International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 6 June. 2026, pp: 240-252 www.ijaem.net ISSN: 2395-5252

Detection of Deepfake Videos Using Long Distance Attention

VadisilaManojKumar[1],YBheemShankarRao[2]
Anil Neerukonda Institute of Engineering and Technology, Visakhapatnam, AP – 531162

Date of Submission: 28-05-2026 Date of Acceptance: 08-06-2026

***Abstract*—Deepfakes shift quickly, advancing at a rapid pace. Nearly indistinguishable artificial faces now emerge through systems like Generative Adversarial Networks, paired with encoderdecoderstructures.Asthesefakesimprove,confidencein online content weakens. Distinguishing what is real from what is fabricatedbecomesharderovertime.Someresearcherstacklethe problembytreatingitlikeayes-or-noquestion.Astrangeglitchin asingleframemightcatchtheireye,whereasdifferentexpertslook at how things move between frames, searching for sudden breaks in flow. Most of these methods fail to consider the whole pictureatonce.Thisresearchhighlightsjusthowfaintdeepfakeflawscan be - spread out, almost invisible. One corner of the face may shift differently than tissue along the chin. In other moments, timing breaks down: speech lacks smoothness, eye closures happen off rhythm. Detecting these missteps requires following fragile hints located in distant areas across the entire video sequence. Neither local convolutional attention with small kernels nor brute-force application of Vision Transformers adequately addresses both dimensions simultaneously. To close this gap, we have introduced the Long Distance Attention (LDA) mechanism. This is a novel patch-level non-local self-attention formulation that computes cross-patch relevance globally and recalibrates the deep feature mapstohighlightpivotalforgeryregions.TheLDAmechanismis instantiated in a Dual-Branch Spatial-Temporal (DBST) network comprising a *spatial branch* that discovers intra-frame artifacts and a *temporal branch* that discovers inter-frame inconsistencies, both unified by the LDA block as the shared core attention module. Training is guided by a multi-task binary cross-entropy loss over frame-level, clip-leveland fused predictions. Extensive experiments on FaceForensics++ (FF++), Celeb-DF v2and the Deepfake Detection Challenge (DFDC) datasets demonstrate that theproposedframeworkconsistentlyoutperformsstate-of-the-art baselines, including MesoNet, Face X-Ray, LipForensics, MACNNandViT-B/16,achievingframe-levelAUCimprovements ofupto1.8%onFF++(c23)and3.3%onCeleb-DFunder cross-datasetevaluation.Ablationstudiesrigorouslyconfirmthat the long-distance attention span, rather than any individual architecturalcomponent,isthekeydriveroftheseimprovements. The paper further provides a theoretical analysis of why long-rangeattentionsubsumesconvolutionalattentionasaspecialcase and why it outperforms LSTM-based temporal modeling in this setting.**

***Index Terms*—Deepfake detection, long distance attention, spatial-temporal model, fine-grained classification, face forgery, self-attention, Vision Transformer, patch-level attention, video forensics, multimedia security, generative adversarial networks, digital content authentication.**

I. INTRODUCTION

THElandscapeofdigitalmediahasundergoneapro-found transformation in the past half-decade. Deepfake technology—the automated synthesis of photorealistic manip-ulatedvideousingdeepneuralnetworks—hasmovedfrom theresearchlaboratorytothepublicdomainwithalarming speed.Theopen-sourceavailabilityofface-swaptools[7]–[9],combinedwithexponentiallyimprovinggenerativemodels [1]–[3], now means that a convincing deepfake video can be producedbyanon-expertwithconsumer-gradehardwarein a matter of hours. Society is already experiencing the down-stream consequences: fabricated political speeches, celebrity impersonations used for fraudand identity theft attacks that bypass biometric authentication systems. The question of how to distinguish a synthetically generated human face from areal one has therefore become one of the most pressing open problems in computer vision and multimedia forensics.Now Deepfake detection struggles most when examining video creation methods. Analysis typically unfolds across three phases. [34] Images emerge one by one from the video frame chain. Each snapshot gets scanned next searching specifically for facial regions. Those found spots later feed into swap operations,drivenbysystemssuchasautoencodersorGANs. [1]Adigitalfacemimicsthemovementsofthesource human, while also adapting changes in light conditions - yet pulls characteristics from another person entirely. In the end, the altered version blends smoothly into the scene where it began. Step by step, side products emerge during the process. While elements develop, variations tend to appear between sections.Lookclose-smallflawsmayrevealthemselves as images render, like mismatched tones where facial zones meet surroundings, or eyes seeming slightly misaligned with adjacent features. [10],[12] Gradually, combining outputs introduces new complications. Because pictures are typically processed individually, [18] without attention to the overall sequence rhythm, [20] oddities start showing: blinks that seem off, mouth motions lacking fluidity, pulse indicators shifting abruptly across frames [17],[21].

A. LimitationsofPriorArt

One way scientists spot fake videos focuses only on still imageswithinthem.Thesemethodsexaminetinydetailsin pixels where fakes often slip up. Performance holds upwhen test examples match what they learned from. But once things change even slightly, accuracy drops fast. Why doesthis happen? Most times, a single example is all they recall, missing signs that point to deception. Because real insightisn't built on repetition, recognition fails when appearances shift slightly. Tracking movement across frames helps certain modern methods spot deepfakes, relying on pulse patterns or memory driven systems. Though tougher to deceive, these approachesoverlookdetailsvisibleinindividualimages.When dealingwithstaticmanipulatedpictureslackingtemporal indicators they falter without delay. Odd blinking behaviors break their logic just as swiftly. A different approach applies attention mechanisms to manage intricate patterns. Much like, the current work, research by Zhao and colleagues frames forgeryidentificationasadetailedcategorizationtask,applying a convolutional architecture enhanced with targeted attention layers to spot subtle facial

inconsistencies. Instead of merging wide-rangingfeaturerepresentations,thesystemoperatesvia tiny 1×1 filters that inspect single pixels individually. Because each location gets evaluated in isolation, broader contradictionsstayunnoticedpictureaneyeappearinggenuine yet clashing with lighting cues around the distant cheekbone. Such piecemeal analysis undermines overall coherence despite accurate local observations. Across a full picture, scanning comesnaturallytoVisionTransformersyetstrongperformance demands vast data. When samples run low, outcomes slip, unlikeclassicCNNsthatstaysteadybydesign.Theiradvantage lies in their natural ability to detect even the smallest surface defects.

B. OurApproachandContributions

Afresharrangementsitsattheheartofthisapproach.Instead of picking either localized filters or the full-scale focus. We defineintegrationasitsfoundation.Aconvolutionalframework shows up first and then generates fine-grained signals like hue and surface patterns. On top, the Long Distance Attention (LDA)componentactsdirectlyuponthoseoutcomes.Division happens when the representation separates into regions, after which adaptive weights adjust relationships between them. Where similarities appear between areas, hidden mistakes start showing up. Focus shifts suddenly to places where changes look suspicious. Instead of using standard models that watch pixels closely, examination targets altered sections through transformed features. Original data gets skipped while review follows derived structures more carefully. Centered in the dual path spacetime framework lies the LDA module. One branch examines single frames, whereas the other analyzes complete movement sequences. Data merges within a common LDA backbone. By combining outputs, static views meet motion trends via cohesive evaluation. When heavy compression affectsvideoinput,howdoesyoursystemrespondmixingfine elements into wider contexts without clear separation? This design is motivated by the observation that spatial artifactsand temporal inconsistencies co-occur in deepfake videos and are complementary sources of evidence that, when jointly exploited,yieldsubstantiallybetterdetectionperformancethan either alone.

