Pixels of Faith: Exploiting Visual Saliency to Detect Religious Image Manipulation

Giuseppe Cartella[1], Vittorio Cuculo[1], Marcella Cornia[1], Marco Papasidero[2], Federico Ruozzi[1], and Rita Cucchiara[1]

[1] University of Modena and Reggio Emilia, Italy
name.surname@unimore.it
[2] University of Palermo
name.surname@unipa.it

Abstract. The proliferation of generative models has revolutionized various aspects of daily life, bringing both opportunities and challenges. This paper tackles a critical problem in the field of religious studies: the automatic detection of partially manipulated religious images. We address the discrepancy between human and algorithmic capabilities in identifying fake images, particularly those visually obvious to humans but challenging for current algorithms. Our study introduces a new testing dataset for religious imagery and incorporates human-derived saliency maps to guide deep learning models toward perceptually relevant regions for fake detection. Experiments demonstrate that integrating visual attention information into the training process significantly improves model performance, even with limited eye-tracking data. This human-in-the-loop approach represents a significant advancement in deepfake detection, particularly for preserving the integrity of religious and cultural content. This work contributes to the development of more robust and human-aligned deepfake detection systems, addressing critical challenges in the era of widespread generative AI technologies.

Keywords: Gaze-assisted AI · Human Attention · Deepfake Detection · Religious Studies

1 Introduction

In recent years, the diffusion of generative models has significantly transformed various aspects of everyday activities. Generative models, particularly those based on deep learning architectures, have shown remarkable capabilities in generating realistic data, including images [30, 48, 50], text [12], and audio [35]. These models have enabled a wide range of applications from creative arts to personalized content creation, and have enhanced productivity in fields such as design, entertainment, and education.

The democratization of generative models has enabled individuals and small organizations to leverage sophisticated AI tools previously accessible only to large corporations and research institutions. Platforms and applications powered by generative models allow users to create high-quality digital art, write coherent articles, compose music, and even develop virtual environments with minimal technical expertise. This accessibility fosters innovation and creativity, driving forward the capabilities of AI in augmenting human activities and contributing to economic growth. However, the widespread adoption of generative models also brings significant challenges, particularly in the domain of misinformation and deepfake generation. The capacity of generative models to synthesize highly convincing yet entirely fabricated images, videos, and audio clips has raised concerns about their potential misuse in influencing public opinion, political campaigns, and social media narratives.

In the context of religious studies, these challenges are particularly relevant. Religious images have been and still are of great significance to billions of people worldwide. The history of faith communities that do not apply aniconic prohibitions shows that religious imagery and iconography have played a fundamental role in the dissemination and transmission of faith, devotion, and knowledge about sacred texts, but also had polemic and propaganda purposes in intra- and inter-religious dynamics. In medieval and modern times, images have been modified mainly with the aim of reproducing sacred images to adapt them to different geographical and cultural contexts or to ameliorate the subjects. In the cases when images have been used in overtly propagandist ways – supporting one identity, culture, religious, or political party over another – manipulation of the elements obviously opted for pejorative or ridiculing elements.

More recently, the adoption of images in a digital format and the spread of AI changed the way images are manipulated and their effects: deepfakes make it much more difficult to distinguish real images from those created by AI, and they are spread at unprecedented speed, with chances to "go viral" on a global scale in short time. Such a change has also affected Christian imagery: numerous devotional images produced by AI algorithms, have caused controversy as they depict physical defects. These examples have alarmed believers and various Christian Churches around the world because the images, while apparently suitable for private devotion, exhibit subtle traits of blasphemy. Moving from such specific cases, it has been demonstrated that deepfakes, created ad hoc, generate conflicts with high social costs and have an impact at social, religious, and political levels that cannot be ignored. This is particularly relevant when deepfakes are improperly used for targeting specific groups or individuals based on their race, gender, or religion [5]. This form of discrimination conveys prejudices that heighten existing social tensions and further reinforce harmful stereotypes.

To counterbalance these trends and work on accurate historical and cultural knowledge, interdisciplinary teams of scholars can develop control systems based on robust methods for automatic fake recognition. However, the discrepancy between human and algorithmic deepfake detection capabilities presents an intriguing challenge in this field. As reveled in [36], state-of-the-art algorithms often struggle to detect deepfake videos that human observers find obviously fake. This counterintuitive finding suggests that current models may not be capturing the same visual cues that humans instinctively recognize as indicators of

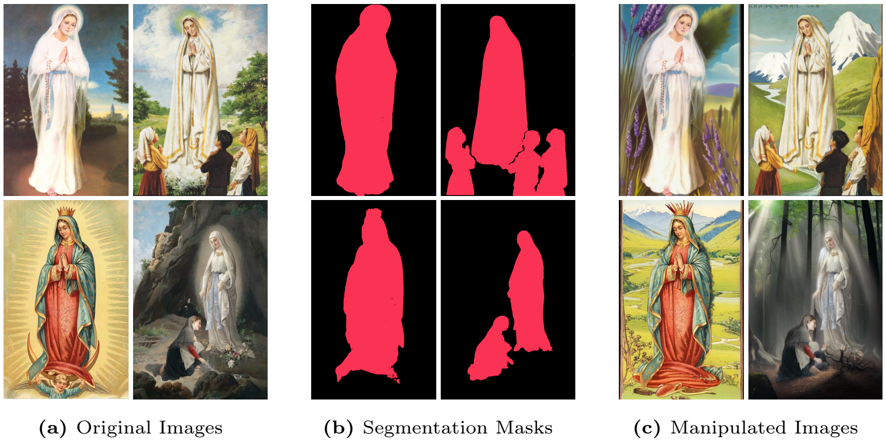


Fig. 1: Qualitative examples of our proposed dataset comprising original religious images and corresponding manipulated versions using segmentation masks to preserve religious figures. From top to bottom and from left to right: Marian apparition of Banneux (Belgium), Marian apparition of Fatima (Portugal), Marian apparition of Guadalupe (Mexico), Marian apparition of Lourdes (France).

