真实头部说话神经模型的少样本对抗学习

Egor Zakharov[1;2] Aliaksandra Shysheya[1;2] Egor Burkov[1;2] Victor Lempitsky[1;2]
[1]Samsung AI Center, Moscow [2]Skolkovo Institute of Science and Technology

图 1：使用从同一个人的不同视频序列（左侧）提取的面部标记轨迹，并使用不同人的面部标记（右侧）来合成头部说话图像的结果。 结果取决于从目标帧获取的界标，而源帧是来自训练集的示例。 左侧的说话的头部模型使用 8 帧进行训练，而右侧的模型则以一次性方式进行训练。

摘 要

最近的几项研究表明，通过训练卷积神经网络来做生成任务，可以获得高度逼真的人体头部图像。为了创建个性化的头部说话模型，这些工作需要对单个人的图像的大量数据进行训练。然而，在许多实际场景中，需要从人的一些图像视图（甚至可能是单个图像）学习这种个性化的头部说话模型。在这里，我们提出了一个具有如此少次数能力的系统。它在大型视频数据集上执行冗长的元学习，之后能够将以前不曾见的人的头部说话神经模型的少次或单次学习作为高容量生成器和判别器的对抗性训练问题。至关重要的是，该系统能够以特定于人的方式初始化生成器和判别器的参数，因此尽管需要调整数千万个参数，但训练可以仅基于少量图像并快速完成。我们表明，这种方法能够学习新人甚至肖像画的高度逼真和个性化的头部说话模型。

1. 介绍

在这项工作中，我们考虑创建个性化的真实照片级头部说话模型的任务，即系统可以合成含有语音表达的合理视频序列和特定个体的模拟。更具体地说，我们考虑了一组面部标记来合成视觉个性化的头部图像的问题，这些标记驱动模型的动画。这种能力能推动远程呈现的实际应用，包括视频会议和多人游戏，以及一些特殊行业。已知合成逼真的头部说话序列面临两种困难。首先，人头具有高光度，几何和运动复杂性。这种复杂性不仅源于建模面部（存在大量建模方法），还来自创造口腔，头发和服装。第二个复杂因素是人类视觉系统对人类头部外观建模中的微小错误（即所谓的神秘谷效应[24]）的敏锐性。对模型错误的这种低容忍度解释了当前在许多实际部署的远程电子激励系统中非真实感卡通式化身的普遍存在。

为了克服这些挑战，已经有几项工作被提出，来通过扭曲单个或多个静态帧来合成关节头部序列。经典的变形算法[5,28]和使用机器学习（包括深度学习）[11,29,40]合成的变形场景都可以用于这样的目的。虽然基于翘曲的系统可以从一个单一的图像创建头部说话部序列，运动量，头部旋转和去除遮挡

但他们处理出没有明显的人工制品痕迹的能力是有限的。

使用经过对侧训练的深度卷积网络（ConvNets）直接（无翘曲）合成视频帧，为具有真实感的说话人合成提供了新的希望。最近，这些系统已经证明了一些非常现实的结果[16,20,37]。然而，为了成功，这些方法必须训练大型网络，其中生成器和判别器对于每个头部说话模型具有数十万个参数。因此，这些系统需要几分钟长的视频[20,37]或大型照片数据集[16]以及数小时的 GPU 训练才能创建新的个性化头部说话模型。虽然这种效果低于使用非常复杂的物理和光学建模构建照片般逼真的头部模型的系统[1]，但对于大多数实际的远程呈现场景来说，它仍然是特别需要的，我们希望用户能够创建他们的个性化头部模型尽可能少的费力。

在这项工作中，我们提出了一个系统，用于从少量照片（所谓的 few-shot 学习）和有限的训练时间创建头部说话模型。事实上，我们的系统可以基于单张照片（one-shot 学习）生成合理的结果，而添加更多照片可以提高个性化的保真度。类似于[16,20,37]，由我们的模型创建的头部说话是深度的卷积网络，它通过一系列卷积操作而不是翘曲直接合成视频帧。因此，我们系统创建的头部说话可以处理各种各样的姿势，这些姿势超越了基于翘曲的系统的能力。

通过对与不同说话者相对应的大型头部对话视频进行广泛的预训练（元学习）来获得少样本学习能力。在元学习过程中，我们的系统模拟了少样本的学习任务，并学习了具有里程碑意义的给定标记转换成逼真的人物化照片，给予这个人一套小的训练图像。在那之后，一些新人的照片通过元学习预先训练了高容量生成器和判别器，从而形成了新的对抗性学习问题。新的对抗性问题会聚合到经过几个训练步骤后生成逼真和个性化图像的状态。

在实验中，我们通过定量测量和用户研究提供了由我们的系统与替代神经头部说话模型[16,40]创建的头部说话图片的比较，其中我们的方法生成足够的现实主义和个性化保真度的图像来欺骗研究参与者。我们演示了我们的头部说话模型的几种用途，包括使用从同一人的视频序列中提取的标记轨迹的视频合成，以及木偶操作（基于不同人的面部标记轨迹的某人的视频合成）。

2. 相关工作

大量的作品致力于人脸外观的统计模型[6]，使用经典技术获得了非常好的结果[35]，最近又用深度学习[22,25]（只列举少量工作）。虽然面部建模是谈论头部建模的高度相关的任务，但是这两个任务并不相同，因为后者还涉及对非面部部分进行建模，例如头发，颈部，口腔以及肩部/上衣。这些非面部部件不能通过面部建模方法的一些微小的扩展来处理，因为它们不太适合于配准，并且通常具有比面部部位更高的可变性和更高的复杂性。原则上，面部建模[35]或嘴唇建模[31]的结果可以拼接成现有的头部视频。然而，这种设计不允许完全控制所得视频中的头部旋转，因此不会产生完全成熟的头部说话系统。

我们系统的设计借鉴了最近图像生成建模的进展。因此，我们的架构使用对抗性训练[12]，更具体地说，是条件判别器背后的思想[23]，包括投影判别器[32]。我们的元学习阶段使用自适应实例规范化机制[14]，这被证明在大规模条件生成任务中很有用[2,34]。

与模型无关的元学习者（MAML）[10]使用元学习来获得图像分类器的初始状态，在没有训练样本的情况下，它可以快速收敛到看不见的类的图像分类。我们的方法也使用了这种高级想法，尽管我们的实现方式有很大不同。有几项工作进一步提出将对抗训练与元学习相结合。因此，数据增强 GAN [3]，Meta-GAN [43]，对抗性元学习[41]使用经过对侧训练的网络来为元学习阶段未见的类生成其他示例。虽然这些方法专注于提高少样本分类性能，但我们的方法使用类似的对抗性目标来处理图像生成模型的训练。总而言之，我们将对抗性微调纳入元学习框架。在我们通过元学习阶段获得生成器的初始状态和判别器网络之后应用前者。

最后，与我们的相关的是最近关于文本到语音生成的两篇文章[4,18]。它们的设置（生成模型的少样本学习）和一些组件（独立的嵌入式网络，生成器微调）也在我们的案例中使用。我们的工作在应用领域，对抗性学习的使用，对元学习过程的具体适应以及众多实施细节方面存在差异。

