arXiv:2611.32232v1 [cs.CV] 24 Nov 2026

MYSTIQUE：少样本人脸身份迁移与性能保持

ShihanCheng[1] David Hyde[1] Junting Deng[2] Michael Bao[3]Paul Debevec[3]
RoshniCooper[3] Soshi Shimada[31]范德比尔特大学 [2]卡内基梅隆大学 [3]
Netflix{shihan.cheng, david.hyde.1}@vanderbilt.edu juntingd@andrew.cmu.edu
{mbao, debevec, roshnic, soshis}@netflix.com



图 1。**使用 MYSTIQUE 进行身份迁移。** MYSTIQUE 从目标人物照片中提取身份，同时保留驱动图像中的表情、光照和遮挡物。用户还可以在迁移过程中编辑图像。每组示例左侧展示目标人物的四张照片，右侧展示三个结果，每个结果均以驱动图像作为插图。这些示例涵盖复杂的彩色光照和定向光照（左上）、花朵和相机等遮挡物（右上）、不同的表情，以及来自不同年龄或性别人物的表情（左下），还有对头发和配饰的编辑（右下）。

摘要

在视觉特效中，身份迁移会将驱动图像中的身份替换为目标人物的身份，同时保留该图像中的表情、光照、姿态和场景。配对真实标注很少可用，因此许多方法通过自重建进行训练，在此过程中，模型可能通过从驱动图像中复制身份线索来学习捷径。在推理时，当驱动图像展示的是另一个人时，这种捷径会导致身份泄漏。我们的关键观察是，任何读取驱动图像的输入都可能引入泄漏。因此，我们提出了 *MYSTIQUE*，一种基于扩散的身份迁移方法。该方法从目标人物的一组照片中提取身份，并通过独立流读取表情、光照和场景。它通过多阶段训练，将身份与其他流解耦，并在这些流的输入端采用解耦策略：使用边界不沿面部轮廓分布的随机块掩码；采用自增强，在保持表情不变的同时随机改变表情输入中的脸型；以及人脸随机化，用随机人脸替换表情和光照输入中的人物，同时真实值仍保持为驱动图像。

在推理时，我们还会将光照输入变形到第一遍结果的脸型上，从而减少通过光照流产生的脸型泄漏。由于人脸识别分数和人工判断各有局限，我们结合二者来评估身份，同时还结合面部

对形状已知的渲染人脸进行形状指标评估。在一项用户研究中，评审员认为 *MYSTIQUE* 的结果比各个对比方法的结果更像目标。在这些方法中，*MYSTIQUE* 对目标脸型的还原也最准确；随着提供的目标照片增多，其识别分数的提升幅度也最大。

1. 引言

面部身份迁移会将驱动图像中的身份替换为目标 人物的身份，同时保留驱动图像中的姿态、表情、目光、光照、头发、遮挡物和背景。在影视视觉特效（VFX）中，应用包括特技替身、补拍和数字替身，其中目标演员的肖...

必须重现，同时保留原始表演 [49, 51]。此类镜头可能包含强烈或彩色光照，迁移后的面部必须与之匹配。通常也能获取目标演员的许多照片，将它们结合起来可以提升身份保真度。

实际上，结果通常会保留驱动图像中人物的身份线索，我们将其称为身份泄漏。这源于身份迁移方法的训练方式。直接监督需要包含两位不同人物、且表情、光照和场景都相同的图像，而现实世界的数据集很少包含这样的图像对。因此，许多前馈方法通过自重建进行训练，即模型根据同一人物的照片重建驱动图像 [3, 22, 39, 77]。由于驱动图像已经展现了正确的身份，模型可能会学到一种捷径：从为其提供表情、光照和场景信息的输入中复制身份线索。在推理时，驱动图像展现的是另一个人，而同样的捷径会将该人物的身份带入结果中。然而，模型仍然需要这些输入来保留结果所应具有的表情和光照。

现有方法从单独的输入中获取身份，但它们的其他输入仍然读取驱动图像，并可能携带其中的身份 [56]。基于GAN的方法 [7, 18, 31, 40, 58]注入目标照片的人脸识别嵌入，并重新绘制驱动图像的内部面部。基于扩散的方法 [3, 20, 22, 73, 80] 使用以目标照片的身份嵌入为条件的扩散模型重新生成面部。Face-Adapter [22] 在推理时遏制了这种泄漏：其属性输入编码整个驱动图像，因此将驱动图像的身份令牌作为负条件使用。在我们的实验中，这些方法（包括 Face-Adapter）的结果仍然往往会遵循驱动图像的下颌线。

我们的关键观察是，泄漏可能通过任何读取驱动图像的输入进入，因此我们会在每个输入处减少泄漏。为此，MYSTIQUE 将身份、表情和光照解耦到冻结的扩散变换器的独立流中 [4], , 而场景流读取驱动图像时会将面部掩盖（图 2）。随后，它通过多阶段训练学习将身份与其他输入解耦。阶段 1 在一个独立模型中学习身份表征，该模型能够根据一个人的照片重建此人，因此身份词元独立于驱动图像。阶段 2 在冻结的身份编码器之上学习身份迁移，阶段 3 则学习光照迁移。阶段 2 和阶段 3 通过解耦策略减少这些输入中的泄漏。场景流通过随机块掩码读取驱动图像，该掩码的边界不沿着面部轮廓分布。表情流读取驱动图像时，会随机改变其脸型，同时保留其表情，我们将此称为自增强。表情流和光照流还使用 人脸随机化：它们的输入显示另一个人，而真实值仍然是驱动图像，因此从这些输入中复制身份的作用会减弱。对于表情输入，图像生成器会虚构一个具有不同性别、年龄和面部结构的人。对于光照输入，现有的身份迁移模型 [31]会改变身份并保留光照。由于其结果仍保留部分脸型，因此在推理时，我们会将光照输入变形到第一轮无光照结果的脸型上。

单张照片只能展示目标人物在一种姿态、表情和光照下的样貌，几乎无法揭示其在其他情况下的面部外观。因此，MYSTIQUE 会读取目标人物的无序照片集合，其遵循 Omni-ID 架构的身份编码器 [53], 将这些照片整合为身份令牌。在我们的实验中，与其他方法对照片进行平均相比，增加照片能够更大幅度地提升 MYSTIQUE 的识别分数。由于来自其他输入的泄漏更少，增加的照片也能使脸型更接近目标人物。第二个问题在于，身份很难测量。判断两张面部是否属于同一个人，是一种复杂的人类判断，依据包括脸型、个体特征及其排列方式 [32, 50]。人脸识别嵌入 [11, 37] 能够捕捉这种判断的一部分。即使结果的嘴巴和下方面部形状与目标人物不同 [77], , 它们也可能为该结果给出很高的分数；此外，它们还会对发型和发色作出响应，而这些特征本应保留驱动图像中的样貌 [63]。一些身份迁移工作加入了用户研究或新的身份指标 [20, 40, 77]。沿用这些工作，我们在报告识别分数的同时，也报告人类偏好，并在已知形状的渲染人脸上增加人脸形状指标。

总而言之，我们的贡献如下：

- MYSTIQUE 是一种身份迁移方法，能够从读取驱动图像的输入中学习将身份解耦。该方法通过多阶段训练学习其流，并通过解耦策略减少每个输入中的泄漏。

- 来自一组目标人物照片的身份信息。与其他方法中的取平均相比，增加照片能更大幅度地提高识别分数；同时，由于泄漏更少，这些照片还能使脸型更接近目标人物。

- 由评审员和识别器进行身份评估，并在已知形状的渲染人脸上使用人脸形状指标。在一项用户研究中，与每种对比方法的结果相比，评审员认为 MYSTIQUE 的结果与目标人物更相似；此外，在所有对比方法中，MYSTIQUE 对目标人物脸型的恢复最为准确，且无需目标人物面部的三维模型。

2. 相关工作

身份解耦。自重建会让模型从驱动图像中复制身份信息，而以往的工作通过两种方式减轻这一问题。第一种方式构建跨身份训练对。显示另一个人的输入由

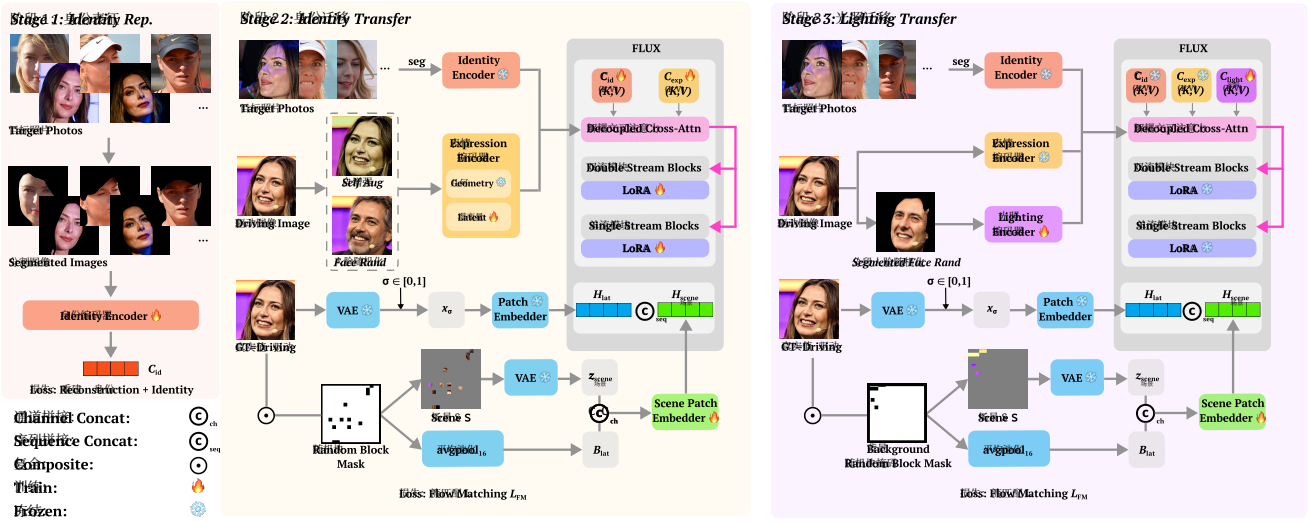


图 2。**分三个阶段进行训练**。阶段 1 从人物照片中学习身份表征；其第二步通过冻结的 Transformer 完成，图中未绘制。阶段 2 在冻结身份编码器和几何分支的情况下学习身份迁移。阶段 3 在冻结身份流和表情流的情况下学习光照迁移。身份、表情和光照词元通过解耦交叉注意力进入每个块，其中每个编码器的投影作为其键和值的一部分绘制（$\boldsymbol{K}$，$\boldsymbol{V}$）；场景词元 $\boldsymbol{H}_{\text{scene}}$ 会附加到噪声潜在标记 $\boldsymbol{H}_{\text{lat}}$上。自增强、人脸随机化、分割、真实值分别缩写为 Self Aug、Face Rand、seg、GT。

表情重演和混合 [77], 通过现有的身份迁移模型 [29, 49, 62, 73], 或由面部标志点驱动的视频模型 [20]生成。当图像对还应在脸型上有所不同时，还会增加额外步骤，例如变形 [29, 73] 或 3DMM 重定向 [20]。第二种方式减少模型所读取内容中的身份线索。BlendFace [63] 在属性被替换的图像上训练其身份编码器，使发型及类似属性不会对其产生影响。在表情重演中，X-NeMo [81] 将姿态和表情压缩为一维潜变量，并使用颜色和脸型被随机改变的驱动图像进行训练，使其携带很少的身份信息。Omni-ID [53] 在训练其身份编码器时，保留待重建图像的一小部分可见，从而使该部分泄漏的身份信息很少。MYSTIQUE 同时采用了这两种方式：随机块掩码和自增强遵循第二种方式，而人脸随机化遵循第一种方式。人脸随机化是域随机化的一种形式 [65]：输入显示另一个人，而真实值仍保持为驱动图像。对于表情输入，图像生成器会虚构一个具有不同性别、年龄和面部结构的人，从而无需额外步骤即可改变脸型。

身份迁移。 DeepFaceLab [44]为每对人物训练一个模型。与主体无关的基于 GAN 的方法将目标的人脸识别嵌入注入驱动图像 [7, 18, 31, 40, 58]的特征中。基于扩散的方法 [3, 22, 36, 73, 80]随后出现，其中一些将该任务视为条件图像修复 [3, 22, 80], , 而近期的方法则使用扩散 Transformer 将其扩展到视频 [20, 49]。ID-Constrained [23] 首先仅使用身份条件进行训练，然后再添加其他条件；MYSTIQUE 在一种独立模型。为控制脸型，许多这类方法会添加目标人物的显式几何信息，例如 3DMM 系数或 3D 关键点 [22, 62, 69, 71, 80]。MYS-TIQUE 不使用目标人物的三维几何信息，而是从其照片中获取脸型。多参考身份迁移会使用目标人物的多张照片 [72], , 由获得专利的基于 StyleGAN 的方法 [35], 对这些照片进行注意力处理，以增强对质量不佳照片的鲁棒性；GSwap [82], 则替换整个头部，并针对每个对象和每段视频优化一个模型。基于 GAN 的方法的嵌入也可以在多张照片上取平均值，就像我们在第 4 节中对基线方法所做的那样。在身份迁移之外，照片集也很常见：个性化方法会对识别嵌入取平均值 [52, 68]或堆叠图像嵌入 [41], ； 少样本人脸重演会汇集主体的多帧画面 [21, 78],；基于参考的修复会关注主体的多张照片 [26, 46], ； 而 Omni-ID [53] 会将照片集编码为固定数量的词元。MYSTIQUE 将目标人物的照片作为一个集合来读取。其他输入造成的泄漏越少，增加照片也能让脸型更接近目标人物。

3. 方法

MYSTIQUE 通过冻结的扩散 Transformer 的独立流读取目标人物照片和驱动图像（第 3.1 节），并分三个阶段对其进行训练，从而仅根据某人的照片学习身份词元（第 3.2 节）。由于其他流会读取驱动图像，三种解耦策略（第 3.3 节）和两阶段推理（第 3.4 节）可减少身份通过这些流发生的泄漏。

3.1. 架构

给定一组 $\mathbb{T} = \{\mathbf{T}_i\}_{i=1}^{M}$ M 目标人物照片和一张驱动图像 $\mathbf{D}$，MYSTIQUE 会生成一张图像，其中呈现目标人物的身份，同时保留 $\mathbf{D}$ 的姿态、表情、光照和场景。所有图像均为人脸裁剪图：每张照片调整为$\mathbf{T}_i \in \mathbb{R}^{448\times448\times3}$，驱动图像调整为$\mathbf{D} \in \mathbb{R}^{H\times W\times3}$，这也是结果图像的分辨率（$H = W = 256$；补充材料 F 和 H 报告了 512 像素模型）。我们将该任务表述为图像修复：二值掩码 $\boldsymbol{B} \in \{0,1\}^{H\times W}$ 标记 $\mathbf{D}$ 的人脸区域，模型会移除该区域并重新生成。生成过程在 FLUX.2 Klein 9B 扩散 Transformer [4], 中运行，其预训练权重保持冻结；在每个块的注意力投影上应用低秩更新 [27] ，使其适应身份迁移。图 2 展示了每个训练阶段的架构。

编码器。 三个编码器将输入转换为宽度为 $d_c = 1280$ 的词元。身份编码器读取目标的 M 人脸图像，每张图像都由人脸解析模型进行掩码处理 [75]，仅保留面部、耳朵和颈部。这些区域是模型在推理时会替换的部分，因此头发不会进入身份词元 $\boldsymbol{C}_{\text{id}} \in \mathbb{R}^{256\times d_c}$。词元数量不取决于 M，因此推理时可以自由选择 M。表情编码器通过两个分支读取驱动图像。几何分支预测面部的 128 个隐式三维关键点，遵循 LivePortrait 的原理 [19], ，并使用一个单独的模型；该模型由我们预训练（补充材料 A.1）且保持冻结；一个线性层将每个关键点的三维位置映射为一个词元。潜变量分支，即 X-NeMo 的表情编码器 [81], ，从人脸图像中读取潜在表情描述符，并通过线性层将其扩展为128 个宽度为 d_c 的词元；阶段 2 使用其预训练权重对其进行微调。层归一化与平均操作将两个分支融合为 $\boldsymbol{C}_{\text{exp}} \in \mathbb{R}^{128\times d_c}$，与单独使用任一分支相比，它能更贴近表情（第 4.5 节）。光照编码器读取经过掩码处理的驱动图像，仅保留皮肤、眉毛、眼睛、鼻子、嘴巴和眼镜，不包括颈部、耳朵或头发。一个小型卷积神经网络将其映射为一个 16×16 网格，其单元成为光照词元 $\boldsymbol{C}_{\text{light}} \in \mathbb{R}^{256\times d_c}$，每个词元与一个潜变量词元对齐。

条件化。 这四种流以两种方式进入 Transformer：身份、表情和光照令牌通过交叉注意力输入，而场景则作为词元序列的一部分输入。Transformer 对 FLUX.2 Klein 自编码器的潜变量进行去噪，该潜变量是一个包含 128 个通道的 16×16 网格，Transformer 将其读取为具有隐藏状态 $\boldsymbol{H}_{\text{lat}} \in \mathbb{R}^{N\times d}$的词元 $N = 256$, $d = 4096$。与 I P-Adapter [74], 一样，每种词元流 $k \in$ {id、exp、light} 都有一个由所有块共享的投影 $\phi_k : \mathbb{R}^{d_c} \to \mathbb{R}^{d}$，并且每个块都有自己的键 $\boldsymbol{K}_k = \phi_k(\boldsymbol{C}_k)\boldsymbol{W}_{\text{K}}^{k}$和值 $\boldsymbol{V}_k = \phi_k(\boldsymbol{C}_k)\boldsymbol{W}^{k}$ v，其中包含 $\boldsymbol{W}_{\text{K}}^{k}$、$\boldsymbol{W}^{k}v\in \mathbb{R}^{d\times d}$。这些流的注意力输出会加到噪声潜变量的隐藏状态中：

$$\boldsymbol{H}_{\text{lat}} \leftarrow \boldsymbol{H}_{\text{lat}} + \sum_{k} \text{Attn}(\boldsymbol{Q}, \boldsymbol{K}_k, \boldsymbol{V}_k), \tag{1}$$

其中 $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$ 是该块中噪声潜在标记的查询。对于光照流，旋转位置嵌入 [64] 将每个键放在其单元格的中心，并将每个查询放在其自身位置，因此注意力取决于相对位置，且每个输出位置都可以关注输入中同一位置的光照。场景$\mathbf{S} = (\mathbf{1} - \boldsymbol{B}) \odot \mathbf{D}$ 是移除了面部的驱动图像。自编码器将其编码为 $\boldsymbol{z}_{\text{scene}}$，并与 $\boldsymbol{B}_{\text{lat}}$连接；掩码 $\boldsymbol{B}$ 经平均池化后达到潜变量分辨率。一个可训练的补丁嵌入副本，即场景图块嵌入，将结果映射为场景令牌 $\boldsymbol{H}_{\text{scene}} \in \mathbb{R}^{N\times d}$。这些令牌沿序列附加到噪声潜在标记之后，并共享其空间位置，因此场景与结果保持对齐。固定的文本提示（见补充材料 A.1）补全了该序列。

3.2. 多阶段训练

我们通过自重建来训练 MYSTIQUE：驱动图像和目标人物照片是同一个人的不同照片，模型重建驱动图像，并将其作为真实值。令 $\boldsymbol{x}_0 \in \mathbb{R}^{n}$ 为真实值展平后的潜变量，其中 $n = 16 \cdot 16 \cdot 128$。给定 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$，从逻辑正态分布 [17], 中抽取的噪声水平 $\sigma \in [0,1]$ 以及 $\boldsymbol{x}_\sigma = (1-\sigma)\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}$，第 2 阶段和第 3 阶段将流匹配损失 [17, 43]最小化

$$L_{\text{FM}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},\sigma}\left\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_\sigma, \sigma) - (\boldsymbol{\epsilon} - \boldsymbol{x}_0)\right\|_2^2, \tag{2}$$

其中， $\boldsymbol{v}_{\boldsymbol{\theta}}$ 是模型根据噪声潜变量和各流预测的速度， $\boldsymbol{\theta}$ 是其可训练参数。损失仅作用于噪声潜在标记。

图 2 所示的三个阶段均从上一阶段的模型开始。 **阶段 1**学习不依赖驱动图像的身份表征。遵循 Omni-ID 的架构 [53], ，我们训练自己的身份编码器，通过掩码解码器重建人物照片。除了重建损失外，身份损失还通过余弦相似度比较每张重建图像与其未掩码照片的 AdaFace [37]嵌入。随后，我们通过冻结的 Transformer，使用公式 2中的流匹配损失进一步训练编码器：Transformer 根据身份令牌生成该人物的其他照片；这一步也会训练身份流的交叉注意力。由于阶段 1 不输入驱动图像，身份令牌学会承载身份信息，包括脸型，而无需借助驱动图像。

阶段 2 在此表征的基础上学习身份迁移。身份编码器和几何分支保持冻结，而潜变量分支、身份流和表情流的投影与交叉注意力，以及场景

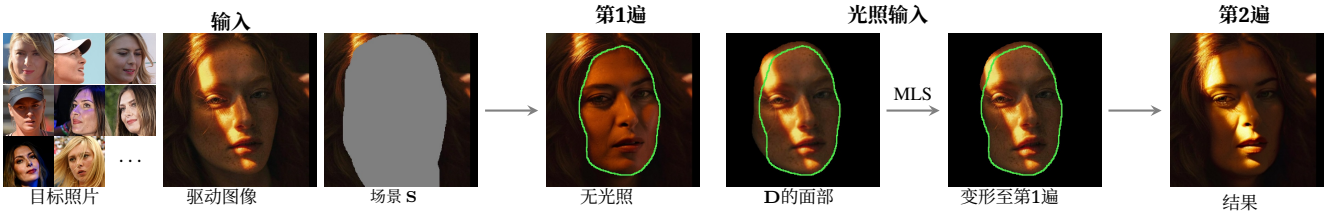


图3. **两阶段推理**。 第1遍在不使用光照流的情况下运行。驱动图像中的面部会变形为第1遍结果（绿色轮廓）的脸型，但发际线处除外；该面部随后作为第2遍的光照输入，而第2遍的其他步骤与第1遍相同。

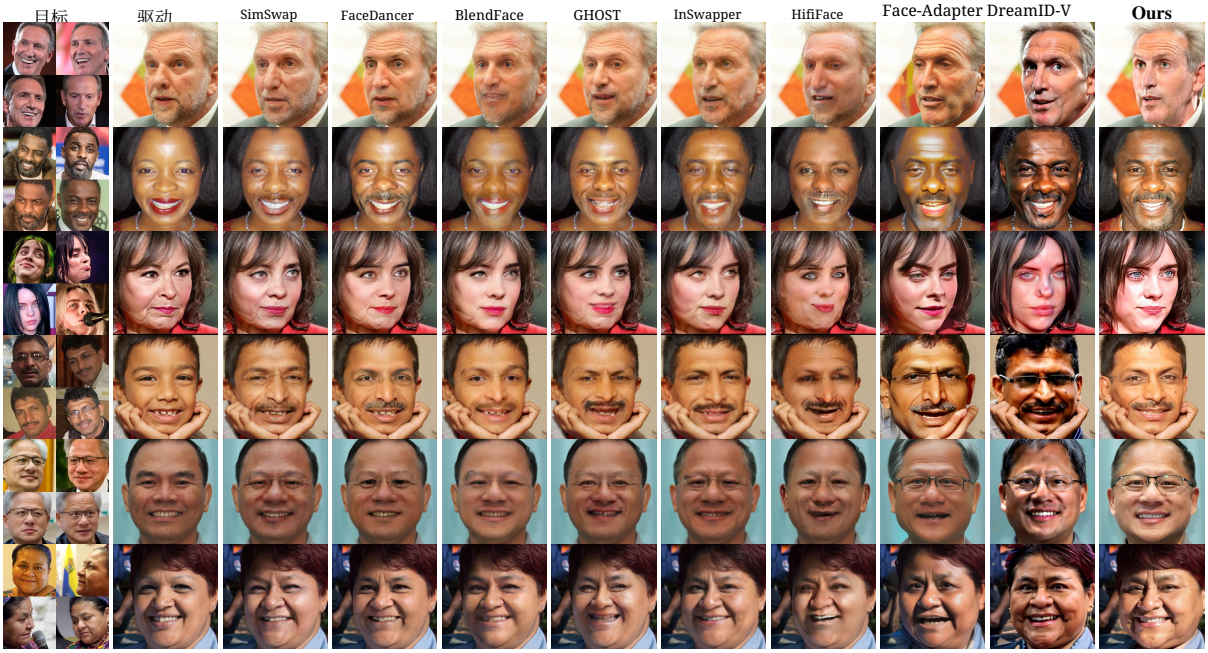


图4. **与先前工作的比较**。每行展示目标对象的四张照片、驱动图像和生成结果。MYSTIQUE 的结果看起来更像目标对象，包括其脸型，同时保留了驱动图像中的头发、表情、光照和背景。

图块嵌入和低秩更新会参与训练。每个样本随机抽取 M 张照片，数量在 1 到 24 张之间，因此模型学会使用任意数量的照片。阶段 2 不使用光照流。

阶段 3 随后学习光照迁移：添加光照流，并将其与场景补丁嵌入一同训练，同时冻结身份流、表情流和低秩更新，因此这一阶段不会改变它们所编码的内容。

3.3. 解耦策略

表情、光照和场景流读取驱动图像，而在自重建中，驱动图像本身就是真实值。因此，这些输入中的每一个都可能将真实值的身份线索传递到结果中：几何分支编码其脸型，跟随面部的掩码向场景流揭示其轮廓，而光照输入则显示出面部本身。复制这些线索是一种能够降低训练损失的捷径，但在推理时，它会将驱动图像中人物的脸型带入结果。第 2 阶段和第 3 阶段通过三种解耦策略减少这种泄漏。

随机块掩码。 基于图像修复的方法通常会在面部掩码内生成面部，无论是在训练阶段 [3] 还是推理阶段，这种掩码的轮廓都会将驱动图像的脸型传递给结果 [22]。REFace 和 SelfSwapper [39] 会在训练时扰动该掩码。阶段 2 则改为使用随机块掩码进行训练 $\boldsymbol{B}$，它会遮挡驱动图像中随机的 95–99% 的 16×16像素块，每个潜变量令牌对应一个块。隐藏区域不沿着面部轮廓分布，因此掩码几乎不包含有关脸型的信息。在阶段 3 中，可见块仅取自背景，因此场景不会呈现人物身上的光照，模型会从光照流中读取相关信息。

表1。 **身份基准测试上的比较**（240 图像对）。每种方法均采用其最佳设置（第 4.1节）。 偏好度：用户研究中（第 4.2节），偏好我们的结果而非该行结果的身份判断所占比例（%）。 *ArcFace*：目标人物的真实照片得分为 0。677；斜体值高于该值，不参与排名。粗体：最佳结果。

方法	偏好度	ArcFace↑	表情↓	姿态↓	光照↑	FaceScore↑
SimSwap [7]	97	0.500	0.243	**2.03**	**0.96**	3.67
FaceDancer [58]	81	0.394	**0.205**	2.38	**0.96**	4.39
BlendFace [63]	92	0.487	0.231	4.42	0.90	3.73
GHOST [18]	99	0.548	0.337	4.60	0.93	4.02
InSwapper [31]	98	*0.715*	0.258	2.70	0.93	3.92
HifiFace [71]	100	0.528	0.345	3.14	0.90	4.02
Face-Adapter [22]	73	0.501	0.332	4.38	0.83	4.09
DreamID-V [20]	79	0.550	0.372	7.52	0.52	4.42
Ours	–	**0.601**	0.253	3.99	0.83	**4.62**

表 2. **形状基准测试中的脸型** (348 图像对）。指标与第 4.2 节相同。每种方法均接收目标对象的 24 张渲染图，并按表 1 中的设置运行。粗体：最佳。 驱动图像：驱动图像本身，作为结果进行评分。

方法	s_{gt} ↓	泄漏↓	3D↓	方法	s_{gt} ↓	泄漏↓	3D↓
SimSwap	0.125	0.95	1.83	InSwapper	0.116	0.81	1.48
FaceDancer	0.122	0.91	1.85	HifiFace	0.138	0.78	2.03
BlendFace	0.119	0.86	1.99	Face-Adapter	0.138	0.60	1.86
GHOST	0.132	0.90	2.21	DreamID-V	0.149	0.52	1.67
Ours	**0.083**	**0.34**	**1.35**	驱动图像	0.128	1.00	3.49



