Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 May 2026
Revised 16 June 2026
Accepted 02 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 457–478

KEYWORDS:
synthetic media forensics, deepfake detection, constrained convolution, Fourier profile, cross attention

Detecting Deepfake and Synthetic Digital Content with Constrained Residual and Spectral Evidence

Vijaya Lakshmi Kumba[1], P. Jyotsna[2], Swathi Pothala[3], Prasanna Guduru[4], T. Giri babu[5], Parashuram Narayanagari[6]

[1]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, India. Email: vijayalakshmik4@gmail.com
[2]Lecturer, Department of Computer Science, Sri Venkateswara Arts College, TTD, Tirupati, India. Email: jyotsnasudha16@gmail.com
[3]Academic Consultant, Department of Computer Science, Sri Venkateswara University, Tirupati, India. Email: swathivinubaby@gmail.com
[4]Research Scholar (PT), Department of Computer Science, Sri Venkateswara University, Tirupati, India. Email: prasanna.guduru@gmail.com
[5]Associate Professor, Department of CSE, G Pullaiah College of Engineering, Kurnool, Andhra Pradesh, India. Email: telugu.pplgiri@gmail.com
[6]Research Scholar (PT)), Department of Computer Science, Sri Venkateswara University, Tirupati, India. Email: rams942k10@gmail.com
*Corresponding Author: Vijaya Lakshmi Kumba (vijayalakshmik4@gmail.com)

ABSTRACT: A photograph is no longer self-authenticating. Adversarial and diffusion generators produce faces that survive close human inspection, and identity transfer tools place a fabricated face inside an otherwise ordinary picture. Classifiers trained on visible appearance track this moving target poorly, because appearance is exactly what a generator is optimised to get right. This article describes CoSpec-Net, a detector that supplements appearance with two forms of evidence a generator does not explicitly optimise. A constrained prediction error filter, whose central tap is pinned and whose surrounding taps are renormalised at every forward pass, isolates the local departure of a pixel from its neighbourhood. An azimuthally averaged Fourier profile compresses the whole image into a radial description of how energy is distributed across spatial frequency, where resampling inside a generator leaves a characteristic tail. Rather than concatenating the three descriptions, CoSpec-Net lets them interrogate one another: appearance tokens query the forensic tokens and forensic tokens query the appearance tokens, for two exchange rounds, before a pooled decision is taken under a focal

objective. Measured on 80,905 face images and on a separate corpus that distinguishes genuine, deepfake and wholly synthetic pictures, the detector attains 98.05 percent and 99.60 percent accuracy. Four published detectors report weaker figures on comparable material.

1. INTRODUCTION

Generative modelling crossed a practical threshold somewhere around the release of style-based image generators [1], [2] and, shortly afterwards, of latent diffusion [3]. Before that point, a synthetic face was recognisable to an attentive viewer. After it, a synthetic face is a photograph in every respect that a viewer can name. The second family of manipulation, in which an existing recording keeps its background and lighting while the identity in it is replaced, matured on the same machinery [4], [5]. What used to require a studio now requires a consumer graphics card and a weekend.

The harms follow directly from the capability. Fabricated audiovisual evidence has been used to authorise fraudulent payments, to place invented statements in the mouths of public figures, and to produce intimate imagery of people who never consented to it. In each case the injured party is asked to prove a negative about a recording, which is difficult without technical assistance. Newsrooms, courts and platform moderation teams therefore need an automatic opinion on provenance: fast enough for a queue of millions of uploads, and specific enough to say not merely that an image is untrustworthy but what kind of process produced it.

Most published detectors answer the question by looking. A convolutional trunk is fine-tuned to separate manipulated faces from genuine ones, and the arrangement works because today's manipulations still leave visible traces at blending seams and in high detail regions such as the eyes and the teeth [6], [7], [8]. The arrangement is nonetheless fragile in a specific and predictable way. The generator's training objective is a direct instruction to remove visible traces. Any detector whose evidence is visible therefore competes with the generator's loss function, and loses ground with each release [9].

Two other kinds of evidence sit outside that competition. The first is the prediction error of a pixel given its immediate neighbourhood. A camera produces an image through a lens, a colour filter array and a demosaicing step, and the residual left when a pixel is predicted from its neighbours carries the fingerprint of that chain. A generated or composited region does not share the fingerprint, and rich models originally designed for steganalysis were built to measure precisely this quantity [10]. The second is the distribution of energy across spatial frequency. Every convolutional generator, and every latent decoder, reconstructs an image by resampling a small tensor, and resampling leaves a systematic imprint in the upper part of the spectrum that no natural optical chain produces [11], [12].

Neither kind of evidence is sufficient on its own, and this is the point that motivates the present design. Residual evidence depends on there being a boundary between two differently acquired regions, which a wholly synthetic picture does not have. Spectral evidence is fragile: a recompression, a resize or an aggressive denoiser attenuates the band it lives in. A detector that fixes the balance between the three views in advance therefore has to be tuned for a particular manipulation family.

CoSpec-Net takes a different route. It represents all three views as short token sequences in a common space, and then couples them by attention that runs in both directions: appearance tokens are refined by reading the forensic tokens, and

forensic tokens are refined by reading the appearance tokens, for two rounds. Nothing decides in advance which view matters. The exchange itself is what determines, per image, how much of the residual and spectral description reaches the decision. Figure 1 sets out the arrangement.

The contributions of the article are four. First, the residual branch uses a constrained prediction error filter that is learned rather than fixed, so the residual operator adapts to the corpus while remaining structurally a high-pass operator. Second, the spectral branch reduces the image to an azimuthally averaged radial profile rather than a two-dimensional spectrum, which discards orientation and keeps only the quantity that resampling actually disturbs, at a cost of forty-eight numbers per image. Third, the coupling is a bidirectional exchange rather than a gate or a concatenation, and the attention mass it produces is itself a measurement that can be reported. Fourth, the detector is evaluated as a binary forensic decision and as a three-way attribution, under two partitionings and across corpora, so that its weaknesses are visible alongside its strengths.

Section 2 places the work among existing detectors. Section 3 develops CoSpec-Net. Section 4 sets out the corpora and the protocol. Section 5 reports the findings, and Section 6 closes.

