arXiv:2303.02659v1 [cs.CR] 5 Mar 2023

深度伪造免疫的网络疫苗

Ching-Chun Chang、Huy Hong Nguyen、Junichi Yamagishi、Isao Echizen

摘要—深度伪造对网络安全构成不断演变的威胁，因此亟需开发自动化对策。尽管大量取证研究致力于深度伪造的检测与定位，但将伪造内容还原为真实内容的解决方案尚未形成。在本研究中，我们提出通过网络免疫赋予抵御深度伪造的免疫力。类似于生物疫苗接种通过注入抗原，在实际病原体感染之前诱导免疫力，网络免疫通过模拟深度伪造并执行对抗训练，构建防御性免疫系统。为了以有限的计算资源构建与攻击无关的免疫力，我们提出使用一种单一的强力攻击——人脸遮挡——来模拟各种深度伪造。所提出的免疫系统由用于诱导免疫力的接种器和用于恢复面部内容的中和器组成。实验评估表明，该系统能够有效抵御人脸替换、人脸重演以及各种类型的失真。

索引词—**网络安全、深度伪造、取证、免疫力**

I. 引言

DEEPFAKES 作为一种新兴的网络安全威胁，利用人工智能和机器学习创建合成媒体。这项技术是一把双刃剑，既能带来无害的娱乐，也会对经济、政治和社会造成潜在的危害，包括但不限于市场操纵、选举影响、未经同意的色情内容、诽谤性指控、伪造证据和身份欺诈。俗话说“眼见为实”，这种自然倾向助长了深度伪造所传播的虚假信息。深度伪造的广泛应用和快速发展，带来了持续演变的挑战：如何制定对策以辨别事实与虚构 [1]。

如今，“深度伪造”一词已被泛化，用来指代种类繁多的合成媒体；我们则聚焦于其中引发公众高度关注的一大类——面部操纵 [2]。通常，防御操纵涉及三个基本的（循序渐进）任务：检测、定位和恢复，如图 1 所示。计算取证在检测给定图像是否被深度伪造算法篡改 [3]–[8]，并进一步定位遭篡改的部分 [9]–[12]方面取得了丰硕成果。然而，恢复欺诈性内容的解决方案尚未开发出来。与可依赖被动诊断的检测和定位不同，可靠的恢复往往需要主动防护。在数字通信中，例如，

本研究部分得到了日本学术振兴会（JSPS）科研费补助金（JP16H06302、JP18H04120、JP20K23355、JP21H04907 和 JP21K18023）以及日本科学技术振兴机构（JST）CREST 资助项目（JPMJCR18A6 和 JP-MJCR20D3）的支持。C.-C. Chang、H. H. Nguyen、J. Yamagishi 和 I. Echizen 均任职于日本东京的国立情报学研究所（电子邮件：ccchang@nii.ac.jp；nhhuy@nii.ac.jp；jyamagis@nii.ac.jp；iechizen@nii.ac.jp）。数字对象标识符 xx.xxxx/2023



检测 定位 恢复

图 1：网络取证的目标：检测、定位、恢复。

纠错码用于在消息通过噪声信道传输之前对其进行保护。接种则是另一种类比：它将抗原注入身体，引发免疫反应，从而在实际病原体造成感染之前，形成保护性免疫力。

对抗训练类似于病原体特异性疫苗的概念，是一种直观的应对安全漏洞的方法 [13]–[15]。这种方法通过在模拟对手存在的情况下训练模型来实现，使计算模型能够像生物免疫系统一样，学会防御可能的攻击。尽管对抗训练可能是一种解决方案，但它存在若干固有局限。首先，准备和执行多种深度伪造算法，可能会耗费大量（计算）时间和资源。此外，深度伪造作为一种网络病毒，会随时间演变；在实践中，几乎不可能考虑到每一种可能性。这就像针对某一种特定病原体的疫苗，通常无法抵御另一种病原体，除非两者非常相似。因此，具备防御各种（甚至是未预见到的）攻击的能力至关重要，这有时被称为攻击无关性。

本文研究一种用于恢复深度伪造的攻击无关方法。所提出的系统由用于诱导免疫力的接种器、用于恢复面部内容的中和器、用于区分已接种疫苗和未接种疫苗媒体的验证器，以及用于模拟深度伪造的对手组成。我们不试图穷尽所有可能的深度伪造，而是考虑一种单一且强大的对手模型——掩码人脸模型，旨在以有限的计算资源建立攻击无关能力。

II. 预备知识

“深度伪造”一词是“深度学习”和“伪造媒体”的合成词，据认为源自一位名为‘deepfakes’的匿名用户，他曾在社交媒体平台上发布换脸视频。人脸替换是深度伪造最早的示例，它使用自动编码器模型生成逼真的换脸视频。该模型由一个共享编码器和两个

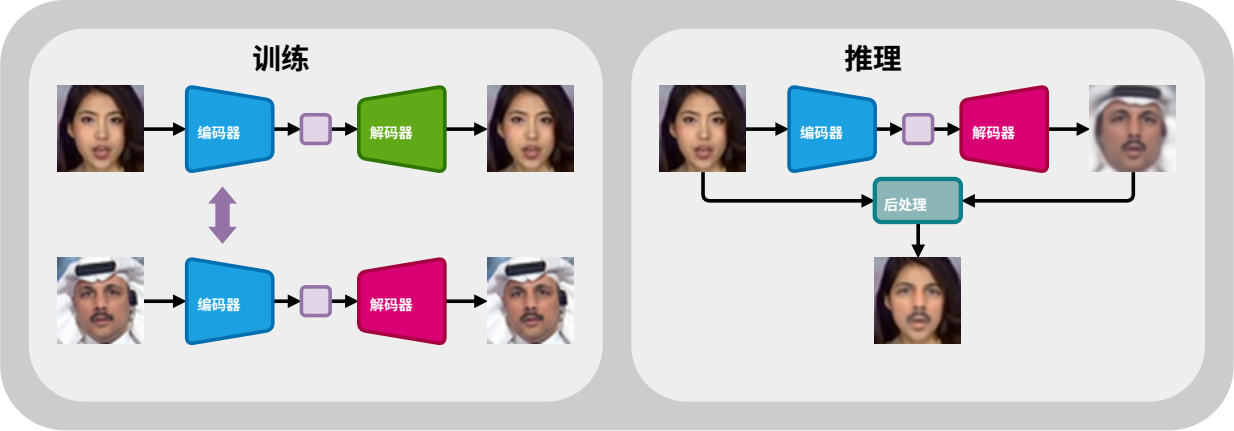


图 2： 用于人脸替换的深度伪造自编码器。

解码器组成，分别对应两个身份（源身份和目标身份），如图 2 所示。自动编码器是一类神经网络模型，它使用编码器降低数据维度，或将数据投射到紧凑的潜在空间中，再通过解码器从潜在特征重建数据。这一过程模拟了将信息挤过瓶颈，从而保留对预测有用的信息。换脸自动编码器的一个启发式构想是：共享编码器学习提取特征，例如面部表情和姿势等变化特征；而每个解码器则学习利用这些特征以及对应身份的不变特征来重建视频帧。训练完成后，将目标帧输入共享编码器，并使用源身份的解码器重建视频帧，即可生成合成源帧。随后，将合成源帧中的人脸区域与目标帧中的非人脸区域融合，生成换脸视频。

