arXiv:2302.12133v1 [cs.MM] 23 Feb 2023

常见社交媒体平台和照片存储服务如何处理上传图像的实证分析

Duc-Tien Dang-Nguyen[1,2],Vegard Velle Sjøen[1],Dinh-Hai Le[3,4], Thien-Phu Dao[3,4], Anh-Duy Tran[3,4],和 Minh-Triet Tran[3,4]

[1] 挪威卑尔根大学[2] 挪威克里斯蒂安尼亚大学学院[3] 越南胡志明市科学大学[4] 越南国家大学胡志明市分校，越南

摘要。 本研究深入探讨了当今社会上传至三大社交媒体平台和图像存储服务的数字图像所发生的变化：Facebook、Flickr 和 Google Photos。除了提供有关不断变化的不同社交媒体网络数字指纹的最新数据外，对这些社交网络文件名惯例的深入分析还提出了两种新方法：(i) 估算 Flickr 照片的真实上传日期，无论日期是否曾被用户更改，也无论图像是否可供公众访问或已从平台删除；(ii) 仅根据上传至 Facebook 的照片的文件名，揭示该照片的照片 ID。

关键词： 数字图像取证 · 图像指纹识别 · 图像来源验证 · 社交网络 · 命名约定

1 引言

多年来，多媒体研究界已经开发出各种数字图像取证工具和技术，用于图像篡改检测、 [21], 深度伪造检测 [12,22],以及来源识别[18], 等方面，旨在回答两个主要问题：“该图像是否由声称用于获取它的设备拍摄？”以及“该图像是否仍然呈现其原始内容？” [16]然而，这些工具和技术只有在实验室实验等受控环境中才能发挥良好效果，在实际环境中往往无法提供可靠的答案。导致这些工具性能不佳的一个关键因素，是社交媒体平台或在线存储服务对上传图像进行的处理 [15]。这一过程可能会通过压缩改变文件名、图像质量以及图像的属性。因此，对从社交网络中获取的图像进行取证分析时，可能会受到上述处理过程的影响。因此，了解上传图像经过了怎样的处理，将在提升数字取证工具的性能和准确性方面发挥关键作用。

在本研究中，我们旨在调查社交网络在上传过程中如何改变跨平台发布的图像，即，**研究图像上传至常见平台或照片存储服务的过程中会发生什么，以及图像文件会发生何种改变**。了解这些修改后，我们将尝试预测哪些数字取证和验证工具最适合用于追溯这些图像的来源，以及检测篡改和操纵。

研究表明，社交媒体平台通常会在上传的图像本身留下某些数字指纹 [7, 21]，这些指纹可能表现为根据时间码更改图像名称（如 Facebook），或在图像底部添加新的部分以表明其来源平台（如 Reddit）。这些特征可用于识别图像的真实来源。由于任何图像修改都会改变其属性，图像取证技术的有效性取决于图像经过了何种处理。本文收集的数据将用于确定如何运用不同技术来预测图像的上传来源。因此，能够识别图像经过了怎样的修改，在尝试运用不同的图像取证技术时会带来显著优势。

社交媒体和在线存储服务是普通用户发布日常生活图像和上传照片的热门场所。*Photutorial* [5] 的数据表明，2021 年全球拍摄了 1.2 万亿张照片。2022 年，Facebook 上每天分享 21 亿张图像，Flickr 上每天分享 100 万张，使 Facebook 和 Flickr 成为图像分享领域领先的社交媒体平台，与 WhatsApp、Instagram 和 Snapchat 并列 [5]。Google 相册是一款用于存储和管理个人照片与视频的在线应用。2020 年，Google 报告称，该平台的数据库中有超过 4 万亿张照片，每周新增 280 亿张图片和视频 [3]。根据这些数据，我们决定选取两个社交媒体平台 Facebook 和 Flickr，以及一项在线存储服务 Google 相册，研究上传图像的变化。考察这些受欢迎的平台有助于全面了解在线图像的处理方式。

本研究的主要贡献如下：

– 提供关于热门媒体网络和照片存储服务，尤其是 Facebook、Flickr 和 Google 相册如何处理用户上传图像的最新数据；– 提出一种新方法，用于估算已上传至 Flickr 平台的图像的实际上传日期；– 提出一种新方法，用于对上传至 Facebook 的图像的图像标识符进行逆向工程。

