

基于块间光照不一致性的深度伪造检测算法

吴文轩, 周文柏, 张卫明, 俞能海

(中国科学技术大学网络空间安全学院, 安徽 合肥 230026)

摘要: 深度伪造技术的迅速发展和广泛传播引起了社会的广泛关注, 但深度伪造技术的恶意应用也给社会带来了潜在威胁。因此, 如何检测出此类深度伪造内容成为热门研究课题。以往的多数深度伪造检测算法着重于捕捉像素级别的细微伪造痕迹。目前的深度伪造算法大多忽略了伪造前后的光照信息, 导致原始人脸与伪造人脸之间存在一定的光照不一致性, 这为使用光照不一致性来检测深度伪造提供了可能。从引入光照不一致性信息和为特定任务设计网络结构模块两个角度设计了对应的算法。针对光照的引入, 通过设计对应的通道融合算法, 将更多的光照不一致信息提供给网络特征提取层, 从而衍生出新的网络结构。为了保证该网络结构的可移植性, 将特征通道融合的过程置于网络提取信息之前, 从而使所提算法能够完整移植至常见的深度伪造检测网络。针对网络结构的设计, 从网络结构和损失函数设计两个角度出发, 提出了基于块间相似性的光照不一致性深度伪造检测算法。对于网络结构, 基于伪造图像篡改区域和背景区域不一致的特性, 在网络特征层中对提取特征进行分块, 通过对比块间余弦相似度得到特征层相似矩阵, 使网络拟合重心更偏向于光照不一致性。在此基础上, 基于特征层相似性对比方案, 通过将输入图像与该图像的未篡改图像进行块间真伪性对比, 为这一任务设计了独立的真实数据参考及损失函数。实验结果表明, 与基线算法相比, 所提算法对于深度伪造检测的准确性有明显提升。

关键词: 深度伪造检测; 光照不一致性; 色彩恒常性; 块间相似度对比

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023011

Deepfake detection method based on patch-wise lighting inconsistency

WU Wenxuan, ZHOU Wenbo, ZHANG Weiming, YU Nenghai

School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China

收稿日期: 2022-03-10; 修回日期: 2022-05-19

通信作者: 周文柏, welbeckz@ustc.edu.cn

基金项目: 国家自然科学基金(62002334, U20B2047, 62121002); 安徽省自然科学基金(2008085QF296); 统筹推进世界一流大学和一流学科建设专项资金(YD3480002001); 中央高校基本科研业务费专项资金(WK2100000011)

Foundation Items: The National Natural Science Foundation of China(62002334, U20B2047, 62121002), The Natural Science Foundation of Anhui Province(2008085QF296), Special Funds for Promoting the Construction of World-class Universities and First-class Disciplines(YD3480002001), Fundamental Research Funds for the Central Universities(WK2100000011)

引用格式: 吴文轩, 周文柏, 张卫明, 等. 基于块间光照不一致性的深度伪造检测算法[J]. 网络与信息安全学报, 2023, 9(1): 167-177.

Citation Format: WU W X, ZHOU W B, ZHANG W M, et al. Deepfake detection method based on patch-wise lighting inconsistency[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 167-177.

Abstract: The rapid development and widespread dissemination of deepfake techniques has caused increased concern. The malicious application of deepfake techniques also poses a potential threat to the society. Therefore, how to detect deepfake content has become a popular research topic. Most of the previous deepfake detection algorithms focused on capturing subtle forgery traces at pixel level and have achieved some results. However, most of the deepfake algorithms ignore the lighting information before and after generation, resulting in some lighting inconsistency between the original face and the forged face, which provided the possibility of using lighting inconsistency to detect deepfake. A corresponding algorithm was designed from two perspectives: introducing lighting inconsistency information and designing a network structure module for a specific task. For the introduction of lighting task, a new network structure was derived by designing the corresponding channel fusion method to provide more lighting inconsistency information to the network feature extraction layer. In order to ensure the portability of the network structure, the process of feature channel fusion was placed before the network extraction information, so that the proposed method can be fully planted to common deepfake detection networks. For the design of the network structure, a deepfake detection method was proposed for lighting inconsistency based on patch-similarity from two perspectives: network structure and loss function design. For the network structure, based on the characteristic of inconsistency between the forged image tampering region and the background region, the extracted features were chunked in the network feature layer and the feature layer similarity matrix was obtained by comparing the patch-wise cosine similarity to make the network focus more on the lighting inconsistency. On this basis, based on the feature layer similarity matching scheme, an independent ground truth and loss function was designed for this task in a targeted manner by comparing the input image with the untampered image of this image for patch-wise authenticity. It is demonstrated experimentally that the accuracy of the proposed method is significantly improved for deepfake detection compared with the baseline method.