manipulation. Several factors could contribute to this phenomenon: (i) humans may rely on subtle contextual or semantic cues that AI models are not attuned to; (ii) algorithms might overfit to specific artifacts rather than learning generalizable features; and (iii) there could be a mismatch between fake images used to train algorithms and those shown to human participants. Moreover, human perception can be influenced by expectations, context, and cognitive biases, while algorithms apply their detection criteria more consistently [41]. This discrepancy reveals gaps in our understanding of both human and machine perception of deepfakes, highlighting the need for further research to bridge this divide.

Building on previous studies that have demonstrated how visual attention is subtly influenced by the authenticity or manipulation of observed images [14], this paper presents the first attempt to exploit the potential of human visual attention for the automatic detection of partially manipulated images, with a specific focus on the field of religious studies. By analyzing where humans focus their attention when identifying deepfakes, we create human-derived saliency maps that highlight the most perceptually relevant regions for deepfake detection. In this context, saliency maps can highlight areas that might have been manipulated, as these regions often exhibit subtle inconsistencies that draw attention. By incorporating this visual attention information into visual backbones, we aim to guide the feature extraction process toward areas that are more likely to show signs of manipulation. This human-in-the-loop approach is shown to enhance the performance of deepfake detection systems, especially on images that are visually obvious to human observers but challenging for current algorithms.

Contributions. To sum up, the main contributions of this paper are as follows:

- We exploit human attention and visual saliency for the task of deepfake detection on partially manipulated images.

- We introduce a new testing dataset specifically designed to address the distinctive challenges posed by religious images and, more broadly, partially modified ones (Fig. 1).
- Our experiments show how the integration of visual attention into the training process of a model for the deepfake detection task contributes positively to its performance. Moreover, we show that even in the presence of a small amount of data recorded by eye-trackers, they are of value and beneficial compared to the exclusive use of generated data from saliency models.

2 Related Work

Gaze-assisted AI. Human visual attention provides valuable insights into the perceptual processes that guide human decision-making [8, 17, 33, 47]. Understanding its mechanisms holds significant potential for advancing the development of efficient and generalizable AI systems. Human attention guidance, obtained from eye-tracking data or computational models [3, 7, 19, 40], has proven beneficial in solving several intriguing challenges [13], and the integration into AI models can enhance their performance by guiding the focus of the model to the most informative parts of the data. This has implications across various domains, ranging from visual recognition [57], natural language processing [52], and vision-and-language understanding [32] to sound source localization [45] and deepfake detection [11]. Some examples include the works presented in [1,42] that adopted a saliency-guided image enhancement approach to reduce distractors while shifting the attention towards the most interesting objects in the scene. On another line, some works [26,37] proposed new solutions for advancing automatic graphic design. Others [15,18,53,54], instead, harnessed the potential of human attention for visual question answering, image captioning, and reasoning.

However, human attention is not only used as a tool for specific applications but also in enhancing the generalizability of deep learning models. Boyd *et al.* [10] annotated a small amount of training data with human saliency, where salient regions indicate the most discriminative parts to make a decision. They demonstrated that this type of human-aided training improves the performance of deep learning models, especially in a setting with limited training data. Unfortunately, human saliency annotation is quite expensive. As a consequence, Crum *et al.* [22] addressed this problem by devising a teacher-student framework involving three main steps. First, a small amount of available human-annotated data is adopted to train a teacher model. Then, the teacher model generates saliency maps for a large amount of new training data. Finally, a student model is trained on the samples annotated by the teacher. It is proved that the student model surpasses the performance of models trained only with limited human saliency.

The principle of incorporating human supervision can also be particularly beneficial in tasks where subtle visual cues are crucial, such as in the detection of deepfakes. Following this direction, recent studies have started the integration of human visual attention into deepfake detection models. Korshunov *et al.* [36] highlighted the discrepancies between human and algorithmic detection capabilities, suggesting that models could benefit from mimicking human perceptual strategies. Cartella *et al.* [14] conducted an eye-tracking experiment demonstrating that humans are particularly susceptible to the alteration of images. Their findings suggest a difference in the human observational pattern when looking at real or fake images. Building on these recent works, in this paper, we focus on the detection of deepfakes within a religious context. The application of human attention and saliency models in detecting partially manipulated religious images is still underexplored. In fact, religious images hold significant cultural and emotional value, making the detection of alterations critical to preserving the integrity of the cultural heritage.