深度伪造的另一种形式称为面部重演，它将一个人的身份变成一具木偶。顾名思义，这种技术会操纵目标人物的面部动作，例如表情 [16]–[18], 视线 [19]–[21], 姿势 [22]–[26] 以及身体 [27]–[31]。口部重演也称为唇形同步，会使目标人物的唇部动作与语音音轨相匹配 [32]–[35]。面部编辑或修饰也是一种常见的深度伪造操纵形式，通常出于娱乐目的改变目标人物的外貌 [36]–[38]。

III. 方法论

我们首先介绍网络免疫的基本概念模型，并讨论自然解决方案的局限性。随后，我们提出相应的解决方案，以及构建免疫系统的流程。

A. 深度伪造采样

网络免疫可以视为一种通信形式。通信系统通常由编码器（位于源端）和解码器（位于目的端）组成 [39]。其目标是借助一对编码器/解码器，通过噪声信道将消息从源端准确传输到目的端。由于通信系统使用一对编码器和解码器进行纠错，网络接种系统则使用一对接种器和中和器进行操纵逆转。可以在中间加入攻击模型，联合训练一对神经网络。在训练过程中随机采样攻击模型，可能使系统能够适应各种有害条件。不过，这种方法也面临若干可能限制其实际应用的挑战。

从工程角度来看，构建一个包含多种攻击模型、能够实时纳入训练过程的通用深度伪造工具包十分困难。通用工具包面临的阻碍包括但不限于：不同的输入要求（例如照片尺寸、人脸位置、肖像构图、辅助数据）、不同程度的泛化能力（即身份特定或身份无关），以及不同的预处理／后处理流程。此外，不同的攻击模型会导致截然不同程度和形式的失真（例如人脸重演和人脸替换），因此，训练损失能否在合理的时间范围内收敛仍值得商榷。更进一步而言，要证明实验室疫苗能够可靠地迁移，从而抵御野外病毒变体，是一项极具挑战性的工作。

B. 人脸遮挡

为克服这些问题，我们既不局限于数量有限的深度伪造算法，也不试图穷尽所有可能性，而是尝试考虑一种单一的攻击模型。该模型应当与身份无关，并且能够实时执行。最重要的是，应对这一极具破坏性的攻击，就意味着能够应对各种深度伪造攻击。几乎所有深度伪造操纵都有一个共同点：人脸区域在不同程度上变得不再可信。作为一种极端情况，人脸遮挡可以成为满足上述所有目的的理想攻击模型。

如果我们遮挡人脸区域，并尝试根据其余上下文信息对其进行重建，这就成为一个图像修复问题 [40]–[42]。尽管我们可以期待缺失部分得到合理且逼真的重建，但在大多数情况下，原始内容都无法被绝对确定地恢复

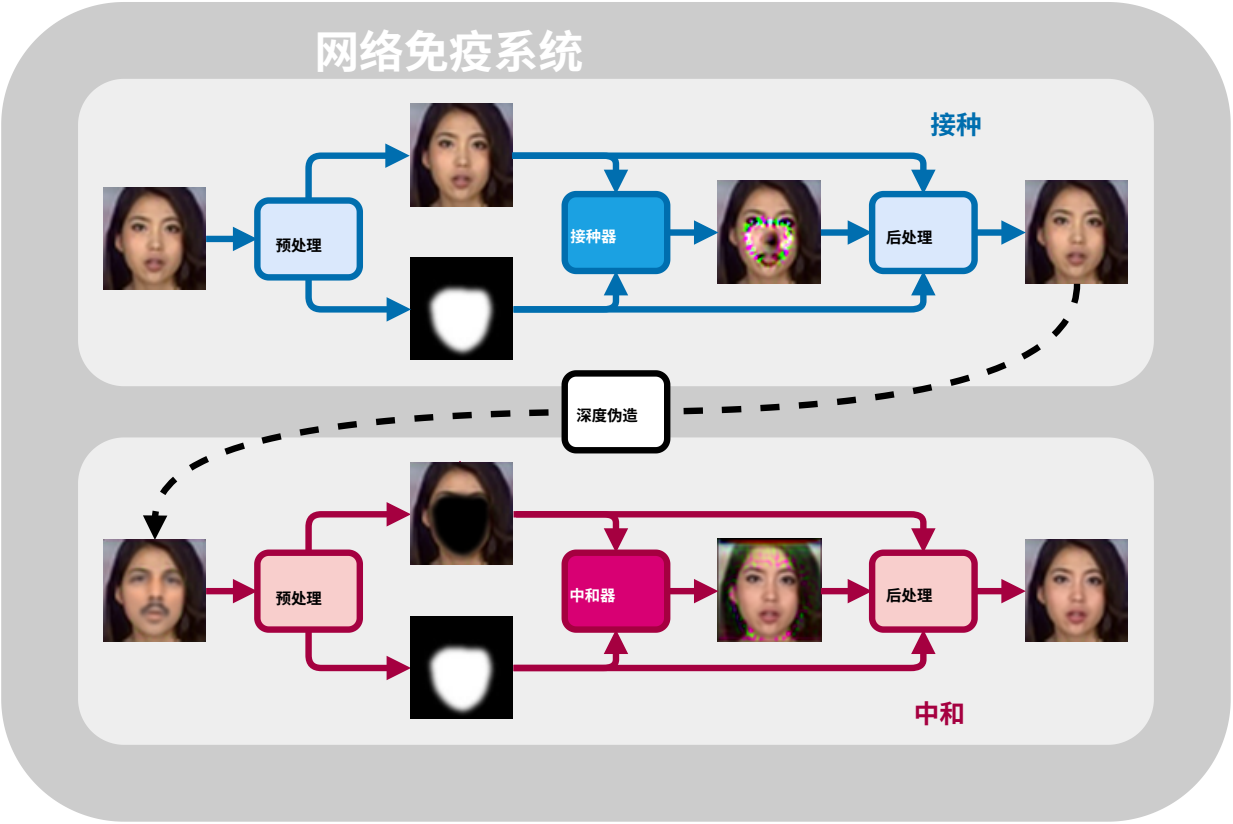


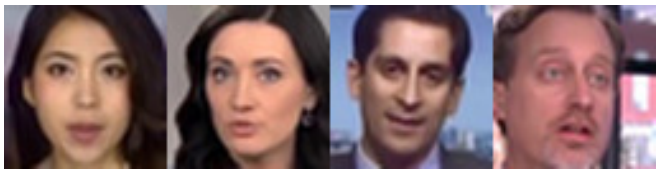
图3： 用于深度伪造恢复的网络免疫系统，包含一对接种器和中和器。

。如果允许在人脸遮挡前进行不可察觉的修改，我们便可以使用隐写算法，将有关人脸区域的重建信息嵌入非人脸区域。不过，这种隐写方案（也称为自嵌入）在实际应用时，通常需要对参数进行相当复杂的手动调整 [43]–[48]。考虑到嵌入失真约束会限制隐写容量，为了将嵌入的数据量控制在容量上限以内，可以使用信源编码压缩数据，但这会涉及编码效率与重建质量之间的权衡。此外，数字图像可能会经历多种变换，例如模糊度、亮度、对比度、饱和度、色调等。由于隐写技术较为脆弱，可以采用信道编码纠正错误，以确保可靠地提取数据，但这会涉及编码冗余与纠错能力之间的权衡。此外，某些依赖内容的算法参数可能需要针对每张图像分别优化。尽管已有基于学习的隐写方法，旨在以自动且稳健的方式嵌入消息 [49]–[51]，但通常假设消息是随机二进制数字序列或秘密图像，而且无法指定隐藏位置。在我们的场景中，消息就是肖像图像本身的人脸区域，而隐藏位置则是非人脸区域。如果编码器端检测到的人脸区域与解码器端检测到的区域不一致，也可能导致同步问题。