Keywords: Deepfake detection, lighting inconsistency, color constancy, patch-similarity comparison

0 引言

随着深度生成模型的发展,如生成对抗网络(GAN)^[1]和变分自动编码器(VAE)^[2],深度伪造技术成为近期热门的研究课题。逼真的深度伪造图像或视频对人眼来说是难以分辨的,它们很容易被滥用于恶意的目的,最终导致不良的后果,如欺诈、诽谤和虚假新闻^[3-6]。

由不完善的合成算法产生的早期伪造人脸会引入明显的视觉伪影,许多基于人工设计特征的检测算法很容易将其区分出来^[7-10]。一些算法将原始的RGB图像作为输入,并以二分类输出方式自动训练检测模型^[11-12]。在这种方式下,模型很可能过拟合于特定的数据集,而当面对复杂多变的实际情况时,表现却不尽如人意。因此,学者设计了新的网络结构,引入了更有效的模块,如注意力机制^[13]、纹理特征^[14-15]、音视频相关性^[16]和频谱^[17],为深度伪造检测捕捉更有力的特征。

然而,这些算法大多是以手工特征的方式设计的,忽略了原始人脸和伪造人脸之间潜在的语义不一致性。

大多数的深度伪造技术没有考虑光照的因素。如图1所示,换脸前/后的照明变化可以明显看出,深度伪造换脸不能处理光照细化问题,在交换后的脸部(保留源的光照信息)和目标脸部之间留下明显的光照不一致。尽管伪造的人脸可以交换源图像的身份并保留目标的属性,但人脸区域的光照信息条件被交换了,这导致人脸不自然的效果。这种异常的光照通常不被现有的检测算法所捕获,而这可能是深度伪造检测的有效辅助信息。

本文提出了基于块间光照不一致性的深度伪造检测算法,它利用人脸图像的光照信息来辅助检测。本文将色彩恒常性应用于人脸图像,获得预处理后的人脸图像,并提取光照和反射引起的全局光照信息,此光照信息将被直接送入主干网

作为训练的辅助信息。此外，本文针对这一任务专门设计了块间相似性对比模块，使所提算法能够将深度伪造检测网络分类器的检测点聚焦在光照不一致性上。

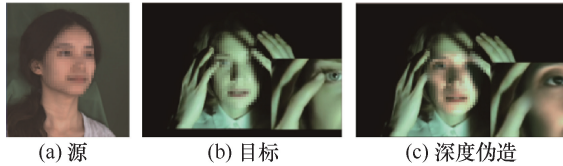


图1 换脸前后的照明变化示意

Figure 1 Before/after deepfake lighting change diagram

为了证明基于块间光照不一致性的深度伪造检测算法的有效性，本文在 FaceForensics++^[12]、Celeb-DF^[13]和 DFDC^[18]数据集上进行了大量的实验，可视化结果表明本文算法有很好的可解释性。

本文的主要贡献如下。

1) 提出了基于块间光照不一致性的深度伪造检测算法，该算法将语义光照信息作为辅助输入，块间相似性对比作为网络附加模块来提高检测准确率。

2) 分析了原始人脸和伪造人脸之间光照差异的可视化结果，所提算法表现出良好的可解释性。

3) 实验结果表明，所提算法在各种数据集上的表现优于基线算法。

1 相关工作

1.1 深度伪造

基于深度生成技术，深度伪造技术在近几年取得了巨大的突破。目前，深度伪造的算法大致分为两种：面部交换算法和脸部属性编辑算法。面部交换算法通常在源和目标之间交换脸部信息，而脸部属性编辑算法主要是修改部分脸部属性，如将脸部表情和姿势从一个肖像转移到另一个肖像。Face2Face^[19]是一种典型的脸部属性编辑算法，它从源图像中捕捉脸部表情并将其转移到目标图像中。DeepFakes^[20]和 FaceSwap^[21]是两种基于深度生成模型的人脸互换算法。FaceShifter^[22]提出了两阶段的高保真和闭塞感知的人脸互换框架。Head2Head^[23]对特定的人像视频使用 3D 建模，使姿势和表情可控，用于面部动作重现。

为了促进深度伪造检测算法的发展，许多学者

公布大量的公开数据集，如 FaceForensics++^[12]、Celeb-DF^[13]和 DFDC^[18]。然而，这些数据集是针对高保真度的深度伪造视频提出的，它们通常关注与修复细微的像素级别的视觉伪影，忽略了全局光照信息的一致性。AOT^[24]首次注意到光照条件在产生真实的深度伪造人脸中的重要性。它采用了外观最优传输模型来修复身份互换过程中源人像和目标人像之间的光照度和皮肤颜色的外观差距，且并没有被广泛用于深度伪造数据集。

1.2 深度伪造检测

深度伪造技术存在潜在的社会安全隐患，如何准确有效地检测深度伪造是网络安全领域中的一大重要任务。为了解决这个问题，学者提出了许多算法^[25-29]，如利用视觉上的伪影，即部分手工设计的特征，如眨眼、不一致的头部姿势^[7]、细微的面部表情变化^[27]等。XceptionNet^[12]是直接通过提取空间特征，来进行深度伪造检测的。多注意力机制^[13]首次将细粒度分类任务引入深度伪造检测中，并在特定数据集上获得了超越 baseline 的表现。Face X-ray^[28]观察了深度伪造产生的混合边界伪影，并取得了出色的可迁移性表现。空间相位浅层学习^[29]专注于被伪造图片的频域，在可迁移性上取得了令人信服的表现。然而，几乎所有的算法只是关注像素级别的伪影，光照不一致性这一语义信息往往被忽略。现有的深度伪造检测算法在光照的语义层面表现不佳，因此光照的语义信息成为深度伪造检测任务中有用的辅助信息。

1.3 色彩恒常性

人类有一种不因光源或外部环境因素而改变对特定物体的颜色判断的心理倾向，这种倾向就是色彩恒常性。计算机视觉中的色彩恒常性处理则是类比人类的心理特征产生的仿真算法，可以将原始图像中的光源及外部环境对图像的影响去除，完成光照信息的提取。

