# 使用地标位置曝光GAN合成的面孔

杨欣，李月尊，齐红岗，吕四维

美国纽约州立大学奥尔巴尼分校计算机科学系

中国科学院大学计算机与控制工程学院

## 摘要

生成对抗网络（GAN）最近导致了非常逼真的图像合成结果。在这项工作中，我们描述了一个新的，使用面部标志性点位置的方法，来曝光GAN合成的图像。我们的方法基于观察：通过GAN模型生成的面部部位的配置，由于缺乏全局性约束而与真实面孔所具有的差别。我们进行的实验证明了这种现象，并且表明使用面部标志点位置进行训练的SVM分类器，足以实现对GAN合成面孔良好的分类性能。

## 关键词

图像取证，生成对抗网络，面部标志

**ACM参考格式：**

杨欣，李月尊，齐红岗†，吕四维。2019年。使用地标位置曝光GAN合成的面孔。在美国纽约ACM共6页。

## 1引言

人工智能技术的飞速发展，以及大量在线图像和视频和高通量计算硬件的可用性增加，使用于生成视觉逼真的图像和视频的工具，发生了革命性的变化。这些技术变得越来越高效，更多用户可以使用。在最近的深神经网络[1，3，4，21] 发展中，尤其是生成对抗网络（GAN）[5]，催生了一种新型的图像合成方法，它能够产生具有高度真实感的图像。图1显示了一些GAN合成人脸的示例，使用最近基于GAN的方法[7 ,8]获得的效果非常出色。

GAN合成图像的复杂性的提升，也会造成具有虚假视觉媒体的负面影响，并且最具破坏性的例子，可能是捏造或操纵人的脸，因为脸带有一个人最可识别的信息。带有GAN合成面孔的伪造媒体的广泛传播，引发了重大的道德，法律和安全问题，并且在图像和视频中可以检测GAN合成人脸的方法是迫切需要的。

与以前的图像/视频操作方法不同，逼真的图像完全是由随机噪声通过深度神经网络生成的。当前的检测方法基于低级别特征，例如色差[10，13]，或使用整张图像作为输入，到神经网络中，以提取整体特征 [19]。

在这项工作中，我们开发了一种新的GAN合成人脸检测方法，这是一种基于语义上更有意义的特征的方法

，即面部标志点的位置。这是因为GAN合成脸部在脸部标志位置中表现出某些/确切异常。具体来说，基于GAN的人脸合成算法可以产生脸部部位（例如，眼睛，鼻子，皮肤和嘴等），带有大量的真实细节，但在脸上的这些部分的位置没有明确的约束。打个比方，基于GAN的人脸合成方法就像在Fukuwarai游戏中的玩家一样，它具有脸部所有部位，但缺乏使他们像真实面孔一样自然而连贯的方法。

（Fukuwarai是新年期间在日本玩的传统游戏。一个Fukuwarai玩家的眼睛被蒙上，并被要求将脸部的一部分（即眼睛，眉毛，鼻子和嘴巴），通常印在纸上，留在一张空白的脸上。）

我们发现这些在GAN合成的脸的面部的配置的异常，可以使用，在脸部自动检测的面部标志点的位置显示（例如，眼睛，鼻子和嘴）。为了适应不同形状，方向和比例的变化，我们将进一步将所有面部标志标准化为相同的标准的坐标系统。然后，我们使用这些面部标志的标准化位置作为简单SVM分类器的功能/特征。基于SVM分类器的标志位置的在，用最先进的基于GAN的面部合成PGGAN [7]脸部生成，中测试，其中，它显示出合理的分类性能，而仅使用低尺寸特征和带有较少参数的光照模型。

## 2相关工作

### 2.1基于GAN的人脸合成方法

自[[5](https://translate.googleusercontent.com/translate_f#6)]的创始工作以来，GAN彻底改变了图像合成方法。一个GAN模型由两个神经网络组成，被称为发生器和编码器，它们串联训练。发生器将随机噪声作为输入并合成图像，该图像被送到鉴别器，旨在从真实图像区分合成图像。这两个网络是经过训练可以互相竞争：发生器旨在创造更具真实感的图像可击败分类器，同时鉴别器网络经过培训，可以更有效地区分两种类型的图像。当两个网络达到游戏的平衡时，训练结束。原始GAN模型曾经历了许多改进。特别地，为了改善训练中的稳定性，Radford等人通过介绍深度卷积GAN（DCGAN）[15]优化了网络结构。耦合生成对抗网络（COGAN）学习来自不同领域的联合分布，进一步改善了合成图像的真实性[11]。但是，训练过程中存留[2，6，9，17，18]不稳定性，它传播到合成样本并将模型限制为仅合成低分辨率，请参见图1。

