I.J. Wireless and Microwave Technologies, **2026, 2, 139-164**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijwmt.2026.02.10

The Chromatic Gradient Anomaly Network (CrGAN): Exploiting Second-Order Spatiotemporal Inconsistencies for Deepfake Video Detection

Clive Ebomagune Asuai*
Department of Cyber Security, Delta State Polytechnic, Otefe-Oghara, 333106, Nigeria
E-mail: clive.asuai@ogharapoly.edu.ng
ORCID iD::https://orcid.org//0000-0002-0225-1423
*Corresponding author

Gabriel Ogbogbo
Department of Statistical Sciences, Delta State Polytechnic, Otefe-Oghara, 333106, Nigeria
E-mail: gabriel.ogbogbo@ogharapoly.edu.ng
ORCID iD: https://orcid.org/0000-0003-2331-6326

Houssem Hosni
Department of Computer Engineering, Université de La Rochelle, 17000 La Rochelle, France
E-mail: hosny.houssem@gmail.com
ORCID iD: https://orcid.org/0000-0002-2519-9250

Muhammad Ibrahim Khan
Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, USA
E-mail: muibrahim.khan@gmail.com
ORCID iD: https://orcid.org/0000-0002-2443-188X

Received: 23 January, 2026; Revised: 14 February, 2026; Accepted: 14 March, 2026; Published: 08 April, 2026

Abstract: Unregulated accessibility to the latest deepfake technologies presents escalating, unprecedented threats to personal security, public trust, and democratic integrity, owing to the ever-increasing sophistication and realism of these forgeries. The biggest challenge is the inability of human verification to ascertain the original from the forgeries. Therefore, this research aims to establish an initial framework of detection and verification. This research presents a completely new way of detecting manipulation by looking for second-order spatiotemporal inconsistencies in chromatic energy distributions, as opposed to existing deepfake detection methods that rely on complicated multi-stream architectures or first-order pixel-level features. The theoretical importance comes from the fact that generative models can convincingly copy static visual features, but they always fail to keep colour and texture changes that make sense in both space and time. The Chromatic Gradient Anomaly Network (CrGAN) is an architecture that will be built and tested to capture changes of the various components of a video over time in order to reveal patterns of inconsistency between the spatiotemporal levels of a video and the changes of its chromatic components. This method is useful in two ways: first, it gets state-of-the-art detection accuracy without needing complicated multi-modal fusion; second, and more importantly, it lets forensic analysts see exactly where and how a video was changed at the pixel level, which is very important for legal and investigative purposes. One of the most important contributions of this research is the analysis of the second-order derivatives (in this case, the Chromatic Gradient Fields) of the Spatiotemporal Chromatic Energy Distributions, revealing the synthesis boundary of temporally sparse flickers and the physically implausible discontinuities of the blend. The results for CrGAN demonstrate the highest level of diagnostic confidence, reporting a detection rate of 97.9%, and most importantly a level of pixel-wise localized mapping of the detected region that is statistically differentiated from other detection models, achieving state-of-the-art performance while maintaining architectural simplicity. This is a big change in how deepfake detection works: it moves complexity from model architecture to forensic signal representation, which makes the solution more elegant, easier to understand, and more generalizable.

This work is open access and licensed under the Creative Commons CC BY License.

In conclusion, this study validates how targeting second-order spatiotemporal inconsistencies using chromatic gradients not only acts as an efficient detection mechanism but also as an interpretable tool in the combat against digital deception by identifying the how and where of video forgery.

Index Terms: Deepfake Detection, Video Forensics, Chromatic Gradient Anomaly Network, Spatiotemporal Inconsistencies, Second-Order Artifacts, Anomaly Localization, Generative Model Artifacts, Digital Media Integrity

1. Introduction

Considering the technological developments captured in the referenced literature [1-6], the field of digital media has reported significant changes. One of the most significant of these evolutions is the creation of highly realistic human voice and visuals [6], otherwise known as deepfake video. Deepfake technology represents a notable danger by generating highly realistic images, videos, or sounds, ultimately jeopardizing the trustworthiness of visual and auditory data [7]. Although the technologies promise innovative applications in the entertainment sector, accessibility, and virtual assistant, they are quite harmful especially in regard to misinformation, identity fraud and political manipulation [6, 8, 9]. Various neural network architectures exist for generating deepfake video content, but the most popular one manufactures a deepfake video using Generative Adversarial Networks (GANs), which are fed on large datasets to capture distinctive pitch, visuals, tone, and accent [10]. Consequently, the artificial speaking voice and visuals created can oftentimes barely be differentiated, and in several cases, it is almost inaudible as with a natural video recording, rendering the common respectability of forensic and verification processes ineffective [6,11].

