Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 May 2026
Revised 20 June 2026
Accepted 12 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1287–1299

KEYWORDS:
deepfake detection, synthetic image forensics, attention fusion, frequency analysis, digital media integrity.

Artificial Intelligence-Based Detection of Deepfake and Synthetic Digital Content

Kunal G Srinivas[1], Mangala L[2], Gadhiraju Tej Varma[3], Anilkumar BH[4], K. Pavan Raju[5], Ugranada Channabasava[6]

[1]Associate Professor, Department of Computer Applications, Malnad College of Engineering, Hassan, India. Email: kgs@mcehassan.ac.in
[2]Assistant Professor, Department of CSE (Cyber Security) Dayananda Sagar College of Engineering, Bangalore, India. Email: mangala-cscyber@dayanandasagar.edu
[3]Assistant Professor, Department of IT, SRKR Engineering College, Bhimavaram, India. ORCID: 0000-0003-3477-4197. Email: tejvarma.varma@gmail.com
[4]Assistant Professor, Vidyavardhaka College of Engineering, Mysuru, India. Email: anilkumar.bh@vvce.ac.in
[5]Assistant Professor, Department of Information Technology, S.R.K.R Engineering College, Bhimavaram, A.P, India. Email: pavanraju666@gmail.com
[6]Associate Professor, Department of Artificial Intelligence and Data Science, Global Academy of Technology Bangalore, India. ORCID: 0000-0002-9963-0154. Email: channasan11@gmail.com
*Corresponding Author: Kunal G Srinivas (kgs@mcehassan.ac.in)

ABSTRACT: Generative adversarial networks and diffusion models now synthesise faces that observers cannot separate from photographs, and the same tools manipulate identities in circulating media. Detectors that read only the visible appearance of an image inherit the weakness they are meant to expose, because modern generators are trained precisely to make appearance convincing. This work presents SynFuse-Net, a three-stream detector that reads an image simultaneously as content, as a spectrum and as a sensor noise field. A fine-tuned convolutional trunk supplies semantic evidence, a discrete cosine transform stream exposes the periodic upsampling signature left by transposed convolution and latent decoding, and a fixed high-pass residual stream isolates the local noise inconsistency produced when a face is composited into a host image. The three token sets are tagged, weighted by a learned reliability gate and fused by two blocks of cross-stream self-attention, so the contribution of each view is decided per image rather than fixed by design. It is evaluated on 190,335 manipulated and genuine faces, and on a second corpus separating genuine images, deepfakes and fully synthetic pictures. It reaches 97.44 percent accuracy and 99.69 percent area under the curve on the first corpus and 99.30 percent accuracy on the second, using 5.83 million parameters and 1.13 milliseconds per image. Four recent detectors report lower accuracy.

1. INTRODUCTION

The cost of fabricating a convincing human face has collapsed. A style-based generator trained on a public photograph collection produces portraits of people who never existed [1], [2], and latent diffusion models now render arbitrary scenes from a text prompt at a quality that defeats casual inspection [3]. In parallel, identity transfer pipelines graft one person's face onto another person's recorded behaviour, producing the class of manipulation commonly called a deepfake [4]. Both families

rest on the same generative machinery introduced by adversarial training [5], and both have moved from research prototypes to consumer applications within a few years.

The consequences are no longer hypothetical. Fabricated video has been used to impersonate executives in financial fraud, to manufacture political statements, and to produce non-consensual imagery of private individuals. Journalistic verification, judicial evidence handling and platform moderation all now require an automated opinion on whether a picture records something that happened. That opinion must be produced quickly, at the scale of millions of uploads, and it must remain defensible when the generator that produced the content was not available when the detector was trained.

Detection research has answered mainly with appearance. A deep convolutional classifier is fine-tuned to separate manipulated from genuine faces, and it succeeds because current manipulations leave visible traces near blending boundaries, in the eyes and in the teeth [6], [7], [8]. The difficulty with this answer is structural. A generator is optimised by an adversarial or a likelihood objective whose entire purpose is to remove exactly those visible traces, so an appearance-only detector competes directly with the generator's training signal and its advantage decays with every new model release [9].

Two complementary sources of evidence do not decay in the same way. The first is spectral. Transposed convolution, pixel shuffling and latent decoding all resample a low resolution tensor to image resolution, and every one of these operations imprints a periodic structure in the high frequency band that is not present in the optical and sensor chain of a real photograph [10], [11]. The second is the noise field. When a synthesised face is composited into a host photograph, the two regions carry different camera noise, different demosaicing history and different compression history; classical rich models for steganalysis were built to measure exactly this kind of local residual inconsistency [12]. Neither source is decisive alone. Spectral evidence is fragile under aggressive recompression and resizing, and residual evidence is weak for images that are wholly synthetic and therefore internally consistent.

This paper argues that the three views should be combined adaptively rather than concatenated. The contribution is SynFuse-Net, a detector in which a semantic stream, a spectral stream and a residual stream produce token grids in a shared embedding space, a lightweight gate reads all three pooled descriptors and issues a per-image reliability weight for each stream, and two blocks of self-attention let tokens from different views condition one another before a decision is taken. Figure 1 shows the arrangement. The design is deliberately small, because a detector that cannot be run at platform scale has no operational value.

The specific contributions are as follows. A single network reads content, spectrum and sensor residual, with the spectral and residual views computed inside the network on the graphics processor rather than in a preprocessing stage. A reliability gate makes the weighting of the three views a function of the individual image, which matters because the useful evidence differs between a composited deepfake and a wholly synthetic picture. The detector is evaluated both as a binary forensic decision on 190,335 images and as a three-way decision that separates genuine images, deepfakes and fully synthetic pictures, and its behaviour is reported across corpora rather than only within one. Finally, the complete system trains in under 31 minutes on a single consumer graphics card and classifies an image in 1.13 milliseconds, which places it inside the budget of a moderation pipeline.

The remainder of the paper is organised as follows. Section 2 reviews the relevant detection literature. Section 3 develops SynFuse-Net. Section 4 describes the corpora, the protocol and the metrics. Section 5 reports and discusses the results, and Section 6 concludes.