PGGAN [7] 是合成高辨率真实人脸图像的重大突破。从发生器生成4×4分辨率图像开始，它同时提高了发生器和鉴别器。生成的图像和训练图像一起，调整为相同比例大小，输入鉴别器。网络稳定后，三层块（类似于剩余块），生成高度和宽度加倍，消失/褪色于网络。这些模型稳定化通过训练和较高分辨率图层淡入交替带出，直到生成图像的1024×1024分辨率达到为止。这种方法不仅提高了训练速度和稳定性，而且还合成了带有前所未有细节的、高分辨率的人脸图像（1024×1024）。PGGAN模型通过style-transfer GAN（STGAN）[8[]](https://translate.googleusercontent.com/translate_f#6)进一步完善，STGAN处理面部合成问题是将一张脸的样式转换为另一张脸。然而，在，以现有图像而不是直接来自随机噪音，为条件，生成最佳质量的图像，的方面，STGAN与以前基于GAN的图像综合模型有根本不同。因此，在这项工作中，我们不考虑检测STGAN生成的图像。



图1：多年来，GAN模型在其合成的面孔质量上已得到显着改善。这里我们展示一些不同的GAN模型的例子（a）GAN [5]，（b）DCGAN [15]，（c）COGAN [11]，（d）PGGAN [7]，（e）Style-GAN [8]。



图2：由PGGAN合成的基于GAN的面部异常[7]。（ a）眼睛的大小和位置不对称。（b）嘴相对于鼻子向左移动。（c）左眼眼角（眼睛的内角）的尖角和向下外侧。

### 2.2 GAN 合成图片的检测方法

与流行的用GAN的合成人脸图像探索策略相比，区分真实的与合成图像的方法远不能令人满意。李等人[10]观察到：真实的与GAN生成的图像中H，S，V和Cb，Cr，Y通道之间的颜色不匹配。同样，麦克洛斯基和奥尔布赖特确定GAN生成图像的饱和像素的频率和彩色图像稳定度与由相机[13]捕获的图像是不同的。但是，这种色差很容易在图像合成后通过后期处理将其移除。另一方面，Mo等[14]和Tariq [19]设计了伪造暴露的深卷积神经网络分类器，通常需要GPU进行培训和测试，并且无法显示分类背后的机制。

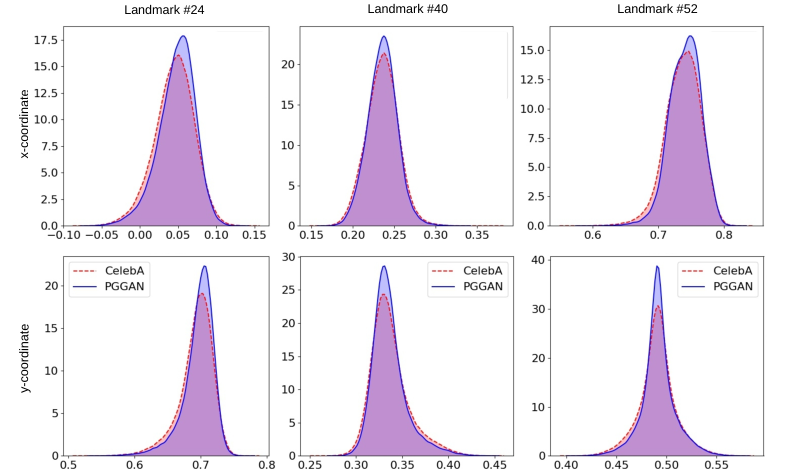


图3：真实（CelebA）和GAN合成的假（PGGAN）面孔在x-，y-坐标上的标准化面孔标志位置的密度分布。真实的面孔来自CelebA数据集，具有200k +张图像，假脸来自PGGAN数据集，具有100k张图像。