C. 网络免疫系统

机器学习开辟了一条道路，使参数配置基本摆脱人工干预。设想肖像图像通过人脸掩蔽信道，在接种器和中和器之间传输。接种器和中和器都是神经网络，其目标是在确保中和后图像质量的同时，保留免疫图像的质量。此外，还需要能够区分已接种疫苗的对象和未接种疫苗的对象。在实际应用中，具备抵御常见图像处理操作的鲁棒性也颇具价值。如图3所示，接种和中和均由预处理、中间处理和后处理步骤组成。

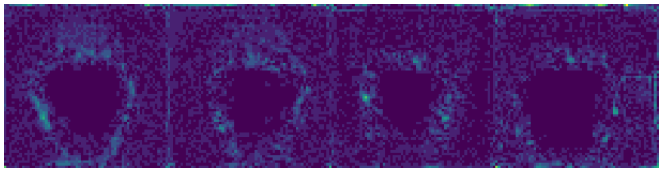
接种： 给定视频片段或静态图像，初步步骤是检测人脸区域（使用任何可用的人脸检测算法），裁剪出肖像（对每个视频帧）并调整其尺寸以符合输入分辨率要求（例如 64×64 像素）。为简化起见，不考虑单张图像中存在多张人脸的情况。接种的预处理步骤通过使用现成的人脸对齐算法检测面部关键点，并为人脸轮廓内外的像素分配二进制数字，从而生成掩码。中间处理步骤将肖像和掩码一并输入接种器，并获得原始输出。后处理步骤根据掩码，将原始输出中的人脸区域替换为原始肖像中的对应区域，从而生成免疫图像。从图4可以看出，接种



未接种样本



已接种样本



残差

图4： 未接种与已接种图像样本之间残差的可视化。

会在非人脸区域引入不可察觉的扰动。

中和：中和过程的预处理步骤会生成一个掩码，并用它遮盖给定图像中的人脸区域，该图像可以是已接种疫苗的，也可以是未接种疫苗的。在实际推理阶段，给定图像中的人脸区域可能会被任意深度伪造算法操纵，从而导致中和过程中面部关键点出现轻微错位。在训练阶段，无需使用深度伪造算法，因为可以对接种过程中生成的掩码应用随机仿射变换来模拟这种错位。我们还应用随机颜色变换，以增强对常见图像失真的鲁棒性。中间处理步骤将掩码图像及其掩码输入中和器，并生成原始输出。后处理步骤将原始输出中的人脸区域合并到掩码图像中，从而得到已中和图像。

验证与损失函数：损失函数对于训练神经网络至关重要。对于网络免疫，我们的目标是实现不可感知性、可逆性和可验证性。我们通过比较免疫图像与原始图像的非人脸区域的相似度来评估不可感知性，并通过比较已中和图像与原始图像的人脸区域的相似度来评估可逆性。有多种可行的方法可以使系统具备可验证性，其中一种可行的做法是让中和器对未接种输入不作响应。换言之，未接种疫苗时，输出应保持为掩码肖像。这样，既可以轻松地用肉眼识别，也可以通过数据驱动的视觉分类器自动识别。令 $\boldsymbol{x}_\circ$ 和 $\boldsymbol{x}_\bullet$ 分别表示未接种疫苗的图像和已接种疫苗的图像，$\boldsymbol{y}_\circ$ 和 $\boldsymbol{y}_\bullet$ 分别表示它们对应的中和图像。此外，令 $\boldsymbol{m}$ 表示一个掩码，其中人脸区域赋值为1，非

算法1 训练（网络免疫系统）

输入： $\boldsymbol{x}_\circ \in \mathcal{D}$ ▷来自数据集的图像
输出： [接种器，中和器]

▷ 接种
$\boldsymbol{m} =$ 掩码检测器$(\boldsymbol{x}_\circ)$ ▷ 预处理
$\boldsymbol{x}_\bullet^{\text{raw}} =$ 接种器$(\boldsymbol{x}_\circ, \boldsymbol{m})$ ▷ 中间处理
$\boldsymbol{x}_\bullet = \boldsymbol{x}_\bullet^{\text{raw}} \cdot \boldsymbol{m} + \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}$ ▷ 后处理
$\mathcal{L}_{\text{imp}} =$ 距离$(\boldsymbol{x}_\bullet, \boldsymbol{x}_\circ)$ ▷ 损失

▷ 中和
$\boldsymbol{m}^{\text{rnd}} =$随机仿射变换$(\boldsymbol{m})$ ▷ 预处理
$\boldsymbol{x}_\bullet^{\text{rnd}} =$随机变换$(\boldsymbol{x}_\bullet)$ ▷ 数据增强
▷ 已接种疫苗病例
$\boldsymbol{y}_\bullet^{\text{raw}} =$ 中和器$(\boldsymbol{x}_\bullet^{\text{rnd}},\ \boldsymbol{m}^{\text{rnd}})$ ▷ 中间处理
$\boldsymbol{y}_\bullet = \boldsymbol{y}_\bullet^{\text{raw}} \cdot \boldsymbol{m}^{\text{rnd}} + \boldsymbol{x}_\bullet \cdot \bar{\boldsymbol{m}}^{\text{rnd}}$ ▷ 后处理
$\mathcal{L}_{\text{rev}} =$ 距离$(\boldsymbol{y}_\bullet,\ \boldsymbol{x}_\circ)$ ▷ 损失
▷未接种疫苗病例
$\boldsymbol{y}_\circ^{\text{raw}} =$ 中和器$(\boldsymbol{x}_\circ,\ \boldsymbol{m}^{\text{rnd}})$ ▷ 中间处理
$\boldsymbol{y}_\circ = \boldsymbol{y}_\circ^{\text{raw}} \cdot \boldsymbol{m}^{\text{rnd}} + \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}^{\text{rnd}}$ ▷ 后处理
$\mathcal{L}_{\text{val}} =$ 距离$(\boldsymbol{y}_\circ,\ \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}^{\text{rnd}})$ ▷ 损失

▷ 反向传播
$\mathcal{L} = \mathcal{L}_{\text{imp}} + \mathcal{L}_{\text{rev}} + \mathcal{L}_{\text{val}}$ ▷ 损失
反向传播$(\mathcal{L},$ [接种器，中和器]$)$ ▷ 更新

算法 2 推理（接种）

输入： $\boldsymbol{x}_\circ$ ▷图像（未接种疫苗的）
输出： $\boldsymbol{x}_\bullet$ ▷图像（已接种疫苗的）
$\boldsymbol{m} =$ 掩码检测器$(\boldsymbol{x}_\circ)$ ▷ 预处理
$\boldsymbol{x}_\bullet^{\text{raw}} =$ 接种器$(\boldsymbol{x}_\circ, \boldsymbol{m})$ ▷ 中间处理
$\boldsymbol{x}_\bullet = \boldsymbol{x}_\bullet^{\text{raw}} \cdot \boldsymbol{m} + \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}$ ▷ 后处理