Themaincontributionsofthisworkareasfollows:

1) **Long Distance Attention Mechanism:** We propose a novel patch-level non-local self-attention module that,in a single forward pass, computes pairwise relevance betweenall$N$patchesofadeepfeaturemap.Weprovide aformalproofthatstandard$1k$ convolutionalattention is a special constrained case of LDA, establishing LDA as a strict generalization. We further prove that theLDAattentionscorematrixsubsumestheglobalaverage pooling operation as the uniform attention limiting case, demonstratingthatLDAcangracefullydegradetoglobal pooling when all patches are equally informative.

2) **Dual-Branch Spatial-Temporal Framework:** We in-troduce the DBST network that jointly processes per-frame spatial evidence (spatial branch) and multi-frame temporal evidence (temporal branch) using the LDA mechanism as the core module in both branches. The temporal branch extends LDA to the spatio-temporal domain by treating each (t,i) patch-frame token jointly, enabling direct attention between spatially and temporally distant tokens without the sequential bottleneck of recurrent models.

3) **Theoretical and Empirical Justification of Long-Range Attention for Forgery Detection:** We providean analysis, supported by ablation experiments, showing that the *effective attention span*—the spatial distance between the two most strongly attending patches—is a monotonicallyincreasingfunctionofdetectionaccuracy. This provides a principled, quantitative argument forwhy long-distance attention is superior to both short-range convolutional attention and recurrent temporal aggregation for fine-grained forgery classification.

This paper follows a standard structure. Section II looks at previousstudiesinthefield.SectionIIIexplainsthetheoryand whythisresearchmatters.SectionIVdescribesabouttheLDA mechanism and the DBST architecture. Section V details how we set up our experiments and perform. Section VItells about our findings and what they mean. Section VII describes about the results, where the study falls shortand ethics. Section VIII givestheconclusionwithideasforwherethisworkgoesnext.

II. RELATEDWORK

A. DeepfakeGeneration:ABriefTaxonomy

Toseewhyspottingfakesissodifficult.Wehavetolook athowtheyaremade.[1].introducedGANs,asystemwhere a generator$G$ andadiscriminator$D$ trainagainsteachother. Theywillkeepgoinguntil$G$Thisprocessmakesthefinalresult hard to distinguish from the truth. The progressive training strategy of Karras et al. [3] allowed GAN training to scaleto 1024×1024 resolution, dramatically raising the visual qualityofsynthesizedfaces.KingmaWelling[2]introduced theVariationalAutoencoders(VAEs),thebackboneofmany early face-swap pipelines. More recent methods such as StyleGAN2 [4],SimSwap[5]. FSGAN [6] produces the near-perfect face swaps that are imperceptible to the naked eye. The implication for detection is that generation artifacts are becoming increasingly subtle, and a detector that relies on any fixed artifact signature is bound to become obsolete as generation quality improves. This motivates attention-guided, generalizable approaches.

B. Spatial-DomainDeepfakeDetection

Maternetal.[10]wereamongthefirsttosystematically cataloguethevisualartifactfingerprintsofdeepfakegenerators—irregulareyecolor,skin-tonediscontinuitiesandgeometric misalignments—achievingAUCupto0.866withclassical featureextractors.Afcharetal.[11]proposedMesoNet, a shallow CNN focusing on mesoscopic image properties(mid-level patterns between pixel-level noise and semantic content), achieving >98% detection on early Deepfake and Face2Face datasets. Yang et al. [12] exploited facial landmark inconsistencies to flag GAN-generated faces. Zhou et al. [13] fused an RGB stream with a steganalysis feature stream, while Bayar and Stamm [14] proposed a constrained convolutional layer that suppresses

image content to focus on manipulation residuals. Li et al. [15] introduced Face X-Ray, which learnsto detect the blending boundary present in any face composite, achieving strong cross-dataset generalization. Dang et al. [16] Adifferentpathemergedwhenattentionmapsguidedlearning through a CNN structure. Supervision took the shape not by labels alone, but via spatial focus cues shaped during the training. The backbone adapted as the directional signals has highlighted relevant regions. Guidance came from the focused areas, not just the final outputs. Learning shifted where the emphasis was placed across the images. Attention steered the process while convolution layers extracted the features. This mixisdefinedasthemethod'sfoundation.Oneofthefrequent shortcoming of strictly spatial methods is how much they depend on visible hints.A single frame appears, yet soon gives way what to follow moments. Hidden within these shifts lies a story of fragments out of sync, missed without warning. These are marked by time's false impressions.

C. AttentionMechanismsandTransformersforVision

Examining about the topic, Dang et al. (2020) combined the qualitative insights with numerical data. Their approach relied on the methods that captured both the depth and measurable patterns. While one of the part looked at the meanings people assigned and another counted recurring trends. Instead of choosing just one path, they have merged styles to strengthen the findings. Because of that each of the method compensated fortheother'slimits,resultsandgainedbalance.Thisway the interpretation rested on more than numbers alone. Theself-attention mechanism [24] computes for each of the Apiece within the chain appears never by itself, yet formed through varied contributions carrying the distinct importance. When any location alters it influence shifts too and adjustshow the parts connect as the moments pass. Even if bonds changeintensity,thecorestructureremainsconstantthroughout weights are determined by pairwise relevance scores. This enables *O(1)*-hop information flow between any two elements, regardlessofdistance.Dosovitskiyetal.[25]appliedpureself-attention to image patches (ViT), demonstrating strong image classification performance when pre-trained on large corpora. Non-LocalNeuralNetworks[26]introducedspatialnon-local means blocks into CNNs for video understanding, with particulareffectivenessforcapturinglong-rangespatiotemporal dependencies.SqueezeandExcitationNetworks[27]has introducedthechannelwiseattention,recalibratingfeature for forensic feature maps rather than semantic recognition, operating at the patch level rather than the pixel or channel level.

D. Attention-BasedDeepfakeDetection

Zhao et al. [29] reformulated deepfake detection as fine-grainedclassificationandproposedMACNN,usingmulti-ple11 convolutional attention× heads to highlight facial discriminative regions. While conceptually aligned with our work, MACNN's attention has zero receptive field and cannot model non-local patch relationships. Coccomini et al. [30] combined EfficientNet with ViT by feeding CNN features as token sequences to a transformer encoder, showing that hybrid architectures outperform pure ViT on small forensic datasets. Wodajo and Atnafu[31] applied a convolutional ViT directly to video frames. Guo et al. [32] proposed a local-sensitive deepfake attention mechanism focusing on high-frequency artifacts. Our work differs from all of the above in that we: (a) provideaformaltheoreticalanalysisofLDAasageneralization ofconvolutionalattention;(b)integrateLDAintoadual-branch spatial-temporaldesign;and(c)extendLDAtospatio-temporal tokensinthetemporalbranch,whereaspriortransformer-based methods treat only spatial patches.

III. THEORETICALBACKGROUNDANDMOTIVATION

A. WhyDeepfakeDetectionisaFine-GrainedProblem

Definition 1 (Fine-Grained Classification)**.** A classification problemisfine-grainediftheinter-classvisualdifference is smaller than the intra-class visual variation, necessitating localized, discriminative feature extraction guided by global context.