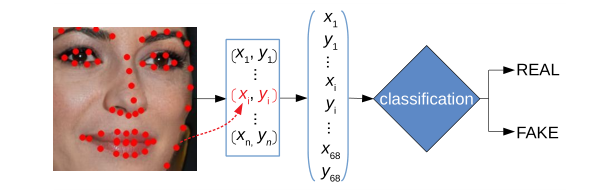


图4：用于分类的图像管道。从变形为标准配置的面部图像中检测到68个标记，然后将标志位置平铺到分类器的136D 向量。

## 3方法

正如我们在简介中所述，GAN合成的面孔可能由于较弱的全局限制而显示出面部部位的不一致构造。这种现象的几个例子如图2所示，使用最新的PGGAN方法[7] 合成了高分辨率的人脸图像。在（a）中，我们观察到合成了两只眼睛，鼻子和上唇不对称。在（b）中，右眼变形，相对于鼻子的尖端嘴巴向左偏。在（c）中，脸部显示不自然外眦（左眼的内角尖锐和向下）以及大小不同的两只眼睛。

为了量化这种不一致，我们比较了在GAN合成的真实面孔上检测到的面部标志位置。我们先运行面部检测器并提取面部标志，如图4。通过最小化对齐误差的微型仿射变换，检测的标志将在[0，1]×[0，1]的区域，变形为标准配置。为了减少脸型对对齐结果的影响，我们按照标准程序估算，仅使用除脸部轮廓区域以外的，脸部中央区域标志进行的变形。图3显示了，沿x-和y-图像坐标方向分布的，真实的和GAN合成的的面孔，对齐的标志位置的差异。这些结果表明，GAN生成的面孔的地标的边界分布显示出一些一致的差异，当我们考虑人脸标志点的集合的所有坐标上的联合分布时，这种差异更加突出。因此，我们可以使用，通过将所有这些地标位置矢量化而形成的矢量，作为特征向量以建立用于区分GAN合成的面孔和真实面孔的分类系统，如图4。

将此类功能/特性用于分类任务有三个优点。首先，此功能/特性的尺寸相对较小（它是我们从每个面孔提取的地标数量的两倍）。这有助于构建更简单的分类方案。第二，地标位置与图像大小无关，因此无需在训练和使用获得的分类方法时重新缩放图像，这也可以避免，在图像分辨率方面，由于调整大小操作，导致分类器捕获人为差异的不良的副作用。第三，面部标志位置的异常归因于底层GAN图像合成的基本机制，因此，不引入更复杂的约束到GAN框架中，这可能不是可以轻松修复的。

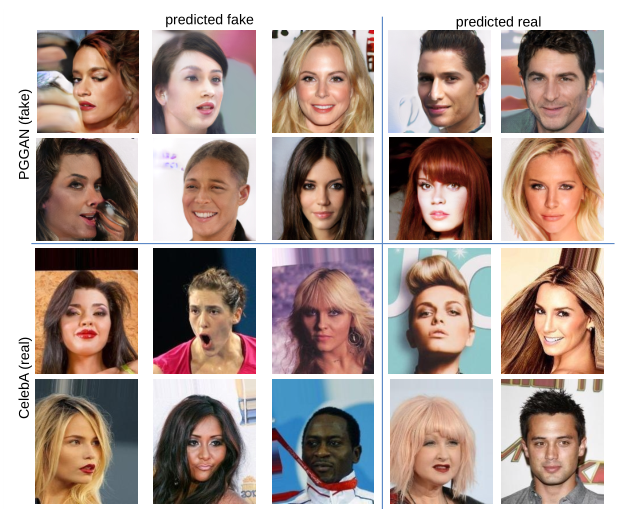


图5：关于CelebA和PGGAN数据集的正确和不正确预测的示例。

## 4实验

在本节中，我们报告，使用标记位置作为特征，以从由GAN合成的图像中分辨出真实图像，进行的实验评估。

**选择基于GAN的人脸合成方法。**

虽然有许多基于GAN的面合成方法[5，7， 8，11，15]，我们在这项工作中选择了最新的PGGAN进行构建和评估，基于分类方法的标志位置，来分类当前最先进的高质量GAN合成面孔。这个选择是出于以下原因。早期基于GAN的人脸合成方法[ 5， 11，15]产生低质量的低分辨率的面部图像，因此它们不能代表这门艺术的状态。另一方面，最新的STGAN无法像基于GAN的其他方法一样从噪声中合成人脸图像，但可视其为样式转移问题。