算法 3 推理（中和）

输入： $\boldsymbol{x}$ ▷图像（遭受攻击/未遭受攻击；已接种疫苗/未接种疫苗）
输出： $\boldsymbol{y}$ ▷图像（已中和）
$\boldsymbol{m} =$ 掩码检测器$(\boldsymbol{x})$ ▷ 预处理
$\boldsymbol{y}^{\text{raw}} =$ 中和器$(\boldsymbol{x} \cdot \bar{\boldsymbol{m}},\ \boldsymbol{m})$ ▷ 中间处理
$\boldsymbol{y} = \boldsymbol{y}^{\text{raw}} \cdot \boldsymbol{m} + \boldsymbol{x} \cdot \bar{\boldsymbol{m}}$ ▷ 后处理

人脸区域为零。其反向掩码记为 $\bar{\boldsymbol{m}}$，其中二元赋值与之相反。请注意，实际应用中，掩码不一定由二进制数字组成，也可以由带有柔和边缘的实数构成，以提供更平滑的融合效果。用于优化接种器和中和器的损失函数是三个损失项的（加权）总和：

$$\mathcal{L} = \mathcal{L}_{\text{imp}}(\boldsymbol{x}_\bullet, \boldsymbol{x}_\circ) + \mathcal{L}_{\text{rev}}(\boldsymbol{y}_\bullet, \boldsymbol{x}_\circ) + \mathcal{L}_{\text{val}}(\boldsymbol{y}_\circ, \boldsymbol{x}_\circ \cdot \bar{\boldsymbol{m}}). \quad (1)$$

在我们的实现中，每个损失项由平均绝对误差（MAE）和结构相似性指数



图5： 比较 接种与未接种情况下已中和图像样本之间的差异。数值表示潜在空间余弦相似度 ity.

（SSIM）组成。训练和推理（接种和中和）的算法流程分别如算法1、2和3所示。为了训练一个自动验证器来区分已接种疫苗的图像和未接种疫苗的图像，我们使用中和后的图像作为输入，并将二元交叉熵作为损失函数。

IV. 评估

我们从不可感知性、可逆性、鲁棒性和可验证性等方面评估所提出的网络接种系统。我们还开展了案例研究，以展示深度伪造免疫力，并讨论潜在的局限性。

A. 实验设置

这些模型在 FaceForensics++数据集上进行训练和测试，该数据集包含一千个带有可追踪人脸的视频片段 [52]。对于接种器和中和器，我们均采用 U-Net 作为骨干架构 [53] ，并使用 OpenAI 推出的先进版本，该版本采用残差连接和多头注意力机制 [54]，，已广泛用于各类计算机视觉任务中的扩散概率模型 [55]–[57]。我们将残差块数量设为 3，并将注意力分辨率设为 4、8 和 16。对于验证器，我们测试了从早期开创性架构到当代架构的多个代表性分类模型，包括 LeCun 等人的多层感知机（MLP） [58],神经网络（LeNet） [59]，、残差神经网络（ ResNet） [60]，、视觉变换器（ViT） [61]以及新一代卷积神经网络（ConvNeXT） [62]。我们通过依赖掩码和不依赖掩码的深度伪造，测试对人脸替换的免疫力。对于前者，我们采用原始深度伪造方法，并训练多个自编码器对，以实现不同身份之间的互换。对于后者，我们使用经过预训练的身份无关 SimSwap 模型 [63]。我们还使用经过预训练的 X2Face 模型 [64]，测试对人脸重演的免疫力。

B. 网络疫苗的消融研究

我们通过与不含接种器的变体系统进行比较，评估网络疫苗所产生的效果。换言之，该变体系统仅由一个用于填充遮挡的人脸区域的中和器组成，其方式类似于图像修复。为比较接种与未接种情况下的中和结果，我们测量身份相似度。具体而言，我们使用 FaceNet 将图像投影到潜在空间中，并计算潜在空间向量之间的余弦相似度 [65]。从图5可以看出，尽管基于图像修复的方法能够重建出逼真的人像，但其结果在视觉上与原始样本存在差异，且身份相似度远低于基于接种的方法。对于200个测试样本，使用疫苗时的平均身份相似度为0.99，而不使用疫苗时为0.57，这表明接种通常能够带来较高的身份相似度。

C. 不可感知性与可逆性

我们要求接种器将失真控制在不可感知的程度，并要求中和器以高保真度重建原始内容。不可感知性和可逆性分别指与原始图像相比，免疫图像和中和图像的视觉质量。峰值信噪比（PSNR）是图像质量的一种常见客观评估指标。图6展示了200个视频样本的PSNR值，每个值均为所有帧的平均值。在每个颜色通道为8位的条件下，有损图像/视频压缩的典型PSNR值介于30至50分贝（dB）之间。免疫视频和