本文内容安排如下：第 2 节简要介绍背景信息和相关研究。第 3 节讨论研究方法。第 4 节报告研究结果。第 5 节探讨潜在应用，第 6 节则总结全文。

2 背景与相关工作

2.1 背景

数字图像取证（DIF）与“数字水印”协同工作，用于识别和对比恶意图像篡改 [21]。为完成数字图像取证的任务，可以采用从被动到主动的多种方法。最简单的一种方法是通过检查图像元数据进行分析。这类证据能够概述特定图像曾经过哪些处理，并为后续分析提供起点。

除了用于渲染数字图像的数据外，许多图像格式还会保留其他数据，用于描述文件的内容或特征。这类数据称为元数据。下面简要介绍一些常用的元数据：

– **Exif 数据**。可交换图像文件格式（Exif）是一项标准，用于规定由数码相机或移动电话拍摄的照片或其他媒体的详细信息。它还可以存储相机曝光、图像拍摄日期/时间，甚至 GPS 位置等关键信息。

– **IPTC 数据**。IPTC 元数据对于有效的图像管理和保存至关重要。对于新闻机构和记者、图像档案库以及其他需要完整呈现图像详细信息的组织而言，它尤其具有价值。与关注图像技术层面的 EXIF 数据不同，IPTC 标准关注照片的内容，以及其所有权、权利和许可状态。

– **XMP 数据**。可扩展元数据平台（XMP）是由 Adobe 开发的文件标记系统，可在文件开发流程的各个阶段向文件中插入元数据，例如标题和描述、可搜索的关键词，以及最新的作者和版权信息。

– **当前 IPTC 摘要**。Harvey 的 ExifTool [9] 提供一种哈希函数，可将任意大小的数据映射为固定大小的值。该值称为当前 IPTC 摘要。当前 IPTC 摘要的主要用途是与其他 IPTC 摘要进行比较，以判断文件是否已被修改。这些哈希值通过对旧版 IPTC IIM 值计算 MD5 哈希生成。随后，这些哈希值会与元数据中的 IPTC 摘要进行比较。然而，由于许多图片管理程序已转而采用 IPTC XMP 标准，IPTC IIM 格式已不再积极维护。

2.2 相关工作

本节探讨数字图像取证中的一些相关技术和挑战，包括平台溯源、来源验证和日期/时间验证。

– **平台溯源：**验证用户生成内容最重要的步骤之一是溯源。此过程旨在确定该内容最近上传至哪个互联网平台或社交网络。

Bharati 等人 [4] 提出了一种方法，利用非内容类信息提取特定图像在互联网上传播的路径，同时避免产生大量计算开销。Siddiqui 等人 [19] 提出了一种模型，利用基于深度学习的方法查找这些独特的痕迹，以识别图像的来源社交网络，重点在于确定特定图像是从哪个社交网络下载的。

– **来源验证：**虽然溯源可追溯至图像的首位上传者，但来源验证旨在确定其创建背景（例如，拍摄该照片的人）。这一过程在数字取证中也发挥着关键作用，因为上传者和创作者未必总是同一人。Kee 等人 [10] 分析了 JPEG 头部，以识别由量化表、霍夫曼编码、缩略图和 Exif 信息构成的相机特征签名。他们表明，该特征签名在涵盖 773 台相机和手机的 130 万张图像中具有高度区分性。Marra 等人 [11] 提出了一种盲法基于 PRNU 的图像聚类算法，利用噪声残差对图像来源进行聚类。Mullan 等人 [14] 研究了利用 JPEG 头部信息关联现代智能手机拍摄照片的可能性。他们的研究表明，Exif 元数据的变化与实际的苹果硬件关联较弱，而与 iPhone 操作系统 iOS 版本的变化关联更为密切。

– **日期与时间验证：**在数字取证中，图像拍摄或上传的时间可能是对事实核查人员有价值的信息 [13]。例如，某张图像被声称是在特定事件中拍摄的，但实际上它早在该事件发生之前就已上传。因此，这一说法是错误的。然而，事实证明，此类信息的提取难度要大得多 [17]。来源信息可以帮助确定图像首次上传的时间，但无法帮助确定相关材料是在何时收集的。虽然可以提取元数据，但其并不可靠，因为使用正确的应用程序可以轻易篡改这些数据 [1]。