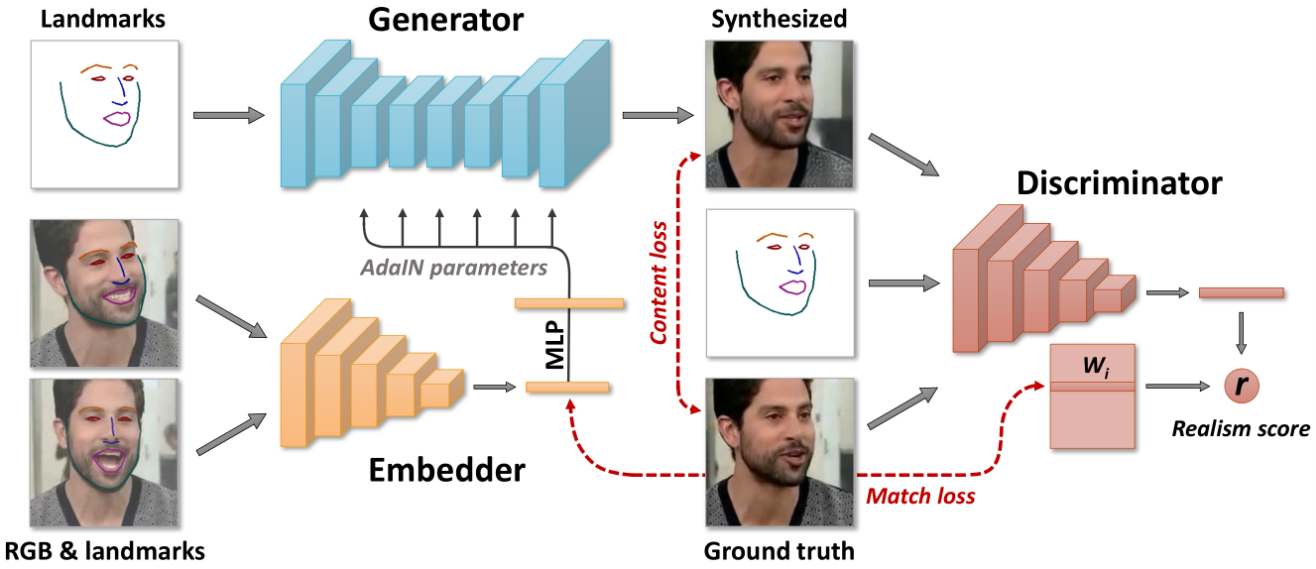

图 2：我们的元学习架构涉及将头部图像（带估计的面部标记）映射到嵌入向量的嵌入式网络，嵌入向量包含与姿势无关的信息。 生成器网络通过卷积层集将输入面部标记映射到输出帧中，卷积层通过自适应实例归一化由嵌入向量调制。 在元学习期间，我们通过嵌入器传递来自同一视频的帧集，将得到的嵌入平均并使用它们来预测生成器的自适应参数。 然后，我们通过生成器传递不同帧的标记，将得到的图像与完全真实图像进行比较。 我们的目标函数包括感知和对抗性损失，后者通过条件投影判别器实现。

3. 方法

3.1. 模型架构和标记

我们的方法的元学习阶段假定 M 视频序列的可用性，包含不同人的头部说话。 我们用$\mathbf{x}_i$表示第 i 个视频序列，用$\mathbf{x}_i(t)$表示第 t 帧。 在学习过程中以及在测试时间期间，我们假设所有帧的面部标记位置的可用性（我们使用现成的面部对齐代码[7]来获得它们）。 使用预定义的一组颜色将标记光栅化为三通道图像，以将某些标记与线段连接。 我们用$\mathbf{y}_i(t)$表示为$\mathbf{x}_i(t)$计算的结果标记图像。

在我们方法的元学习阶段，训练了以下三个网络（图 2）：

- 嵌入器$E(\mathbf{x}_i(s), \mathbf{y}_i(s); \phi)$采用视频帧$\mathbf{x}_i(s)$，相关联的标记图像$\mathbf{y}_i(s)$并将这些输入映射成 N 维向量$\hat{\mathbf{e}}_i(s)$。 这里，φ表示在元学习阶段中学习的网络参数。 通常，在元学习期间，我们旨在学习φ使得向量$\hat{\mathbf{e}}_i(s)$包含对于姿势不变并且在特定帧 s 中模仿的视频特定信息（诸如人的身份）。 我们将由嵌入器计算的嵌入向量表示为$\hat{\mathbf{e}}_i$。
- 生成器$G(\mathbf{y}_i(t), \hat{\mathbf{e}}_i; \psi, \mathbf{P})$获取嵌入器未看到的视频帧的标记图像$\mathbf{y}_i(t)$，预测视频嵌入$\hat{\mathbf{e}}_i$并输出合成大小的视频帧$\mathbf{x}_i(t)$。 训练生成器以最大化其输出和完全真实帧之间的相似性。 生成器的所有参数分为两组：人物通用参数ψ和人物特定参数$\hat{\psi}_i$。 在元学习期间，仅 ψ 被直接训练，而$\hat{\psi}_i$使用可训练投影矩阵$\mathbf{P}$：$\hat{\psi}_i = \mathbf{P}\hat{\mathbf{e}}_i$从嵌入向量$\hat{\mathbf{e}}_i$来预测。
- 判别器$D(\mathbf{x}_i(t), \mathbf{y}_i(t), i; \theta, \mathbf{W}, \mathbf{w}_0, b)$采用视频帧$\mathbf{x}_i(t)$，相关的标记图像$\mathbf{y}_i(t)$和训练序列 i 的索引。 这里，θ，W，$\mathbf{w}_0$和 b 表示与判别器相关的可学习参数。 判别器包含 ConvNet 部分$V(\mathbf{x}_i(t), \mathbf{y}_i(t); \theta)$，其将输入帧和标记图像映射为 N 维向量。 判别器预测单个标量（真实性得分）r，其表示输入帧$\mathbf{x}_i(t)$是否是第 i 个视频序列的真实帧以及它是否与输入姿势$\mathbf{y}_i(t)$匹配，基于其 ConvNet 部分的输出和参数 W，$\mathbf{w}_0$，b

3.2. 元学习阶段

在我们方法的元学习阶段，所有三个网络的参数都以对抗的方式进行训练

它通过模拟 K-shot 学习的场景来完成（在我们的实验中 K = 8）。 在每一批次中，我们随机地绘制了训练视频序列 i 和该序列中的单帧 t。 除了 t 之外，我们随机绘制了附加的 K 帧 $s_1, s_2, \ldots, s_K$ 来自同一序列。 然后，我们通过简单地平均为这些附加帧预测的嵌入 $\hat{\mathbf{e}}_i(s_k)$ 来计算第 i 个视频嵌入的估计 $\hat{\mathbf{e}}_i$：

$$\hat{\mathbf{e}}_i = \frac{1}{K}\sum_{k=1}^{K} E\left(\mathbf{x}_i(s_k), \mathbf{y}_i(s_k);\, \phi\right) . \tag{1}$$

然后，基于估计的嵌入 $\hat{\mathbf{e}}_i$，计算第 t 帧的重建 $\mathbf{x}_i(t)$：

$$\hat{\mathbf{x}}_i(t) = G\left(\mathbf{y}_i(t), \hat{\mathbf{e}}_i;\, \psi, \mathbf{P}\right) . \tag{2}$$

然后优化嵌入器和生成器的参数以最小化包括内容项，对抗项和嵌入匹配项的以下目标：

$$\begin{aligned}\mathcal{L}(\phi, \psi, \mathbf{P}, \theta, \mathbf{W}, \mathbf{w}_0, b) = \mathcal{L}_{\text{CNT}}(\phi, \psi, \mathbf{P}) + \\ \mathcal{L}_{\text{ADV}}(\phi, \psi, \mathbf{P}, \theta, \mathbf{W}, \mathbf{w}_0, b) + \mathcal{L}_{\text{MCH}}(\phi, \mathbf{W}) .\end{aligned} \tag{3}$$

在（3）中，内容丢失项 $\mathcal{L}_{\text{CNT}}$ 使用感知相似性度量[19]测量完全真实图像 $\mathbf{x}_i(t)$ 与重建 $\hat{\mathbf{x}}_i(t)$ 之间的距离，与 VGG19 [30]相对应。 网络被训练用于 ILSVRC 分类和 VGGFace 网络 [27]被训练用于面部验证。 损失计算式为这些网络的特征之间的 L1 损失的加权和。

（3）中的对抗项对应于需要最大化的判别器计算的真实性得分，以及使用判别器计算的特征匹配项[38]，其基本上是感知相似性度量（它有助于训练的稳定性）：

$$\begin{aligned}\mathcal{L}_{\text{ADV}}(\phi, \psi, \mathbf{P}, \theta, \mathbf{W}, \mathbf{w}_0, b) = \\ -D(\hat{\mathbf{x}}_i(t), \mathbf{y}_i(t), i; \theta, \mathbf{W}, \mathbf{w}_0, b) + \mathcal{L}_{\text{FM}} .\end{aligned} \tag{4}$$

根据投影判别器的想法[32]，矩阵 W 的列包含对应于各个视频的嵌入。 判别器首先将其输入映射到 N 维向量 $V(\mathbf{x}_i(t), \mathbf{y}_i(t); \theta)$，然后计算真实度得分为：

$$\begin{aligned}D(\hat{\mathbf{x}}_i(t), \mathbf{y}_i(t), i;\, \theta, \mathbf{W}, \mathbf{w}_0, b) = \\ V\left(\hat{\mathbf{x}}_i(t), \mathbf{y}_i(t);\, \theta\right)^T (\mathbf{W}_i + \mathbf{w}_0) + b ,\end{aligned} \tag{5}$$

