片，需要研究的问题变为 Face-to-Parameter (F2P)



二、主要模型

Imitator G(X)

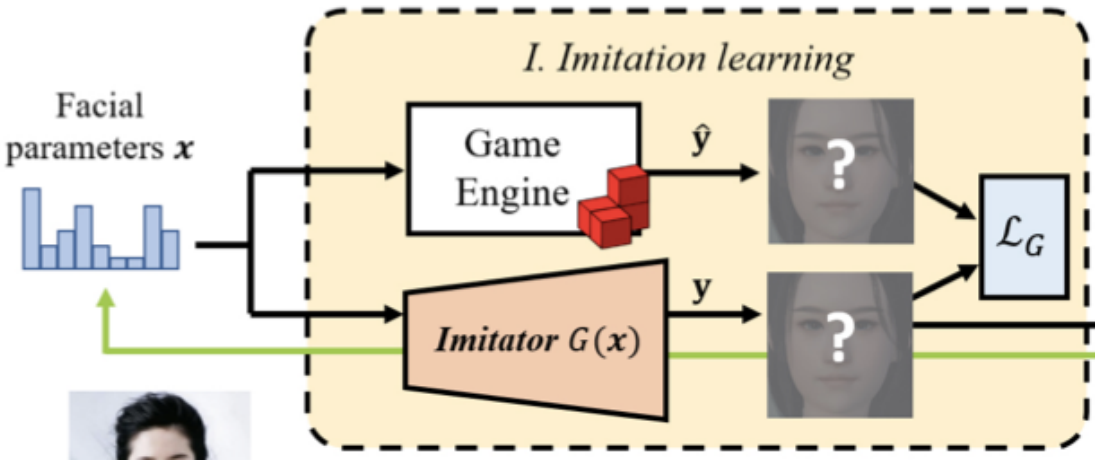
用于模仿游戏引擎渲染出的人物角色的面部图片，输入 X 是和游戏引擎的输入一样为一些代表的脸部特征的个性化参数。



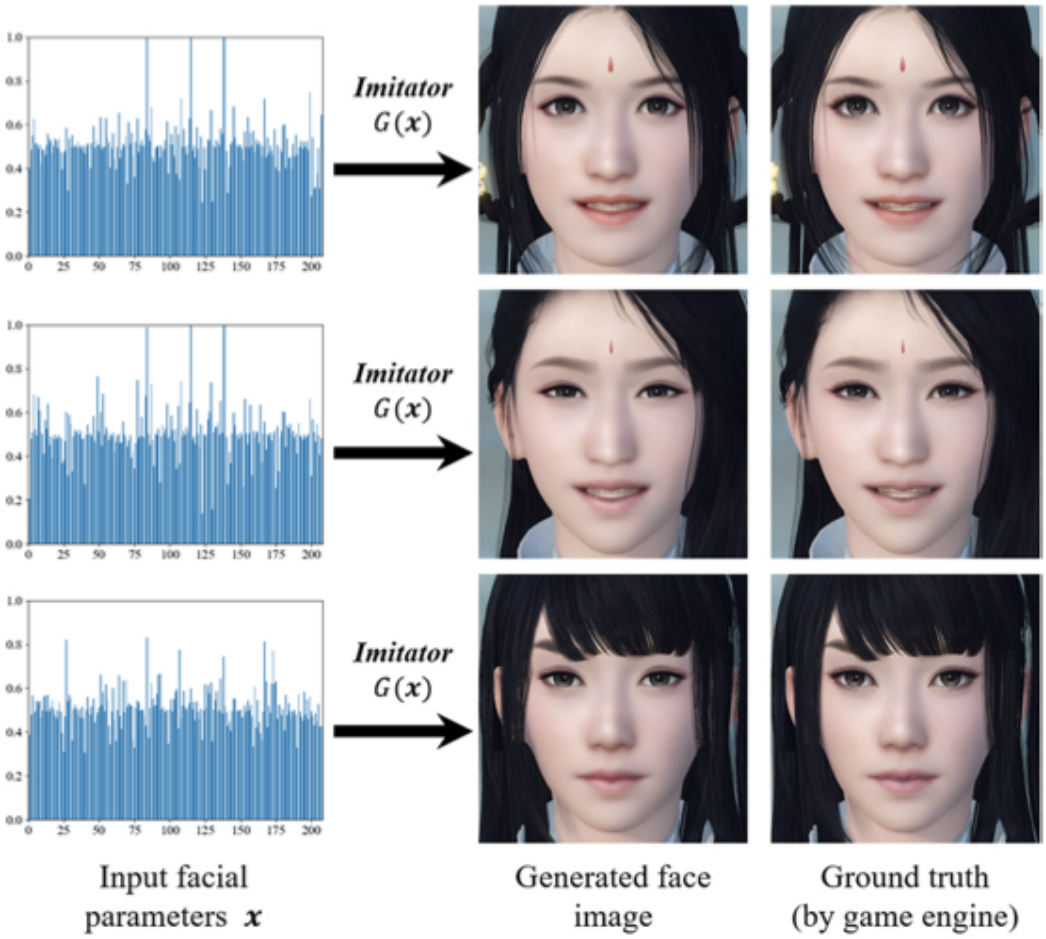
网络结构类似于 DCGAN，训练时随机产生 20,000 组参数，然后将游戏引擎生成的图片作为 groundtruth，生成器的 loss 为 G(x) 和 groundtruth 的 l1 loss，从而进行回归