3 种方法

本研究旨在分析 Facebook、Flickr 和 Google 相册如何处理上传图像：在将图像上传至社交媒体之前，通过进行各种操纵，构造大量特定场景。如图 1 所示，整个过程可分为两个主要阶段：上传至社交媒体，以及从社交媒体下载图像。第一阶段中，我们对特定图像进行裁剪、调整大小，或保留原图。随后，将这张（经修改的）图像转换为 PNG 格式，再进行压缩。最终图像通过多种方式上传至这三个平台，包括使用 iOS 和 Android 上的原生应用程序，以及通过不同个人电脑上的 Chrome 浏览器上传。我们还使用虚拟专用网络，考察从不同位置上传图像的影响。第二阶段中，我们通过上述平台和虚拟专用网络，从社交媒体下载对应图像，并与上传前的图像进行比较，以检查发生的变化。各项操纵过程的细节如下所述。

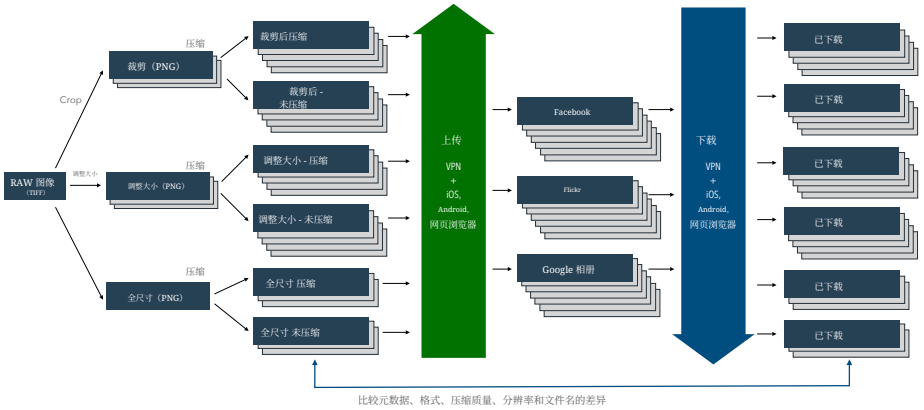


图 1. 分析流程。从 RAISE 1K 数据集 [6] 中选取一张 TIFF 格式的 RAW 图像，我们首先通过裁剪和调整大小生成不同的图像，并将格式转换为 PNG（以便上传至各个平台）。然后，使用不同的质量因子压缩每张新生成的图像。最后，通过 Android 和 iOS 应用程序，以及 PC 网页浏览器，将这些图像上传至不同的平台和服务。我们还尝试使用虚拟专用网络连接到不同国家的不同服务器，从不同的位置进行上传。随后，我们研究上传前后的差异。

– **裁剪：** 为分析在线平台对比原图经过裁剪的图像有何反应，我们以三种方式将图像裁剪至原始尺寸的 50%–80%：左上、右下和中心。

– **调整大小：** 与裁剪类似，我们也将图像尺寸调整为原始尺寸的 200%、150% 和 50%，以检查平台的反应。

– **压缩：** 我们还会调整图像质量，以考察社交媒体的压缩会如何处理上传的图像。我们将目标图像的质量因子设为 80、60、40 和 20 进行压缩。

– **虚拟专用网络：** 为检查社交媒体是否会根据位置改变其对上传图像的处理方式，我们使用虚拟专用网络操控图像的源位置。我们在不同大洲的多个国家进行了测试，包括越南、挪威、美国和澳大利亚。

4 结果

4.1 数据与实验设置

数据集： 我们使用了 RAISE 1K 数据集中的 15 张 RAW 图像 [6]。我们按照图 1 所述的流程，采用了五种不同的尺寸和分辨率（范围从 3430×2278 到 858×570），以及五种不同的压缩质量（未压缩、80、60、40 和 20）。最终，我们获得了 2475 对上传前后的图像。

工具： 随后使用 IBSNtools [20] 分析文件名、压缩表和分辨率，并将其与原始图像进行比较，同时检查 Exif 和

IPTC 数据（使用 ExifTool [9]）。我们分析了文件名，寻找可能提供图像额外信息的规律。我们分析了图像分辨率，以查看图像在上传过程中是否调整过大小或重新压缩。我们分析了 Exif 和 IPTC 数据，以查看元数据是否被删除或修改。