现已有多种基于传统数理统计的色彩恒常性算法被提出，如 White-Patch 算法^[30]、灰度世界假说法^[31]、灰度边缘算法^[32]、灰度阴影算法^[33]、亮度像素法^[34]、灰度像素法^[35]和灰度指数算法^[36]。随着时间的推移，学者提出了大量基于网络的色彩恒常性算法。

色彩恒常性提取算法大致可以分为两类：基

于神经网络学习的算法和基于传统数理统计的算法。基于神经网络学习的算法在某些特定任务上表现得更好,因为它能更好地拟合特定任务下的角度误差特征,然而,这一角度误差特征在深度伪造检测算法中并不重要,基于传统数理统计的算法反而拥有更好的可扩展性与可迁移性。另外,对于运用和计算来说,基于传统数理统计的算法能够以更小的计算代价完成对大部分深度伪造检测算法的辅助和移植。所以,本文选择使用基于传统数理统计的自动色彩均衡算法^[37]作为光照信息的提取算法。

2 基于光照不一致性的深度伪造检测

在传统的深度伪造检测算法中,光照信息的细微纹理特性是不被考虑的,通常在完成深度伪造重构后的图像容易产生光照层面的伪像。因而,引入光照信息的不一致性有助于提升深度伪造检测器的检测性能,包括但不限于检测的准确性、检测的可解释性以及检测的可迁移性。如何高效地获取可信的光照信息以及有效地使用额外产生的光照信息是难点。

本文将深度神经网络检测深度伪造流程分为两部分:输入部分和计算部分。第一部分引入额外的光照信息作为深度神经网络的前级辅助输入;第二部分进行针对性的修改,约束深度神经网络更聚焦在深度伪造图像的光照不一致性上。

2.1 光照不一致性信息获取

2.1.1 光照信息提取算法

本文获取的深度伪造图像的光照信息,即通过图像提取出原始场景中的光照信息。色彩恒常性提取作为计算机视觉中的一种方案,能够还原出物体在环境光照影响后信息原本的真实颜色,因此,选用色彩恒常性算法提取光照信息。

自动色彩均衡算法源自 Retinex^[38]算法,主要考虑在图像的基本颜色和亮度之间的相对空间关系,在局部进行特征自适应的滤波,从而完成对环境光照条件影响的物体颜色和对比度的调整。

首先,对图像的空间域进行颜色调整,完成图像的色差校正,得到在空间域上的重构图像。

$$R_c(p) = \sum_{j \in \text{Subset}, j \neq p} \frac{r(I_c(p) - I_c(j))}{d(p, j)} \quad (1)$$

其中, R_c 为中间结果, $I_c(p) - I_c(j)$ 为两个不同的像素点之间的亮度差, $d(p, j)$ 为距离测量函数。 $r(*)$ 为亮度奇函数, $r(*)$ 可以将差距小的部分放大,同时将差距大的部分对应缩小,并根据局部的特定内容选择扩大或缩小动态范围。 $r(*)$ 可表示为

$$r(x) = \begin{cases} 1, & x < -T \\ x/T, & -T \leq x \leq T \\ -1, & x > T \end{cases} \quad (2)$$

在完成对图像的空间域重构之后,还需要对图像进行动态的扩展。鉴于一般的图像均含有 R、G、B 这 3 个通道,需要对这 3 个通道分别进行动态扩展的处理。

本文使用的动态扩展算法的表达式如下。

$$O_c(p) = \text{round}[127.5 + s_c R_c(p)] \quad (3)$$

其中, s_c 为线段 $[(m_c, 0), (M_c, 255)]$ 的斜率。

$$\begin{aligned} M_c &= \max[R_c(p)] \\ m_c &= \min[R_c(p)] \end{aligned} \quad (4)$$

动态扩展完成后,通过自动色彩均衡算法对输入图像完成一次色彩恒常性处理,从而得到不含环境光照影响的增强图像。

2.1.2 光照信息提取分析

本文算法能够很容易地通过观察捕捉到部分深度伪造算法对于光照信息重构不完全而留下的特殊伪像,在成对的情况下进行对比,区别尤为明显。然而,如何捕捉到细微的光照信息,通过分析得到更具广泛性的光照不一致性,是需要探讨的问题。本节通过深入分析深度伪造图像与原始图像在提取之后光照信息上的不同,说明提取的光照信息在统计学上存在可分的特征。

本文引入基于颜色的视觉增强算法来使光照信息的视觉特征更容易被发现,算法如下。

算法 1 视觉增强算法

输入 原始图像

输出 视觉增强后的图像

function enhancement(RGB input)

result $\leftarrow$ 0

for every pixel do

result.r $\leftarrow$ 255 - input.r*factor

result.g $\leftarrow$ 255 - input.g*factor

```

    result.b ← 255 - input.b*factor
  end for
  return result
end function

```

本文对由自动色彩均衡算法进行色彩恒常性处理后的图像与原始图像绝对值做差，得到差分的光照信息，同时使用上述算法进行颜色增强。如图2所示，第三行（即颜色增强后的图像）的原始图像与深度伪造后的图像存在显著区别。尤其当把注意力聚焦在面颊这一特定区域，不难发现，深度伪造后的图像比原始图像红色部分更少。事实上，图中第三列中出现的颜色信息，是在色彩恒常性处理中被去除的部分，这说明在深度伪造的图像中，本应存在的非线性光照信息被破坏，并没有在色彩恒常性处理时认定为光照信息。