As deepfake generation quality improves, the visual gap betweenrealandfakefacesshrinks:modernGANs[4]produce faces that are globally photorealistic. Defects, when theyexist, are confined to small spatial regions and are often only discriminative in the context of their surroundings. A single artifact-freeregionisuninformative;itbecomesdiscriminative onlywhencontrastedwithagloballyinconsistentcontext.This is precisely the scenario that fine-grained recognition methods are designed forand it motivates the adoption of attention mechanisms that combine global semantic context with local feature discriminability.

B. LimitationsofConvolutionalAttention

Let$\mathbf{F}\in\mathbb{R}^{C\times H\times W}$beadeepfeaturemap.Astandard $k\times k$convolutional attention module produces an attention map $\mathbf{M}\in\mathbb{R}^{H\times W}$by applying a $k\times k$linear filter followed by a sigmoid:

$$\mathbf{M}_{h,w}=\sigma\left(\sum_{c=1}^{C}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\mathbf{W}_{c,i,j}\mathbf{F}_{c,h+i,w+j}+b\right), \quad (1)$$

channelsbymodelingtheinterdependenciesbetweenthem.

The Dual Attention Network [28] has introduced the position attention and the channel attention modules for the scene segmentation. Our LDA mechanism draws inspiration from non-localmeans[26] andViT[25] butisspecifically designed

1 1

where$\mathbf{W}_e \in \mathbb{R}$ and$\mathbf{b}_e \in \mathbb{R}^d$arelearnable.The

C. WhyLDASubsumesConvolutionalAttention

Proposition1. Standard convolutional attention(Eq.(1)) is a special case of LDA (Eq. (**??**)) in which all cross-patch attentionweightsbetweennon-adjacentpatchesareconstrained to zero.

*Proofsketch.*Partition$\mathbf{F}$intopatchesofsize$p \times p$($p=1$ forpixel-levelgranularity).InLDA,theattentionweight$\mathbf{A}_{ij}$ betweenpatch$i$andpatch$j$isunrestricted.Inconvolutional attention,theequivalentweightisnon-zeroonlywhenpatch j falls within the r-neighborhood of patch i, i.e., $\mathbf{A}_{ij} = 0$ whenever the spatial distance between patch centers $d(i,j) > r$. Setting$\mathbf{A}_{ij}=0$for$d(i,j)>r$andrestrictingtheremaining weightstobetranslation-invariant(sharedacrossall$i$)recovers Eq.(1).SinceLDAimposesneitheroftheseconstraints,itis a strict generalization. □

D. WhyLDAOutperformsLSTM-BasedTemporalModeling

Let$\{f_1,\ldots,f_T\}$beaframesequence.AnLSTMprocesses thissequencerecurrently:thehiddenstate$h_t$encodesa summaryof$\{f_1,\ldots,f_t\}$.Theattentionbetweenframef_t andframe$f_{t'}$(where $t' \ll t$) passesthrough $t-t'$recurrent transitions, each of which introduces a multiplicative factor of the gradient, leading to exponential vanishing for large$|t-t'|$[41].InthetemporalLDAformulation(SectionIV-D),the attention weight between a token at (t_1,patch$_i$) and a token at (t_2,patch$_j$) is computed *directly* as a single scaled dot-product, regardlessof$|t_1-t_2|$.Thismeansthatlong-rangetemporal dependencies receive the same gradient flow as short-range ones, resolving the vanishing gradient pathology at the cost of quadratic attention complexity.

intoaspatialscoregrid$\hat{\mathbf{M}} \in \mathbb{R}^{(H/p)\times(W/p)}$andupsampled

IV. PROPOSEDMETHODOLOGY

A. ProblemFormulation

Let$V=\{f_1,f_2,\ldots,f_T\}$beavideoof$T$frames,witheach frame$f_t \in \mathbb{R}^{H\times W\times 3}$.AfacedetectorD(MTCNN[38]inour implementation)producesalignedfacecrops$x_t \in \mathbb{R}^{h\times w\times 3}$. DefineX=$\{x_1,\ldots,x_T\}$asthefaceclip.Thedetection

$c=1\,i=-r\,j=-r$

where$r=k/2$.[Thecriticalobservationisthattheattention score at location (h,w) is a *function only of the local neighborhood*$\mathbf{F}_{h\pm r,w\pm r}$.Twolocations$(h_1,w_1)and(h_2,w_2)$ are*independent*in$\mathbf{M}$if$|h_1-h_2|>2r$or$|w_1-w_2|>2r$.

Consequently,ifaforgeryartifactat(h,w)isonly*detectable* when contrasted against a reference region at (h_2,w_2) far away—as is the case for mismatched eye colors, uncoordi-nated facial texture gradients, or lighting inconsistencies—convolutional attention with small $k$is structurally incapableof detecting it.

single face crop x .$_t$We partition $\mathbf{F}$ into $N=\frac{H'}{p}\times\frac{W'}{p}$non-overlapping spatial patches of size $p \times p$pixels(infeature-mapcoordinates),yieldingpatchtensors $p_1, p_2, \ldots, p_N$with$p_i \in \mathbb{R}^{C\times p\times p}$.Eachpatchislinearlyprojectedtoa d-dimensional embedding:

$$\mathbf{e}_i=\mathbf{W}_e\,\mathrm{vec}(p_i)+\mathbf{b}_e,\ \mathbf{e}_i\in\mathbb{R}^d,\quad i=1,\ldots,N, \tag{2}$$

where$\mathbf{W}_e \in \mathbb{R}^{d\times Cp^2}$ and$\mathbf{b}_e \in \mathbb{R}^d$arelearnable.The embeddingmatrixis$\mathbf{E}=[\mathbf{e}_1,\ldots,\mathbf{e}_N]^\top \in \mathbb{R}^{N\times d}$.

2) ScaledDot-ProductCross-PatchAttention: Weproject $\mathbf{E}$intoqueries,keysandvalues:

$$\mathbf{Q}=\mathbf{E}\mathbf{W}_Q,\quad \mathbf{K}=\mathbf{E}\mathbf{W}_K,\quad \mathbf{V}=\mathbf{E}\mathbf{W}_V, \tag{3}$$

where$\mathbf{W}_Q,\mathbf{W}_K \in \mathbb{R}^{d\times d_k}and\mathbf{W}_V \in \mathbb{R}^{d\times d_v}$.The$N \times N$ LDAweightmatrixis:

$$\mathbf{A}=\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\in\mathbb{R}^{N\times N}. \tag{4}$$

Entry$\mathbf{A}_{ij}\in[0,1]$ quantifies the *global relevance of patch j to patch i*, regardless of their spatial separation. The output context matrix is:

$$\mathbf{Z}=\mathbf{A}\mathbf{V}\in\mathbb{R}^{N\times d_v}. \tag{5}$$

Toenhanceexpressiveness,weuse$H$attentionheads.For head h:

$$\mathbf{A}^{(h)}=\mathrm{softmax}\left(\frac{\mathbf{Q}^{(h)}\mathbf{K}^{(h)\top}}{\sqrt{d_k/H}}\right) \tag{6}$$

$$\mathbf{Z}^{(h)}=\mathbf{A}^{(h)}\mathbf{V}^{(h)}, \tag{7}$$

andthemulti-headoutputis$\mathbf{Z}=\mathrm{Concat}(\mathbf{Z}^{(1)},\ldots,\mathbf{Z}^{(H)})\mathbf{W}_O$, where $\mathbf{W}_O \in \mathbb{R}^{Hd_v\times d}$.

