

使用可解释的图文基础模型增强变形攻击检测

Sushrut Patwardhan^{1,2}, Raghavendra Ramachandra¹[0000-0003-0484-3956], and
Sushma Venkatesh²[0000-0002-8557-0314]

¹ Norwegian University of Science and Technology (NTNU), Norway.

² MOBAI AS, Norway.

Email: raghavendra.ramachandra@ntnu.no

Abstract. 变形攻击检测已成为面部识别系统中的一个重要组成部分，确保证场景的可靠性。在本文中，我们提出了一种多模态学习方法，可以为变形攻击检测提供文本描述。我们首先展示了使用对比语言图像预训练 (CLIP) 对所提出框架进行零样本评估，不仅能够实现可推广的变形攻击检测，还能预测出最相关的文本片段。我们对十个不同的文本提示进行了广泛分析，这些提示包括短文本和长文本提示。这些提示是通过考虑人类可理解的文本片段来设计的。我们在利用公开可用的人脸生物识别数据集开发的面部变形数据集上进行了广泛的实验。我们展示了在五种不同的变形生成技术的零样本评估中，与所提出的框架一起评估了最先进的预训练神经网络，这些技术是在三种不同的介质中捕获的。

1 介绍

人脸识别系统 (FRS) 广泛应用于高安全性应用中，如边境控制，以实现无缝且准确的验证。然而，它们易受各种攻击的影响，其中变形攻击在实际的边境控制场景中构成了显著的威胁 [?]，[?]，[?]，[?]，[?]。变形涉及混合多个人的面部图像以创建一个新图像，该图像保留了每个个体的特征，这使得人类操作者和自动化 FRS 都难以将其与合法面孔区分开来，正如在 [?] 中所示。这在身份文件场景中构成了严重风险，变形的图像可能会通过将文件链接到多个人来破坏文件的完整性，可能导致滥用。因此，检测变形图像对于确保可靠的用户验证至关重要。

面部变形攻击检测 (MAD) 已经在文献中进行了广泛研究，并可以大致分为 (a) 基于单一图像的 MAD (S-MAD) 和 (b) 基于差异的 MAD (D-MAD)。S-MAD 方法使用单一图像，而 D-MAD 则使用两幅图像（一个来自护照，另一个来自受信设备）来确定变形。在这两种方法中，基于 S-MAD 的方法报告的错误率比 D-MAD 的方法更高 [?]。在我们的研究中，我们专注于 S-MAD 方法，因为这些方法在解释和检测方面提出了独特的挑战。虽然基于深度学习的 MAD 算法表现出合理的检测准确性，但其缺乏解释性限制了其可靠性。因此，大量研究已针对基于 S-MAD 的解释技术进行 [?]。

早期关于可解释的 MAD 的研究主要集中在视觉化手工特征，这些特征可以指示真实和变形图像的特征特征。关于手工特征解释的首次工作在 [?] 中提出，展示了用于 S-MAD 的二进制统计图像特征 (BSIF) 方法中的特征。在 [?] 中展示了五种不同手工特征的解释，这些特征可以指示与真实和变形对应的不同纹理模式。在 [?] 中讨论了使用颜色通道作为 S-MAD 的解释技术，展示了颜色通道的相关性。然而，使用手工特征作为解释并不能为非技术人员，如边境守卫，

提供有用信息，以确定 S-MAD 技术决策背后的理由。关于深度学习方法用于 S-MAD 解释的首次工作在 [?]，[?] 中提出。梯度图用于解释来源于 CNN 不同层的深度特征，这表明了变形和真实图像的独特特征。在 [?] 中，针对深度学习和手工方法的残余噪音基于 S-MAD 方法的解释被提出。关于残余深度网络的解释在 [?] 中提出，其中使用热图来呈现 S-MAD 的解释。在 [?] 中提出了使用集中逐层相关传播 (FLRP) 的像素级解释方法。使用基于 VGG-A 的 DNN 架构训练变形攻击检测器，以评估 FLRP，以指示决策的像素级解释性。

基于前面的讨论，热图通常用作决策过程的解释工具。然而，需要注意的是，使用 GradCAM [?] 或类似方法生成的热图可能仅突出图像中的特定区域，并且可能呈现噪声特性。因此，没有技术专业知识的人，特别是在边境控制机构工作的人，可能会发现难以解释这些方法。因此，这些视觉解释有几个缺点，它们不是精确的，细粒度的，并且可能存在解释的主观性。

在这项工作中，我们提出了一种多模态方法来解释 S-MAD 的决策。据我们所知，这是首次使用基于深度学习的方法来提供对变形攻击决策的文本描述。我们建议使用多模态方法对比语言-图像预训练 (CLIP) [?]，该方法可以预测给定变形或真实面部图像最相关的文本片段。我们展示了在 CLIP 上使用不同文本提示的零样本学习，以获得人类理解水平的解释。以下是本工作的主要贡献：