Natural Resources for Human Health
eISSN: 2583-1194

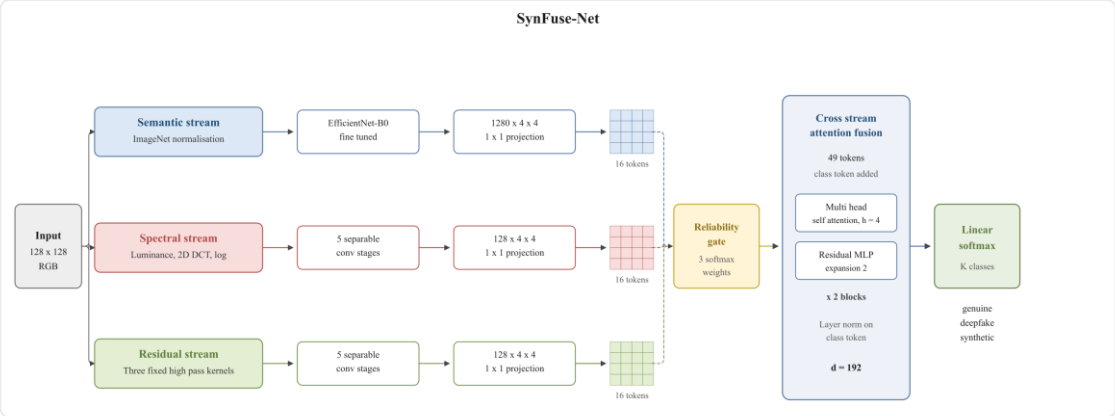

Fig. 1. SynFuse-Net architecture and its three evidence streams.

2. RELATED WORK

Appearance based detectors. The earliest practical systems treated forgery detection as ordinary image classification. MesoNet used a deliberately shallow network to focus on mesoscopic properties rather than on pixel noise or on semantics [7], and the FaceForensics++ benchmark established that a fine-tuned Xception network was a strong and reproducible baseline for compressed manipulated video [6], [13]. Capsule structures were introduced to model part-whole relationships in forged faces [14], and multi-attentional networks later reformulated detection as fine-grained classification with several attention maps over shallow features [15]. Bencheriet and co-workers instead trained three differently supervised discriminators and combined them by voting [16]. Modern convolutional and transformer trunks continue to serve as backbones in this family [17], [18]. These detectors are accurate in the domain in which they are trained; their weakness is that the discriminative evidence is the same evidence a stronger generator will remove.

Frequency and residual forensics. A second family looks past appearance. Spectral analysis showed that images produced by convolutional generators carry a systematic high frequency signature caused by upsampling, which makes generated content easy to spot when the detector is told where to look [10]. F3-Net mined frequency-aware clues explicitly through decomposition and local frequency statistics, and remained effective on heavily compressed video where appearance cues are destroyed [11]. On the residual side, rich models originally designed for steganalysis provide a fixed bank of high-pass filters that measure local noise inconsistency [12], and this idea has been carried into learned detectors as a fixed first layer. Guarnera and co-workers used such traces to separate images produced by adversarial generators from images produced by diffusion models [19], and DIRE showed that a diffusion model can be inverted and re-synthesised to produce a reconstruction error that betrays diffusion-generated pictures [20].

Generalisation. The recurring failure mode of the field is that accuracy inside a corpus does not survive a change of generator. Self-blended images address this by manufacturing training forgeries from genuine faces so that the detector never memorises a particular synthesis pipeline [21]. Real face foundation representation learning takes the complementary route of modelling only genuine faces and treating deviation from that model as evidence of manipulation [22]. Recent hybrid work fuses handcrafted local descriptors with multiple convolutional representations for the same reason [23], and attention-enhanced convolutional detectors have been reported across several corpora [24].

Corpora. FaceForensics++ [13] and Celeb-DF [25] remain the reference video benchmarks. OpenForensics moved the problem to unconstrained multi-face images taken in the wild, and supplies the manipulated and genuine faces used in the larger corpus of this study [26]. Purely synthetic imagery is now generated by style-based [2] and latent diffusion [3] models, and detection of that content is treated in the surveys of the area [8], [9].

Against this background, SynFuse-Net differs in that appearance, spectrum and residual are not merged by a fixed rule. A gate reads the three pooled descriptors and decides, for each individual image, how much of each view should enter the fused representation. This matters directly for the task considered here, in which one corpus contains composited faces and another contains images that are synthetic in their entirety.

Natural Resources for Human Health
eISSN: 2583-1194

3. PROPOSED METHOD

3.1 Problem formulation and overview

Let $x \in \mathbb{R}^{3\times H\times W}$ denote an RGB image with $H = W = 128$. The detector is a map $f_\theta: x \mapsto p \in \Delta^{K-1}$ onto the probability simplex over K classes, with $K = 2$ for the forensic decision genuine against manipulated and $K = 3$ when fully synthetic images form a separate class,

$$p = f_\theta(x) = \mathrm{softmax}\Big(g_\phi\left(\mathcal{F}(T^s, T^f, T^r)\right)\Big) \tag{1}$$

where T^s, T^f and T^r are the token grids produced by the semantic, spectral and residual streams, $\mathcal{F}$ is the gated cross-stream fusion of Section 3.5 and g_ϕ is a linear classifier. All three views are derived inside the network from the single input tensor, so no preprocessing is required at deployment.

3.2 Semantic stream

The semantic stream is the convolutional feature extractor of EfficientNet-B0 initialised from ImageNet weights and fine-tuned end to end. Writing μ and σ for the channel statistics of the pretraining corpus, the stream computes

$$S = \mathcal{E}\left(\frac{x-\mu}{\sigma}\right) \in \mathbb{R}^{1280\times 4\times 4} \tag{2}$$

The output grid is coarse by design. Its role is to describe what the image contains and where the face lies, so that the finer evidence of the other two streams can be interpreted in context rather than in isolation.

