I-SPAI: A Human Perception Dataset for Deepfake Detection and Perception-Aware Auxiliary Supervision

Anja Hrvatič[1], Peter Peer[1], Vitomir Štruc[2], Borut Batagelj[1]
[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
ah40785@student.uni-lj.si

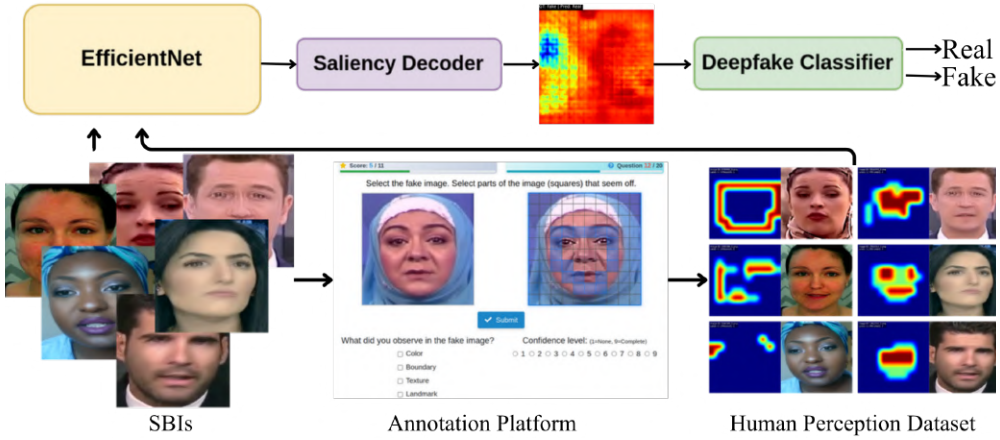

Figure 1. **Overview of the I-SPAI dataset collection and perception-aware detection framework.** Top: The consecutive saliency decoder architecture. An EfficientNet-B4 backbone extracts features which are passed to a Saliency Decoder, producing a predicted saliency map. The map features are then passed to the Deepfake Classifier for binary real/fake prediction. Bottom: The I-SPAI annotation pipeline. SBI-generated image pairs (left) are presented to participants via a custom annotation platform (center), where they identify the fake image, mark suspicious regions on a 10×10 grid, select inconsistency types, and rate their confidence. The resulting annotations (right) produce aggregated spatial saliency maps used as auxiliary supervision during training.

Abstract

*Deepfake detectors typically overfit to low-level artifacts of the generative models seen during training, leading to poor generalization on unseen manipulation techniques. This work explores human visual perception as an alternative and more stable source of supervision. We introduce I-SPAI (**I**nconsistency, **S**aliency, and **P**erception **A**nnotations for Deepfake **I**dentification), the first publicly available deepfake dataset with human perception annotations, consisting of 2,880 SBI-generated image pairs labeled by 435 participants with inconsistency types, spatial saliency maps, and confidence scores. Using I-SPAI, we propose a multitask detection framework and conduct a systematic evaluation of three types of human perceptual supervision: categorical inconsistency labels, spatial saliency map prediction, and indirect saliency-based loss supervision via CYBORG. Results on four benchmark datasets show that not all human perception signals are equally effective. Spatial saliency maps incorporated via a consecutive decoder provide the most consistent cross-dataset improvement, while categorical and indirect supervision offer limited gains. This suggests that spatial localization is the most transferable form of human perceptual knowledge for deepfake detection, and that the form of the supervision signal matters as much as its source.*

1. Introduction

In recent years, the rise of manipulated facial images has posed significant challenges to digital security and public trust. The rapid advancement of generative modeling, specifically Generative Adversarial Networks (GANs) and diffusion models, has fundamentally challenged the integrity of visual media. These models enable the synthesis of highly realistic facial manipulations, or deepfakes, which can convincingly alter both identity and expression. In the

era of easy access to such technology and widespread social media consumption, manipulated content can quickly reach large audiences, fueling misinformation and posing a direct threat to biometric authentication systems. Consequently, robust and generalizable deepfake detection has become a critical research challenge [6, 7, 28].

Deepfake detection has traditionally relied on detecting low-level architectural fingerprints and pixel-level inconsistencies left behind during the synthesis process. While many existing frameworks [1, 17, 22, 47] achieve high intra-dataset accuracy, they often struggle with a significant generalization gap. These methods tend to overfit to specific artifacts that are characteristic of the generative models used to produce training images. As deepfake technology improves, those artifacts are removed, leading to a decline in performance in cross-dataset scenarios. This cycle shows the need for more robust, manipulation-agnostic detection methods that depend on stable features rather than specific flaws of synthesis algorithms [41].

An alternative and potentially more stable source of supervision lies in human visual perception [4, 5, 9]. Humans rely on holistic reasoning and an intuitive understanding of physical and social norms, making them sensitive to high-level semantic inconsistencies such as unnatural hairlines, illogical lighting, or overly smooth skin textures. Importantly, these perceptual judgments are not tied to the specific mechanisms used to generate the fake content, but the semantics and quality of the facial imagery, making them inherently more stable than the low-level artifacts targeted by traditional detectors.

However, it is not obvious which aspects of human perception are most useful for automated detection. Human observers can provide different types of signals: (1) *categorical judgments* about the type of artifact present, (2) *spatial attention maps* indicating where they look, or (3) *confidence scores* reflecting their certainty. Whether these signals are equally useful, and how best to incorporate them into a detection framework, remains an open question.