4.2 Google 相册上的结果

Exif。上传到 Google 相册后，所有 Exif 数据均保持完整。

IPTC 数据。使用 Google 相册上传 JPEG 图像且关闭压缩设置时，所有图像的 IPTC 数据均保持完整，因此其 IPTC 摘要哈希与原始图像完全一致。开启压缩设置后，所有图像都产生了新的且唯一的 IPTC 摘要哈希，这意味着它们仍然包含 IPTC 数据，但每张图像中的数据都经过了修改。

XMP 数据 上传图像并进行重新压缩时，Google 相册保留了大部分 XMP 数据。不进行重新压缩时，所有 XMP 数据都与原始数据完全一致。在这两种情况下，“原始文档 ID”和“保留的文件名”条目均被保留，并且与原始文件相匹配。

分辨率。所有图像的分辨率都与原始图像一致；无论 Google 相册中的压缩设置是启用还是禁用，图像尺寸均未被修改。

压缩。与其他社交网络不同，上传的 RAW 图像没有转换为 JPEG，也完全没有经过压缩，而是与原始文件完全相同。JPEG 文件经过了重新压缩，但由于 Google Photos 提供了完全不修改文件的选项，因此这取决于 Google 账户的设置。

文件名。所有图像的文件名也都保持不变。

4.3 Facebook 上的结果

Exif。在所有情况下，无论是 JPEG 还是 RAW 原始文件，Facebook 都会从图像中删除所有 Exif 数据，但“Artist”和“Copyright”标签除外。

IPTC 数据。上传 RAW 文件时，Facebook 会保留所有图像中相同的 IPTC 数据，因此所有图像的 IPTC 摘要都相同。JPEG 文件中的 IPTC 数据则被全部删除。

XMP 数据。上传 JPEG 和 RAW 图像时，Facebook 均未保留任何 XMP 数据。

压缩。在所有情况下，Facebook 都使用相似但略有差异的压缩表压缩图像。平均误差率小于 2%。

分辨率。Facebook 对尺寸较大的部分图像进行了调整大小，观察到的阈值为 2048。图像的宽度和高度均可能触发调整大小。这与 Giudice 等人在论文《社交数据图像弹道分类引擎》 [8]中的研究结果一致。

文件名。无论原始图像格式和用于上传图像的平台是什么，所有文件都按照特定的命名约定进行了重命名。随后，我们进一步研究了这些平台如何命名上传图像。

Facebook 命名约定： Facebook 使用 Haystack 照片基础设施 [2], 存储照片；当用户上传照片时，系统会为其分配一个唯一的 64 位 ID。随后，该标识符会与相册关联，为用户生成下载图像的链接。2012 年 7 月之前，Facebook 按照五个数字的模式为图像命名：

$$aa_bb_cc_dd_ee_n$$

其中，aa 是照片 ID（fbid），bb是相册 ID，cc 是上传图片用户的个人资料 ID；dd 和 ee 未公开。2012 年 7 月，Facebook 改用三个数字的模式：$aa_bb_ee_n$，即移除了个人资料 ID。

在本研究开展时，Facebook 上上传图片的名称遵循四个数字的模式：

$$xx_yy_zz_n_aa$$

其中，aa 仍然是照片 ID（fbid）；如果图像隐私设置为公开，或者某个帐户通过 Facebook 好友关系获得了访问权限，则可以使用该 ID 直接访问图像。图 2 展示了如何通过其 fbid 访问图片的示例。

如果用户通过在浏览器中右键单击并选择“另存为”来下载图像，他们只能获取文件名，fbid 会被移除：

$$xx_yy_zz_n$$

我们在 2475 对图像中观察到，第二部分（yy）的前六位数字与 fbid（aa）的前六位数字相同。为了探究这一发现，我们又分析了来自同一 Facebook 账户的 1998 张图像（即该用户的图像总数），以及来自另一个账户的 200 张图像，进行了另一项实验。结果证实了相同的规律。

我们提出一个假设：仅使用 yy 就可以恢复 aa。为了进一步测试这一假设，我们尝试通过一些运算比较这两个数字。用 yy 减去 aa （见图 2 中的图像），得到 6666666。对其他图像进行相同的减法运算后，得到的数字包括 3333333、6666666、9999999、13333332 和 19999998，这表明 yy 与 aa 之间存在某种规律：它们的差始终是 3333333 的倍数。

为了了解需要多少次迭代才能枚举 yy，直到其达到与 aa 相同的数字（每次减去 3333333），我们得到的最高数字为 213333312（3333333×64）。**这意味着，对于每张图片，最多只需遍历 64 次迭代，就能从文件名中找到真实的 fbid**。