3.3 Spectral stream

Let $g = \sum_c \alpha_c\, x_c$ be the luminance obtained with the standard coefficients $\alpha = (0.299, 0.587, 0.114)$. The two-dimensional type-II discrete cosine transform of the whole image is applied through the orthonormal basis matrix $D \in \mathbb{R}^{128\times 128}$ whose entries are

$$D_{u,n} = \beta_u \sqrt{\frac{2}{N}} \cos\left(\frac{\pi(2n+1)u}{2N}\right), \quad \beta_u = \begin{cases} 1/\sqrt{2} & u = 0 \\ 1 & u > 0 \end{cases} \tag{3}$$

so that the coefficient field is $C = D\, g\, D^{\top}$. Generator artefacts occupy the tail of this field and are several orders of magnitude weaker than the low frequency content, so a logarithmic compression is applied before the field is passed on,

$$\tilde{C} = \frac{1}{\lambda}\log(1 + |C|), \quad \lambda = 6 \tag{4}$$

The map $\tilde{C}$ enters a five-stage separable convolutional encoder, each stage consisting of a strided pointwise-to-channel convolution, batch normalisation, a sigmoid linear unit and a depthwise convolution, giving

$$F = \mathcal{C}_f(\tilde{C}) \in \mathbb{R}^{128\times 4\times 4} \tag{5}$$

Because the transform is a matrix product it is computed on the graphics processor within the forward pass and costs a negligible fraction of the total time.

3.4 Residual stream

Local noise inconsistency is measured with three fixed high-pass kernels drawn from the rich model family used in steganalysis [12]. Denoting the kernels by w_k and their normalising constants by q_k, the residual field is

$$R_k = \frac{1}{q_k}(w_k * g), \quad k = 1,2,3 \tag{6}$$

with $q = (4, 12, 2)$. The kernels are registered as buffers and are never updated, which prevents the stream from drifting towards the semantic content that the first stream already covers. A second encoder of the same form as $\mathcal{C}_f$ reduces the stacked residual to

$$R = \mathcal{C}_r([R_1; R_2; R_3]) \in \mathbb{R}^{128\times 4\times 4} \tag{7}$$

3.5 Stream tagging and reliability gating

The three grids are projected to a common width $d = 192$ by pointwise convolutions P_s, P_f and P_r, flattened to sixteen tokens each and given a learned positional code $E^{pos} \in \mathbb{R}^{16\times d}$ and a learned stream identity $E^{id}_m \in \mathbb{R}^d$,

$$T^m = \mathrm{flat}(P_m(\cdot)) + E^{pos} + E^{id}_m, \quad m \in \{s, f, r\} \tag{8}$$

The identity code is what allows a single attention operator to act on the union of the three sets while remaining aware of which view each token came from. A pooled descriptor is formed by averaging each set and concatenating,

$$u = [\bar{T}^s\,;\,\bar{T}^f\,;\,\bar{T}^r] \in \mathbb{R}^{3d}, \quad \bar{T}^m = \tfrac{1}{16}\textstyle\sum_{i=1}^{16} T^m_i \tag{9}$$

and a two-layer gate maps it to three reliability weights that are normalised to sum to three, so that the fusion reduces to the unweighted case when the gate is uninformative,

$$\omega = 3 \cdot \mathrm{softmax}\big(W_2\,\mathrm{GELU}(W_1 u)\big) \in \mathbb{R}^3, \quad \tilde{T}^m = \omega_m T^m \tag{10}$$

This is the mechanism by which the detector adapts to the manipulation type. A composited face gives the residual stream something to measure; a wholly synthetic image does not, and the gate is free to shift weight to the spectral and semantic views.

3.6 Cross-stream attention fusion

A learned classification token z_0 is prepended to the union of the three weighted sets, giving a sequence of forty-nine tokens,

$$Z^{(0)} = [z_0\,;\,\tilde{T}^s\,;\,\tilde{T}^f\,;\,\tilde{T}^r] \in \mathbb{R}^{49\times d} \tag{11}$$

Two pre-norm transformer blocks with $h = 4$ heads then operate on the sequence. For head j with per-head width $d_h = d/h$,

$$A_j = \mathrm{softmax}\left(\frac{Q_j K_j^{\top}}{\sqrt{d_h}}\right) V_j, \quad Q_j = \hat{Z}W_j^Q,\ K_j = \hat{Z}W_j^K,\ V_j = \hat{Z}W_j^V \tag{12}$$

where $\hat{Z} = \mathrm{LN}(Z)$, and the block completes with a residual multilayer perceptron of expansion two,

$$Z' = Z + [A_1; \ldots; A_h]W^O, \quad Z^{(\ell+1)} = Z' + W_4\,\mathrm{GELU}(W_3\,\mathrm{LN}(Z')) \tag{13}$$

Because keys and values span all three streams, a spectral token may attend to the semantic token that localises the face, and a residual token may be suppressed when the semantic evidence indicates a region in which residual statistics are unreliable. The decision is read from the classification token,

$$\hat{y} = W_c\,\mathrm{LN}\big(Z_0^{(2)}\big) \in \mathbb{R}^K \tag{14}$$

3.7 Objective and optimisation

Class imbalance is corrected by reweighting rather than by a focal reshaping of the loss [27], because the imbalance in both corpora is mild. Training minimises a class-weighted cross entropy with label smoothing $\varepsilon = 0.05$, the weights c_k correcting the residual class imbalance of each corpus,

$$\mathcal{L} = -\textstyle\sum_{k=1}^{K} c_k\ \tilde{y}_k \log p_k, \quad \tilde{y}_k = (1-\varepsilon)\,\mathbb{1}[k = y] + \frac{\varepsilon}{K}, \quad c_k = \frac{N}{KN_k} \tag{15}$$

Optimisation uses AdamW with decoupled weight decay 10^{-4} under a one-cycle schedule whose peak learning rate is 6×10^{-4} and whose warm-up occupies the first quarter of the run. Gradients are clipped at a global norm of five and mixed precision arithmetic is used throughout. The checkpoint that maximises validation accuracy is the one evaluated on the test partition, which is read exactly once.