This work addresses this question through a systematic study and makes the following main contributions:

- It introduces **I-SPAI**, a publicly available deepfake dataset of SBI-generated image pairs [39] annotated by 435 participants with spatial saliency maps, inconsistency type labels, and confidence scores, collected via a custom annotation platform.
- It introduces a **multitask detection framework** that incorporates human perceptual annotations as auxiliary supervision signals alongside binary deepfake classification.
- It presents a **systematic evaluation** of three types of human perceptual supervision (spatial, categorical, and indirect), showing that among the three spatial saliency maps provide the most consistent cross-dataset improvement, while categorical and indirect signals offer limited gains within the studied setup.

Unlike prior work that primarily focuses on designing complex architectures and elaborate loss functions for training, the objective of this study is to isolate and evaluate the contribution of different forms of human perceptual supervision under a fixed backbone architecture.

2. Related Work

Deepfake Generation. Face manipulation techniques can be classified into four categories: full-face synthesis, face swapping, attribute manipulation, and expression transfer [16]. Face swapping replaces the identity of a person in an image or video with another, attribute manipulation alters certain facial characteristics (e.g., expression, age, makeup, etc.), full face synthesis generates novel facial images of known or unknown identities, while expression transfer (face reenactment) maps source expressions onto a target face while preserving their identity.

Early face swap methods used autoencoder-based frameworks with a shared encoder and identity-specific decoders, forming the basis of tools like DeepFaceLab [16, 32]. GAN-based approaches [29, 37] later improved realism and generalisation, with FaceShifter [29] introducing subject-agnostic training and occlusion refinement. For expression transfer, Face2Face [46] pioneered real-time expression transfer using coarse 3D face reconstructions, while ReenactGAN [48] improved structural fidelity through facial boundary transfer in a latent space. A more comprehensive review of modern deepfake generation techniques can be found in recent surveys on this topic, e.g. [8, 34]

Deepfake Detection. Deepfake detection methods broadly fall into three categories: spatial-domain, frequency-domain, and time-domain techniques [35]. Early approaches relied on detecting manually identified artifacts/cues such as incorrect lighting [44] or face warping artifacts [30], before CNN-based methods demonstrated that learned features outperform manually defined ones.

A central challenge is cross-dataset generalization: models trained on images produced by one generation method often fail on unseen ones. Self-Blended Images (SBIs) [40] address this by simulating forgery artifacts from authentic samples, training classifiers on generalized, manipulation-agnostic representations. Building on this, M-Task-SS [3] extends the self-blending paradigm with a multitask self-supervised objective to jointly predict artifact presence and location. SeeABLE [27] formalizes detection as an out-of-distribution problem, generating soft discrepancies to localize altered facial regions. WMamba [35] combines wavelet-domain analysis with a Mamba-based architecture to model facial contours and capture long-range spatial

forgery clues, achieving strong cross-dataset generalization. For temporal localization in video, W-TDL [14] fuses audio and visual signals to identify fake segments. Despite this progress, these detectors remain vulnerable to adversarial attacks such as the Degrading and Restoring (DR) attack [38], which removes high-frequency deepfake artifacts while restoring natural image statistics, and still exhibit suboptimal cross-dataset performance.

Human Perception of Deepfakes. Despite human sensitivity to facial irregularities such as unnatural textures or inconsistent expressions [15, 18, 23], deepfake detection accuracy remains modest, typically between 57.6% and 75.43% [43], with participants frequently overconfident in their judgments [42]. Demographic factors such as gender, education, income, and profession show little influence on detection performance [2, 33, 45], though some evidence suggests older participants perform worse [13]. At the stimulus level, face swaps are generally easier to detect than expression transfers [33]. Interventions such as prior familiarisation or written tips have shown limited benefit [25,26,42], whereas immediate feedback during the task consistently improves performance [10, 21].

Human-Perception-Aided Detection. Human perceptual signals have been used across computer vision to improve model generalization. Boyd *et al.* [4] demonstrated that human saliency maps can guide data augmentation for improved generalization on smaller datasets. Huang *et al.* [20] incorporated over 200,000 human reaction time measurements into a psychophysical loss function to align model processing time with human behavioral patterns. MENTOR [9] uses human saliency maps in a two-stage pretraining framework, first training an autoencoder to reproduce human attention, then fine-tuning for classification.

In deepfake detection specifically, the CYBORG loss function [5] penalizes divergence between a model's Class Activation Map and human-annotated saliency, directly guiding spatial attention during training. For multi-face scenarios, Hu *et al.* [19] identified four human perceptual cues (scene-motion coherence, inter-face appearance, gaze alignment, and face-body consistency) and incorporated them into the HICOM framework, improving detection and generalization. These works collectively motivate the integration of human perceptual signals into detection pipelines, which is the approach adopted in this work.

3. Methodology

3.1. Human Perception Dataset

To the best of our knowledge, at the time of this research no publicly available deepfake dataset included human perception annotations. We therefore constructed and publicly release the **I**nconsistency, **S**aliency, and **P**erception **A**nnotations for Deepfake **I**dentification dataset (I-SPAI) as part of this work (code and data: https://github.com/AnjaH4/I-SPAI). I-SPAI, illustrated in Figure 1, is unique in combining three complementary types of human perceptual signal (spatial saliency maps, categorical inconsistency labels, and confidence scores) collected from a diverse non-expert population under controlled annotation conditions. This makes it suitable not only for auxiliary supervision in detection models, but also as a resource for studying how humans perceive and reason about deepfake artifacts.