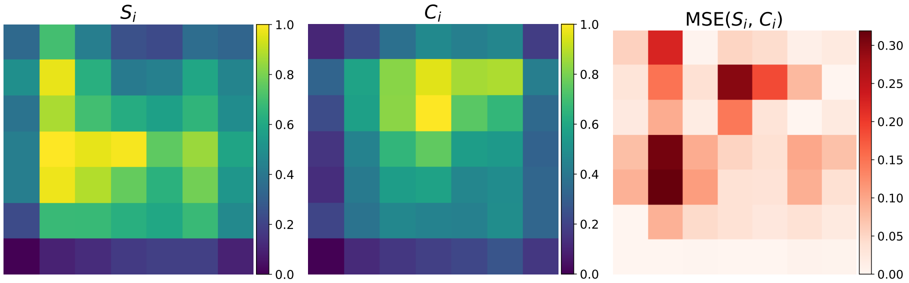
Deepfake Detection. The field of deepfake detection has seen significant advancements in recent years, primarily driven by the continuous evolution of image generation techniques. Early efforts in this area focused on detecting GAN-generated faces [49, 55, 59]. As image generation methods evolved to include diffusion models [44, 48], detection techniques expanded accordingly. Researchers expanded their focus to include natural images beyond the facial domain [2,4,25]. This adaptation marked a significant shift, demonstrating the need for versatile detection methods capable of handling various generation techniques.

In this context, a distinct line of research has utilized frequency domain analysis, capitalizing on the unique spectral characteristics that differentiate real from generated images [20, 27]. Another approach [58] focused instead on the discrepancies between input images and their reconstructions produced by pre-trained diffusion models. To increase the robustness of deepfake detection methods to unseen generators, recent strategies have tackled the deepfake detection problem by leveraging CLIP [46] to extract visual features for the task [2, 16, 21, 43, 51]. These approaches emphasize the visual patterns identified by the CLIP backbone rather than its semantic text-image alignment properties.

As previously mentioned, only a few attempts have been made to include human visual attention in the deepfake detection task. Among them, Boyd *et al.* [11] proposed a training strategy to address the task by incorporating human visual annotations into a loss function. The acquisition of the training set required subjects to manually annotate regions in the presented images, while providing an answer on which image is either the synthetic or real one, in a two-alternative forced choice manner. Following the major trend in standard deepfake detection literature [6,38,49,60], this work [11] focused on the recognition of face manipulation. Differently from previous research, we tackle deepfake detection by relying on training data directly collected from an eye-tracking experiment [14] and present a novel solution to detect partially manipulated religious images, a domain never addressed before in existing literature.

3 Proposed Approach

The main goal of our study is to assess, from an objective standpoint, the potential benefits of incorporating visual attention information into the design of models for the automatic identification of manipulated images.



(a) Saliency Map **(b)** Class Activation Map **(c)** Mean Squared Error

Fig. 2: Visualization of mean squared error between saliency map (S_i) and Class Activation Map (C_i) for a manipulated image x_i.

Starting with the basic formulation, let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be a set of N input images and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ their corresponding labels, where $y_i \in \{0, 1\}$ indicates whether the i-th image is genuine (0) or manipulated (1). The model employs a visual backbone to extract features, denoted by function $f_\theta(\cdot)$, where θ represents the learnable parameters. The feature vector extracted from the backbone for image x_i is given by $z_i = f_\theta(x_i)$. A classifier $g_\phi(\cdot)$ with parameters ϕ is then applied to these features, producing a probability distribution over the two classes: $p_i = g_\phi(z_i)$. The model is therefore trained to minimize the binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1 - y_i)\log(1 - p_i)], \tag{1}$$

where p_i represents the predicted probability of the i-th image being manipulated. The parameters θ of the visual backbone and ϕ of the classification head are optimized jointly using stochastic gradient descent to minimize the binary cross-entropy loss over the training dataset.

Inspired by [11], we incorporate human generalization capabilities with insights from visual attention models within the aforementioned framework. In particular, given an image x_i we aim to focus on specific regions that are likely to be informative for the classification task. This information can be represented as an empirical fixation density S_i^h, obtained by convolving the fixation locations of all the observers on x_i by an isotropic bidimensional Gaussian function, or via an ideal saliency model, defined as $S_i^m = \mathcal{M}(x_i)$, where $\mathcal{M}(\cdot)$ is a function that maps an input image to a saliency map based on low-level visual features and higher-order statistics.

To leverage this information, we compute Class Activation Maps (CAMs) [61] for each image, denoted as $C_i = h_\psi(z_i)$, where $h_\psi(\cdot)$ is the CAM generation function with parameters ψ. The model is therefore trained to minimize a combined loss function:

$$\mathcal{L} = \lambda\mathcal{L}_{BCE} + (1 - \lambda)\mathcal{L}_{Sal}, \tag{2}$$

where $\mathcal{L}_{Sal}$ is the saliency loss that measures the discrepancy between the saliency map and the CAM. Formally, this is defined as follows:

$$\mathcal{L}_{Sal} = \frac{1}{N} \sum_{i=1}^{N} D(S_i, C_i), \tag{3}$$

where $D(\cdot, \cdot)$ is a distance function (*e.g.* mean squared error or KL divergence) between $S_i \in \{S_i^h, S_i^m\}$ and the CAM C_i related to the image x_i, while $\lambda \in [0, 1]$ is an hyperparameter that balances the two loss components.

To provide a graphical representation of the distance function, we report in Fig. 2 the mean squared error obtained from a generic manipulated image x_i. Here S_i refers to the saliency map while C_i is the 7×7 pixels feature map obtained from one of the last layers of a CNN. Minimizing Eq. (2) encourages the model to learn features that effectively distinguish between real and fake images (through the classification loss $\mathcal{L}_{BCE}$) while focusing on image regions that are likely to be informative for the classification task, guided by the saliency map and the CAM similarity term.