- 我们提出了一种用于人脸篡改攻击检测的文本可解释方法，该方法利用对比语言-图像预训练 (CLIP) 的零样本学习，特别是在零样本设置下。据我们所知，这是第一个预测文本片段和相应分数以进行篡改检测的工作。
- 我们分析了十个不同的提示，既平衡了连贯的文本解释又兼顾了变形攻击检测性能。这为在零样本环境中识别合适的提示提供了基础。
- 进行了广泛的实验，使用五种不同类型的变形生成技术和三种不同类型的媒介（数字、高质量打印扫描和低质量打印扫描）。

本文的其余部分组织如下：第 2 节介绍了可解释的变形攻击检测的提出方法，第 3 节展示了十个不同提示的定性和定量结果，第 ?? 节得出结论。

2 提出的方法

提出的方法的目标是实现基于单一图像的变形攻击检测的文本描述。因此，我们计划采用基于对比语言-图像预训练 (CLIP) 的多模态框架。CLIP 是一个使用大量图像-文本对基于对比学习框架训练的图像-语言模型。CLIP 模型的主要优势之一是其在零样本学习中的良好表现，这使其在多种应用中得到了使用。特别是，最近的研究表明，CLIP 可以有效用于检测面部呈现攻击 [?]。在 [?] 中，CLIP 在跨数据场景中表现出合理的检测性能，并预测了反映决策的文本片段，特别是在零样本学习的背景下。这些结果促使我们探索使用 CLIP 在单一图像基础变形攻击检测中的零样本设置中的应用。

图 1 展示了用于基于单一图像的变脸攻击检测的多模态模型文本可解释性建议框架的框图。该建议框架基于 CLIP 在零样本设置中操作，从图像编码器获得检测准确性，并利用文本编码器对决策进行文本解释。我们首先介绍 CLIP 模型的详细信息，随后是其在可解释变脸攻击检测中的应用。

2.1 对比语言-图像预训练 (CLIP)

对比语言-图像预训练 (CLIP) 神经网络 [?] 通过使用互联网上现成的自然语言监督下的数百万生动图像进行训练。通过这种方法，CLIP 模型在共享特征空间

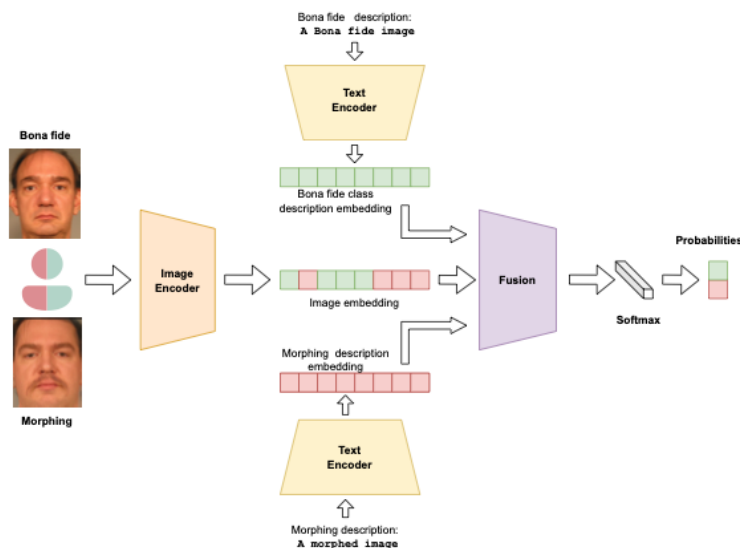


Fig. 1: 所提出的框架的框图，该框架使用零样本可解释单图像基础的变形攻击检测

中有效地编码了图像及其对应的描述。CLIP 模型包括两个主要组件：图像编码器和文本编码器。

图像编码器 CLIP 模型的图像编码器有两种变体：ResNet-50 [?] 和 Vision Transformer [?]。两种变体在架构上有微小调整，对于我们的方法，我们使用基于 Vision Transformer 的图像编码器，其表现出了更优异的性能，如 [?] 中详细描述。此外，使用预训练视觉变压器还显示出相对于其他 S-MAD 技术的改进泛化能力 [?]。图像编码器由生成图像嵌入的变压器模块组成，然后使用线性层将其投射到与文本描述的共享特征空间中。

文本编码器 在 CLIP 中实现的文本编码器采用了一种基于变换器的架构，并进行了一些修改，如 [?] 中所讨论。编码器在字节对编码（BPE）表示上运行，所有字符均为小写，以确保更准确的编码。最大序列长度限制为 76，以优化性能。与图像编码器类似，文本编码器生成的嵌入被投影到共享的特征空间上，使用线性层进行投影。

2.2 可解释的 S-MAD 的零样本评估

零样本学习（ZSL）是一种机器学习方法，使模型能够识别和分类它以前从未遇到过的对象。为了使用 CLIP 模型在拟议框架下执行单图像形变攻击检测，我们将面部图像通过图像编码器并获得相应的特征。同时，我们将真实和变形图像的文本描述通过文本编码器并获得文本特征。然后使用基于图像和文本特征