4.4 Flickr

Exif。如果上传的图像是 JPEG，Flickr 会完整保留所有 Exif 数据。不过，如果上传的是 RAW 文件，则部分条目会被修改、删除或更新。

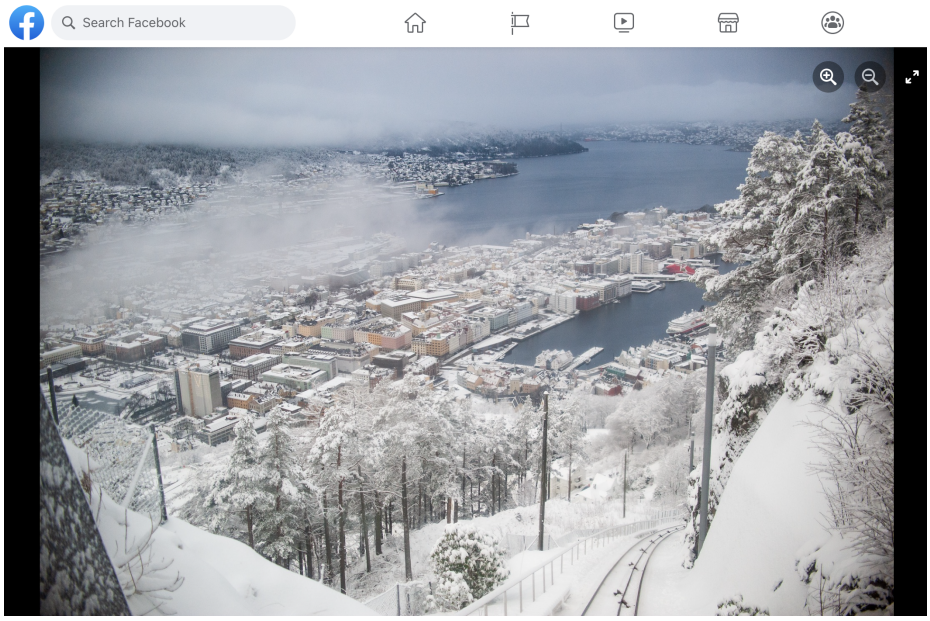


图 2. 这张 Facebook 上的照片的 fbid 为 130119042955747。用户可以通过 https://www.facebook.com/photo/?fbid=130119042955747 访问它。下载的照片文件名为："280372071_130119036289081_1523944611184590851_n.jpeg"，其中不包含 fbid。我们的研究结果表明，只需知道文件名的第二部分（130119036289081），即可恢复 fbid。

表 1. Facebook照片ID（fbid - aa）与文件名第二部分（yy）之间关系的示例。我们观察到，所有差值都是 3333333 的倍数。

yy	fbid（aa）	差值因子 yy		fbid（aa）	差值因子	
1507255906343130	1507256119676442	213333312	64	1507251046343616	1507251119676942	73333326	22
1507250843010303	1507250999676954	156666651	47	144877834756144	144877908089470	73333326	22
1507250899676964	1507251053010282	153333318	46	144877091422885	144877164756211	73333326	22
1507250853010302	1507250996343621	143333319	43	144877431422851	144877504756177	73333326	22
144877701422824	144877828089478	126666654	38	145475774696350	145475848029676	73333326	22
1507250789676975	1507250913010296	123333321	37	145475684696359	145475758029685	73333326	22
1507250933010294	1507251056343615	123333321	37	1507251473010240	1507251543010233	69999993	21
1507250846343636	1507250966343624	119999988	36	144877624756165	144877694756158	69999993	21
1507250849676969	1507250963010291	113333322	34	145475731363021	145475801363014	69999993	21
1507250949676959	1507251059676948	109999989	33	144877661422828	144877728089488	66666660	20
144877774756150	144877884756139	109999989	33	144877641422830	144877704756157	63333327	19
144877738089487	144877841422810	103333323	31	144877658089495	144877721422822	63333327	19
1507250816343639	1507250909676963	93333324	28	144877381422856	144877444756183	63333327	19

IPTC 数据。 上传到 Flickr 的 JPEG 图像显示的 IPTC 摘要哈希与原始文件相同。然而，上传到 Flickr 的 RAW 图像，其当前 IPTCDigest 哈希会被修改，但仍保留原始 IPTC 摘要哈希。