其中 $\mathbf{W}_i$ 表示矩阵 W 的第 i 列。同时，$\mathbf{w}_0$ 和 b 不依赖于视频索引，因此这些项对应于 $\mathbf{x}_i(t)$ 的一般现实性及其与具有里程碑意义的标记图像 $\mathbf{y}_i(t)$。

因此，在我们的系统中存在两种视频嵌入：由嵌入器计算的视频嵌入，以及与分解器中的矩阵 W 的列对应的视频嵌入。（3）中的匹配项 $\mathcal{L}_{\text{MCH}}(\phi, \mathbf{W})$ 通过惩罚 $\hat{\mathbf{e}}_i$ 和 $\mathbf{W}_i$ 之间的 L1 距离来提升两种类型的嵌入的相似性。

当我们更新嵌入器的参数和生成器的参数时，我们也会更新参数θ，W，$\mathbf{w}_0$，判别器的 b。 更新由以下铰链损失的最小化驱动，这促使真实图像 $\mathbf{x}_i(t)$ 上的真实性得分的增加及该分数在合成图像 $\hat{\mathbf{x}}_i(t)$ 上的减少：

$$\begin{aligned}\mathcal{L}_{\text{DSC}}(\phi, \psi, \mathbf{P}, \theta, \mathbf{W}, \mathbf{w}_0, b) = \\ \max(0, 1 + D(\hat{\mathbf{x}}_i(t), \mathbf{y}_i(t), i;\, \phi, \psi, \theta, \mathbf{W}, \mathbf{w}_0, b)) + \\ \max(0, 1 - D(\mathbf{x}_i(t), \mathbf{y}_i(t), i;\, \theta, \mathbf{W}, \mathbf{w}_0, b)) .\end{aligned} \tag{6}$$

因此，目标（6）比较假实例 $\hat{\mathbf{x}}_i(t)$ 和实例 $\mathbf{x}_i(t)$ 的真实性，然后更新判别器参数以将这些分数分别推到 1 以上和-1 以下。 通过更改嵌入器和生成器的更新来进行训练，最大限度地降低 $\mathcal{L}_{\text{CNT}}$，$\mathcal{L}_{\text{ADV}}$ 和 $\mathcal{L}_{\text{MCH}}$ 的损失，使用判别器的更新来最小化 $\mathcal{L}_{\text{DSC}}$ 的损失。

3.3. 通过微调来进行少样本学习

一旦元学习融合，我们的系统就可以学习在元学习阶段不曾见的新人的头部说话序列。 和以前一样，合成是以标记图像为条件的。 系统是以少样本方式学习的，假设 T 训练一组图像 $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)$（例如，相同视频的 T 帧）并且给出 $\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(T)$ 是相应的标记图像。 注意，帧数 T 不必等于元学习阶段中使用的 K.

当然，我们可以使用元学习的嵌入器来为新的头部说话序列进行嵌入：

$$\hat{\mathbf{e}}_{\text{NEW}} = \frac{1}{T}\sum_{t=1}^{T} E(\mathbf{x}(t), \mathbf{y}(t);\, \phi) , \tag{7}$$

重用在元学习阶段估计的参数φ。 然后，生成与新标记图像相对应的新帧的简单方法是使用估计嵌入编码 $\hat{\mathbf{e}}_{\text{NEW}}$ 和元变化参数ψ以及投影矩阵 P 来应用生成器。通过这样做，我们发现生成的图像看似合理且逼真，但是，对于大多数旨在实现高个性化程度的应用而言，通常存在相当大的不可识别身份的缺陷。

这种身份识别缺陷通常可以通过微调阶段弥合。 微调过程可视为元学习的简化版本，具有单个视频序列和

更少的帧数。 微调过程涉及以下组件：

• 生成器$G(\mathbf{y}(t), \hat{\mathbf{e}}_{\text{NEW}};\ \psi, \mathbf{P})$现在用$G'(\mathbf{y}(t);\ \psi, \psi')$代替。如前所述，它采用标记图像$\mathbf{y}(t)$并输出合成帧$\hat{\mathbf{x}}(t)$。重要的是，人物身份标识的生成器参数，我们用$\psi'$表示，现在直接与人物通用参数ψ一起优化。 我们仍然使用计算嵌入$\hat{\mathbf{e}}_{\text{NEW}}$和在元学习阶段估计的投影矩阵 P 来初始化$\psi'$，即我们从$\psi' = \mathbf{P}\hat{\mathbf{e}}_{\text{NEW}}$开始。

• 判别器$D'(\mathbf{x}(t), \mathbf{y}(t); \theta, \mathbf{w}', b)$，如前所述，计算真实性得分。 其 ConvNet 部分$V(\mathbf{x}(t), \mathbf{y}(t); \theta)$和偏差 b 的参数θ被初始化为元学习阶段的结果。 下面讨论$\mathbf{w}'$的初始化。

在微调期间，判别器的真实性得分以与元学习阶段类似的方式获得：

$$D'(\hat{\mathbf{x}}(t), \mathbf{y}(t);\ \theta, \mathbf{w}', b) = V\left(\hat{\mathbf{x}}(t), \mathbf{y}(t);\ \theta\right)^T \mathbf{w}' + b\,. \tag{8}$$

从表达式（5）和（8）的比较可以看出，矢量$\mathbf{w}'$在微调阶段的作用与元学习阶段中矢量$\mathbf{W}_i + \mathbf{w}_0$的作用相同。 对于初始化，我们无法访问$\mathbf{W}_i$的新模拟项（因为此人不在元学习数据集中）。 然而，元学习过程中的匹配项$\mathcal{L}_{\text{MCH}}$确保了判别器视频嵌入与嵌入器计算的向量之间的相似性。 因此，我们可以将$\mathbf{w}'$初始化为$\mathbf{w}'$和$\hat{\mathbf{e}}_{\text{NEW}}$的总和。

一旦建立了新的学习问题，微调阶段的损失函数就直接来自元学习变量。 因此，优化生成器参数ψ和ψ'以最小化简化的目标：

$$\mathcal{L}'(\psi, \psi', \theta, \mathbf{w}', b) = \mathcal{L}'_{\text{CNT}}(\psi, \psi') + \mathcal{L}'_{\text{ADV}}(\psi, \psi', \theta, \mathbf{w}', b)\,, \tag{9}$$

其中$t \in \{1 \ldots T\}$是训练样本的编号。 判别器参数θ，$\mathbf{w}_{\text{NEW}}$，b 通过最小化与（6）中相同的铰链损失来优化：

$$\begin{aligned}\mathcal{L}'_{\text{DSC}}(\psi, \psi', \theta, \mathbf{w}', b) = \\ \max(0, 1 + D(\hat{\mathbf{x}}(t), \mathbf{y}(t);\ \psi, \psi', \theta, \mathbf{w}', b)) + \\ \max(0, 1 - D(\mathbf{x}(t), \mathbf{y}(t);\ \theta, \mathbf{w}', b))\,.\end{aligned} \tag{10}$$

在大多数情况下，微调生成器提供了更好的训练序列拟合。 通过元学习阶段初始化所有参数也是非常重要的。 正如我们在实验中所展示的那样，这种初始化注入了一个强大的逼真的头部说话模型，这允许我们的模型推断和预测具有不同头部姿势和面部表情的姿势的逼真图像。

3.4. 实现细节

我们将生成器网络$G(\mathbf{y}_i(t), \hat{\mathbf{e}}_i; \phi, \mathbf{P})$建立在 Johnson 等人提出的图像到图像转换架构上[19]，但用[2]类似的方法用残差块代替下采样和上采样层（批量归一化[15]用实例归一化[36]代替）。人物身份标识的参数$\hat{\psi}_i$作为实例归一化层的仿射系数，遵循[14]中提出的自适应实例归一化技术，尽管我们仍然在编码标记图像$\mathbf{y}_i(t)$的下采样块中使用标准（非自适应）实例归一化层。