的余弦距离的 softmax 层计算概率。由于是零样本评估，因此预训练网络的权重在没有任何更改的情况下被使用。

零样本评估的关键在于选择合适的文本提示。[?] 中展示的研究结果强调了在优化 CLIP 模型的零样本评估性能时提示工程的重要性。为此，我们在这项工作中提出了十个不同的提示，考虑到了人类理解和检测能力。这些提示在变形检测适用性方面进行了检查。关于不同提示及其各自检测性能的具体细节在 3 中提供。

3 实验和结果

在本节中，我们将展示并讨论所提出的 S-MAD 可解释框架的定量和定性结果，该框架利用了 CLIP 的零样本评估。首先，我们将深入探讨本研究使用的变形数据集。然后，我们讨论可以用于比较不同类型提示的零样本评估的各种评估协议。最后，我们呈现并分析每个提示的定量结果。

3.1 变形数据集

在本节中，我们介绍了形变数据集 (MD)，该数据集取样自公开可用的 FRGC V2 数据集 [?]。我们选择了 143 名带有中性表情和姿势的个体，遵循 [?] 的建议。用于形变的对应不同个体的图像对选择也基于来自 [?] 的列表。形变图像是通过五种不同的面部形变技术生成的：基于标记点的 (LMA-I) [?]，带后处理的标记点 (LMA-II) [?]，MIPGAN-2 [?]，MorDiff [?] 和 PIPE [?]。

MD 数据集由三种不同类型的介质组成：数字和两种类型的打印-扫描 (PS)。数字版本包含常规的变形图像，而 PS 变形图像是这些数字图像的重新数字化版本。包含 PS 版本的动机是为了反映护照签发场景，其中接受打印的护照图像。在这项工作中，我们创建了高质量和低质量的打印-扫描图像，以反映现实场景。高质量 PS 图像是使用 DNP 打印机 [?] 打印数字正式和变形图像创建的（我们称之为 PS-1），而低质量 PS 图像（我们称之为 PS-2）是使用 RICOH 打印机创建的。在这两种情况下，打印的图像都使用办公扫描仪扫描，以实现 300dpi 的分辨率，遵循 ICAO 的建议 [?]。

图 2 展示了与“数字”和“PS”数据集相对应的 MD 数据集的示例。值得注意的是，图像的质量因“PS”而略有下降。MMD 数据集包含 1276 个真实样本（分别为数字和变形）和 2526 个变形图像（分别为五种不同的变形技术，包括所提的方法，并分别用于数字和“PS”）。因此，MD 数据集包含 $1276 \times 2 = 2552$ 个真实样本和 $2526 \times 5 \times 2 = 12630$ 个变形图像。

3.2 结果与讨论

提出的框架的定量结果通过 ISO/IEC SC 37 30107 指标展示，包括变形攻击分类错误率 (MACER) 和真实呈现分类错误率 (BPCER)。在这项工作中，我们展示了在 $BPCER @APCER = 10\%$ 下的结果，针对本研究中的所有十个提示进行 S-MAD 研究。在这项工作中，我们提出了三种不同的实验评估协议：(a) 实验-1：本实验旨在通过平衡人类理解和检测准确性来评估不同的文本提示。图 3-12 讨论了本研究中设计的提示的定量准确性。实验-2：本实验旨在提供依据提示类型（短或长）、变形生成算法和变形媒介的影响进行量化检测的见解。图