4 Testing Dataset

4.1 Dataset Construction

To evaluate the effectiveness of the proposed approach, we introduce a novel testing dataset specifically designed to address the distinctive challenges posed by religious images and, more broadly, partially modified ones.

Our proposed dataset is composed of images representing the iconography attested within Marian devotion and the phenomena of Marian apparitions over the centuries, that could be used in the study of devotional phenomena and the perception of religious images in the contemporary age. In particular, we collect 50 images of Marian apparitions related to 12 different religious representations. Examples of the collected images are depicted in Fig. 1a. Our objective is to define a manipulation process that exclusively focuses on the context of the religious representation. Indeed, the partial modification of the original image, coupled with the use of a conservative mask, precludes the introduction of awkward elements or distortions on the sacred figure, thus making the detection task particularly challenging also from a semantic standpoint.

In detail, given an image x_i, we extract the segmentation mask s_i (see Fig. 1b) pertaining to the individuals depicted in the scene through the SAM model [34]. In the editing phase, we employ the inpainting pipeline of the Stable Diffusion XL model (SDXL [44])[3] to generate manipulated images. This process takes as input the original image x_i, a segmentation mask representing the inpainting regions (*i.e.* in our case the inverse of s_i since we want to preserve the religious figures), and a textual prompt t_i which guides the regeneration of the image context while keeping the main subjects unchanged. Specifically, t_i is randomly

[3] `https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1`

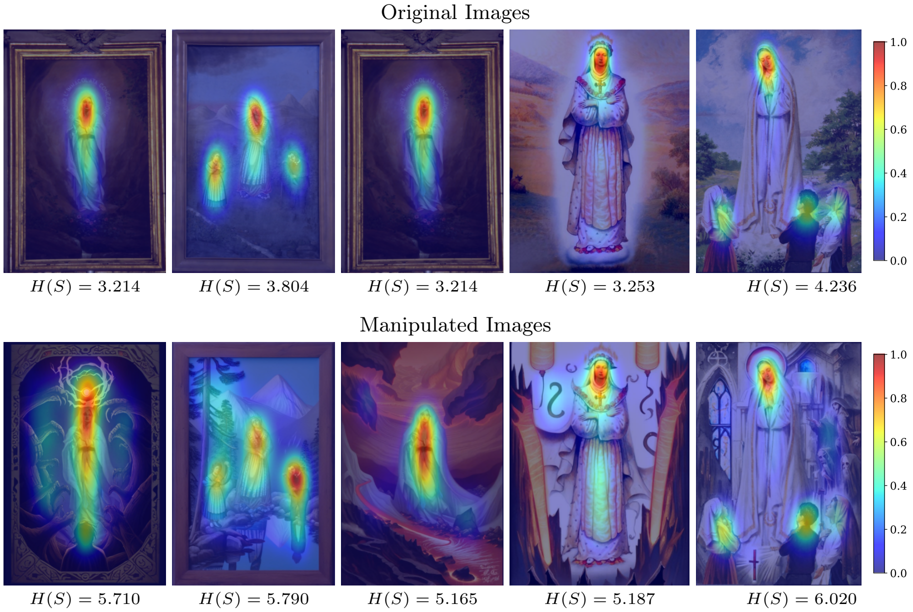


Fig. 3: Examples of the computed saliency maps before (top) and after (bottom) the manipulation. H represents the entropy while S is the saliency map predicted by [3].

sampled from a template of possible prompts that we manually devise for the scope. Examples of t_i are: *"Peaceful orchard with fruit-laden trees and buzzing bees"*, *"Tranquil garden with blooming flowers and greenery"*. Each of the 50 original images is linked to six distinct contextual modifications, resulting in a total of 350 examples (Fig. 1c).

4.2 Dataset Analysis

Considering the manipulation pipeline previously described, we aim to understand the variation in the focus of attention between a real image and its altered counterpart. We demonstrate that the difference in the attentive pattern across the images of our constructed dataset is of considerable importance. By utilizing TempSAL [3], a state-of-the-art saliency predictor, we generate saliency maps for both real and altered images, to simulate human visual attention. The qualitative analysis presented in Fig. 3 provides evidence of a discernible attentive pattern across the two image categories. Compared to the real samples, the generation of the surrounding context through a diffusion-based model causes a significant shift towards the inpainted regions.

To quantify this observed difference in attention distribution, we employ entropy as a measure of dispersion for our saliency maps. Entropy, in this context, provides an indication of how focused or dispersed the attention of the model is across the image. A lower entropy value suggests that the attention of the model is concentrated on specific regions, while a higher entropy indicates a

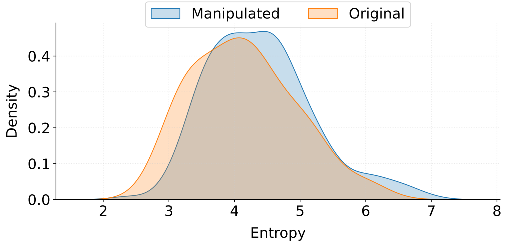


Fig. 4: Comparison of saliency map entropy distributions between original and manipulated images.

more uniform distribution of attention across the image. Given a saliency map S of size $m \times n$, we first normalize it to form a probability distribution:

$$P(i,j) = \frac{S(i,j)}{\sum_{x=1}^{m}\sum_{y=1}^{n} S(x,y)}, \tag{4}$$

where $P(i,j)$ represents the normalized saliency value at position (i,j). The entropy H of the saliency map is then calculated as:

$$H = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(i,j)\log_2 P(i,j). \tag{5}$$