$$\begin{aligned}\mathcal{L}_G(\boldsymbol{x}) &= E_{\boldsymbol{x}\sim u(\boldsymbol{x})}\{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|_1\} \\ &= E_{\boldsymbol{x}\sim u(\boldsymbol{x})}\{\|G(\boldsymbol{x})-\text{Engine}(\boldsymbol{x})\|_1\},\end{aligned}$$

$$G^\star = \arg\min_G \mathcal{L}_G(\boldsymbol{x}).$$



训练后的效果



Feature Extractor F(y)

将不同 domain 的图片（真实人脸图片和 G(x) 生成的游戏人脸图片）映射到同一特征空间中，从而衡量两张图片的相似性，在其基础上定义定义两种损失。

Discriminative Loss

一种是 Discriminative Loss，保证两张图片的全局特征相似。使用 Light CNN-29 v2 面部识别模型 F1，可以提取图片中的 256 个特征，然后计算其 cosine 距离。

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{x}, \boldsymbol{y}_r) &= 1 - \cos(F_1(\boldsymbol{y}), F_1(\boldsymbol{y}_r)) \\ &= 1 - \cos(F_1(G(\boldsymbol{x})), F_1(\boldsymbol{y}_r)),\end{aligned}$$

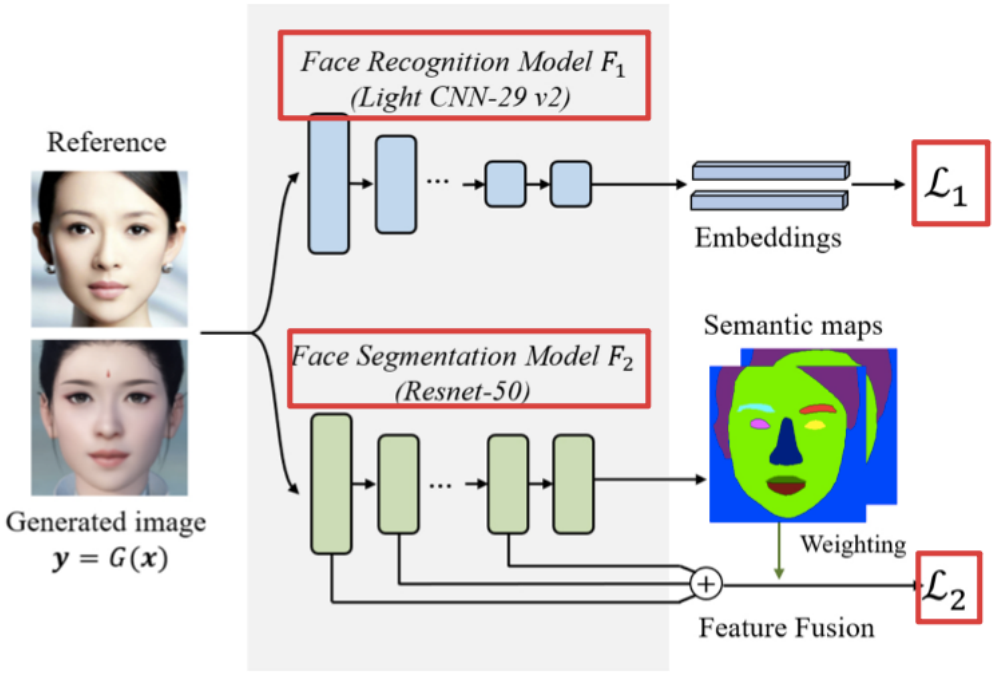
$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{< \boldsymbol{a}, \boldsymbol{b} >}{\sqrt{\|\boldsymbol{a}\|_2^2 \|\boldsymbol{b}\|_2^2}}.$$

yr 为玩家输入的真实 2 维图片

Facial Content Loss

另一个是 Facial Content Loss，计算 pixel-wise 的误差，使用的特征提取模型是基于 Resnet-50 的脸部分割网络，记作 F2，在 Helen 人脸数据集上进行训练，并且为了增加位置敏感性，使用分割得到的结果 class-wise probability maps 作为 feature maps 的权重从而得到 loss 函数：

$$\mathcal{L}_2(\boldsymbol{x}, \boldsymbol{y}_r) = \|\omega(G(\boldsymbol{x}))F_2(G(\boldsymbol{x})) - \omega(\boldsymbol{y}_r)F_2(\boldsymbol{y}_r)\|_1,$$