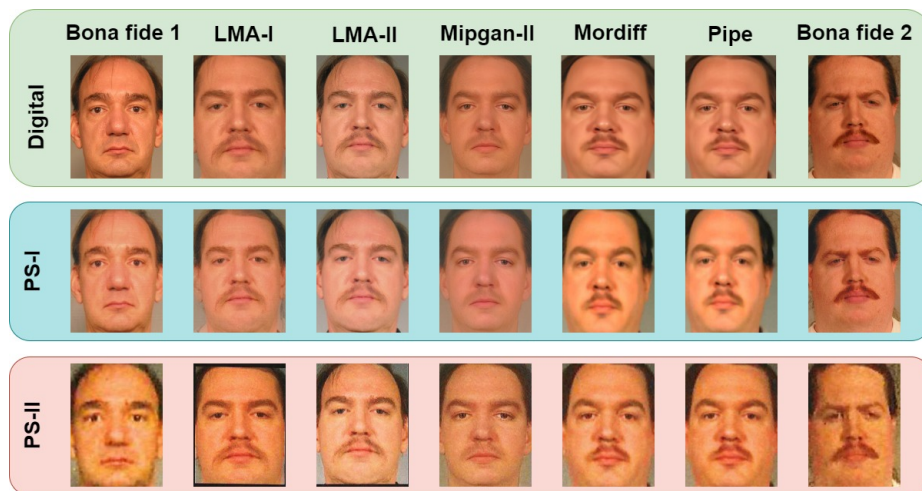


Fig. 2: 来自 MD 数据集的不同介质的变形样本插图。

13、14、15 和 16 展示了定量结果。实验-3: 本实验旨在分析所提提示的个体表现, 而不论变形过程的生成和媒介。图 18 展示了定量结果。

图 3 至 12 显示了对应于各个提示词及其在三种不同数据媒介下对应定量结果的定量结果。在十种不同的提示词中, 前五种提示词(提示 # 1- # 5) 被设计为短句, 以反映在训练 CLIP 时使用的类似(或同义) 句子。最后五个提示词(提示 # 6 - # 10) 被设计为更具描述性, 因此这些是长句子。我们展示了五种不同的变形生成方法的检测结果, 分别是: 基于标记的方法 LMA-I [?] 和 LMA-II [?], 以及生成方法如 MIPGAN-2 [?]、Mordiff [?] 和 PIPE [?]。每个提示词独立的 S-MAD 的检测性能在图 3-12 中显示, 其中对于每个图像使用推理的提示词以及对应的结果用柱状图表示。柱状图中的 x 轴表示不同的变形生成方法, y 轴表示 $BPCER @ MACER = 10\%$ 。为了更好地理解获得的结果, 我们进行了如下所述的附加分析。

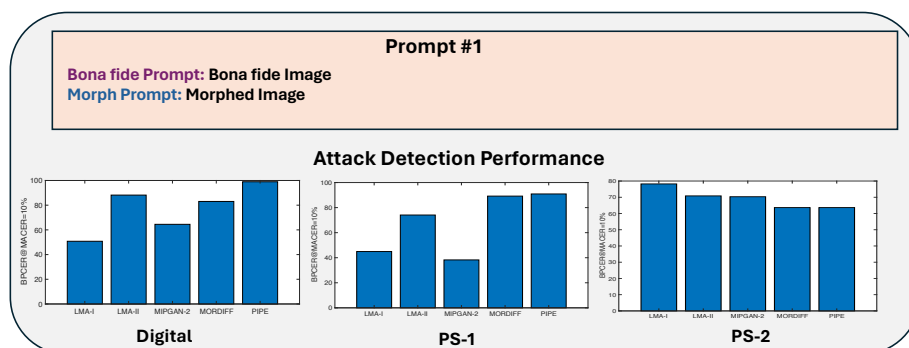


Fig. 3: 使用拟议框架和提示 # 1 的 S-MAD 定量结果。

图 13 展示了不同变形生成技术在三种不同媒介上的平均检测性能。总体而言，数字媒介的检测性能优于 PS-1 和 PS-2 媒介。此外，不同的变形生成类型表现出不同的性能。在数字媒介中，MIPGAN-II 和 LMA-I 取得了最佳结果，而在 PS-1 中观察到了相似的性能。然而，在 PS-2 中，LMA-I 和 LMA-II 的性能优于用于变形攻击的生成方法。因此，媒介对检测准确率有影响，具体取决于使用的变形生成技术类型。

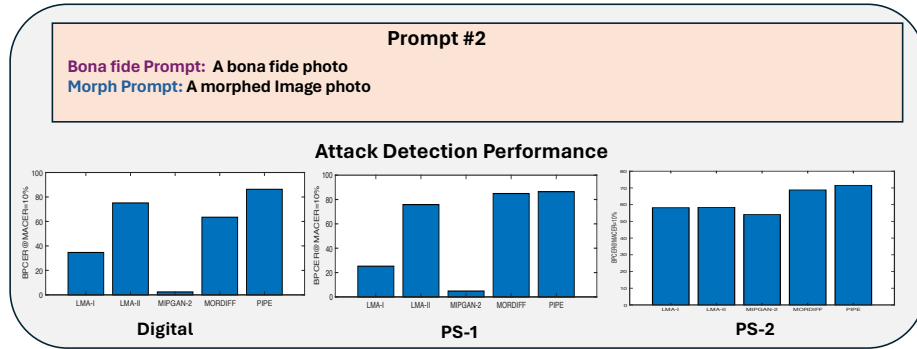


Fig. 4: 提出的 S-MAD 框架与提示 # 2 的定量结果。

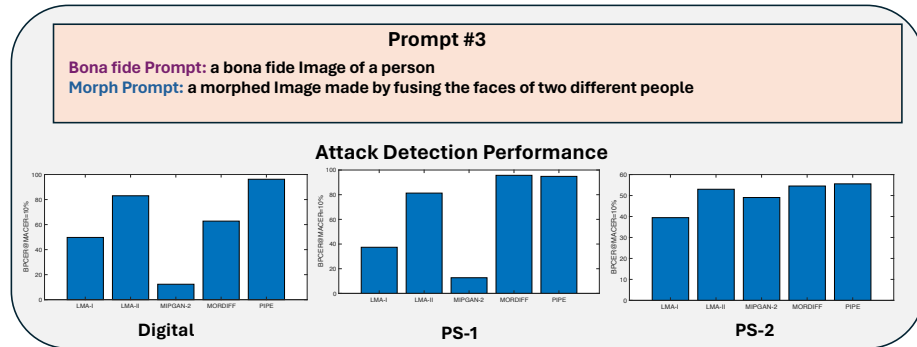


Fig. 5: 提出的 S-MAD 框架与提示 # 3 的定量结果。

图 17 展示了通过对变形生成算法和提示的检测性能进行平均计算得出的每个媒介的单独箱线图。总体而言，显然较短的提示通常比较长的提示具有更好的检测性能。数字和 PS-1 媒介在短提示和长提示下表现相似。然而，使用短提示时，PS-2 数据媒介在所有三种数据媒介中表现相似，但使用长提示时性能有所下降。