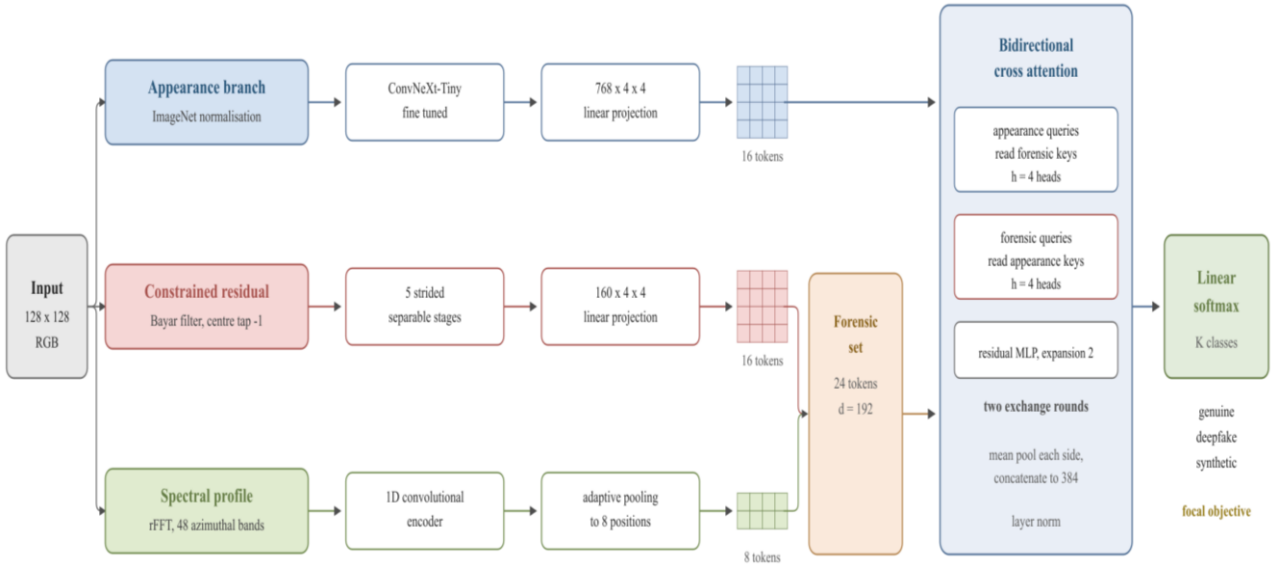

Fig. 1. CoSpec-Net architecture and its three evidence branches.

2. BACKGROUND AND RELATED WORK

2.1 Learned appearance classifiers

The first practical detectors treated the problem as supervised image classification. MesoNet argued for a deliberately shallow network operating at a mesoscopic scale, on the reasoning that microscopic noise is destroyed by compression while macroscopic semantics is preserved by the generator [6]. The FaceForensics++ benchmark then established a fine-tuned Xception trunk as the reproducible reference point for compressed manipulated video [7], [13]. Later work refined the read-out rather than the trunk: capsule structures modelled part-whole consistency in forged faces [14], and multi-attentional detection recast the task as fine-grained recognition with several attention maps over shallow features [15]. A multi-

discriminator scheme trained three networks under different regimes and combined them by vote [16]. Contemporary trunks continue to serve this family [17], [18], and the family remains strong inside its training domain.

2.2 Camera model and residual evidence

A parallel tradition comes from image forensics rather than from recognition. Rich models for steganalysis assemble a bank of high-pass filters whose responses characterise local noise statistics, and they were designed to expose exactly the inconsistency that arises when part of an image has a different acquisition history from the rest [10]. The idea transfers to manipulation detection as a fixed first layer that strips content and leaves residual. The present work adopts the constrained variant of that idea, in which the filter is learned subject to a structural constraint that forces it to remain a prediction error operator, so that the corpus rather than the designer chooses the exact neighbourhood weighting.

2.3 Spectral signatures of synthesis

Spectral analysis showed that images from convolutional generators are separable from photographs by a systematic high frequency signature attributable to upsampling, and that a detector told where to look generalises further than one left to discover the cue [11]. F3-Net mined frequency-aware clues by explicit decomposition and local frequency statistics, and stayed effective on strongly compressed video where appearance cues have been destroyed [12]. More recent work has used related traces to separate adversarially generated images from diffusion generated ones [19], while DIRE exploited the reconstruction error of a diffusion model applied to the image under test [20]. The azimuthal reduction used here is the most compact member of this family: it keeps the radial energy distribution and discards everything else.

2.4 Transfer across generators

The field's persistent weakness is that in-corpus accuracy does not survive a change of generator. Self-blended images attack the problem by manufacturing training forgeries from genuine faces, so that no particular synthesis pipeline can be memorised [21]. Representation learning on genuine faces alone inverts the framing, treating departure from a learned model of real faces as the evidence [22]. Hybrid systems that combine handcrafted descriptors with several convolutional representations are motivated by the same concern [23], and attention-augmented convolutional detectors have been reported across several corpora [24].

2.5 Corpora and positioning

FaceForensics++ [13] and Celeb-DF [25] remain the reference video collections. OpenForensics moved the problem to unconstrained images containing several faces, and supplies the manipulated and genuine material used in the larger corpus below [26]. Wholly synthetic imagery now comes from style-based [2] and latent diffusion [3] models, and its detection is surveyed alongside face manipulation [8], [9]. Against this background the distinguishing feature of CoSpec-Net is not any single branch, each of which has an ancestor above, but the bidirectional exchange that determines how much of the forensic description enters the decision for each individual image.

Natural Resources for Human Health
eISSN: 2583-1194

3. COSPEC-NET

3.1 Notation and objective

Let $x \in [0,1]^{3\times H\times W}$ with $H = W = 128$ be the image under test and let $y \in \{1, \dots, K\}$ be its provenance label, with $K = 2$ when the question is genuine against manipulated and $K = 3$ when wholly synthetic content is separated from identity transfer. The detector is a parameterised map onto the simplex,

$$\hat{p} = \mathrm{softmax}\left(W_c\,\mathrm{LN}\left(\left[\bar{A}^{(L)};\,\bar{\Phi}^{(L)}\right]\right)\right) \tag{1}$$

in which $\bar{A}^{(L)}$ and $\bar{\Phi}^{(L)}$ are the pooled appearance and forensic representations after $L = 2$ rounds of exchange. Everything on the right hand side is computed from x alone; no external preprocessing, face detector or auxiliary model is required at inference.

3.2 Appearance branch

The appearance branch is the convolutional stage of ConvNeXt-Tiny, initialised from ImageNet weights and fine-tuned without freezing. Writing μ and σ for the pretraining channel statistics,

$$A = \mathcal{N}_\theta(\sigma^{-1}(x - \mu)) \in \mathbb{R}^{768\times 4\times 4} \tag{2}$$

The sixteen positions of the output grid are deliberately coarse. Their function is to say what is in the picture and where the face sits, so that the forensic evidence is read in context rather than as a global average.