This formulation yields entropy values in bits. A saliency map with perfectly uniform attention would have a maximum entropy of $\log_2(mn)$, while a map with all attention focused on a single pixel would have a minimum entropy of 0. As a consequence of the attention shift towards inpainted regions in synthetic images, the entropy of the corresponding saliency maps presents higher values, as indicated in Fig. 4 where the empirical distributions of the saliency entropy values are compared to the real case. This higher entropy in synthetic images suggests a more dispersed attention pattern, likely due to the model detecting inconsistencies or artifacts introduced by the inpainting process across a wider area of the image. In contrast, the lower entropy observed in real images indicates more focused attention, possibly on natural salient features of the original, unaltered image, as the faces of the people featured. This quantitative entropy analysis corroborates our qualitative observations, providing a numerical basis for the difference in attention patterns between real and synthetic images. It underscores the potential of entropy as a discriminative feature in distinguishing between original and manipulated images, offering insights into the spatial characteristics of model attention.

5 Experiments

5.1 Experimental Setting

Datasets. As previously described, the training procedure entails the calculation of a saliency loss, which employs images and associated saliency maps representative of the region of attention. Consequently, a dataset comprising both types of data, collected in the context of partially modified images is essential for effective learning. The Unveiling the Truth (UTruth) dataset presented in [14] represents the only collection of both real and manipulated images paired with visual attention data obtained through eye-tracking. The dataset consists of 400 images: 100 authentic and 300 partially modified using diffusion models. The modifications were carefully executed to preserve the realism, semantics, and context of the original images. During an eye-tracking experiment, each image was viewed by five distinct subjects, resulting in a saliency map that combines all the recorded fixations for every image in the dataset. To achieve a balanced ratio between real and manipulated images, we supplement this initial set with 200 authentic images and their corresponding saliency maps from CAT2000 [9], a widely-used saliency dataset that features images from different categories.

However, the limited number of available images resulting from the onerous process of data acquisition using an eye-tracker led us to integrate into the training dataset $2,000$ images collected in D^3 [4], a comprehensive dataset designed for large-scale deepfake detection. D^3 includes 9.2 million generated images created using four state-of-the-art diffusion model generators and their original counterpart. In particular, we randomly selected $1,000$ original samples and $1,000$ corresponding generated versions among all the images with at least 400 pixels on each side. Given that images from D^3 do not include human fixations, we equip them with computational saliency maps obtained from TempSAL [3], a recent saliency prediction model which can well simulate human attention.

Architecture and Training Details. Our approach utilizes a ResNet-50 architecture [29] as the visual backbone, pre-trained on ImageNet-1k [23]. We modify the final fully connected layer to output two classes, corresponding to "real" and "manipulated" images. The model is trained using stochastic gradient descent with a momentum of 0.9 and weight decay of 1×10^{-6}. We employ a step learning rate scheduler, initially setting the learning rate to 0.005 and reducing it by a factor of 0.1 every 12 epochs. The training process spans 50 epochs with a batch size of 32. As described in Eq. (2), our loss function is a weighted combination of two components: binary cross-entropy loss for classification $\mathcal{L}_{BCE}$ and mean squared error as saliency loss $\mathcal{L}_{Sal}$ prediction. The best combination of the two loss components is obtained with $\lambda = 0.9$.

To visualize the focus of the model, we implement Class Activation Mapping (CAM) [61] by extracting features from the final convolutional layer of ResNet-50. The CAM is computed by weighting these features with the corresponding class weights from the final fully connected layer. This allows us to generate heatmaps highlighting the regions most influential in the model decision-making process. All input images are resized to 224×224.

Evaluation Metrics. Our evaluation methodology is tailored to address the significant imbalance in our testing dataset, which comprises 50 original images and 400 fake images. This 1:8 ratio necessitates a careful selection of evaluation metrics that are robust to class imbalance.

We prioritize Average Precision (AP) as our primary metric. AP provides a single-value summary of the precision-recall curve, offering a concise yet compre-

Table 1: Accuracy results on our testing dataset comparing our solution with state-of-the-art deepfake detection methods. Precision and recall are not in bold as they are not individually significant in the case of an unbalanced dataset. Ours w/o $\mathcal{L}_{Sal}$ means our approach with $\lambda = 1$, while Ours w/ $\mathcal{L}_{Sal}$ means our approach with $\lambda = 0.9$ which represents our best configuration.

	AP	**AUC-PR**	**Precision**	**Recall**	**AUC-ROC**
Grag2021 [28]	87.81	87.69	100.00	0.25	51.68
Wang2020 [56] (Blur+JPEG 0.5)	90.44	90.38	100.00	0.25	56.78
Wang2020 [56] (Blur+JPEG 0.1)	92.28	92.25	100.00	0.50	61.66
DIRE [58] (LSUN/ADM)	85.83	85.65	75.00	3.00	43.14
DIRE [58] (LSUN/PNDM)	87.79	87.94	89.04	95.50	47.15
Ours w/o $\mathcal{L}_{Sal}$	92.97	92.95	92.79	48.25	62.13
Ours w/ $\mathcal{L}_{Sal}$	**95.35**	**95.34**	94.76	63.25	**72.93**

hensive measure of model performance. It is particularly informative in our imbalanced scenario, as it emphasizes the model ability to detect the minority class (*i.e.* original images) without being influenced by the majority class (*i.e.* manipulated images). Complementing AP, we utilize the Area Under the Precision-Recall Curve (AUC-PR) that offers a more detailed view of the precision-recall trade-off across different thresholds. Given the prevalence of manipulated images in our dataset, Precision becomes a critical metric. It quantifies the proportion of correct fake detections among all images classified as fake, thereby measuring our model ability to avoid false alarms. This is particularly important given the potential real-world implications of misclassifying original images as fake. Complementing Precision, we also consider Recall (or sensitivity). Recall indicates the proportion of actual fake images successfully identified by our model. With the large number of fake images in our dataset, a high recall ensures that we detect a significant portion of the manipulated content.