图 5. **遮挡物和眼镜**。 MYSTIQUE 会生成手部周围以及眼镜后方的面部，同时仍遵循光照。在掩码中，红色区域为生成内容，手部（蓝色）和镜框（黄色）保留自驱动图像，而透明镜片后方（绿色）的眼睛则会重新生成。

自增强。 表情流应传递驱动图像中的表情与姿态，但不应传递其脸型。因此，阶段 2 会向表情编码器提供一份驱动图像副本，其中脸型经过随机改变，而表情保持不变：移动最小二乘法 [60] 会将面部轮廓移动最多 12%，并将面部宽高比按$[0.8, 1.25]$中的随机因子缩放。真实值仍为原始驱动图像。这样一来，复制表情输入的脸型对于重建真实值的帮助就会减小，而遵循其表情与姿态仍然有帮助。

人脸随机化。 自增强会改变表情输入中的脸型，但该输入与光照输入一样，仍然显示真实值中的人物，而在推理过程中，两者显示的都是目标人物以外的其他人。人脸随机化将这种情况引入训练：表情输入和光照输入显示的是另一个人，而真实值仍然是驱动图像，因此从这些输入中复制身份会使结果趋向错误的人物，而损失会对此进行惩罚。对于阶段 2 中的表情输入，图像生成器需要将驱动图像转换为具有不同性别、年龄和面部结构的人物，同时保留表情、姿态、光照、肤色以及图像的其他部分。对于阶段 3 中的光照输入，现有的身份迁移模型 [31] 会用随机选择的身份替换原身份，同时保留光照。其结果保留了部分真实值的脸型，而双遍推理会将这部分从我们的结果中排除（见 3.4）。

3.4. 推理

两阶段推理。 两阶段推理可避免在包含脸型泄漏的输入上训练的光照流将驱动图像中人物的脸型传递到结果中（图 3）。第一阶段不使用光照流进行推理。模型支持这样做，因为训练时每个流都会被随机丢弃（补充材料 A.2），因此该阶段结果的脸型不受光照输入影响。随后，我们使用移动最小二乘法将驱动图像的面部变形到该结果上 [60],，第二阶段则以变形后的面部作为光照输入，重复第一阶段的处理，从而使其携带驱动图像的光照，并保留第一阶段结果的脸型。

掩码与背景。在推理时，掩码 $\boldsymbol{B}$ 决定图像中哪些部分会被生成，哪些部分会保留驱动图像中的内容。默认情况下， $\boldsymbol{B}$ 是一个人脸解析掩码 [75]，覆盖面部、耳朵和颈部，也就是身份编码器读取的区域（第3.1节），因此头发会保留驱动图像中的内容。面部前方的遮挡物（例如手或珠宝）不包含在掩码中，因此也会保留；眼镜框同样如此；透明镜片后的眼睛则会重新生成（图5）。若要编辑头发，人脸+头发掩码也会覆盖头发，并通过描述编辑内容的提示词引导生成新的头发，如图1所示。掩码外的潜变量会在每个采样步骤中替换为驱动图像的潜变量，并加上与当前级别相对应的噪声，这与混合潜变量扩散方法相同， [1], 从而确保掩码外的所有内容都与驱动图像一致。

4. 实验

4.1. 设置

实现细节。 除非另有说明，结果均使用 256 像素模型、面部掩码和目标人物的 M=24 照片。我们使用约 3 万人的面部照片进行自重建训练，并通过面部识别模型 [37] 按人物分组。对于身份、表情和

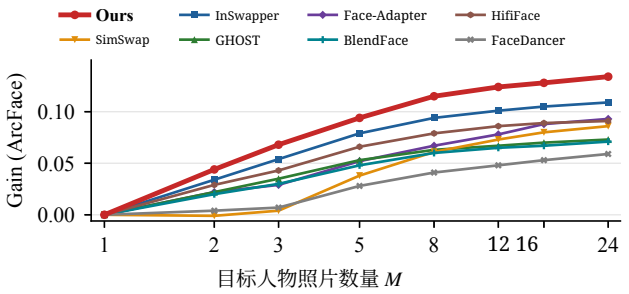


图6。**更多目标照片**。身份增益（ArcFace，与24 张留出照片的平均余弦相似度）相比 M=1，在身份基准的240 对图像上。基线方法对其身份输入的 M 张照片取平均值；MYSTIQUE 则将这些照片作为一个集合读取。绝对分数见补充材料。I。

光照流（补充材料A.4和K.3）。

基线方法。 我们比较了十二种方法。论文正文报告了其中八种：基于GAN的 SimSwap [7],FaceDancer [58], BlendFace [63], GHOST [18], InSwap-per [31] 和 HifiFace [71], , 以及基于扩散的 Face-Adapter [22] 和 DreamID-V [20]。补充材料 F报告了另外四种方法：DiffSwap [80], REFace [3], DiffFace [36] 和 E4S [47], , 并介绍了两种商业图像模型，以及各方法的运行方式。每种基线方法均采用其最佳设置：对于需要单个身份嵌入的方法，输入目标人物照片的平均嵌入，这会提高其识别分数（第 4.4节）。

4.2. 评估

我们为本文构建了两个基准测试：一个基于真实照片的身份基准，人工评审者在该基准上评判身份；以及一个基于渲染人脸的形状基准，其真实面部形状是已知的。

身份基准。 身份基准包含来自 VGGFace2-HQ 的 240个目标对象，VGGFace2-HQ 是一个我们未用于训练的 [5, 9], 公开数据集。每种方法都会接收目标对象的 24张照片，并另外保留 24 张其他照片用于身份评分。每个目标对象都与数据集中另一人的一张驱动图像配对。这些图像对分为四组，每组包含 60 对。在其中一组中，两人的性别和族裔相同，年龄也相近。在另外三组中，他们分别在其中一项属性上存在差异（补充材料 B.1）。

身份测量指标。 通常使用的身份得分，即结果与目标人物的人脸识别嵌入之间的余弦相似度，存在已知局限。它反映的是两张照片是否呈现同一个人，而不是两张面部之间有多相似 [38]。由于识别器经过训练，能够对图像质量保持稳健性 [37],，因此该指标可能忽略细微细节 [29, 71]，并且对脸型的响应很小。它还偏向于使用同一识别器的身份损失进行训练的方法 [63]。因此，近期研究开始根据人工判断来学习人脸相似度 [28, 38]。人工判断也存在局限：在我们的研究中，评审员之间超出偶然水平的一致性为 $\kappa = 0.34$，而相似度和真实感的回答有 87% 的试验相符（补充材料E）。鉴于二者均存在局限，我们同时报告这两项指标。在研究的每次试验中，一名评审员会看到目标人物照片中的三张照片、驱动图像，以及按平衡顺序呈现的两张未标注结果图：我们的方法和一个基线方法的结果；评审员需要选出更像目标人物的结果，以及更真实的结果。研究分为两个批次进行，每批次有 30名评审员（共 42人），每批次每人评判约 50 次试验。针对表 1 中八种基线方法的 1,726次试验使用了身份基准中的目标对象和名人目标；补充材料 E 报告了其他方法。身份得分采用 ArcFace [11]，并以 24 张留出照片为依据，而我们的任何损失都未使用这些照片；目标人物的真实照片得分为 0.677。二者可能不一致：InSwapper 保留了驱动图像的大部分下颌线（表 2），但得分为 0.715；相比 InSwapper 的结果，评审员在 98% 的试验中更偏好我们的方法。

形状基准测试。 真实照片无法提供目标对象的面部在驱动图像中应呈现何种样貌的真实值，因此我们渲染具有已知形状的面部。形状基准测试包含 29 个 MetaHuman [16] 角色，这些角色是在 Unreal Engine 5 中渲染的。每个角色作为目标对象对应 12 组图像对，其驱动图像展示其他角色，共形成 348 组图像对。真实值是使用驱动图像的相机、表情和光照渲染出的目标对象；每种方法还会接收目标对象的 24 张其他渲染图。我们报告三个面部形状指标（补充材料 B.2）。轮廓距离 s_{gt}是结果与真实值之间下颌线的平均距离。泄漏指标衡量结果的下颌线在多大程度上遵循驱动图像中的下颌线，取值范围为 0（完全不遵循）至 1 （完全复制）。三维误差比较 Deep3DFaceRecon [12]从结果和真实值重建出的面部形状。

表情、姿态、光照和真实感。 在身份基准上，我们还测量每个结果与驱动图像的接近程度。表情是 EmoNet 对两幅图像预测的效价和唤醒度之间的距离，姿态则是 HopeNet [66]预测的两幅图像头部旋转之间的测地角 [59]。光照是两幅图像在人脸区域内低通亮度之间的相关性。真实感是 FaceScore [42], , 一种无需参考图像、用于人脸质量的学习型奖励。其他三种表情度量，以及用于姿态的 6DRepNet [24] , 对这些方法的排序相似（补充材料F）。

4.3. 与先前工作的比较

表 1 比较了各方法在身份基准上的表现，表 2则比较了它们在形状基准测试上的表现。图 4 展示了示例。

身份。 在 90% 的试验中，评审员认为我们的结果比目标更相似；与 Face-Adapter 相比，这一比例为 73%，而与 HifiFace 相比则达到 100%。在 ArcFace 评估下，我们的结果得分为 0.601，仅次于 InSwapper，其得分高于

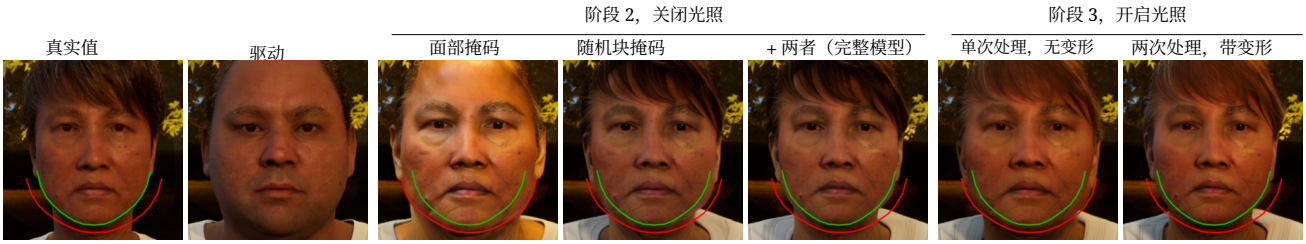


图 7。 **消融实验中的脸型**。 表 3 的各个变体在形状基准测试中一对角色上的结果。第一列为真实值：使用驱动图像的相机、表情和光照渲染的目标。每个变体都会接收目标的 24 张其他渲染图。线条：真实值的下颌线（绿色）和驱动图像的下颌线（红色）。

目标人物的真实照片（第 4.2节）。

脸型。六种基线方法都紧随驱动图像的下颌线，泄漏值为 0.78 至 0.95（图 4）。Face-Adapter 和 DreamID-V 对其跟随程度较低（泄漏值分别为 0.60 和 0.52），但它们的轮廓距离（ 0.138 和 0.149）高于驱动图像的轮廓距离（0.128），HifiFace 也是如此（0.138）。这些结果改变了驱动图像的下颌线，却没有朝目标方向改变。我们的方法轮廓距离最低，为 0.083；最接近的基线方法 InSwapper 为 0.116。我们的方法泄漏值也最低，为 0.34。此外，我们的三维误差最低，为 1.35，而 InSwapper 为 1.48。

表情、姿态、光照与真实感。我们的方法在表情和姿态上的误差处于基线方法的范围内：EmoNet 距离为 0.253（基线方法为 0.205 至 0.372），姿态误差 $3.99°$，而 DreamID-V 为 2.03 至 $4.60°$ 和 $7.52°$ 。光照流将我们的光照得分从 0.66 提高到 0.83（第 4.5 节）。以驱动图像为参照进行评分时，这些测量指标会倾向于保留其面部的结果（补充材料 D.1）。FaceScore 对我们结果的真实感评分为 4.62，是表1中的最高分。

4.4. 目标照片数量的影响

更多照片会提高所有采用身份嵌入的基线方法的识别分数，其中我们的方法提升幅度最大（图6）。每种方法均接收固定顺序中的前 M 张照片，该顺序会改变姿态；基线方法对身份输入取平均，而只使用一张照片的 DreamID-V 不纳入比较。从 $M{=}1$ 到 24，基线方法的 ArcFace 得分提高了 0.059 至 0.109；我们的方法将照片作为一个集合进行整合，得分提高了 0.134。新增照片也让我们的方法生成的脸型更接近目标：轮廓距离从 0.090 降至 0.083。这只需要少量泄漏。In-Swapper 的轮廓距离变化不到 0.001；而不采用这些策略的模型（表 3 第一行）的结果仍接近驱动图像（ 0.124 对比 0.128），尽管其 ArcFace 得分从 0.470 升至 0.599。

4.5. 消融实验

表 3 对解耦策略（第3.3节）、表情编码器和两阶段推理（第3.4节）进行了消融研究。这些策略和两阶段推理主要作用于面部

表 3. **消融实验**。 每一行代表一个模型。左：身份基准测试上的身份、表情、姿态和光照，详见表 1。右：形状基准测试上的脸型，详见表 2。两组使用不同的数据。

	身份基准				形状基准测试		
变体	身份↑	表情↓	姿态↓	光照↑	s_{gt} ↓	泄漏↓	3D↓
阶段 *2*，关闭光照							
面部掩码	0.599	0.268	3.44	0.74	0.124	0.82	1.51
随机块掩码	0.610	0.248	3.68	0.67	0.098	0.54	1.37
+ 自增强	0.610	0.256	3.98	0.66	0.082	0.38	1.34
+人脸随机化	0.618	0.256	3.78	0.67	0.087	0.45	1.28
+ 两者（完整模型）	0.616	0.256	3.97	0.66	0.081	0.33	1.26
仅几何分支	0.634	0.375	5.11	0.57	0.093	0.23	1.47
仅潜变量分支	0.613	0.271	4.52	0.63	0.082	0.33	1.30
阶段 *3*，启用光照							
单次处理，无变形	0.599	0.248	3.42	0.87	0.095	0.60	1.45
使用变形进行两次处理	0.601	0.253	3.99	0.83	0.083	0.34	1.35

形状，而表情编码器作用于表情与姿态。

脸型。 每种策略以及两阶段推理都会降低泄漏（图 7）。使用推理阶段的人脸解析掩码进行训练后，模型会遵循该掩码的轮廓（泄漏 0。82）。随机块掩码将泄漏降低至 0。54，而自增强与人脸随机化结合可将其进一步降低至 0。33。光照流将光照得分从 0。66提升至 0。83，但在单次处理中，它会使驱动图像的脸型恢复（泄漏 0。60）；两次处理将泄漏保持在0。34。沿着这些行，身份和表情变化很小；姿态误差增加了约 $0.5°$，部分原因是姿态估计器也会对脸型产生响应（补充材料D.1）。

表情与姿态。 表情编码器的两个分支协同工作时效果最佳。仅使用几何分支会使表情误差和姿态误差升至 0.375 和 $5.11°$；仅使用潜变量分支会使表情误差升至 0.271，并使光照得分降至 0.63。两者都无法让脸型更接近目标。

5. 结论

我们介绍了 MYSTIQUE，这是一种身份迁移方法，能够学习将身份与读取驱动图像的输入解耦。它将身份、表情和光照分别分配到冻结的扩散 Transformer 的独立流中，并通过多阶段训练、三种解耦策略和两次推理，减少这些输入造成的泄漏。因此，面部形状主要来自目标人物照片，添加照片后，面部形状会更接近目标。在一项用户研究中，评审员认为我们的结果比各个对比方法的结果更像目标；在渲染人脸上，我们的方法生成的脸型最接近目标。

大语言模型使用声明

我们使用 Claude 在作者指导下编写并运行训练和评估代码，以及起草和润色文本。除论文所描述的结果图和面部随机化图像外，没有任何图由 AI 模型生成。在方法和评估过程中，Qwen2.5-VL 负责命名遮挡物，闭源图像模型创建训练输入，多模态 GPT 模型标注族裔以构建基准测试图像对，而 GPT 和 Gemini 图像模型则生成用于补充材料中对比的图像。

参考图像

[1] Omri Avrahami、Ohad Fried 和 Dani Lischinski。混合潜扩散。ACM TOG，42(4):149:1–149:11，2023年。6，16[2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song 等。Qwen2.5-VL 技术报告。arXiv 预印本 arXiv:2502.13923，2025年。15[3] Sanoojan Baliah、Qinliang Lin、Shengcai Liao、Xiaodan Liang 和 Muhammad Haris Khan。真实高效的人脸交换：一种基于扩散模型的统一方法。载于 WACV，2025年。2，3，5，7，20[4] Black Forest Labs。FLUX.2 [klein]。https : / / huggingface.co/black- forest- labs/FLUX. 2-klein-base-9B，2026年。开放权重模型，2026年1月发布。2，4[5] Qiong Cao、Li Shen、Weidi Xie、Omkar M. Parkhi 和 Andrew Zisserman。VGGFace2：一个跨姿态和年龄识别面部的数据集。载于 FG，2018年。7，17，20[6] Di Chang、Yufeng Yin、Zongjian Li、Minh Tran 和 Mohammad Soleymani。LibreFace：用于深度面部表情分析的开源工具包。载于 WACV，2024年。21[7] Renwang Chen、Xuanhong Chen、Bingbing Ni 和 Yanhao Ge。SimSwap：一种高保真人脸交换的高效框架。载于 ACM多媒体会议，2020年。2，3，6，7，20[8] Wei-Ting Chen、Gurunandan Krishnan、Qiang Gao、Sy-Yen Kuo、Sizhuo Ma 和 Jian Wang。DSL-FIQA：通过双集合退化学习和地标引导 Transformer 评估面部图像质量。载于 CVPR，2024年。17，20，21[9] Xuanhong Chen、Bingbing Ni、Yutian Liu、Naiyuan Liu、Zhilin Zeng 和 Hang Wang。SimSwap++：迈向更快的以及高质量身份交换。IEEE TPAMI，46(1)：576–592，2024年。7, 17[10] Mehdi Cherti、Romain Beaumont、Ross Wightman、Mitchell Wortsman、Gabriel Ilharco、Cade Gordon、Christoph Schuh- mann、Ludwig Sch midt 和 Jenia Jitsev。对比式语言-图像学习的可复现缩放定律。载于 CVPR，2023年。12[11] Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou。ArcFace：用于深度人脸识别的加性角度间隔损失。载于 CVPR，2019年。2, 7, 20[12] Yu Deng、Jiaolong Yang、Sicheng Xu、Dong Chen、Yunde Jia 和 Xin Tong。采用弱监督学习的精准三维人脸重建：从单张图像到图像集。载于 CVPRW，2019年。7, 18, 20[13] Donglin Di、He Feng、Wenzhang Sun、Yongjia Ma、Hao Li、Wei Chen、Lei Fan、Tonghua Su 和 Xun Yang。DH-FaceVid-1K：用于人脸视频生成的大规模高质量数据集。载于 ICCV，2025年。12[14] Paul Ekman 和 Wallace V Friesen。测量面部动作。《环境心理学与非言语行为》，1 (1)：56–75，1976年。18[15] Paul Ekman 和 Dacher Keltner。情绪面部表情的普遍性：一场旧有争论与新发现。载于《非言语交流：自然与文化的交汇》，第27–46页。新泽西州马瓦：Lawrence Erlbaum Associates，1997年。18[16] Epic Games。MetaHuman。https://www.metahuman.com，2021年。7, 18[17] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel、Dustin Podell、Tim Dockhorn、Zion English 和 Robin Rombach。扩展整流流变换器以实现高分辨率图像合成。载于 ICML，2024年。4[18] Alexander Groshev、Anastasia Maltseva、Daniil Chesakov、Andrey Kuznetsov 和 Denis Dimitrov。GHOST——一种面向图像与视频领域的新型人脸交换方法。IEEE Access，10：83452–83462，2022年。2, 3, 6, 7, 20[19] Jianzhu Guo、Dingyun Zhang、Xiaoqiang Liu、Zhizhou Zhong、Yuan Zhang、Pengfei Wan 和 Di Zhang。LivePortrait：通过拼接和重定向控制实现高效肖像动画。arXiv预印本 arXiv:2407.03168，2024年。4, 12[20] Xu Guo、Fulong Ye、Xinghui Li、Pengqi Tu、Pengze Zhang、Qichao Sun、Songtao Zhao、Xiangwang Hou 和 Qian He。OmniFace：通过扩散变换器弥合图像到视频的差距，实现高保真人脸交换。载于 ECCV，2026年。代码和模型以 DreamID-V 的形式发布。2, 3, 6, 7, 20[21] Sungjoo Ha、Martin Kersner、Beomsu Kim、Seokjun Seo 和 Dongyoung Kim。MarioNETte：保留未见目标身份的少样本人脸表演迁移。载于 AAAI，2020年。3[22] Yue Han、Junwei Zhu、Keke He、Xu Chen、Yanhao Ge、Wei Li、Xiangtai Li、Jiangning Zhang、Chengjie Wang 和 Yong Liu。Face-Adapter：面向预训练扩散模型的细粒度身份与属性控制。载于 ECCV，2024年。2, 3, 5, 6, 7, 20[23] Dailan He、Xiahong Wang、Shulun Wang、Hao Shao、Bingqi Ma、Guanglu Song、Yu Liu 和 Hongsheng Li。高保真

采用身份约束面部条件的人脸扩散交换。载于 CVPR，2026年。3[24] Thorsten Hempel、Ahmed A. Abdelrahman 和 Ayoub Al-Hamadi。用于非受限头部姿态估计的六维旋转表示。载于国际图像处理会议，2022年。7, 12, 19, 21[25] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv预印本 arXiv:2207.12598，2022年。14[26] Chi-Wei Hsiao、Yu-Lun Liu、Cheng-Kun Yang、Sheng-Po Kuo、Yucheun Kevin Jou 和 Chia-Ping Chen。ReF-LDM：用于基于参考的人脸图像复原的潜空间扩散模型。载于 NeurIPS，2024 年。3[27] Edward J. Hu 、 Yelong Shen 、 Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。载于 ICLR，2022年。4[28] Ying Huang、Wencan Zhang 和 Brian Y. Lim。AlignFace：具可解释概念关系的人类对齐人脸相似度度量。载于 ACM多媒体会议，2026年。7[29] Ziyao Huang、Fan Tang、Yong Zhang、Juan Cao、Chengyu Li、Sheng Tang、Jintao Li 和 Tong-Yee Lee。通过双代理生成模型实现身份保持的人脸交换。ACM TOG，43(5)：161:1–161:19，2024年。3, 7[30] InsightFace。InsightFace：二维和三维人脸分析项目。https://github.com/deepinsight/insightface，2017年。17[31] InsightFace。InSwapper：inswapper 128人脸交换模型。h t t p s://github.com/deepinsight/insightface/tree/master/examples/in_swapper，2022年。随 InsightFace Python软件包 v0.7 一同发布。2, 3, 6, 7, 14, 20[32] Kamila M. Jozwik、Jonathan O'Keeffe、Katherine R. Storrs、Wenxuan Guo、Tal Golan 和 Nikolaus Kriegeskorte。可变形模型和可由图像计算的模型中的表征距离可以预测人脸差异性判断。《美国国家科学院院刊》，119(27)：e2115047119，2022年。2[33] Tero Karras、Samuli Laine 和 Timo Aila。生成对抗网络的基于风格的生成器架构。载于 CVPR，2019年。15, 18, 24, 27[34] Yury Kartynnik、Artsiom Ablavatski、Ivan Grishchenko 和 Matthias Grundmann。通过移动GPU上的单目视频实时获取人脸表面几何。载于 CVPRW，2019年。18[35] Dae-Shik Kim 和 Jae-Hyeok Lee。使用注意力机制，通过单张/多张源图像进行人脸交换的系统和方法。美国专利 12,354,195 B2，2025年。申请公开为美国专利申请公开US 2023/0260176 A1。3[36] Kihong Kim、Yunho Kim、Seokju Cho、Junyoung Seo、Jisu Nam、Kychul Lee、Seungryong Kim 和 KwangHee Lee。DiffFace：带面部引导的基于扩散的人脸交换。Pattern Recognition，163：111451，2025年。3, 7, 20[37] Minchul Kim、Anil K. Jain 和 Xiaoming Liu。AdaFace：人脸识别的质量自适应间隔。载于 CVPR，2022年。2, 4, 6, 7, 13, 20[38] Haruka Kumagai、Leslie W¨ohler、Satoshi Ikehata 和 Kiyoharu Aizawa。PerFace：通过感知人脸相似度度量学习增强人脸匿名化。载于国际图像处理会议，2025年。7[39] Jaeseong Lee、Junha Hyung、Sohyun Jeong 和 Jaegul Choo。SelfSwapper：通过形状无关掩码自编码器实现自监督人脸交换。载于 ECCV，2024年。2，5[40] Lingzhi Li、Jianmin Bao、Hao Yang、Dong Chen 和 Fang Wen。推进用于伪造检测的高保真身份交换。载于 CVPR，2020年。2，3[41] Zhen Li、Mingdeng Cao、Xintao Wang、Zhongang Qi、Ming-Ming Cheng 和 Ying Shan。PhotoMaker：通过堆叠式身份嵌入定制逼真照片。载于 CVPR，2024年。3[42] Zhenyi Liao、Qingsong Xie、Chen Chen、Hannan Lu 和 Zhijie Deng。FaceScore：人类生成中的人脸质量基准测试与提升。arXiv 预印本 arXiv:2406.17100，2024年。7，21[43] Yaron Lipman、Ricky T. Q. Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。生成建模的流匹配。载于 ICLR，2023年。4[44] Kunlin Liu、Ivan Perov、Daiheng Gao、Nikolay Chervoniy、Wenbo Zhou 和 Weiming Zhang。DeepFaceLab：集成式、灵活且可扩展的人脸交换框架。Pattern Recognition，141:109628，2023年。3[45] Shilong Liu、Zhaoyang Zeng、Tianhe Ren、Feng Li、Hao Zhang、Jie Yang、Qing Jiang、Chunyuan Li、Jianwei Yang、Hang Su、Jun Zhu 和 Lei Zhang。Grounding DINO：将 DINO 与基于定位的预训练相结合，用于开放集目标检测。载于 ECCV，2024年。15[46] Siyu Liu、Zheng-Peng Duan、Jia OuYang、Jiayi Fu、Hyunhee Park、Zikun Liu、Chun-Le Guo 和 Chongyi Li。FaceMe：具有个人身份识别能力的鲁棒盲人脸修复。载于 AAAI，2025年。3[47] Zhian Liu、Maomao Li、Yong Zhang、Cairong Wang、Qi Zhang、Jue Wang 和 Yongwei Nie。通过区域 GAN 反演实现细粒度人脸交换。载于 CVPR，2023年。7，20[48] Camillo Lugaresi、Jiuqiang Tang、Hadon Nash、Chris McClanahan、Esha Uboweja、Michael Hays、Fan Zhang、Chuo-Ling Chang、Ming Guang Yong、Juhyun Lee、Wan-Teh Chang、Wei Hua、Manfred Georg 和 Matthias Grundmann。MediaPipe：用于构建感知流水线的框架。arXiv 预印本 arXiv:1906.08172，2019年。18，20[49] Zekai Luo、Zongze Du、Zhouhang Zhu、Hao Zhong、Muzhi Zhu、Wen Wang、Yuling Xi、Chenchen Jing、Hao Chen 和 Chunhua Shen。保持源视频真实感：面向电影级质量的高保真人脸交换。载于 CVPR，2026年。2，3[50] Daphne Maurer、Richard Le Grand 和 Catherine J. Mondloch。构形加工的多重面貌。Trends in Cognitive Sciences，6(6):255–260，2002年。2[51] Jacek Naruniec、Leonhard Helminger、Christopher Schroers 和 Romann M. Weber。用于视觉特效的高分辨率神经人脸交换。Comput. Graph. Forum，39(4):173–184，2020年。2[52] Foivos Paraperas Papantoniou、Alexandros Lattas、Stylianos Moschoglou、Jiankang Deng、Bernhard Kainz 和 Stefanos Zafeiriou。Arc2Face：面向身份一致人脸的基础模型。载于 ECCV，2024年。3[53] Guocheng Qian、Kuan-Chieh Wang、Or Patashnik、Negin Heravi、Daniil Ostashev、Sergey Tulyakov、Daniel Cohen-Or、