3.1.1 Image Dataset

Fake and real image pairs were generated using the Self-Blended Images (SBI) method [40] applied to frames extracted from the raw authentic videos of the FaceForensics++ (FF++) dataset [36]. FF++ contains 1,000 authentic videos from which 4,000 manipulated videos were created using four forgery methods (Face2Face, FaceSwap, DeepFakes, NeuralTextures). The authentic videos are split into 720 training, 140 validation, and 140 testing videos and provided at multiple H.264 compression levels. Only the raw videos were used in this work. Four frames were extracted per training video and processed with SBI, yielding 2,880 fake and 2,880 authentic training images. Four frames per video were selected to keep the total number of training images within the annotation capacity of the study. For validation, where no human annotation was needed, 32 frames per video were extracted, producing 4,480 fake and 4,480 authentic images.

SBI generates fake images entirely from authentic samples by blending a transformed version of the image, I_s, onto the original, I_t, using a deformed facial mask M. This avoids overfitting to the artifacts of any specific generation method. The use of SBI images for the annotation survey also means that the human-annotated cues reflect general blending artifacts rather than method-specific fingerprints, which is desirable for the auxiliary supervision signal. The final self-blended image I_{SB} is obtained as:

$$I_{SB} = I_s \odot M + I_t \odot (1 - M), \tag{1}$$

where I_s is the transformed source image, I_t is the target image, $M \in [0, 1]^{H \times W}$ is the deformed facial mask, and $\odot$ denotes element-wise multiplication. It has to be noted that the transformed image is generated using various process, including blurring, compression, color transforms and others, as detaile in [40].

3.1.2 Annotation Task

Participants were presented with pairs of fake and authentic images and asked to identify the fake. After selecting an image, they were required to indicate which of the four

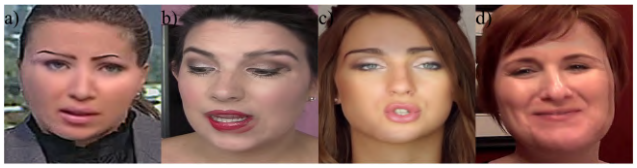

Figure 2. Examples of the four inconsistency types used in the annotation task: a) texture, b) landmark, c) colour, and d) boundary.

predefined inconsistency types they observed: texture (inconsistent or unnatural texture across the face), landmark (duplicated or misaligned facial features such as eyes, eyebrows, mouth, or nose), color (variations in skin tone or colour across different facial regions), and boundary (visible seams or blending artifacts where two facial regions meet), as illustrated in Figure 2. Participants could select any combination but were required to choose at least one. They then marked inconsistent regions on a 10×10 grid overlaid on the selected image, producing a binary saliency map. Finally, they rated their confidence on a scale from 1 to 9. Response time per question was also recorded.

3.1.3 Survey Procedure

Each participant completed 20 questions, annotating a unique set of 20 image pairs. To monitor participant reliability, the 9th and 16th questions were fixed attention checks, drawn from a pool of 11 easily distinguishable image pairs. Only participants who answered both attention checks correctly were retained, which automatically excluded those who completed fewer than 16 questions. At the start of the survey, participants were shown an introduction explaining deepfakes, the four inconsistency types with examples, and an example of a solved task with the expected answer. This introduction was included because participants were not required to have prior experience with deepfakes. After each submission, participants received immediate feedback on whether their answer was correct. This kept participants engaged and was expected to improve classification accuracy over the course of the survey, consistent with prior work [10,21]. There was no restriction on repeat participation. To reach a diverse population, the survey was distributed via email, Discord, Reddit, and SurveyCircle. In total, 435 participants contributed 6,665 responses.

3.1.4 Dataset Statistics

A total of 69.0% of participants passed both attention checks, resulting in 5,998 valid responses (90.0% of all collected annotations). Overall, 4,974 responses (87.8%) were correct. Annotations were collected for 2,780 fake images (96.5%) and 688 authentic images (23.9%), as authentic images were only annotated when a participant incorrectly selected them as fake. Saliency heatmaps are present in 5,615 responses (82.2%), as heatmap annotation was optional.

Mean classification accuracy across participants was 0.840 ± 0.010, higher than the 57.6–75.43% range typically reported for human deepfake detection [43]. This is likely because SBI artifacts are more visible than those in real deepfakes, and because the survey included an introduction to inconsistency types and immediate feedback after each response. Mean confidence was 6.75 ± 0.08, with a negatively skewed distribution indicating a tendency toward high confidence responses. Wrong answers had a mean confidence of 4.95 ± 0.12, suggesting overconfidence in incorrect decisions. Texture inconsistency was the most frequently reported category (45%), followed by landmark (34%), colour (32%), and boundary (29%). Single-type annotations were more common than any combination, indicating that participants typically identified one dominant artifact rather than multiple co-occurring ones. The dataset is publicly available under the MIT License and provides a foundation for future research on human-guided deepfake detection and the study of human perceptual behavior in the context of facial manipulations.