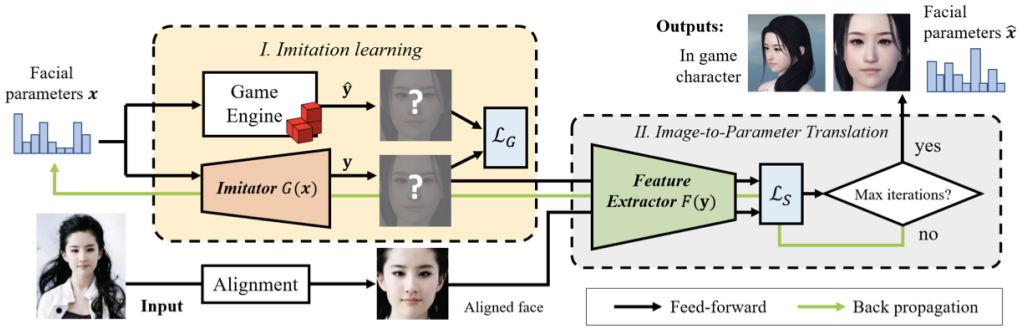


最终的损失函数为两个 loss 的加权和

$$\begin{aligned}\mathcal{L}_S(\boldsymbol{x}, \boldsymbol{y}_r) &= \alpha\mathcal{L}_1 + \mathcal{L}_2 \\ &= \alpha(1 - \cos(F_1(G(\boldsymbol{x})), F_1(\boldsymbol{y}_r))) \\ &+ \|\omega(G(\boldsymbol{x}))F_2(G(\boldsymbol{x})) - \omega(\boldsymbol{y}_r)F_2(\boldsymbol{y}_r)\|_1,\end{aligned}$$

完整的优化过程为：
Stage 1：训练 imitator G，面部识别网络 F1，面部分割网络 F2
Stage 2：保持训练好的 G，F1，F2 不变，初始化面部参数 x，前向计算得到 G(x)，然后将其和玩家输入的真实图片 yr 分别输入到 F1 和 F2 中计算得到 Ls 损失，使用梯度下降法和链式法则优化 x，直到达到最大迭代次数后停止，得到最终的 x。

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \mu \frac{\partial \mathcal{L}_S}{\partial \boldsymbol{x}} \text{ } (\mu\text{: learning rate}).$$
$$\text{Project } x_i \text{ to } [0,1]\text{: } x_i \leftarrow \max(0, \min(x_i, 1)).$$



三、Ablation Studies

1. Discriminative Loss

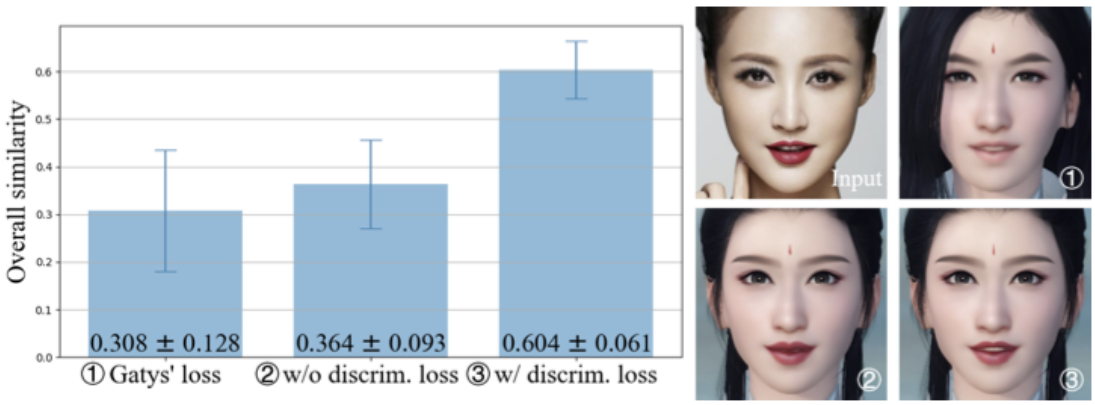
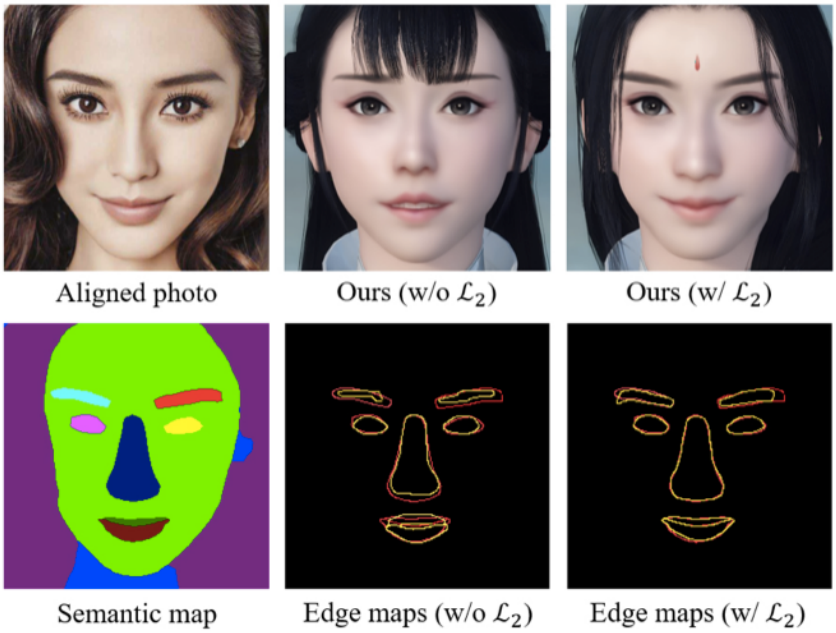