以及 Kfir Aberman。Omni-ID：为生成任务设计的整体身份表征。载于 CVPR，2025年。2，3，4，12[54]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移的视觉模型。载于 ICML，2021年。12[55] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman Radle、Chloe Rolland、Laura Gustafson 等。SAM 2：分割图像和视频中的任何目标。载于 ICLR，2025年。15¨Xiaohang Ren、Xingyu Chen、Pengfei Yao、Heung-Yeung Shum 和 Baoyuan Wang。无需跳跃连接的人脸交换强化解耦。载于 ICCV，[56] 。George Retsinas、Panagiotis P. Filntisis、Radek Danˇeˇcek、Victoria F. Abrevaya、Anastasios Roussos、Timo Bolkart 和 Petros Maragos。通过神经分析合成实现三维面部表情。载于 CVPR，2024年。20

2023 2 Felix

[57] Rosberg、Eren Erdal Aksoy、Fernando Alonso-Fernandez 和 Cristofer Englund。FaceDancer：姿态与遮挡感知的高保真人脸交换。载于 WACV，2023年。2，3，6，7，20[58] Nataniel Ruiz、Eunji Chong 和 James M. Rehg。无需关键点的细粒度头部姿态估计。载于 CVPRW，2018年。7，21[59] Scott Schaefer、Travis McPhail 和 Joe Warren。使用移动最小二乘法进行图像变形。ACM TOG，25(3):533–540，2006年。6，14，15[60] Florian Schroff、Dmitry Kalenichenko 和 James Philbin。FaceNet：用于人脸识别与聚类的统一嵌入。载于 CVPR，2015年。20[61] Hao Shao、Shulun Wang、Yang Zhou、Guanglu Song、Dailan He、Zhuofan Zong、Shuo Qin、Yu Liu 和 Hongsheng Li。VividFace：鲁棒且高保真的视频人脸交换框架。载于 NeurIPS，2025年。3[62] Kaede Shiohara、Xingchao Yang 和 Takafumi Taketomi。BlendFace：重新设计用于人脸交换的身份编码器。载于 ICCV，2023年。2，3，6，7，20[63] Jianlin Su、Murtadha Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu。RoFormer：采用旋转位置嵌入的增强型 Transformer。Neurocomputing，568:127063，2024年。4，12[64] Josh Tobin、Rachel Fong、Alex Ray、Jonas Schneider、Wojciech Zaremba 和 Pieter Abbeel。用于将深度神经网络从仿真迁移到真实世界的域随机化。载于 IROS，2017年。3[65] Antoine Toisoul、Jean Kossaifi、Adrian Bulat、Georgios Tzimiropoulos 和 Maja Pantic。在自然条件下从面部估计连续的效价与唤醒度水平。Nature Machine Intelligence，3(1):42–50，2021年。7，20，24[66] Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。MEAD：用于情感说话人脸生成的大规模音视频数据集。载于 ECCV，2020年。12，24[67] Qixun Wang、Xu Bai、Haofan Wang、Zekui Qin、Anthony Chen、Huaxia Li、Xu Tang 和 Yao Hu。InstantID：数秒内实现零样本身份保持生成。arXiv 预印本 arXiv:2401.07519，2024年。3[69] Runqi Wang、Yang Chen、Sijie Xu、Tianyao He、Wei Zhu、Dejia Song、Nemo Chen、Xu Tang 和 Yao Hu。Dynamic-Face：利用可组合的三维面部先验实现高质量且一致的图像与视频人脸交换。收录于 ICCV，2025年。3[70] Xintao Wang、Yu Li、Honglun Zhang 和 Ying Shan。面向真实场景的盲人脸修复：生成式面部先验。收录于 CVPR，2021年。17[71] Yuhan Wang、Xu Chen、Junwei Zhu、Wenqing Chu、Ying Tai、Chengjie Wang、Jilin Li、Yongjian Wu、Feiyue Huang 和 Rongrong Ji。HifiFace：由三维形状与语义先验引导的高保真人脸交换。收录于 IJCAI，2021年。3、6、7、20[72] Ruizheng Wu、Xin Tao、Yingcong Chen、Xiaoyong Shen 和 Jiaya Jia。超越共性，凸显个性：利用多个参考图像进行无配对身份迁移。收录于 ECCV，2020年。3[73] Fulong Ye、Miao Hua、Pengze Zhang、Xinghui Li、Qichao Sun、Songtao Zhao、Qian He 和 Xinglong Wu。DreamID：通过三元组身份组学习实现高保真、快速的基于扩散的人脸交换。收录于 SIGGRAPH Asia会议论文集，2025年。2、3[74] Hu Ye、Jun Zhang、Sibo Liu、Xiao Han 和 Wei Yang。IP-Adapter：兼容文本的图像提示适配器，适用于文本到图像扩散模型。arXiv 预印本 arXiv:2308 06721，2023年 4 . . [75] Changqian Yu 、 Jingbo Wang 、 Chao Peng 、 Changxin Gao、Gang Yu 和 Nong Sang。BiSeNet：用于实时语义分割的双边分割网络。收录于 ECCV，2018年。4、6、12、17、21[76] Jianhui Yu、Hao Zhu、Liming Jiang、Chen Change Loy、Wei- dong Cai 和 Wayne Wu。CelebV-Text：大规模面部文本视频数据集。收录于 CVPR，2023年。12[77] Ge Yuan、Maomao Li、Yong Zhang 和 Huicheng Zheng。ReliableSwap：通过可靠监督提升通用人脸交换效果。arXiv 预印本 arXiv:2306.05356，2023年。2、3[78] Egor Zakharov、Aliaksandra Shysheya、Egor Burkov 和 Victor Lempitsky。逼真神经说话头模型的少样本对抗学习。收录于 ICCV，2019年。3[79] Richard Zhang、Phillip Isola、Alexei A. Efros、Eli Shechtman 和 Oliver Wang。深层特征作为感知度量的非凡效用。收录于 CVPR，2018年。13[80] Wenliang Zhao、Yongming Rao、Weikang Shi、Zuyan Liu、Jie Zhou 和 Jiwen Lu。DiffSwap：通过三维感知掩码扩散实现高保真、可控的人脸交换。收录于 CVPR，2023年。2、3、7、20[81] Xiaochen Zhao、Hongyi Xu、Guoxian Song、You Xie、Chenxu Zhang、Xiu Li、Linjie Luo、Jinli Suo 和 Yebin Liu。X-NeMo：通过解耦的潜变量注意力实现富有表现力的神经运动重演。收录于 ICLR，2025年。3、4、12、13[82] Jingtao Zhou、Xuan Gao、Dongyu Liu、Junhui Hou、Yudong Guo 和 Juyong Zhang。GSwap：利用动态神经高斯场实现逼真的头部交换。IEEE TVCG，32(7)：5595–5607，2026年。3[83] Hao Zhu、Wayne Wu、Wentao Zhu、Liming Jiang、Siwei Tang、Li Zhang、Ziwei Liu 和 Chen Change Loy。CelebV-HQ：大规模视频面部属性数据集。收录于 ECCV，2022年。12

MYSTIQUE：少 -镜头面部身份迁移与性能保持

补充材料

A. 方法与实现细节

本节按照第3节和第4.1节的顺序，详细介绍复现 MYSTIQUE 所需的内容。除非另有说明，这些内容均针对 3节中提到的 256px 模型；第 A.4 节最后介绍 512px 模型的改动。

A.1. 架构

表 4 列出了每个流的输入、编码器和词元，以及用于训练该流的阶段。下方各段落按照表中的行依次展开。

输入裁剪图。 所有输入均为面部周围的裁剪图像。根据五个面部标志点（眼睛、鼻尖和嘴角）拟合一个边界框，将其放大。0倍，并向上移动其高度的0。2；图像外部的部分以黑色填充。目标照片调整为448像素，驱动图像调整为256像素。

身份编码器。 身份编码器遵循Omni-ID的架构 [53]，，并由我们自行训练（参见 A.2）。人脸解析模型 [75]首先将照片中面部、耳朵和颈部之外的每个像素设为黑色，因此编码器不会看到头发。一个从LAION-2B权重 14 [10, 54],初始化的CLIP ViT-H/14 [10, 54],会读取每张448像素照片，我们取其倒数第二层中不含类别词元的1024个图像块令牌。所有 M 照片的图像块令牌构成一个键和值序列，键上的照片索引使用一维旋转位置嵌入 [64]，以区分不同照片。256 个可学习查询随后经过一个自注意力模块以及两对交叉注意力和自注意力模块，每个模块均具有16 个注意力头，最后通过层归一化得到 $\boldsymbol{C}_{\text{id}} \in \mathbb{R}^{256\times1280}$。由于查询是固定的，词元数量不依赖于 M。

几何分支。 几何分支以与 LivePortrait [19] 相同的方式，从驱动图像中提取 128 个面部隐式三维关键点。我们将其作为独立模型进行预训练，与 LivePortrait 有两点不同：它预测 128 个关键点而非 21 个，并以 6D 表示头部旋转。在这项预训练中，ViT生成器根据关键点和外观特征渲染人脸，6DRepNet [24] 则通过测地线损失监督头部旋转。此外还加入了两种一致性损失：身份一致性损失用于惩罚同一人的两张图像之间身份向量的差异；表情一致性损失则用于惩罚规范空间中同一表情不同视角之间表情向量的差异。数据包括 CelebV-Text [76], CelebV-HQ [83], DH-FaceVid-1K [13],MEAD [67] 以及约 8 万张生成的人脸图像，并且编码器训练了约 1.03M 步。MYSTIQUE 仅使用冻结的关键点提取器。其驱动关键点经归一化至 $[-1, 1]$ 后，通过一个从 3 维映射至 1280 维的线性层，并为每个关键点添加一个可学习的位置嵌入，该嵌入初始化为零。

潜变量分支。 潜变量分支使用 X-NeMo 的表情编码器，从人脸图像中提取潜变量表情描述符。X-NeMo [81], 是一个约有 6500 万个参数的网络，接受 224 像素输入。我们使用 X-NeMo 的权重初始化该编码器，不使用其边界框条件；在阶段 2 对其进行微调，并在阶段 3 将其冻结。它输出一个 512 维向量。一个线性层将该向量扩展为 128 个宽度为 16 的词元，与几何分支的关键点词元数量相同；随后添加固定的正弦位置编码，再由第二个线性层（其权重初始化为零）将每个词元映射为宽度 1280。然后融合这两个分支：对每组词元进行层归一化，将两组词元取平均，再对平均结果进行层归一化，从而得到 $\boldsymbol{C}_{\text{exp}} \in \mathbb{R}^{128\times1280}$。每个样本都会启用这两个分支。

光照编码器。 光照编码器读取驱动图像，并将面部以外的所有像素设为零；这里的面部包括皮肤、眉毛、眼睛、鼻子、嘴巴和眼镜，不包括耳朵、颈部和头发。四个步幅为 2 的卷积阶段（通道数分别为 32、64、128 和 256，并采用组归一化和 SiLU）将 256 像素输入缩小为一个 16×16 网格，因此每个网格单元覆盖一个潜变量词元。对每个网格单元进行层归一化，再通过一个共享线性层将其投影为宽度 1280，并为每个单元添加一个可学习的位置嵌入，从而得到 $\boldsymbol{C}_{\text{light}} \in \mathbb{R}^{256\times1280}$。

交叉注意力。 身份、表情和光照词元会输入 Transformer 的全部 32 个块，包括 8 个双流块和 24 个单流块。每个流共享的投影ϕ_k 是一个两层多层感知机，包含 GELU，维度从 1280 映射到 4096，随后进行层归一化。在每个块中，大小为 4096×4096 的线性层生成该流的键和值，且键会按头部进行均方根归一化。查询和输出投影沿用 Transformer 的组件，并保持冻结。值层初始化为零，因此新流在开始训练时不会改变模型的输出。各流的输出以权重 1 加到噪声潜在标记的隐藏状态中，不会加到其他词元中。

场景令牌与文本。 场景中的人脸区域填充为中灰色，在自编码器的输入范围 $[-1, 1]$中对应 0，因此$\mathbf{S} = (\mathbf{1} - \boldsymbol{B}) \odot \mathbf{D}$ 在此处成立。掩码经平均池化至 16×16 潜变量网格，并附加到 128 个潜变量通道上，作为第 129 个通道。一个将 129 维映射到 4096 维的线性层，其权重由

表4。**MYSTIQUE 的流**（256 像素模型）。词元：数量 × 宽度。场景令牌进入 Transformer 的序列；其他流通过交叉注意力进入每个块。

流	输入	编码器	词元	训练阶段
身份	M 目标照片，448像素；面部，耳朵和颈部	CLIP ViT-H/14 和注意力池-ing	256×1280	阶段 1
表情（几何）	驱动图像，256 像素	关键点提取器，预先单独训练	128×1280	冻结；嵌入位于阶段 2
表情（潜变量）	驱动图像，224 像素	X-NeMo 表情编码器 [81]	128×1280	阶段 2
光照	驱动图像中的面部，256 像素	16×16 上的卷积神经网络 grid	256×1280	阶段 3
场景	不含面部的驱动图像，以及 $\boldsymbol{B}$	自编码器和补丁嵌入	256×4096	第 2 阶段和第 3 阶段

Transformer 的补丁嵌入初始化，并为掩码对应的列设为零，由此生成场景令牌。它们携带旋转时间偏移 1，使其与处于相同空间位置的噪声潜在标记区分开来；其调制使用干净图像的时间步。训练和推理时，序列中都会加入固定提示词“a photorealistic portrait of a human face”以补全序列，且从不丢弃。

低秩更新。 低秩更新的秩为 64，缩放因子为 64。在双流块中，它会调整图像标记的查询、键、值和输出投影；在单流块中，它会调整融合输入投影和输出投影。文本标记的投影不会进行调整。

A.2. 训练

表5给出了各阶段的调度安排。每个阶段都从前一阶段的权重开始，并且所有阶段每步在每个图形处理器上从 M均匀采样 $\{1, \ldots, 24\}$一次。

阶段 1。阶段 1 分两个步骤训练身份编码器。在第一步中，掩码解码器根据身份词元重建该人物的其他照片，与 Omni-ID 中的做法相同。损失由区域加权 L1 项（权重 1.0）、带 VGG 特征的 LPIPS [79]项（权重 0.5）以及身份项（权重 0.1）组成；其中，$1-\cos$ 表示重建图像与照片的 AdaFace [37]嵌入之间的余弦相似度，两者均使用五个面部标志点对齐至 112px。前 925k 步中的前 45k 步不启用身份项。每个样本包含一名人物，其中 4 张照片用于重建，M 张照片用于身份编码器。学习率在 500 步内预热至 10^{-4}，随后采用余弦调度衰减。训练曾在 775k 步时恢复一次，并将调度延长至 1M 步，使学习率升至 1.2×10^{-5}，之后再次衰减；我们使用其 925k 步时的检查点。在第二步中，编码器和身份交叉注意力通过冻结的 Transformer，使用公式 2 的流匹配损失进行训练：Transformer 根据身份词元生成该人物的其他照片，不使用驱动图像。CLIP 主干网络在两个步骤中都进行微调。阶段 1 的样本中，99% 来自我们的数据集，1% 来自 MEAD。

阶段 2。 阶段 2 通过在我们的数据集上进行自重建来训练身份迁移，其中照片按人物分组（第 4.1 节）。驱动图像是从某一组中均匀抽取的一张照片，目标人物的照片最多为同组中的 24 张其他照片，这些照片依次从不同头部姿态的簇中抽取；模型接收其中前 M 张。驱动图像本身绝不会出现在目标人物的照片中。该阶段分为两个阶段，每个阶段包含 100k 步。第一阶段使用面部掩码进行训练，该掩码覆盖面部解析区域、耳朵和颈部，并扩张 12px；同时不使用解耦策略；其对应于表 3 的第一行模型。第二阶段从此基础上继续训练，并加入随机块掩码、自增强和人脸随机化（第 A.3 节）。在两个阶段中，潜变量分支的输入还会接收颜色抖动（亮度、对比度和饱和度 ±0.2、色调 ±0.05）、最多 $\pm3\%$ 的平移以及最多 $\pm5\%$ 的缩放，不进行旋转。这些增强在每次阶段 2 运行中均启用，包括表 3 的第一行，因此不属于自增强。两个阶段均使用 0.01 的权重衰减、500 步预热以及 1.0 的梯度裁剪。

阶段 3。 阶段 3 加入光照流，并训练50k 步，包含 500 步预热和余弦调度。身份流、表情流和低秩更新保持冻结。光照输入是进行人脸随机化后的驱动图像（第 A.3节），面部之外的每个像素都被设为零，场景中可见的块仅来自背景。在此阶段，自增强和人脸随机化都不会改变表情输入。阶段 3 使用阶段 2 的数据集和帧采样方式；驱动图像是为一个样本采样的帧中最后一帧。

流丢弃与引导。 身份流、表情流和光照流分别以 0.1 的概率被丢弃；被丢弃的流，其键和值均设为零。三次抽样彼此独立，并且每个批次大小的所有样本在每一步仅抽样一次。场景和文本提示永不丢弃。在阶段 2 中，所有条件流同时关闭的步约占 1%（两个流）；在阶段 3 中约占 0.1%（三个流）。这使模型能够训练预测——

表 5. **训练计划**（256 像素模型）。批次大小：所有 GPU 上每步的样本数。学习率：学习率。所有阶段均使用混合精度（bfloat16）。

阶段	Step	步数	LR	批次大小	GPUs	训练对象
1，身份	重建	925k	10^{-4}	16 人	8 H200	身份编码器，掩码解码器
	流匹配	410k	10^{-5}	8	8 H200	身份编码器，身份交叉注意力
2，身份迁移	固定人脸掩码	100k	10^{-5}	16	8 B200	潜变量分支；投影和交叉注意力 身份流和表情流；场景图像块 嵌入；低秩更新
	随机块掩码 以及两种策略	100k	10^{-5}	16	8 H200	如阶段 2 的第一步
3，光照	–	50k	2×10^{-5}	32	16 H200	光照编码器和交叉注意力；场景图块 嵌入

时不使用第一遍所用光照流的预测，以及无分类器引导的无条件预测 [25]，后者会同时丢弃三个流（见第A.4节）。公式 2 中的噪声水平为 $\sigma = \text{sigmoid}(u)$，其中 $u \sim \mathcal{N}(0,\ 1)$。

A.3. 解耦策略

随机块掩码。阶段 2 的随机块掩码会为每个样本重新绘制。每 r 步在每个图形处理器上从 $[0.95, 0.99]$ 均匀抽取一次，并且256 个 16×16 像素块会以概率 r独立隐藏。平均而言，约有3% 的块（约8个）保持可见，分布在图像各处，也包括面部。阶段 3 会隐藏所有与人物相交的块，包括头发，只保留背景中的块，概率相同；因此场景中人物没有光照，模型会从光照流中读取光照信息（第3.3节）。在我们的紧裁剪图像中，约有0.8% 的块保持可见，每张图像约2 个；约四分之一的样本中没有可见块。训练掩码不会屏蔽遮挡物；这只会在推理时发生。

自增强。自增强会对阶段 2 第二阶段每个样本的表情输入进行变形，即对两个分支的输入都进行变形。移动最小二乘法 [60] 首先以 0.88 到 1.12 之间的随机因子移动人脸轮廓，随后以 0.8 到 1.25 之间的随机因子缩放面部宽高比，不进行平移。场景、光照输入和真实值不进行变形。

人脸随机化。人脸随机化会在阶段 2 第二阶段的 30% 样本中，将表情输入中的人物替换掉。对于这些样本，一个闭源图像生成器会根据下方的提示词，将照片编辑为一个新人物。它接收我们数据集中的整张方形照片，其中面部约占边长的 18%，而不是训练裁剪图（约 55%），因此有足够空间改变脸型。随后，根据在生成图像上检测到的面部标志点，使用与真实照片相同的裁剪方式，从其输出中裁剪出表情输入。表 6 检验生成图像是否保留了表情与姿态。目标人物和真实值人物的照片仍然是驱动图像中人物的真实照片，因此从表情输入中复制身份会使结果偏离真实值。提示词原文如下：

> 将照片中的人物替换为一名性别和表观年龄都不同的其他人物，使任何人脸识别系统都无法将其与原人物匹配。改变底层骨骼结构：颅骨和脸颊骨的形状、面部宽度和长度、下颌与下巴的形状、鼻子、嘴唇、眼睛的形状和间距，以及眉毛的形状。重要提示：新人物必须与原人物具有完全相同的肤色、肤质、底色和皮肤颜色——不得在任何位置提亮、变暗或改变皮肤的色调，并且必须完全保留皮肤上的每一道阴影和高光。其他所有内容都必须保持像素级一致：头部姿态和相机角度、面部表情以及每块肌肉的位置（眉毛、眼睑开合程度、嘴巴形状、下颌张合程度）、目光方向、光照方向／颜色／强度以及面部上的所有阴影和高光、肤色和色彩分级、衣物、背景、取景，以及头部在帧中的尺度和位置，还有图像清晰度和电影颗粒。不得重新调整姿态、光照、取景，不得裁剪、缩放、拉直或美化任何内容。输出相同的分辨率。

我们请求生成 40,000 张图像，每张图像都来自从我们数据集中随机选取的一张照片；1,487 张图像被生成器的内容过滤器拦截，另有 38,513 张图像成功生成。关键点检查会移除生成器移动或替换了人物的图像（面部距离原始位置超过 0.5个眼间距），并保留其中37,883张，约涉及30k 人。照片与生成图像之间的身份相似度会被记录，但不会用于筛选。图8 展示了生成器处理公开照片的效果。对于阶段 3 的光照输入，InSwapper [31]会将每个样本的身份替换为从我们数据集的固定身份池中随机抽取的一个，池中共有 511 个身份，且不进行变形。

表6 检查训练所使用的生成图像，其中包括 500组随机配对，来自通过关键点检查的 37,883 组图像。它们改变了身份和脸型：ArcFace 与照片的相似度平均为 0.39，低于同一人物照片配对中 99.6%的平均值（均值为0.73），而下颌线移动了 0.14个单位s_{gt}，与自增强的典型变形幅度大致相当。在我们检查的 15 组16 组随机配对中，性别发生了变化。

表6. **人脸随机化改变和保留的内容**，在 500 张训练用随机照片上进行。每列将照片与另一张图像进行比较：阶段 2 生成的图像、同一人的另一张照片、另一个人的照片，以及阶段 3 的 InSwapper 结果。ArcFace：余弦相似度。下颌线：两条下颌线之间的距离，单位为 s_{gt}，每张图像均单独裁剪。EmoNet 和 6DRepNet：如第 D节所述的表情与姿态；以照片作为驱动图像。对于形状基准测试中表情与姿态相同的两个角色，它们的值为 0.26 和 2.62°（表 12）。光照：光照相关性。 $|\Delta L^*|$：皮肤亮度的绝对差值。

	生成的	同一人	其他人	InSwapper
身份和脸型：应发生变化				
ArcFace	0.39	0.73	0.01	0.03
下颌线	0.14	0.51	0.72	0.09
表情、姿态和光照：应保持不变				
EmoNet	0.20	0.37	0.48	0.31
6DRepNet（度）	3.6	27.0	37.5	3.1
光照	0.77	0.49	0.12	0.92
$\|\Delta L^*\|$	3.1	5.3	13.7	1.3

新人物仍然保留了照片中的一些特征，因为提示词保留了皮肤的肤色以及皮肤上的每一道阴影和高光，而且生成器不会改变所有特征：相似度仍高于与另一人物照片的相似度（0.01）。它们保留了表情与姿态：EmoNet 距离为 0.20，低于该估计器对形状基准测试中两个具有相同表情与姿态的角色给出的距离（0.26，第 D.1 节）；姿态误差 3.6°比它们的 1.0° 高。光照相关性为 0.77，皮肤略微变暗（-2.4 在 L^*中）。在这些图像上进行训练并不会削弱表情与姿态：在随机块掩码中加入人脸随机化（表3）会使 EmoNet 距离变化 $+0.008$（95% 区间为 $[-0.007, 0.022]$），并使 6DRepNet 姿态误差变化 $+0.2°$（$[0.05, 0.43]$），小于仅改变脸型对该估计器造成的变化（从 1.4° 到 1.6°，表12）。InSwapper 更好地保留了光照，但下颌线移动幅度更小（0.09），因此其结果保留了照片部分脸型（第3.3节）。

A.4. 推理

两阶段推理。 第一阶段将光照流的键和值设为零。为了构建第二阶段的光照输入，我们使用 InsightFace 在驱动图像和第一阶段结果中检测 106 个面部标志点，并使用人脸解析模型对面部进行解析。变形的控制点包括内部标志点、从面部中心发出的 72 条射线与面部边界的交点，以及耳朵和颈部边缘上的点。基于相似性的移动最小二乘变形 [60] 将驱动图像的面部映射到第一阶段结果上，然后再次对变形图像进行解析。光照输入保留与第一阶段结果中的面部重叠或位于眉毛上方的变形面部，但不包括耳朵。



图 8. **公开照片上的人脸随机化**。每组展示一张来自 FFHQ [33] 的照片（左）以及图像生成器根据训练提示词生成的图像（右）。训练中的关键点检查保留了全部八张照片。生成的人物在性别和年龄上有所不同，而姿态、表情、光照和背景则与照片相近。

如果未检测到面部，或有效射线少于 12 条，第二阶段会读取驱动图像中的面部，而不进行变形。第二阶段使用与第一阶段相同的随机种子及所有其他输入。**掩码与遮挡物**。 面部掩码是面部、耳朵和颈部的人脸解析区域膨胀 12 px 后，再减去膨胀 3 px 的遮挡物所得的区域。遮挡物通过开放词汇检测器链在每张驱动图像中检测一次。Grounding DINO [45]会查找固定对象列表（手、手指、麦克风、面罩、食物、冰淇淋、手机、相机、鲜花、吸管、香烟、杯子、瓶子、围巾和拳头）。Qwen2.5-VL-7B-Instruct [2]会进一步识别覆盖或接触中央面部的其他物体，但不包括眼镜、头发、帽子以及图像中的其他人物；语言模型和 Grounding DINO 会在较低阈值下分别定位每个名称。SAM 2.1 [55] 会对所有框进行分割；对于其他人物，在驱动图像中该人物的眼睛、鼻子和嘴巴上执行负点击，以排除该面部。眼镜、头发、帽子、衣物和面部部位的名称会被丢弃；将其余掩码合并并膨胀 3×3 像素，同时排除驱动图像中人物的眼睛后，所得结果即为遮挡物图。面部+头发掩码还会覆盖头发，使用相同的膨胀设置，并且不会遗漏任何遮挡物。

这种默认设置是自动完成的，速度很快。对于21 组240组图像对中（见表格中的行）1，自动生成的映射标记范围过大或过小，因此我们手动修正了掩码或遮挡物映射；

其余219组图像对保留自动生成的映射。在电影制作中，艺术家可以通过转描或使用更强大的检测和分割模型进行此类修正，而模型本身不会改变。对于全部240 组图像对均使用自动生成的映射时，我们在表1 中的得分变化最多为0。004 （ArcFace）和0。02（FaceScore），且各列排名均未下降。表1和3中的其他行使用相同的映射；基线方法则使用各自的默认设置。

眼镜。在电影中，演员和特技替身经常戴眼镜，透明镜片仍能看见镜片后面的眼睛。因此，眼镜的处理方式不同于上一段中的遮挡物，后者会保留驱动图像中的内容：透过透明镜片，眼睛必须变成目标的眼睛。眼镜在每次处理中都遵循自己的规则。眼镜区域（包括镜框和镜片）会在前 60% 的采样步骤中生成。此后，每一步都会将其潜变量设为驱动图像经过加噪的潜变量，而最后一步之后则设为干净的潜变量，因此镜框会保留驱动图像中的样子。如果镜片的平均亮度低于 0.31 倍皮肤亮度，且低于皮肤亮度的 0.72 倍，则镜片为深色；深色镜片会保持原样。透过透明镜片时，还会进行一次额外的采样运行：从该轮处理的输出开始，并使用新的随机种子，重新生成镜片内部，从而让结果中的眼睛透过镜片显现。因此，一个结果需要进行两次采样运行；若有透明镜片，则需要四次。在身份基准测试的 240 张驱动图像中，有 10 张戴着眼镜，其中 9 张戴的是透明镜片。

采样。我们使用 28 个欧拉步数，在 FLUX.2 的偏移噪声调度上进行采样，其偏移量根据由 256 个潜变量令牌组成的序列长度而定。引导将两种预测组合为 $\boldsymbol{v} = \boldsymbol{v}_{\mathrm{u}} + 4(\boldsymbol{v}_{\mathrm{c}} - \boldsymbol{v}_{\mathrm{u}})$。无条件预测 $\boldsymbol{v}_{\mathrm{u}}$ 将身份、表情和光照流的键和值设为零，并将文本替换为下方的负面提示词；两种预测共享场景令牌。每一步都会将掩码外的潜变量替换为驱动图像的潜变量，并用初始噪声将其加噪至当前级别；最后再写回驱动图像的干净潜变量 [1]。在我们的评估中，每一对样本在两次处理过程中都使用同一个固定随机种子。