3.2. Core Architecture

All models use an EfficientNet-B4 backbone pretrained on ImageNet, with the original classifier replaced by a binary real/fake head. Input images are resized to 380×380 pixels. Training uses the SAM optimizer with AdamW as the base optimizer, a batch size of 32, and up to 100 epochs. Models incorporating human perception data from I-SPAI are further pretrained on SBI images before additionally training with the task-specific heads. During additional training, a 5-epoch warm-up freezes the backbone, with separate learning rates for backbone and heads to stabilize optimization. All auxiliary heads are discarded at inference, and only the binary classifier is used for prediction.

3.3. Data Preprocessing

Inconsistency type annotations were encoded as four-dimensional binary vectors, where each index corresponds to one inconsistency type (texture, landmark, colour, boundary). For authentic images, all vector elements were set to zero. For fake images with multiple annotations, the vectors were averaged across participants to produce soft labels. Saliency maps were created from the 10×10 binary grids marked by participants. Each grid was upsampled to 380×380 via bilinear interpolation and smoothed with a Gaussian filter ($\sigma = 5$) to produce continuous saliency regions. Maps were then averaged across all participants who annotated the same image and normalized to $[0, 1]$ with a small offset $\epsilon = 10^{-14}$ for numerical stability.

Confidence scores were normalized across participants to account for differences in difficulty between image pairs. In some model configurations, these scores were used to

weight the per-sample loss during training. The weight w assigned to each image pair is defined as:

$$w = \begin{cases} \bar{c} & \text{if all correct} \\ \max(0.001,\, c_{\text{cor}} - c_{\text{inc}}) & \text{if mixed} \\ 0.001 & \text{if all incorrect,} \end{cases} \quad (2)$$

where $\bar{c}$ is the mean confidence across respondents, and c_{correct} and $c_{\text{incorrect}}$ are the mean confidence scores of correct and incorrect respondents respectively. The weighting scheme emphasizes samples where participants were both correct and confident, and reduces the influence of samples where human judgment was unreliable.

4. Experimental Setup

4.1. Evaluation Datasets

Models are evaluated on four commonly used benchmark datasets covering a range of generation methods and scenarios, i.e., Celeb-DF v2, DFDC, DFDC Preview and FFIW. Celeb-DF v2 [31] contains 5,639 high-quality face swap videos of celebrities, generated with improved blending and temporal smoothing. The DFDC [11] and DFDC Preview [12] datasets were produced for the DeepFake Detection Challenge and cover a wide range of GAN-based and non-learning manipulation methods filmed in diverse real-world conditions. FFIW [49] presents a multi-face challenge where only a subset of faces per frame are manipulated. None of these datasets were used during training, making the evaluation a cross-dataset generalization test.

4.2. Evaluation Metrics

All models are evaluated at the video level. For each test video of the FF++ dataset, 32 frames are sampled, faces are extracted and scored by the model, and frame-level scores are averaged to produce a single video-level prediction. Performance is reported as ROC-AUC, as it is widely used in related work and enables direct comparison across methods. Standard error (SE) is estimated via bootstrap resampling with 1,000 repetitions to quantify uncertainty in the reported scores.

4.3. Multitask Framework and Auxiliary Heads

The goal of this study is to systematically evaluate which types of human perceptual supervision are most effective, under the hypothesis that joint training forces the backbone to learn more robust, manipulation-agnostic representations. All model variants share the EfficientNet-B4 backbone and the binary classifier. Auxiliary heads are attached during training only and discarded at inference. The following auxiliary components are evaluated:

1. Inconsistency type classification: A linear head outputs four scores corresponding to texture, landmark, colour, and boundary inconsistencies. By training the model to recognize the type of artifact present, the auxiliary task encourages the backbone to learn features that are semantically meaningful to human observers. The head is trained against soft labels derived from aggregated participant annotations, where each value reflects the proportion of respondents who identified that inconsistency type. Since multiple types can co-occur, a Binary Cross-Entropy loss is applied independently per output, treating each inconsistency type as a separate binary prediction. When confidence weighting is applied, the per-sample loss is scaled by the transformed confidence score described above.

2. Saliency map prediction: A lightweight decoder transforms the 1792-channel EfficientNet feature map into a single-channel saliency map via two transposed convolution blocks (1792→256→128 channels), bilinear upsampling to 380×380, and a final 3×3 convolution. The decoder is trained with an MSE loss against the aggregated human saliency maps from I-SPAI. This encourages the backbone to focus on the same regions that humans find suspicious. Two configurations are evaluated, as illustrated in Figure 3. In the *parallel* configuration, the decoder operates alongside the classifier, both receiving the same backbone features, with classification and saliency losses combined with equal weighting. In the *consecutive* configuration, the decoder precedes the classifier and passes its output features forward, under the assumption that fake and real images produce structurally different saliency activations, which the classifier can then exploit. Both configurations use an SBI-pretrained backbone.

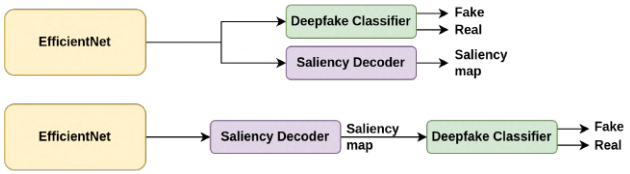

Figure 3. Saliency decoder configurations. Top: parallel configuration where the decoder and classifier receive the same backbone features. Bottom: consecutive configuration where the decoder precedes the classifier, passing its output features forward.