对于嵌入器$E(\mathbf{x}_i(s), \mathbf{y}_i(s); \phi)$和判别器$V(\mathbf{x}_i(t), \mathbf{y}_i(t); \theta)$的卷积部分，我们使用类似的网络，其由残差下采样块组成（与生成器中使用的相同，但没有标准化层）。与嵌入器相比，判别器网络在末端具有额外的残差块，其以 4×4 空间分辨率操作。为了获得两个网络中的矢量化输出，我们在空间维度上执行全局求和池化，然后是 ReLU。

我们对所有网络中的所有卷积和完全连接层使用频谱归一化[33]。我们还使用自注意块[2]和[42]。它们以 32×32 空间分辨率插入网络的所有下采样部分，并在生成器的上采样部分以 64×64 分辨率插入。

对于$\mathcal{L}_{\text{CNT}}$的计算，我们评估了 Conv1,6,11,20,29 VGG19 层和 Conv1,6,11,18,25 VGGFace 层激活之间的 L 1 损失，用于真实和假图像。我们将这些损失加权，对于 VGG19，权重等于$1 \cdot 10^{-2}$，对于 VGGFace 项，权重等于$2 \cdot 10^{-3}$。我们对这两种网络使用 Caffe [17]训练版本。对于$\mathcal{L}_{\text{FM}}$，我们在判别器网络的每个残余块之后使用激活，并且权重等于$1 \cdot 10^{1}$。最后，对于$\mathcal{L}_{\text{MCH}}$，我们将权重设置为$8 \cdot 10^{1}$。

我们将卷积层中的最小通道数设置为 64，并将最大通道数以及嵌入向量的大小 N 设置为 512。总共，嵌入器有 1500 万个参数，生成器有 3800 万个参数。判别器的卷积部分有 2000 万个参数。使用 Adam [21]优化网络。我们将嵌入器和生成器网络的学习速率设置为5×10^{-5}并且判别器设置为2×10^{-4}，在后者中使用两个更新步骤，采用[42]的方法。

4. 实验

两个带有头部说话视频的数据集用于定量和定性评估：VoxCeleb1 [26]（256p 视频，1 fps）和 VoxCeleb2 [8]（224p 视频，25 fps），后者大约有 10 倍的视频数量

Method (T)	FID↓	SSIM↑	CSIM↑	USER↓
VoxCeleb1				
X2Face (1)	45.8	**0.68**	**0.16**	0.82
Pix2pixHD (1)	**42.7**	0.56	0.09	0.82
Ours (1)	43.0	0.67	0.15	**0.62**
X2Face (8)	51.5	**0.73**	**0.17**	0.83
Pix2pixHD (8)	**35.1**	0.64	0.12	0.79
Ours (8)	38.0	0.71	**0.17**	**0.62**
X2Face (32)	56.5	**0.75**	0.18	0.85
Pix2pixHD (32)	**24.0**	0.70	0.16	0.71
Ours (32)	29.5	0.74	**0.19**	**0.61**
VoxCeleb2				
Ours-FF (1)	**46.1**	0.61	**0.42**	**0.43**
Ours-FT (1)	48.5	**0.64**	0.35	0.46
Ours-FF (8)	42.2	0.64	**0.47**	0.40
Ours-FT (8)	**42.2**	**0.68**	0.42	**0.39**
Ours-FF (32)	40.4	0.65	**0.48**	0.38
Ours-FT (32)	**30.6**	**0.72**	0.45	**0.33**

表 1：具有多个少样本学习设置的不同数据集上的方法的定量比较。有关更多详细信息和讨论，请参阅文本。

比前者。 VoxCeleb1 用于与基线和消融研究进行比较，而通过使用 VoxCeleb2，我们展示了我们方法的全部潜力。

指标。对于定量比较，我们对在元学习（或预训练）阶段期间未见过的人物的大小为 T 的少本学习集上的所有模型进行微调。在少样本学习之后，评估在相同序列的保持部分上执行（所谓的自我重演场景）。为了评估，我们从 VoxCeleb 测试集中均匀采样了 50 个视频，并为每个视频均匀采样了 32 个保持帧（微调和保持部分不重叠）。

我们使用多个比较指标来评估生成图像的照片真实性和身份保存性。也就是说，我们使用 Frechet-inception distance（FID）[13]，主要测量感知真实感，结构相似性（SSIM）[39]，测量与完全真实图像的低级相似性，以及编码之间的余弦相似度（CSIM）用于测量身份不匹配的最先进的人脸识别网络[9]的向量（注意，该网络具有与训练期间内容损失计算中使用的 VGGFace 完全不同的架构）。

我们还进行了一项用户研究，以评估人类受访者所看到的结果的感知相似性和真实性。我们向人们展示了从三个不同视频序列中获取的同一个人的图像的三元组。其中两个图像是真实的，一个是假的，由其中一种方法生成，正在进行比较。我们要求用户找到假图像，因为所有这些图像都是同一个人这评估了照片真实性和身份保存性，因为用户可以从两个真实图像中推断出身份（并且即使生成的图像完全真实，也可以发现身份不匹配）。我们使用用户准确度（成功率）作为我们的指标。这里的下限是三分之一的准确性（当用户不能根据非现实性或身份不匹配发现假货并且必须随机猜测时）。通常，我们认为与 FID，SSIM 或 CSIM 相比，此用户驱动度量（USER）可以更好地了解方法的质量。

方法。在 VoxCeleb1 数据集上，我们将我们的模型与另外两个系统进行比较：X2Face [40]和 Pix2pixHD [38]。对于 X2Face，我们使用了作者提供的模型以及预训练的权重（在原始论文中，它也在 Vox-Celeb1 数据集上进行了训练和评估）。对于 Pix2pixHD，我们在整个数据集上从头开始训练模型，与我们的系统进行相同数量的迭代，而不对作者提出的训练方法进行任何更改。我们选择 X2Face 作为基于翘曲的方法的强基线，并选择 Pix2pixHD 作为直接合成方法。

在我们的比较中，我们通过改变在少样本学习中使用的帧数 T 来评估几种情况下的模型。 X2Face 作为一种前馈方法，只需通过训练帧进行初始化，而 Pix2pixHD 和我们的模型在少样本集上进行了 40 个时期的微调。值得注意的是，在比较中，X2Face 使用在完全真实图像上计算的密集对应场来合成生成的对应场，而我们的方法和 Pix2pixHD 使用非常稀疏的标记信息，这可能会给 X2Face 带来不公平的优势。

比较结果。我们在三种不同的设置中与基线进行比较，在微调集中有 1 帧，8 帧和 32 帧。如前所述，测试集由 50 个测试视频序列中的每一个的 32 个保持帧组成。此外，对于每个测试帧，我们从同一个人的其他视频序列中随机采样两帧。这些框架用于伴随有假图片的三元组以供用户学习。

正如我们在表 1-Top 中所看到的，基线在我们的两个相似性度量上始终超出我们的方法。我们认为这是方法本身固有的：X2Face 在优化期间使用 L2 损失[40]，这导致了良好的 SSIM 得分。另一方面，Pix2pixHD 仅仅最大化感知度量，没有身份保存损失，导致 FID 最小化，但具有更明显的身份不匹配问题，如从 CSIM 列中看到的。此外，这些指标与人类感知并不完全相关，因为这两种方法都会产生不可思议的山谷假象，可以从定性比较图 3 和

图 3：VoxCeleb1 数据集的比较。 对于每种比较方法，我们对在元学习或预训练期间未见过的人的视频进行一次和几次学习。 我们将训练帧的数量设置为等于 T（最左边的列）。 其中一个训练框架显示在源列中。 接下来的列显示了从视频序列的测试部分获取的完全真实图像，以及比较方法的生成结果。

用户学习结果中看出。另一方面，余弦相似性与视觉质量相关，但仍然有利于模糊，不太逼真的图像，这也可以通过将表 1-Top 与图 3 中显示的结果进行比较来看出。

虽然在客观指标方面的比较尚无定论，但用户研究（包括 4800 个三元组，每个显示给 5 个用户）清楚地揭示了我们的方法实现的更高的再现性和个性化程度。

我们还对我们的系统进行了消融研究，并对少数几个学习时间进行了比较。两者均在补充材料中提供。

大规模的结果。然后，我们扩展可用数据并在更大的 VoxCeleb2 数据集上训练我们的方法。在这里，我们训练我们方法的两个变体。 FF（前馈）变量训练了 150 个时期而没有嵌入匹配损失 $\mathcal{L}_{\text{MCH}}$，因此，我们只使用它而不进行微调（通过简单地预测自适应参数 ψ' 通过嵌入的投影 $\hat{\mathbf{e}}_{\text{NEW}}$）。FT 变量仅训练一半（75 个时期），但 $\mathcal{L}_{\text{MCH}}$ 允许微调。我们对这两种模型进行了评估，因为它们可以将少样本学习速度与结果质量进行权衡。它们都取得了很高的分数的成就，与在 VoxCeleb1 上训练的小型模型相比。值得注意的是，对于 T = 32 设置中的用户学习准确度，FT 模型达到 0.33 的下限，这是一个完美的分数。我们在图 4 中给出了这两个模型的结果，并且在补充材料和图 1 中给出了更多结果（包括结果，其中动画是由来自同一个人的不同视频的标记驱动的）。