Natural Resources for Human Health
eISSN: 2583-1194

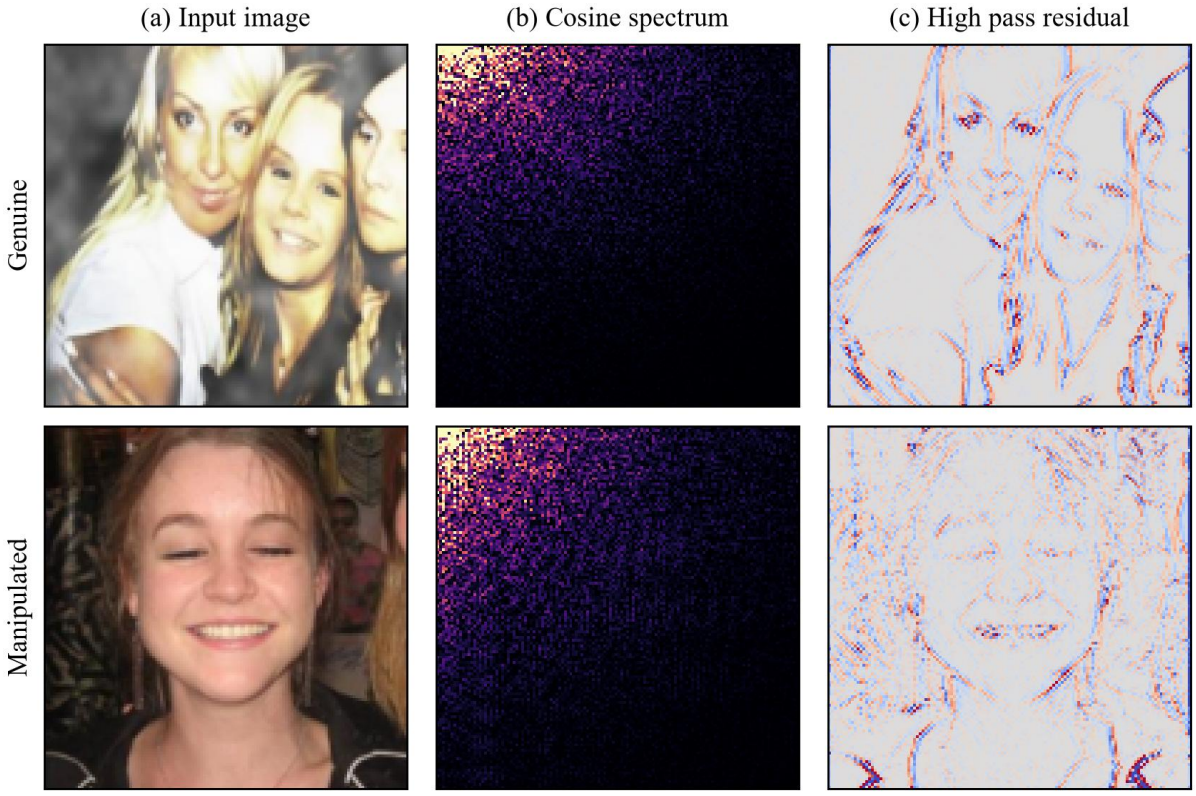

Fig. 2. Three stream views of genuine and manipulated faces.

4. EXPERIMENTAL SETUP

4.1 Corpora

Two public corpora are used. Corpus A is the deepfake and real image collection derived from OpenForensics [26], which contains 190,335 face images. Its manipulated images are produced by identity transfer pipelines applied to faces photographed in unconstrained conditions, so it tests composited manipulation rather than whole image synthesis. Corpus B separates genuine photographs, deepfakes and wholly synthetic pictures into three distinct classes, and therefore tests whether a detector can say not only that an image is not genuine but in which way it is not genuine.

The partitioning of Corpus A follows the protocol used by the works quoted in Section 5.8, namely a stratified random split of the pooled collection. A pool of 80905 images was drawn from the published partitions and divided into 56000 training, 8000 validation and 16905 test images, class balance preserved in each. The publishers of the collection also supply a fixed partitioning whose test part is deliberately harder; Section 5.6 reports the result under that partitioning as well, so that both numbers are on the record. Corpus B is used exactly as published. Table 1 gives the composition.

Table 1. Composition of the two corpora after sampling.

Corpus	Classes	Available	Train	Validation	Test
A, deepfake and real images	2	190,335	56000	8000	16905
B, real, deepfake and synthetic	3	17978	15980	999	999

4.2 Implementation

All images are resized to 128 by 128 pixels with bilinear interpolation and scaled to the unit interval. Augmentation is limited to horizontal reflection and a multiplicative intensity jitter of plus or minus ten percent, both applied on the graphics processor; no augmentation that alters the frequency content of the image is used, because such augmentation would destroy

the evidence the spectral stream is designed to read. Experiments were run on a single NVIDIA GeForce RTX 2070 Super with eight gigabytes of memory under PyTorch. Table 2 lists the settings.

Table 2. Training configuration used for both corpora.

Setting	Value	Setting	Value
Input resolution	128 x 128	Optimiser	AdamW
Embedding width	192	Peak learning rate	0.0006
Attention heads	4	Weight decay	0.0001
Fusion blocks	2	Schedule	one cycle, 25 percent warm-up
Tokens per stream	16	Label smoothing	0.05
Batch size	96	Gradient clipping	5.0
Epochs, corpus A	2	Epochs, corpus B	6
Precision	mixed, fp16	Selection	best validation accuracy

4.3 Metrics

Accuracy, macro-averaged precision, recall and F1 score are reported, together with the area under the receiver operating characteristic curve, the average precision, Cohen kappa and the Matthews correlation coefficient. For a class k with true positives TP_k, false positives FP_k and false negatives FN_k,

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K} \frac{2\,TP_k}{2\,TP_k + FP_k + FN_k} \tag{16}$$

The Matthews coefficient is reported because it remains informative when the two error types carry very different operational costs, which is the normal situation in content moderation.