负面提示词。负面提示词内容为“面部彩绘、纹身、皮肤上的迷彩图案、标记、污渍、面部彩色斑块、文本、说明文字、字幕、字母、水印、徽标、韩文字符、手、手指、触碰面部的手、电缆、绳线”。我们发现，它降低了模型在生成区域绘制伪影的可能性，例如纹身、说明文字或手等标记。模型在训练过程中从未看到该提示词。在身份基准测试中，移除该提示词会使 ArcFace 从 0.601 降至 0.592，使姿态误差从 3.99 度升至 4.23 度，并使光照得分从 0.83 降至 0.82，而 EmoNet 距离变化为 0.003。提示词中只列出了伪影，因此我们将这些变化理解为干扰得分的伪影减少，而不是更好的——



图 9. **负面提示词**。使用相同随机种子时，模型不使用和使用负面提示词的结果。不使用负面提示词时，脸颊上会出现污渍（上图），嘴巴处会出现一个物体（下图）。使用负面提示词后，ArcFace 分数从 0.53 升至 0.58，以及从 0.60 升至 0.64；姿态误差则从 7.0° 降至 5.4°，以及从 5.0° 降至 3.6°，变化幅度是身份基准测试平均变化的四至七倍；EmoNet 距离从 0.14 降至 0.09，以及从 0.37 降至 0.24。除伪影外，使用和不使用提示词时的结果看起来相同。

ter 身份迁移（图9）。

编辑。编辑会将条件预测中的固定提示词替换为描述该编辑的提示词，并使用面部+头发掩码。图1 中的编辑使用了“一幅头发紧密卷曲的人脸照片级真实感肖像”、“一幅拥有铂金金发的人脸照片级真实感肖像”以及“一幅头发上戴着鲜花花冠的人脸照片级真实感肖像”，并使用相同的负面提示词。训练仅使用固定提示词。

运行时间。256 像素模型在一块 H200 GPU 上生成结果需要 17 到 21 秒，这是身份基准测试中两次遍历的平均值；戴眼镜和不戴眼镜的结果没有分别计时。

512 像素模型。512 像素模型从 256 像素阶段 2 模型开始训练，训练 100k 步，学习率为 10^{-5}，批次大小为 16，使用 8 个 GPU，并采用阶段 2 的数据和策略。低秩更新的秩从 64 补零扩展到 256，因为 512 像素图像的潜变量令牌数量是其四倍，缩放因子设为 256；随机块掩码的块大小为 32 像素。其光照流以 2×10^{-5} 的学习率训练 50k 步，使用 32×32 光照网格，因此每个单元再次覆盖一个潜变量令牌，即 16×16 像素。推理时，面部掩码膨胀 24 像素，遮挡物膨胀 3 像素。评分时，将两个模型的结果都调整为 448像素。图 10 展示了各模型原始分辨率下的结果：512 像素模型能够渲染更精细的细节，例如单颗牙齿和胡须中的毛发。

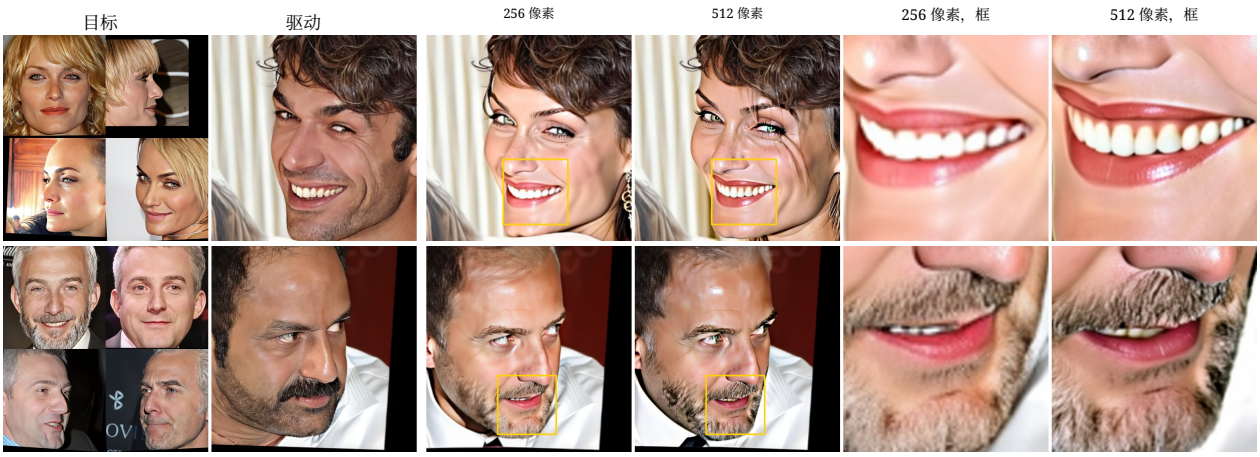


图 10。 **256 像素模型和 512 像素模型**。 两个模型在相同目标、驱动图像和随机种子下的结果，每个结果均采用自身的分辨率；256 像素结果通过双三次插值放大至 512 像素。最后两列放大了框选区域。512 像素模型展示了单颗牙齿（顶部）和胡须中的单根毛发（底部）。目标：其 24 张照片中的四张。

B. 基准测试

B.1. 身份基准测试

照片。 身份基准由 VGGFace2- HQ 构建 [9]，这是 VGGFace 的一个版本2 [5] 其照片对齐至 512像素，并使用 GFPGAN 修复 [70]。我们选取约 1k 人的子集，并清理每个人照片中的身份噪声。我们为照片构建一个 k近邻图（k=8），当 ArcFace 和 AdaFace 的平均余弦相似度至少为 0。30时，在两者之间连边，并保留其最大的连通分量。在这些照片中，我们保留 AdaFace 与该人物正面照片质心的相似度至少为 0。75的照片，阈值根据偏航角进行缩放。每张保留的照片还必须具有至少 [8]的面部质量分数 0。5，且鼻子和嘴巴清晰可见、仅有一张面部，并且不戴太阳镜。

给定照片和留出照片。 每个人有 24 张供方法接收的照片，以及 24 张用于身份评分的留出照片；清理后子集中的全部 999 人均达到这些数量。当一张留出照片与任意给定照片的平均余弦相似度达到 0。90 或更高时，将跳过该留出照片，以确保留出照片不是给定照片的近重复；如果这样会导致照片数量过少，则将阈值提高到 0。95，然后提高到 0。97。给定照片遵循固定顺序；采用 M 张照片的方法接收前 M 张照片，因此，较小数量照片对应的照片集是较大数量照片集的子集。该顺序首先从一张近正面照片开始：从偏航角最小的三张照片中选取一张并通过人工目检确认；其余照片则按照偏航角和俯仰角的最远点顺序排列，因此每增加一张照片，都会展示差异最大的姿态。使用单张照片的方法接收在24张照片（第 C节）中人脸质量分数最高的近正面照片，而不是该顺序中的第一张照片。这些照片来自不同于我们训练数据的来源，因此两组之间没有共享图像。为

表7. **不包括与我们的训练身份最接近的目标对象**。在身份基准的全部240对图像上使用ArcFace；并排除嵌入与训练数据中的某个身份的余弦相似度至少为0.5（剩余170对）或0.4（剩余165对）的目标对象。斜体值超过目标人物真实照片的值。

方法	All	< 0.5	< 0.4
SimSwap	0.500	0.496	0.495
FaceDancer	0.394	0.386	0.386
BlendFace	0.487	0.471	0.473
GHOST	0.548	0.545	0.545
InSwapper	*0.715*	*0.714*	*0.714*
HifiFace	0.528	0.525	0.529
Face-Adapter	0.501	0.505	0.505
DreamID-V	0.550	0.558	0.558
Ours	0.601	0.603	0.603

为检查我们的结果是否依赖于可能也出现在训练数据中的人物，我们将每个目标对象八张照片的平均人脸嵌入，与训练数据中人物的身份进行比较（对 AdaFace 和 ArcFace 取平均）。在这70 个240 个目标对象中，达到至少0.5 的余弦相似度、与其中一人相似的目标对象被排除后，表1 中各方法的 ArcFace 得分变化均小于0.02，且顺序保持不变（表7）。排除这些目标对象以及32 个 176 名人目标对象后，评审员在相似度评测中有84.4% 的情况下更偏好我们的结果，而不是83.6%（全部14 种方法）。

条件。该基准测试包含240 个目标对象，分为四种条件，每种60个，依据目标对象与驱动图像中的人物之间的差异程度划分（表8）。性别和年龄数据来自 In-sightFace [30]，，数值为每个人八张照片的平均值；皮肤亮度则是人脸解析模型 [75]在0–255 尺度上测得的皮肤中位亮度。族裔类别由多模态语言模型对每个人的三张

张照片进行投票得出；我们仅用它来构建图像对。

表 8. **身份基准测试的条件**。 目标与驱动图像中人物之间的差异。年龄差以年为单位；皮肤亮度差采用 0–255 的尺度。–：不受限制。每种条件下 60 对图像。

条件	Sex	族裔	年龄差	亮度差
A	same	same	≤ 15	≤ 30
B	不同	same	≤ 15	≤ 30
C	same	不同	≤ 15	≥ 30
D	same	same	≥ 30	–

驱动图像。 每个目标对象都与同一子集中另一人的一张照片配对，该照片作为其驱动图像。每个人在所有目标对象和驱动图像中仅出现一次。每种方法接收相同的驱动图像，即一张围绕面部裁剪的 448 像素裁剪图。

名人目标。 用户研究（第 4.2节）还包括176 名人目标对象，其面部可能为评审员所熟知。我们从维基共享资源中通过跨领域和族裔搜索收集了他们的照片；每张照片均保留其自身许可证。每位名人有10 至 24 张照片（其中60位拥有24张），且没有留出照片，因此每种方法都会接收某位名人的全部照片，而名人配对仅进入用户研究。他们的驱动图像展示的是其他人，这些人并非名人：125张来自 FFHQ [33]，另有 51 张来自身份基准的驱动图像，并依据四种条件的规则进行分配。

B.2. 形状基准测试

角色与渲染图。 形状基准测试包含在 Unreal Engine 5.8 中渲染的 29 个 MetaHuman [16] 角色。这些角色是不同性别、年龄和族群的 MetaHuman 默认预设，并以最高质量设置构建。每个角色均以 2048×1080px 分辨率渲染，包含 53 种表情：一个中性表情、面部动作编码系统的 45 个动作单元 [14]以及 7 种基本情绪 [15]。光照来自 30 个 4K 分辨率的 HDRI 环境，涵盖室内和室外，以及白天、傍晚和夜间；每种环境分别进行单独渲染，并在添加光源后再次渲染，从而得到 60 种光照条件。摄像机覆盖前半球的 102 个视点（$\pm90^\circ$ 方位角和仰角）；我们使用方位角范围 $\pm70^\circ$ 和仰角范围 $\pm30^\circ$，在该范围内面部始终完全可见。每个角色都是 12 对图像的目标。驱动图像展示该集合中的其他角色，呈中性表情、面向相机，并处于一种固定光照下。真实值展示目标角色在相同相机、表情和光照下的结果。每种方法都会接收目标角色的另外 24 张渲染图。它们的光照、表情和相机分别从 60 种光照条件、45 个动作单元和 7 种情绪，以及上述相机范围中随机抽取。渲染图按照固定顺序排列：最接近正面的渲染图（头部旋转低于 20°）首先出现，接下来是面部质量模型评分最高的图像，其他图像则按照采样顺序排列。对身份输入取平均的方法会使用全部 24 张渲染图，而只接收一张照片的 DreamID-V 则使用第一张图。对于 14 组 348 图像对，驱动角色的头发遮挡了结果中的下颌线点，因此我们用该集合中的其他角色替换了它们的驱动图像。图 11 展示了其中四个角色及其渲染图。

轮廓距离与泄漏。 我们使用 MediaPipe [34, 48]沿下颌线检测 21 个点，并排除真实值或驱动图像中被头发遮挡的点，以确保所有方法都在相同的点上评分。结果、驱动图像和真实值使用同一台相机，因此我们在图像坐标中对它们进行比较。轮廓距离 s_{gt} 是结果中的下颌线点到真实值下颌线的平均距离，以两眼间距的一半为单位。泄漏是指在某个目标的各张驱动图像中，结果的下颌线随驱动图像下颌线变化的斜率；该斜率沿垂直于真实值下颌线的方向测量（0：无泄漏；1：完全复制）。

三维误差。 三维误差比较 Deep3DFaceRecon [12] 从结果和真实值重建的面部形状。在重建之前，我们在两张图像中将真实值或驱动图像里被头发遮挡的区域涂成灰色。三维误差是两个重建面部在去除表情与姿态后的平均距离，计算范围为涂色区域之外的顶点，单位为两眼间距的百分比。

参考值。 同一角色在其他光照下渲染时，轮廓距离得分为 0。024，三维误差为 1。14，这就是两个指标的噪声水平。驱动图像本身的得分分别为 0。128 和 3。49。

到达比率。 作为对同一组点的进一步解读，到达比率 $s_{\text{drv}}/(s_{\text{drv}}+s_{\text{gt}})$ 使用结果的下颌线点与驱动图像下颌线之间的距离 s_{drv}。对于位于驱动图像下颌线上的结果，该值为 0；对于位于真实值下颌线上的结果，该值为 1。

三维误差的版本。三维误差的两个更简单版本比较的是80重建结果的身份系数，而不是网格：(a) 使用原始图像，以及 (b) 将头发区域涂成灰色后。表9比较了这三个版本在噪声水平、驱动图像和纯脸型变化下的表现。三者都能将驱动图像与噪声水平区分开来，差异倍数为2.3 到 3.1倍。仅对下颌线进行变形时，网格距离仅略高于其噪声水平；而改变面部宽高比时，网格距离则接近或高于驱动图像。因此，三维误差主要反映面部的比例，并可与读取下颌线的轮廓距离互为补充。

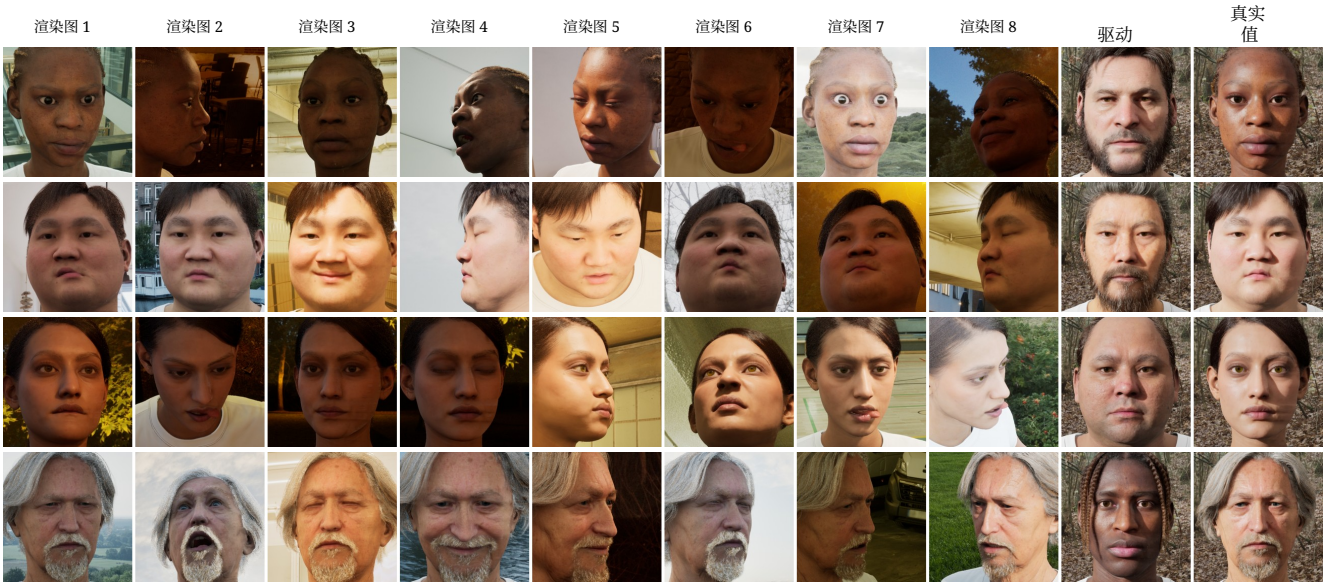


图 11. **形状基准测试**。 29 个角色中的 4 个。左：方法接收的 24 张渲染图中，按固定顺序排列的前 8 张。右：一张驱动图像（另一个面带中性表情、面向相机、处于固定光照下的角色），以及该配对的真实值（目标在相同相机、表情和光照下的图像）。

表 9。**形状基准测试中的三维误差版本**。(a)、(b)：80结果与真实值重建的身份系数之间的 L2 距离，分别对应未将头发区域涂成灰色和涂成灰色的情况。(c)：第B.2节中的网格距离，以眼睛之间距离的百分比表示。噪声：同一角色在另一种光照下的情况（29 对图像）。驱动：将驱动图像作为结果进行评分（348）。变形：仅在脸型上对每个角色的真实值进行变形，将下颌线缩放为0.88或1.12，或将宽高比缩放为0.80 或1.25 (各29 对)。

			下颌线变形		比例变形	
版本	噪声	驱动	0.88	1.12	0.80	1.25
(a) 系数	2.56	6.32	2.84	2.62	4.07	3.68
（b）系数，已绘制	2.59	6.05	2.12	1.79	4.12	3.82
（c）网格（我们的方法）	1.14	3.49	1.46	1.22	4.03	3.38

C. 基线方法

表10 列出了每种方法所使用的代码、检查点、工作分辨率以及目标人物照片的使用情况。每种方法接收相同的驱动图像、Sec. 448px 的裁剪图像，以及目标人物的原始照片。对于其人脸检测器需要更多上下文的方法，会使用带有黑色边框的裁剪图像，黑色边框宽度为图像尺寸的 B.1，并将其结果粘贴回 0.25 至 0.3 的 448px 帧中。其他分辨率的输出会调整为 448px。

对身份输入取平均值。SimSwap、FaceDancer、BlendFace、GHOST、InSwapper、HifiFace 和 Face-Adapter 使用一张照片的身份嵌入。我们将它们扩展到 M 张照片：对嵌入取平均值，并按各方法自身的方式进行标准化，

HifiFace 和 Face-Adapter 还会对与嵌入一同使用的三维身份系数取平均值。此扩展由我们提出，并能提高它们的识别分数（第4.4节），因此它们的最佳设置会使用全部 24 张照片。HifiFace 没有官方代码。我们使用公开的重新实现，并在评分前依据一条固定规则，在1.0的裁剪比例下，对40 组留出配对进行选择。

使用单张照片的方法。 DreamID-V、DiffSwap、REFace、DiffFace、E4S 和 Gemini-Pro 接收一张照片：在所有照片中，使用 6DRepNet 估计的头部旋转 [24], 低于20°阈值的照片中人脸质量分数最高的照片；如果没有符合条件的照片，则使用最正面的照片。DreamID-V 是一个视频模型。我们将驱动图像重复为包含 21 帧的静态视频，并保留第 18帧。在形状基准测试中，使用平均结果的方法采用全部 24个渲染图，而其他方法采用第一个渲染图（Sec. B.2）。

商业图像模型。 我们提示两个商业图像模型进行身份迁移：GPT-Image 使用驱动图像和目标的前 8 张照片，Gem- ini 3 Pro Image（Nano Banana Pro）使用驱动图像和一张按上方方式选择的照片。两者接收相同的提示词，其中 GPT-Image 使用 n=9，Gemini-Pro 使用 n=2:

> “将图像 1 中的面部替换为图像 2 至 $\{n\}$ 中人物的面部。使用所有参考照片，尽可能忠实地再现其脸型、前额、眉毛、眼睛、鼻子、嘴巴、下颌、耳朵、肤色、年龄和面部毛发。保留图像 1 的头部姿态、目光、表情、光照、背景，以及图像 1 中的发型和发色（头发须与图像 1 相同；不要复制参考图像中的头发或头饰）。不要美化；保持照片级

表 10. **基线方法**。 各方法的公开代码和检查点、生成面部的分辨率，以及使用目标人物照片的方式。24 张照片均值：身份输入取自第 24 张照片的平均值，见第 B.1节。单张照片：人脸质量模型 [8]评分最高的近正面照片。

方法	代码和检查点	分辨率	目标人物照片
SimSwap [7]	“people”模型，人脸解析回贴	224	24 个 ArcFace 嵌入的均值，重新归一化
FaceDancer [58]	配置C，高清	256	24 个 ArcFace 嵌入的均值
BlendFace [63]	使用 BlendFace编码器的 AEI-Net	256	24 个单位嵌入的平均值
GHOST [18]	U-Net 生成器，2 个块，无超分辨率	256	24 个嵌入的平均值
InSwapper [31]	inswapper 128，InsightFace buffalo l	128	24 个归一化 ArcFace 嵌入的均值
HifiFace [71]	maum-ai 的重新实现，基于 VGGFace2 训练	256	24 个 ArcFace 嵌入和三维系数的均值
Face-Adapter [22]	发布的检查点，sd-vae-ft-mse	512	24 个 ArcFace 嵌入和三维系数的平均值
DreamID-V [20]	搭载 DWPose 的 Wan-1.3B 模型	832 × 480	一张照片
DiffSwap [80]	去噪扩散隐式模型，200 步，目标尺度 0.01	256	一张照片
REFace [3]	FFHQ 配置	512	一张照片
DiffFace [36]	4 次迭代	256	一张照片
E4S [47]	官方流程	1024	一张照片
GPT-Image	gpt-image-2.5-sunburst，图像编辑端点	1024	前 8 张照片
Gemini-Pro	gemini-3-pro-image (Nano Banana Pro)，批处理端点	1024	一张照片

真实感。”

Gemini-Pro 会在每张图像前附上说明：“图像 1（驱动照片，编辑此图）：”和“图像 2（目标人物）：”。我们从七种提示词中选出这一种，依据是结果保留驱动图像中头发的效果，以及与目标人物的相似程度。对于 GPT-Image，我们在 1、3、5、8 和 15 张照片上进行测试，对48 组图像对的测试结果表明，使用超过 3 张照片并未提升身份相似度，因此我们使用8张。对于 Gemini-Pro，使用三张照片的结果不如使用一张。Gemini-Pro 使用默认采样设置。两种模型也都在形状基准测试上运行：GPT-Image 使用目标的前 8 张渲染图，Gemini-Pro 使用第一张（表21）。

缺失结果。有些方法对少数驱动图像未能返回结果，每张表都会报告评分结果的数量 n；缺失结果不会重试，也不会进行插补。REFace、DiffFace 和 E4S 对相同的 240 组图像对中的 17 组未能返回结果，因为它们的人脸对齐器无法在驱动图像中检测到面部；DiffFace 使用 REFace 的对齐器。DiffSwap 对 14 组未能返回结果，其中 12 组属于上述 17 组。GPT-Image 对 16 组未返回图像；未记录每组未返回图像的原因。

D. 指标

本节定义表 1 及补充材料中的指标，并说明每项指标可能产生误导之处。所有指标均在身份基准的 240 组图像对上计算，并根据各方法返回的结果进行评分。

身份。身份是结果的识别嵌入与目标的24张留出照片的识别嵌入之间的平均余弦相似度。我们使用三种识别器：ArcFace [11] （IResNet-50）、AdaFace [37] （IR-101）和 FaceNet [61]（Inception-ResNet-v1）。我们使用的 FaceNet 模型是在 VGGFace 上训练的2 [5], , 身份基准测试的目标对象也来自该数据集。表11 给出了每种方法各有两个参考点，评分方式相同。第一个是每个目标对象的第一张给定照片，即该人物的真实照片，也是完美结果的标准。第二个是将同一张照片与另一目标对象的留出照片进行比较所得的分数。结果的得分可能高于真实照片，例如 InSwap- per 在 AdaFace 和 ArcFace 下的表现。这样的分数反映的是识别器如何解读结果，并不代表结果比该人物的照片更相似，因此我们用斜体标出，不参与排名。AdaFace 是阶段 1 的身份损失的一部分（第3节），因此正文报告的是 ArcFace 的结果，因为我们的损失函数都没有使用 ArcFace。在名人目标对象上（两者的结果均有报告），使用给定照片而非留出照片进行评分，不会改变方法的排序（表18）。

表 11. **身份得分的参考点**。 与目标对象的 24 张留出照片（240 个目标对象）的平均余弦相似度。

	AdaFace	ArcFace	FaceNet
目标人物的真实照片	0.656	0.677	0.736
另一个人的照片	0.012	0.011	0.012

表情。 表情是 EmoNet [66] 对结果图像和驱动图像预测的效价与唤醒度之间的 L2 距离。EmoNet 读取每张图像中分别检测到的人脸框，并将其裁剪为人脸框 1。2 倍的正方形区域，再调整为 256像素。我们还报告了三项比较面部参数而非情绪的测量指标：SMIRK 的表情误差 [57], , 即其 50 个表情、2 个眼睑和3 个下颌参数的平均绝对差；52 个 MediaPipe 混合形状的平均绝对差 [48]；以及 Deep3DFaceRecon 的 64 个表情系数之间的 L2 距离 [12]。在表 1的九种方法中，它们对方法的排名与 EmoNet 相似，斯皮尔曼相关系数分别为0。93 （SMIRK）、0。97（混合形状）和0。93 （Deep3DFaceRecon）。在

这四项指标中，我们的结果都不如 SimSwap 和 FaceDancer 更贴近表情，但比 GHOST、HifiFace、Face-Adapter 和 DreamID- V 更贴近表情。以驱动图像为基准进行测量时，这些分数会奖励保留驱动图像面部的结果（第 D.1节）。两项更狭义的测量指标的一致性较低：嘴巴张开程度（0。72）以及 LibreFace 的 12个动作单元的强度 [6]（0。72），后者也会受到光照影响。

姿态。 姿态是由两种估计器预测的结果与驱动图像的头部旋转之间的测地角。6DRepNet [24] 读取完整的 448 像素帧，并将其调整为 224像素。其权重是在预训练期间用于监督几何分支头部姿态的权重（第 3.1节），因此主论文报告 HopeNet [59], 读取驱动图像人脸框周围的裁剪区域，两幅图像均使用相同的裁剪区域。HopeNet 的偏航角、俯仰角和翻滚角用于计算旋转。两种估计器对表1中的九种方法给出的排序相同（斯皮尔曼相关系数为 0.97），且两者都将我们的结果排在第五位。

光照。 光照是结果与驱动图像在人脸区域内的低通亮度之间的皮尔逊相关系数。我们将两幅 256像素图像转换为亮度图，并用高斯核进行模糊处理，核大小为 σ=8像素；随后在驱动图像中由人脸解析模型 [75] 识别出的人脸区域内进行比较：皮肤、眉毛、眼睛、眼镜、鼻子、嘴巴和嘴唇，不包括耳朵、颈部和头发。该分数奖励保留驱动图像脸型和皮肤的结果：驱动图像本身的得分为1；SimSwap、FaceDancer、BlendFace、GHOST、InSwapper 和 HifiFace 保留了大部分下颌线（泄漏值为 0.78 至0.95），得分为 0.90 至 0.96。因此，高分本身并不能说明结果在改变身份的同时保留了光照。在相同区域内，表 16 还给出了平均亮度的绝对差值（ΔL，范围为 0–255）以及结果与驱动图像之间平均 CIELAB a^* 和 b^*（Δab）的 L2 距离。

真实感。 真实感无需参考图像即可测量。FaceScore [42] 是针对人脸质量学习得到的奖励函数，而 DSL-FIQA [8] 是一种人脸质量模型，它还会筛选身份基准测试中的照片，并为需要输入照片的方法选出一张照片（第 B.1 节和 C节）；它会对驱动图像中人脸框周围的裁剪区域进行评分。评审员也会在用户研究中评判真实感（第E节）。