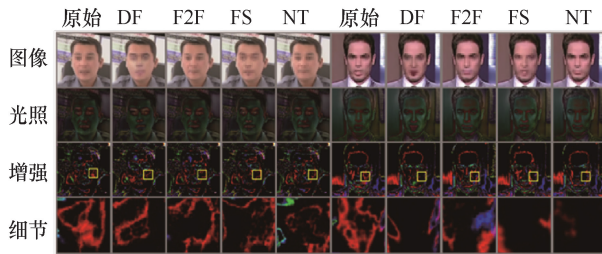


图2 对于光照不一致性的分析
Figure 2 Analysis on the lighting inconsistency

2.1.3 光照信息的辅助引入

在获取光照信息后，需要将这光照信息引入传统的深度伪造检测流程中。在对原始图像进行色彩恒常性处理后，即得到原始图像与去除光照

信息的原始图像。为了获取准确的光照信息，对这两张图像绝对值做差，从而得到由原始图像得出的差分光照信息。至此，获得了两组输入：一是用于检测的原始图像；二是基于原始图像提取出的差分光照图像。

接下来将对这两组输入进行拼接。由于这两组图像均为 R、G、B 这 3 个通道图像，将它们进行横向的拼接，并将拼接后的矩阵输入深度伪造检测器分类模型。需要注意的是，该模型的第一层卷积的输入由原始的 3 通道改为 6 通道，这意味着有额外 3 个通道的 RGB 信息被引入深度伪造检测器的分类模型中。如图3所示，上下两张图分别为分类器提供了 3 通道总计 6 通道的信息，这一步为传统深度伪造检测模型提供了额外的光照信息。

通过通道拼接的方式，为深度伪造检测分类器提供额外的光照信息用于辅助检测，是一种可以被广泛应用于各类检测分类器的算法。这是因为本文对光照信息的使用可以被看作一种特殊的数据增强，拥有非常强的泛化性。在这样的情况下，成功地将传统深度伪造检测中常常被忽视的光照信息引入了绝大多数检测分类架构中。

引入辅助的光照信息，有助于检测器更聚焦特殊的光影特征，从而在检测任务中提升检测器的检测准确性。

2.2 块间相似性对比模块

不同于对深度伪造检测分类器输入进行针对

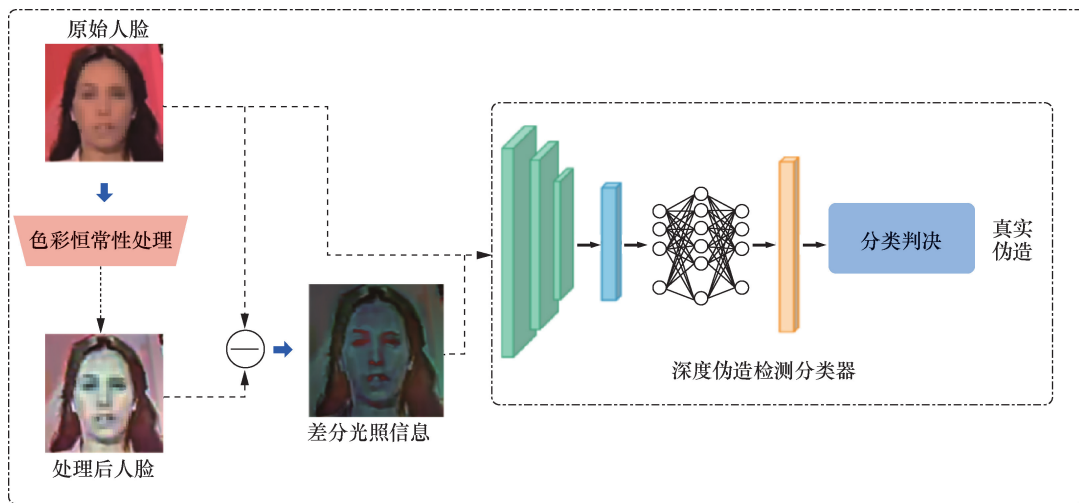


图3 引入光照信息的深度伪造检测网络
Figure 3 Deepfake detecting network with illumination information

性更改, 本节为了提升传统深度伪造检测分类器的性能, 设计了一个单独的模块。这是因为在不同的深度伪造检测方案中, 往往会引入不同的分类器架构, 且这些分类器架构在网络结构、网络层数上均存在不同, 所以如何设计出一种较为通用且有效的网络特征模块是关键。

2.2.1 块间相似性网络模块设计

在深度伪造图像中, 存在真实的区域(背景)和被深度伪造的区域(面部), 而光照不一致性将基于这两者之间存在的差异辅助于深度伪造图像检测。因而, 引入图像块间相似度对比的概念。

具体来说, 设计一个多层的相似性对比模式来明确地模拟不同局部区域之间的对应关系, 并利用成对的关系, 通过余弦相似度建立一个余弦相似度对应表。在这个相似度对应表中, 真实的区域彼此相似, 深度伪造区域也是如此, 但真实的区域和深度伪造的区域彼此并不相似。这一模式能够更好地捕捉到光照不一致性导致的“区域性”差异, 从而使得深度伪造检测分类器对深度伪造更敏感, 并能够提供更好的语义可解释性。