3) Attention Map Generation and Feature Recalibration: Theforgery-importancescoreofeachpatchiscomputedas the row-sum of the attention matrix:

$$s_i=\sum_{j=1}^{N}\mathbf{A}_{ij},\ i=1,\ldots,N. \tag{8}$$

Intuitively,$s_i$measureshowmuchinformationpatch$i$aggre-gates from the entire feature map, serving as a proxy for its discriminative importance. The scores $\{s_i\}_{i=1}^{N}$are reshaped

to the feature map resolution via bilinear interpolation to yield the *LDA attention map* $\mathbf{M} \in \mathbb{R}^{H' \times W'}$. Feature recalibration is then performed by element-wise multiplication:

$$\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{M}, \quad (9)$$

where $\odot$ denotes channel wise broadcast multiplication. The

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 6 June. 2026, pp: 240-252 www.ijaem.net ISSN: 2395-5252

taskistolearnamapping$F_\Theta: X \to [0,1]$,wherethe output is the probability that the clip is a deepfakeandΘ denotesalllearnableparameters.Wetreatthisasafine-grained classificationproblem;thedecisionboundaryisnotdefinedby coarseglobalappearancebutbysubtle,contextuallydependent local artifacts.

B. LongDistanceAttentionMechanism

1) Patch Embedding: Let $\mathbf{F} \in \mathbb{R}^{C \times H' \times W'}$ be the feature mapproducedbyaconvolutionalbackboneappliedtoa duce$\mathbf{F} \in \mathbb{R}^{\quad}$. $_t$

theframe-levelforgeryprobability$\hat{y}^{(s)} \in [0,1]$. $_t$

Thevideo-levelspatialpredictionisthemeanoverall$T$ frames:

$$\hat{y}^{(s)} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(s)}_t. \quad (11)$$

D. TemporalBranch

recalibratedfeaturemap$\tilde{\mathbf{F}}$suppressesnon-discriminative regionsandup-weightspivotalforgerypatches,guidedentirely by long-range global context.

4) ResidualFusion: Toavoidinformationlossfrompoorly calibratedattentionmapsearlyintraining,wefusethe recalibrated and original features via a learnable residual:

$$\hat{\mathbf{F}} = \lambda\tilde{\mathbf{F}} + (1-\lambda)\mathbf{F}, \quad (10)$$

whereλ[0,1 isalearnablescalarinitializedto0.5,allowing thenetworktoadaptivelyblendglobalattentionguidancewith the original local features.

C. SpatialBranch

The spatial branch processes individual frames to captureintra-frame forgery artifacts. The pipeline is as follows:

1) A face crop $x_t$is passed through a CNN backbone (EfficientNet-B4 [37] pre-trained on ImageNet) to pro-

2) TheLDAblock(SectionIV-B)produces$\hat{\mathbf{F}}^{(s)} \in \mathbb{R}^{C \times H' \times W'}$viapatch embedding,attentioncomputationandresidualfusion.

3) Aspatialclassificationheadconsistingofglobalaverage pooling(GAP)→ LayerNormalization→ FC(512)→ ReLU→ Dropout(0.5) → FC(1) → Sigmoid produces

2)Multi-TaskTrainingObjective: Themodelistrained jointly on three predictions:

$$L = L_{BCE}(\hat{y}, y) + L^{(s)}_{BCE}(\hat{y}, y) + L_{BCE}(\hat{y}^{(t)}, y), \quad (15)$$

where$_{BCE}$(p̂, y) = y log p̂ (1y) log(1p̂) and y 0,1 isthegr{und-} truthlabel.Themulti-taskformulationforcesboth branchestoindependentlylearndiscriminativerepresentations, preventing the fusion layer from learning to simply suppress the weaker branch.

F. Algorithm

Algorithm1 DBST-LDATrainingandInference

Require: Face clip $= \{x_1, \ldots, x_T\}$; label$y$; trainedmodel parametersΘ(inferenceonly);learningrateη

Ensure: Prediction$\hat{y}$;updatedΘ(trainingonly)

/*Preprocessing*/

Detect&alignfacesusingMTCNN;resizeto224 224. ×

/*SpatialBranch*/

for $t = 1$ **to** T **do**

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 6 June. 2026, pp: 240-252 www.ijaem.net ISSN: 2395-5252

Thetemporalbranchcapturesinter-frameinconsistenciesby jointly processing a clip of Tconsecutive face crops.

1) Spatio-TemporalTokenSequence: Eachframe$x_t$is encodedbythesameCNNbackbone(sharedweightswith thespatialbranch)toobtain$\mathbf{F}_t \in \mathbb{R}^{C \times H \times W}$.Each frame's featuremapispartitionedinto$N$spatialpatches,yieldingTN totalpatchtokens.The(t,i)-thtokenisembeddedas:

$$\mathbf{e}_{t,i} = \mathbf{W}_e \, \mathrm{vec}(p_i^{(t)}) + \mathbf{b}_e + \mathbf{p}_t + \mathbf{q}_i, \tag{12}$$

where$\mathbf{p}_t \in \mathbb{R}^d$isalearnabletemporalpositionencodingand $\mathbf{q}_i \in \mathbb{R}^d$isalearnablespatialpositionencoding.

2) Spatio-TemporalLDA: TheTNtokensareorganizedinto matrix $\mathbf{E}^{(\mathrm{ST})} \in \mathbb{R}^{TN \times d}$. The spatio-temporal LDA attention matrix is:

$$\mathbf{A}^{(\mathrm{ST})} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(\mathrm{ST})} \mathbf{K}^{(\mathrm{ST})\top}}{\sqrt{d_k}} \right) \in \mathbb{R}^{TN \times TN}, \tag{13}$$

whereprojectionsaredefinedanalogouslytoEq.(3).Entry $\mathbf{A}^{(\mathrm{ST})}_{(t_1,i),(t_2,j)}$ encodestherelevanceofspatialpatch$j$atframe $t_2$to spatial patch $i$at frame t_1, simultaneously capturing both intra-framespatialattentionandcross-frametemporalattention in a single matrix. The output $\mathbf{Z}^{(\mathrm{ST})} \in \mathbb{R}^{TN \times d_v}$is temporally pooled (mean over the Tdimension) and passed through the same classification head architecture as the spatial branch to yield clip-level probability $\hat{y}^{(t)}$.

E. FusionandTraining

1) AdaptiveFusion: Thefinalpredictioncombinesboth branches:

$$\hat{y} = \alpha \hat{y}^{(s)} + (1-\alpha)\hat{y}^{(t)}, \tag{14}$$

where$\alpha \in [0,1]$isalearnablescalarinitializedto0.5.