D.1. 形状变化下的表情、姿态和光照测量指标

仅改变脸型就会影响我们使用的每个表情与姿态估计器。表12 对表情和姿态在构造上相同的图像进行评分。我们仅改变身份基准中每张驱动图像的脸型，使用自增强（第3.3节）中的变形，然后将其与原图进行比较评分。在形状基准测试中，我们比较了两个不同角色的渲染图，它们使用相同的相机、表情和光照，以及同一角色在两种光照下的图像。将下颌线变形12% 会使 EmoNet 达到表45 至56% 中各方法差异范围的水平，另外三个表情测量指标则达到约100 至140%的差异范围。对于每个表情测量指标，具有相同表情、姿态和光照的两个不同角色之间的差异都超过了各方法之间的整个差异范围（1.6 至3.7 倍）；对于两个姿态估计器，差异则达到其42 至 60% ；这分别是同一角色在两种光照下差异的2 至4倍。因此，在这些测量指标上比较两种方法时，会将表情或姿态差异与其结果之间的脸型差异混杂在一起。

光照测量指标的表现也相同（表 12 的最后三行）。相同光照下的两个角色，其光照相关性仅为 0。80，低于表 1 中九种方法里的八种；二者的平均亮度相差 30。3，超过所有这些方法。差异与肤色有关：肤色最接近的三分之一图像对中，差异为 10。4；肤色相差最大的三分之一图像对中，则为 55。0，而相关性仅从0。82 变为 0。79。因此，在光照相同的情况下改变人物的结果会降低光照得分，而 ΔL 和 Δab衡量肤色的程度与衡量光照的程度相当。

在形状基准测试中，真实值是已知的，因此每个结果还可以与真实值进行比较评分，而不是与驱动图像进行比较（表 13）。与驱动图像比较时，真实值本身在表1 的九种方法中，针对每项测量指标都会排在其中五至八种方法之后。与真实值比较时，各方法的排序发生变化：Spearman ρ 两种排序之间的相关性为 0。80 （EmoNet）、0。30（SMIRK）、 -0.23 （HopeNet）、0。43 （6DRepNet）以及0。12 （光照相关性）。与真实值比较时，我们的结果在 SMIRK、HopeNet 和 6DRepNet 上在九种方法中排名第一，在光照上排名第三，在 EmoNet 上排名第四；而与驱动图像比较时，排名分别为第五、第七、第六、第七和第二。该基准测试中的驱动图像面向相机，呈中性表情（第 B.2 节），因此这些分数涵盖的表情与姿态范围较窄。在这一范围内，与驱动图像进行比较会奖励那些保留驱动图像面部特征的结果。

E. 用户研究

人类偏好是我们衡量身份的两项指标之一（第 4.2节），另一项是识别分数。每次试验都会在同一驱动图像上，将我们的结果与某个基线方法的结果进行比较。评分者会看到这些方法所接收的三张目标人物照片、驱动图像，以及并排呈现的两项结果。随后，评分者通过点击其中一个结果回答两个问题：哪项结果看起来更像目标人物（相似度），以及哪项看起来更逼真（真实感）。不提供平局选项。在每种方法的各个图板中，我们的结果交替出现在左右两侧，并在

表12。 **仅改变脸型时的表情、姿态与光照度量**。 每列比较两张表情与姿态相同的图像，使用第 D节中的测量指标。在光照行中，其他人物保持光照不变而更换人物，其他光照保持人物不变而更换光照；此处省略自增强，因为其黑色填充有一部分位于人脸区域内。下颌线：将驱动图像（240）的下颌线通过移动最小二乘法缩放 0。88 或1。12 。自增强：按照训练方式为每张驱动图像生成三个变形，同时也会改变宽高比，并且平均在帧上留下 13% 的黑色填充。其他人物：形状基准测试中处于相同姿态、表情、相机视角和光照下的两个角色（348 图像对）。其他光照：同一角色在两种光照下的图像（29）。方法：表 1中九种方法的最低值和最高值。HopeNet：表 1中的测地角。

	身份基准，已变形			形状基准测试		
测量指标	下颌线 0.88	下颌线 1.12	自增强	其他人物	其他光照	方法
HopeNet（度）	2.30	2.23	3.11	3.32	1.45	2.03–7.52
6DRepNet（度）	1.55	1.40	1.79	2.62	1.06	1.99–8.18
EmoNet	0.094	0.075	0.123	0.263	0.095	0.205–0.372
SMIRK	0.311	0.311	0.396	0.553	0.263	0.420–0.696
混合形状	0.026	0.023	0.027	0.050	0.013	0.036–0.058
Deep3DFaceRecon	1.14	1.04	1.17	2.99	1.29	2.48–3.29
光照相关性 ↑	0.943	0.962	–	0.802	0.930	0.52–0.96
ΔL	5.2	3.4	–	30.3	24.1	2.4–23.9
Δab	1.20	0.46	–	3.93	14.77	0.82–5.88

表13。 **相对于驱动图像（D）或真实值（GT）的表情、姿态和光照** 在形状基准测试（348 图像对）上的结果，涵盖表1中的九种方法。第一行是将真实值与驱动图像进行比较所得的分数。HopeNet 的人脸框和光照相关性的区域取自与结果进行比较的图像。粗体：各列九种方法中的最佳结果。

	EmoNet ↓		SMIRK ↓		HopeNet（度）↓		6DRepNet（度）↓		光照相关性 ↑	
方法	D	GT	D	GT	D	GT	D	GT	D	GT
真实值	0.262	–	0.550	–	3.33	–	2.64	–	0.802	–
SimSwap	0.221	0.256	0.411	0.457	**1.60**	3.42	**1.52**	3.12	0.970	0.847
FaceDancer	0.203	0.263	0.409	0.450	1.90	3.59	1.61	3.37	**0.975**	0.850
BlendFace	**0.142**	0.215	**0.348**	0.440	2.14	3.61	1.75	3.00	0.952	0.847
GHOST	0.389	0.367	0.558	0.539	1.81	3.44	1.74	3.08	0.957	0.832
InSwapper	0.201	**0.189**	0.448	0.385	2.39	3.41	1.82	3.21	0.953	**0.880**
HifiFace	0.384	0.413	0.776	0.736	2.71	3.49	2.49	3.43	0.945	0.826
Face-Adapter	0.303	0.225	0.578	0.454	3.43	3.37	2.73	3.24	0.907	0.876
DreamID-V	0.363	0.320	0.606	0.401	3.52	3.53	3.77	3.96	0.497	0.592
Ours	0.186	0.234	0.498	**0.337**	2.96	**3.28**	2.10	**2.64**	0.915	0.872

图板的构建方式使得我们的结果有一半位于左侧。评审员阅读了以下说明（原文照录）：

> 目标。 根据两项标准评估并比较两种换脸方法：相似度：生成的面部在多大程度上保留了目标人物的身份。真实感：生成的面部看起来有多自然、逼真。图像参考指南。 左上方图像（绿色边框）：展示目标人物身份的照片。右上方图像（红色边框）：原始驱动图像，提供面部表情和背景。底排图像：标记为 LEFT 和 RIGHT 的生成结果，供评估。任务说明。 对于每项标准（相似度优先和真实感优先），从底排中选择一张图像（LEFT 或 RIGHT）。两项标准可以选择同一张图像，也可以分别选择不同图像。请勿从顶排图像中作出选择。每项标准只能选择一张图像。

每种方法最多有 416 个图板。其中，240 个图板展示了身份基准测试的结果，这些结果在表1 和表 16中进行了评分；其余 176 个图板展示了名人目标人物的结果，他们每人有 10 至 24 张照片。对于部分驱动图像未返回结果的方法，图板数量会更少（至少387 个）。每种方法均以其最佳设置运行（第C节）。

该研究在 Labelbox 上分两个批次进行。在第一批次中，30 名评审员分别评判了 50 个图板，并将其与表1中的八种基线方法中的七种进行比较。在第二批次中，30 名评审员中有 18 人也参加了第一批次，他们分别评判了约50 个图板，并将其与 HifiFace 及第 C 节中的另外六种方法进行比较。在两个批次中，共有 42 名评审员给出了 2,983 个相似度标签和 2,988个真实感标签。同一名评审员给出的标签之间存在相关性，因此 95%置信区间通过对评审员进行自助法抽样获得，共进行了5,000 次重抽样。评审员属于该平台的标注人员队伍。在每个批次中，每名评审员都评判一个项目的 50 个图板，而一个项目中的每个图板只被评判一次；

表14. **用户研究**：评审员选择我们结果的试验占比（%），以及对评审员进行自助法得出的95%区间。相似度：哪个结果看起来更像目标。真实感：哪个结果看起来更真实。n：标签。

基线方法	相似度	n	真实感	n
SimSwap	96.5 [93.1、99.1]	200	97.5 [94.8，99.5]	200
FaceDancer	80.7 [74.2，86.8]	207	62.1 [54.4，70.2]	206
BlendFace	92.4 [88.0，96.2]	210	92.0 [87.2, 95.9]	212
GHOST	98.7 [97.1, 100]	226	99.6 [98.6, 100]	226
InSwapper	97.6 [95.3，99.5]	210	97.6 [94.5, 100]	210
HifiFace	100 [100, 100]	235	100 [100, 100]	236
Face-Adapter	73.4 [67.5，79.0]	229	78.6 [73.3，83.8]	229
DreamID-V	79.4 [74.4，84.4]	209	81.3 [75.2，87.2]	209
DiffSwap	91.9 [87.5，95.7]	197	85.8 [77.9，92.2]	197
REFace	81.7 [75.1, 87.5]	218	77.5 [69.6、84.5]	218
DiffFace	90.6 [85.6，94.8]	212	85.8 [79.2、91.7]	212
E4S	89.1 [84.1, 93.6]	211	84.8 [79.1，89.8]	210
GPT-Image	37.4 [29.8，45.6]	211	15.0 [9.5，21.2]	214
Gemini-Pro	60.6 [53.2，67.5]	208	39.2 [31.2，46.9]	209
表1	89.9 [88.2, 91.5]	1726	88.8 [87.2、90.4]	1728
另外六个	75.1 [71.9、78.2]	1257	64.4 [60.4，68.0]	1260
全部十四个	83.6 [81.7，85.5]	2983	78.5 [75.9，80.8]	2988

这些图板分布在 60 个项目中，因此其中 530 个图板被多次评判。每个图板都将我们的结果与某个基线方法的结果进行比较。一次标注在第一批次和第二批次的中位耗时分别为 79 秒和 66 秒。评审员在相似度试验中选择左侧结果的次数为 1,506 次，选择右侧结果的次数为 1,477 次。

评审员认为，我们的方法的结果比每种方法的结果更像目标对象（表14）。与表1中的八个基线相比，他们在89.9%的试验中选择我们的方法的结果，其中相较于Face-Adapter的比例为73.4%，相较于HifiFace则为100%。与另外四种方法相比，该比例介于81.7%（REFace）和91.9%（DiffSwap）之间。两种通用图像模型的比较结果有所不同：在60.6%的试验中，评审员选择我们的方法的结果而非Gemini-Pro的结果；而在62.6%的试验中，他们选择GPT-Image的结果而非我们的结果。在两组目标对象中，GPT-Image都更受偏好；评审员选择我们的方法的结果的比例分别为34.4%（身份基准）和41.6%（名人）。评审员还认为，在85.0%的试验中，GPT-Image的结果更具真实感，并且他们对相似度的回答往往与对真实感的回答一致（见下方）。

每位评审员都表现出这一偏好。第一批的 30名评审员中，每个人在针对八种基线方法的试验中选择我们结果的比例为75.5% 至96.0%（中位数为 90.0%）。第二批的 30 名评审员中，每个人在针对另外六种方法的试验中选择我们结果的比例为55.3%至 90.9%（中位数为 74.4%）。在身份基准和名人目标上的比例也大致相同；评审员可能熟悉这些目标对象的面部（表15）。

真实感是第二个问题。与方法相比，结果也呈现出相似的情况；与图像模型相比，则有所不同。在与八种基线方法的对比中，评审员在 88.8% 的试验中认为我们的结果更真实。最低比例是与 FaceDancer 的对比（62.1%），其余对比的比例介于 78.6% 至 100% 之间。与图像模型相比，评审员在 85.0% 的试验中认为 GPT-Image 的结果更真实，在 60.8% 的试验中认为 Gemini-Pro 的结果更真实。FaceScore 对这两者的评分也高于我们的结果（分别为 5.14 和 4.98，而我们的结果为 4.62；表16）。它给 DiffFace（4.72）和 E4S（4.70）的评分也高于我们的方法，但在与它们的对比中，评审员分别在 85.8% 和 84.8% 的试验中认为我们的结果更真实。按评审员统计，在与八种基线方法（第一批次大小）的对比中，真实感比例介于 78.6% 至 95.0% 之间；在与其他六种方法的对比中，介于 37.5% 至 80.0% 之间，其中 30 名评审员中有 26 名在过半数试验中选择了我们的结果。

在两组目标对象中，图像对的四种条件（B 节）之间的相似度差异很小：针对八种基线方法时，评审员在各条件下选择我们结果的试验比例分别为 90.0%（A）、90.8%（B）、90.0%（C）和 88.8%（D）。就真实感而言，条件 A 的比例最低（85.2%，而 B–D 为 89.7% 至 90.3%）。在由不止一名评审员进行评判的图板中，两名评审员在相似度评判中选择相同结果的图像对比例为 82.0%（530 个图板），在真实感评判中为 83.0%（532 个图板）。由于大多数标签都偏向我们的结果，两名评审员仅凭偶然机会也会在 72.6% 和 66.3% 的图像对中达成一致，因此，超出偶然一致的程度（κ）分别为 0.34 和 0.50。

这两种回答往往一致：在 86.6% 的试验中，评审员对相似度和真实感选择了相同的结果。当他们认为我们的方法的结果更真实时，在 94.7% 的这些试验中，他们认为该结果与目标更相似（$n = 2{,}344$）；当他们认为另一结果更真实时，这一比例为 43.0%（$n = 637$）。与 GPT-Image 相比，评审员在 211 次试验中有 179 次认为它的结果更真实；在这些试验中，他们有 28.5% 的时候认为我们的方法的结果与目标更相似。

F. 与先前工作的比较

表 16 列出了十二种对比方法、两种商业图像模型以及我们模型的四个变体在身份基准测试上的所有指标。表 17 按条件细分了身份分数，而表18 列出了用户研究中的名人目标指标（第E节）。第 D节定义了这些指标。

身份。 在 AdaFace 和 ArcFace 下，我们的方法的结果在低于目标人物的真实照片的结果中得分最高。FaceNet 在 VGGFace2 上训练，而基准测试的目标对象也来自该数据集（第 D 节）；在其下方，InSwap- per（0。706）、Face-Adapter（0。666）和 DreamID-V（0。652）的得分高于我们的方法（0。647）。使用面部+头发掩码时，我们的方法的结果在 AdaFace 和 ArcFace 下得分最高，并且在 FaceNet 下与 InSwapper 并列。给定一张照片而不是24张照片时，每种平均基线方法在 ArcFace 中的得分都会损失 0。06到 0。11 分，而其表情、姿态、光照和真实感得分基本保持不变，因此平均是其最佳设置。我们的得分在四种条件下变化很小（表17）。

表15. **按目标对象组别划分的用户研究**：评审员选择我们结果的试验占比（%），以及通过对评审员进行自助法计算得出的95%区间。IB：身份基准。名人：名人目标。

	相似度		真实感	
基线方法	IB	名人	IB	名人
表 1	90.3 [88.2，92.1]	89.3 [86.6，91.9]	88.9 [86.9，91.0]	88.6 [85.5，91.3]
其他六项	75.1 [70.7，79.3]	75.1 [71.4，79.0]	63.1 [58.9, 66.9]	66.1 [60.8，71.3]
全部十四个	84.0 [81.6, 86.3]	83.2 [80.4，85.6]	78.3 [75.7、80.7]	78.8 [75.5，82.2]

表情、姿态、光照和真实感。SMIRK 表情误差和混合形状距离对方法的排名与主论文中的 EmoNet 距离相同，而 6DRep-Net 对它们的排名与 HopeNet 相同。我们的结果与驱动图像差异最大的是平均亮度（ΔL 23.9，相比于2.4 到 17.3）。 ΔL 肤色与光照的变化程度相当（第 D.1节），并且我们的结果使皮肤变得更亮（第 L节）。在表1的方法中，我们的结果具有最高的 FaceScore；DiffFace、E4S 和两个图像模型的得分更高，而 GPT-Image 的 DSL-FIQA 也最高。在名人目标上（表 18），在 ArcFace 指标下，InSwapper 的得分再次高于其他方法，无论是对给定照片还是对留出集照片。

F.1. 定性比较

图 12和 13 展示了全部十四种方法在12 组用户研究图像对上的结果，每种条件各三组。

G. 表情、光照和遮挡物

G.1. 表情和头部姿态

我们从 MEAD 中获取驱动图像 [67]，，其中演员展示八种情绪，并同时由七台摄像机进行拍摄。我们使用阶段 1 的两个步骤均留出的四名受试者，因此他们的面部对于身份编码器而言是全新的；该编码器在阶段 1 的样本中仅见过 MEAD 的1%；几何分支则在全部 MEAD受试者上进行了预训练（第 A.1 节）。对于每名受试者，我们针对每种情绪从正面相机拍摄的画面中选取一帧：包括中性情绪和七种情绪在最高强度下的画面，具体选取 EmoNet [66] 判定为最清晰体现该情绪的帧；此外，还从每台相机中选取一帧惊讶表情图像。一名受试者仅由七台摄像机中的三台拍摄，因此得到52 张驱动图像。我们使用模型、InSwapper、Face-Adapter 和 DreamID-V，按照表1中的设置，将身份基准测试中的四个目标对象（其身份均与我们的训练数据中的任何身份都不相近）迁移到这些帧中。

表19 比较了每个结果与其驱动图像的差异。对于情绪，我们的结果最常属于驱动图像的 EmoNet 类别（0.68，而其他方法为 0.35 至 0.54）；对于相机视角（1.00，而其他方法为 0.86 至 0.89），我们的结果也最常匹配驱动图像，并且效价-唤醒度距离最低。InSwapper 在情绪方面的 SMIRK 表情误差略低，而且与表1一样，ArcFace 得分最高。在侧视角和俯视角下，我们的 ArcFace 得分保持在 0.53。

图14和15 展示了另外四位 MEAD 受试者，其中两位肤色较浅，两位肤色较深；每位受试者都与身份基准测试中同性别且肤色相近的目标配对。在图14中，受试者展示了八种情绪；在图 15中，则是在七台摄像机下呈现中性表情。我们的结果呈现了惊讶和恐惧时张开的嘴、厌恶和愤怒时露出的牙齿，以及悲伤时的皱眉；同时，无论从一个侧面转向另一个侧面，还是从上方或下方观察，都保留了目标的面部特征。图16 比较了其中两位受试者上的各方法：InSwapper 和 Face-Adapter 的结果表情较弱，而 DreamID-V 则重新生成了头发和头部。对于四位受试者的各种情绪，我们的结果有66% 的案例与驱动图像的 EmoNet 类别一致，而50% 的案例中 InSwapper 和 DreamID-V 达到这一结果，Face-Adapter 则为30% 。

G.2. 光照

在电影中，面部常常会受到硬光或彩色光的照射：一道阳光条纹、穿过树叶的光、投影图案，或有色侧光。图 17 展示了这种光照下的结果，驱动图像来自网络和 FFHQ [33]，目标图像来自 VGGFace2-HQ；它们既不是身份基准测试中的目标对象，也与我们的训练数据中的任何身份都不接近（第 B 节）。我们的模型采用默认设置运行，使用变形进行两次处理（第 3.4 节）。InSwapper、DreamID-V 和 Face-Adapter 采用表 1 中的设置。

我们的方法将阳光条纹、穿过树叶的光、投影图案和彩色光迁移到面部上。InSwapper 也保留了驱动图像的光照，Face-Adapter 保留了其中一部分，而 DreamID-V 则生成了处于自身光照条件下的面部。

最后一列使用了第二个光照流，在光照更接近与身份更清晰之间提供了权衡；用户可以针对每项任务在两个流之间进行选择。它读取更大的输入区域，除了驱动图像中的面部，还涵盖耳朵和颈部。它在阶段 3 使用另一个第2阶段模型进行训练，并原样用于我们的方法。它能更贴近地呈现光照，例如投影图案、彩虹，以及深色皮肤上的彩色光（第 3、5、7 和 8 行）。在身份基准测试中，其光照相关性从 0.83 提升至 0.84，而 ArcFace 得分则从 0.601降至 0.589 （表 20）。人脸随机化

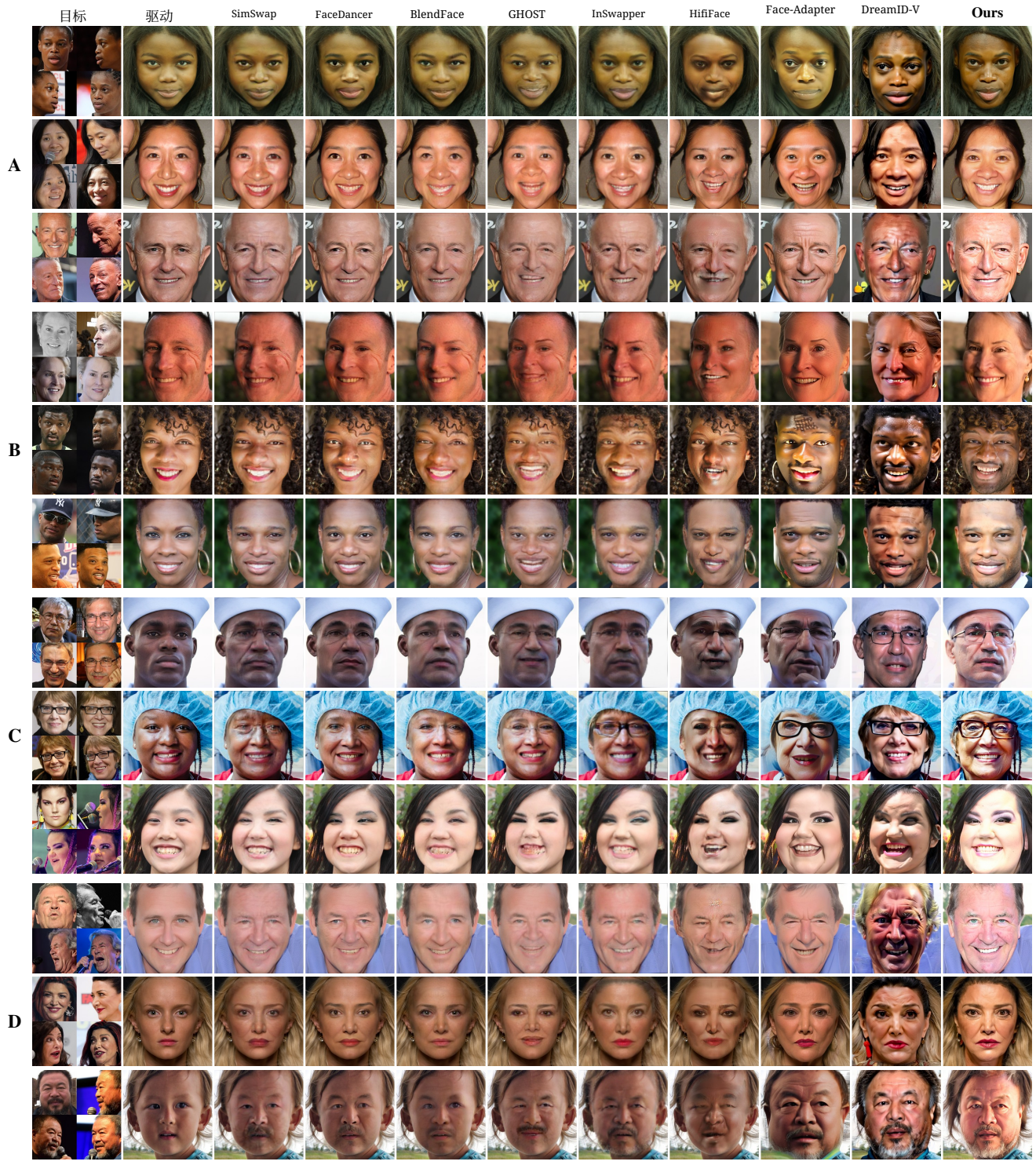
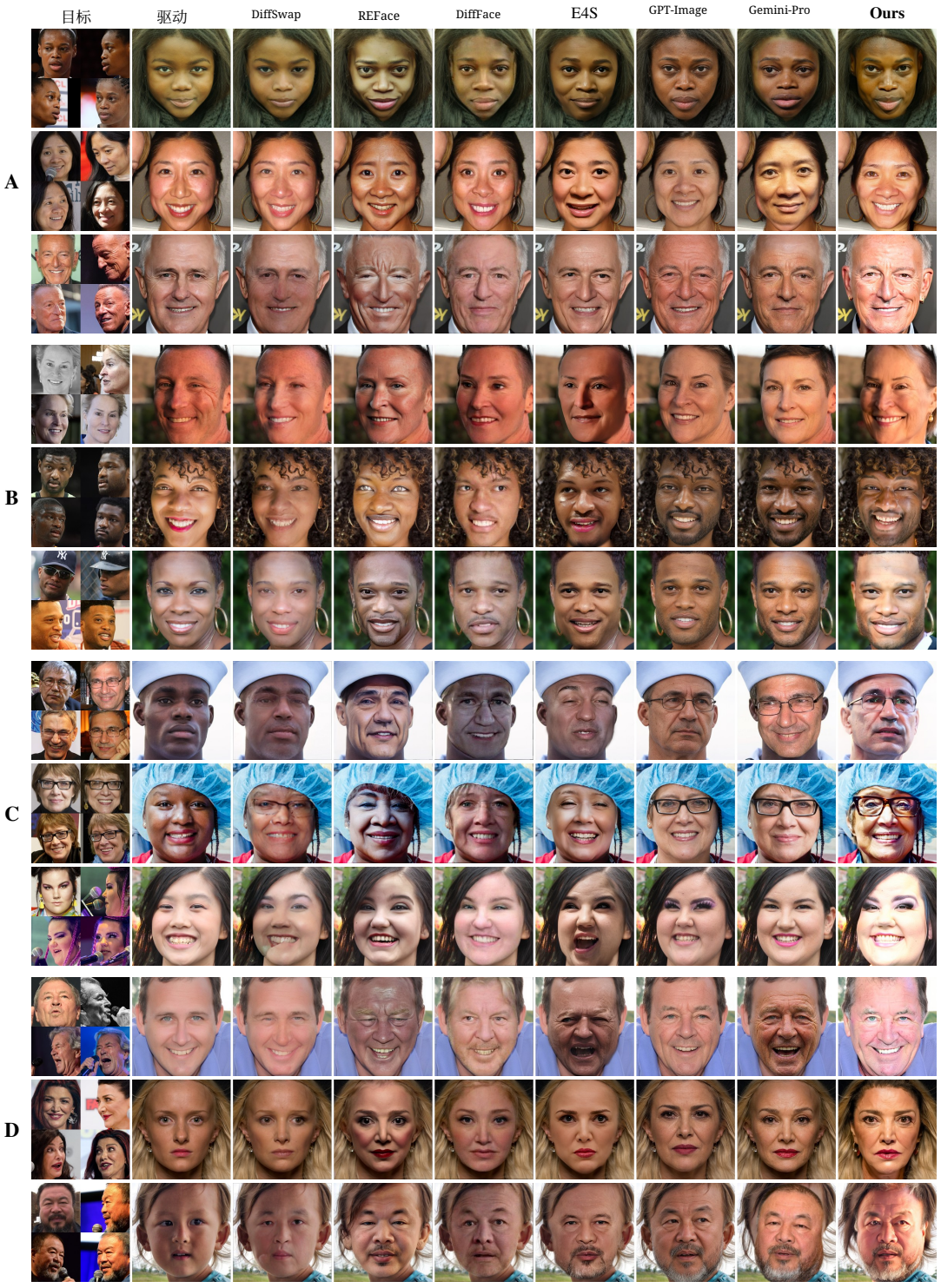


图12. **与表中八种方法的比较** 1，针对用户研究中的 12 对图像，每种表中条件各三对 17（A：匹配；B：不同性别；C：不同族裔和皮肤亮度；D：年龄差至少为 30 岁）。目标：提供给这些方法的四张照片之一。每种方法均以其最佳设置运行。我们从 416 对中挑选了这些图像对，以涵盖不同目标对象。