**数据集。**

为了暴露GAN合成的图像，SVM分类器的训练和测试基于两个数据集：（a）CelebFaces属性数据集（CelebA）[12]，包含超过200K个固定分辨率为216×178像素的真实面部图像。（b）PGGAN数据集[7 ]，由100K个分辨率为1024×1024像素的PGGAN合成的面孔图像组成，用作假脸。两个数据集的75％被合并为用于训练的正样本和负样本，其余25％用于测试。

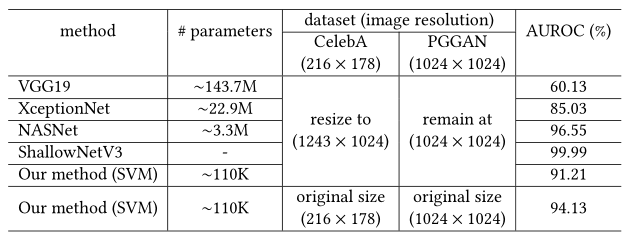
**预处理和训练。**

使用的归一化的所有脸部标志位置作为特征，我们可以开发一个简单的分类方案，来区分真实面孔和GAN合成的面孔（使用标准分类方法（例如SVM或神经网络）），如图4。在这项研究中，归一化的每个脸（∈R 68×2）的标志位置将展平为一个向量（∈R 136×1），即，通过减去平均值并除以所有训练样本的标准偏差来标准化。我们使用，在超参数使用5倍交叉验证进行格点搜索的，径向基函数（RBF）内核，训练了SVM分类器。两类的损失通过反比例调整样本损失，归类训练数据集中的频率来平衡。

**性能。**

图5 显示了一些在PGGAN和CelebA数据集上预测结果的示例。PGGAN人工合成的脸可以正确地被预测为假脸，而被错误地预测为真实的脸几乎没有可见的缺陷。对于CelebA数据集中的真实面部图像，有些面部被错误地预测为是假的。这可能是由于难以准确估计具有强烈面部表情和遮挡的脸部的标志位置所致，如图5左下方所示。