Algorithm 1 Spatio-TemporalDeepfakeDetection

Input: Video frames $\{x_1, \ldots, x_T\}$
Output: Fake probability $\hat{y}$

for $t = 1$ **to** T **do**
  $\mathbf{F}_t^{(s)} \leftarrow \mathrm{CNN}_{\mathrm{backbone}}(x_t)$
  $\{\mathbf{e}_i\}_{i=1}^{N} \leftarrow \mathrm{PatchEmbed}(\mathbf{F}_t^{(s)})$ //Eq.(2)
  $\mathbf{A} \leftarrow \mathrm{ScaledDotProd}(\mathbf{Q}, \mathbf{K})$ //Eq.(4)
  $\tilde{\mathbf{F}}_t^{(s)} \leftarrow \mathrm{Recalibrate}(\mathbf{F}_t^{(s)}, \mathbf{A})$ //Eqs.(8)–(10)
  $\hat{y}_t^{(s)} \leftarrow \mathrm{ClassHead}(\tilde{\mathbf{F}}_t^{(s)})$
endfor
$\hat{y}^{(s)} \leftarrow \frac{1}{T} \sum_t \hat{y}_t^{(s)}$ //Eq.(11)
*/*TemporalBranch*/*
$\mathbf{E}^{(\mathrm{ST})} \leftarrow \text{ST-PatchEmbed}(x_1, \ldots, x_T)$ //Eq.(12)
$\mathbf{A}^{(\mathrm{ST})} \leftarrow \mathrm{ScaledDotProd}(\mathbf{Q}^{(\mathrm{ST})}, \mathbf{K}^{(\mathrm{ST})})$ //Eq.(13)
$\hat{y}^{(t)} \leftarrow \mathrm{ClassHead}(\mathrm{TemporalPool}(\mathbf{Z}^{(\mathrm{ST})}))$
*/*Fusion*/*
$\hat{y} \leftarrow \alpha \hat{y}^{(s)} + (1 - \alpha)\hat{y}^{(t)}$ //Eq.(14)
if training **then**
  $L \leftarrow$ Eq.(15)
  $\Theta \leftarrow \Theta - \eta \nabla_\Theta L$
endif
return $\hat{y}$

V. EXPERIMENTALSETUP

A. Datasets

We evaluate the proposed method on three widely used benchmark datasets selected to cover different levels of difficulty, manipulation diversity, and dataset scale.

FaceForensics++ (FF++) [34]: This dataset contains a 1,000 original YouTube videos along with 4,000 manipulated videos whichhasbeengeneratedusingfourfacemanipulationmethods. The methods include DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The videos are provided at three compression settings: raw (uncompressed), lightcompression(`c23`),andheavycompression(`c40`).FF++ is commonly used as a standard benchmark for evaluating in-distribution deepfake detection performance. We report results at`c23`and`c40`toassessrobustnesstocompression.

Celeb-DF v2 [35]: A more challenging dataset containing 590 realcelebrityinterviewvideosand5,639high-qualitydeepfake videos synthesized with a significantly improved pipeline over first-generation tools. The higher visual quality makes Celeb-DF an important test of generalization beyond FF++.

Deepfake Detection Challenge (DFDC) [36]: The largest publicly available deepfake dataset (>100,000 videos), featuring diverse actors, lighting conditions, backgroundsandmultiplecutting-edgegenerationmethods.DFDCwasexplicitly designed to challenge overfitted detectors and is the closest proxy for real-world deployment conditions.

B. ImplementationDetails

Facepreprocessing: FacesaredetectedusingMTCNN[38], aligned to 5 canonical landmarksand cropped to 224224.For FF++ we use 32 frames per video sampled uniformly; for Celeb-DF and DFDC we use 16 frames due to variable video length.

CNNbackbone: EfficientNet-B4[37]pre-trainedonImageNet; thelastclassificationlayerisremoved.Theoutputfeaturemap is$C = 1792$channelsat$H' = W' = 7$.

LDA hyperparameters: Patch size $p = 1$ (treating each spatial location as a patch, yielding $N = 49$ tokens per frame); embedding dimension $d = 512$; key/query dimensiond_k=64; value dimension d_v=64; $H = 8$ attention heads; $d_k/H = d_v/H = 64/8 = 8$perhead.

Temporalbranch: Cliplength$T$=8frames,giving $TN = 392$ spatio-temporal tokens. To manage the quadratic complexity$O((TN)^2) = O(392^2)$153kattentionentries perclip,weapplyFlashAttention[40]formemory-efficient computation.

Optimization: AdamW[39] with learning rate 10^{-4}, weightdecay 5×10^{-4}, batch size 32, 30 training epochsand cosineannealing with warm restarts. Backbone is frozen for the first5 epochs then fine-tuned with a 10 × reduced learning rate.

Data augmentation: Random horizontal flip, random JPEG compression(quality50–100),Gaussianblur(σ:0–1.5),brightness/contrast jitter (±15%)and random erasing.

Hardware: Two NVIDIA A100 80GB GPUs; total training time approximately 14 hours per model on FF++.

C. BaselineMethods

We compare the proposed DBST-LDA against the following baselines spanning all three detection families:

*Spatial-domain:*MesoNet[11],Two-Stream[13],FaceX-Ray [15], EfficientNet-B4 (fine-tuned, no attention module).

*Temporal-domain:*LipForensics[23],recurrentCNN(ResNet-18 + GRU) [22].

*Attention/Transformer:*MACNN[29],ViT-B/16fine-tuned[25], EfficientNet+ViThybrid[30].

D. EvaluationMetrics

Following standardpractice inthe deepfake detectionlitera-ture, we report:

- **Accuracy(Acc)**:overallcorrectclassificationrateat threshold 0.5.

- **Precision(P)**:TP/(TP+FP).
- **Recall(R)**:TP/(TP+FN).
- **F1-Score**:harmonicmeanofPandR.
- **AUC**:areaundertheROCcurve(threshold-free).

For cross-dataset experiments, only AUC is reported to allow fair comparison with published results.

VI. RESULTSANDANALYSIS

A. In-DistributionPerformanceonFaceForensics++

TableIpresentsframe-levelresultsonFF++(c23compres-sion).

TABLEI
IN-DISTRIBUTIONPERFORMANCEONFF++(c23).FRAME-LEVEL EVALUATION.

Method	Year	Acc(%)	P(%)	R(%)	F1(%)	AUC(%)
MesoNet[11]	2018	83.1	82.8	83.5	83.1	85.2
Two-Stream[13]	2017	86.9	85.4	87.6	86.4	88.0
EfficientNet-B4(noattn)	–	92.4	91.9	92.8	92.3	94.8
FaceX-Ray[15]	2020	91.3	90.5	91.9	91.2	93.5
MACNN[29]	2021	93.1	92.7	93.4	93.0	95.4
LipForensics[23]	2021	90.5	89.8	91.1	90.4	92.7
ViT-B/16[25]	2021	91.7	90.9	92.3	91.6	94.0
Effic+ViT[30]	2022	93.6	93.1	94.0	93.5	95.9
DBST-LDA(Ours)	–	**95.6**	**95.2**	**95.9**	**95.5**	**97.2**

TheproposedDBST-LDAmodelachievesthehighestscores across all five metrics, with an AUC of 97.2%—a gain of1.8% over the strongest baseline (Effic+ViT, 95.9%) and 1.3% over MACNN (95.4%). The margin over the backbone-only EfficientNet-B4(94.8%)quantifiesthecontributionoftheLDA attention mechanism in isolation: the 2.4% AUC improvement is attributable entirely to the attention module.

Table IIpresents outcomes observed with intense c40compression applied Here, challenges grow more pronouncedduetothefrequentpresenceofsuchfactorsin actual scenarios Most platforms alter video content significantly once it's uploaded.

TABLEII
IN-DISTRIBUTIONPERFORMANCEONFF++(c40).FRAME-LEVELAUC(%).