以 XceptionNet 为例, 对该深度伪造检测分类器的特征提取层分层采样。此处令 $\tilde{U}^l \in \mathbf{R}^{H^l \times W^l \times C^l}$ 表示网络分类器在第 l 层的输出特征图。如图 4 所示, 分别在 XceptionNet 的 Entry Flow、Middle Flow 和 Exit Flow 3 层的卷积层之后对特征图进行采样。由于这 3 层分别对应整体卷积活动中的浅层、中层和深层, 用 $\tilde{U}^{\text{low}}$ 、 $\tilde{U}^{\text{mid}}$ 和 $\tilde{U}^{\text{high}}$ 来进行标记。

由于卷积的不断进行, 这 3 次采样所取得的特

征图在矩阵形状上存在不一致, 以 Exit Flow 卷积后的特征图 $\tilde{U}^{\text{high}}$ 大小作为标准, 将前两次采样的特征图进行降维, 使得 3 次对特征图的采样均为同样大小的向量矩阵, 并将这 3 个特征图在第一维向量特征上进行拼接, 得到块间相似性特征图。选择多次采样并降维合并, 主要原因如下。

1) Entry Flow、Middle Flow 和 Exit Flow 实际上代表了网络的浅层、中层和深层, 网络提取到的特征在这几层有着相应的不同, 在更浅的层中, 网络倾向于提取并学习的特征, 是更细粒度的信息; 而在更深的层中, 网络倾向于提取并学习的特征是更整体的信息。在网络捕捉的全过程中均加强其对于光照信息的捕捉, 在细粒度上, 希望其对于光照的特征像素更敏感; 在整体上, 希望其对于光照的语义含义有更好的拟合。

2) 在块间相似性比较的过程中, 引入块间的运算, 这一运算将获取的特征图的特征维度以平方的形式进行, 会加大计算资源的计算量和内部存储空间, 因而在保留一定数据量的前提下, 如何合理地将捕捉的特征图进行降维, 是非常重要的。本文采用的算法不仅将前级特征降维, 而且将 3 次特征合并为一次, 有效降低了计算资源的负载, 提升了网络拟合的效率。

为了完成块间相似性的网络特征提取, 在完成对综合特征图 $\tilde{U}^{\text{total}}$ 的提取后, 分为若干个小块, 每个小块 u_i 满足 $\tilde{u}_i \in \mathbf{R}^{\tilde{h} \times \tilde{w} \times \tilde{c}}$, 其中 $\tilde{c}$ 为 $\tilde{U}^{\text{high}}$ 的通道数量。对于每一个小块 $\tilde{u}_i$ 和 $\tilde{u}_j$ 进行块间相似度对比, 使用余弦相似度计算, 即

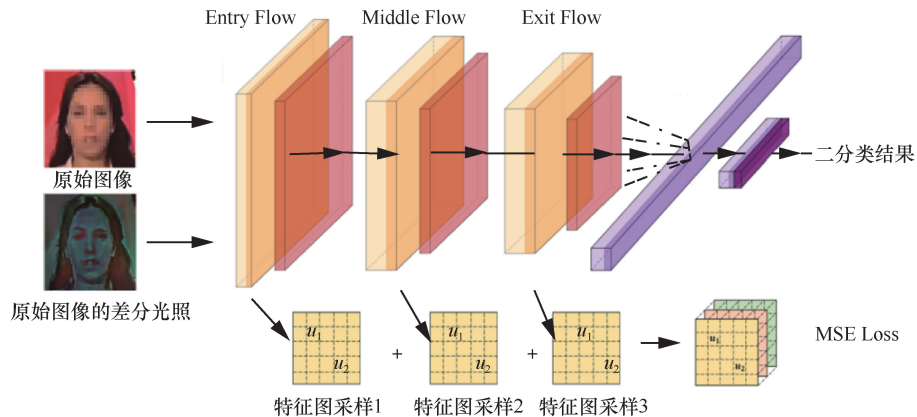


图 4 块间相似性比较采样流程
Figure 4 Sampling procedure of patch-wise similarity comparison

$$\hat{S}_{i,j} = \frac{\left\langle \frac{u_i}{\|u_i\|_2}, \frac{u_j}{\|u_j\|_2} \right\rangle + 1}{2} \quad (5)$$

通过对每一个小块的遍历，可以得到特定特征融合图的余弦相似度信息 $\hat{S}$ ，此信息由网络的浅层、中层、深层对应的特征层合并后提取，能够较好地表达深度伪造检测网络中提取的特征信息的块间相似性。

2.2.2 块间相似性损失函数设计

为了将获取的特征图余弦相似度信息作为网络反向传导收敛的一部分，为块间相似性对比设计了一套损失函数模型。基于块间相似性，希望得到的内容满足：真实的区域彼此相似，深度伪造区域也是如此，但真实的区域和深度伪造的区域彼此并不相似。所以用于损失函数对比的真实数据参考应当更多聚焦于深度伪造人脸与原始人脸对比后被篡改的区域。因此，本节设计了特有的真实数据参考获取模式。

在数据集中，将原始的真实人脸图像与该图像衍生出的深度伪造图像两两配对成一组，针对每一张伪造图像，都可以溯源至真实图像。此时，将对深度伪造后的图像与其原始人脸图像做逐像素的差分处理，取绝对值后得到逐像素的差分图像。