Lastly, we include the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for completeness. While AUC-ROC is a popular metric that measures the model discriminative ability across various classification thresholds, we interpret it cautiously in our context. AUC-ROC can potentially overestimate performance on imbalanced datasets like ours, as it is less sensitive to class imbalance compared to precision-recall based metrics.

5.2 Comparison with Standard Deepfake Detection Methods

We assess the performance of our model in detecting real and fake images on our proposed testing dataset. In particular, we compare our model against state-of-the-art approaches, namely the model proposed by Gragnaniello *et al.* (Grag2021) [28], the one introduced by Wang *et al.* (Wang2020) [56], and the DIRE architecture [58]. The results for deepfake detection are reported in Table 1 and plotted in Fig. 5. Significantly, our fake detection approach surpasses state-of-the-art detectors across all key metrics, including AP, AUC-PR, and AUC-ROC. Although the considered approaches achieve higher absolute precision or recall values, it is important to note that in the context of an imbal-

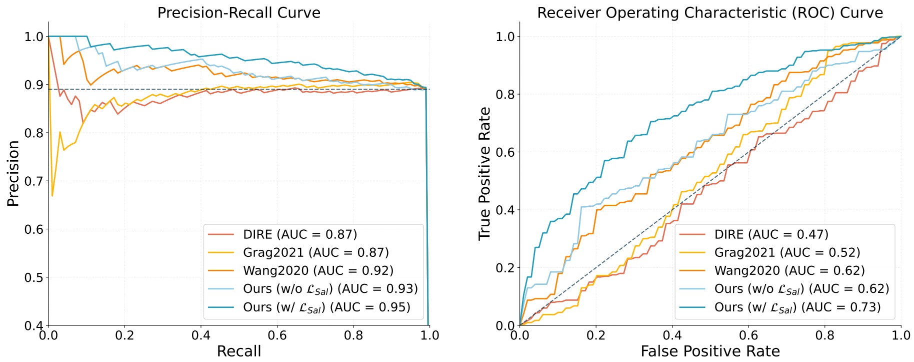


Fig. 5: Performance comparison of image manipulation detection methods on our testing dataset. We report the Precision-Recall curve (left) and ROC curve (right). The dashed line indicates the performance of a random classifier.

anced dataset, precision and recall individually may not provide a comprehensive evaluation. In addition to these results, in Table 1 we also report a comparison of our model with and without the saliency loss (cf. Eq. (2)). It is clear that augmenting training with saliency maps representing the human focus of attention leads to better performance. Saliency maps represent a guide for the detector that during training learns to adjust its internal attention representation towards regions that are the most discriminative for the real-fake classification, thus mimicking the human behavior.

5.3 Ablation Studies

Analyzing the Impact of the Training set. In the previous section, we demonstrated the efficacy of our method in detecting manipulated images depicting religious representations. The integration of saliency within the fake detection pipeline proved to be fundamental. In this section, we conduct an ablation study to shed light on the contributions of human attention in contrast to the attention maps generated by a predictive model. In Table 2, we compare different training strategies considering three combinations of training sets. The first row refers to a setting where we employ 300 real and 300 fake images from the UTruth dataset presented in [14] and CAT2000 *et al.* [9], along with their corresponding human saliency maps. In the second case, we train our model on $1,000$ real and $1,000$ fake images derived from D^3 [4]. Although training on D^3 achieves superior performance, the approach based entirely on human attention remains competitive. We attribute this performance gap to the different sizes of the two training sets. Indeed, focusing on the latter training setting in the table, where the dataset includes a combination of real and generated saliency maps, the performance gain is maximized. Such results prove that the integration of human attention is beneficial for the detection of partially manipulated images.

Assessing the Generalizability across Different Backbones. We conduct extensive experiments across various visual backbones to evaluate the robustness of the proposed method. In detail, we report in Table 3 the results for four

Table 2: Performance comparison of different training set strategies.

Human Saliency		**Computed Saliency**					
UTruth [14]	CAT2000 [9]	D^3 [4]	**AP**	**AUC-PR**	**Precision**	**Recall**	**AUC-ROC**
✓	✓	-	89.97	89.93	89.72	96.00	56.06
-	-	✓	93.57	93.55	94.40	29.50	64.31
✓	✓	✓	**95.35**	**95.34**	94.76	63.25	**72.93**

Table 3: Performance comparison of various visual backbone architectures.