Method	AUC(%)
MesoNet[11]	70.3
FaceX-Ray[15]	79.5
MACNN[29]	83.8
LipForensics[23]	82.4
Effic+ViT[30]	86.1
DBST-LDA(Ours)	**88.4**

Evenifitismuchsmallernow,DBST-LDAremainsas the part of the ranking order. Its position persists as thedespite of the shrinkage. It hasn't vanished with size of thereduction. It has still present in the line, though it scaled downsignificantly.Thesequenceincludesit,evenwhenpared backsharplyThoughhitting88.4%AUC,thisapproachjustout

performsthenextbestat86.1%.YetOddlyenough,distinctions remain clear even when elements are far apart in the sequence Thoughsqueezedtight,subtletracesstillshowaresultofslow, steady filtering. Fading happens, but intricacy remains hidden under polished surfaces. Clarity slips away, though delicate differences stay present in the gaps among dots. Reduced size means less accuracy; even so, gentle rhythms emerge softly after multiple passes Occasionally, devices designed to detect spatial characteristics rely on

B. Cross-Dataset Generalization

Table III displays AUC values in percent (%) , measured throughevaluationsonvariousdatacollections:modelsTrained onFF++(c23)EvaluatedUsingCeleb-DFandDFDC*method gets tested without changes applied*. Evaluation here stands as the most demanding type of review detector generalization.

TABLEIII
CROSS-DATASETGENERALIZATIONEVALUATEDONCELEB-DFV2AND DFDC.AUC(%).

Method	Celeb-DFv2	DFDC
MesoNet[11]	54.8	55.9
Two-Stream[13]	61.4	61.0
FaceX-Ray[15]	74.2	70.5
MACNN[29]	79.6	75.1
LipForensics[23]	82.4	73.5
ViT-B/16[25]	77.3	72.8
Effic+ViT[30]	83.1	77.4
DBST-LDA(Ours)	**85.7**	**79.3**

DBST-LDA improves cross-dataset Celeb-DF AUC by 2.6% over Effic+ViT and by 6.1% over MACNN. overall improvementthroughtheDFDCis1.9%overEffic+ViT.These results suggest through long-distance patch attention learns more universal forgery manners indicators that transfer better overallacross generationmethods than artifact-specific spatial features.

C. AblationStudy

Table IV systematically contribution of each component by their progressive construction on Seeds.

TABLEIV
ABLATIONSTUDYONFF++(C23).AUC(%).

Configuration	FF++	Celeb-DF	DFDC
Backboneonly(noattention)	94.8	73.0	66.4
+Localconv.attn.(1×1)	95.0	73.3	66.7
+Localconv.attn.(3×3)	95.2	73.9	67.1
+LDA,N=16(largepatches)	95.6	76.1	68.9
+LDA,N=49(mediumpatches)	96.5	80.2	72.4
+LDA,N=196(fine-grained)	96.8	82.1	74.1
+Temporalbranch(LDAonly)	96.3	81.5	73.8
FullDBST-LDA(spatial+temporal)	**97.2**	**85.7**	**79.3**

SeveralkeyinsightsemergefromTableIV:

(1)

(2) Localattentionprovidesmarginalgain. Adding1 × 1 or 3×3Looking closely at the test results, three distincttrends come into view. When different kinds of input are applied,subtletendenciesstarttosurface.Incaseswhere data lacks clarity, a drop in effectiveness shows up each time. Looking at the outcomes explains why local attention has minimal impact.+0.2–0.4% Gains on new, unseen data hardly shifted regardless of those modifications. Limited receptive fields cannot grasp relevant structures effectively. What truly counts is recognizing relationships across distant parts of an image. For small filters, context remains out of reach. Looking far apart reveals more telling signs than what's right next to you. Improvement in spotting patterns begins when attention moves outward instead of staying near.

(3) LDAgainincreaseswithfinerpatchgranularity. Though tinypatchesboostLDAperformance,gainsemergeonlybeyond a threshold. As numbers rise N=16 to N=196 results climb step by step across all data sets. Finer facial divisions make smalldefectsstandoutwithintightzones.Whenregionsstretch wider, subtle mismatches begin to merge. These faint signs may vanish entirely if space expands too much.

(4) Both branches contribute. Ultimately,integratingtemporal patterns with spatial data turns out critical. 96.8% FF++ AUC and82.1%Celeb-DFAUC,Sincedependingsolelyonlocation-based traits. Still, outcomes improved noticeably 96.3% and 81.5%, respectively. The full dual-branch model (97.2% and 85.7%) after fusing both streams. What stands out is how odd visuals and stiff motion point in different directions. When combined, these signs make spotting fakes easier - more so than either clue by itself.

D. AnalysisofAttentionSpanvs.DetectionAccuracy

Surprisingly,thegapbetweentop-twoactivepatchesrevealed patterns during validation. From this point on, wide focus seemed tied to better counterfeit spotting. Using the 90th percentile carved out a clear boundary - this line then defined how far attention was allowed to spread. Afterward came six modelversions,noneabletoattendpastfixedlimits.Eachbuilt under tighter constraints, yet followed the same core design. Fromapointbeyonddistance$r$,removingconnectionsshowed clearly the reliance on far-off information. Each version's resultsOnlinedmeasuredusingFF++AUC.Asaccuracyrises, interaction between more separated regions is allowed. Better detection emerges when widely spaced parts of faces connect. AUC rises from 94.8% at r=0 (degenerate, equivalent to their attention)to97.2%at$r$=(fullLDA).Everyadditional 2-unit increase in r yields a signs through the significant AUC improvement (p <0.05, paired t-test allover 5 seeds).

E. QualitativeAttentionMapAnalysis

Thebrightzonestendtoclusteraroundthefacialedgesjaws, cheeks and necks when it is spotting synthetic layouts. These patterns appear repeatedly in the LDA attention maps across the different samples. If there is Strong attention concentrated aroundtheeyeboundariesandnostrilregionsthenthere will be inconsistencies such as uneven blending and texture discontinuities are more likely to occur. In real videos, visual information is usually spread more naturally across the frame. The manipulated content tends to produce the localized attention near the facial edges.

Abruptthemotionvariationsarealsonoticeableintheforged sequences. Even if there is some of the distant frames appear visually sharp, nearby temporal inconsistencies can still reveal the signs of the manipulation. The proposed approach learns these through structural relationships directly from the data without relying on manually on the defined rules. Unlike the MACNNwhichmayloseglobalcontextduetothefragmented feature extraction. The proposed method preserves the broader spatialrelationships.Italsoremainsasmorestablethantheviт-based approaches in situations which involves the reflections, lighting variation, or noisy textures.

VII. Discussion

A. WhyLDAOutperformsCNNandLSTM:ASummary

The experimental results that corporate the theoretical anal-ysis of the Section III. To summarize the argument concretely: **vs. CNN:** A k×$k$conversion towards the attention module canonly*internally*comparefeaturesthemselveswithina k×$k$window. Stacking L such layers expands the effective receptive field to $O(kL)$, but: (a) the dependence between distant locations is *indirect*, mediated through a chain of intermediatelayers;(b)eachlayerappliesthesametranslation-invariantfiltertoalllocations,unabletocompute*location-pair-specific* relevance. LDA computes explicit pairwise relevance between all N patches in a *single* attention step, enabling $O(1)$-depth global reasoning.

vs.LSTM: Recurrentmodelssufferfrom:(a)sequential processingthatpreventsparallelism;(b)vanishinggradientsfor long-range dependencies (Section III); (c) inability to capture *spatial* non-local dependencies within frames; (d) sequential bottleneck that cannot attend directly from frame $t_1$to frame$t_2$withoutpassingthroughallintermediatehiddenstates.LDA processes all TN tokens jointly and in parallel, with direct gradientflowbetweenanytwotokensregardlessoftheirspatio-temporal distance.

B. Limitations

Like most of theresearch systems, the proposed framework also has a few practical limitations.