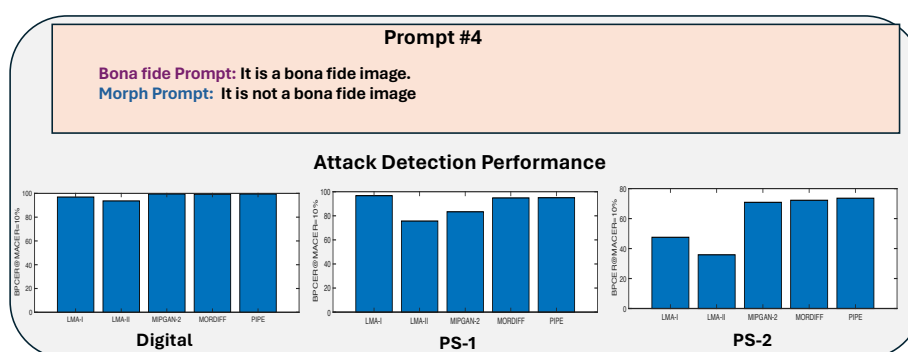


Fig. 6: 提出的带提示 # 4 的 S-MAD 框架的定量结果。

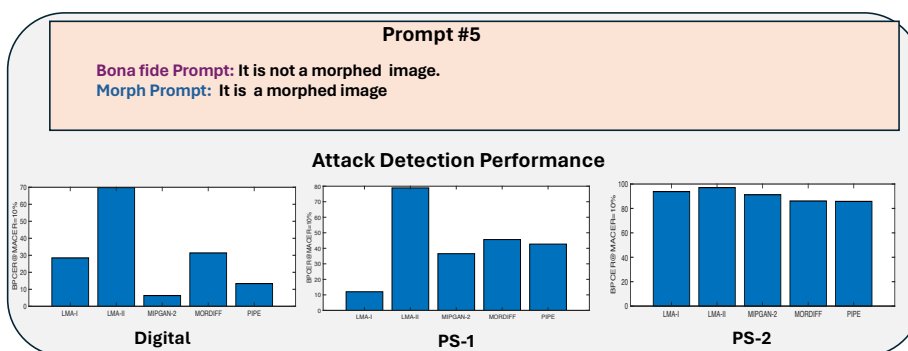


Fig. 7: 提出的 S-MAD 框架与 Prompt # 5 的定量结果。

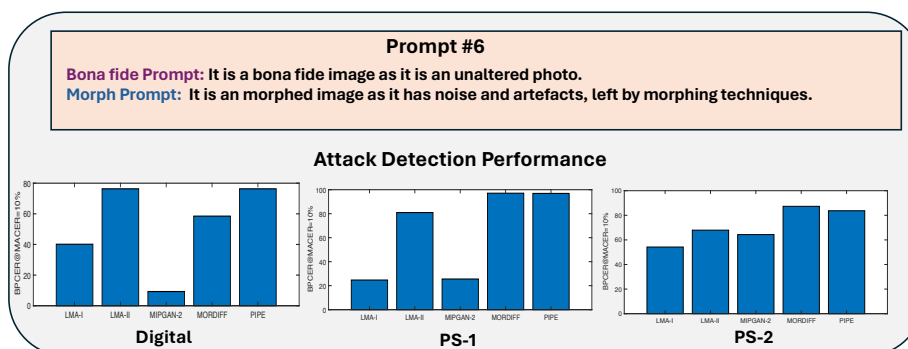


Fig. 8: 所提出的 S-MAD 框架与提示 # 6 的定量结果。

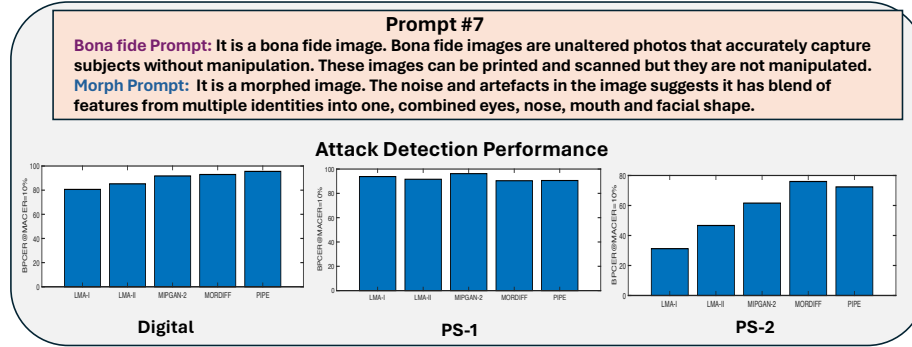


Fig. 9: 使用提示 # 7 的建议 S-MAD 框架的定量结果。

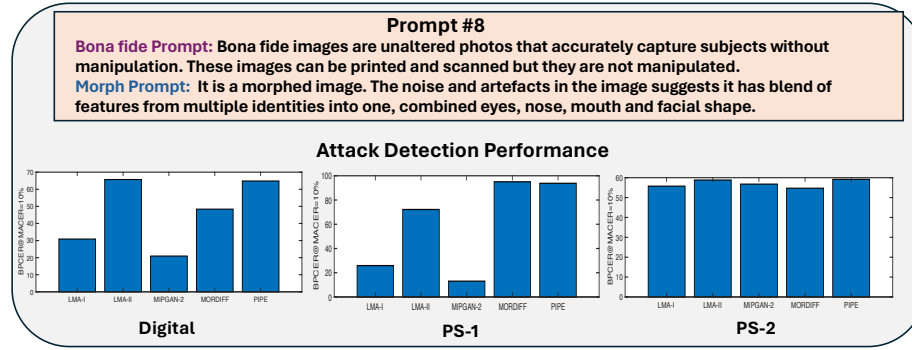


Fig. 10: 所提出的 S-MAD 框架与 Prompt # 8 的定量结果。

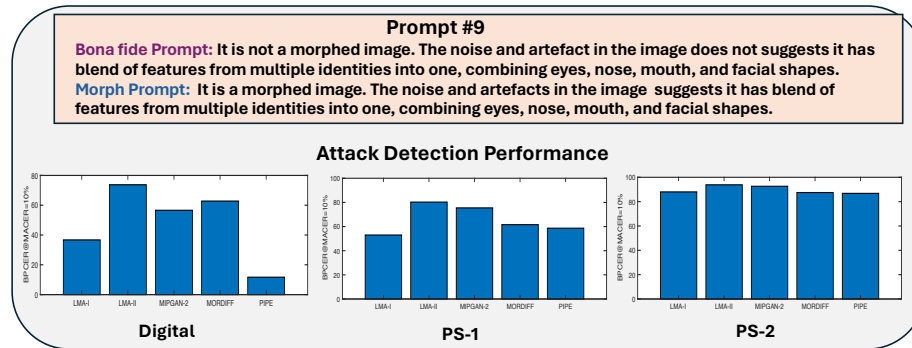


Fig. 11: 提出的具有提示 # 9 的 S-MAD 框架的定量结果。

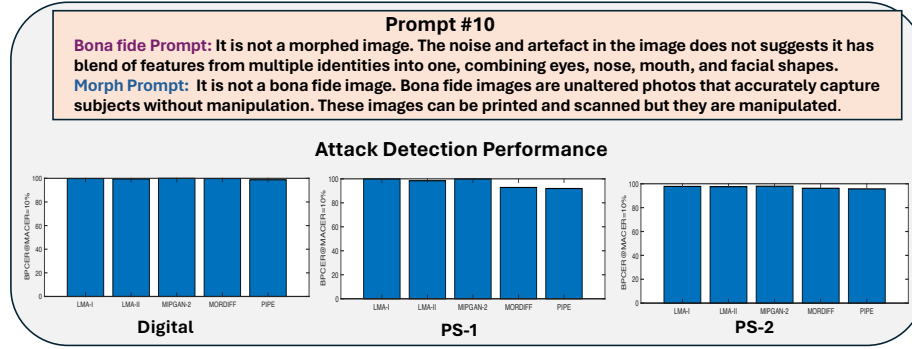


Fig. 12: 所提出的 S-MAD 框架与提示 # 10 的定量结果。

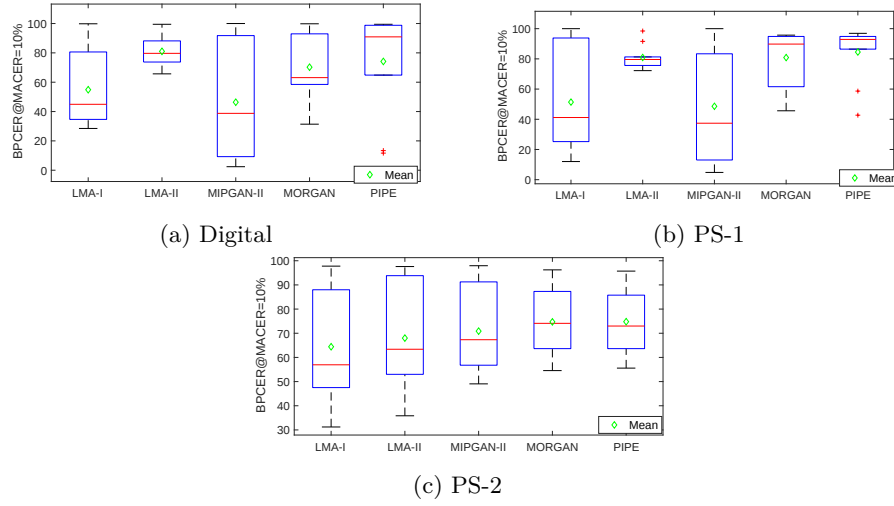