	$\mathcal{L}_{Sal}$	**AP**	**AUC-PR**	**Precision**	**Recall**	**AUC-ROC**
ResNet-18	✗	91.29	91.26	92.00	34.50	57.67
	✓	**91.75**	**91.72**	90.32	56.00	**58.00**
ResNet-50	✗	92.97	92.95	92.79	48.25	62.13
	✓	**95.35**	**95.34**	94.76	63.25	**72.93**
DenseNet-201	✗	92.35	92.33	91.60	57.25	61.96
	✓	**95.17**	**95.16**	98.00	24.50	**71.73**
Swin-Tiny	✗	90.54	90.50	89.81	35.25	54.86
	✓	**92.40**	**92.38**	94.26	28.75	**57.94**

backbones, each trained with ($\lambda = 0.9$) and without ($\lambda = 1$) the incorporation of saliency maps. In addition to reporting the performance of ResNet-50, we consider ResNet-18 [29], DenseNet-201 [31], and a Vision Transformer-based backbone [24] like Swin-Tiny [39]. As it can be seen, comparing the results of the considered backbones, ResNet-50 leads to the best results according to all evaluation metrics. Most importantly, the adoption of the saliency loss leads to better detection performance across all the key metrics, regardless of the backbone employed thus further demonstrating the effectiveness of human attention for the deepfake detection task.

Analyzing the Impact of the λ Parameter. Fig. 6 illustrates the impact of varying the lambda parameter from Eq. (2) on the accuracy of detecting real and manipulated images, as well as the balanced accuracy between the two. Here, we consider the best configuration of our approach obtained with a ResNet-50 backbone. As lambda increases from 0.01 to 1.0 (representing less weight given to the saliency loss), we observe that the accuracy for real images steadily increases, reaching a peak around $\lambda = 0.9$. The accuracy for manipulated images sharply decreases, especially for λ values above 0.7. The balanced accuracy shows a gradual improvement up to $\lambda = 0.9$, after which it declines with $\lambda = 1.0$ which corresponds to a training without the saliency loss.

Notably, the performance in detecting manipulated images drops significantly as lambda increases, indicating that reducing the influence of the saliency loss negatively impacts the model ability to identify manipulated content. This underscores the importance of combining both class loss and saliency loss for optimal performance. The balanced accuracy peaks at $\lambda = 0.9$, suggesting this is the optimal value for balancing the detection of both real and manipulated images. Beyond this point, corresponding to a global loss entirely based on the classification loss, the overall performance degrades.

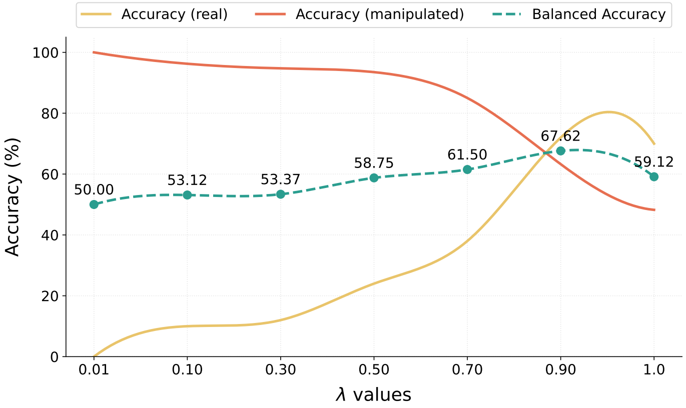


Fig. 6: Impact of lambda parameter on image classification accuracy. The optimal balance between real and manipulated image detection is achieved at $\lambda = 0.9$, demonstrating the benefit of combining classification and saliency losses.

6 Conclusion

This study demonstrates the efficacy of integrating human visual attention into the automatic detection of manipulated religious images. Our experiments strongly support the hypothesis that human-derived saliency maps can significantly enhance the performance of deepfake detection models, particularly for images that are visually obvious to human observers but challenging for current algorithms. The introduction of our novel testing dataset, specifically designed for religious visual data, has provided valuable insights into the unique challenges posed by partially modified images in this domain. Notably, our results show that even a small amount of eye-tracking data can yield substantial improvements compared to relying solely on generated data from saliency models. This finding underscores the value of human-in-the-loop approaches in AI development and suggests that expanding this dataset would be highly beneficial for further advancements in the field. Future work should prioritize the collection of more extensive eye-tracking data across diverse scenarios and cultural contexts to enhance the robustness and generalizability of these models. As generative AI technologies continue to evolve, the approach presented in this paper offers a promising direction for future research, suggesting that further exploration of human-AI collaboration could lead to more robust, context-aware, and ethically aligned deepfake detection systems.

Acknowledgments

This work has been supported by the PNRR project “Italian Strengthening of Esfri RI Resilience (ITSERR)” funded by the European Union - NextGenerationEU (CUP B53C22001770006).