Quadratic attention∞ complexity: The main computational costcomesfromthefullcross-patchattentionmechanism.The spatio-temporal LDA matrix scales as $O((TN)^2)$. The current setup with T=8 frames and N=49 patches per frame. The model processes around 153k attention entries per clip, which remains manageable using FlashAttention[40]. Increasing the clip length significantly raises memory usage. For example using T=32 results in more than the 2.4 million entries which makes the computation much more demanding. Future work may explore the linear attention methods [42] or sparse the attention strategies to reduce this overhead while maintaining the forensic accuracy.

Sensitivitytocompression: Strong c40compression signifi-

cantlyreducestheoverallperformance.ThiscanlowertheAUC from97.2%to88.4%onFF++.Althoughtheproposed method still performs better than all evaluated baselines. The AUC of 88.4% may not be sufficient for high-risk applications such as the large-scale content moderation. The legal verification orthemediaauthentication.Themainchallengeisthataggressive video compression removes many of the fine texture details. Theattentionmechanismreliesonthedetectingmanipulations. Once this information is lost during the encoding, it cannot be fully recovered by the model.

Adversarialdeepfakes: Theresultsreportedinthisworkwere obtainedusingthevideoscreatedwithouttargetingtheproposed detector. In the practical scenarios, this assumption may not always the hold. In the Previous studies include the work by Neekhara[43] have shown that the attackers can generate manipulated videos which are designed specifically to bypass thedetectionsystems.Thecurrentstudydoesnotinclude the adversarial robustness experiments, which remains an the important limitation. A determined attacker with the access to the detector's behaviour or the architecture could potentially reduce the detection of the performance beyond the results shown in Table I. Incorporating these adversarial examples duringthetrainingmayhelptoimprovetherobustnessagainst such attacks in future work.

Scopeofmanipulation: Theproposedpipelinefocuseson a specific type of manipulation. This consits of fake human face inserted into real video frame. This represents a major category of the deepfake content, several other forms of manipulation which are not addressed in the current work. These includes voice cloning with minimal facial motion, full-body motion transfer and audio-driven lip synchronisation techniques.Limitingthestudytofacialforgerydetectionhelped keep the problem focused and manageable. The system should not be considered a complete solution for all the types of deepfake media.

C. EthicalConsiderations

Thedevelopmentanddeploymentofthedeepfakedetection systems also raise the important ethical concerns. Theseissues influence the public trust, personal reputation and the responsibleuseoftheAItechnologies.Discussingthemopenly is necessary to understand their broader social impact.

False-positive harm: An incorrect deepfake prediction can seriously affect an individual's reputation. A genuine video thatismistakenlyflaggedasmanipulatedmaycreatesuspicion even after the error is corrected. In many cases, public doubt spreadsfasterthantheclarificationandthedamagecanpersist long after the original claim is withdrawn. In Such situations may also affect the journalism, legal investigations and public discourse, where the inaccurate conclusions can lead to unfair consequences.

Forthisreason,deployingDBST-LDAinreal-worldsettings requires careful calibration and reliable uncertainty estimation methods, such as temperature scaling or Bayesian posterior analysis. In sensitive applications, model predictions should always be reviewed by human experts before final decisions are made.

What a detector fails to capture is the influences of how findingsareseen.Whenlimitationsareknown,thetrust grows.Underspecificcircumstances,errorsemergemoreoften. Awareness of failure modes helps avoid misuse. Reliability depends on recognizing where function ends.

Arms-race dynamics: Publishing a detection architecture Within this section, every detail appears exactly as open practices usually require. Because of the demands reveal what traits matter most is the rivals start noticing patterns. Whatone prioritizes is the signals weaknesses others might exploit. Focus shifts when goals highlight specific qualities. Attention followswheretheeffortisspent.Cluesemergethroughchoices which are made under pressure Move along the sensor's edge. We find this acceptable because: (a) long-term societal benefit ofacommunity-reviewed,reproducibledetectorWhatremains significant outweighs fleeting shortcomings; our observations consist of

Some of these components differ too much to be uniformly reduced without affecting how sharp the image appears.

Privacy and consent: Training datasets (FF++, Celeb-DF, DFDC)involverealindividuals'facevideos.Wehaveusedonly publicly released benchmarks with documented institutional review board approvals. Any downstream deployment that involves collecting new face videos for detector retraining must obtain explicit informed consent from the subjects.

Misuse for censorship or surveillance: A deepfake detector couldbeco-optedtobuildsurveillancetoolsortofalsely flag authentic journalist videos as fake. We strongly caution againsttheseapplicationsandadvocateforopenlicensingwith use-case restrictions.

Bias and fairness: Deepfake detectors trained primarily on datasets of celebrity faces may have systematically lower accuracy for demographic groups underrepresented in training data. Future work must include explicit demographic bias audits.

VIII. CONCLUSION

This paper addressed the challenge of deepfake video detectionfromanoveltheoreticalandarchitecturalperspective. We observed that contemporary deepfakes introduce forgery artifactsthatareinherently*fine-grainedandnon-local*:theyare onlydiscriminativewhenasuspiciouslocalregionisevaluated in the context of globally inconsistent surroundings. This motivated the design of the Long Distance Attention (LDA) mechanism, a patch-level non-local self-attention module that computespairwiserelevanceacrossthe*entire*featuremap inasingleattentionstep,withoutthelocalityconstraintsof convolutional attention or the sequential bottleneck of recurrentmodels.WeprovidedaformalproofthatLDAstrictly generalizes local convolutional attention and a theoretical analysis of why it resolves the vanishing-gradient pathologyof LSTMs for long-range temporal modeling.

The LDA mechanism was embedded in a Dual-Branch Spatial-Temporal (DBST) network that simultaneously ex-ploits intra-frame spatial artifacts and inter-frame temporal inconsistencies—thetwoprimaryevidencestreamsforforgery detection—through a unified attention framework.Training was guided by a multi-task cross-entropy loss that independently

supervisesbothbranches,ensuringeachlearnscomplementary discriminative representations.