Fig图 13。 **与其他方法和两种图像模型进行比较**，使用与图 12相同的图像对。

表16。**身份基准测试中的所有指标**（240组图像对）。每种方法均以其最佳设置运行；标记为 $M{=}1$ 的行向平均基线方法提供目标人物的一张照片，而不是24张。 M：提供给方法的目标人物照片。 n：已评分的结果；其余方法未返回图像或人脸对齐失败（部分估计器的评分结果较少，详见正文）。身份：与目标人物24张留出照片的平均余弦相似度。表情：EmoNet 的效价和唤醒度、SMIRK 表情误差以及 MediaPipe 混合形状，均为与驱动图像之间的距离。姿态：结果与驱动图像的头部旋转之间的角度（度）。光照：低通亮度的相关性、平均亮度差异（0–255）以及人脸区域内平均颜色的差异（a^*b^*）。真实感：FaceScore 和 DSL-FIQA，不使用参考图像。我们的方法对应的行：分辨率和推理掩码；面部掩码覆盖面部、耳朵和颈部，而面部+头发掩码还覆盖头发，随后由模型生成这些区域；关闭光照：仅进行第一遍处理。斜体数值超过目标人物真实照片的相应数值，因此不参与排名。粗体：在各自行的最佳设置下表现最佳。

			身份↑			表情↓			姿态↓		光照			真实感↑	
方法	M	n	AdaFace	ArcFace	FaceNet	EmoNet	SMIRK	Blendsh.	HopeNet	6DRepNet	相关性↑	ΔL↓	Δab↓	FaceScore	DSL-FIQA
SimSwap	24	240	0.399	0.500	0.573	0.243	0.422	**0.037**	**2.03**	2.04	**0.96**	**2.4**	2.03	3.67	0.556
$M{=}1$	1	240	0.326	0.414	0.506	0.241	0.424	0.036	2.08	2.04	0.96	2.3	2.01	3.64	0.555
FaceDancer	24	240	0.314	0.394	0.618	**0.205**	0.420	**0.037**	2.38	**1.99**	**0.96**	2.5	1.21	4.39	0.711
$M{=}1$	1	240	0.260	0.335	0.544	0.208	0.421	0.037	2.31	1.97	0.96	2.3	1.15	4.39	0.711
BlendFace	24	240	0.425	0.487	0.537	0.231	0.466	0.041	4.42	3.97	0.90	6.3	1.39	3.73	0.640
$M{=}1$	1	240	0.353	0.416	0.479	0.222	0.463	0.041	4.39	3.86	0.90	6.2	1.39	3.70	0.640
GHOST	24	240	0.473	0.548	0.615	0.337	0.569	0.057	4.60	4.43	0.93	4.4	**0.82**	4.02	0.570
$M{=}1$	1	240	0.401	0.475	0.561	0.341	0.576	0.057	4.45	4.43	0.93	4.3	0.82	4.00	0.569
InSwapper	24	240	*0.690*	*0.715*	**0.706**	0.258	0.498	0.045	2.70	2.61	0.93	2.6	1.28	3.92	0.492
$M{=}1$	1	240	0.566	0.606	0.637	0.275	0.495	0.045	2.62	2.60	0.93	2.6	1.27	3.92	0.492
HifiFace	24	240	0.444	0.528	0.588	0.345	0.696	0.057	3.14	3.04	0.90	6.2	2.11	4.02	0.682
$M{=}1$	1	240	0.354	0.437	0.523	0.369	0.717	0.061	3.42	3.42	0.89	6.1	2.12	3.96	0.680
Face-Adapter	24	240	0.446	0.501	0.666	0.332	0.624	0.052	4.38	4.33	0.83	6.9	2.59	4.09	0.811
$M{=}1$	1	240	0.355	0.408	0.580	0.343	0.618	0.054	4.34	3.98	0.83	7.1	2.64	4.10	0.817
DreamID-V	1	240	0.516	0.550	0.652	0.372	0.689	0.059	7.52	8.18	0.52	15.2	5.88	4.42	0.732
DiffSwap	1	226	0.166	0.208	0.255	0.213	**0.416**	**0.037**	2.86	2.46	0.94	6.0	2.65	4.42	0.654
REFace	1	223	0.279	0.346	0.507	0.338	0.596	0.059	3.83	3.52	0.84	15.4	4.27	4.28	0.795
DiffFace	1	223	0.315	0.381	0.527	0.325	0.531	0.052	4.07	3.79	0.93	2.7	3.43	4.72	0.634
E4S	1	223	0.171	0.218	0.385	0.376	0.631	0.066	4.79	4.63	0.81	13.5	3.00	4.70	0.790
GPT-Image	8	224	0.369	0.414	0.567	0.295	0.557	0.049	3.94	3.75	0.86	14.9	6.15	**5.14**	**0.846**
Gemini-Pro	1	240	0.329	0.371	0.510	0.379	0.603	0.058	7.35	7.39	0.75	17.3	5.31	4.98	0.844
我们的方法，256 像素，面部	24	240	0.573	0.601	0.647	0.253	0.550	0.042	3.99	3.51	0.83	23.9	4.23	4.62	0.668
512 像素，面部	24	240	0.586	0.609	0.641	0.239	0.548	0.041	4.09	3.52	0.84	18.4	3.70	4.50	0.750
256 像素，面部+头发	24	240	**0.618**	**0.638**	**0.706**	0.257	0.558	0.043	4.30	3.41	0.75	29.2	5.40	4.75	0.654
256像素，面部，关闭光照	24	240	0.590	0.616	0.672	0.256	0.577	0.044	3.97	3.76	0.66	25.4	7.39	4.62	0.675
目标人物的真实照片	–	240	0.656	0.677	0.736										
另一个人的照片	–	240	0.012	0.011	0.012										
驱动图像	–	240											4.23	0.772	

阶段 3 的阶段会用另一个人的面部替换光照输入的面部，但所用模型不会改变耳朵和颈部。因此在训练过程中，这些区域会呈现真实值中的人物，从而为流提供一条可从输入中复制内容的路径。当目标人物与驱动图像中的人物在细节上有所不同时，驱动图像中的某些纹理便可能出现在结果中，例如年长者面部的皱纹（图 18）。图1 （左上）的光照示例使用了这一流。

图 19 说明了为什么第二次处理会使光照输入发生变形。单次处理时，光照流读取驱动图像中的面部，因此结果会带有其部分形状特征：第一行中从下方打光的圆脸，以及下面两行中男性宽阔的下颌。经过变形后，光照输入会跟随第一次处理的面部，因此在相同光照下，结果的面部会更接近目标。图中的轮廓展示了这一变化：经过变形后，三行中的下颌都更窄。

G.3. 遮挡物

在视觉特效中，决定保留驱动图像中哪些内容的掩码，是艺术家经常针对每个镜头进行调整的一项输入。我们的默认设置会通过通用检测链找出遮挡物（第 A.4节）；对于细小或不常见的物体，更精细的检测器、另一种掩码设计或人工参与都可以对其进行改进，而无需更改模型。图 20 展示了使用 FFHQ 驱动图像的结果 [33]，这些图像中的某些物体遮住了部分面部，例如手、麦克风、食物、手机、围巾或头饰；目标图像来自 VGGFace2-HQ，这些目标既不属于身份基准中的目标，也与我们的训练数据中的身份不相近（第 B节）。第 6 行是一组用户研究中的名人配对，其驱动图像中人物戴着透明防护眼镜。我们的模型使用默认设置运行；作为此类改进的示例，在第 7 行和第 8 行，检测器还会接收细小物体（发带、细绳）的名称。InSwapper、DreamID-V 和 Face-Adapter 使用表 1中的设置。

我们的结果保留了驱动图像中的遮挡物，使其位于

表17。 **身份基准测试中各条件下的身份表现**。 结果与目标人物的 24 张留出照片之间的平均余弦相似度。A：同性别、同族裔，年龄差不超过 15 岁，皮肤亮度差不超过 30 （0–255）。B：同A，但性别不同。C：同A，但族裔不同，且皮肤亮度差至少为 30。D：同性别、同族裔，年龄差至少为 30 岁。性别、年龄、族裔和皮肤亮度由模型估算。每种条件下有 60 对图像。 M， n 以及我们的方法中的各行与表 16相同。斜体值高于目标人物真实照片的值，不参与排名。粗体：最佳。

			AdaFace↑					ArcFace↑					FaceNet↑				
方法	M	n	A	B	C	D	All	A	B	C	D	All	A	B	C	D	All
SimSwap	24	240	0.441	0.373	0.384	0.399	0.399	0.535	0.480	0.491	0.493	0.500	0.621	0.548	0.571	0.551	0.573
FaceDancer	24	240	0.335	0.297	0.312	0.311	0.314	0.401	0.387	0.395	0.391	0.394	0.630	0.614	0.625	0.605	0.618
BlendFace	24	240	0.461	0.431	0.408	0.400	0.425	0.528	0.478	0.479	0.463	0.487	0.601	0.521	0.529	0.497	0.537
GHOST	24	240	0.493	0.468	0.473	0.457	0.473	0.571	0.542	0.550	0.530	0.548	0.654	0.614	0.611	0.580	0.615
InSwapper	24	240	*0.691*	*0.684*	*0.696*	*0.688*	*0.690*	*0.718*	*0.709*	*0.722*	*0.712*	*0.715*	*0.727*	0.693	0.712	0.694	**0.706**
HifiFace	24	240	0.456	0.445	0.444	0.429	0.444	0.539	0.525	0.536	0.513	0.528	0.623	0.578	0.600	0.550	0.588
Face-Adapter	24	240	0.440	0.434	0.446	0.464	0.446	0.491	0.491	0.506	0.517	0.501	0.666	0.653	0.678	0.666	0.666
DreamID-V	1	240	0.511	0.508	0.516	0.529	0.516	0.550	0.543	0.557	0.552	0.550	0.647	0.641	0.672	0.649	0.652
DiffSwap	1	226	0.202	0.154	0.167	0.144	0.166	0.237	0.205	0.202	0.191	0.208	0.329	0.240	0.249	0.205	0.255
REFace	1	223	0.319	0.256	0.286	0.257	0.279	0.383	0.324	0.362	0.316	0.346	0.562	0.451	0.550	0.465	0.507
DiffFace	1	223	0.341	0.315	0.304	0.299	0.315	0.408	0.379	0.372	0.365	0.381	0.546	0.527	0.548	0.486	0.527
E4S	1	223	0.193	0.159	0.170	0.164	0.171	0.234	0.205	0.222	0.211	0.218	0.418	0.367	0.376	0.379	0.385
GPT-Image	8	224	0.347	0.359	0.380	0.390	0.369	0.382	0.412	0.423	0.435	0.414	0.554	0.568	0.574	0.571	0.567
Gemini-Pro	1	240	0.345	0.320	0.334	0.318	0.329	0.376	0.365	0.382	0.363	0.371	0.527	0.516	0.509	0.490	0.510
我们的方法，256 像素，面部	24	240	0.581	0.556	0.578	0.577	0.573	0.609	0.587	0.609	0.601	0.601	0.653	0.631	0.658	0.646	0.647
512 像素，面部	24	240	0.592	0.577	0.587	0.589	0.586	0.616	0.601	0.612	0.607	0.609	0.650	0.628	0.656	0.632	0.641
256 像素，面部+头发	24	240	**0.607**	**0.618**	**0.621**	**0.623**	**0.618**	**0.632**	**0.640**	**0.645**	**0.636**	**0.638**	**0.690**	**0.707**	**0.721**	**0.706**	**0.706**
256 像素，面部，关闭光照	24	240	0.592	0.575	0.595	0.598	0.590	0.622	0.602	0.623	0.618	0.616	0.672	0.650	0.690	0.675	0.672
目标人物的真实照片	–	240	0.642	0.650	0.663	0.668	0.656	0.671	0.669	0.687	0.682	0.677	0.724	0.731	0.750	0.739	0.736

表18。 **用户研究中名人目标的指标** （176 图像对）。大多数名人除了提供给方法的照片之外，几乎没有其他照片，因此 ArcFace 会计算两次：一次是针对给定照片（池），这有利于复制这些照片的方法；另一次是针对留出照片（留出集），仅针对至少有四张照片的 40 位名人。其他列与表 16相同； n 统计由 ArcFace（池）评分的结果。粗体：最佳结果。

			ArcFace↑		表情↓		姿态↓	光照↑	真实感↑	
方法	M	n	Pool	留出集	EmoNet	SMIRK	HopeNet	相关性	FaceScore	DSL-FIQA
SimSwap	24	176	0.513	0.501	0.277	0.421	**1.70**	**0.96**	4.34	0.533
FaceDancer	24	176	0.404	0.403	0.240	0.429	1.81	0.95	5.12	0.696
BlendFace	24	176	0.550	0.544	**0.226**	**0.362**	1.86	0.92	4.56	0.624
GHOST	24	176	0.562	0.572	0.378	0.551	2.27	0.92	4.62	0.545
InSwapper	24	176	**0.733**	**0.724**	0.274	0.491	2.27	0.91	4.57	0.469
HifiFace	24	176	0.521	0.521	0.371	0.700	2.61	0.89	4.33	0.671
Face-Adapter	24	176	0.495	0.466	0.402	0.635	3.72	0.81	4.66	0.816
DreamID-V	1	176	0.599	0.588	0.409	0.667	5.50	0.47	4.69	0.655
DiffSwap	1	173	0.217	0.227	0.259	0.418	2.09	0.93	4.86	0.634
REFace	1	173	0.334	0.328	0.405	0.625	3.22	0.82	4.77	0.783
DiffFace	1	173	0.360	0.341	0.432	0.589	3.31	0.92	5.09	0.609
E4S	1	173	0.269	0.257	0.437	0.665	4.19	0.71	4.95	0.785
GPT-Image	8	163	0.461	0.447	0.347	0.594	4.23	0.82	**5.48**	**0.831**
Gemini-Pro	1	172	0.479	0.461	0.408	0.643	5.43	0.70	5.30	0.827
Ours	24	176	0.619	0.603	0.258	0.568	3.18	0.83	5.11	0.703
驱动图像	–	176							5.15	0.775

目标脸部正面的遮挡物。InSwapper 也保留了这些遮挡物，因为它只改变内部面部。DreamID-V 移除了手，并替换了麦克风和食物；Face- Adapter 则丢失了大部分手部，并改变了食物。

表19。**由 MEAD帧驱动的结果**。四个目标对象，四名留出集 MEAD受试者。匹配：EmoNet 类别与驱动图像类别相同的结果所占比例。VA：EmoNet 效价-唤醒度距离。SMIRK：SMIRK 表情误差。ArcFace：目标对象的 24 张留出照片的平均余弦相似度。在所有估计器均检测到面部的结果上计算平均值（对于情绪，每种方法为 105 至 116 个结果；对于视角，为 65 至 80 个结果）。

	情绪（前置摄像头）				相机视角			
方法	匹配↑	VA↓	SMIRK↓	ArcFace↑	匹配↑	VA↓	SMIRK↓	ArcFace↑
InSwapper	0.54	0.34	**0.59**	*0.718*	0.87	0.17	0.66	*0.708*
Face-Adapter	0.35	0.35	0.73	0.464	0.89	0.21	0.72	0.443
DreamID-V	0.41	0.39	0.90	0.510	0.86	0.30	0.78	0.509
Ours	**0.68**	**0.20**	0.63	**0.567**	**1.00**	**0.12**	**0.62**	**0.527**

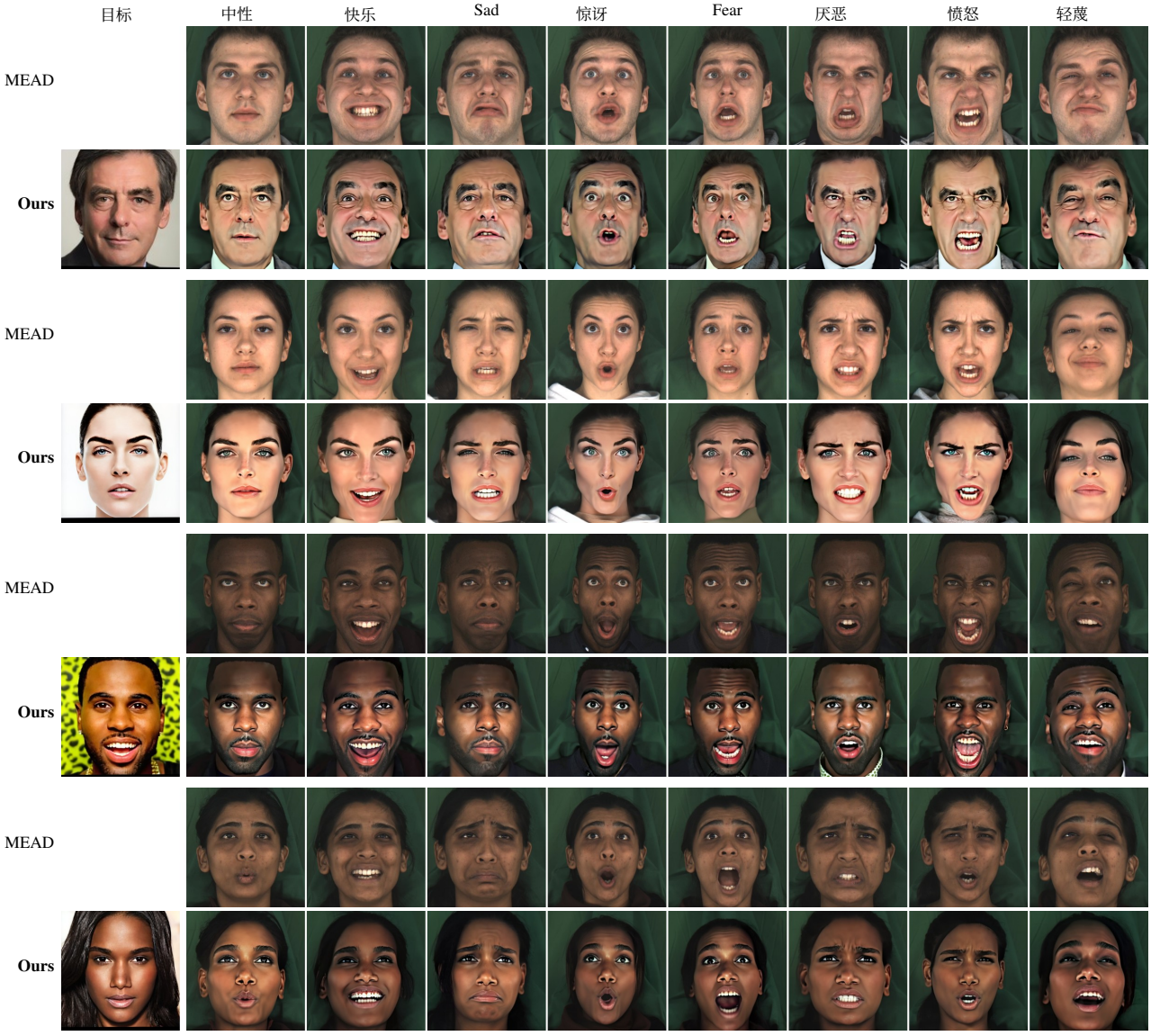


图14. **我们的结果遵循 MEAD 的情绪**。 四位 MEAD 受试者（从上到下：一位浅色皮肤男性、一位浅色皮肤女性、一位深色皮肤男性和一位深色皮肤女性），每位都与一位同性别且肤色相近的身份基准目标配对。对于每位受试者，上排展示前置摄像头拍摄的帧，下排展示我们的结果；第一列是目标的照片。

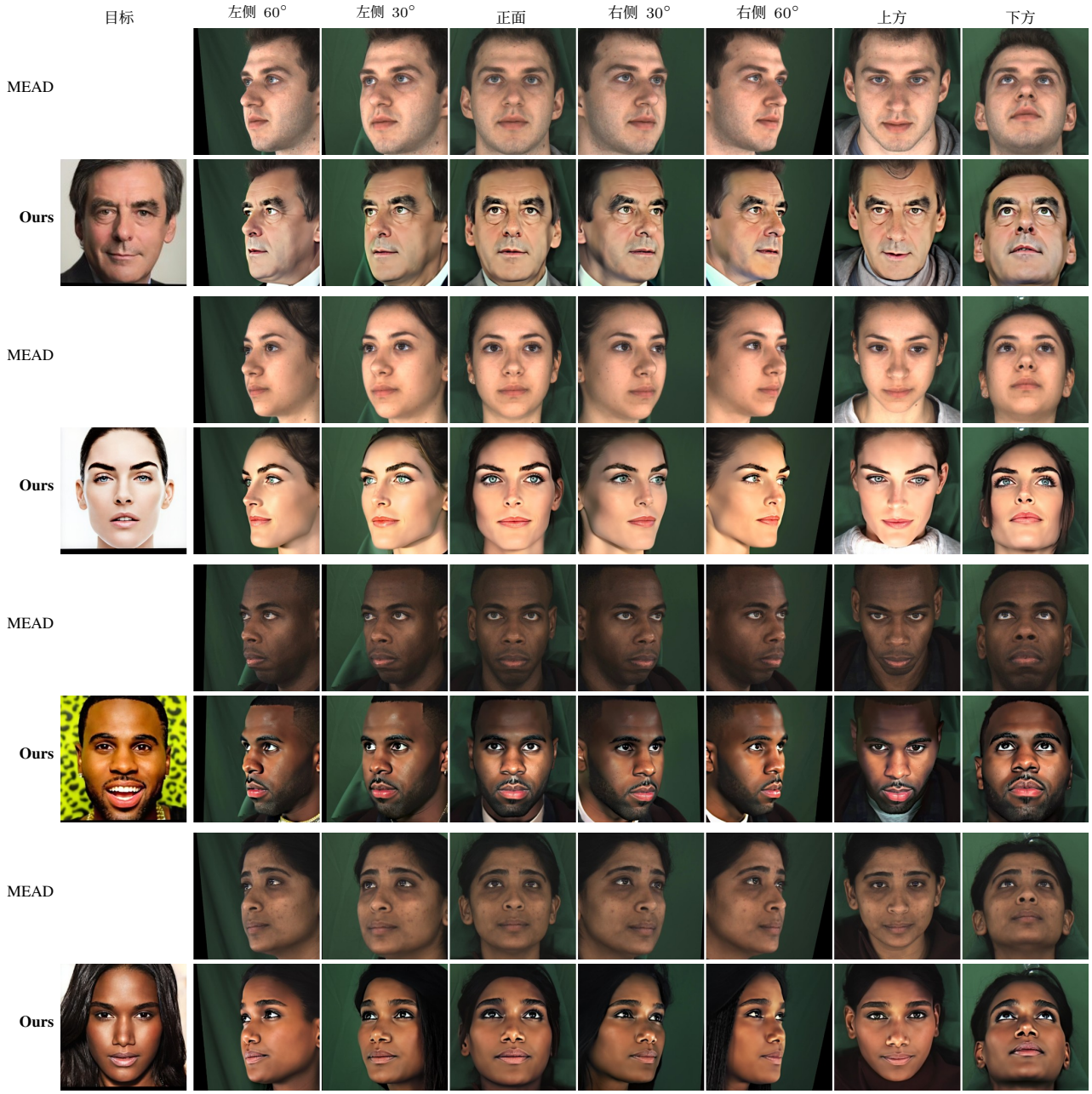


图15。 **我们在 t 下的结果　　MEAD 的七台摄像机。** 图 14 中的主体和目标对象，其中中性表情由七台摄像机在同一　　n 时刻拍摄。

表20。 **身份基准测试上的两种光照流** （240 对图像）以及形状基准测试上的两种光照流（348 对图像）。测量指标与表 1 和2相同；随机种子和设置相同，仅光照流不同。

流的输入	ArcFace↑	AdaFace↑	表情↓	姿态↓	光照↑	FaceScore↑	s_{gt} ↓	泄漏↓
面部（默认）	**0.601**	**0.573**	0.253	3.99	0.83	**4.62**	**0.083**	0.34
Face, 耳朵和颈部	0.589	0.562	0.245	3.85	**0.84**	4.49	0.086	0.34

H. 脸型

图 21 展示了表1中各方法在形状基准测试中的六组图像对，这些图像对按照驱动图像的下颌线与

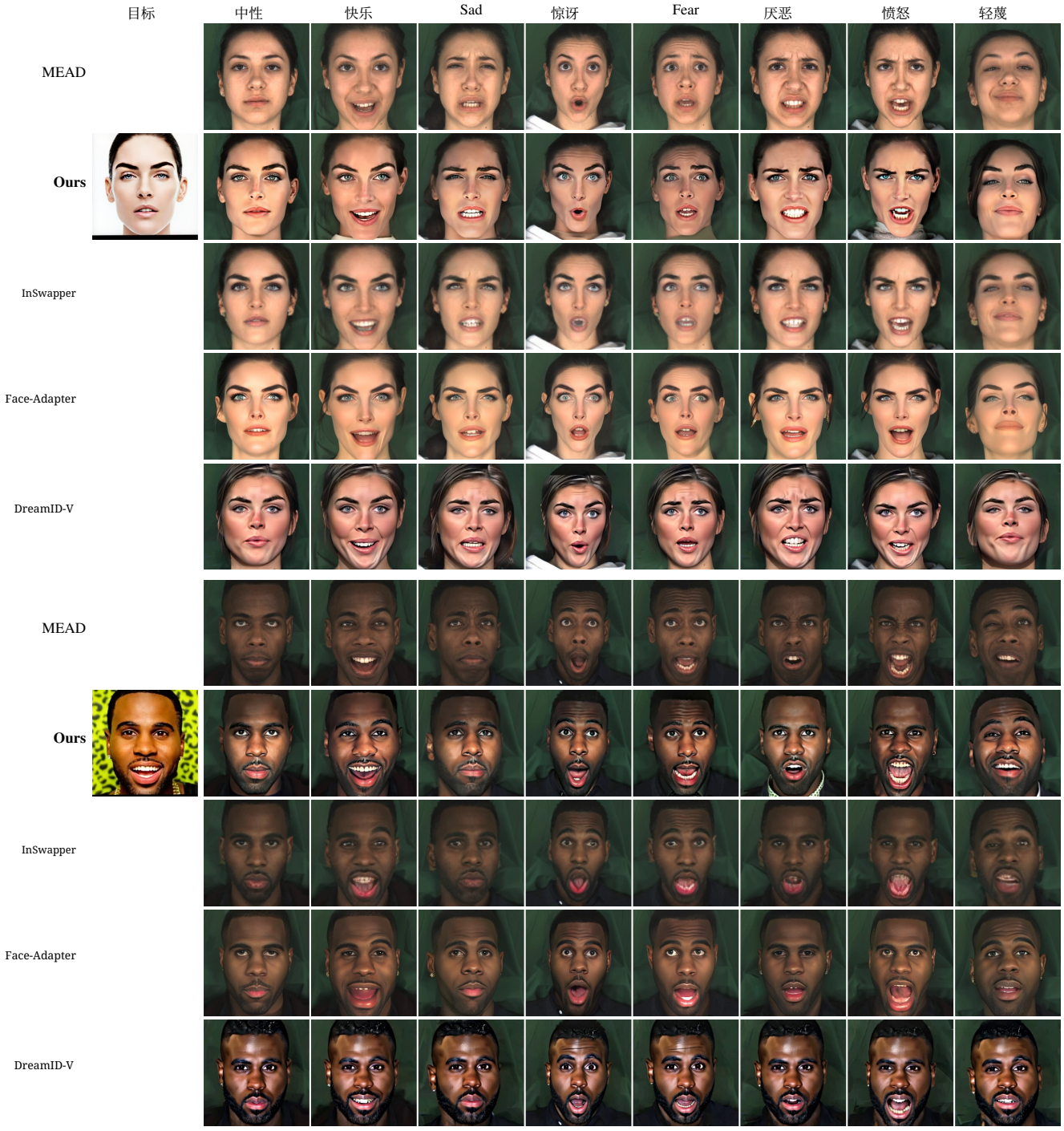


图 16。 **MEAD 的情绪，与先前工作进行比较。** 图 14 中的两个主体（上：一位浅色皮肤的女性；下：一位深色皮肤的男性）。每种方法均采用表 1中的设置。

真实值的下颌线之间的距离排序。基线方法的结果都接近驱动图像的下颌线，因此随着该距离增大，它们会偏离真实值：八种基线方法的平均 s_{gt} 从顶行的 0.08 上升到最底行的 0.37，而

我们的结果保持在 0。01 和 0。07之间。在整个基准测试中也是如此。表 22 根据真实值与驱动图像之间的下颌线距离，将 348 个图像对分为三等份。从最近的三分之一到最远的三分之一，表 s_{gt} 中八种基线方法的均值1上升