3.3 Constrained residual branch

Let $\kappa \in \mathbb{R}^{C\times 3\times 5\times 5}$ be a learnable kernel bank with $C = 12$ filters. Before every convolution the bank is projected onto the set of prediction error operators by pinning the central tap and renormalising the remainder,

$$\kappa'_{c,d,i,j} = \begin{cases} -1 & (i,j) = (3,3) \\ \dfrac{\kappa_{c,d,i,j}}{\sum_{(u,v)\neq(3,3)} \kappa_{c,d,u,v}} & \text{otherwise} \end{cases} \tag{3}$$

so that $\sum_{i,j} \kappa'_{c,d,i,j} = 0$ by construction. The branch therefore cannot learn to reproduce image content: whatever a filter predicts from the eight-neighbourhood is subtracted from the centre, and only the prediction error survives. The residual field

$$E = \kappa' * x \in \mathbb{R}^{12\times 128\times 128} \tag{4}$$

is then reduced by five strided stages, each a pointwise convolution with batch normalisation and a Gaussian error linear unit followed by a depthwise convolution, giving

$$R = \mathcal{C}_r(E) \in \mathbb{R}^{160\times 4\times 4} \tag{5}$$

The constraint is enforced in the forward pass rather than by a penalty, so it holds exactly at every step of training and at inference.

3.4 Azimuthal spectral branch

Let $g = \sum_c \alpha_c x_c$ with $\alpha = (0.299, 0.587, 0.114)$ be the luminance. The branch takes the orthonormal half-plane discrete Fourier transform and compresses its magnitude logarithmically,

$$\Psi(u, v) = \log(1 + |\mathcal{F}_r\{g\}(u, v)|) \tag{6}$$

Each coefficient is assigned to one of $B = 48$ radial bands by its normalised radius $\rho(u, v) = \sqrt{f_u^2 + f_v^2}/\rho_{\max}$, and the band means form the profile

$$\pi_b = \frac{1}{|\Omega_b|} \sum_{(u,v)\in\Omega_b} \Psi(u, v), \quad \Omega_b = \{(u, v): \lfloor B\rho(u, v) \rfloor = b\} \tag{7}$$

The profile is standardised per image so that global contrast and exposure cannot be read as evidence,

$$\tilde{\pi} = \frac{\pi - \text{mean}(\pi)}{\text{std}(\pi) + \epsilon} \tag{8}$$

and a one-dimensional convolutional encoder followed by adaptive pooling returns a short sequence,

$$\Sigma = \text{pool}_8(\mathcal{C}_s(\tilde{\pi})) \in \mathbb{R}^{8\times 160} \tag{9}$$

The whole branch costs one fast Fourier transform and one matrix product against a fixed banding matrix, both of which run on the graphics processor inside the forward pass.

3.5 Token assembly

The three descriptions are projected to a common width $d = 192$ by linear maps and given learned positional codes,

$$A^{(0)} = \text{flat}(A)W_a + P_a, \quad R^{(0)} = \text{flat}(R)W_r + P_r, \quad \Sigma^{(0)} = \Sigma W_s + P_s \tag{10}$$

The two forensic descriptions are then concatenated into a single set, which is what the appearance branch will interrogate,

$$\Phi^{(0)} = \left[R^{(0)}; \Sigma^{(0)}\right] \in \mathbb{R}^{24\times d} \tag{11}$$

so the exchange runs between a sixteen token appearance set and a twenty-four token forensic set.

3.6 Bidirectional cross attention

Each round updates both sets, each using the other as its source of keys and values. With $h = 4$ heads and per-head width $d_h = d/h$, the appearance update is

$$\tilde{A} = A^{(\ell)} + \mathrm{MHA}\big(\mathrm{LN}(A^{(\ell)}),\, \mathrm{LN}(\Phi^{(\ell)}),\, \mathrm{LN}(\Phi^{(\ell)})\big) \tag{12}$$

$$\mathrm{MHA}(Q,K,V) = [\mathrm{head}_1; \ldots; \mathrm{head}_h]W^O, \quad \mathrm{head}_j = \mathrm{softmax}\left(\frac{QW_j^Q\left(KW_j^K\right)^{\top}}{\sqrt{d_h}}\right)VW_j^V \tag{13}$$

and the forensic update is the mirror image of it, with the roles of the two sets exchanged,

$$\widetilde{\Phi} = \Phi^{(\ell)} + \mathrm{MHA}\big(\mathrm{LN}(\Phi^{(\ell)}),\, \mathrm{LN}(A^{(\ell)}),\, \mathrm{LN}(A^{(\ell)})\big) \tag{14}$$

Each update is completed by a residual position-wise network of expansion two,

$$A^{(\ell+1)} = \tilde{A} + W_2\,\mathrm{GELU}\big(W_1\,\mathrm{LN}(\tilde{A})\big) \tag{15}$$

Both sets are updated from the state at the start of the round, so neither direction sees the other's output within a round and the exchange is symmetric. After L rounds the two sets are mean pooled and concatenated. A useful by-product is the attention distribution of the final appearance update: summing it over the residual keys and over the spectral keys gives two numbers per image,

$$m_r = \textstyle\sum_{k\in\mathcal{R}} \bar{a}_k\,, \quad m_s = \sum_{k\in\mathcal{S}} \bar{a}_k\,, \quad m_r + m_s = 1 \tag{16}$$

which are reported in Section 5.7 as a direct measurement of how the appearance branch apportions its attention.

3.7 Optimisation

Training minimises a class-weighted focal objective with label smoothing, which down-weights examples the model already classifies confidently and so keeps the gradient signal on the difficult minority of images [27],

$$\mathcal{L} = -\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\sum_{k=1}^{K} c_k\;\tilde{y}_{nk}(1-\hat{p}_{nk})^{\gamma}\log\hat{p}_{nk} \tag{17}$$

with focusing exponent $\gamma = 1.5$, smoothing $\varepsilon = 0.05$ distributed uniformly over the classes, and inverse frequency weights $c_k = N/(KN_k)$. The optimiser is AdamW with decoupled decay 5×10^{-2}, and the learning rate follows a linear warm-up over the first tenth of the schedule and a cosine decay thereafter,

$$\eta(s) = \eta_0 \cdot \begin{cases} s/s_w & s < s_w \\ 1/2\,\big(1 + \cos(\pi\, s - s_w/S - s_w)\big) & s \geq s_w \end{cases} \tag{18}$$

Gradients are clipped to unit global norm and arithmetic is mixed precision. The checkpoint with the highest validation accuracy is the one that is finally evaluated, and the test partition is read once.