XMP 数据。上传 JPEG 文件时，Flickr 会保留所有 XMP 数据。上传 RAW 图像时，Flickr 会保留部分 XMP 数据。原始文档 ID 会被保留，并与原始文件中的 ID 相匹配。

压缩 Flickr 会重新压缩所有上传的 RAW 文件，但不会重新压缩 JPEG 文件。

分辨率 无论原始文件类型如何，上传到 Flickr 后，所有图像都会保留原始分辨率。

文件名和命名约定在 Android 和浏览器上，登录账户后通过“下载相册”功能下载图像时，Flickr 会保留原始文件名作为前缀，并按照特定模式在后面添加标识符作为后缀：

$$filename_identifier_o.jpg$$

在 iOS 上，图像会以不同顺序添加类似的标识符并重新命名：

$$identifier_filename_o.jpg$$

账户已登出时，也就是说，不是上传图像的所有者，并使用内置的下载功能下载图像，无论使用哪个平台（Android、iOS 或网页浏览器），图像的文件名都是 $identifier_filename_o.jpg$。

Flickr 允许用户在上传图像时为其选择标题。更改标题，或为通过 iPhone 应用程序上传的文件添加标题后，在账户已登录的情况下使用下载相册功能下载图像，该标题会显示为文件名的前缀。此外，如果相册中只有一张图像，使用下载相册功能时，文件名中不会包含标题。

表 2. Flickr 上传后如何为图像“r05d9d749t.jpg”命名的示例。

图像标题	状态	平台	下载的文件名
No	已登录	Android 或个人电脑	r05d9d749t_52027420848_o.jpg
No	已登录	iOS	52027420848_r05d9d749t_o.jpg
No	已登出	Android、iOS、个人电脑	52027420848_4efc66e8a4_o.jpg
“测试”	已登录	Android、iOS、个人电脑	测试_52027420848_o.jpg
"测试"	已登出	Android、iOS、个人电脑	52027420848_4efc66e8a4_o.jpg

与 Facebook ID 类似，通过修改以下 URL 中的 ID 参数，即可访问公开可用的图像：www.flickr.com/photo.gne?id=**52026001712**。此外，访问图像后，URL 会变为显示上传该图像的用户 ID：www.flickr.com/photos/**194832707@N04**/52026001712/。

解读标识符。为了了解 Flickr 如何生成标识符，我们又进行了一些实验。

首先，我们研究上传位置是否会产生影响。为此，我们在连接到虚拟专用网络（分别位于挪威、澳大利亚、美国和越南）的情况下重新上传图像。我们没有观察到不同位置之间存在任何差异。有趣的是，所有新上传图像的标识符都以 5202 开头，而之前图像的标识符则以 5197 或 5198 开头，这似乎表明标识符会根据时间或平台上的图像数量递增。为了测试时间是否会影响标识符，我们在第二天又上传了更多图像，并发现这些标识符比前一天的大。表 3 展示了我们测试中的一些示例。

表 3.上传至 Flickr 的图像日期和时间（上传之间等待特定时长）

第一天	Date	Time	次日	Date	Time
52026301587	04.24.22	09:08:51 晚上	52028946936	04.25.22	03:42:30 下午
52027901120	04.24.22	09:08:53 晚上	52027942297	04.25.22	04:01:27 下午
52027390871	04.24.22	晚上09:08:54	52027951357	04.25.22	下午04:26:05
52027420848	04.24.22	晚上 09:08:52	52029084033	04.25.22	下午 05:44:54
52027606999	04.24.22	09:08:51 p.m	52028186187	04.25.22	05:45:47 p.m.

除这项测试外，还调查了平台上的较早图像。这些于 2017 年上传到平台的图像，其标识符的位数相同，但开头是三个数字而非五个：36070463433、36880336625、36880336625。2012 年上传的图像标识符以 7 开头，但比其他标识符少一位：7542009332。