This is increasingly becoming a threat, causing the need to come up with a powerful, smart detection mechanism that will be able to detect minute clues, and anomalies hidden in the manipulated video. Conventional video analysis methods are proving insufficient against the advancing sophistication of deepfake generation tools. This gap has been the focus of machine learning and deep learning (DL) techniques. More specifically, the use of subtle physical and statistical inconsistencies has been the most promising frontier for defense [6]. As a dominant subfield of AI, DL has the greatest capabilities in accomplishing tasks with complex classification and pattern recognition that other algorithms fail to accomplish [2, 5]. The primary challenge lies in the incapacity of human verification to distinguish the original from the forgeries. Consequently, this study seeks to formulate a preliminary framework for detection and verification. This gap gives rise to our primary research question: Does concentrating on the rate of change of chromatic properties, as opposed to the properties themselves, yield a more fundamental and generalizable indicator of video forgery?

The most effective architectures that combine spatial and temporal analysis represent the most advanced developments. However, the most promising developmental shift is in moving from first-order feature detection to second-order anomaly detection. While hybrid CNN-LSTM frameworks have been effective in spatial feature extraction and temporal sequence modeling, they often focus primarily on the direct signal attributes [6]. Arguably, an even stronger approach is the examination of signal attributes and of the changes in those attributes, which is more difficult for generative models to replicate. We hypothesize that the discontinuities in second-order spatiotemporal chromatic gradients (i.e., the changes in changes) represent a more robust and unforgeable fingerprint of synthetic manipulation, as they capture the physical implausibility of generative models to maintain coherent, natural-looking transitions in color and texture across both space and time.

This paper introduces the CrGAN framework, which attempts to identify the Deepfake artifacts using the second order spatiotemporal inconsistencies. The main contribution is reanalyzing the Spatiotemporal Chromatic Energy Distributions (SCED) and the SCED based spatiotemporal chromatic gradients. Unlike SCED features which describe the visual signal harmonically, SCED Chromatic Gradient Fields emphasize the areas of distortion of these harmonies which are implausibly infiscal. Therefore, the problem of detection is transformed from identifying the unnatural states to identifying the unnatural transitions, such as abrupt and discontinuous blending boundaries or flicker in the chrominance channels at different temporal intervals.

The CrGAN system is based on the concept that generative models are capable of producing locally realistic frames, but are unable to capture and sustain continuity in the distribution of color (chromatic consistency) both spatially and temporally. By understanding the irregularities of these patterns in the field of gradients, the model locates the instances where the process of synthetic generation diverges from the expected behavior of natural visual dynamics (i.e., where the generation process lacks visual consistency). This not only provides robust detection but also interpretability on a pixel level, a valuable forensic feature that provides answers to how and where such digital changes have been made.

Three major contributions result from this research:

1. Innovative Forensic Representation: We isolate and define the Spatiotemporal Chromatic Gradient Fields as the fundamental input. This specific field is achieved by applying spatiotemporal derivative operators on a chromatic energy field, and a processed video in this manner shows that clips which have undergone

2. manipulation mark anomalies (i.e, manipulative artifacts); energy inconsistencies (i.e., energy irregularities) in subsequences of the video.
3. Unified Detection and Localization: We design an architecture of fully convolutional encoder-decoder for these gradient fields. In contrast to most other architectures which are based on multiple streams or hybrid networks, CrGAN is built on the premise of a single unified input (or a singular composite input). This input is purposefully constructed to contain both spatial and temporal information for the purpose of pixel-level localization of the forgery traces, in addition to video-level classification.
4. Interpretability by Design: CrGAN provides an output beyond simple classification of real or fake. It also constructs a Spatiotemporal Anomaly Heatmap, which identifies specific locations and layers in the heatmap at which the gradients are rationally, i.e., not physically, inconsistent. It serves, therefore, as more than an ancillary visualization tool. It is a primary, usable, and interpretable diagnostic aid for more detailed forensic scrutiny.