3. CYBORG loss: Rather than predicting saliency maps explicitly, CYBORG [5] incorporates human saliency maps directly into the loss. It replaces the standard BCE classification loss with a weighted combination of a saliency matching term and a classification term, applied to the final convolutional layer:

$$L = (1-\alpha)\left\|\mathbf{s}^{(\text{human})} - \mathbf{s}^{(\text{model})}\right\|^2 - \alpha \log p(y \in \mathcal{C}_c), \quad (3)$$

where $\mathbf{s}^{(\text{human})}$ and $\mathbf{s}^{(\text{model})}$ are the human and model saliency maps respectively, $p(y \in \mathcal{C}_c)$ is the predicted probability for the correct class, and $\alpha = 0.5$ balances the two

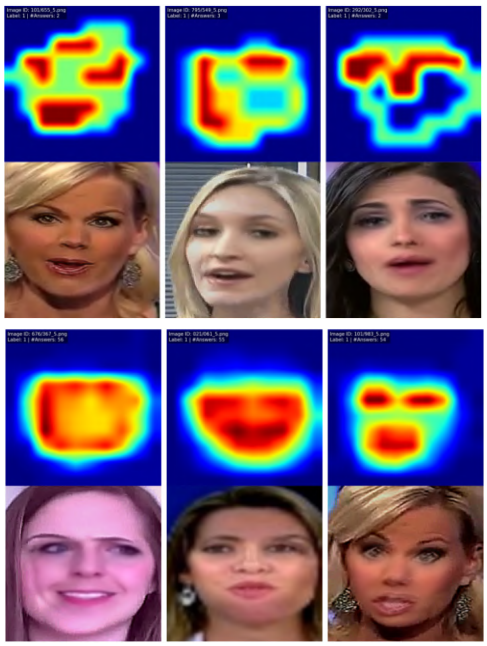

Figure 4. Aggregated saliency maps. Top: sparse maps from standard training images with approximately two annotations per image, showing geometric grid artifacts. Despite the sparsity, the training maps still localize the manipulation to the correct facial region. Bottom: dense maps from attention check images with over 50 annotations, showing smooth activations concentrated on facial regions. Red indicates high and blue indicates low importance.

terms. No additional head is required.

4. Dual auxiliary tasks: The backbone is extended with both the inconsistency type head and the saliency decoder simultaneously, with the expectation that the two supervision signals are complementary, as inconsistency types provide semantic category information while saliency maps provide spatial localization. All three losses are combined using uncertainty-based weighting [24]:

$$\mathcal{L} = \frac{1}{2\sigma_1^2}\mathcal{L}_1 + \frac{1}{2\sigma_2^2}\mathcal{L}_2 + \frac{1}{\sigma_3^2}\mathcal{L}_3 + \log \sigma_1\sigma_2\sigma_3, \quad (4)$$

where $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ are the classification, inconsistency, and saliency losses, respectively. σ_1, σ_2, σ_3 are learned uncertainty parameters that balance their contributions.

4.4. Saliency Map Quality

Prior to model evaluation, the quality of the aggregated human saliency maps from I-SPAI used as training targets was assessed. Most training images received approximately two annotations, resulting in sparse maps that retain a geometric structure reflecting the 10×10 selection grid. The attention check images, which accumulated over 50 annotations each due to repeated exposure, serve as a reference for what the maps look like with dense annotation. Figure 4 illustrates this difference. With sparse annotations, the maps exhibit higher variance and sharper grid edges. With dense annotations, the averaging process eliminates grid artifacts and produces smooth, semantically meaningful activations concentrated on facial regions such as the eyes and mouth.

When sufficient annotations are available, the maps converge toward meaningful distributions, confirming that the aggregation method produces valid localization targets. Despite the sparsity of the training maps, the highlighted regions still correspond to the correct facial areas, providing a usable localization signal for the model. The blocky appearance and high variance remain a limitation, as they introduce noise into the saliency supervision signal, which likely contributes to the limited gains observed with the saliency-based auxiliary tasks.

Table 1. Cross-dataset deepfake detection performance (ROC-AUC, higher is better) of all model variants trained on I-SPAI with SBI pretraining and evaluated on four unseen benchmark datasets. EN: EfficientNet-B4, IH: inconsistency type head, CSD: consecutive saliency decoder, PSD: parallel saliency decoder. **Bold** indicates the best result per column.

Model	CDF	DFDCP	DFDC	FFIW	Mean
EN baseline	0.872	0.870	0.684	0.821	0.812
EN + IH	**0.889**	**0.876**	0.684	0.818	0.817
EN + CSD	0.887	0.863	**0.697**	**0.843**	**0.822**
EN + PSD	0.884	0.869	0.687	0.821	0.815
EN CYBORG	0.864	0.867	0.685	0.798	0.804
EN + IH + PSD	0.873	0.873	0.688	0.826	0.815

5. Results

The quantitative results for all model variants are reported in Table 1. All models share the same EfficientNet-B4 backbone and binary classifier. Models incorporating I-SPAI annotations use an SBI-pretrained backbone as the starting point. Direct comparison to methods such as SBI [40] or WMamba [35] is not meaningful, as these methods train on substantially larger datasets. The goal of this work is not to achieve state-of-the-art performance, but to systematically evaluate which forms of human perceptual supervision are most beneficial, holding training data constant across all variants.