通常，根据比较结果（表格 1-Bottom）和视觉评估来判断，FF 模型对于少样本学习（例如单样本）更好，而 FT 模型对于更大的 T 实现更高的质量通过对抗微调。

木偶操作结果。最后，我们展示了照片和绘画的木偶操作结果。为此，我们根据 VoxCeleb2 数据集的测试视频评估模型，在单样本设置中训练。我们使用 CSIM 指标对这些视频进行排名，在原始图像和生成的图像之间进行计算。这使我们能够找到具有相似标记几何形状的人物并将其用于木偶操纵。结果可以在图 5 和图 1 中看到。

图 4：VoxCeleb2 数据集上我们最佳模型的结果。 训练帧的数量再次等于源列中显示的 T（最左列）和示例训练帧。 下一栏显示了完全真实图像和 Ours-FF 前馈模型的结果，以及 Ours-FT 模型在微调之前和之后的结果。 虽然前馈变体允许对新化身进行快速（实时）少量学习，但微调最终会提供更好的真实性和保真度。

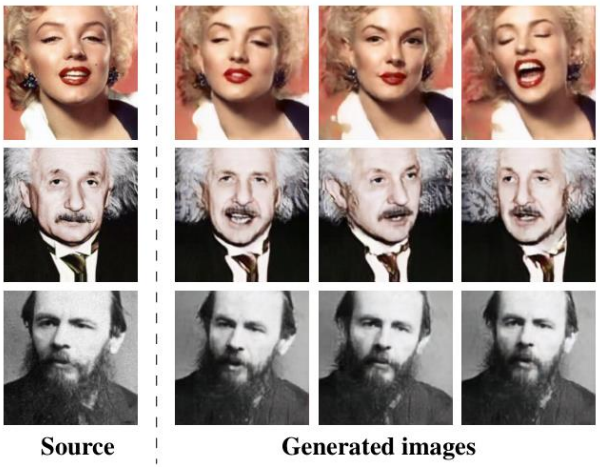

图 5：将静态照片带入生活。 我们展示了从源列中的照片中学习的单样本模型的木偶操作结果。 驱动姿势取自 VoxCeleb2 数据集。 推荐数码变焦。

5. 结论

我们已经提出了一个广泛的生成模型的元学习框架，它能够以深度生成器网络的形式训练高度逼真的虚拟头部说话模型。至关重要的是，创建新模型只需要少量照片（只需一张），而在 32 幅图像上训练的模型在我们的用户研究中获得完美的真实感和个性化评分（对于 224p 静态图像）。

目前，我们的方法的关键限制是模拟表示（特别是，当前的一组标记不代表任何方式的凝视）和缺乏具有里程碑意义的适应。使用来自不同人的标记会导致明显的人物身份不匹配。因此，如果想要创建没有这种不匹配的“假”木偶操纵视频，则需要一些具有里程碑意义的适应性。然而，我们注意到，许多应用程序不需要木偶操纵不同的人，而只需要能够驱动自己的头部说话。对于这种情况，我们的方法已经提供了一种高度真实的解决方案。

参考文献

[1] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, and P. Debevec. The Digital Emily project: Achieving a photorealistic digital actor. IEEE Com-puter Graphics and Applications, 30(4):20–31, 2010. 2

[2] K. S. Andrew Brock, Jeff Donahue. Large scale gan training for high fidelity natural image synthesis. arXiv:1809.11096, 2018. 2, 5

[3] A. Antoniou, A. J. Storkey, and H. Edwards. Augmenting image classifiers using data augmentation generative adver-sarial networks. In Artificial Neural Networks and Machine Learning - ICANN, pages 594–603, 2018. 2

[4] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. Neural voice cloning with a few samples. In Proc. NIPS, pages 10040–10050, 2018. 2

[5] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. ACM Transactions on Graphics (TOG), 36(6):196, 2017. 1

[6] V. Blanz, T. Vetter, et al. A morphable model for the syn-thesis of 3d faces. In Proc. SIGGRAPH, volume 99, pages 187–194, 1999. 2

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 1021–1030, 2017. 3

[8] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In INTERSPEECH, 2018. 5

[9] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 6

[10] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Proc. ICML, pages 1126–1135, 2017. 2

[11] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manip-ulation. In European Conference on Computer Vision, pages 311–326. Springer, 2016. 1

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014. 2

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish-wanathan, and R. Garnett, editors, Advances in Neural In-formation Processing Systems 30, pages 6626–6637. Curran Associates, Inc., 2017. 6

[14] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proc. ICCV, 2017. 2, 5

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32Nd International Conference on In-ternational Conference on Machine Learning - Volume 37, ICML'15, pages 448–456. JMLR.org, 2015. 5

[16] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proc. CVPR, pages 5967–5976, 2017. 2

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Gir-shick, S. Guadarrama, and T. Darrell. Caffe: Convolu-tional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014. 5

[18] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proc. NIPS, pages 4485–4495, 2018. 2

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV, pages 694–711, 2016. 4, 5

[20] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Perez,´ C. Richardt, M. Zollhofer,¨ and C. Theobalt. Deep video portraits. arXiv preprint arXiv:1805.11714, 2018. 2

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014. 5

[22] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. ACM Transactions on Graphics (TOG), 37(4):68, 2018. 2

[23] S. O. Mehdi Mirza. Conditional generative adversarial nets. arXiv:1411.1784. 2

[24] M. Mori. The uncanny valley. Energy, 7(4):33–35, 1970. 1

[25] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agar-wal, J. Fursund, H. Li, R. Roberts, et al. paGAN: real-time avatars using dynamic textures. In SIGGRAPH Asia 2018 Technical Papers, page 258. ACM, 2018. 2

[26] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In INTER-SPEECH, 2017. 5

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In Proc. BMVC, 2015. 4

[28] S. M. Seitz and C. R. Dyer. View morphing. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 21–30. ACM, 1996. 1

[29] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Para-gios, and I. Kokkinos. Deforming autoencoders: Unsuper-vised disentangling of shape and appearance. In The Eu-ropean Conference on Computer Vision (ECCV), September 2018. 1

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Proc. ICLR, 2015. 4

[31] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from au-dio. ACM Transactions on Graphics (TOG), 36(4):95, 2017. 2

[32] M. K. Takeru Miyato. cgans with projection discriminator. arXiv:1802.05637, 2018. 2, 4

[33] M. K. Y. Y. Takeru Miyato, Toshiki Kataoka. Spec-tral normalization for generative adversarial networks. arXiv:1802.05957, 2018. 5

[34] T. A. Tero Karras, Samuli Laine. A style-based gen-erator architecture for generative adversarial networks. arXiv:1812.04948. 2

[35] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reen-actment of RGB videos. In Proceedings of the IEEE Con-ference on Computer Vision and Pattern Recognition, pages 2387–2395, 2016. 2

[36] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. CoRR, abs/1607.08022, 2016. 5

[37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. arXiv preprint arXiv:1808.06601, 2018. 2

[38] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion, 2018. 4, 6

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. Trans. Img. Proc., 13(4):600–612, Apr. 2004. 6

[40] O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In The European Conference on Computer Vision (ECCV), September 2018. 1, 2, 6

[41] C. Yin, J. Tang, Z. Xu, and Y. Wang. Adversarial meta-learning. CoRR, abs/1806.03316, 2018. 2

[42] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. arXiv:1805.08318, 2018. 5

[43] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song. Metagan: An adversarial approach to few-shot learning. In NeurIPS, pages 2371–2380, 2018. 2