表1：我们的方法和其他深度神经网络方法[19 ]在PGGAN和CelebA调整大小变成不同的图像宽度后的AUCROC性能



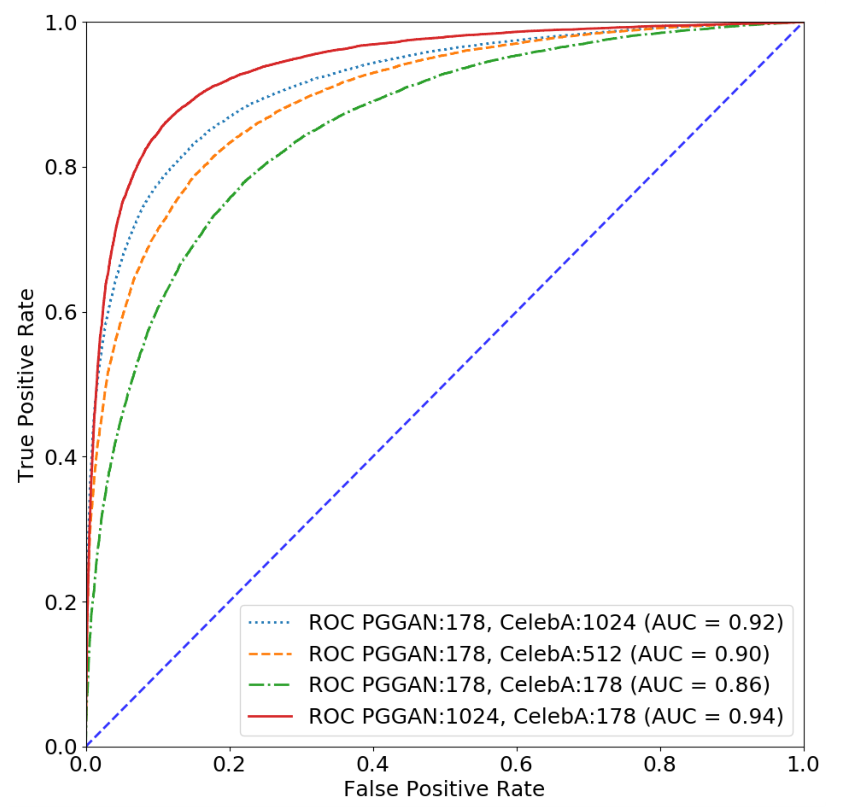


图6：通过改变PGGAN和CelebA图片的宽度得到的性能。每个数据集的调整大小图像的宽度在图例中。

我们的方法，在ROC下的区域（AUROC）方面，的定量的结果如表1所示。作为比较，我们还包括，在同一数据集上，来自[19]的不同神经网络架构的性能。请注意，[19]中的所有方法都将图像作为输入。为了适应CelebA（216×178）和PGGAN（1024×1024）数据集之间不同输入图像的尺寸，图像被调整为相同大小，并且表1中列出了1243×1024的celebA图像放大后的结果。

结果表明，SVM分类器的AUROC达到了94.13％，并且胜过几种基于一些其他方法（例如VGG19和XceptionNet）的深度神经网络。基于分类精度更高的方法的两个深度神经网络，需要大量的参数。更重要的是，这些结果是通过调整大小后的图像获得的，没有学术研究对调整大小对最终分类（向上采样图像导致某些伪像）的影响做过研究。目前尚不清楚多少高性能可以归因于两种类型的图像的本质不同。如前所述，该基于面部标志的位置的特征与图像大小无关；另一组实验中，我们比较了调整两类图像大小的效果，如图6 所示，证明调整大小操作，对分类性能相对无所谓。我们还要强调这些所有的CNN模型需要GPU进行培训和测试，而我们的方法的参数少得多，只需要CPU进行训练和测试。

## 5 FACE FORENSICS数据集的测试结果

尽管我们提出的特征最初是为基于GAN的人脸合成设计的，我们认为其他类型的人脸合成方法也可能表现出类似的异常。为此，我们在FaceForensics数据集上测试我们的方法，FaceForensics数据集包含成对的真实视频和Face2Face [ 20 ]合成的伪造视频。它包含用于训练的740对视频（726,270张图像），150对（151,052张图片）进行验证，并提供150个视频（155,490张图片）进行测试[16]。FaceForensics数据集中的视频帧，宽度在576到1920像素之间，高度在480至1080像素之间，变化。基于地标特征的SVM模型在此数据集上进行相似的训练，我们在图7中报告了性能，用于单帧的分类达到了0.83 AUROC。通过平均对单个视频进行分类预测，AUROC会增加至0.90。

## 6结论

在这项工作中，我们提出使用对齐的面部标志位置作为区分PGGAN合成的假人脸图像的特征。我们的方法基于以下观察：目前的基于GAN算法使用随机噪声作为输入，擅长描述面部零件的细节，但缺乏不同的面部组件配置上的约束。因此，它在面部部位的位置引入了错误，这在GAN模型的修复中很重要。我们进行了实验以证明这种现象，并在此提示下进一步开发了分类模型。结果表明了我们的，低维输入，轻量级模型，以及对比例变化的鲁棒性（也就是强健性），的方法的有效性。