References

1. Aberman, K., He, J., Gandelsman, Y., et al.: Deep saliency prior for reducing visual distraction. In: CVPR (2022)
2. Amoroso, R., Morelli, D., Cornia, M., Baraldi, L., Del Bimbo, A., Cucchiara, R.: Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. ACM TOMM (2024)
3. Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., Süsstrunk, S.: TempSAL - Uncovering Temporal Information for Deep Saliency Prediction. In: CVPR (2023)
4. Baraldi, L., Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R.: Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities. In: ECCV (2024)
5. Berg, A., Valaskivi, K.: Representational silence and racial biases in commercial image recognition services in the context of religion. In: Handbook of Critical Studies of Artificial Intelligence, pp. 607–618. Edward Elgar Publishing (2023)
6. Boccignone, G., Bursic, S., Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R., Patania, S.: DeepFakes Have No Heart: A Simple rPPG-Based Method to Reveal Fake Videos. In: ICIAP (2022)
7. Boccignone, G., Cuculo, V., D'Amelio, A.: How to look next? a data-driven approach for scanpath prediction. In: Formal Methods. FM 2019 International Workshops (2020)
8. Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R.: On Gaze Deployment to Audio-Visual Cues of Social Interactions. IEEE Access **8**, 161630–161654 (2020)
9. Borji, A., Itti, L.: CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. arXiv preprint arXiv:1505.03581 (2015)
10. Boyd, A., Bowyer, K.W., Czajka, A.: Human-Aided Saliency Maps Improve Generalization of Deep Learning. In: WACV (2022)
11. Boyd, A., Tinsley, P., Bowyer, K.W., Czajka, A.: CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning-Based Synthetic Face Detection. In: WACV (2023)
12. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The Revolution of Multimodal Large Language Models: A Survey. In: ACL Findings (2024)
13. Cartella, G., Cornia, M., Cuculo, V., D'Amelio, A., Zanca, D., Boccignone, G., Cucchiara, R.: Trends, Applications, and Challenges in Human Attention Modelling. In: IJCAI (2024)
14. Cartella, G., Cuculo, V., Cornia, M., Cucchiara, R.: Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. IEEE SPL **31**, 820–824 (2024)
15. Chen, S., Jiang, M., Yang, J., Zhao, Q.: AiR: Attention with Reasoning Capability. In: ECCV (2020)
16. Cocchi, F., Baraldi, L., Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis. In: ICIAP (2023)
17. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience **3**(3), 201–215 (2002)
18. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. ACM TOMM **14**(2), 1–21 (2018)

19. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: SAM: Pushing the Limits of Saliency Prediction Models. In: CVPR Workshops (2018)
20. Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., Verdoliva, L.: Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models. In: CVPR Workshops (2023)
21. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the Bar of AI-generated Image Detection with CLIP. In: CVPR Workshops (2024)
22. Crum, C.R., Boyd, A., Bowyer, K., Czajka, A.: Teaching AI to Teach: Leveraging Limited Human Salience Data Into Unlimited Saliency-Based Training. arXiv preprint arXiv:2306.05527 (2023)
23. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR (2021)
25. Epstein, D.C., Jain, I., Wang, O., Zhang, R.: Online Detection of AI-Generated Images. In: ICCV Workshops (2023)
26. Fosco, C., Casser, V., Bedi, A.K., O'Donovan, P., Hertzmann, A., Bylinskii, Z.: Predicting visual importance across graphic design types. In: ACM UIST (2020)
27. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: ICML (2020)
28. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L.: Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In: ICME (2021)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
31. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: CVPR (2017)
32. Ilaslan, M., Song, C., Chen, J., Gao, D., Lei, W., et al.: GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations. In: EMNLP (2023)
33. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. PAMI **20**(11), 1254–1259 (1998)
34. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment Anything. In: ICCV (2023)
35. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: DiffWave: A Versatile Diffusion Model for Audio Synthesis. In: ICLR (2021)
36. Korshunov, P., Marcel, S.: Subjective and objective evaluation of deepfake videos. In: ICASSP (2021)
37. Leiva, L.A., Xue, Y., Bansal, A., Tavakoli, H.R., et al.: Understanding visual saliency in mobile user interfaces. In: ACM MobileHCI (2020)
38. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: CVPR Workshops (2018)
39. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: CVPR (2021)
40. Lou, J., Lin, H., Marshall, D., et al.: TranSalNet: Towards perceptually relevant visual saliency prediction. Neurocomputing **494**, 455–467 (2022)

41. Maiano, L., Benova, A., Papa, L., Stockner, M., Marchetti, M., Convertino, G., et al.: Human versus Machine: A Comparative Analysis in Detecting Artificial Intelligence-generated Images. IEEE Security & Privacy **22**(3), 77–86 (2024)
42. Miangoleh, S.M.H., Bylinskii, Z., Kee, E., Shechtman, E., Aksoy, Y.: Realistic Saliency Guided Image Enhancement. In: CVPR (2023)
43. Ojha, U., Li, Y., Lee, Y.J.: Towards Universal Fake Image Detectors That Generalize Across Generative Models. In: CVPR (2023)
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv preprint arXiv:2307.01952 (2023)
45. Qiao, M., Liu, Y., Xu, M., Deng, X., Li, B., Hu, W., Borji, A.: Joint learning of audio–visual saliency prediction and sound source localization on multi-face videos. IJCV **132**(6), 2003–2025 (2024)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
47. Rensink, R.A.: The dynamic representation of scenes. Visual cognition **7**(1-3), 17–42 (2000)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
49. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: ICCV (2019)
50. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: CVPR (2023)
51. Sha, Z., Li, Z., Yu, N., Zhang, Y.: DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In: ACM CCS (2023)
52. Skerath, L., Toborek, P., Zielińska, A., et al.: Native Language Prediction from Gaze: a Reproducibility Study. In: ACL Workshops (2023)
53. Sood, E., Kögel, F., Müller, P., et al.: Multimodal Integration of Human-Like Attention in Visual Question Answering. In: CVPR Workshops (2023)
54. Takmaz, E., Pezzelle, S., Beinborn, L., Fernández, R.: Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In: EMNLP (2020)
55. Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y.: FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In: IJCAI (2020)
56. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In: CVPR (2020)
57. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019)
58. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for Diffusion-Generated Image Detection. In: ICCV (2023)
59. Yang, X., Li, Y., Lyu, S.: Exposing Deep Fakes Using Inconsistent Head Poses. In: ICASSP (2019)
60. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: CVPR (2021)
61. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)