A. 补充材料

在补充材料中，我们提供了额外的定性结果以及消融研究和我们的方法与推理和训练基线之间的时间比较。

A.1. 时间比较结果。

在表 2 中，我们提供了三种方法的时间比较。此外，我们在比较中包含了我们方法的前馈变体，该变体仅针对 VoxCeleb2 数据集进行了训练。比较是在单个 NVIDIA P40 GPU 上进行的。对于 Pix2pixHD 和我们的方法，通过对大小为 T 的训练集上的 40 个时期进行微调来完成少样本学习。对于大于 1 的 T，我们在 8 个图像的批次上训练模型。每次测量平均超过 100 次迭代。

我们看到，给定足够的训练数据，我们的前馈变量方法可以在很少的训练时间内大大超过所有其他方法，同时在很高的水平上保持个性化保真度和输出的真实性（如图 4 所示）。但是为了在质量方面获得最佳效果，必须进行微调，在 P40 GPU 上需要大约四分半钟才能完成 32 个训练图像。可以根据具体情况或通过我们未执行的训练调度器的介绍来进一步优化迭代周期数以及因此的微调速度。

另一方面，我们的方法的推理速度与其他方法相当或更慢，这是由我们需要编码关于头部说话的先验知识的大量参数引起的。虽然，通过使用更现代的 GPU 可以大大改善这个数字（在 NVIDIA 2080 Ti 上，推理时间可以减少到每帧 13ms，这对于大多数实时应用来说已经足够了）。

Method (T)	Time, s
Few-shot learning	
X2Face (1)	0.236
Pix2pixHD (1)	33.92
Ours (1)	43.84
Ours-FF (1)	**0.061**
X2Face (8)	1.176
Pix2pixHD (8)	52.40
Ours (8)	85.48
Ours-FF (8)	**0.138**
X2Face (32)	7.542
Pix2pixHD (32)	122.6
Ours (32)	258.0
Ours-FF (32)	**0.221**
Inference	
X2Face	0.110
Pix2pixHD	**0.034**
Ours	0.139

表 2：三种模型的少样本学习和推理时间的定量比较。

A.2. 消融研究

在本节中，我们评估与我们在模型训练中使用的损失相关的贡献，以及激励训练程序。我们已经在图 4 中显示了微调对结果质量的影响，因此我们不在此进行评估。相反，我们专注于微调的细节。

我们问的第一个问题是关于通过嵌入器初始化人物身份标识的参数的重要性。我们尝试了嵌入向量 $\hat{\mathbf{e}}_{\text{NEW}}$ 和生成器的自适应参数 $\hat{\psi}$ 的不同类型的随机初始化，但是这些实验在微调之后没有产生任何合理的图像。因此，我们意识到嵌入器提供的生成器的人物身份标识的初始化对于微调问题的收敛是重要的。

然后，我们评估了判别器的人物身份标识参数的初始化的贡献。我们从目标中删除 $\mathcal{L}_{\text{MCH}}$ 项并执行元学习。在我们的最终方法中使用多个训练帧来解决少样本学习问题会导致优化不稳定，因此我们使用了单样本元学习配置，结果证明是稳定的。在元学习之后，我们随机初始化判别器的人物特定向量 $\mathbf{W}_i$。结果可以在图 6 中看到。我们发现随机初始化的结果似乎是合理的，但在现实性和个性化保真度方面引入了明显的差距。因此，我们得出结论，判别器的人物身份标识参数的初始化也有助于结果的质量，尽管其方式不如生成器的初始化。

最后，我们评估了在微调期间对抗项 $\mathcal{L}'_{\text{ADV}}$ 的贡献。因此，我们将其从微调目标中删除，并将结果与我们的最佳模型进行比较（参见图 6）。虽然这些变体之间的差异非常微妙，但我们注意到对抗性的微调会产生更清晰的图像，从而在姿势和图像细节方面更好地匹配基本事实。选择图 6 中的图像以突出这些差异。

A.3. 其他定性结果

图 7 中显示了一次性学习肖像和照片的更多木偶操作结果。我们还显示了从图 9 中的自拍中学习的头部说话的结果。其余图中提供了这些方法之间的其他比较。

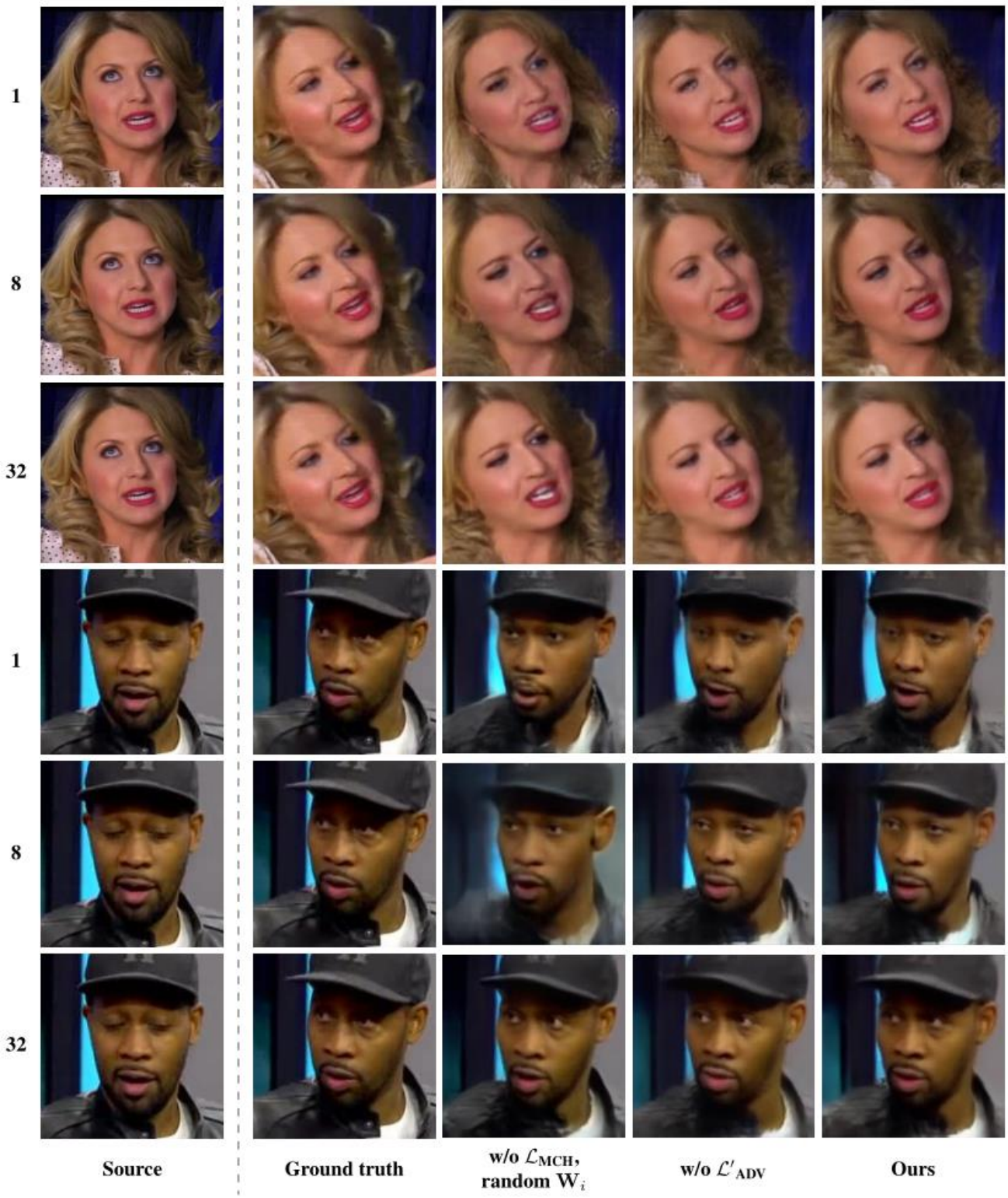

图 6：我们贡献的消融研究。 训练帧的数量再次等于 T（最左边的列），源列中显示的示例训练帧和下一列显示完全真实图像。 然后，我们从元学习目标中移除 $\mathcal{L}_{\text{MCH}}$ 并随机初始化判别器的嵌入向量（第三列）并评估对抗性微调的贡献与目标中没有 $\mathcal{L}'_{\text{ADV}}$ 的常规微调相比（第五列）。 最后一列代表我们最终模型的结果。