由于需要获取的是面部被深度伪造篡改后的区域特征标记信息，因而获取的三维 RGB 图像存在信息冗余，将它从三维的 RGB 信息转换为灰度信息。为了方便计算，选择将这一灰度图转换为二进制的面具信息，从而标记出被深度伪造的人脸图像中被篡改的像素点位置。根据经验，在实际测试中使用 0.15 作为判定的阈值，即令在灰度图中超过 0.15 阈值的像素点被标记为篡改区域，得到面具 M_{xy} ，如下所示：

$$M_{xy} = \begin{cases} 0, & \text{该像素点为真实} \\ 1, & \text{该像素点为伪造} \end{cases} \quad (6)$$

其中， x, y 为像素点所在坐标。

本文希望深度伪造分类器更关注块间光照相似度的信息，因而，在损失函数的真实数据参考部分也应引入余弦相似度信息。

将获取的面具分为 N 个小块，此处的 N 对

应式(5)获取到的特征图余弦相似度信息矩阵大小。按照顺序将每一个小块记作 m_x ，每个像素小块对应的可能造假概率 p_x ，即该区域内所有像素点所含值的平均数。最后，计算出每一个小块被造假概率之间的欧几里得距离，得出用于指导网络拟合的最终真实数据参考面具。

本文为块间相似性对比部分损失函数设计了独有的真实数据参考，如图 5 所示。基于本模块中捕捉的特征图与真实数据参考面具均为二维信息，采用二维的 MSELoss 对这二者进行损失的计算，块间相似性损失 L_{sim} 计算式如下。

$$L_{sim} = \|\mathbf{S} - \hat{\mathbf{S}}\|_2 \quad (7)$$

其中， $\mathbf{S}$ 为原始图像和伪造图像差分得到的真实数据参考矩阵， $\hat{\mathbf{S}}$ 为由深度伪造检测器网络中提取特征计算出的块间余弦相似度矩阵。

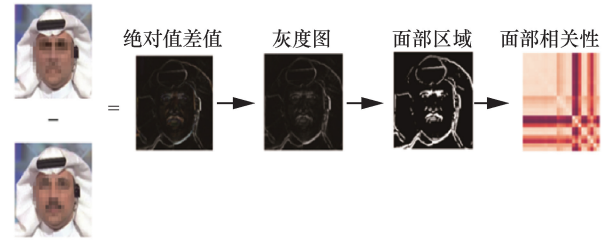


图 5 块间相似性真实数据参考
Figure 5 Ground truth of the patch similarity

至此，本文提出并构造了块间相似性对比网络结构并构造了相应的损失函数，且这一损失函数是独立于原始网络结构中用于判决的二分类损失函数的，也就意味着，所提算法在引入块间相似性对比后，仍然延续了原有的“可移植性”，即便于移植至绝大多数常见的基于深度神经网络进行深度伪造检测的算法。在最终损失函数构造时，可以调节块间相似度对应的损失函数超参数 λ 来得到最终的损失函数。

3 实验与结果

3.1 在常见数据集上的实验结果

本节对基于块间光照不一致性的深度伪造检测算法与基线算法——XceptionNet 算法在不同的视频压缩比下进行了对比。

如表 1 所示，本文算法在特定任务的深度伪造检测准确率上有较为明显的提升，尤其是

在 NT_c23/NT_c40 这种相对含有更少基于像素纹理伪造痕迹的数据集上。这意味着本文算法相较于传统的基于深度神经网络的深度伪造检测算法,能够更好地捕捉到细微的基于语义的伪造痕迹信息。这一结论也可以在低分辨率数据集的实验结果中得出,本文算法能够获得比在高分辨率数据集中更高的准确度提升。这是因为,在低分辨率视频中,视频经过了多次高强度的压缩,常见的深度神经网络聚焦的像素纹理伪造痕迹被大幅度减弱。而本文算法能够获得如此高的检测准确率,是因为本文算法能够更好地捕捉基于光照不一致性所带来的语义层面的伪造痕迹。

表 1 本文算法与 XceptionNet 算法的准确率对比

Table 1 Comparison of accuracy between the proposed algorithm and XceptionNet algorithm

数据集	准确率	
	XceptionNet 算法	本文算法
DF_c23	98.69	99.01
NT_c23	93.67	94.79
F2F_c23	98.7	98.71
FS_c23	99.08	98.94
Total_c23	92.83	96.03
DF_c40	95.21	97.83
NT_c40	78.02	81.3
F2F_c40	90.66	90.92
FS_c40	91.99	93.24
Total_c40	84.75	86.19

将本文算法与隐写分析算法^[39]进行了对比,从表 2 中可以发现,本文算法在检测准确率上优于使用隐写分析算法。这是因为本文算法能够更好地针对深度伪造中缺失的光照一致性进行捕捉,为检测分类器提供分类依据。

此外,本文算法并不会影响原始网络结构,是便于移植的算法。在表 3 中,将色彩恒常性辅助模块移植到其他深度伪造检测分类器模型上,可以看到本文算法准确率得到提升。需要注意的是,这并不意味着块间相似性模块是不便于移植的部分,对于不同的网络,块间相似性模块需要

捕捉的不同深度网络层有所不同,为控制变量唯一,使用了这样的移植比较实验。

表 2 本文算法与隐写分析算法的准确率对比

Table 2 Comparison of accuracy between the proposed algorithm and steganalysis algorithm