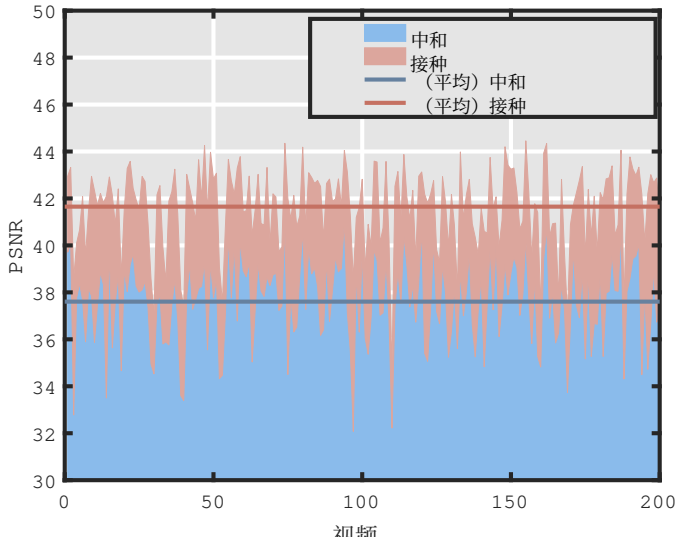


图 6： 不可感知性与可逆性评估。

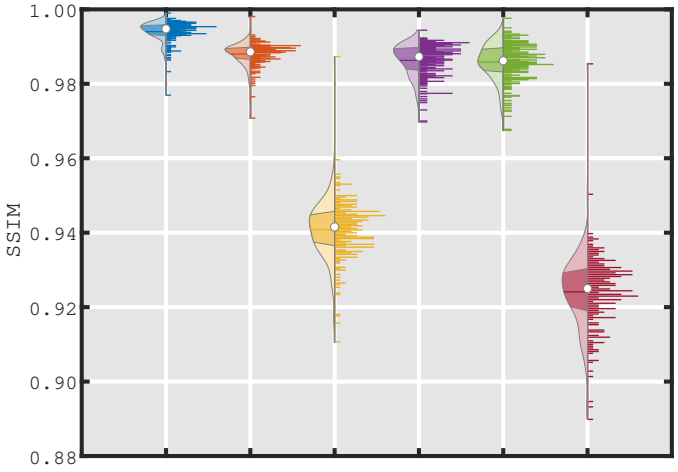


图 7： 针对不同类型失真的鲁棒性评估。

中和视频的平均PSNR值分别高于40和35，表明其不可感知性和可逆性达到可接受的水平。

D. 鲁棒性

将系统应用于实际场景时，鲁棒性是一个重要考量因素。它涉及系统在存在故障、干扰和不同条件下仍能继续运行的程度。我们通过中和视频相对于模糊度、亮度、对比度和色调变化的结构相似性指数，来衡量中和性能。图 7 根据结构相似性指数值的分布，对各个情况进行了比较。退化参数在有限范围内随机采样。正如预期，在无失真和混合失真的情况下，结构相似性指数的平均值分别处于两端。在各个单独的情况下，亮度调整对中和性能的负面影响似乎最大，而与无失真情况相比，其他调整仅引起轻微波动。图 8 对整个退化范围内的各个情况进行了比较。可以看出，该系统能够承受亮度和对比度的显著变化，但在模糊度和色调出现极端调整时，性能会突然下降。总体而言，评估结果表明，在一定范围的退化条件下，该系统可以被认为具有相当的韧性，并且对于某些类型的失真而言是可靠的。

E. 可验证性

为了证明已接种视频能够自动验证，并且很容易与未接种视频区分开来，我们采用了多种规模和架构复杂度各异的神经网络分类器，并评估它们在不同退化条件下的分类性能。如图 9 所示，在所有条件下，各分类器的平均准确率均约为 99%，且真正率仅略低于真负率。总体而言，参数更多、神经元连接更复杂的先进模型似乎能够取得更高的准确率。

尽管如此，值得注意的是，即使是最基础的感知器模型，其分类性能也能与最先进模型相媲美，这表明已接种和未接种视频经中和处理后的结果很容易区分。

F. 深度伪造免疫性与局限性

我们展示了对人脸替换和人脸重演的免疫力，这两类是可能导致严重后果的主要深度伪造类型。我们评估了对依赖掩码和与掩码无关的深度伪造方法的免疫力。总体而言，依赖掩码的深度伪造更容易应对，因为非人脸区域更有可能保持完整。对于依赖掩码的人脸替换，我们采用经典的基于自编码器的方法。从图 10 中可以看出，只要视频在遭受攻击前完成接种，面部区域就能以高保真度得到恢复。相比之下，未接种疫苗病例中可以观察到空白的面部区域，这表明当给定视频未接种疫苗时，系统不会产生免疫反应。中和帧上的边界框标注显示了由验证器模型确定的接种有效性。值得注意的是，系统还表现出对头部位置和遮挡物（例如眼镜）的适应性。不过，系统仍存在一些局限性，如肤色不准确以及化妆色彩不匹配，如图 11 所示。前一个问题可以通过后处理，采用更精细的融合机制加以改善。后一个现象可能是由异常色彩（即分布外数据）导致的，因此一种可能的改进方式是使用从真实来源收集或人工生成的更多样化数据来训练模型。对于与掩码无关的人脸替换，我们使用预训练的 SimSwap 模型。该模型与身份无关，能够使用单个模型交换任意身份，而不需要为每一对身份分别训练一个模型。图 12 显示 SSIM 分数有所提高

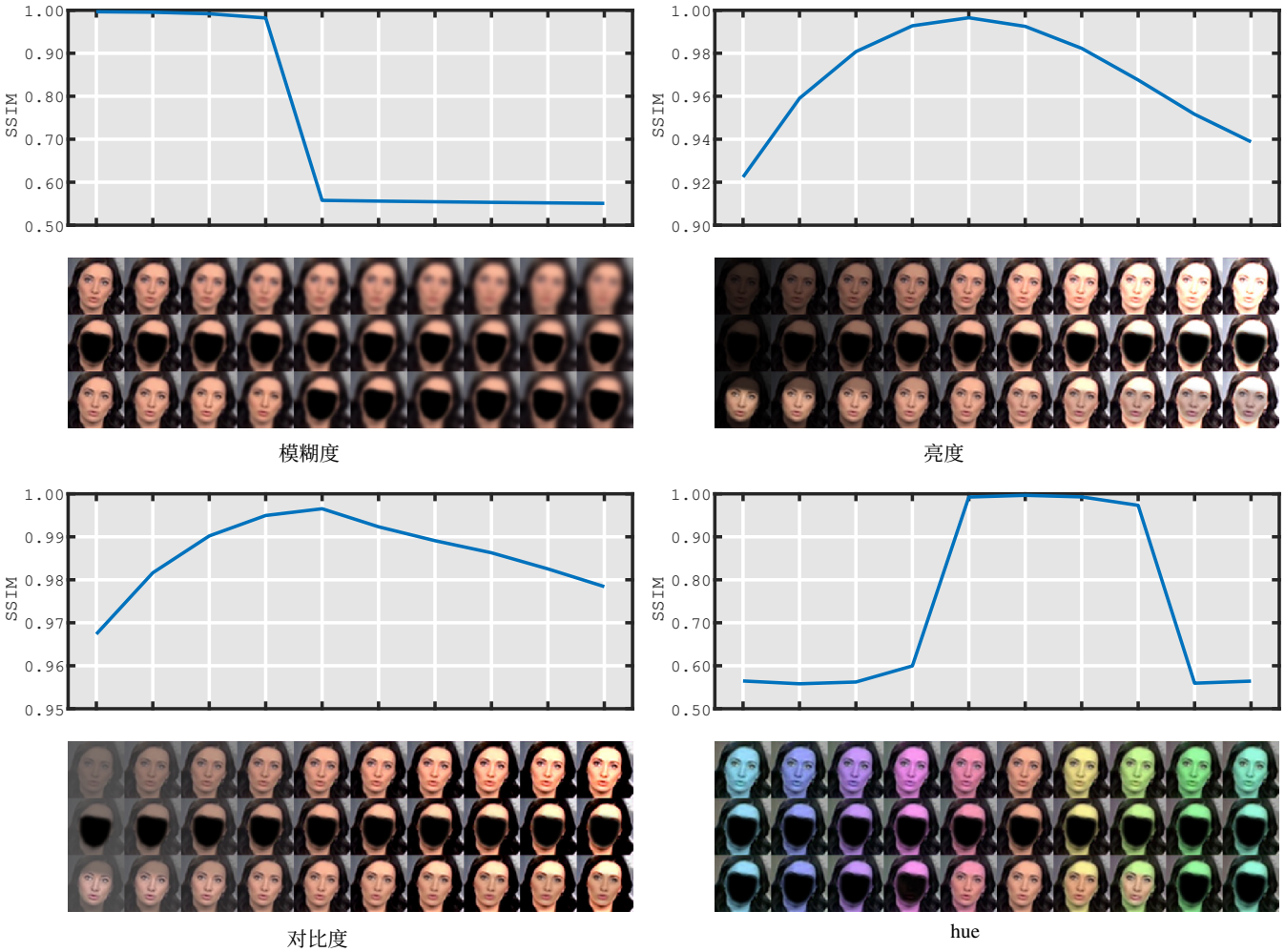


图 8： 针对全频谱退化的鲁棒性评估 础。

在受感染视频和中和视频之间。从图 12 提供的视觉示例可以看出，尽管中和机制有效，但与掩码无关情形下的可逆性不如依赖掩码情形，尤其是在面部区域的边界处（例如眉毛）。我们还使用预训练的 X2Face 模型测试了系统对人脸重演的性能。具体而言，探索姿态变化的影响十分有趣。图 14 的结果表明，在姿态变化较小的情况下，中和是可行的；然而，当出现较大的姿态变化时，感染是不可逆的。