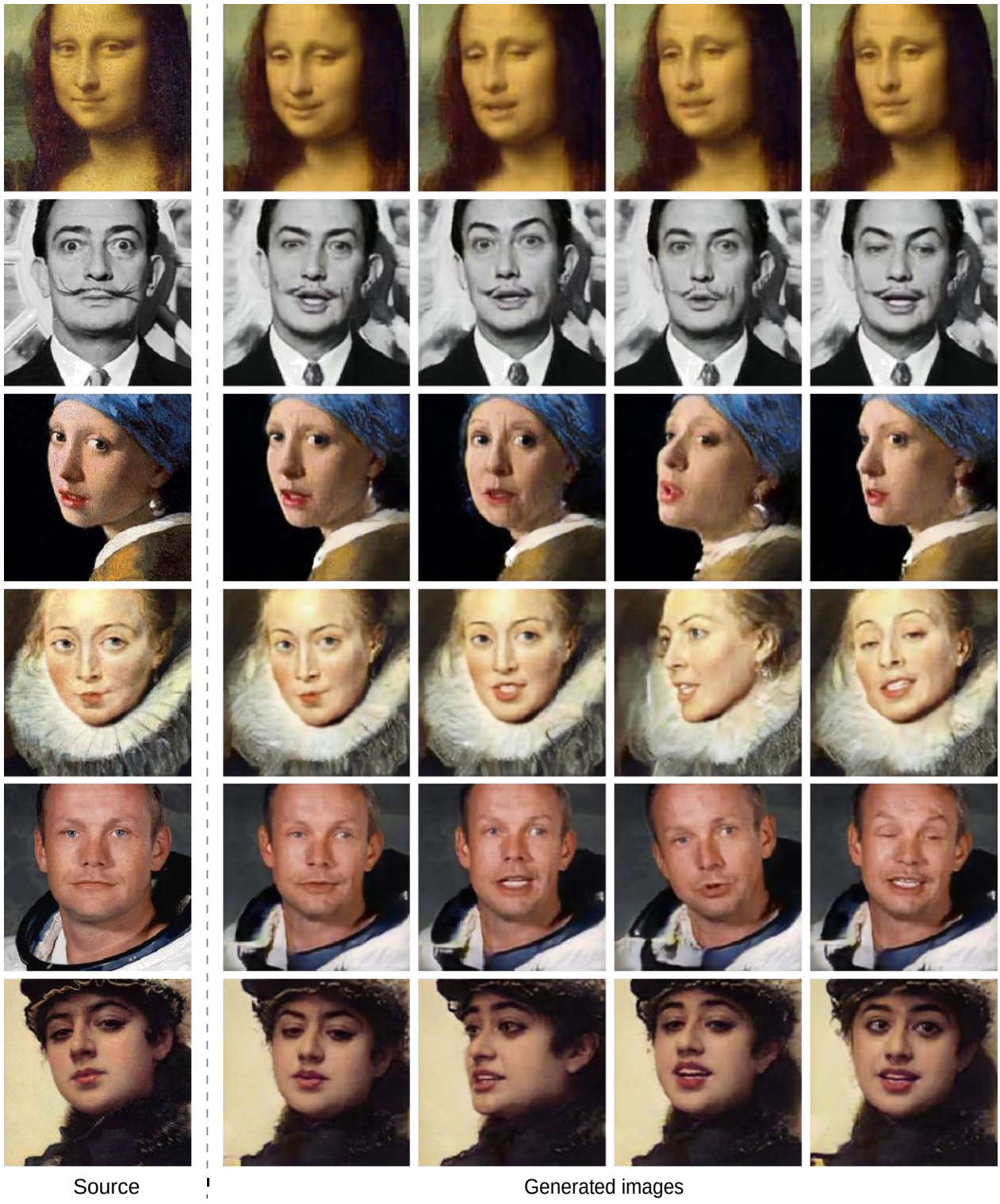

图 7：在单样本设置中训练的谈话头部模型的更多木偶操作结果。 用于一次性训练问题的图像位于源列中。 接下来的列显示生成的图像，这些图像以不同人的视频序列为条件。

图 8：以八样本设置训练的谈话头部模型的结果。 示例训练框架位于源列中。 接下来的列显示生成的图像，这些图像以从同一个人的不同视频序列拍摄的姿势轨迹为条件。

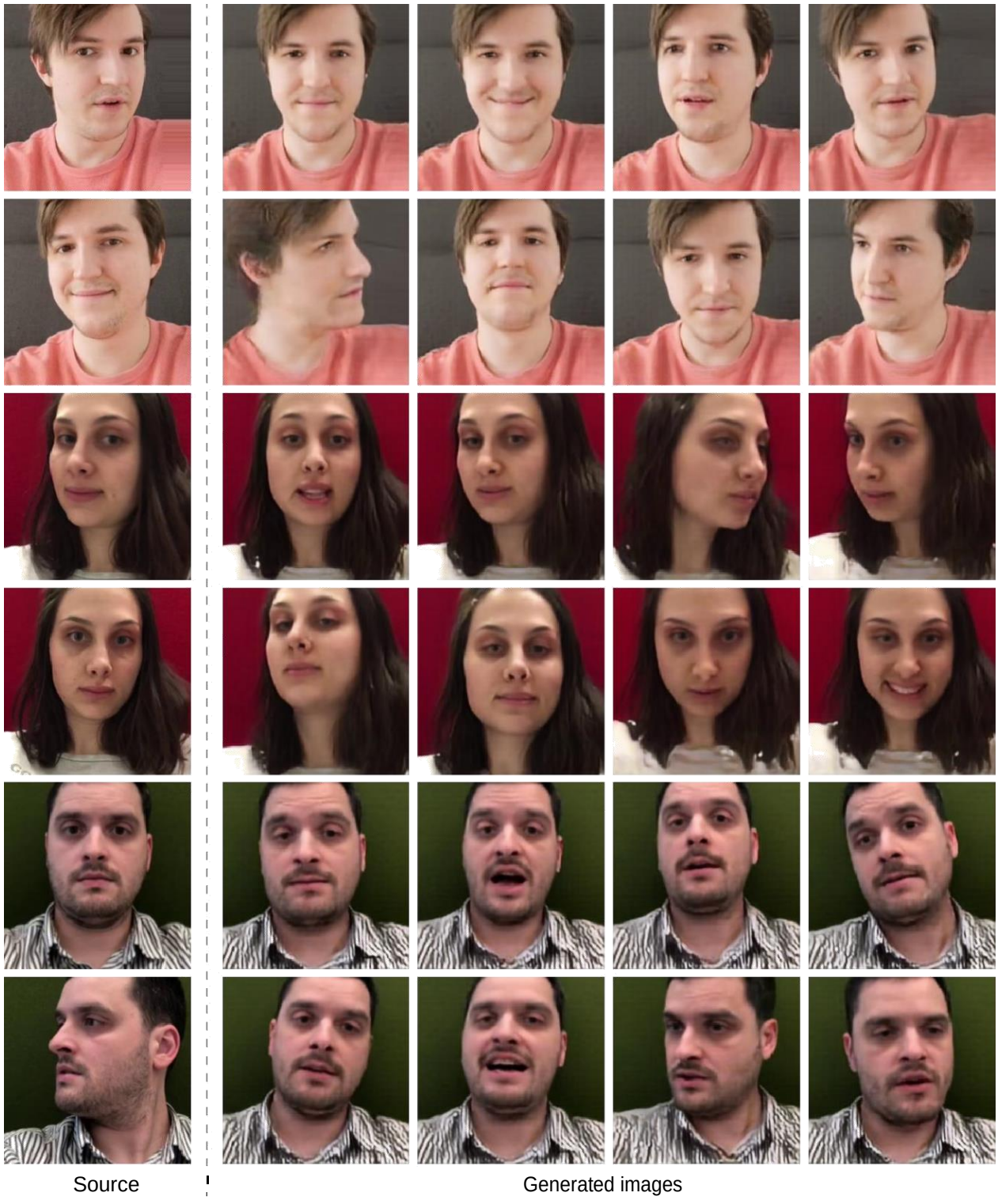

图 9：在自拍照片上以 16 样本设置训练的谈话头部模型的结果，其中驱动标记取自同一人的不同视频。 示例训练框架显示在源列中。接下来的列显示生成的图像，这些图像以同一个人的不同视频序列为条件。