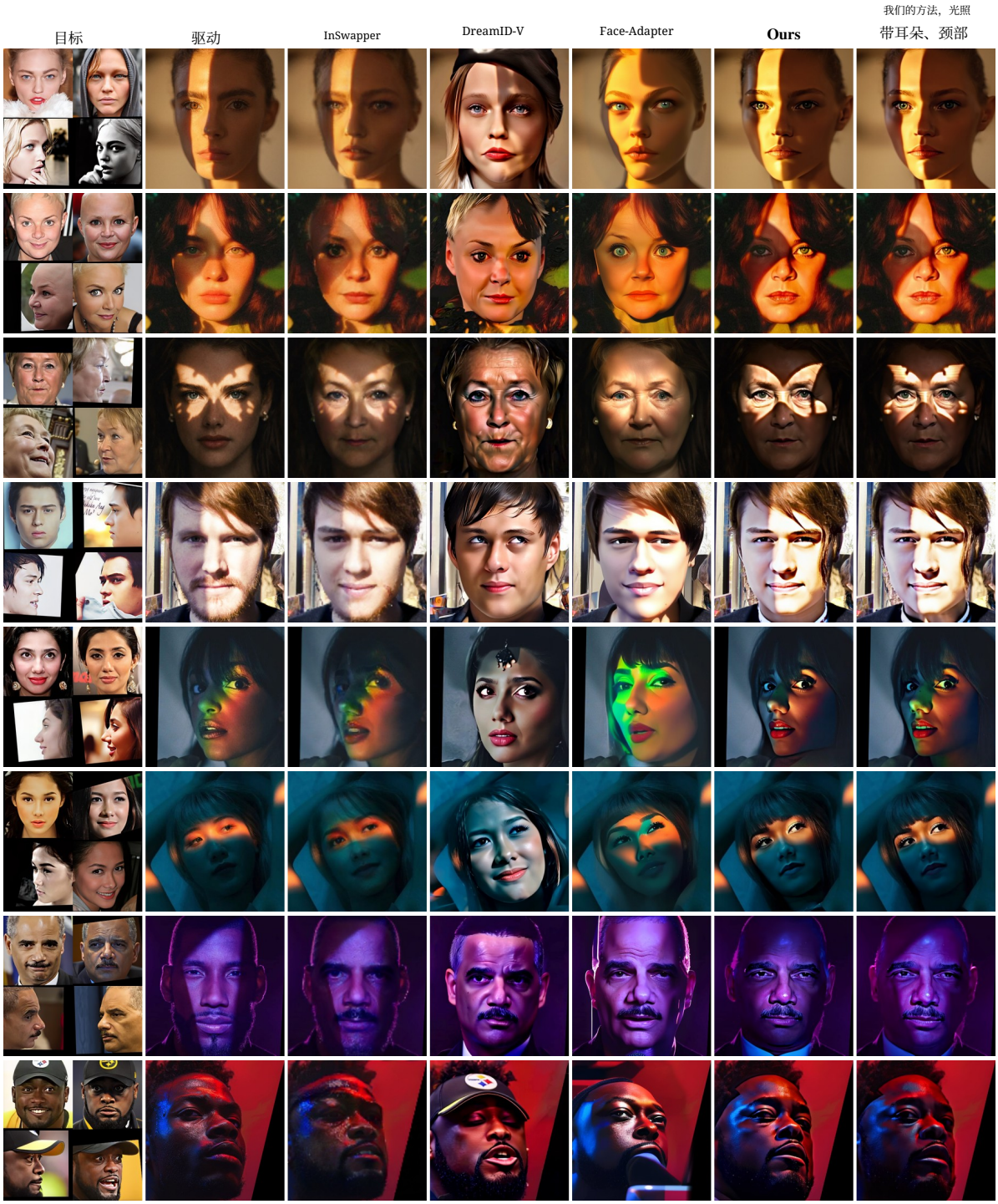


图17. **强光照下的结果**。 第1–4行：阳光条纹、透过树叶的阳光、投影图案和强侧光；第5–8行：彩色光。我们的方法：默认模型，使用变形进行两次处理；最后一列：输入也涵盖耳朵和颈部的光照流。目标：其24张照片中的四张。每种方法均采用表1中的设置。

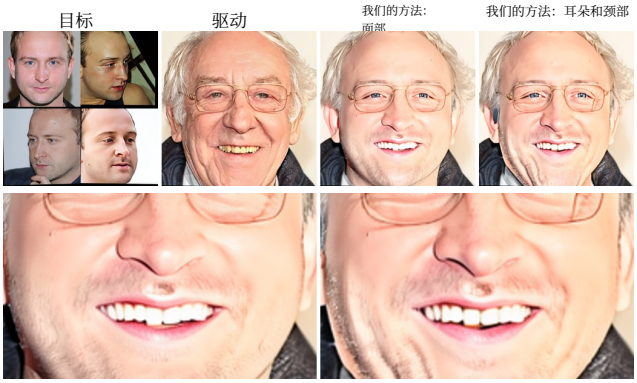


图18。**驱动图像中的纹理**。年轻的目标人物搭配年长者的驱动图像，并使用两种光照流。底部：两种结果的脸颊。采用同时读取耳朵和颈部的流时，年长者面部的皱纹会出现在脸颊和下颌上。



图19. **第二次处理的变形**。我们的单次处理结果（不使用变形）和两次处理结果（使用变形，为默认设置）。每行的两张图像都带有相同的两组面部轮廓：单次处理（红色）和两次处理（绿色）。上：从下方照明的驱动图像。中间和下方：女性目标人物与男性驱动图像。不使用变形时，结果会呈现出驱动图像的部分脸型。

从 0.076 到 0.116 和 0.197，而我们的方法从 0.065 到 0.085 和 0.099。当两条下颌线接近时，驱动图像的下颌线几乎是正确的：驱动图像本身的得分为 0.052，十二种基线方法中的六种以及 Gemini-Pro 的得分略低于我们的方法（0.053 到 0.062）。

在两个商业图像模型中，Gemini-Pro 的轮廓距离是基线方法中最低的（0.103），且与 DreamID-V 并列具有最低的泄漏值（0.47 和 0.52）：它复制了驱动图像中约一半的下颌线。该基准测试中的驱动图像均面向相机，表情中性

表21。**形状基准测试上的脸型，所有方法**（348 图像对）。s_{gt}：结果下颌线与真实值下颌线之间的距离。泄漏：当结果忽略驱动图像的下颌线时为 0，当结果复制该下颌线时为 1。三维：结果与真实值的重建三维人脸形状在头发以外区域的距离。在另一种光照下，真实值的得分为 0。024 (s_{gt}) 和1。14 (三维)。每种方法接收目标的 24 张渲染图，并在表 16 的设置下运行；GPT-Image 使用前8 张，Gemini-Pro 使用第一张，在每种方法的 347 组图像对上进行测试，因为其安全过滤器拦截了一组。我们的方法对应的行：分辨率和推理掩码；面部掩码覆盖面部、耳朵和颈部，而面部+头发掩码还覆盖头发，随后模型会生成头发。粗体：最佳结果。

方法	s_{gt} ↓	泄漏↓	3D↓
SimSwap	0.125	0.95	1.83
FaceDancer	0.122	0.91	1.85
BlendFace	0.119	0.86	1.99
GHOST	0.132	0.90	2.21
InSwapper	0.116	0.81	1.48
HifiFace	0.138	0.78	2.03
Face-Adapter	0.138	0.60	1.86
DreamID-V	0.149	0.52	1.67
DiffSwap	0.125	0.92	2.53
REFace	0.135	0.94	2.18
DiffFace	0.120	0.84	1.92
E4S	0.131	0.77	1.93
GPT-Image	0.127	0.75	1.62
Gemini-Pro	0.103	0.47	1.49
我们的方法，256 像素，面部	0.083	0.34	1.35
512 像素，面部	**0.082**	**0.30**	**1.34**
256 像素，面部+头发	**0.082**	0.32	1.35

；Gemini-Pro 保持了这些图像中的姿态：其 6DRep-Net 误差 2.4° 达到了完美结果的水平（真实值的得分为 2.6°，见 D节.1）。以照片作为驱动图像时，在身份基准上，它的 EmoNet 距离最大，姿态误差仅次于 DreamID-V（表 16）。GPT-Image 保留了驱动图像中的大部分下颌线（泄漏值为 0.75），其轮廓距离为 0.127，与驱动图像本身的轮廓距离（0.128）相近。与 DreamID-V 一样，它会生成整幅图像，且头部移动幅度超过大多数方法。对于 GPT-Image，双眼中点相对于驱动图像中的位置，在 s_{gt} 单位下移动了 0.07；对于 DreamID-V，则移动了 0.14；其他基线方法的移动幅度为 0.02 至 0.08，我们的方法为 0.04。真实值中的双眼中点偏移了 0.03，因为其中的眼睛属于目标。为了将下颌线形状与头部位置区分开来，我们还测量了经过平移和缩放后的 s_{gt}，使每个结果中的眼睛与真实值中的眼睛对齐。此时，我们的方法得分为 0.065，最接近的基线方法分别为 0.096（Gemini-Pro）、0.100（InSwapper）和 0.101（DreamID-V）。

掩码大小。推理时，面部掩码会膨胀 12px（第A.4节），为驱动图像中的面部周围留出空间，以容纳更宽或更长的脸型。表24 展示了在形状基准测试中对该膨胀量的调整。较小的掩码（6px）留出的空间较少，因此驱动图像中的更多脸型

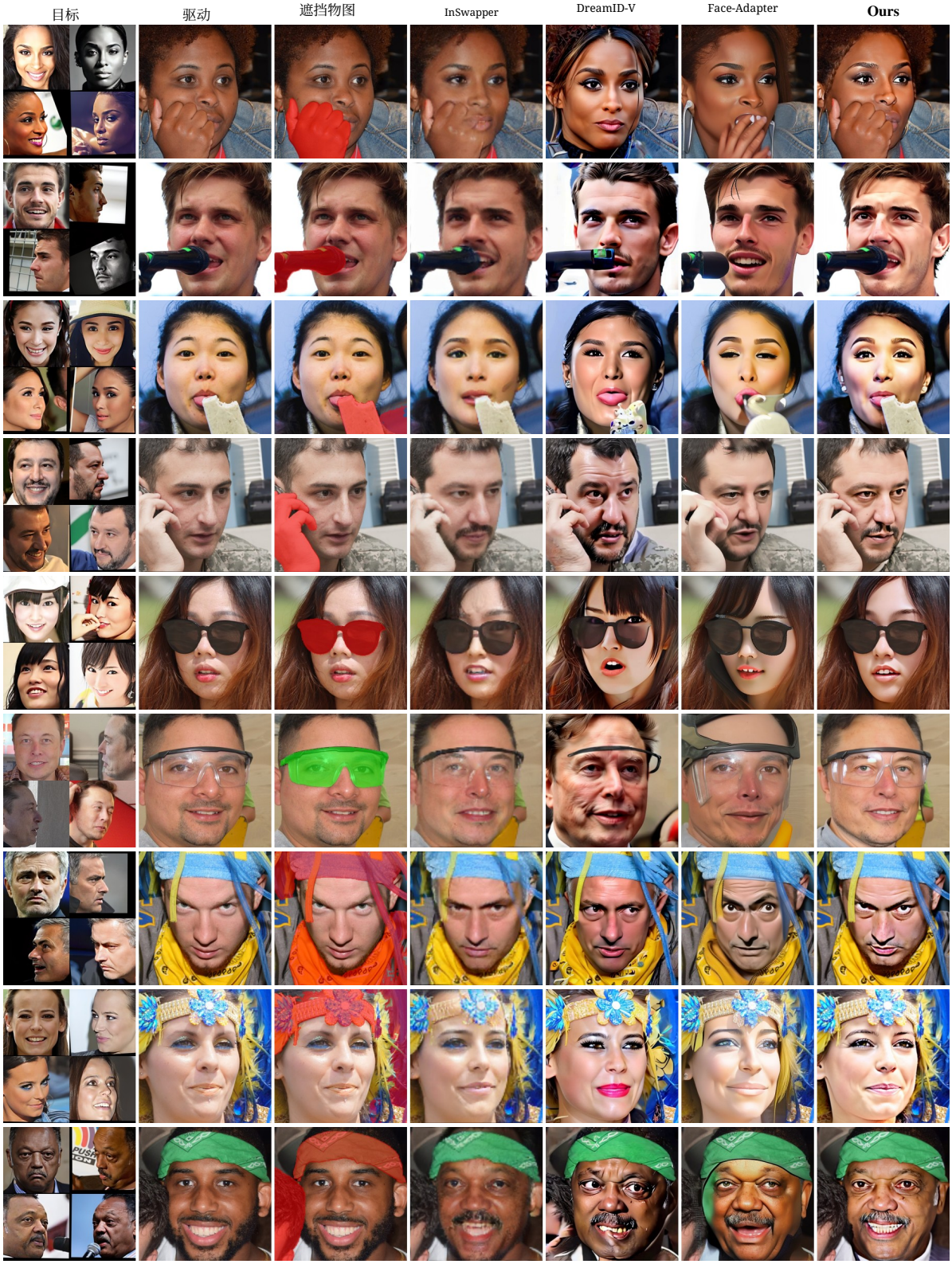


图 20. **有遮挡物时的结果**。 各行依次为：手、麦克风、食物、手机和手、太阳镜、透明防护眼镜、围巾和发带、头饰，以及头巾。遮挡物图：在驱动图像中检测到的遮挡物（红色）和透明眼镜（绿色），如第 G.3 节所示。第 6 行的透明眼镜不是遮挡物；它们遵循眼镜规则（第 A.4 节），其后的眼睛会重新生成。目标：其中四张照片。每种方法均采用表 1 中的设置运行。

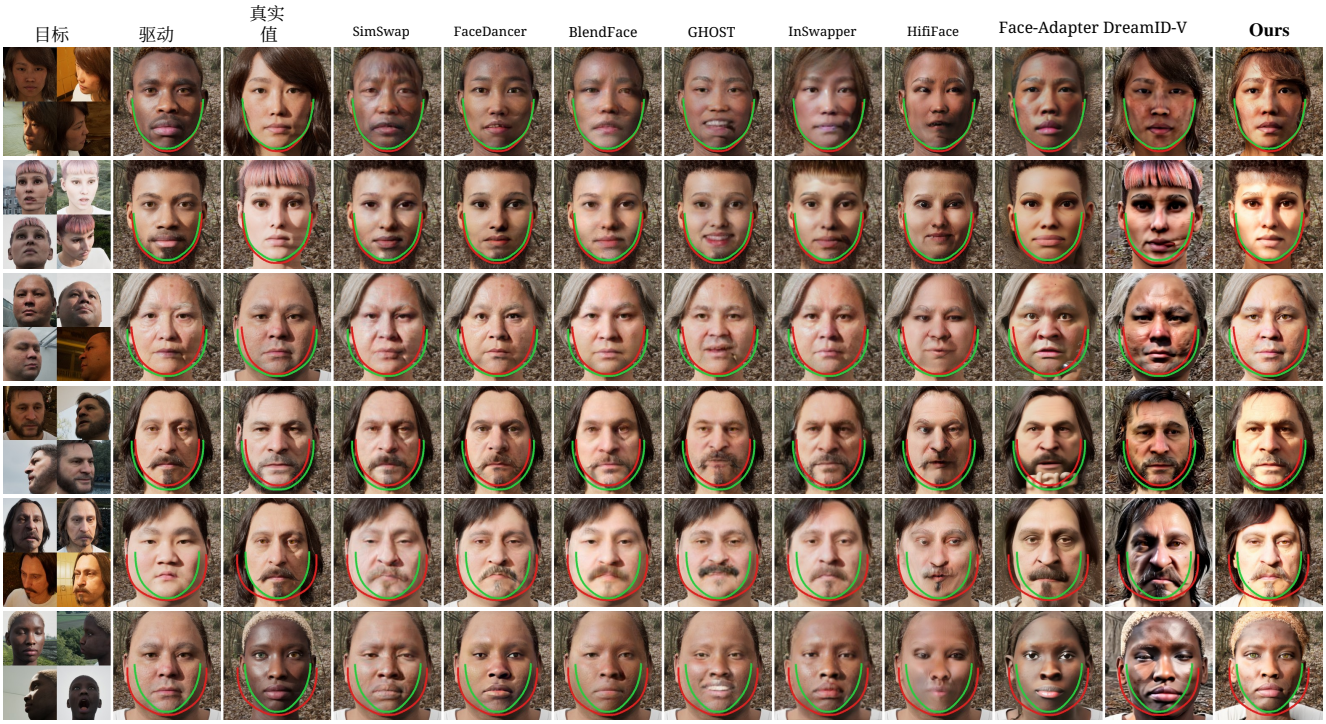


图21. **形状基准测试中的脸型。** 表 1中的方法的六组图像对，按真实值与驱动图像的下颌线之间的距离从小到大（从上到下）排序。目标：其 24 张渲染图中的四张。线条含义同图 7：真实值的下颌线（绿色）和驱动图像的下颌线（红色）。

图像仍保留在结果中（泄漏值为0.41，而非0.34）。更大的掩码并不会使面部填满掩码区域：从12像素增至24像素时，掩码边缘向外移动了12像素，但结果中的下颌仅移动了约2像素。在这一空间下，结果与驱动图像的下颌线（泄漏值为0.28）偏离得更远，与真实值的ArcFace相似度上升；然而，下颌末端略微超出目标中的下颌，这可能是因为更大的掩码还会重新生成面部周围更多的头发和颈部，从而固定驱动图像中头部的尺度。因此，轮廓距离和三维误差会略有上升。使用覆盖整幅图像的掩码时，头发和颈部完全不会被保留，面部会进一步变大（s_{gt} 0.113），而泄漏值保持在0.28。在每种膨胀设置下，我们的结果在三项形状指标上均优于表21中的所有基线方法；使用覆盖整幅图像的掩码时，Gemini-Pro的轮廓距离更低（0.103，而非0.113）。默认的12像素设置在所有掩码中具有最低的轮廓距离和三维误差，而18像素与其相近。

I. 更多目标照片

本节补充完成第 4.4 节。实验方案相同：每种方法都接收固定顺序中最前面的 M 张照片（顺序见第 B.1 节），基线方法则对这些照片的身份输入取平均。

身份。图 22 展示了 ArcFace 得分本身、目标人物真实照片的得分水平，以及在 AdaFace 和 FaceNet 下相对于 $M{=}1$的提升。在全部三种识别器下，我们的方法从 $M{=}1$ 张增加到 24 张时提升最大：AdaFace 下提升 0.146，而基线方法为 0.054 至 0.124；FaceNet 下提升 0.095，而基线方法为 0.054 至 0.086。ArcFace 下，我们的得分从 0.467 升至 0.601。InSwapper 的起始得分较高，为 0.606，并且从 $M{=}5$ 张起就超过了目标人物真实照片的得分水平。InSwapper 的身份编码器并非我们的 ArcFace 模型，而是在其他数据上训练的同类网络（ResNet-50，采用 ArcFace 损失）；两者对相似度的判断也相近：在 141 张照片上，它们的相似度矩阵相关系数为 0.98。因此，针对 InSwapper 编码器拟合的结果，在我们的 ArcFace 下也可能得分很高。

脸型。在形状基准测试中，我们的轮廓距离在 $M{=}1$、2、4、8 和 24 时分别为 0.090、0.085、0.086、0.089 和 0.083。 $M{=}1$ 和 24 之间的差异很明显，为 0.007，图像对上的 95%区间为 $[0.003, 0.012]$，尽管距离并非每一步都会下降。InSwapper 的轮廓距离在 $M{=}1$ 时为 0.115，在 $M{=}24$时为 0.116。若不采用解耦策略（表 3 第一行），我们的模型保持在 0.125 和 0.124，接近驱动图像的 0.128。

将照片放在一起读取。 身份编码器一次读取全部 M 照片：其查询会关注每张照片的图像块令牌

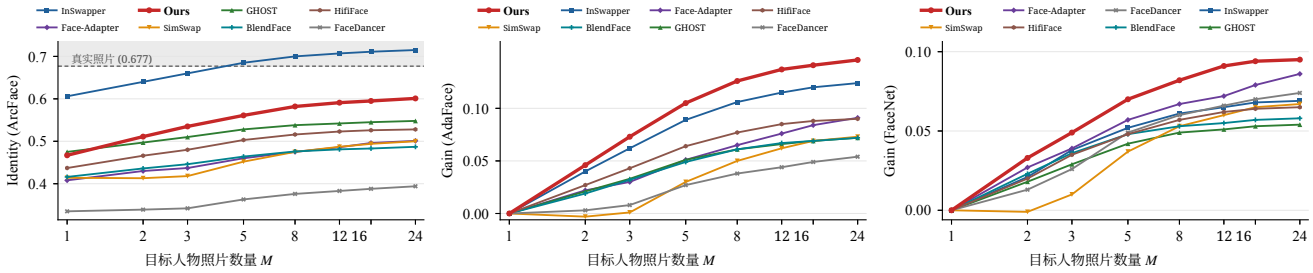


图 22。**身份与目标人物照片数量的关系**。左：ArcFace 得分，以及目标人物真实照片的水平（虚线）。中间和右侧：AdaFace 和 FaceNet 下相对于 $M=1$ 的增益。身份基准中的 240 对图像；基线方法对其身份输入中的 M 照片取平均值。

表22. 根据真实值与驱动图像的下颌线距离划分的脸型（形状基准测试）。按此距离（全部 21 个下颌点的平均距离，单位为 s_{gt}）将 348 对图像分为三组：0.03–0.10、0.10–0.18、0.18–0.44，每组 116 对；GPT-Image 和 Gemini-Pro 各缺少一组（表21）。每组三组的平均 s_{gt}。驱动图像行对驱动图像本身进行评分。我们的方法各行与表21相同。粗体：最佳结果。

	第三			
方法	最近	中间	最远	All
驱动图像	0.052	0.105	0.226	0.128
SimSwap	**0.053**	0.100	0.222	0.125
FaceDancer	0.056	0.101	0.208	0.122
BlendFace	0.054	0.098	0.205	0.119
GHOST	0.069	0.108	0.218	0.132
InSwapper	0.055	0.103	0.191	0.116
HifiFace	0.088	0.122	0.203	0.138
Face-Adapter	0.107	0.156	0.152	0.138
DreamID-V	0.128	0.141	0.178	0.149
DiffSwap	0.057	0.100	0.218	0.125
REFace	0.070	0.107	0.228	0.135
DiffFace	0.054	0.096	0.209	0.120
E4S	0.095	0.111	0.188	0.131
GPT-Image	0.072	0.100	0.208	0.127
Gemini-Pro	0.062	0.094	0.153	0.103
我们的方法，256 像素，面部	0.065	0.085	0.099	0.083
512 像素，面部	0.067	**0.083**	0.096	**0.082**
256 像素，面部+头发	0.067	0.084	**0.095**	**0.082**

（第 A.1 节）。使用多张照片的一种更简单方法，是分别编码每张照片，然后按位置对身份令牌取平均。以这种方式对 24 张照片取平均后，ArcFace 在身份基准上的得分从 0.601 降至 0.477，接近单张照片的 0.467，并且 99% 的结果得分更低。平均后的令牌也比编码器生成的令牌更短（平均范数为 16，而不是 22）。将每个令牌重新缩放到其所在位置的单张照片令牌的平均范数，可挽回约一半的下降幅度（ArcFace 为 0.534）；但仍有 97% 的结果得分低于编码器同时读取这些照片时的结果。使用平均令牌时，脸型也略差一些（s_{gt} 约为 0.09，而不是 0.083）。在图 23 中，平均后的令牌，无论是否重新缩放，

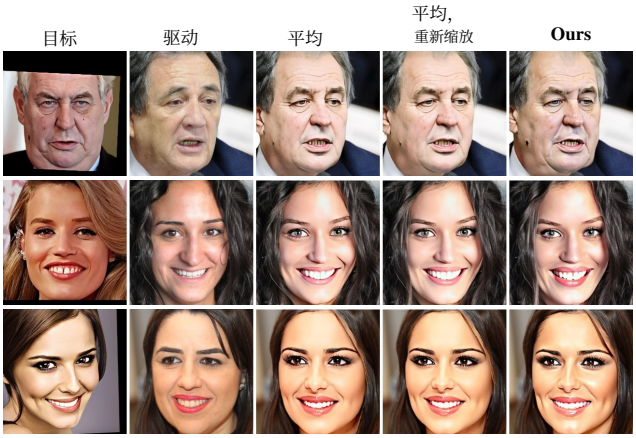


图23。平均身份令牌。我们的方法：身份编码器同时读取目标对象的 24 张照片。平均：每张照片单独进行编码，然后按位置对词元求平均。平均，重新缩放：使用相同的词元，并将其重新缩放至单张照片词元的范数。行：图 25 中的三个目标对象，分别具有脸颊上的痣、门牙间隙和酒窝。目标：一张能展现该细节的照片。三组图像对的 ArcFace，从上到下依次为：0.645、0.668、0.681；0.440、0.510、0.616；0.593、0.629、0.689。

没有，只会将脸颊上的痣画得很淡，并且遗漏门牙间隙。

示例。图24展示了身份基准测试中的四个目标对象，它们都与我们的训练数据（第B.1节）中的任何身份都相去甚远，分别位于 $M=1$、3 和24。每一行中，$M=1$ 处给出的照片展示的某些特征属于照片本身，而非人物：眼镜、刘海、帽子下的长发，或黑白胶片效果。仅凭这张照片，我们的结果会将这些特征视为身份的一部分；而在$M=3$处添加的两张照片中看不到这些特征，因此它们也不会出现在结果中。添加的照片也能展现第一张照片中被遮挡的内容。第三行中，只有第三张照片展示了驱动图像中可见的面部侧面，而该侧脸颊上的妆容也出现在 $M=3$处的结果中。InSwapper 的结果随 M变化较小。

表 23。 **驱动图像与目标之间的差异所对应的脸型**（形状基准测试）。 s_{gt} 以及三维，如表 21 所示。性别和年龄由面部分析模型根据每个角色的渲染图进行估计，皮肤亮度取人脸区域中央部分的平均光度（0186、162、177、171、222、1262553015 21–）。皮肤亮度不同：差异至少为；年龄不同：差异至少为年。每组图像对从左到右依次为；GPT-Image 和 Gemini-Pro 各缺少一组图像对（表）。驱动图像行表示驱动图像本身的得分。我们的方法的各行如表 21所示。粗体：最佳。

	s_{gt} ↓						3D↓					
	皮肤亮度		Sex		Age		皮肤亮度		Sex		Age	
方法	sim.	差异	same	差异	sim.	差异。	sim.	差异。	same	差异。	sim.	差异。
驱动图像	0.135	0.120	0.129	0.127	0.122	0.139	3.35	3.66	3.26	3.74	3.28	3.88
SimSwap	0.132	0.117	0.129	0.121	0.116	0.141	1.73	1.94	1.78	1.88	1.72	2.03
FaceDancer	0.128	0.114	0.125	0.118	0.115	0.134	1.77	1.93	1.87	1.82	1.75	2.01
BlendFace	0.126	0.110	0.122	0.116	0.110	0.134	1.90	2.09	1.89	2.09	1.87	2.20
GHOST	0.133	0.130	0.133	0.130	0.126	0.141	2.19	2.23	2.16	2.26	2.17	2.29
InSwapper	0.120	0.111	0.119	0.113	0.105	0.136	1.43	1.54	1.44	1.52	1.42	1.58
HifiFace	0.139	0.137	0.139	0.136	0.130	0.151	2.01	2.05	2.00	2.06	1.99	2.08
Face-Adapter	0.138	0.139	0.148	0.129	0.131	0.151	1.83	1.89	1.93	1.78	1.81	1.94
DreamID-V	0.148	0.151	0.147	0.152	0.138	0.170	1.64	1.71	1.71	1.63	1.65	1.71
DiffSwap	0.132	0.117	0.126	0.123	0.118	0.137	2.43	2.63	2.51	2.54	2.45	2.66
REFace	0.141	0.128	0.135	0.135	0.129	0.146	2.09	2.29	2.10	2.27	2.09	2.33
DiffFace	0.127	0.111	0.123	0.116	0.111	0.134	1.86	2.00	1.91	1.93	1.90	1.97
E4S	0.138	0.123	0.138	0.124	0.124	0.144	1.92	1.93	1.92	1.94	1.81	2.14
GPT-Image	0.133	0.120	0.132	0.122	0.121	0.138	1.55	1.70	1.66	1.58	1.58	1.70
Gemini-Pro	0.109	0.096	0.110	0.095	0.098	0.112	1.46	1.54	1.51	1.48	1.45	1.56
我们的方法，256 像素，面部	0.084	0.082	**0.085**	0.081	0.078	0.092	**1.27**	1.44	**1.32**	1.38	1.33	**1.39**
512 像素，面部	0.086	**0.078**	0.086	**0.078**	0.079	**0.088**	1.29	**1.40**	**1.32**	**1.36**	**1.28**	1.44
256 像素，面部+头发	**0.083**	0.081	**0.085**	0.080	**0.077**	0.091	1.29	1.43	1.33	1.38	1.33	1.40