V. 结论

在本研究中，我们提出了一种网络疫苗接种机制，以赋予内容抵御深度伪造的免疫力。研究表明，接种造成的失真通常难以察觉，并且在各种损坏条件下都能有效实现中和。此外，使用种类广泛的神经网络分类器即可轻松验证接种的有效性。不过，某些情况下仍会出现颜色错位，可通过后处理和数据增强加以改善。尽管该方法依赖掩码，但研究结果证明，它对与掩码无关的人脸替换也能提供免疫力。对于会改变姿势并因此大幅改变非人脸区域的人脸重演，值得进一步研究具有更强鲁棒性的新机制。我们期待网络疫苗在未来取得进一步进展，以应对深度伪造带来的更多真实环境中的威胁。

参考文献

[1] S. Greengard，“深度伪造会造成严重破坏吗？”*ACM* 通讯，第 63 卷，第 1 期，第 17–19 页，2019。[2] Y. Mirsky 和 W. Lee，“深度伪造的生成与检测：综述”，*ACM* 计算调查，第 54 卷，第 1 期，第 1–41 页，2021，文章编号 7。[3] F. Marra、D. Gragnaniello、D. Cozzolino 和 L. Verdoliva，“社交网络中 GAN 生成的虚假图像检测”，载于*IEEE* 多媒体信息处理与检索会议（*MIPR*）论文集，美国佛罗里达州迈阿密，2018，第 384–389 页。[4] D. Afchar、V. Nozick、J. Yamagishi 和 I. Echizen，“MesoNet：紧凑型人脸视频伪造检测网络”，载于*IEEE* 信息取证与安全国际研讨会（*WIFS*）论文集，中国香港，2018，第 1–7 页。[5] Y. Li、M.-C. Chang 和 S. Lyu，“In ictu oculi：通过检测眨眼揭露人工智能生成的虚假视频”，载于*IEEE* 信息取证与安全国际研讨会（*WIFS*）论文集，中国香港，2018，第 1–7 页。[6] D. G¨uera 和 E. J. Delp，“使用循环神经网络检测深度伪造视频”，载于 *IEEE* 高级视频与信号监控国际会议（*AVSS*）论文集，新西兰奥克兰，2018，第 1–6 页。

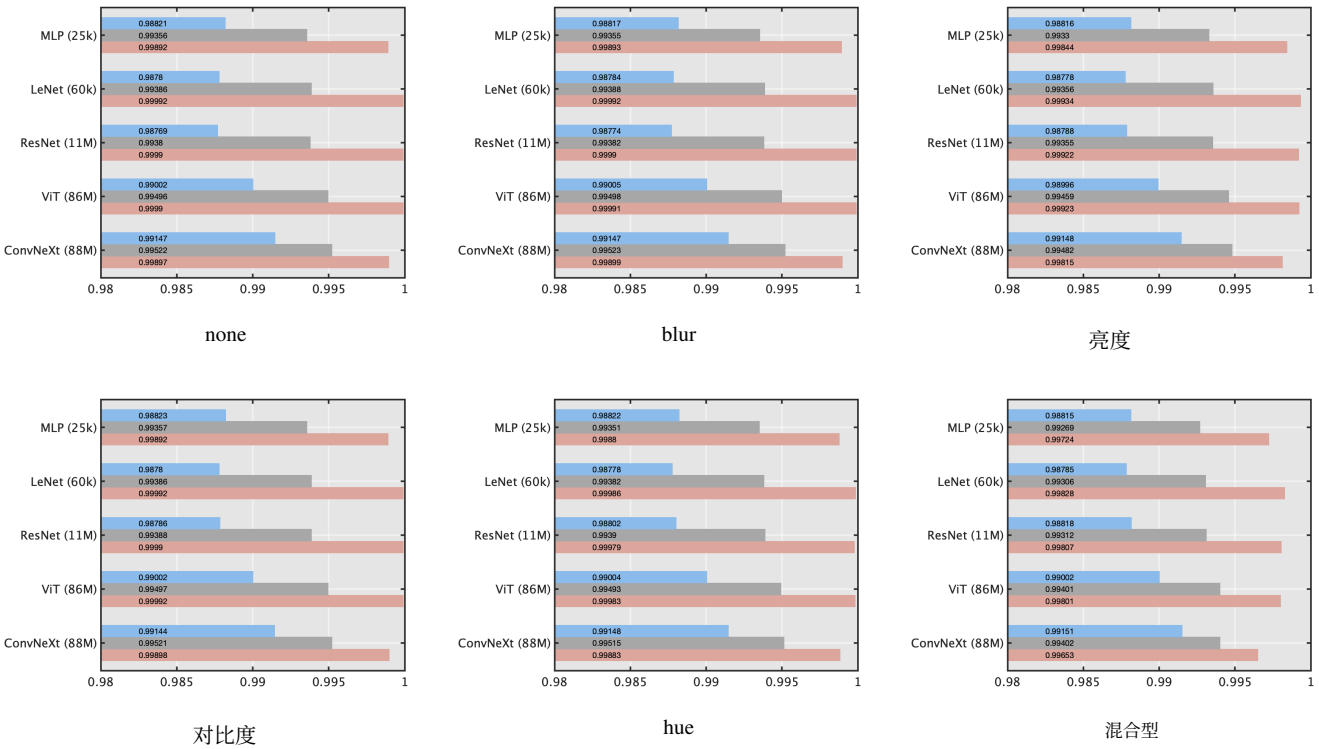


图 9： 不同神经网络分类器的可验证性，分别以真正率（顶部条）、准确率（中间条）和真负率（底部条）表示。



图 10： 原始深度伪造对依赖掩码的人脸替换的免疫力演示。上排（从左到右）：未接种视频、感染视频和中和视频。下排（从左到右）：已接种疫苗的视频、感染视频和中和视频。

[7] H. H. Nguyen，《用于更通用的人脸伪造检测的人脸X射线》，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，布莱顿，英国，2019年，第2307–2311页。

[8] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo，《用于更通用的人脸伪造检测的人脸X射线》，载于论文集*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），美国华盛顿州西雅图，2020年，第5000–5009页。

[9] J. H. Bappy、A. K. Roy-Chowdhury、J. Bunk、L. Nataraj 和 B. Manjunath，《利用空间结构定位篡改图像区域》，载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，意大利威尼斯，2017年，第4980–4989页。