Natural Resources for Human Health
eISSN: 2583-1194

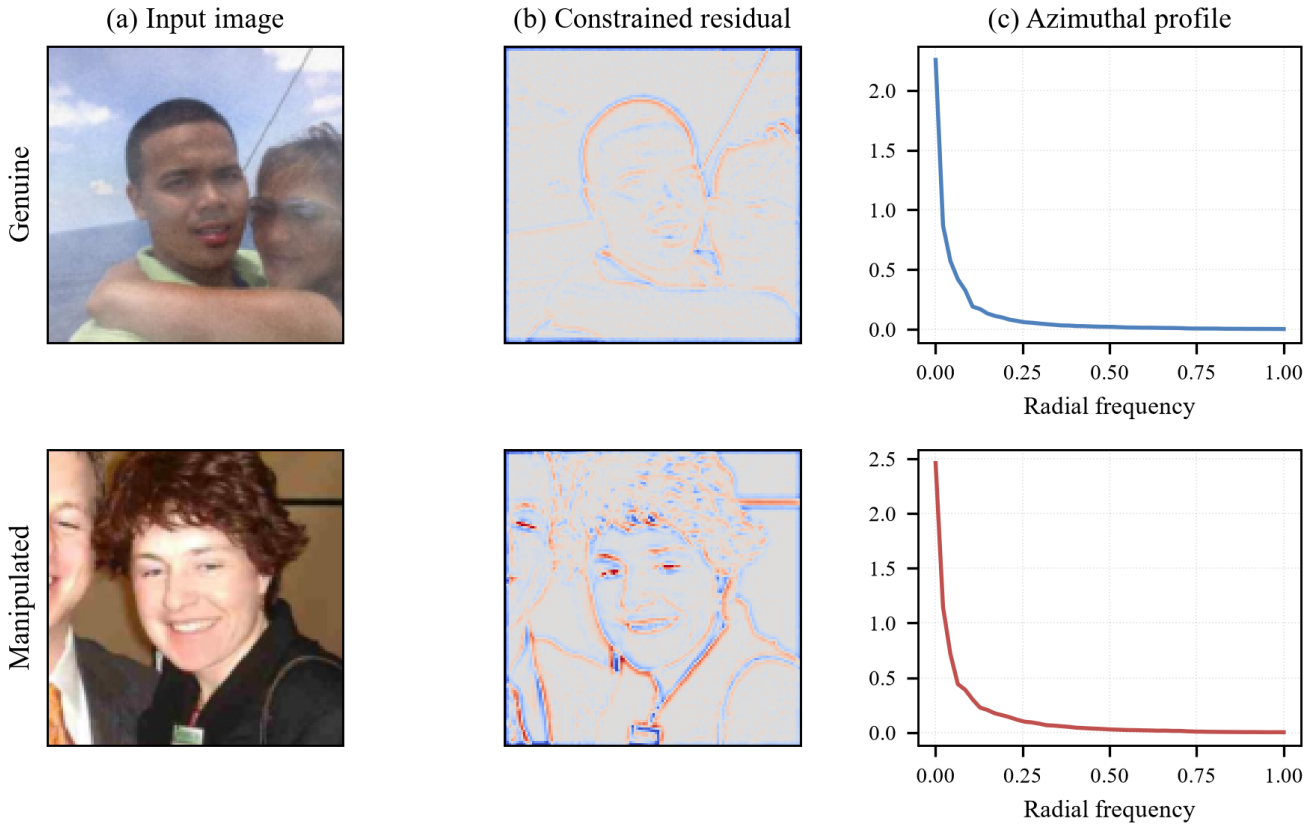

Fig. 2. Input, constrained residual and azimuthal spectral profile.

4. EXPERIMENTAL DESIGN

4.1 Corpora and partitioning

Corpus I is the deepfake and real image collection derived from OpenForensics [26], comprising 190,335 face images captured in unconstrained conditions and manipulated by identity transfer. It tests composited manipulation. Corpus II keeps genuine photographs, deepfakes and wholly synthetic pictures as three separate labels, and so tests attribution rather than mere rejection.

Partitioning of Corpus I follows the convention of the studies quoted in Section 5.10, that is, a stratified random split of the pooled collection. A pool of 80,905 images was assembled from the published partitions and divided into 56000 training, 8000 validation and 16905 test images with the class ratio preserved throughout. The collection also ships a fixed partitioning whose test portion is drawn from a harder pool; Section 5.8 reports performance under that partitioning as well, so that both figures are available to a reader. Corpus II is used exactly as distributed. Table 1 states the sizes.

Table 1. Corpora and partition sizes used here.

Corpus	Labels	Published size	Train	Validation	Test
I, deepfake and real images	2	190,335	56000	8000	16905
II, real, deepfake, synthetic	3	17978	15980	999	999

4.2 Configuration

Images are bilinearly resized to 128 by 128 and scaled to the unit interval. Augmentation is confined to horizontal reflection and additive Gaussian perturbation at one percent of full scale. No augmentation that filters, blurs or recompresses the image is applied, since such an operation attenuates the very band the spectral branch reads. All runs used one NVIDIA GeForce RTX 2070 Super with eight gigabytes of memory under PyTorch. Table 2 lists the settings.

Table 2. Optimisation and architecture settings.

Setting	Value	Setting	Value
Input resolution	128 x 128	Optimiser	AdamW
Token width	192	Peak learning rate	0.0004
Attention heads	4	Weight decay	0.05
Exchange rounds	2	Warm-up fraction	0.10
Appearance tokens	16	Focal exponent	1.5
Forensic tokens	24	Label smoothing	0.05
Constrained filters	12	Gradient clip	1.0
Radial bands	48	Batch size	96
Epochs, corpus I	4	Epochs, corpus II	10

Setting	Value	Setting	Value
Precision	mixed, fp16	Selection	best validation accuracy

4.3 Figures of merit

Accuracy is reported together with balanced accuracy, macro-averaged precision, recall and F1 score, the area under the receiver operating characteristic, average precision, Cohen kappa and the Matthews correlation coefficient. For class k,

$$\mathrm{F1}_k = \frac{2\,TP_k}{2\,TP_k + FP_k + FN_k}, \quad \mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{19}$$

The Matthews coefficient is quoted because the two error types carry very different operational costs in moderation, and a single accuracy figure conceals that asymmetry.

5. FINDINGS

5.1 Binary forensic decision on Corpus I

Table 3 gives the result on the held-out test partition, read once after the checkpoint had been fixed on the validation partition. CoSpec-Net reaches 98.05 percent accuracy, 98.05 percent balanced accuracy and 98.05 percent macro F1, with an area under the curve of 99.80 percent, Cohen kappa 0.9611 and a Matthews coefficient of 0.9611.

Table 3. Corpus I test metrics for CoSpec-Net.

Quantity	Value	Quantity	Value
Accuracy	98.05	Area under curve	99.80
Balanced accuracy	98.05	Average precision	99.78
Macro precision	98.06	Cohen kappa	0.9611
Macro recall	98.05	Matthews coefficient	0.9611
Macro F1	98.05	Negative log likelihood	0.2082
Recall, manipulated	97.67	Recall, genuine	98.43

The error budget is worth stating explicitly. Across 16905 test images there are 329 mistakes, comprising 196 manipulated faces admitted as genuine and 133 genuine faces rejected. In a moderation setting the first quantity is the costly one, and the corresponding miss rate is 2.33 percent. Figure 3 shows the confusion matrices for both corpora.