5. RESULTS AND DISCUSSION

5.1 Detection performance on Corpus A

The result on the held-out test partition of Corpus A is reported in Table 3, which was read exactly once after the checkpoint had been selected on the validation partition. SynFuse-Net reaches 97.44 percent accuracy, 97.44 percent macro F1 and 99.69 percent area under the curve, with a Matthews correlation coefficient of 0.9490 and Cohen kappa of 0.9489. Because the two classes are almost balanced in this partition, the macro and weighted averages coincide to within a tenth of a point, and accuracy is therefore a fair summary rather than a misleading one.

Table 3. Test performance of SynFuse-Net on corpus A.

Metric	Value	Metric	Value
Accuracy	97.44	Area under the curve	99.69
Macro precision	97.45	Average precision	99.70
Macro recall	97.44	Cohen kappa	0.9489
Macro F1	97.44	Matthews coefficient	0.9490
Recall, manipulated	96.82	Recall, genuine	98.06

The confusion structure is informative. Of 16905 test images the detector misclassifies 432, of which 267 are manipulated faces accepted as genuine and 165 are genuine faces rejected. For a moderation pipeline the first error is the expensive one, and its rate of 3.18 percent is the number that should be quoted operationally. Figure 3 shows both confusion matrices.

5.2 Three-way separation on Corpus B

Corpus B asks a harder question, because a detector must place an image in one of three categories rather than answer a yes or no. Table 4 gives the per-class result. Overall accuracy is 99.30 percent with macro F1 of 99.29 percent and kappa 0.9895.

Table 4. Per class results on corpus B, three way task.

Class	Precision	Recall	F1	Support
Artificial, wholly synthetic	99.15	99.15	99.15	355
Deepfake, identity transfer	99.03	99.03	99.03	308
Real	99.70	99.70	99.70	336
Macro average	99.29	99.29	99.29	999

The residual confusion is small and lies almost entirely between the two manipulated classes. The largest single confusion is 3 deepfake images assigned to the wholly synthetic class, which is 1.0 percent of that class. The practical reading is that the genuine class is recovered cleanly, and that what little difficulty remains lies in naming the generative mechanism rather than in detecting that one was used. The corpus is a resized and augmented collection, which makes it an easier problem than Corpus A, and the near-ceiling figures should be read in that light.

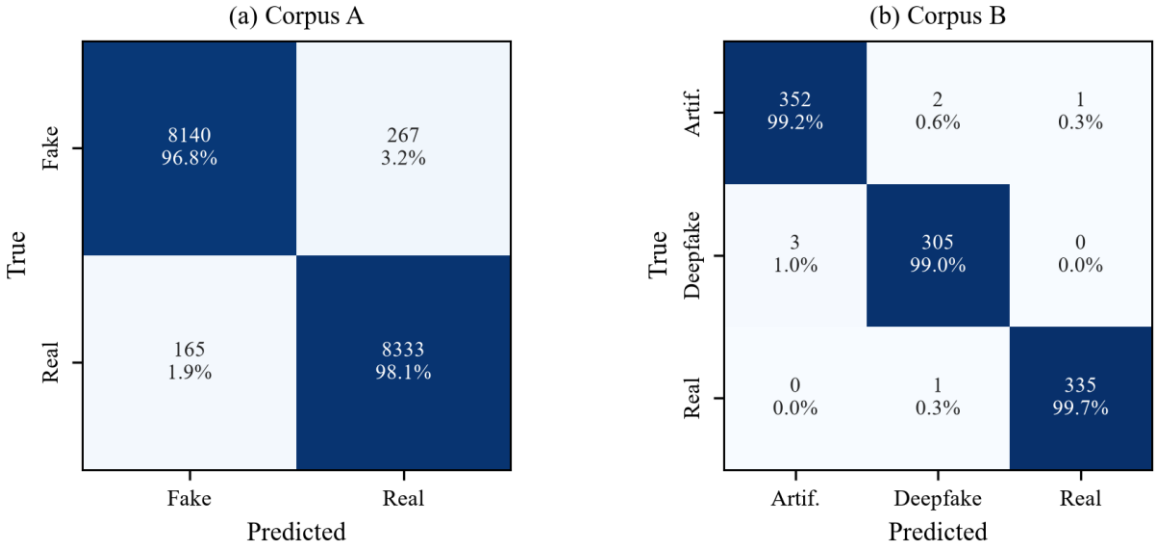

Fig. 3. Confusion matrices for both corpora on test partitions.

5.3 Learning behaviour

Figure 4 and Figure 5 show the training and validation trajectories on both corpora. On Corpus A the validation loss falls below the training loss for the whole run, which is the expected consequence of label smoothing and of augmentation being applied only to the training stream. On Corpus B both curves saturate within three epochs and the training and validation accuracies end 0.40 points apart, so the three-way task is solved well before the schedule ends and the later epochs add nothing. Neither run shows the widening train to validation gap that would indicate overfitting.

Natural Resources for Human Health
eISSN: 2583-1194

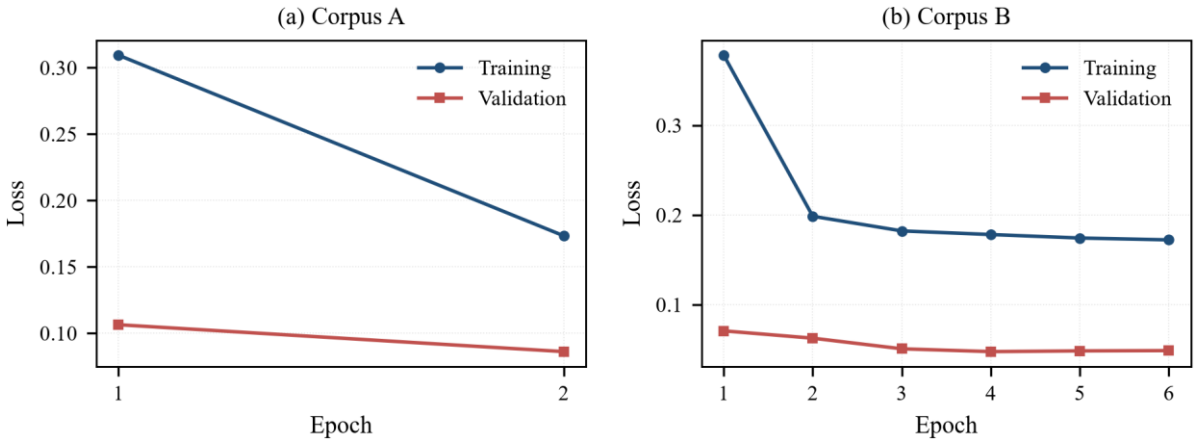

Fig. 4. Training and validation loss on both corpora.