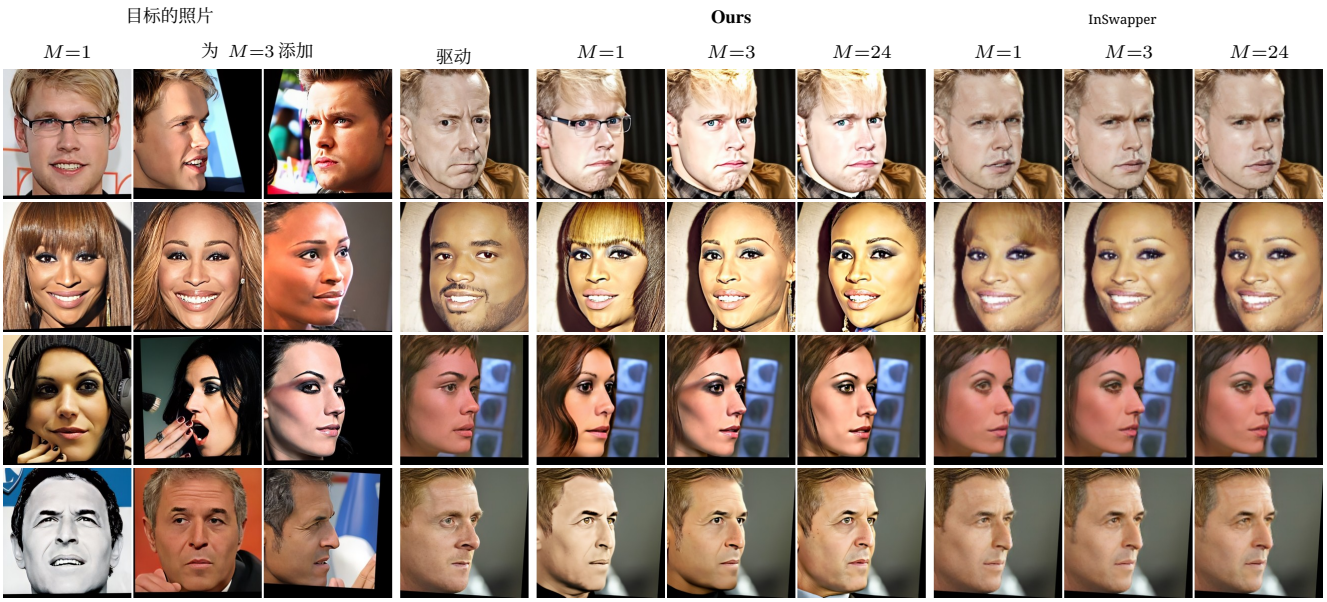


图 24. **使用更多目标照片的结果**。 身份基准测试中的四个目标对象，每个对象对应一张驱动图像。 $M=1$ 给出固定顺序中的第一张照片， $M=3$ 再加入接下来的两张， $M=24$ 则加入所有照片；InSwapper 对相同照片的身份嵌入取平均值。 $M=1$ 所示的照片从上到下分别呈现：眼镜、刘海、帽子下的长发，以及黑白胶片； $M=3$ 加入的照片没有这些特征，因此这些细节不会出现在我们的结果中。第三行中，第三张照片是唯一一张展示了驱动图像中所见面部侧面的照片，而该脸颊上的妆容在 $M=3$时出现在我们的结果中。

在图24中，随着照片不断添加，第一张照片中的细节会从结果中消失。图25展示了会融入结果的细节：特征—

面部中第一张照片遮挡而只有添加的照片显示的特征。对于另外三个人，我们对照片进行排序，使得第一张照片遮挡面部的某个细节，而在 {v1} 添加的两张照片中显示该细节。

表 24. **推理掩码的大小**(形状基准测试，348 对图像对）。面部掩码在 256像素处膨胀；整幅图像：掩码覆盖整幅图像。 s_{gt}，泄漏和三维同表 21。ArcFace：结果与真实值（目标的渲染图）之间的余弦相似度；这些角色没有留出照片。粗体：最佳。

膨胀	$s_{\text{gt}}\downarrow$	泄漏↓	3D↓	ArcFace↑
6 px	0.090	0.41	1.36	0.761
12 像素（默认）	**0.083**	0.34	**1.35**	0.772
18 像素	0.084	0.32	1.36	0.778
24 像素	0.091	**0.28**	1.38	0.782
整幅图像	0.113	**0.28**	1.36	**0.792**

表 25. **自增强，拆分为两种变形**，基于完整模型，关闭光照。移动最小二乘：对人脸轮廓进行移动最小二乘法变形；宽高比：宽高比的变化。各列同表 3。

变形	Id.↑	表情↓	姿态↓	光照↑	$s_{\text{gt}}\downarrow$	泄漏↓	3D↓
两者（完整模型）	0.616	0.256	3.97	0.66	**0.081**	**0.33**	1.26
仅移动最小二乘	0.615	0.259	3.96	0.66	0.084	0.37	1.27
仅使用宽高比	0.616	0.251	3.94	0.67	0.085	0.39	1.26

我们的结果从 $M{=}3$ 中呈现了痣和门牙间隙，并在 $M{=}24$ 处呈现了酒窝。InSwapper 的结果也在 $M{=}24$ 处显示了酒窝，但在任何 M 中都没有显示痣或门牙间隙，尽管在这三个示例中的每个 M 上，其 ArcFace 得分都高于我们的方法，并且在 $M{=}24$ 上高于所提供照片在 $M{=}1$ 处的得分。在最后一行中，对于身份基准数据集中的一位年龄更大的目标，随着 M 增大，我们的结果中的笑纹逐渐加深，而 InSwapper 的结果则保持面部更加平滑。

J. 消融实验

本节完成第 4.5 节。表 3 的第一行是阶段 2 第一阶段的模型；阶段 2 的其他行均以此为基础继续训练 100k 步（表 5）。

自增强、拆分。 自增强会对表情输入应用两次变形：先在人脸轮廓上使用移动最小二乘法，然后改变宽高比（第 A.3 节）。表 25 仅使用其中一种变形训练完整模型。两者都不会改变身份、表情、姿态或光照。单独使用每种变形都能保留部分脸型增益：仅使用移动最小二乘法时，泄漏值为 0。37，仅使用宽高比变形时为 0。39，而两者同时使用时为0。33；轮廓距离分别为 0。084 和 0。085，而两者同时使用时为 0。081。在这两个测量指标上，两种单独变形的结果都不同于完整模型，但彼此无法区分。

脸型变化下的姿态和表情。当图像的脸型发生变化时，姿态和表情度量也会随之变化（第D.1节）。因此，完整模型的姿态误差高于面部掩码（表-格3）的部分原因可能来自估计器本身，因为完整模型对脸型的改变更大。

图 26 展示了图21 中的图像对，以及表 3中的变体。

J.1. 光照流

表 3 的最后几行将光照流加入完整模型。在使用变形的两次处理中，光照相关性从 0.66 上升至 0.83，ArcFace 从 0.616 降至 0.601，而脸型基本保持不变（泄漏值为 0.34，相比之下为 0.33）。在不使用变形的单次处理中，结果更加趋向于光照（0.87），但也更多地呈现驱动图像的脸型（泄漏值为 0.60；图 19）。第 G.2 节展示了强光照下的结果，并比较了另一种同时读取耳朵和颈部的光照流（表20）。

K. 应用

K.1. 身份遍历

身份令牌 $\boldsymbol{C}_{\text{id}}$ 将目标身份带入结果，因此更改它们会改变身份，而驱动图像保持不变。我们利用这一点来改变同一个人的表观年龄。在固定的驱动图像上，我们根据此人年轻时拍摄的照片计算出 $\boldsymbol{C}^A_{\text{id}}$，并根据数十年后拍摄的照片计算出 $\boldsymbol{C}^B_{\text{id}}$，然后对二者进行线性插值：

$$\boldsymbol{C}_{\text{id}}(\tau) = (1-\tau)\,\boldsymbol{C}^A_{\text{id}} + \tau\,\boldsymbol{C}^B_{\text{id}}, \qquad \tau \in [0,1]. \quad (3)$$

每一帧都使用相同的驱动图像、随机种子和两次推理（第3.4节），因此帧与帧之间只有身份令牌发生变化。

图 27展示了四个人。每组照片包含5 至 11张照片，驱动图像是此人自己的照片之一，来自其中一组。随着 τ 从 0 变为1，表观年龄和面部形状逐渐变化，而表情、头部姿态、头发和背景则与驱动图像保持相近。身份编码器也会读取耳朵，因此在某组的多张照片中出现的耳环，可能会在靠近该组的一端出现（第二行和第四行）。

K.2. 使用掩码和提示词进行编辑

掩码决定图像中哪些部分会被生成（第 3.4 节），编辑提示词则会引导生成内容（第 A.4 节）。图28 使用一个模型和四种掩码。使用面部掩码（我们的默认设置）时，只有面部会发生变化。使用面部+和头发掩码，并输入关于头发的提示词时，头发也会遵循提示词生成，如图 1 所示。仅使用头发掩码时，只会生成头发，面部则保持为驱动图像中的样子。使用整幅图像掩码并输入关于背景的提示词时，场景会重新生成，而表情和姿态仍遵循驱动图像。此时，头发和服装取自目标的照片，面部上的光照仍遵循驱动图像，因此新背景对其影响很小。提示词采用第

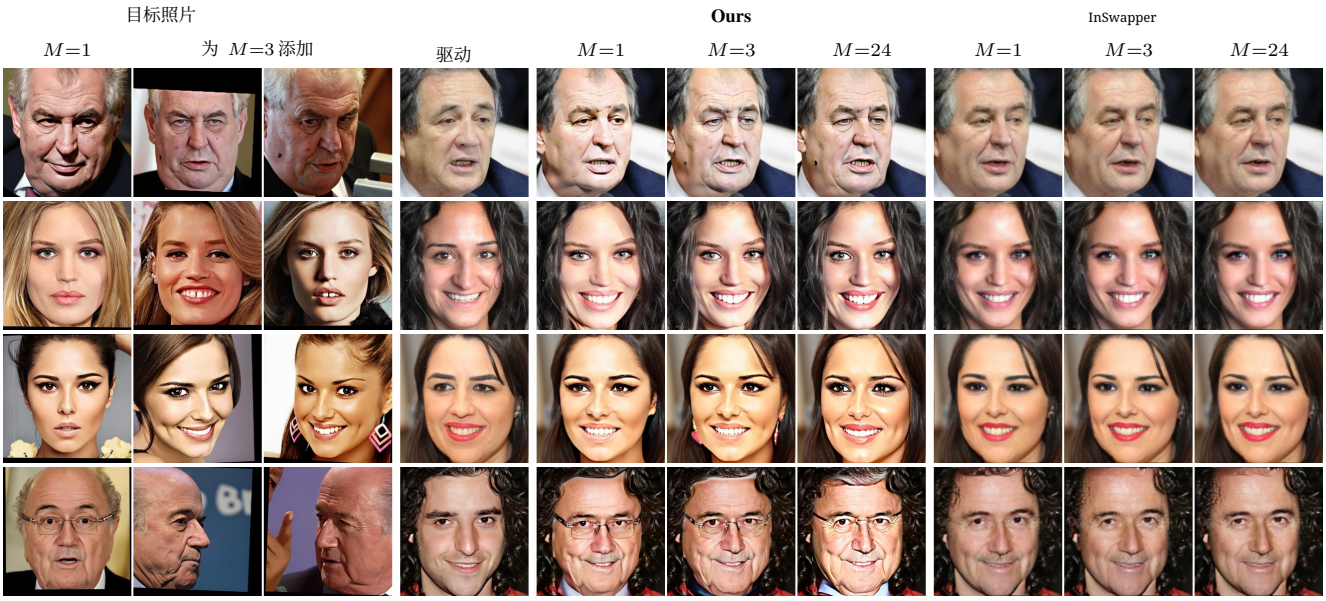


图 25. **增加照片带来的细节**。 第 1–3 行：VGGFace2-HQ 中的三个人物，搭配 FFHQ 的驱动图像；这些人物的照片按如下方式排列：在 $M=1$ 提供的照片中，面部的一处细节被遮住，而在 $M=3$ 添加的两张照片中则能看到该细节：脸颊上的痣、门牙间隙和酒窝。第 4 行：身份基准测试中的一个目标人物，其驱动图像带有微笑，照片顺序固定。InSwapper 对同一组照片中的身份嵌入取平均值。



图 26。 **消融实验中的脸型**。 图 21 与表 3中的变体对应的图像对，从固定人脸掩码到完整模型，以及阶段 3 的两种推理变体。线条说明同图7：真实值（绿色）和驱动图像（红色）的下颌线。

A.4 节中的形式，例 如“一个带辫子的照片级真实感人像” 或“一个……的照片级真实感人像” 日落时海滩上的人脸”。



图27。**身份遍历**。每一行在一组较年轻的照片与一组较年长的照片之间，对同一个人的身份令牌进行插值（每组展示四张），并使用固定的驱动图像。100 帧中的 6 帧；标题给出 τ，四舍五入到小数点后一位。

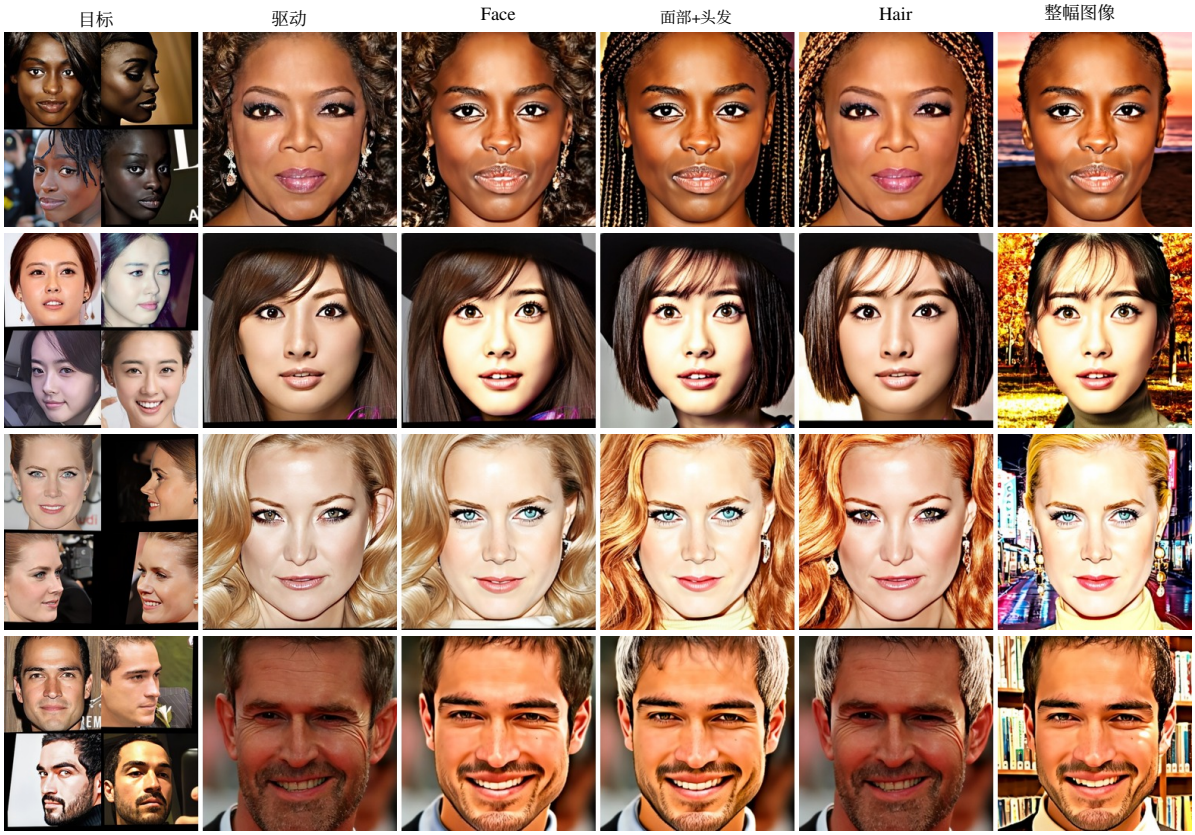


图28. **使用掩码和提示词进行编辑**。 一个模型，四种掩码。面部：默认面部掩码。面部+头发：使用针对头发的提示词。头发：仅包含头发的掩码，使用相同的提示词；面部保持为驱动图像中的面部。整幅图像：使用针对背景的提示词。提示词，第 1–4 行：辫子和日落时的海滩；短波波头和秋日公园；红色卷发和夜晚的霓虹灯街道；银灰色头发和图书馆。目标：其 24 张照片中的四张。



图29。**对各个流的控制**。每一行改变一个流的尺度，另外两个流的尺度均为 4。第 1 行和第 2 行：身份；第 1 行使用目标的 13 张展示眼镜的照片。第 3 行和第 4 行：表情。第 5 行和第 6 行：光照。目标：其中四张照片。

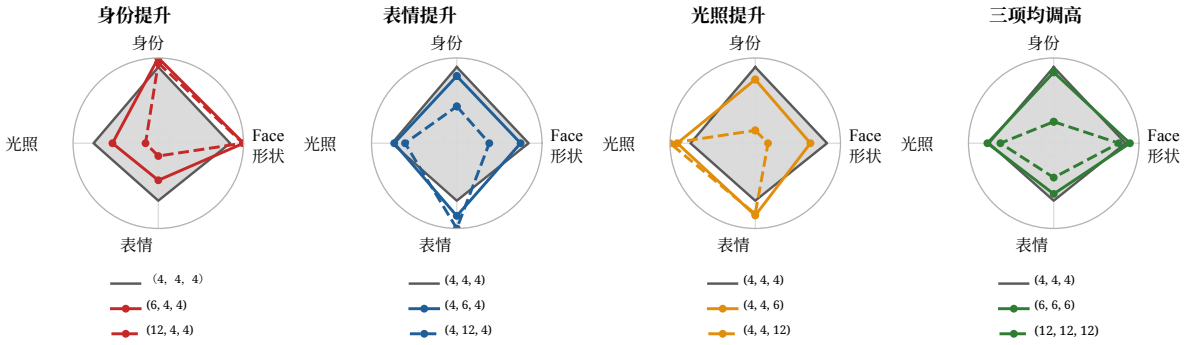


图30。**基准测试中对各个流的控制**。每个面板将一个流的尺度提高到 6（实线）或 12（虚线），另两个流保持在 4；最后一个面板则提高全部三个流。灰色：(4，4，4)，即我们的默认设置。身份：ArcFace；脸型：形状基准测试上的轮廓距离 s_{gt}；表情：EmoNet 误差；光照：光照相关性。每个坐标轴均按九种设置线性缩放，范围从最差值（最小的内圆）到最佳值（最外圆），因此细微差异看起来会很大；表 26 列出了具体数值。

K.3. 对每个流的控制

身份、表情和光照通过独立流输入模型，并且每个流在训练时都会独立丢弃

表26。 **对各个流的控制，测量结果**。 尺度： $(w_{\text{id}}, w_{\text{exp}}, w_{\text{light}})$。身份基准测试的 240 对图像及形状基准测试的 348 对图像的均值。每个均值下方：与 95% 自助法区间，表示其与 (4,4, 4)的差值，按图像对配对（泄漏指标按 29 个目标对象配对）。粗体：显著优于 (4, 4,4)；斜体：显著较差。 ΔL^*：结果的皮肤亮度减去驱动图像的皮肤亮度；越接近 0 越好（第 L节）。

	身份基准测试						形状基准测试		
尺度	ArcFace↑	表情。↓	姿态↓	光照↑	FaceScore↑	ΔL^*	s_{gt} ↓	泄漏↓	3D↓
(4, 4, 4)	0.601	0.253	3.99	0.833	4.62	7.8	0.0830	0.337	1.35
(6, 4, 4)	**0.615**	*0.266*	*4.18*	*0.811*	4.61	*8.2*	**0.0809**	0.327	**1.26**
	[+0.011, +0.016]	[+0.003, +0.023]	[+0.07, +0.33]	[−0.027, −0.017]	[−0.04, +0.02]	[+0.2, +0.7]	[−0.0039, −0.0005]	[−0.023, +0.004]	[−0.11, −0.08]
(12, 4, 4)	**0.608**	*0.281*	*4.47*	*0.773*	*4.47*	*8.6*	0.0810	0.333	**1.28**
	[+0.002, +0.011]	[+0.011, +0.045]	[+0.26, +0.69]	[−0.072, −0.050]	[−0.21, −0.09]	[+0.2, +1.3]	[−0.0052, +0.0009]	[−0.033, +0.028]	[−0.11, −0.04]
(4, 6, 4)	*0.589*	**0.244**	**3.87**	0.831	*4.56*	**7.0**	0.0843	*0.377*	*1.42*
	[−0.015, −0.011]	[−0.020, −0.0003]	[−0.22, −0.01]	[−0.007, +0.001]	[−0.09, −0.03]	[−1.0, −0.6]	[−0.0008, +0.0034]	[+0.024, +0.056]	[+0.05, +0.08]
(4, 12, 4)	*0.544*	**0.236**	3.83	*0.818*	*4.40*	**4.8**	*0.0894*	*0.418*	*1.62*
	[−0.062, −0.052]	[−0.032, −0.002]	[−0.34, +0.02]	[−0.022, −0.009]	[−0.27, −0.16]	[−3.4, −2.5]	[+0.0022, +0.0107]	[+0.047, +0.117]	[+0.24, +0.31]
(4, 4, 6)	*0.583*	**0.244**	*4.11*	**0.849**	*4.57*	7.6	*0.0857*	0.340	*1.50*
	[−0.021, −0.015]	[−0.016, −0.002]	[+0.02, +0.22]	[+0.011, +0.019]	[−0.07, −0.01]	[−0.4, +0.1]	[+0.0013, +0.0041]	[−0.007, +0.014]	[+0.13, +0.17]
(4, 4, 12)	*0.509*	0.245	4.15	**0.857**	*4.33*	**7.1**	*0.0925*	0.349	*1.85*
	[−0.098, −0.086]	[−0.022, +0.006]	[−0.03, +0.34]	[+0.017, +0.031]	[−0.35, −0.23]	[−1.3, −0.2]	[+0.0063, +0.0128]	[−0.011, +0.035]	[+0.46, +0.54]
(6, 6, 6)	*0.594*	0.257	4.05	0.835	*4.51*	7.6	0.0823	0.352	*1.40*
	[−0.010, −0.005]	[−0.007, +0.015]	[−0.10, +0.23]	[−0.003, +0.006]	[−0.14, −0.07]	[−0.4, +0.02]	[−0.0033, +0.0019]	[−0.007, +0.041]	[+0.02, +0.06]
(12, 12, 12)	*0.522*	0.268	*4.35*	*0.820*	*4.18*	**7.1**	0.0841	0.364	*1.59*
	[−0.085, −0.074]	[−0.003, +0.032]	[+0.12, +0.60]	[−0.021, −0.006]	[−0.49, −0.38]	[−1.1, −0.3]	[−0.0029, +0.0050]	[−0.012, +0.067]	[+0.20, +0.27]

（见第A.2节）。因此，每个流在推理时都可以使用各自的引导尺度。当引导尺度为 w_{id}、 w_{exp} 和w_{light} 时，预测结果为

$$\boldsymbol{v} = \boldsymbol{v}_{\text{u}} + 4\,(\boldsymbol{v}_{\text{c}} - \boldsymbol{v}_{\text{u}}) + \sum_{s} (w_s - 4)\,(\boldsymbol{v}_{\text{c}} - \boldsymbol{v}_{\text{c}\setminus s}), \quad (4)$$

其中 s 在三个流上运行，而$\boldsymbol{v}_{\text{c}\setminus s}$ 是丢弃流 s时的条件预测，与训练时相同。三个尺度均设为4时，这就是第A.4节中的引导。第一次传递不包含光照流，因此仅应用身份项和表情项。三个尺度为艺术家提供了每个流一个控制项。视觉特效镜头可能需要更多演员的面部细节、更强烈的表演，或更贴合片场光照的效果；艺术家可以针对每个镜头权衡这些因素。

图 29每次只改变一个尺度，另外两个保持为4。较高的身份尺度会让目标的特征更加突出，例如第 1 行的眼镜和皱纹，以及第 2 行额头上的纹路。在第 2 行中，尺度为4 和6 时的结果看起来都像目标，艺术家可以选择这些特征呈现得多么明显。较高的表情尺度会让笑容更强烈（第 3 行）。第 4 行中，驱动图像里的人物眯起眼睛，眉头紧皱：眉间的纹路在4时显现，并在6时加深；在12 时，纹路和眯眼程度都比驱动图像中更明显。较高的光照尺度会让横跨眼睛的红色光带更加清晰（第 5 行），而第 6 行中，默认设置已能跟随阳光。在24时，每一行的结果都会变差。

表26 和图30 测量了这些控制项在身份基准测试（240 图像对）和形状基准测试（348图像对）上的表现。每个尺度都会改变其所属流的测量指标。在6时，ArcFace 从0.601 升至0.615 ，而轮廓距离从 0.0830 降至 0.0809。表情误差从 0.253 降至 0.244，皮肤亮度也更接近驱动图像（第 L 节）。光照相关性从 0.833 升至 0.849。每项提升都要以牺牲其他流的部分表现为代价。尺度为 12 时，代价增大，身份方面的提升也有所回落（ArcFace 为 0.608）。同时提高三个尺度并不能叠加这些收益。在（6, 6, 6）时，结果与（4, 4, 4）相近，但有些指标略有下降，例如 ArcFace 为 0.594。在（12, 12, 12）时，大多数测量指标都更差（ArcFace 为 0.522）。八种设置中，没有一种在所有测量指标上都至少与（4, 4, 4）一样好。我们将三个流的默认值都设为 4，以保持测量指标的平衡；对于有特定需求的镜头，艺术家可以据此调整。

L. 局限性与更广泛影响

局限性。MYSTIQUE 会分别处理每张驱动图像：它没有时序模型，将其扩展到视频是未来的工作。在身份基准测试的 88% 图像对中，我们的结果比驱动图像中的肤色更浅，平均相差7。8 个 L^*（CIELAB），而 InSwapper、SimSwap 和 Face-Adapter 的结果与其相差不超过1 个单位。不同肤色的偏差大致相同，因此深肤色目标的结果与目标肤色相差最远（最深肤色的三分之一目标对象在+21 个体色型角上的偏差，而最浅肤色目标对象为 +7 ）。即使不使用光照流，这种现象依然存在。降低引导尺度可以减轻这一偏差：在这些图像对中，针对63 % 的图像对，将尺度从4 降至2后，偏差从7。6降至4。0，代价是身份一致性略有下降（ ArcFace −0.014）。此外，光照流无法完全将肤色与光照分离（图31，上），有时还会使面部过曝，遮蔽细微的纹理细节。身份流从不



图31. **局限性**。 上： 肤色和光照。目标人物肤色较浅，而驱动图像中人物肤色较深，且处于阳光下。不使用光照流（推理的第1遍）时，我们的方法在平坦光照下保留目标人物的肤色。使用光照流后，结果会呈现驱动图像中的光照，也会带有部分驱动人物的肤色。中： 头发来自驱动图像。使用面部掩码时，我们的方法会在目标人物的面部下方保留驱动图像中儿童的头发。使用面部+头发掩码时，头发由模型生成：此处头发为白色，与目标人物的头发相同；但身份流不会看到头发，因此不会迁移头发。下： 半透明面纱。它未被检测为遮挡物，因此位于掩码内。我们的模型在第1遍就会生成没有面纱的面部，而掩码外的面纱部分则会保留。

看到头发。因此，即使头发不适合目标对象，也会保留驱动图像中的头发（图31，中）；或者使用面部+头发掩码，生成头发而非进行迁移。覆盖在面部的半透明面纱既不是不透明的遮挡物，也不是眼镜。如果它未被检测为遮挡物，我们的模型会生成不带面纱的面部，而掩码外的面纱部分则会保留（图31，下）。最后，形状基准测试只评估渲染角色的下颌线；对于并非真实人物的面部而言，下颌线只是脸型的一个方面。

更广泛的影响。 身份迁移可能被滥用于制作未经同意的影像、传播虚假信息，以及盗用他人的肖像；能够更忠实地渲染身份的方法会加剧这些风险。MYSTIQUE 旨在用于影视制作中的视觉特效，且须获得影像中人物的同意。