这些结果表明，这些 ID 可能代表平台上的图像总数。为了进一步测试这一假设，我们进行了第二项补充实验：从 ID 52026001712 开始，通过遍历 ID 号并抓取网站内容来提取图像的上传日期，共爬取了 3000 个图像 URL。有些图像已被删除，访问时会返回“页面未找到”错误；另一些图像则设为私密。在 52026001712 至 52026004712 的范围内，共有 1035 张图像于 2022 年 4 月 24 日上传。据报道，ID 为 52026002941 的一张图像于 4 月 23 日上传，ID 为 52026003572 的另一张图像于 4 月 25 日上传，另外两张则于 2022 年 2 月 26 日上传（ID 52026001946 和 ID 52026001956）。Flickr 上的“拍摄日期”选项可以修改为 1825 年，而“上传日期”则可以更改为用户加入 Flickr 之后的任意日期。因此，调查其他 ID 相邻图像的时间数据，比单纯依赖这些可修改的数据更有说服力。

5 种潜在应用

如果其他社交网络也采用与 Facebook 平台类似的调整大小标准，因为它们会调整所有超过特定尺寸的图像的大小，那么图像分辨率就可能成为数字指纹识别中的一项有用数据。这一点与

表 4。研究结果摘要。

属性 Flickr		**Facebook**	**Google 相册**
Exif	仅修改过的 RAW 图像——ages	删除了所有字段，只保留“艺术家”和“版权”tags	保留
IPTC	保留	删除了 JPEG 上的所有内容，保留了一些信息关于 RAW 图像	仅在以下情况下修改 重新压缩已启用
XMP	仅修改了 RAW 图像-ages	全部移除	仅在.....时修改 重新压缩已启用
分辨率	保留	调整大小后的图像大于任一维度上的2048像素	保留
仅压缩压缩后的 RAW 图像-ages		重新压缩 无论图像格式	如果重新压缩已启用，则重新压缩 压缩已启用 用
文件名	已重命名。包含递增的将心智编号作为标识符。这个编号可用于检索原始上传时间。加载时间。	已重命名。 The 原始照片 ID (fbid) 可以从 the file 名称中恢复。	保留

量化表等其他数据结合起来，可能有助于进行数字指纹识别。不过，分辨率无法有效地将图像追溯到 Google 相册或 Flickr 等平台，因为它们在上传过程中不会更改图像尺寸。

只要 Flickr 和 Facebook 对文件名所做的修改没有在其他地方被改动，这些修改就会遵循各自平台特有的模式，有助于判断上传平台。在 Flickr 上，可以利用文件名中的 ID，通过特定 URL 访问相应的 Flickr 平台图像，但前提是该图像公开可见。无论图像是否可见，都可以将 ID号与同一 ID 范围内的其他图像进行比较，从而确定上传日期。在 Facebook 上，也可以采用与 Flickr 相同的方法：使用文件名中可能存在的 ID 访问图像（如果该图像公开可见）。无论文件名中是否存在此 ID，都可以利用文件名的另一部分轻松反向推导出该 ID。

考虑到 Flickr 每天上传的图像数量，估算新近上传至 Flickr 的图像的上传日期，要比估算较早上传的图像容易得多。根据本研究的发现，可以准确估算当前及历史 Flickr 图像的上传日期。

由于可以根据从 Facebook 平台下载的图像文件名逆向推导出 Facebook 图像的图像 ID，因此也有可能从相同的文件名中逆向推导出个人资料 ID 和其他被隐藏的数据。如果能够从文件名中逆向推导出个人资料 ID，或许有助于在图像无法公开访问时追溯其上传来源。当然，我们也必须考虑隐私和伦理方面的问题。

6 结论

本文的综述表明，数字图像在上传过程中会发生多种变化，而且这些变化因上传平台以及图像文件的类型、元数据和其他属性而有很大差异。根据平台的不同，元数据可能保持不变、经过修改，或被彻底删除。根据原始图像的文件类型和 Google 相册中的设置，文件上传到社交网络时可能会重新压缩，也可能不会。Facebook 会调整宽度或高度超过 2048 px 的所有图像的大小，而其他社交网络至少在宽度或高度不超过 4928 px（本项目中使用的最大尺寸）时不会更改图像尺寸。Facebook 和 Flickr 都会以各自独特的方式更改文件名。这些文件名包含与图像上传来源有关的信息，而 Flickr 的文件名还包含上传日期。这些发现总结于表 4。基于这些发现提出的方法，使我们能够几乎仅凭上传照片的文件名来恢复隐藏信息。

致谢

本研究的结果基于挪威卑尔根大学智能信息系统（I2S）研究组与越南国家大学胡志明市国际大学之间的咨询协议。本研究由欧洲地平线 2020 项目资助，资助编号为 825469。