5.1. Baseline

The ImageNet-pretrained EfficientNet-B4 with a binary classification head achieves a mean ROC-AUC of 0.812 ± 0.006 across all four benchmarks, with the strongest result on CDF (0.872) and the weakest on DFDC (0.6844). The lower DFDC performance aligns with its known challenges, including its scale, variety of manipulation methods, and real-world recording conditions.

5.2. SBI Pretraining and Confidence Weighting

Table 2 shows the effect of backbone initialization and confidence weighting on the inconsistency type head. With-

Table 2. Effect of backbone pretraining strategy and confidence weighting on inconsistency type auxiliary supervision (ROC-AUC, higher is better). Models are evaluated on four unseen benchmark datasets. EN: EfficientNet-B4, IH: inconsistency type head, Conf.: confidence weighting, Pretrain: backbone initialization strategy. **Bold** indicates the best result per column.

Model	Pretrain	Conf.	CDF	DFDCP	DFDC	FFIW	Mean
EN baseline	ImageNet		0.872	0.870	0.684	**0.821**	0.812
EN + IH	ImageNet		0.858	0.867	0.682	0.786	0.798
EN + IH	ImageNet	✓	0.868	0.860	0.680	0.766	0.793
EN + IH	SBI		0.889	**0.876**	**0.684**	0.818	**0.817**
EN + IH	SBI	✓	**0.892**	0.874	0.683	0.811	0.815

out SBI pretraining, the auxiliary supervision is counterproductive: the ImageNet-initialized model achieves a mean ROC-AUC of 0.798 ± 0.006, below the baseline of 0.812. Initializing from the SBI-pretrained backbone recovers this gap and improves to 0.817 ± 0.006, a gain of $+0.018$. All subsequent models therefore use the SBI-pretrained backbone as the starting point.

Confidence weighting does not improve performance in either pretraining setting, with the mean ROC-AUC decreasing slightly to 0.815 ± 0.006. The weighting scheme penalizes uncertain or incorrect samples, causing the most challenging fakes to receive near-zero weight, which are precisely the samples where additional supervision would be most valuable. Given these results, confidence weighting was not applied to any subsequent configurations.

5.3. Categorical Supervision: Inconsistency Type Classification

The inconsistency type head, which provides categorical supervision, ensures only marginal improvement over the SBI baseline, achieving a mean ROC-AUC of 0.817 ± 0.006 (Table 1). The strongest result is on CDF (0.889 ± 0.015), but gains do not transfer consistently across datasets. Categorical supervision is inherently limited by the nature of the annotation task: the inconsistency categories are not mutually exclusive and their boundaries are imprecise, leading to inconsistent annotations across participants for the same image. As a result, the model does not consistently learn discriminative features corresponding to the annotated artifact categories. This suggests that categorical labels, while intuitive to collect, are too coarse and noisy to provide reliable guidance for a detection model. More structured annotation protocols with clearer category boundaries may be needed for this type of supervision to be effective.

5.4. Spatial Supervision: Saliency Map Prediction

The saliency decoder configurations produce the strongest results overall, as seen from Table 1. The parallel decoder achieves a mean ROC-AUC of 0.815 ± 0.006, while the consecutive decoder reaches 0.822 ± 0.006, the highest mean across all tested models and an improvement of $+0.010$ over the baseline. The consecutive decoder improves over the baseline on three of out of the four datasets, achieving the highest scores on DFDC (0.697) and FFIW (0.843), with the only exception being CDF.

Saliency maps provide pixel-level localization of suspicious regions, directly guiding the model toward the areas that human observers find inconsistent. Inconsistency category labels, on the contrary, apply to the entire image and do not carry spatial information. Without knowing where the artifact is, the model cannot learn to focus on relevant regions, which explains why categorical supervision produces weaker and less consistent improvements across datasets.

The performance gap between the parallel and consecutive configurations further supports this interpretation. In the parallel setup, the saliency loss only regularizes the shared backbone representation indirectly. In the consecutive setup, the classifier receives features already shaped by the saliency decoder, forcing it to operate on representations that directly reflect human localization patterns. This architectural difference provides a stronger inductive bias, which explains why the consecutive configuration outperforms the parallel one across the conducted experiments.

An inspection of the saliency maps produced by the consecutive decoder (Figure 5) shows a consistent decision pattern. For images predicted as real, high activations appear in the background rather than on the face, suggesting the model uses the absence of facial artifacts as a proxy for authenticity. For images predicted as fake, activations are less consistent and do not reliably localize the manipulated regions. Despite this, the spatial inductive bias introduced by the decoder is sufficient to improve generalization, even without precise artifact localization. The checkerboard pattern visible in the outputs is a known artifact of strided transposed convolutions and could be mitigated by replacing them with upsample-then-convolution blocks.