Figure 8. A performance comparison between different objective functions.

2. Facial content loss



3. Subjective evaluation

Ablations		
Discrim. $\mathcal{L}_1$	Facial-Sem. $\mathcal{L}_2$	Selection Ratio
✓	×	$13.47\% \pm 0.38\%$
×	✓	$36.27\% \pm 0.98\%$
✓	✓	$\mathbf{50.26\% \pm 0.40\%}$

Table 1. Subjective evaluation results of two technical components of our method 1) discriminative loss $\mathcal{L}_1$, 2) facial content loss $\mathcal{L}_2$ on our dataset. A higher selection ration indicates better.

四、结果比较

1. 和 NST 方法进行比较：
   使用同一个性别中所有人脸的平均图像作为 style reference，然后使用不同的 neural style transfer methods （global style method and local style method）进行生成。发现不适用于游戏中的人物模型生成。

2. 和 monocular 3D face reconstruction 方法进行比较：
   使用 3DMM-CNN，发现只能生成相似的面部轮廓。

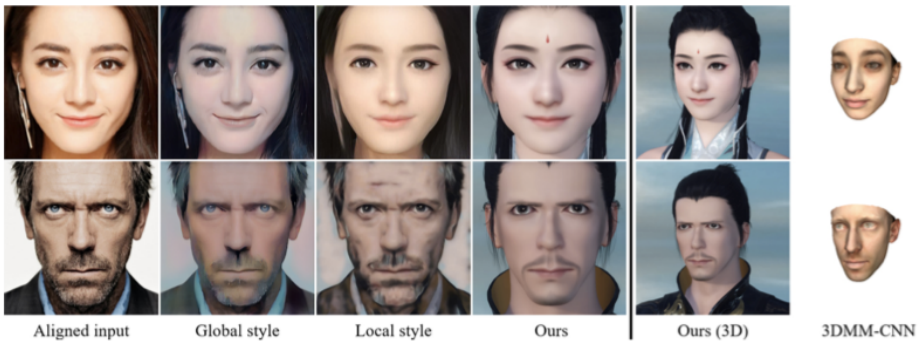


Figure 10. A comparison with other NST methods: Global style [12] and Local style [16], and we use the "average face" of each gender as the style reference of these NST methods. We also compare with a popular monocular 3D face reconstruction method: 3DMM-CNN [39].

3. 量化指标

Table 2. The style similarity and speed performance of different methods. (A higher Mode Score or a lower FID indicates better)

Method	Global style [12]	Local style [16]	3DMM-CNN [39]	Ours
Mode Score	1.0371 ± 0.0134	1.0316 ± 0.0128	–	$\mathbf{1.1418 \pm 0.0049}$
Fréchet Inception Distance	0.0677 ± 0.0018	0.0554 ± 0.0025	–	$\mathbf{0.0390 \pm 0.0018}$
Time (run on TITAN Xp)	22s	43s	15s	16s

五、优势与劣势

对于模糊和光线不足的输入图片有较好的鲁棒性，但对于不同姿势的面部图片则较为失败，因为 Facial Content Loss 定义在局部。

对于除了真实图片之外的其他图片，如素描和卡通图像，仍然可以作为输入的索引，因为相似性的比较不是基于 pixel，而是基于面部特征。

Input sketch image Generated character Input Caricature Generated character