参考文献

1. Acker, A.：数据制作：社交媒体元数据的操纵。数据与社会研究所**13** (2018) 2. Beaver, D., Kumar, S., Li, H.C., Sobel, J., Vajgel, P.：大海捞针：Facebook 照片存储。载于：第九届 USENIX 操作系统设计与实现研讨会（OSDI 10）(2010) 3. Ben-Yair, S.：更新 Google 相册存储政策，为未来做好准备。https: // blog.google/products/photos/storage-changes/ (2020) 4. Bharati, A., Moreira, D., Brogan, J., Hale, P., Bowyer, K., Flynn, P., Rocha, A., Scheirer, W.：超越像素：利用元数据进行图像来源分析。载于：2019 年 IEEE 冬季计算机视觉应用会议（WACV）。第 1692–1702 页。IEEE (2019)

5. Broz, M.：照片数量（2022）：统计数据、事实与预测。https: // photutorial.com/photos-statistics/ (2022) 6. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G.：Raise：用于数字图像取证的原始图像数据集。载于：第六届 ACM 多媒体系统会议论文集。第 219–224 页 (2015) 7. Ferreira, W.D., Ferreira, C.B., da Cruz Júnior, G., Soares, F.：数字图像取证综述。《计算机与电气工程》**85**, 106685 (2020) 8. Giudice, O., Paratore, A., Moltisanti, M., Battiato, S.：用于社交数据图像弹道学的分类引擎。载于：国际图像分析与处理会议。第 625–636 页。施普林格 (2017) 9. Harvey, P.：Phil Harvey 的 ExifTool (2013) 10. Kee, E., Johnson, M.K., Farid, H.：通过 JPEG 头部进行数字图像认证。《IEEE 信息取证与安全汇刊》**6**(3), 1066–1075 (2011) 11. Marra, F., Poggi, G., Sansone, C., Verdoliva, L.：用于来源识别的盲式基于 PRNU 的图像聚类。《IEEE 信息取证与安全汇刊》**12**(9), 2197–2211 (2017) 12. Mehta, V., Gupta, P., Subramanian, R., Dhall, A.：Fakebuster：用于视频会议场景的深度伪造检测工具。载于：第二十六届国际智能用户界面会议伴会。第 61–63 页 (2021) 13. Moltisanti, M., Paratore, A., Battiato, S., Saravo, L.：Facebook 上的图像操纵及其取证证据。载于：国际图像分析与处理会议。第 506–517 页。施普林格 (2015) 14. Mullan, P., Riess, C., Freiling, F.：利用 JPEG 图像头部进行取证来源识别：以智能手机为例。《数字调查》**28**, S68–S76 (2019) 15. Pasquini, C., Amerini, I., Boato, G.：社交媒体平台上的媒体取证：综述。《EURASIP 信息安全期刊》**2021**(1), 1–19 (2021) 16. Redi, J.A., Taktak, W., Dugelay, J.L.：数字图像取证：初学者手册。《多媒体工具与应用》**51**(1), 133–162 (2011) 17. Riggs, C., Douglas, T., Gagneja, K.：通过元数据进行图像映射。载于：2018 年第三届智能城市、工业控制系统与通信安全国际会议（SSIC）。第 1–8 页。IEEE (2018) 18. Sandoval Orozco, A., Arenas González, D., Rosales Corripio, J., García Villalba, L.J., Hernandez-Castro, J.C.：结合传感器缺陷与小波变换的移动设备来源识别。《计算》**96**(9), 829–841 (2014) 19. Siddiqui, N., Anjum, A., Saleem, M., Islam, S.：使用深度卷积神经网络进行基于社交媒体来源的图像追踪。载于：2019 年第五届国际图像信息处理会议（ICIIP）。第 97–101 页。IEEE (2019) 20. Sjøen, V.V.：IBSNtools——社交网络中的图像弹道学。21. Tran, C.H., Tran, Q.T., Long-Vu, Q.C., Nguyen, H.S., Tran, A.D., Dang-Nguyen, D.T.：Dedigi：用于图像取证的隐私设计平台。载于：第三届 ACM 智能跨数据分析与检索研讨会论文集。第 58–62 页 (2022) 22. Vo, N.H., Phan, K.D., Tran, A.D., Dang-Nguyen, D.T.：针对深度伪造检测器的对抗攻击：实践分析。载于：国际多媒体建模会议。第 318–330 页。施普林格 (2022)

n