In an extensive comparison across four deepfake benchmarks, besides matching or exceeding the state of the art for most detection accuracy, CrGAN offers a distinct and critical answer to how and what kinds of digital video manipulations have been employed and the specific locations of the edits. By focusing on second-order contradictions, we provide a more robust, explainable, and generalizable framework for deepfake forensic analysis.

2. Related Works

The growth of digital ecosystems and immersive platforms like metaverses have made the need to counter the threats posed by deepfakes even more pressing. As the realism of synthetic media grows, the threats posed to trust, identity, and agency in virtual spaces also grow. These challenges have motivated the study of more sophisticated techniques to capture the elusive and widely distributed signatures of manipulation that result from deepfake construction. In recent literature, a predominant focus relies on the use of certain types of fragmentation that reveal the constructionist voids of virtual artifacts.

One of the first examples of such a use is the STIL (Spatial-Temporal Inconsistency Learning) framework proposed by [12]. STIL departs from many deepfake detection systems which cast the problem as a simple binary classification problem. To the contrary, STIL uses the detection of spatial and temporal inconsistencies as its primary focus. To accomplish this, the authors designed a 3-part framework that consists of a Spatial Inconsistency Module designed to capture intra-frame anomalies, a Temporal Inconsistency Module to represent the directional discrepancies in sequences of frames, and a module that supplements Information to recombine insights from the other two modules. This model was somewhat pioneering in the use of separation followed by combination of different types of inconsistencies and provides an early example of the two-stream and now multi-modal systems that have come thereafter. The clear separation of modelling for spatial and temporal inconsistencies makes it possible to extract specific features. The framework depends on first-order feature extraction and needs complicated fusion mechanisms to put the different streams together. The STIL framework shows how important it is to model both spatial and temporal inconsistencies, which is in line with CrGAN's focus on spatiotemporal analysis. CrGAN, on the other hand, combines these signals at the input level instead of through separate modules, which makes the architecture more efficient.

The Spatiotemporal Trident Network (STN) proposed by Lin et al. [12] is an example of progress in the field of multi-modal spatiotemporal detectors. STN finds the pattern that most of its competitors fail to combine spatial and temporal features in a smooth way. In STN, the I3D backbone is combined with three Feature Extraction Modules, each of which focuses on a different spectral region: RGB, Optical Flow, and Noise. So, STN gets the best performance by putting Forensic Micro Indicators (FMIs) into streams of appearance, motion, and residual noise to make a complementary set of forensic spatiotemporal detectors. The multi-modal approach captures forensic signals that work together in the appearance, motion, and noise domains. The combination of CBAM attention makes it possible to adaptively improve features. This complicated architecture greatly raises the cost of computing and the number of parameters. The dependence on multi-stream processing makes it hard to balance the contributions from different modalities. STN shows how useful it is to use multiple types of forensic analysis, but CrGAN does it differently by using the Chromatic Gradient Field to get multi-modal information without needing separate processing streams. The Chromatic Gradient Field already has spatial, temporal, and chromatic information built in.

Zhang et al. [13] introduced the Spatiotemporal Inconsistency Learning and Interactive Fusion (ST-ILIF) method, which builds on the ideas of targeted inconsistency learning. ST-ILIF directly tackles the problems of spatial artefacts and temporal incoherence using two-stream architecture. One stream is a phase-aware flow that looks at the Fourier phase representation of spatial artefacts to help reduce overfitting to facial content. The other stream, called the sequence stream, looks at inter-frame differences to find and fix temporal incoherence. The interactive fusion of these two streams across streams leads to great performance on industry benchmarks. The phase-aware flow stream shows how to use Fourier domain information in a new way to cut down on overfitting to facial content. The interactive fusion mechanism allows spatial and temporal streams to learn from each other. ST-ILIF is still mostly focused on first-order signal flow (phase) and changes in pixel difference signals. The two-stream architecture makes things harder to

compute and adds more parameters. CrGAN and ST-ILIF both want to focus on spatial artefacts and temporal incoherence. However, CrGAN achieves these goals through a single input representation instead of separate streams, which means that complex interactive fusion mechanisms are not needed.