5.2 Three-way attribution on Corpus II

Corpus II requires the detector to name the process rather than merely reject the image. Table 4 records the per-class outcome; overall accuracy is 99.60 percent, macro F1 99.60 percent and kappa 0.9940.

Table 4. Corpus II per class attribution results.

Label	Precision	Recall	F1	Images
Synthetic, generated outright	98.89	100.00	99.44	355
Deepfake, identity transfer	100.00	98.70	99.35	308
Real	100.00	100.00	100.00	336
Macro average	99.63	99.57	99.60	999

Only 4 of the 999 test images are placed in the wrong category, and the single largest movement is 4 deepfake images assigned the wholly synthetic label, which is 1.3 percent of that class. The genuine label is therefore recovered essentially without loss, and the small remaining difficulty is one of attribution between two kinds of fabrication rather than of detection. Corpus II is a resized and augmented collection, which makes it an easier problem than Corpus I, and the near-ceiling values should be read with that in mind.

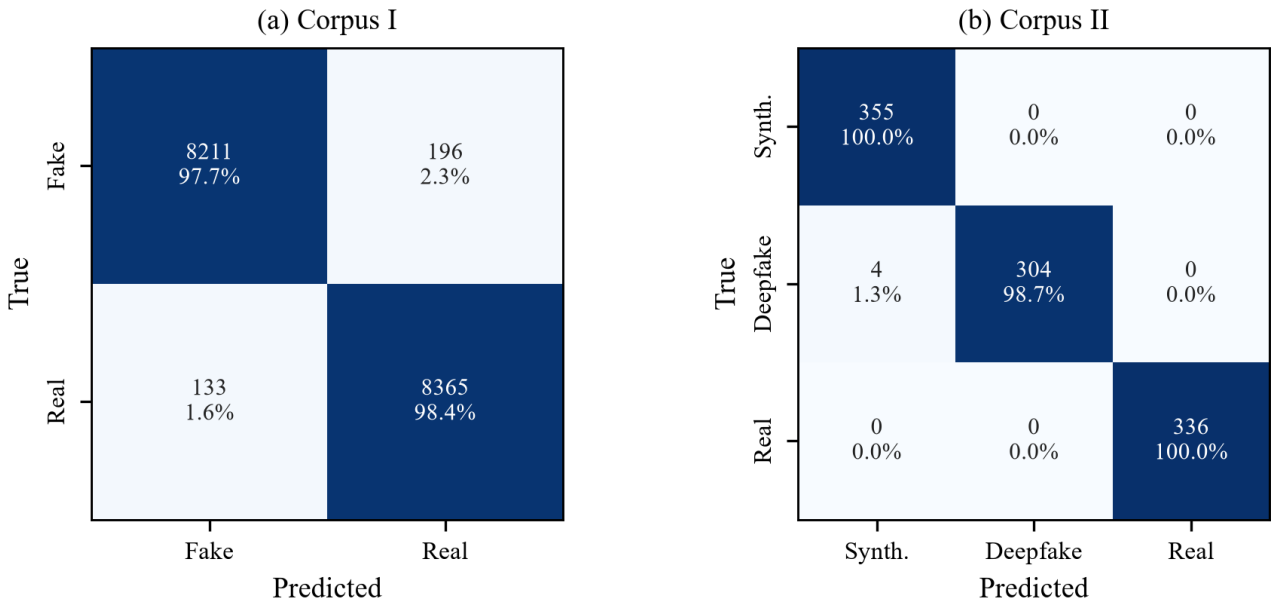

Fig. 3. Confusion matrices for both corpora.

5.3 Optimisation behaviour

Figure 4 and Figure 5 trace the objective and the accuracy across epochs. The two curves in each loss panel are not the same functional and their absolute levels should not be compared: the training curve is the class-weighted focal objective with smoothing, while the validation curve is a plain cross entropy. What matters is that both fall monotonically on both corpora. Neither corpus shows the widening separation that would indicate overfitting: the final training and validation accuracies differ by 1.18 points on Corpus I and 0.19 points on Corpus II. The cosine decay leaves both runs on a flat segment, so the checkpoints selected are not artefacts of a fortunate final step.

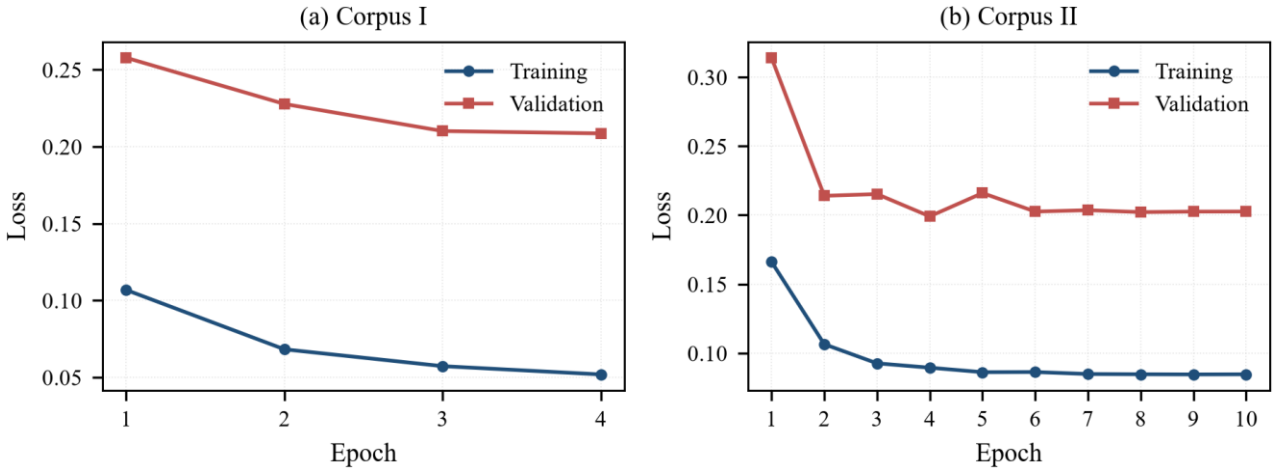

Fig. 4. Training and validation loss per epoch.