5.5. Indirect Supervision: CYBORG Loss

The CYBORG configurations show limited benefit. The SBI-initialized CYBORG model achieves a mean ROC-AUC of 0.804 ± 0.006, below the baseline. CYBORG represents an indirect form of spatial supervision. Rather than explicitly predicting saliency maps, it penalizes divergence between the model's class activation maps and human

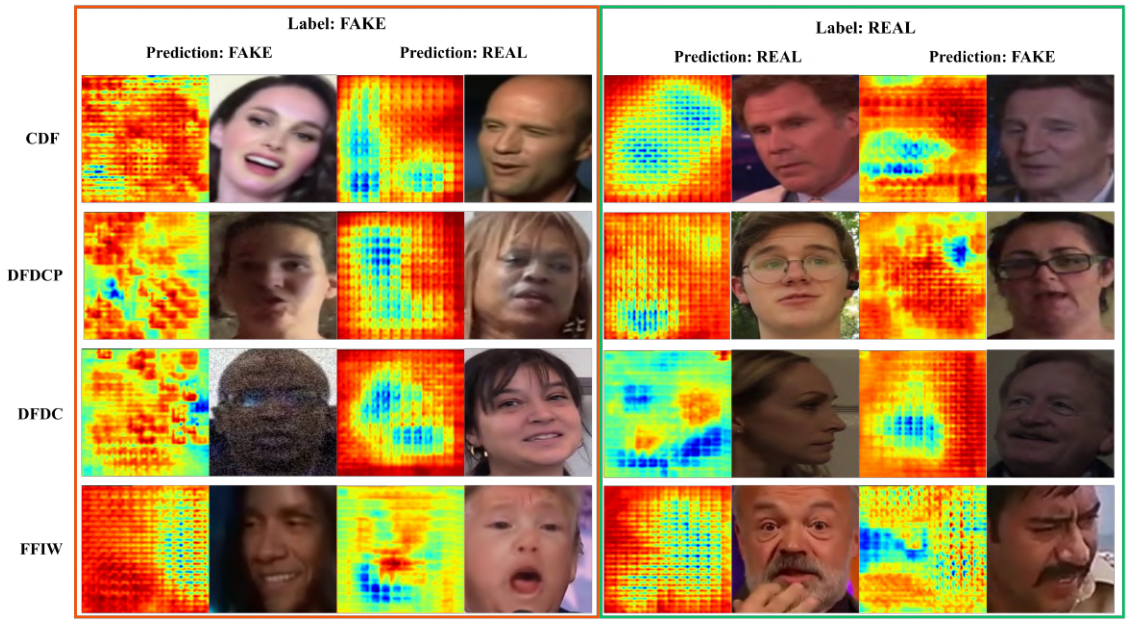

Figure 5. Saliency maps produced by the consecutive saliency decoder for images from four benchmark datasets (rows, top to bottom: CDF, DFDCP, DFDC, FFIW). Each row contains two groups: images with a red border are ground truth fake and images with a green border are ground truth real. Within each group, the left image was correctly classified and the right was misclassified. Warmer colors (red/yellow) indicate higher model activation and cooler colors (blue) indicate lower activation. For real images, high activations appear in the background rather than on the face, suggesting the model uses the absence of facial artifacts as a proxy for authenticity.

saliency maps through the loss function. This indirect signal appears less suitable for meaningfully reshaping the backbone's representations in our setup, particularly given the sparsity and noise in the I-SPAI saliency maps. The comparison with the consecutive saliency decoder suggests that explicit spatial prediction is necessary for the human localization signal to be effectively transferred to the model. Adding inconsistency-type pretraining improves CDF performance to 0.893, the highest single-dataset result across all models, but this gain does not transfer to other datasets, yielding a mean of 0.806 ± 0.006.

5.6. Combined Supervision

Combining the inconsistency type head and the saliency decoder simultaneously yields a mean ROC-AUC of 0.815 ± 0.006, only marginally above the baseline and below the consecutive saliency decoder alone. The three tasks appear to interfere with each other during training, reducing the regularization benefit of each auxiliary head. This suggests that, when auxiliary signals are noisy, combining them does not compound their benefits and may instead introduce conflicting gradients that hurt overall performance. The consecutive saliency decoder alone therefore remains the most effective configuration for incorporating human perceptual knowledge into the detection framework.

6. Conclusion

This work presented a systematic study of human perceptual supervision for deepfake detection. We introduced I-SPAI, a publicly available dataset of SBI-generated image pairs annotated by 435 participants with spatial saliency maps, inconsistency type labels, and confidence scores, and proposed a multitask detection framework that incorporates these annotations as auxiliary supervision signals.

The central finding is that not all human perception signals are equally effective. Categorical labels and indirect supervision via the CYBORG learning objective were shown to offer limited and inconsistent performance gains in our evaluation setup, while spatial saliency maps incorporated via a consecutive decoder provided the most consistent cross-dataset improvement. Saliency maps outperformed categorical labels because they provide pixel-level localization that directly guides the model toward suspicious regions, while categorical labels were shown to carry no spatial information and to be noisy due to imprecise category boundaries. This suggests that the form of the human perceptual signal matters as much as its source.

Acknowledgements

This work was supported in parts by the ARIS Research Programmes P2–0250 (B) and P2-0214 (B) as well as the ARIS project J2-50065 (A) DeepFake DAD.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *IEEE International Workshop on Information Forensics and Security (WIFS) 2018*, pages 1–26, Hong Kong, China, 12 2018. 2

[2] S. Ahmed. Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, 182:111074, 2021. 3

[3] B. Batagelj, A. Kronovšek, V. Štruc, and P. Peer. Robust cross-dataset deepfake detection with multitask self-supervised learning. *ICT Express*, 2025. 2

[4] A. Boyd, K. W. Bowyer, and A. Czajka. Human-aided saliency maps improve generalization of deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2735–2744, January 2022. 2, 3

[5] A. Boyd, P. Tinsley, K. W. Bowyer, and A. Czajka. Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6108–6117, January 2023. 2, 3, 5