Fig. 13: 在每种媒介上，各提示下不同变形生成算法的平均表现。

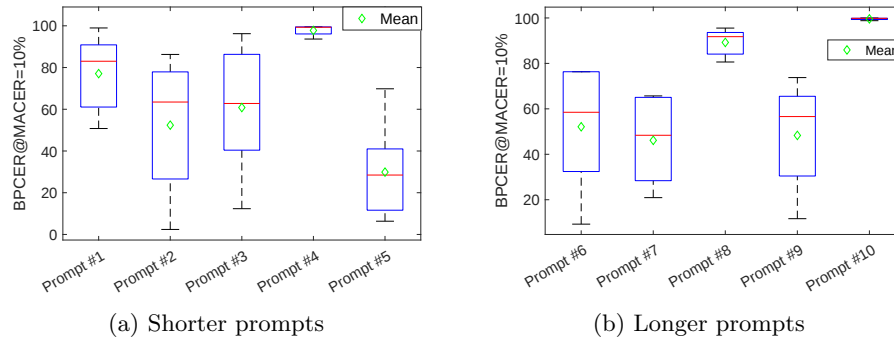


Fig. 14: 不同形态生成算法在较短和较长的提示与数字媒介上的平均表现。

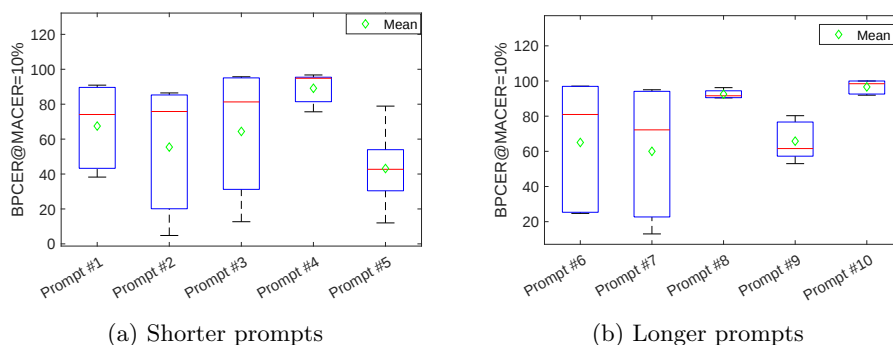


Fig. 15: 不同形变生成算法在更短和更长提示词以及 PS-1 中等性质上的平均表现。

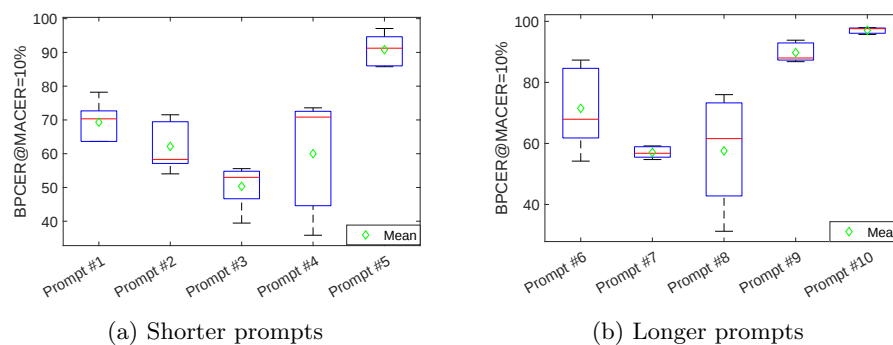


Fig. 16: 不同变形生成算法在较短和较长的提示上使用 PS-2 媒介的平均表现。

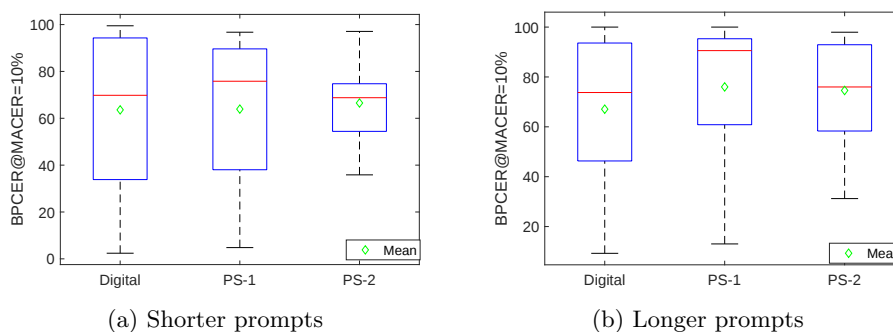


Fig. 17: 不同变形算法在所有提示和变形生成算法中分别在各自媒介上的平均性能。

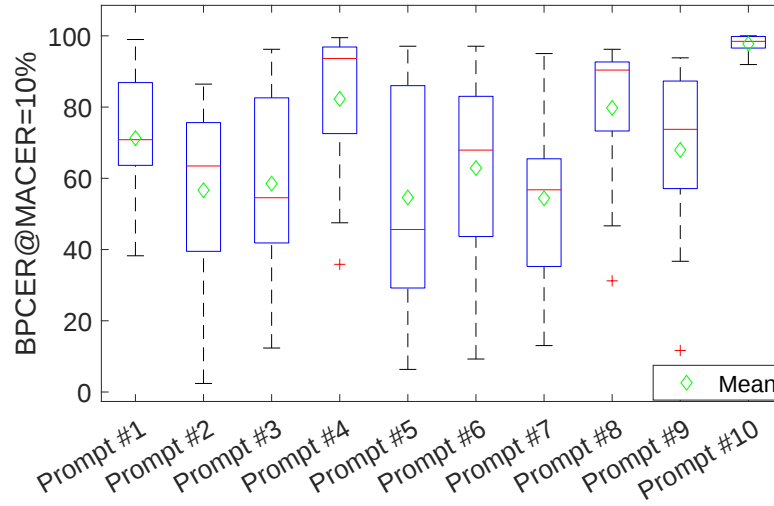


Fig. 18: 不同变形算法在所有三种媒介上对各个提示的平均表现。

图 18 显示了在两种变形生成算法和介质中带有单个提示的平均检测性能。这些结果提供了在一种介质和生成类型未知的实际场景中的零样本评估概览。结果表明使用短提示 # 5 的性能最佳，显示出使用简短提示的优势。

MAD Algorithms	Data Medium	BPCER @ MACER = 10 %