数据集	准确率	
	隐写特征算法	本文算法
DF_c23	83.61	99.01
NT_c23	73.92	94.79
F2F_c23	83.75	98.71
FS_c23	79.67	98.94
Total_c23	76.74	96.03
DF_c40	76.6	97.83
NT_c40	68.9	81.3
F2F_c40	76.72	90.92
FS_c40	71.26	93.24
Total_c40	70.72	86.19

表 3 本文算法与 ResNet-50 算法的准确率对比

Table 3 Comparison of accuracy between the proposed algorithm and ResNet-50 algorithm

数据集	准确率	
	ResNet-50	本文算法
DF_c23	97.55	97.68
NT_c23	91.33	92.16
F2F_c23	98.25	98.14
FS_c23	97.94	97.53
Total_c23	92.06	95.41
DF_c40	96.42	97.26
NT_c40	76.58	78.65
F2F_c40	88.36	88.92
FS_c40	88.69	90.84
Total_c40	84.11	85.47

为进一步证明本文算法的有效性,在 Celeb-DF 和 DFDC 数据集上将本文算法与 XceptionNet 算法进行了对比。与 FaceForensics++ 相比,这两个数据集的视频具有更高的视频质量,同时更难被检测。DFDC 是最具挑战性的 Deepfake 检测数据集,因为其中的假视频很难被人眼识别。表 4 中的结果表明,本文的算法在 Celeb-DF 和 DFDC 数据集上获得了更好的性能。

表4 其他数据集上本文算法与
XceptionNet 算法的准确率对比

Table 4 Comparison of accuracy between the proposed algorithm and XceptionNet algorithm on other datasets

数据集	准确率	
	XceptionNet 算法	本文算法
DFDC	78.26	80.47
Celeb-DF	98.24	98.96

3.2 消融实验对比

本文算法主要可以分为辅助光照信息引入及块间相似度对比两个部分, 本文设计了相应的消融实验。同样地, 在 FaceForensics++ 的高分辨率数据集和低分辨率数据集中与 XceptionNet 进行了对比实验。

在实验中可以发现, 无论是辅助光照信息还是块间相似度对比模块, 在整体上均能明显提升深度伪造检测分类器的分类性能或维持分类器原本水平, 如表 5 所示, 这表明本文算法能够明显提升深度伪造检测网络的分类准确率。在 NeuralTextures 等语义信息较为丰富的数据集以及低分辨率数据集中, 块间相似性对比模块对检测准确率带来的帮助比光照信息的引入所带来的帮助小, 这是因为光照信息直接为深度伪造检测分类器带来了额外的光照不一致性信息补充, 而块间相似性比较则使得网络更聚焦于原始区域与被伪造区域之间的差异。

表5 消融实验下本文算法与
XceptionNet 算法的准确率对比

Table 5 Comparison of accuracy between the proposed algorithm and XceptionNet algorithm in ablation experiment

数据集	准确率			
	XceptionNet	Light	Sim	Light+Sim
DF_c23	98.69	98.92	98.7	99.01
NT_c23	93.67	94.37	94.1	94.79
F2F_c23	98.7	98.34	98.68	98.71
FS_c23	99.08	97.75	99.02	98.94
Total_c23	92.83	95.74	94.86	96.03
DF_c40	95.21	97.41	96.07	97.83
NT_c40	78.02	80.37	78.14	81.3
F2F_c40	90.66	90.68	90.47	90.92
FS_c40	91.99	92.93	92.26	93.24
Total_c40	84.75	85.52	85.37	86.19

4 结束语

本文提出了基于块间光照不一致性的深度伪造检测算法。本文工作相比以往的深度伪造检测算法, 由于能够更多地捕获来自光照不一致性所带来的潜在语义特征, 在准确率上获得了更好的表现。此外, 本文提出的块间相似性模块能够使深度伪造检测分类器更多地关注深度伪造人脸区域的光照不一致性, 从而进一步提升算法的鲁棒性。实验证明所提算法在深度伪造检测, 尤其在低分辨率检测上有明显的准确率提升。

此外, 本文算法着重研究的是深度伪造生成算法中缺少对光照信息的约束而导致的 inconsistency。在攻防升级的未来, 深度伪造生成算法将会逐步将光照信息加入质量约束中, 这也是可以延伸的下一步工作, 即通过被篡改的生成图片中的“人为伪造光照”痕迹来进行鉴别。

参考文献:

- [1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [2] HUH M, LIU A, OWENS A, et al. Fighting fake news: Image splice detection via learned self-consistency[C]//Proceedings of Computer Vision – ECCV 2018. 2018: 106-124.
- [3] KWON P, YOU J, NAM G, et al. KoDF: A large-scale Korean deep-fake detection dataset[J]. arXiv preprint arXiv:2103.10094, 2021.
- [4] LEE S. Detecting handcrafted facial image manipulations and GAN-generated facial images using shallow-fakefacenet[J]. Applied Soft Computing, 2021, 105: 107256.
- [5] NGUYEN T T, NGUYEN C M, NGUYEN D T, et al. Deep learning for deepfakes creation and detection: a survey[J]. arXiv: 1909.11573, 2019.
- [6] SUN K, LIU H, YE Q X, et al. Domain general face forgery detection by learning to weight[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2638-2646.
- [7] YANG X, LI Y Z, LYU S W. Exposing deep fakes using inconsistent head poses[C]//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 8261-8265.
- [8] ZHOU P, HAN X T, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 1831-1839.
- [9] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: A com-