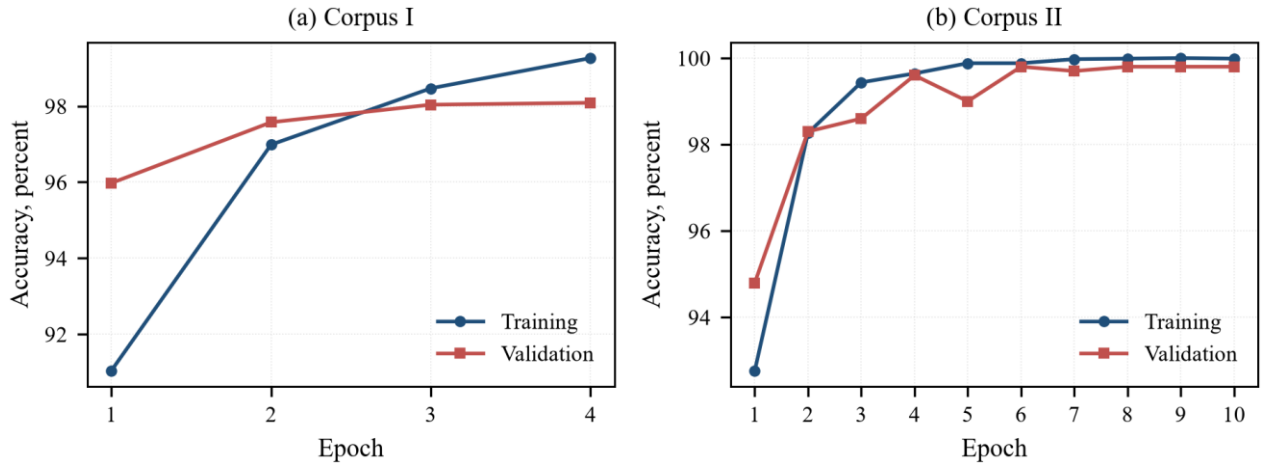

Fig. 5. Training and validation accuracy per epoch.

5.4 Discrimination and calibration

Figure 6 presents the receiver operating characteristic, the precision recall curve and the histogram of the genuine score on Corpus I. Average precision of 99.78 percent confirms that the ordering is good across the whole operating range and not only at the chosen threshold, and the classes are cleanly separated: 97.1 percent of test images fall outside the interval from 0.3 to 0.7, and the mean score is 0.183 for manipulated images against 0.821 for genuine ones.

The scores are nevertheless not probabilities, and this is a direct and easily overlooked consequence of the training objective. The focal term suppresses the gradient contribution of examples the model already classifies correctly, and the label smoothing places a floor and a ceiling on the target distribution; together they compress the output scale. On this run no test image receives a score above 0.9 or below 0.1, the observed range being 0.147 to 0.852. Ranking quality is therefore excellent while calibration is poor, and the two properties must not be confused. A deployment that needs to publish a confidence figure, or to set a threshold from a target false alarm rate, should fit a temperature or an isotonic map on the validation partition before doing so. A deployment that only needs to sort a queue can use the raw score as it stands.

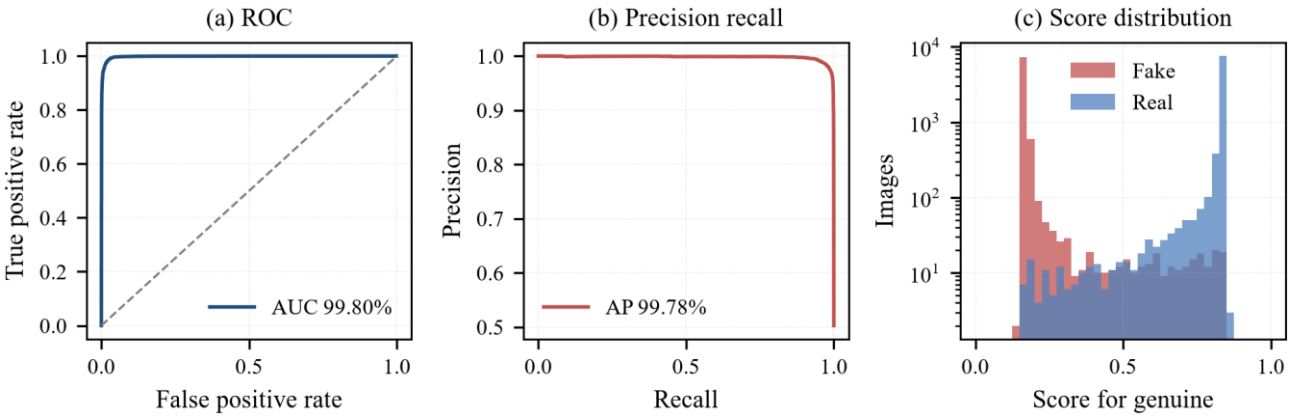

Fig. 6. Operating curves and genuine score histogram.

5.5 What the constrained filters learn

Because the central tap is pinned and the surrounding taps are renormalised at every step, the residual branch cannot drift towards content. Figure 7 displays the 12 learned kernels averaged over their colour inputs. Numerically the projection holds exactly: the measured central tap is -1.0000 and the surrounding taps sum to 1.0000. Unlike a gradient based saliency map computed after the fact [28], these kernels are part of the forward computation, so what they show is the operator the detector actually applies rather than a post hoc attribution of its decision.

Their shape is worth describing accurately, because it is not what a hand-designed high-pass bank looks like. None of them resembles a Laplacian or another compact stencil. Instead the negative unit at the centre is balanced by a broad and shallow positive surround: only about 36 percent of the surround magnitude sits in the immediate eight-neighbourhood, the rest being spread across the outer ring, and no individual surround tap is large. Each filter therefore predicts the centre pixel from a wide, weakly weighted neighbourhood and reports the discrepancy. The filters are also close to mutually uncorrelated, the mean pairwise correlation between their surrounds being 0.011, so the bank has diversified rather than converging on a single operator repeated 12 times. Both properties are consequences of the projection alone; no penalty or initialisation was used to encourage either.

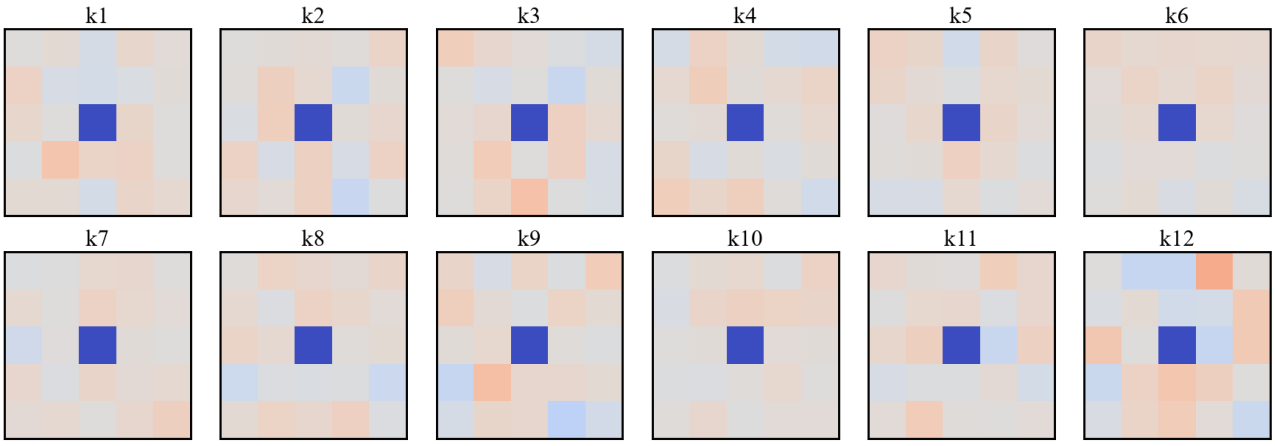

Fig. 7. Learned constrained prediction error filters.