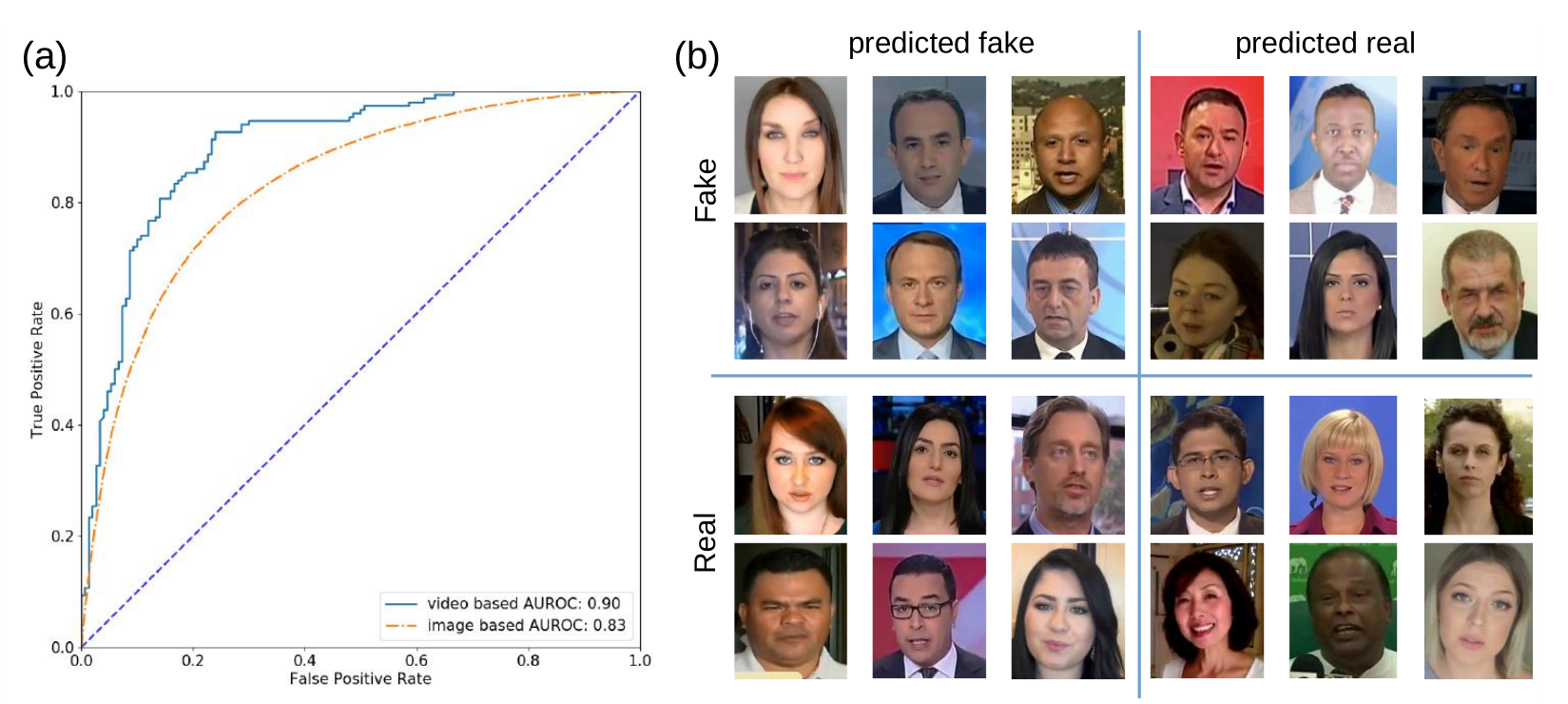


图7：FaceForensics测试数据集的分类。（a）ROC曲线和AUROC分数。（b）对于两个类正确和不正确的预测的例子。

## 参考资料

[1] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866 (2017).

[2] David Berthelot, Thomas Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017).

[3] Li Deng. 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing 3 (2014).

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In NIPS.

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In Advances in Neural Information Processing Systems. 5767–5777.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017).

[8] Tero Karras, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948 (2018).

[9] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. 2017. How to train your DRAGAN. arXiv preprint arXiv:1705.07215 2, 4 (2017).

[10] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. 2018. Detection of deep network generated images using disparities in color components. arXiv preprint arXiv:1808.07276 (2018).

[11] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In Advances in neural information processing systems. 469–477.

[12] Ziwei Liu, Ping Luo,Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In Proceedings of International Conference on Computer Vision (ICCV).

[13] Scott McCloskey and Michael Albright. 2018. Detecting GAN-generated Imagery using Color Cues. arXiv preprint arXiv:1812.08247 (2018).

[14] Huaxiao Mo, Bolin Chen, and Weiqi Luo. 2018. Fake Faces Identification via Convolutional Neural Network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. ACM, 43–47.

[15] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015).

[16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018).

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In Advances in neural information processing systems. 2234–2242.

[18] Tim Salimans and Durk P Kingma. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In Advances in Neural Information Processing Systems. 901–909.

[19] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. 2018. Detecting both machine and human created fake face images in the wild. In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. ACM, 81–87.

[20] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2387–2395.

[21] Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. 2018. A survey on deep learning for big data. Information Fusion 42 (2018), 146–157.