ResNet50 [?]	Digital	90.15
	PS-1	75.98
	PS-2	35.78
VGG-19 [?]	Digital	52.48
	PS-1	54.54
	PS-2	30.56
ViT [?]	Digital	95.64
	PS-1	84.45
	PS-2	75.78
CLIP (Proposed Framework)	Digital	29.87
	PS-1	43.15
	PS-2	90.87

Table 1: 对所提出的可解释性 S-MAD 与通过零样本实验评估的 SOTS 进行定性评价。

表格 1 显示了所提出的方法和 SOTA 神经网络在零样本评估下的定量性能。为简单起见，我们包含了与提示 # 5 相对应的结果，因为它是所提框架中表现最好的提示。获得的结果显示：(a) 基于 CNN 的预训练网络（ResNet50 和 VGG-19）的零样本评估显示在检测使用 PS-2 构建的变形攻击中表现良好。(b) 变压器方法在所有三种媒介中检测变形攻击的表现不佳。(c) 所提框架在数字和 PS-1 媒介中表现出最佳结果。因此，表明在零样本评估中性能最好。



Fig. 19: 在 CLIP 视频编码器上对真实样本和变形样本进行 Lime 可解释性分析。左图显示了真实样本的正确和不正确检测的 LIME 结果。右图显示了变形样本的正确和不正确检测。