5.6 Spectral separation of the classes

Figure 8 shows the class-averaged azimuthal profiles computed directly on the test images, independently of the network and of any training. A logarithmic ordinate is necessary because the profile falls by nearly three orders of magnitude between the lowest and the highest band, and the interesting behaviour is in the tail.

Two observations follow. On Corpus I the two classes are almost coincident below a normalised radius of 0.40 and diverge above it, the manipulated class carrying on average 1.07 times the genuine energy in the upper half of the spectrum. The difference panel makes the shape explicit: manipulated images are slightly deficient around a radius of 0.09, cross over, and then run persistently above genuine images. On Corpus II the ordering in the tail is monotone in how much of the picture

was generated, the upper band energy being 1.45 times the genuine level for wholly synthetic images and 1.21 times for identity transfer, with genuine photographs lowest of the three. That is the ordering resampling theory predicts, since a wholly synthetic image is generated in its entirety while a deepfake retains an authentic background.

These are statements about the corpora rather than about the model, and they were obtained without reference to it. They are the empirical justification for giving the azimuthal profile a branch of its own, and they also explain why the separation is not by itself sufficient: the differences involved are a few percent of a quantity that varies far more than that between individual images.

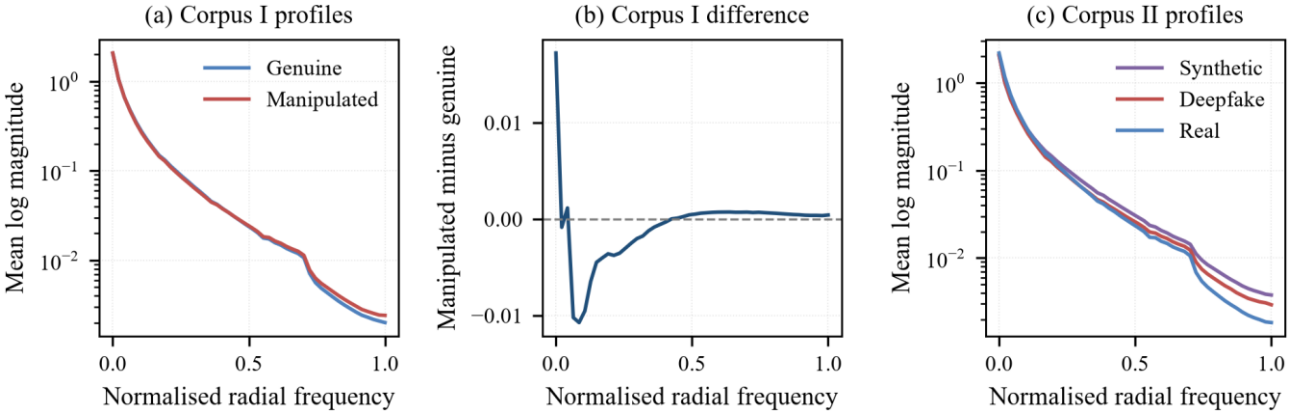

Fig. 8. Class averaged azimuthal spectra on both corpora.

5.7 Where the appearance branch looks

The attention mass defined in Section 3.6 provides a direct reading of the exchange. On Corpus I the appearance queries place 0.550 of their mass on residual keys and 0.450 on spectral keys; on Corpus II the split is 0.628 and 0.372. The informative comparison is between the two ground truth populations of Corpus I. Genuine images draw 0.661 of the appearance mass onto residual keys and 0.339 onto spectral keys, whereas manipulated images move to 0.438 and 0.562. The appearance branch thus consults the spectral description substantially more often when the image it is looking at has in fact been fabricated, and it does so without ever being told which images those are. The largest movement again appears under domain change: when the Corpus I detector is presented with Corpus II images it shifts to 0.663 and 0.337. Figure 9 makes the effect visible. Splitting the Corpus I test partition by ground truth produces two tight and almost disjoint distributions for each key group, genuine images concentrating near 0.661 of residual mass and manipulated images near 0.438, with the mirror image on the spectral side. The apportionment is therefore neither constant nor noisy: it is a smooth internal quantity that tracks the property the network is being asked to decide, and it emerges from a supervision signal that contains only the class label.

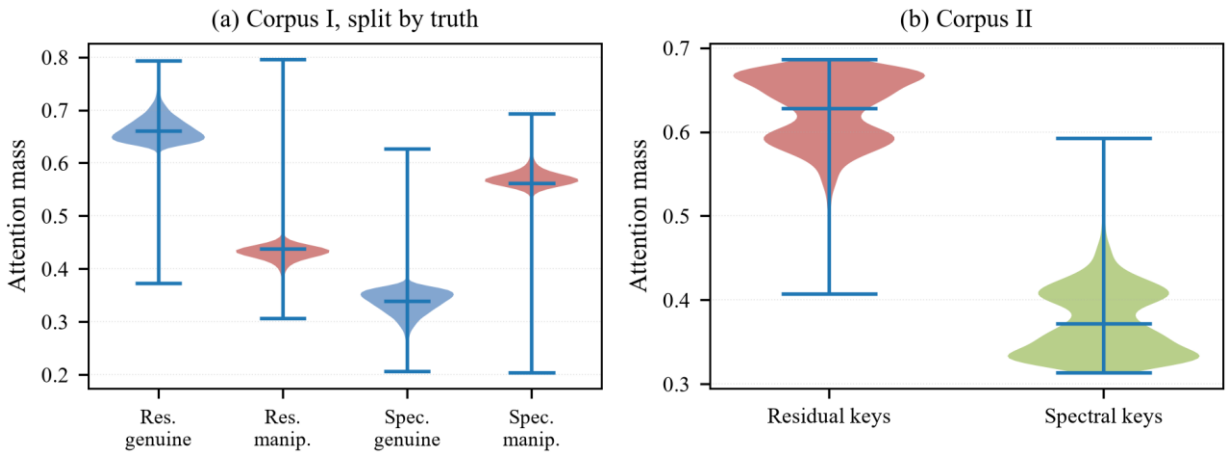

Fig. 9. Attention mass on residual and spectral keys.