- pact facial video forgery detection network[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. 2018: 1-7.
- [10] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 2307-2311.
- [11] LI Y Z, LYU S W. Exposing DeepFake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.
- [12] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: manipulated facial images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 1-11.
- [13] ZHAO H, ZHOU W, CHEN D, et al. Multiattentional deepfake detection[J]. arXiv preprint arXiv:2103.02406, 2021.
- [14] LIU Z Z, QI X J, TORR P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 8057-8066.
- [15] SUN X W, WU B T, CHEN W. Identifying invariant texture violation for robust deepfake detection[J]. arXiv preprint arXiv: 2012.10580, 2020.
- [16] CHUGH K, GUPTA P, DHALL A, et al. Not made for each other audio visual dissonance-based deepfake detection and localization[C]//Proceedings of MM'20: Proceedings of the 28th ACM International Conference on Multimedia. 2020: 439-447.
- [17] ZHANG X, KARAMAN S, CHANG S F. Detecting and simulating artifacts in GAN fake images[C]//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security. 2019 1-6.
- [18] DOLHANSKY B, HOWES R, PFLAUM B, et al. The DeepFake detection challenge (DFDC) preview dataset[EB].
- [19] THIES J, ZOLLHÖFER M, STAMMINGER M, et al. Face2Face: real-time face capture and reenactment of RGB videos[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2387-2395.
- [20] DeepFakes[EB].
- [21] FaceSwap[EB].
- [22] LI L Z, BAO J M, YANG H, et al. FaceShifter: Towards high fidelity and occlusion aware face swapping[EB].
- [22] LI L Z, BAO J M, YANG H, et al. FaceShifter: Towards high fidelity and occlusion aware face swapping[J]. arXiv preprint arXiv: 1912.13457, 2019.
- [23] KOUJAN M R, DOUKAS M C, ROUSSOS A, et al. Head2Head: Video-based neural head synthesis[C]//Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. 2020: 16-23.
- [24] ZHU H, FU C, WU Q, et al. AOT: appearance optimal transport based identity swapping for forgery detection[C]//NeurIPS, IEEE, 2020.
- [24] ZHU H, FU C Y, WU Q Y, et al. AOT: appearance optimal transport based identity swapping for forgery detection[C]//Proceedings of NeurIPS. 2020.
- [25] MASI I, KILLEKAR A, MASCARENHAS R M, et al. Two branch recurrent network for isolating deepfakes in videos[C]//Proceedings of Computer Vision – ECCV 2020, 2020: 667-684.
- [26] WU X, XIE Z, GAO Y T, et al. SSTNet: detecting manipulated faces through spatial, steganalysis and temporal features[C]//Proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. 2020: 2952-2956.
- [27] SHRUTI AL, HANY F, YUMING G, et al. Protecting world leaders against deep fakes[C]//Proceedings of CVPR Workshops. 2019.
- [28] LI L Z, BAO J M, ZHANG T, et al. Face X-ray for more general face forgery detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 5000-5009.
- [29] LIU H G, LI X D, ZHOU W B, et al. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 772-781.
- [30] BRAINARD D H, WANDELL B A. Analysis of the retinex theory of color vision[J]. Journal of the Optical Society of America A, Optics and Image Science, 1986, 3(10): 1651-1661.
- [31] BUCHSBAUM G. A spatial processor model for object colour perception[J]. Journal of the Franklin Institute, 1980, 310(1): 1-26.
- [32] VAN DE WEIJER J, GEVERS T, GIJSEN J A. Edge-based color constancy[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2007, 16(9): 2207-2214.
- [33] FINLAYSON G D, TREZZI E. Shades of gray and colour constancy[C]//Proceedings of Color and Imaging Conference. 2004.
- [34] JOZE H R V, DREW M S, FINLAYSON G D, et al. The role of bright pixels in illumination estimation[C]//Proceedings of Color and Imaging Conference. 2012.
- [35] YANG K F, GAO S B, LI Y J. Efficient illuminant estimation for color constancy using grey pixels[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2254-2263.
- [36] QIAN Y L, KÄMÄRÄINEN J K, NIKKANEN J, et al. On finding gray pixels[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 8054-8062.
- [37] GATTA C, RIZZI A, MARINI D. ACE: an automatic color equalization algorithm[C]//Proceedings of Conference on Colour in Graphics. 2002.
- [38] LAND E H, MCCANN J J. Lightness and retinex theory[J]. Journal of the Optical Society of America, 1971, 61(1): 1-11.
- [39] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[作者简介]



吴文轩（1997- ），男，浙江金华人，中国科学技术大学硕士生，主要研究方向为深度伪造检测。



张卫明（1976- ），男，河北定州人，中国科学技术大学教授、博士生导师，主要研究方向为信息隐藏、多媒体内容安全、人工智能安全。



周文柏（1992- ），男，安徽合肥人，中国科学技术大学特任副研究员，主要研究方向为信息隐藏与人工智能安全。



俞能海（1964- ），男，安徽无为市人，中国科学技术大学教授、博士生导师，主要研究方向为多媒体信息检索、图像处理与视频通信、数字媒体内容安全。