图 19 展示了在 CLIP 模型中使用的 ViT 基础上视频编码器的基于 LIME 的可解释性结果。对真实和变形的例子进行了定性结果展示，包括正确和错误分类。如图 19 所示，(a) 真实图像的正确分类可归因于整幅图像中选定的特征，而错误分类则是由于大量使用了发型特征；(b) 正确分类主要是由于脸部眼睛区域特征的重要性，而错误分类则是由于非面部特征以及脸部区域特征的分散性。

因此，基于可解释性结果，很难确定哪些特征对面部变形攻击检测器的整体决策做出了贡献。然而，所提出的多模态方法还可以预测文本片段，这可能为正确决策提供附加信息。在本文中，我们证明了在零样本实验中使用的图像-语言基础模型可以表现出改进的泛化性能以进行变形攻击检测。我们展示了提示工程对提高可泛化检测准确性的重要性。我们分析了包括短版和长版在内的十种不同的提示，并对变形检测性能进行了基准测试。在零样本设置中，对使用五种不同变形生成方法和三种不同数据介质（数字、高质量打印扫描和低质量打印扫描）构建的变形数据集进行了广泛实验。此外，我们对比了图像-语言模型与仅使用图像模型在变形攻击检测性能上的基准测试。获得的结果表明，图像-语言模型具有最佳性能，尤其是在使用短提示时。未来，我们计划探索视觉模型在变形攻击检测中的微调，并整合最佳的视觉模型以提高泛化能力。这项工作得到了欧盟委员会的支持 [资助 101120657 ”欧洲灯塔展示值得信赖和绿色的人工智能” - ENFIELD]。