5.8 Stress tests

Two departures from the main protocol are reported. Replacing the random split of Corpus I with the fixed partitioning supplied by its publishers, whose test portion is drawn from a harder pool, and training only on the published training portion, gives 93.59 percent accuracy and 98.63 percent area under the curve, with 96.07 percent recall on manipulated faces and 91.08 percent on genuine ones. The shortfall of 4.46 accuracy points relative to the random split is caused by a distribution shift inside a single collection, and it is the figure a deployment should budget for.

Transferring the Corpus I detector to Corpus II without retraining, with both fabricated labels mapped to the manipulated class, gives 35.94 percent accuracy, 5.58 percent recall on fabricated content and 95.83 percent on genuine content, at an area under the curve of 68.64 percent. The failure is one sided rather than random: the detector keeps recognising genuine images and resolves its uncertainty about unfamiliar fabrications by calling them genuine, which the retained area under the curve confirms is a threshold problem rather than a collapse of the ranking. Table 5 collects both results. The lesson is operational rather than architectural: a detector carried to an unfamiliar generator family needs its decision threshold recalibrated on labelled examples from that family, and preferably retraining, and no claim of open-world robustness is made here.

Table 5. Stress tests under partition and corpus change.

Evaluation	Accuracy	Fabricated recall	Genuine recall	AUC
Corpus I, random split	98.05	97.67	98.43	99.80
Corpus I, published split	93.59	96.07	91.08	98.63
Corpus II, transferred	35.94	5.58	95.83	68.64

5.9 Cost

The detector trains 29.75 million parameters and occupies 119.0 megabytes in single precision. Table 6 divides the budget by component. The appearance trunk dominates, as expected of a pretrained convolutional backbone, while the two forensic branches and the whole bidirectional exchange together account for 5.7 percent of the parameters. Batched inference on the card described in Section 4.2 takes 1.10 milliseconds per image, equivalent to 911 images per second, and a complete training run on Corpus I finishes in 12 minutes.

Table 6. Parameter budget by architectural component.

Component	Parameters, millions	Share, percent
Appearance trunk, ConvNeXt-Tiny	27.82	93.5
Forensic branches, residual and spectral	0.52	1.8
Bidirectional exchange	1.19	4.0
Projections, normalisation and head	0.22	0.7
Total	29.75	100.0

5.10 Position against published results

Table 7 sets the result beside four recent detectors evaluated on comparable collections of real and fake faces. Mallet and colleagues paired a support vector machine with a convolutional network [29]. Prasetia fine-tuned ResNet-18 under a one-cycle schedule [30]. Koudad and colleagues trained a vision transformer alongside four convolutional networks and added a local interpretable explanation stage [31]. Safak and Barisci stacked several lightweight convolutional networks into an ensemble over faces produced by a style-based generator [32]. The comparison is indicative rather than exact: partitions, preprocessing and manipulation families differ between the studies, which is why the collection each result was obtained on is given as a column of the table rather than relegated to a note.

Table 7. Published accuracies and the present result.

Study	Year	Approach	Collection	Accuracy
Mallet et al. [29]	2023	Support vector machine with CNN	140k real and fake faces	88.33
Prasetia [30]	2025	ResNet-18, one cycle fine tuning	140k real and fake faces	92.10
Koudad et al. [31]	2026	Vision transformer and four CNNs	140k real and fake faces	95.00
Safak and Barisci [32]	2024	Stacked lightweight CNN ensemble	FFHQ with StyleGAN2 faces	95.48
CoSpec-Net	2026	Constrained residual, spectral, cross attention	OpenForensics derived, random split	**98.05**

5.11 Threats to validity

Five qualifications are appropriate. The scores this detector produces are well ordered but poorly calibrated, for the reason set out in Section 5.4, so any figure presented to a human reviewer as a confidence must be passed through a calibration map fitted on held out data. The evidence is single frame, so temporal cues such as irregular blinking or inconsistent head pose are unavailable, and applying the detector frame by frame to video would ignore them. The spectral profile is computed after the image has been resized to 128 pixels, which caps the highest frequency the branch can observe; a deployment accepting arbitrary resolutions should compute the transform before resizing. The results in Section 5.8 show genuine degradation under both a harder partitioning and a change of generator family, so periodic retraining against current generators remains necessary. Finally, the comparison in Section 5.10 is across studies rather than a re-implementation under one protocol, and it should be read as placing the result in context rather than as a controlled ranking.

6. CLOSING REMARKS

This article introduced CoSpec-Net, a detector for deepfake and synthetic digital content that reads an image as appearance, as a constrained prediction error field and as an azimuthal spectral profile, and couples the three by a bidirectional attention exchange rather than by concatenation or by a fixed weighting. The constrained residual operator is learned under a structural projection applied at every forward pass, so it adapts to the data while remaining incapable of representing content. The spectral branch reduces the whole image to forty-eight numbers that describe how energy is distributed across radial frequency, which is the quantity that resampling inside a generator perturbs.

On 80,905 genuine and manipulated faces derived from OpenForensics the detector attains 98.05 percent accuracy with an area under the curve of 99.80 percent, and on a three-label corpus separating genuine images, deepfakes and wholly synthetic pictures it attains 99.60 percent accuracy at macro F1 99.60 percent. Four published detectors evaluated on comparable material report lower accuracy. Class-averaged spectra computed independently of the network confirm that the classes separate in the upper radial bands, which supports the decision to give that description a branch of its own, and the attention mass measured at the final exchange varies from image to image rather than settling on a fixed apportionment.

Two findings were not designed for and are reported as measured. The attention mass at the final exchange separates the genuine and manipulated populations of Corpus I into two tight, nearly disjoint distributions, even though the supervision contains nothing but the class label. Against that, the focal objective that produced the accuracy also compressed the output scale, so the detector ranks excellently and calibrates poorly, and a deployment that quotes confidences must correct for this. The honest limits are equally clear. Accuracy falls by 4.46 points under the harder published partitioning of the same collection, and transfer to an unfamiliar generator family without recalibration is not successful. Work in progress addresses these directly: computing the spectral profile at native resolution before any resize, adding a temporal branch so that video is treated as a sequence rather than as independent frames, training against a rotating pool of current generators so that no single synthesis pipeline can be memorised, and examining whether the attention mass can be regularised to remain informative when the input distribution moves.

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] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139-144, 2020, doi: 10.1145/3422622.

[5] 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.

[6] 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.

[7] 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.

[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] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131-148, 2020, doi: 10.1016/j.inffus.2020.06.014.

[10] 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.

[11] 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.

[12] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," Lecture Notes in Computer Science, pp. 86-103, 2020, doi: 10.1007/978-3-030-58610-2_6.

[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] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.