Extensive experiments on FF++, Celeb-DF v2and DFDC demonstrated state-of-the-art performance, with particular improvementsincross-datasetgeneralization(+2.6%AUC on Celeb-DF over the strongest baseline), confirming that LDA-guided features are more universal than the artifact-specific fingerprints learned by local spatial detectors. Wehope this work provides a solid theoretical foundation and strong empirical baseline for these future directions.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Inf.Process. Syst.*, vol. 27, Montre´al, Canada, 2014.
[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in*Proc. Int. Conf. Learn. Represent.*, Banff, Canada, 2014.
[3] T. Karras, T. Aila, S. Laineand J. Lehtinen, "Progressive growing ofGANs for improved quality, stabilityand variation," in *Proc. Int. Conf.Learn. Represent.*, Vancouver, Canada, 2018.
[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinenand T. Aila,"Analyzing and improving the image quality of StyleGAN," in *Proc.IEEE/CVF CVPR*, Seattle, WA, USA, 2020, pp. 8107–8116.
[5] R. Chen, X. Chen, B. Niand Y. Ge, "SimSwap: An efficient frameworkfor high fidelity face swapping," in *Proc. ACM Int. Conf. Multimedia*,Seattle, WA, USA, 2020, pp. 2003–2011.
[6] Y. Nirkin, Y. Kellerand T. Hassner, "FSGAN: Subject agnostic faceswapping and reenactment," in *Proc. IEEE/CVF ICCV*, Seoul, Korea,2019, pp. 7184–7193.
[7] "Deepfakes,"https://github.com/deepfakes/,AccessedSep.2019.
[8] "FakeApp,"http://www.fakeapp.com/,AccessedFeb.2020.
[9] "FaceSwap,"https://github.com/MarekKowalski/,AccessedSep.2019.
[10] F. Matern, C. Riessand M. Stamminger, "Exploiting visual artifacts toexpose deepfakes and face manipulations," in *IEEE Winter Conf. Appl.Comput. Vision Workshops*, Waikoloa, USA, 2019, pp. 83–92.
[11] D.Afchar,V.Nozick,J.YamagishiandI.Echizen,"MesoNet:Acompactfacial video forgery detection network," in *IEEE Int. Workshop Inf.Forensics Security*, Hong Kong, 2018, pp. 1–7.
[12] X.Yang,Y.Li,H.QiandS.Lyu,"ExposingGAN-synthesizedfacesusinglandmark locations," in *Proc. ACM Workshop Inf. Hiding MultimediaSecurity*, Paris, 2019, pp. 113–118.
[13] P. Zhou, X. Han, V. I. Morariuand L. S. Davis, "Two-stream neuralnetworks for tampered face detection," in *IEEE CVPR Workshops*,Honolulu, USA, 2017, pp. 1831–1839.
[14] B. Bayar and M. C. Stamm, "A deep learning approach to universalimagemanipulationdetectionusinganewconvolutionallayer,"in*Proc.4thACM WorkshopInf.HidingMultimediaSecurity*,Vigo,Spain,2016, pp.5–10.
[15] L.Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wenand B. Guo, "FaceX-Ray for more general face forgery detection," in *Proc. IEEE/CVFCVPR*, Seattle, WA, USA, 2020, pp. 5001–5010.
[16] H. Dang, F. Liu, J. Stehouwer, X. Liuand A. K. Jain, "On the detectionof digital face manipulation," in *Proc. IEEE/CVF CVPR*, Seattle, WA,USA, 2020, pp. 5781–5790.
[17] U. A. Ciftci, I. Demirand L. Yin, "FakeCatcher: Detection of syntheticportrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach.Intell.*, 2020, doi: 10.1109/TPAMI.2020.3009287.
[18] Y.Li,M.-C.ChangandS.Lyu,"InIctuOculi:ExposingAIcreatedfakevideos by detecting eye blinking," in *IEEE Int. Workshop Inf. ForensicsSecurity*, Hong Kong, 2018, pp. 1–7.
[19] M.Li,B.Liu,Y.HuandY.Wang,"Exposingdeepfakevideosbytrackingeyemove ments,"in*Proc.25thInt.Conf.PatternRecognit.*,Milan,Italy,2021, pp. 5184–5189.
[20] C.-Z. Yang, J. Ma, S. Wangand A. W.-C. Liew, "Preventing deepfakeattacks on speaker authentication by dynamic lip movement analysis,"*IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1841–1854, 2021.
[21] S. Fernandes *et al.*, "Predicting heart rate variations of deepfake videosusingNeuralODE,"in*IEEE/CVFICCVWorkshop*,Seoul,Korea,2019, pp.1721–1729.
[22] E. Sabir *et al.*, "Recurrent convolutional strategies for face manipulationdetectioninvideos,"in*Proc.IEEE/CVFCVPRWorkshops*,LosAngeles,US A, Jun. 2019.
[23] A. Haliassos, K. Vougioukas, S. Petridisand M. Pantic, "Lips don't lie:A generalisable and robust approach to face forgery detection," in *Proc.IEEE/CVF CVPR*, Virtual, 2021, pp. 5039–5048.
[24] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process.Syst.*, vol. 30, Long Beach, CA, USA, 2017.
[25] A. Dosovitskiy*et al.*, "An image is worth 16x16 words: Transformersfor image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*,Virtual, 2021.
[26] X.Wang,R.Girshick,A.GuptaandK.He,"Non-localneuralnetworks,"in *Proc. IEEE/CVF CVPR*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.
[27] J. Hu, L. Shenand G. Sun, "Squeeze-and-excitation networks," in *Proc.IEEE/CVF CVPR*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.
[28] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc.IEEE/CVF CVPR*, Long Beach, CA, USA, 2019, pp. 3146–3154.
[29] J. Zhao, L. Han, L. Shaoand C. Bao, "Multi-attentional deepfakedetection,"in*Proc.IEEE/CVFCVPR*,Nashville,TN,USA,2021, pp.2185–2194.
[30] D.A.Coccomini*etal.*,"CombiningEfficientNetandVisionTransformersfor video deepfake detection," in *Proc. Int. Conf. Image Anal. Process.*,Lecce, Italy, 2022, pp. 219–229.
[31] D.WodajoandS.Atnafu,"Deepfakevideodetectionusingconvolutionalvision transformer," *arXiv preprint arXiv:2102.11126*, 2021.
[32] Z. Guo *et al.*, "Controllable guide-space for generalizable face forgerydetection," in *Proc. IEEE/CVF ICCV*, Paris, France, 2023, pp. 22211–22221.
[33] J.Fu,H.ZhengandT.Mei,"Lookclosertoseebetter:Recurrentattentionconvoluti onalneuralnetworkforfine-grainedimagerecognition,"in*Proc.IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 4438–4446.
[34] A. Ro¨ssler*et al.*, "FaceForensics++: Learning to detect manipulatedfacialimages,"in*Proc.IEEE/CVFICCV*,Seoul,Korea,2019,pp.1–11.
[35] Y. Li *et al.*, "Celeb-DF: A large-scale challenging dataset for deepfakeforensics,"in*Proc.IEEE/CVFCVPR*,Seattle,WA,USA,2020,pp.3207–3216.
[36] B.Dolhansky*etal.*,"Thedeepfakedetectionchallenge(DFDC)dataset,"*arXiv preprint arXiv:2006.07397*, 2020.
[37] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling forconvolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, LongBeach, CA, USA, 2019, pp. 6105–6114.
[38] K.Zhang,Z.Zhang,Z.LiandY.Qiao,"Jointfacedetectionandalignmentusingmu ltitaskcascadedconvolutionalnetworks,"*IEEESignalProcess.Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016.
[39] I.LoshchilovandF.Hutter,"Decoupledweightdecayregularization,"in*Proc. Int. Conf. Learn. Represent.*, New Orleans, LA, USA, 2019.
[40] T. Dao, D. Y. Fu, S. Ermon, A. Rudraand C. Re´, "FlashAttention: Fastand memory-efficient exact attention with IO-awareness," *Adv. NeuralInf. Process. Syst.*, vol. 35, New Orleans, LA, USA, 2022.
[41] Y. Bengio,P.SimardandP.Frasconi,"Learninglong-termdependencieswithgradientdescentisdifficult,"*IEEETrans.NeuralNetw.*,vol.5,no. 2, pp. 157–166, 1994.
[42] A. Katharopoulos, A. Vyas, N. Pappasand F. Fleuret, "Transformers areRNNs: Fast autoregressive transformers with linear attention," in *Proc.Int. Conf. Mach. Learn.*, Virtual, 2020, pp. 5156–5165.
[43] P.Neekhara,B.Dolhansky,J.BittonandC.Canton-Ferrer,"Adversar-ial threats to DeepFake detection: A practical perspective," in *Proc.IEEE/CVF CVPR Workshops*, Virtual, 2021.
[44] S. Adinarayana, D. Ravikiran, U. Sesadri, "Efficient deep fake detection using a hybrid deep learning model in *Proc.IEEE* Virtual,2025