图 11： 原始视频（上排）与中和视频（下排）之间颜色错位的案例研究。

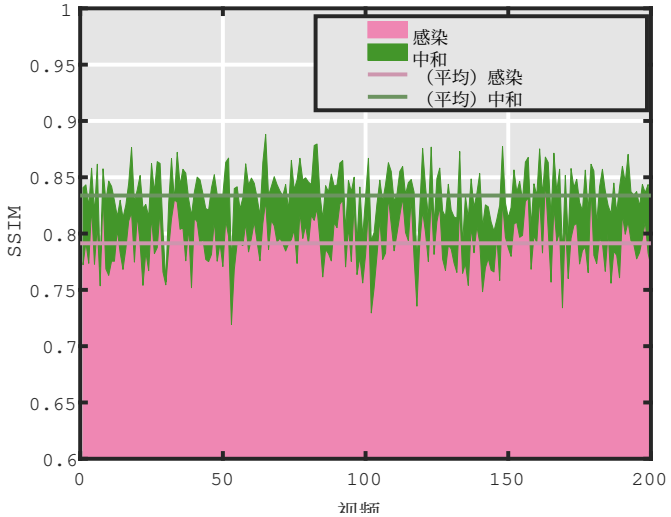


图 12： SimSwap 感染可逆性的评估。

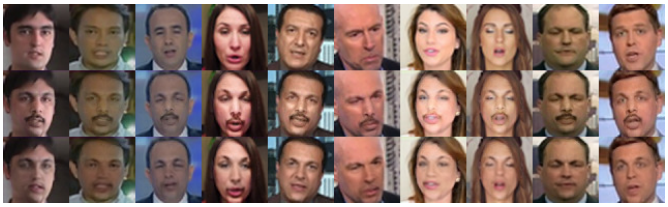


图 13： SwimSwap 对与掩码无关的人脸替换的免疫力演示。上：原始图像。中：感染图像。下：中和图像。

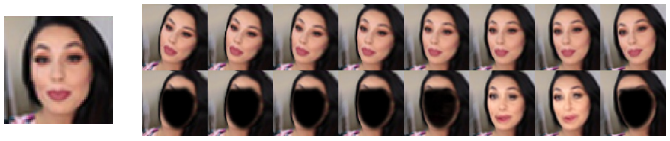


图 14： X2Face 对人脸重演的免疫力演示。左：已接种疫苗的图像。右：感染图像（上排）和中和图像（下排）。

[10] P. Zhou、X. Han、V. I. Morariu 和 L. S. Davis，《学习用于图像篡改检测的丰富特征》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，美国犹他州盐湖城，2018年，第1053–1061页。[11] J. H. Bappy、C. Simons、L. Nataraj、B. S. Manjunath 和 A. K. Roy-Chowdhury，《用于检测图像伪造的混合LSTM与编码器–解码器架构》，*IEEE*《图像处理汇刊》，第28卷，第7期，第3286–3300页，2019年。[12] H. H. Nguyen、F. Fang、J. Yamagishi 和 I. Echizen，《用于检测和分割篡改人脸图像与视频的多任务学习》，载于*IEEE*国际生物识别会议论文集理论、应用与系统（BTAS），美国佛罗里达州坦帕，2019，页1–8。[13] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，“迈向能够抵御对抗攻击的深度学习模型”，载于《国际学习表征会议》（ICLR）论文集，加拿大不列颠哥伦比亚省温哥华，2018，页1–23。[14] F. Tram`er、A. Kurakin、N. Papernot、I. J. Goodfellow、D. Boneh 和 P. D. McDaniel，“集成对抗训练：攻击与防御”，载于《国际学习表征会议》（ICLR）论文集，加拿大不列颠哥伦比亚省温哥华，2018，页1–20。[15] F. Tram`er 和 D. Boneh，“针对多种扰动的对抗训练与鲁棒性”，载于《神经信息处理系统国际会议》（NeurIPS）论文集，加拿大不列颠哥伦比亚省温哥华，2019，页5866–5876。[16] J. Thies、M. Zollh¨ofer、M. Nießner、L. Valgaerts、M. Stamminger 和 C. Theobalt，“用于人脸重演的实时表情迁移”，ACM《图形学汇刊》，第34卷，第6期，页1–14，2015，文章编号183。[17] H. Kim、P. Garrido、A. Tewari、W. Xu、J. Thies、M. Niessner、P. P´erez、C. Richardt、M. Zollh¨ofer 和 C. Theobalt，“深度视频肖像”，ACM《图形学汇刊》，第37卷，第4期，页1–14，2018，文章编号163。[18] J. Thies、M. Zollh¨ofer、M. Stamminger、C. Theobalt 和 M. Nießner，“Face2Face：RGB视频的实时人脸捕捉与重演”，ACM 通讯，第62卷，第1期，页96–104，2018。[19] D. Kononenko、Y. Ganin、D. Sungatullina 和 V. Lempitsky，“利用机器学习实现照片级真实感的单目视线重定向”，IEEE模式分析与机器智能汇刊，第40卷，第11期，页2696–2710，2018。[20] Y. Yu、G. Liu 和 J.-M. Odobez，“通过视线重定向合成改进少样本用户专属视线适应”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，美国加利福尼亚州长滩，2019，页11 929–11 938。[21] Y. Ganin、D. Kononenko、D. Sungatullina 和 V. Lempitsky，“Deep-Warp：用于视线操控的照片级真实感图像再合成”，载于《欧洲计算机视觉会议》（ECCV）论文集，荷兰阿姆斯特丹，2016，页311–326。[22] L. Tran、X. Yin 和 X. Liu，“通过旋转人脸进行表征学习”，IEEE模式分析与机器智能汇刊，第41卷，第12期，页3007–3021，2019。[23] A. Siarohin、E. Sangineto、S. Lathuili`ere 和 N. Sebe，“用于基于姿态的人体图像生成的可变形GAN”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，美国犹他州盐湖城，2018，页3408–3416。[24] G. Balakrishnan、A. Zhao、A. V. Dalca、F. Durand 和 J. Guttag，“合成未见姿态下的人体图像”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，美国犹他州盐湖城，2018，页8340–8348。[25] Y. Hu、X. Wu、B. Yu、R. He 和 Z. Sun，“姿态引导的照片级真实感人脸旋转”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，美国犹他州盐湖城，2018，页8398–8406。[26] N. Neverova、R. Alp G¨uler 和 I. Kokkinos，“密集姿态迁移”，载于《欧洲计算机视觉会议》（ECCV）论文集，德国慕尼黑，2018，页128–143。[27] L. Liu、W. Xu、M. Zollh¨ofer、H. Kim、F. Bernard、M. Habermann、W. Wang 和 C. Theobalt，“人类演员视频的神经渲染与重演”，ACM《图形学汇刊》，第38卷，第5期，页1–14，2019，文章编号139。[28] R. D. Bem、A. Ghosh、A. Boukhayma、T. Ajanthan、N. Siddharth 和 P. Torr，“一种自然场景中人物的条件深度生成模型