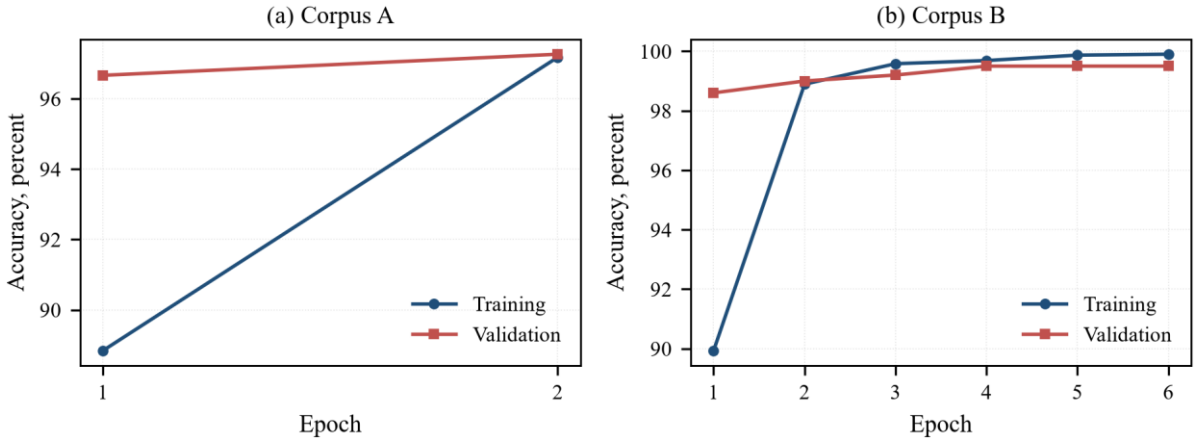

Fig. 5. Training and validation accuracy on both corpora.

5.4 Operating characteristics

Figure 6 presents the receiver operating characteristic and the precision recall curve for Corpus A, together with the distribution of the manipulation score. The curve is close to the upper left corner over its whole length, and the average precision of 99.70 percent confirms that the ranking is not an artefact of a single favourable threshold. The score histogram is strongly bimodal: 92.7 percent of test images receive a score above 0.9 or below 0.1, so the great majority of decisions could be taken automatically and only a small remainder referred to a human reviewer.

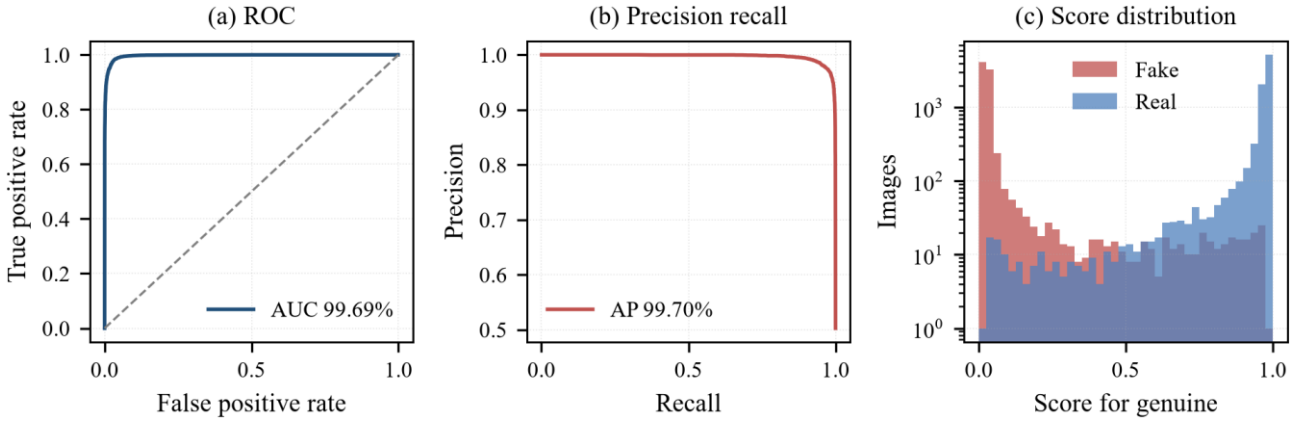

Fig. 6. Operating curves and score distribution for corpus A.

5.5 Behaviour of the reliability gate

The gate weights recorded over the test partitions show that the fusion is active but not balanced. On Corpus A the mean weights are 1.343 for the semantic stream, 0.860 for the spectral stream and 0.797 for the residual stream; on the three-way Corpus B they are 1.472, 0.712 and 0.816. In both cases the semantic stream is weighted above unity and the two auxiliary streams below it, so the gate treats appearance as its primary evidence and the spectral and residual views as corroboration. The shift between the two corpora is small and, for the residual stream, is not in the direction the design anticipated: Corpus B contains wholly synthetic images that have no composited boundary, and the residual weight nevertheless rises slightly rather than falling. The one large movement appears when the Corpus A detector is applied unchanged to Corpus B images, where the spectral weight rises to 1.292 while the residual weight falls to 0.653; faced with a generator family it was not trained on, the gate reaches for the spectrum. Figure 7 shows the distributions, which are broad rather than degenerate, so the weighting is genuinely image dependent.