[6] M. Brodarič, M. Ivanovska, D. K. Jain, P. Peer, and V. Štruc. Hcsi-net: Hierarchical cross-stream interaction for generalizable deepfake detection. In *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2026. 2

[7] M. Brodarič, W. Scheirer, D. K. Jain, P. Peer, and V. Štruc. Learning Forgery Signals from Synthetic Depth Targets for Generalizable Deepfake Detection. In *Proceedings of the International Joint Conference on Biometrics (IJCB)*, 2026. 2

[8] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu, F. S. Khan, and M. Shah. Deepfake media generation and detection in the generative ai era: A survey and outlook. *ACM Computing Surveys*, 2026. 2

[9] C. R. Crum and A. Czajka. Mentor: Human perception-guided pretraining for increased generalization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7470–7479, 2025. 2, 3

[10] A. Diel, A. Bäuerle, and M. Teufel. Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy. *Preprint*, 2024. 3, 4

[11] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 5

[12] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 5

[13] C. Doss, J. Mondschein, D. Shu, T. Wolfson, D. Kopecky, V. A. Fitton-Kane, L. Bush, and C. Tucker. Deepfakes and scientific knowledge dissemination. *Scientific reports*, 13(1):13429, 2023. 3

[14] L. Dragar, P. Rot, P. Peer, V. Štruc, and B. Batagelj. W-tdl: Window-based temporal deepfake localization. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 24–29, 2024. 3

[15] H. Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), Sep. 2022. 3

[16] T. Fernando, D. Priyasad, S. Sridharan, A. Ross, and C. Fookes. Face deepfakes–a comprehensive review. *arXiv preprint arXiv:2502.09812*, 2025. 2

[17] D. Güera and E. J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 2

[18] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4:223–233, 6 2000. 3

[19] J. Hu, S. Fan, and T. Sim. Seeing through deepfakes: A human-inspired framework for multi-face detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14517–14527, 2025. 3

[20] J. Huang, D. Prijatelj, J. Dulay, and W. Scheirer. Measuring human perception to improve open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11382–11389, 2023. 3

[21] N. Hulzebosch, S. Ibrahimi, and M. Worring. Detecting cnn-generated facial images in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, June 2020. 3, 4

[22] T. Jung, S. Kim, and K. Kim. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, 8:83144–83154, 2020. 2

[23] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997. 3

[24] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. 6

[25] N. C. Köbis, B. Doležalová, and I. Soraperra. Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), nov 2021. 3

[26] R. S. S. Kramer, M. O. Mireku, T. R. Flack, and K. L. Ritchie. Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, 4(1):28, jul 2019. 3

[27] N. Larue, N.-S. Vu, V. Štruc, P. Peer, and V. Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21011–21021, October 2023. 2

[28] N. Larue, V. Štruc, P. Peer, and N.-S. Vu. Learning the manifold of authenticity: Hybrid-curvature representation learning for generalizable deepfake detection. *IEEE Access*, pages 1–14, 2026. 2

[29] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5074–5083, 2020. 2

[30] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2

[31] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2020. 5

[32] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141:109628, 2023. 2

[33] R. Nichols, C. Rathgeb, P. Drozdowski, and C. Busch. Psychophysical evaluation of human performance in detecting digital face image manipulations. *IEEE Access*, 10:1–1, 01 2022. 3

[34] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, and D. Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 58(11):1–41, 2026. 2

[35] S. Peng, T. Zhang, L. Gao, X. Zhu, H. Zhang, K. Pang, and Z. Lei. Wmamba: Wavelet-based mamba for face forgery detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4768–4777, 2025. 2, 6

[36] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2019. 3

[37] Shao-An. faceswap-gan. https://github.com/shaoanlu/faceswap-GAN, 2017. GitHub repository. 2

[38] Z. Shi, C. Lin, Z. Zhao, P. Peer, and C. Shen. Evading deepfake detectors via adversarially degrading and restoring forged images. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. 3

[39] K. Shiohara and T. Yamasaki. Detecting Deepfakes With Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18720–18729, 6 2022. 2

[40] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18720–18729, June 2022. 2, 3, 6

[41] E. Soltandoost, R. Plesh, S. Schuckers, P. Peer, and V. Struc. Extracting local information from global representations for interpretable deepfake detection. In *Proceedings of IEEE/CFV Winter Conference on Applications in Computer Vision - Workshops (WACV-W) 2025*, pages 1–11, Tucson, USA, 2025. 2

[42] K. Somoray and D. J. Miller. Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, 149:107917, 2023. 3

[43] K. Somoray, D. J. Miller, and M. Holmes. Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, 2025(1):1833228, 2025. 3, 4

[44] J. Straub. Using subject face brightness assessment to detect 'deep fakes' (Conference Presentation). In *Real-Time Image Processing and Deep Learning 2019*, volume 10996, page 109960, 2019. 2

[45] R. Tahir, B. Batool, H. Jamshed, M. Jameel, M. Anwar, F. Ahmed, M. A. Zaffar, and M. F. Zaffar. Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. 3

[46] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *IEEE Conference on Computer Vision and Patterm Recognition (CVPR)*, pages 2387–2395, 2016. 2

[47] T. Wang and K. P. Chow. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14548–14556, 6 2023. 2

[48] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *European Conference on Computer Vision (ECCV)*, September 2018. 2

[49] T. Zhou, W. Wang, Z. Liang, and J. Shen. Face Forensics in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5778–5788, 6 2021. 5