图像”，载于IEEE冬季计算机视觉应用会议（WACV）论文集，美国夏威夷州威可洛亚，2019，页1449–1458。[29] C. Chan、S. Ginosar、T. Zhou 和 A. Efros，“现在大家跳舞”，载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，韩国首尔，2019，页5932–5941。[30] Y. Zhou、Z. Wang、C. Fang、T. Bui 和 T. L. Berg，“舞蹈生成：互联网视频的动作迁移”，载于IEEE/CVF国际计算机视觉会议研讨会（ICCVW）论文集，韩国首尔，2019，页1208–1216。[31] K. Aberman、M. Shi、J. Liao、D. Lischinski、B. Chen 和 D. Cohen-Or，“基于深度视频的表演克隆”，《计算机图形学论坛》，第38卷，第2期，页219–233，2019。[32] J. Thies、M. Elgharib、A. Tewari、C. Theobalt 和 M. Nießner，“神经语音木偶：音频驱动的面部重演”，载于《欧洲计算机视觉会议》（ECCV）论文集，英国格拉斯哥，2020，页716–731。[33] S. Suwajanakorn、S. M. Seitz 和 I. Kemelmacher-Shlizerman，“合成奥巴马：从音频中学习唇形同步”，ACM《图形学汇刊》，第36卷，第4期，页1–13，2017，文章编号95。[34] A. Jamaludin、J. S. Chung 和 A. Zisserman，“你是这么说的吗？：从音频合成说话人脸”，《国际计算机视觉期刊》，第127卷，第11期，页1767–1779，2019。[35] K. Vougioukas、S. Petridis 和 M. Pantic，“利用生成对抗网络实现逼真的语音驱动面部动画”，《国际计算机视觉期刊》，第128卷，第5期，页1398–1413，2020。[36] A. Brock、T. Lim、J. M. Ritchie 和 N. Weston，“利用内省式对抗网络进行神经照片编辑”，载于《国际学习表征会议》（ICLR）论文集，法国土伦，2017，页1–15。[37] Z. Shu、E. Yumer、S. Hadap、K. Sunkavalli、E. Shechtman 和 D. Samaras，“利用固有图像解耦进行神经人脸编辑”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，美国夏威夷州檀香山，2017，页5444–5453。[38] Z. He、W. Zuo、M. Kan、S. Shan 和 X. Chen，“AttGAN：仅改变所需属性的人脸属性编辑”，IEEE《图像处理汇刊》，第28卷，第11期，页5464–5478，2019。[39] C. E. Shannon，“通信的数学理论”，《贝尔系统技术期刊》，第27卷，第3期，页379–423，1948。[40] M. Bertalmio、G. Sapiro、V. Caselles 和 C. Ballester，“图像修复”，载于计算机图形学与交互技术年会（SIGGRAPH）论文集，美国路易斯安那州新奥尔良，2000，页417–424。[41] S. Iizuka、E. Simo-Serra 和 H. Ishikawa，“全局与局部一致的图像补全”，ACM《图形学汇刊》，第36卷，第4期，2017，文章编号107。[42] Y. Li、S. Liu、J. Yang 和 M.-H. Yang，“生成式人脸补全”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，美国夏威夷州檀香山，2017，页5892–5900。[43] J. Fridrich 和 M. Goljan，“具备自我纠错能力的图像”，载于IEEE国际图像处理会议（ICIP）论文集，第3卷，日本神户，1999，页792–796。[44] X. Zhang 和 S. Wang，“具有无误差恢复能力的脆弱水印”，IEEE《多媒体汇刊》，第10卷，第8期，页1490–1499，2008。[45] H. He、F. Chen、H.-M. Tai、T. Kalker 和 J. Zhang，“基于块邻域的自恢复脆弱水印方案的性能分析”，IEEE《信息取证与安全汇刊》，第7卷，第1期，页185–196，2012。[46] C. Qin、C.-C. Chang 和 P.-Y. Chen，“基于自适应比特分配机制、具备恢复能力的自嵌入脆弱水印”，《信号处理》，第92卷，第4期，页1137–1150，2012。[47] P. Korus 和 A. Dziech，“利用自嵌入进行内容重建的高效方法”，IEEE《图像处理汇刊》，第22卷，第3期，页1134–1147，2013。[48] C. Qin、P. Ji、C.-C. Chang、J. Dong 和 X. Sun，“基于最优迭代BTC的非均匀水印共享，用于图像篡改恢复”，IEEE《多媒体》，第25卷，第3期，页36–48，2018。[49] S. Baluja，“将图像隐藏于显眼处：深度隐写”，载于《神经信息处理系统进展》（NeurIPS）论文集，第30卷，2017，页2066–2076。[50] J. Zhu、R. Kaplan、J. Johnson 和 L. Fei-Fei，“HiDDeN：利用深度网络隐藏数据”，载于《欧洲计算机视觉会议》（ECCV）论文集，德国慕尼黑，2018，页682–697。[51] M. Tancik、B. Mildenhall 和 R. Ng，“StegaStamp：实体照片中的隐形超链接”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，美国华盛顿州西雅图，2020，页2114–2123。

[52] A. R¨ossler、D. Cozzolino、L. Verdoliva、C. Riess、J. Thies 和 M. Niess-ner，“FaceForensics++：学习检测篡改人脸图像”，载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，韩国首尔，2019年，第1–11页。[53] O. Ronneberger、P. Fischer 和 T. Brox，“U-Net：用于生物医学图像分割的卷积网络”，载于国际医学图像计算与计算机辅助干预会议（*MICCAI*）论文集，尚贝里，2015年，第234–241页。[54] A. Q. Nichol 和 P. Dhariwal，“改进的去噪扩散概率模型”，载于国际机器学习会议（*ICML*）论文集，第139卷，线上会议，2021年，第8162–8171页。[55] J. Ho、A. Jain 和 P. Abbeel，“去噪扩散概率模型”，载于神经信息处理系统国际会议（*NeurIPS*）论文集，加拿大不列颠哥伦比亚省温哥华，2020年，第6840–6851页。[56] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，美国路易斯安那州新奥尔良，2022年，第10 674–10 685页。[57] C. Saharia、J. Ho、W. Chan、T. Salimans、D. J. Fleet 和 M. Norouzi，“通过迭代细化实现图像超分辨率”，*IEEE*模式分析与机器智能汇刊，第1–14页，2022年。[58] D. E. Rumelhart、G. E. Hinton 和 R. J. Williams，“通过误差传播学习内部表征”，载于并行分布式处理：认知微观结构探索。美国马萨诸塞州剑桥：MIT出版社，1986年，第318–362页。[59] Y. Lecun、L. Bottou、Y. Bengio 和 P. Haffner，“应用于文档识别的基于梯度的学习”，《*IEEE*会议录》，第86卷，第11期，第2278–2324页，1998年。[60] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，美国内华达州拉斯维加斯，2016年，第770–778页。[61] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，“一张图像胜过16×16个词：面向大规模图像识别的变换器”，载于国际学习表征会议（*ICLR*）论文集，线上会议，2021年，第1–21页。[62] Z. Liu、H. Mao、C.-Y. Wu、C. Feichtenhofer、T. Darrell 和 S. Xie，“面向2020年代的卷积网络”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，美国路易斯安那州新奥尔良，2022年，第11 966–11 976页。[63] R. Chen、X. Chen、B. Ni 和 Y. Ge，“SimSwap：一种高效的高保真人脸交换框架”，载于*ACM*国际多媒体会议（*MM*）论文集，美国华盛顿州西雅图，2020年，第2003–2011页。[64] O. Wiles、A. S. Koepke 和 A. Zisserman，“X2Face：一种利用图像、音频和姿态编码控制人脸生成的网络”，载于欧洲计算机视觉会议（*ECCV*）论文集，德国慕尼黑，2018年，第690–706页。[65] F. Schroff、D. Kalenichenko 和 J. Philbin，“FaceNet：用于人脸识别和聚类的统一嵌入”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2015年，第815–823页。