图 10：VoxCeleb1 数据集的第一次扩展定性比较。 这里，对每种方法的定性性能和训练数据量影响结果的方式进行比较。 列的符号如主卷中的图 3 所示。

图 11：VoxCeleb1 数据集的第二次扩展定性比较。 在这里，我们比较了三种方法在元学习或预训练中未见过的不同人的定性表现。我们使用了八样本学习问题公式。 列的符号如主卷中的图 3 所示。

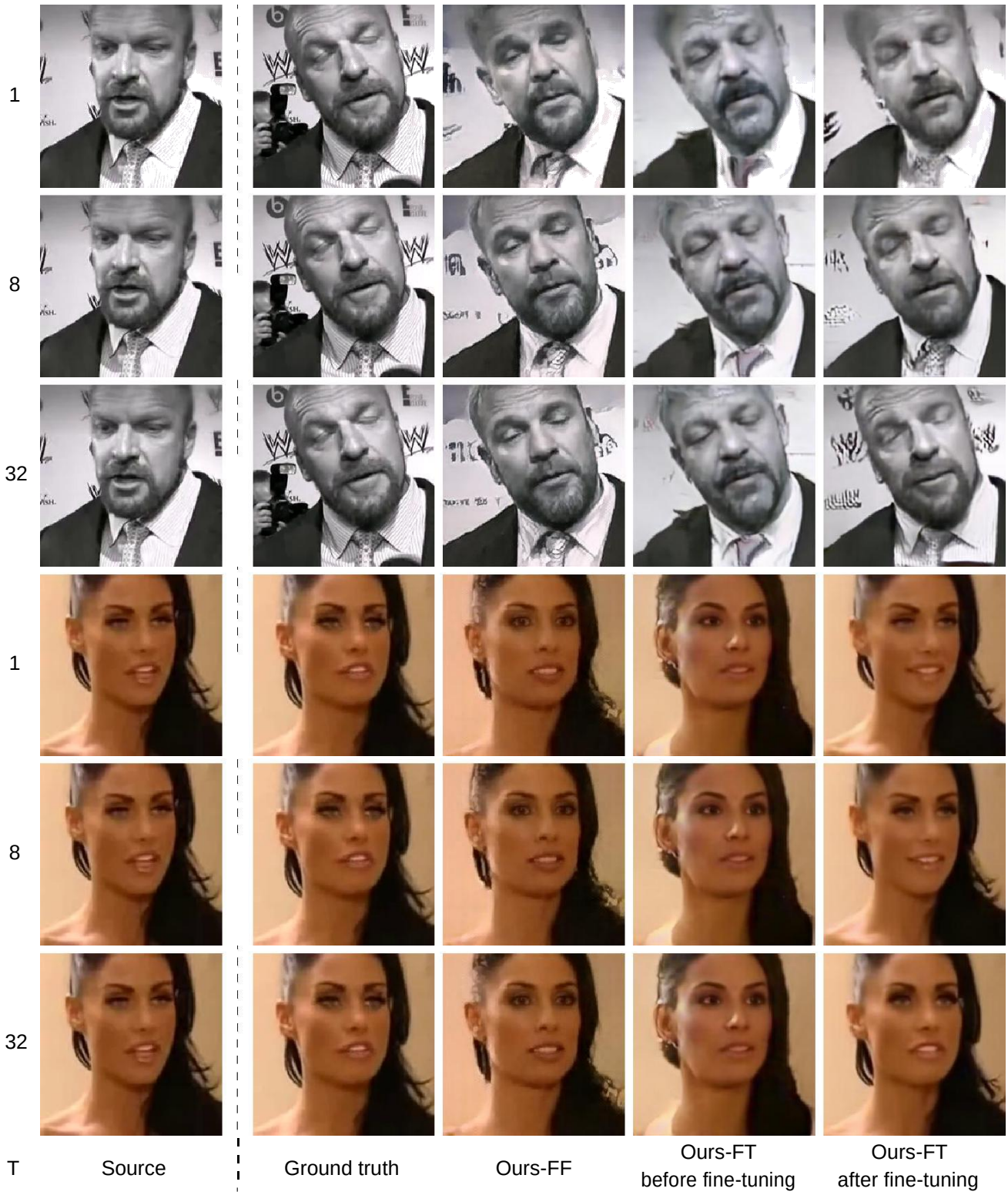

图 12：VoxCeleb2 数据集的第一次扩展定性比较。 这里，对我们方法的每个变体的定性性能和训练数据量对结果的影响方式进行比较。 列的表示法如主要论文中的图 4 所示。

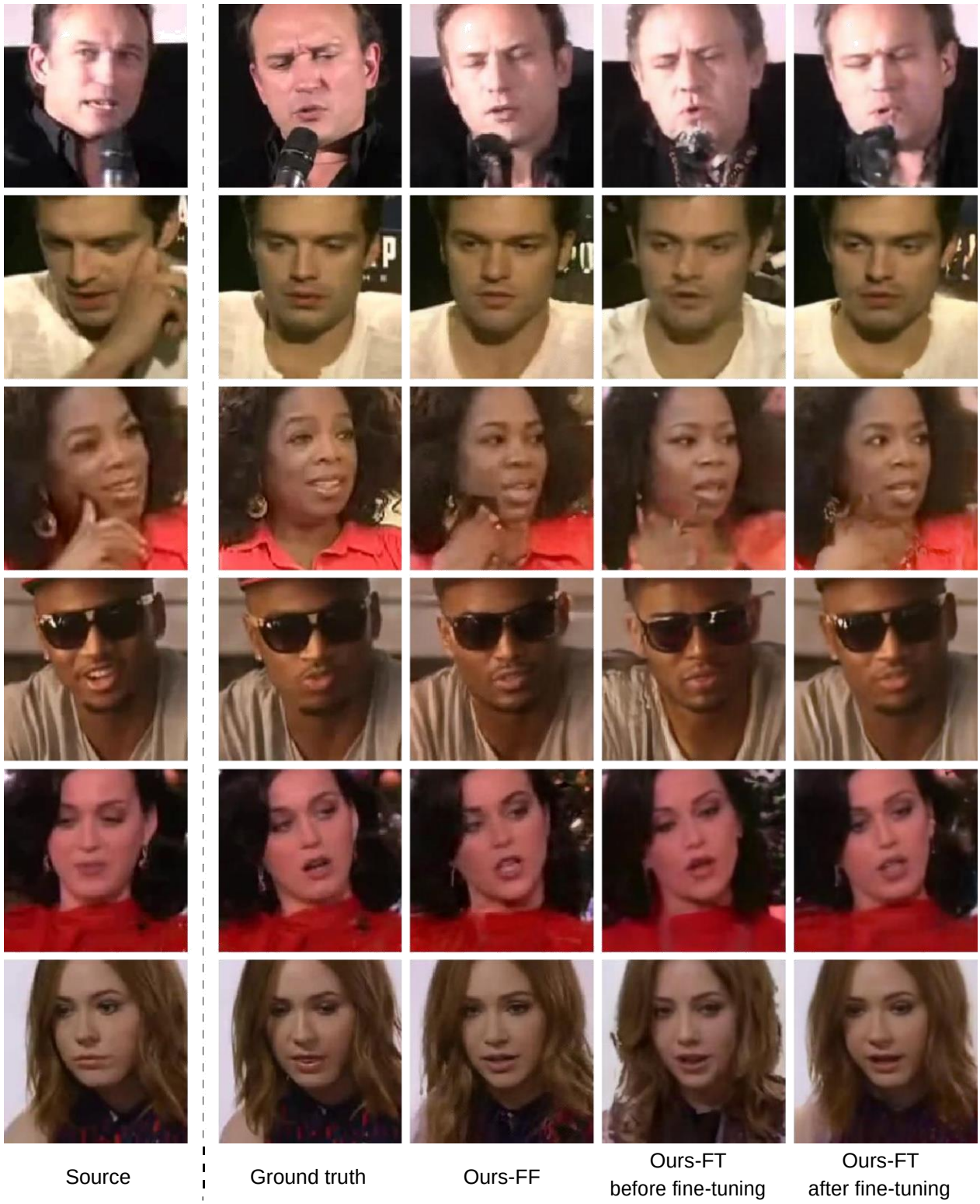

图 13：VoxCeleb2 数据集的第二次扩展定性比较。 在这里，我们比较了我们的方法的三个变体在元学习或预训练中没有看到的不同人的定性表现。 我们使用了八样本学习问题公式。 列的表示法如主要论文中的图 4 所示。