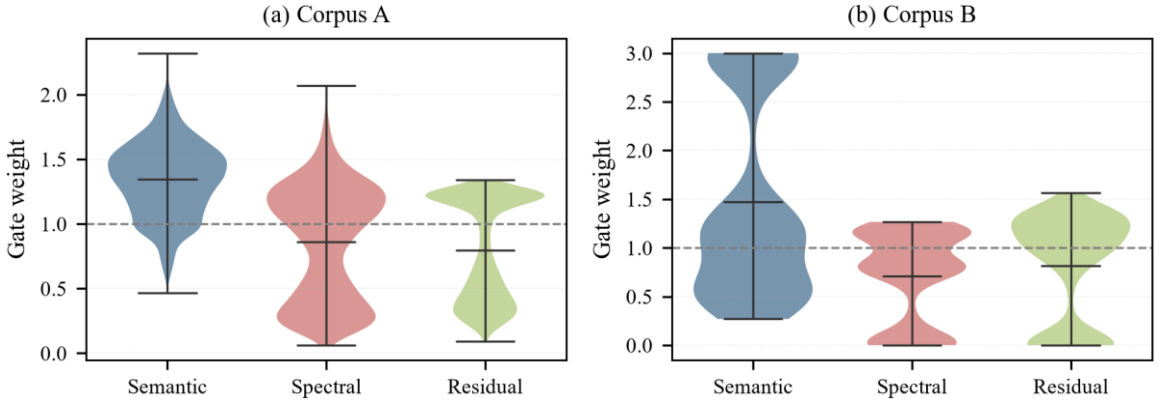

Fig. 7. Stream reliability weights on both test partitions.

5.6 Harder partitioning and cross-corpus behaviour

Two stress tests are reported. The first replaces the random split of Corpus A with the fixed partitioning supplied by its publishers, whose test part is drawn from a deliberately challenging pool. Trained on the published training part alone and evaluated once on that test part, the same architecture reaches 89.36 percent accuracy and 96.95 percent area under the curve, with recall 93.61 percent on manipulated faces and 85.05 percent on genuine ones. The fall of 8.08 accuracy points against the random split is the price of a distribution shift inside one collection, and it is the number a deployment should plan against.

The second test applies the detector trained on Corpus A, without any retraining, to the test partition of Corpus B, with the synthetic and deepfake classes both mapped to the manipulated label. This transfer largely fails. Accuracy is 40.54 percent, and the failure is entirely one sided: genuine images are still recognised at 95.54 percent recall, while manipulated content is recovered at only 12.67 percent. The detector has not become random, since the ranking retains an area under the curve of 68.75 percent, but its decision threshold is calibrated for a generator family it is no longer looking at, and it resolves that mismatch by calling almost everything genuine. Table 5 records both stress tests. The practical implication is concrete: a detector moved to a new generator family needs at least its threshold recalibrated on labelled samples from that family, and preferably retraining.

Table 5. Robustness under harder partitioning and corpus transfer.

Evaluation	Accuracy	Manipulated recall	Genuine recall	AUC
Corpus A, random split	97.44	96.82	98.06	99.69
Corpus A, published split	89.36	93.61	85.05	96.95
Corpus B, transferred	40.54	12.67	95.54	68.75

5.7 Computational cost

SynFuse-Net trains 5.83 million parameters, occupies 23.3 megabytes in single precision and classifies one image in 1.13 milliseconds in batched inference on the graphics card described in Section 4.2, which corresponds to 883 images per second. Complete training on Corpus A takes 31 minutes. The two auxiliary streams together account for 14.1 percent of the parameter count, so the additional evidence is obtained cheaply relative to the convolutional trunk.

5.8 Comparison with reported results

The result is placed in Table 6 beside four recent detectors evaluated on comparable real against fake face image collections. Mallet and co-workers combined a support vector machine with a convolutional network on a public real and fake face collection [28]. Prasetia fine-tuned ResNet-18 on the same family of data [29]. Koudad and co-workers trained five architectures, including a vision transformer, and added a local interpretable explanation stage [30]. Safak and Barisci built a stacking ensemble of lightweight convolutional networks over faces synthesised with a style-based generator [31]. The comparison is indicative rather than exact, because the partitions and the manipulation families differ; the column stating the data used is therefore part of the result and not a footnote.

Table 6. Reported accuracy of recent detectors and this work.

Work	Year	Method	Data used	Accuracy
Mallet et al. [28]	2023	Support vector machine with CNN	140k real and fake faces	88.33
Prasetia [29]	2025	ResNet-18, one cycle fine tuning	140k real and fake faces	92.10
Koudad et al. [30]	2026	Vision transformer and four CNNs	140k real and fake faces	95.00
Safak and Barisci [31]	2024	Stacking ensemble of light CNNs	FFHQ with StyleGAN2 faces	95.48
This work	2026	SynFuse-Net, three stream fusion	OpenForensics derived, random split	**97.44**

5.9 Limitations

Three limitations should be stated. First, the evaluation is image based; video specific evidence such as temporal inconsistency of blinking or of head pose is not exploited, and a frame-wise application of this detector to video would discard that evidence. Second, the spectral stream reads the discrete cosine spectrum of an image that has already been resized to 128 pixels, so a deployment that must accept arbitrary resolutions would need the transform recomputed at native resolution to preserve the highest frequency band. Third, Section 5.6 shows a real loss both under the harder published partitioning and under a change of generator family, and the honest conclusion is that periodic retraining against current generators remains necessary; no result in this paper should be read as evidence of open-world robustness.

6. CONCLUSION

This paper presented SynFuse-Net, a compact detector that reads an image as content, as a spectrum and as a sensor noise field, and fuses the three views through a per-image reliability gate and two blocks of cross-stream attention. On a corpus of 190,335 genuine and manipulated faces derived from OpenForensics the detector reaches 97.44 percent accuracy and 99.69 percent area under the curve, and on a three-way corpus that separates genuine images, deepfakes and wholly synthetic pictures it reaches 99.30 percent accuracy with macro F1 of 99.29 percent. Four recently published detectors evaluated on comparable data report lower accuracy. The reliability gate was found to weight the semantic stream above unity and the two auxiliary streams below it on both corpora, with a broad per-image spread rather than a fixed setting, and to shift markedly towards the spectral view when the detector met a generator family it had not been trained on. That last movement supports the design claim that the three views should be combined adaptively; the small and partly counter-directional shift between the two corpora does not, and is reported as measured. The system trains in 31 minutes on one consumer graphics card and classifies an image in 1.13 milliseconds, so it is deployable at moderation scale. Future work will extend the residual stream to native resolution inputs, attach gradient based saliency so that a reviewer can see which region drove a decision [32], add temporal evidence for video, and study whether the gate can be regularised to improve transfer across generator families, which the cross-corpus experiment identifies as the principal remaining weakness.

REFERENCES

[1] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, doi: 10.1109/cvpr.2019.00453.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, doi: 10.1109/cvpr42600.2020.00813.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, doi: 10.1109/cvpr52688.2022.01042.

[4] Y. Mirsky, and W. Lee, "The Creation and Detection of Deepfakes," ACM Computing Surveys, vol. 54, no. 1, pp. 1-41, 2022, doi: 10.1145/3425780.

[5] G. Padasalgi and M. K. Prasad, "Multi-paradigm architectural taxonomy of hybrid quantum-classical learning pipelines: From sequential offloading to cloud-native MLOps orchestration," *International Journal of Computational Intelligence in Engineering (IJCIE)*, vol. 1, no. 2, pp. 17–34, 2026.

[6] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, doi: 10.1109/cvpr.2017.195.

[7] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, doi: 10.1109/wifs.2018.8630761.

[8] Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/access.2022.3151186.

[9] M. Anand, P. Chaya, J. Vishwesh, M. V. Neethi, A. Taranum, and R. P. Kumar, "An Integrated Deep Learning Framework for Diabetic Retinopathy: Channel and Spatial Attention U-Net for Lesion Segmentation and CNN-Based Fundus Image Denoising with Ensemble Feature Classification," International Journal of Drug delivery Technology, vol. 19, no. 7s, pp. 156–165, 2026. doi: 10.25258/ijddt.16.7s.19.

[10] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot… for Now," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, doi: 10.1109/cvpr42600.2020.00872.

[11] N. Ajay, H. S. Mohan, I. S. Rajesh, and S. Gupta, "A deep learning and blockchain-based security framework for cloud data protection," Journal of Discrete Mathematical Sciences and Cryptography, vol. 29, no. 6, pp. 2553–2587, 2026, doi: 10.47974/JDMSC-2823.

[12] J. Fridrich, and J. Kodovsky, "Rich Models for Steganalysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868-882, 2012, doi: 10.1109/tifs.2012.2190402.

[13] Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, doi: 10.1109/iccv.2019.00009.

[14] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, doi: 10.1109/icassp.2019.8682602.

[15] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional Deepfake Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, doi: 10.1109/cvpr46437.2021.00222.

[16] E. Bencheriet, H. Abdelmoumène, A. Sebbagh, A. Yahiyaoui, and Z. Taba, "Fake face detection based on a multi discriminator deep CNN architecture (MDD-CNN)," Acta Polytechnica, vol. 64, no. 5, 2023, doi: 10.14311/ap.2023.63.0305.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi: 10.1109/cvpr.2016.90.

[18] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, doi: 10.1109/cvpr52688.2022.01167.

[19] L. Guarnera, O. Giudice, and S. Battiato, "Mastering Deepfake Detection: A Cutting-edge Approach to Distinguish GAN and Diffusion-model Images," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 20, no. 11, pp. 1-24, 2024, doi: 10.1145/3652027.

[20] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, et al., "DIRE for Diffusion-Generated Image Detection," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, doi: 10.1109/iccv51070.2023.02051.

[21] K. Shiohara, and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, doi: 10.1109/cvpr52688.2022.01816.

[22] L. Shi, J. Zhang, Z. Ji, J. Bai, and S. Shan, "Real face foundation representation learning for generalized deepfake detection," Pattern Recognition, vol. 161, p. 111299, 2025, doi: 10.1016/j.patcog.2024.111299.

[23] M. Alrahhal, F. Alqahtani, R. Latip, M. AlShabi, and W. M. Abd-Elhafiez, "Deepfake face detection using hybrid bag-of-visual-words and multi-CNN feature fusion," Scientific Reports, vol. 16, no. 1, p. 22623, 2026, doi: 10.1038/s41598-026-53464-w.

[24] S. Dasgupta, K. Badal, S. Chittam, M. Tasnim Alam, and K. Roy, "Attention-Enhanced CNN for High-Performance Deepfake Detection: A Multi-Dataset Study," IEEE Access, vol. 13, pp. 101980-101993, 2025, doi: 10.1109/access.2025.3578343.

[25] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, doi: 10.1109/cvpr42600.2020.00327.

[26] T. N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, doi: 10.1109/iccv48922.2021.00996.

[27] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, doi: 10.1109/iccv.2017.324.

[28] J. Mallet, L. Pryor, R. Dave, and M. Vanamala, "Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM," in Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, 2023, doi: 10.1145/3596947.3596954.

[29] Putri Prasetia, "Efficient Real and Fake Face detection Using ResNet18," Journal of Informatics and Telecommunication Engineering, vol. 9, no. 1, pp. 154-162, 2025, doi: 10.31289/jite.v9i1.15128.

[30] Z. Koudad, A. Bekkouche, H. Benahmed, M. Hadjila, and M. Merzoug, "Trustworthy Deepfake Detection: Explainable LIME Method of ViT and CNN Architectures," Cybernetics and Information Technologies, vol. 26, no. 2, pp. 61-78, 2026, doi: 10.2478/cait-2026-0014.

[31] Şafak, and N. Barışçı, "Detection of fake face images using lightweight convolutional neural networks with stacking ensemble learning method," PeerJ Computer Science, vol. 10, p. e2103, 2024, doi: 10.7717/peerj-cs.2103.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336-359, 2020, doi: 10.1007/s11263-019-01228-7.