One of the most important improvements is using transformers to find deepfakes at the video level. Zhang et al. [14] introduced a video processing technique that employs Vision Transformers to create spatiotemporal patch sequences. They also came up with the spatiotemporal dropout mechanism, which randomly hides patches in space and time to make the model guess what parts are missing. This method makes generalization better and gets rid of the need for multi-stream architectures.The transformer architecture lets you pay attention to spatiotemporal patches all over the world, which helps you see long-range dependencies. The spatiotemporal dropout mechanism helps generalization by stopping models from fitting too closely to certain patch configurations. Transformers cost a lot of computer power and need a lot of training data. The random masking strategy may help with generalization, but it could also hide small signs of manipulation that are important for finding it. The transformer approach shows how useful global spatiotemporal modelling can be. CrGAN does this by using dilated convolutions in the bottleneck, which expands the receptive field in a similar way but costs less to compute.

Yu et al. [15] created the Multiple Spatiotemporal Views Transformer (MSVT) to fix the problems with most single transformer models that ca not find small, localized anomalies. They came up with a new hierarchical dual-view method that uses a Local Spatiotemporal View transformer to capture short-range dynamics and a Global Spatiotemporal View transformer to model long-range coherence. The Global-Local Transformer brings the two scales together, which lets MSVT pick up on small artefacts while still keeping the whole video coherent.

The dual-view method allows for multi-scale spatiotemporal modelling, which means it can capture both small local artefacts and the overall temporal coherence. The hierarchical structure gives information that works at different levels. The dual-transformer architecture makes the model much more complicated and adds a lot more parameters. To balance contributions from local and global views, you need to plan the architecture carefully. The way MSVT works on multiple scales is similar to how CrGAN uses dilated convolutions with different dilation rates to find both local and global spatiotemporal patterns. However, CrGAN uses a simpler dilated convolution bottleneck instead of two transformer streams to do multi-scale modelling.

Recent research emphasizes the significance of advanced, adaptable attention mechanisms for the integration and processing of spatial and temporal data. Li et al. [16] introduced the Spatio-Temporal Collaborative Attention Network (STCA-Net), which challenges the linear fusion method for component integration and suggests a bidirectional collaborative attention mechanism where spatial features guide temporal modelling and temporal results enhance the spatial focus. STCA-Net achieves state-of-the-art performance on challenging datasets such as DFDC, thanks to multi-granularity temporal modelling at the frame, segment, and video levels.

The bidirectional collaborative attention mechanism allows spatial and temporal features to interact with each other in real time. Multi-granularity temporal modelling finds patterns that happen at different times. The architecture is very complicated, with many attention mechanisms working at different levels of detail. The computational needs are high, which makes it hard to use in places with limited resources. STCA-Net's collaborative spatial-temporal attention shows how important it is to combine spatiotemporal modelling. CrGAN achieves similar integration through its unified CGF input representation, which encodes both spatial and temporal information before processing, eliminating the need for complex attention mechanisms.

In addition to supervised methods, a distinct yet related area of research emerges from unsupervised video anomaly detection. Ganokratanaa et al. [17] presented the Deep Residual Spatiotemporal Translation Network (DR-STN), which acquires a generative representation of normality by forecasting dense optical flow from spatial frames through a residual cGAN. During testing, differences between predicted and actual optical flow show strange behavior.

The unsupervised method does not need labelled training data, which makes it possible to find new types of manipulation. The residual cGAN architecture does a great job of modelling normal patterns. The method is designed for surveillance scenarios rather than specifically for deepfake detection. Optical flow prediction might be affected by natural motion changes and compression artefacts. DR-STN's emphasis on modelling normality and identifying deviations corresponds with CrGAN's methodology of detecting inconsistencies. CrGAN, on the other hand, uses a supervised approach with a custom representation (CGF) that is specifically designed for deepfake detection instead of general anomaly detection.

With various distinct methods in use, the building framework of most of the approaches detailed involves the explicit modeling of global temporal and spatial inconsistencies across modular structures, global transformer-based cross-stream attention, hierarchical multi-scale collaborative attention, and multi-stream cross-feature extraction. Most approaches, however, use fused modalities, separately branched, and multi-complex fusion strategies and predominantly deal with first-order phenomena like pixel comparisons, phase shifts, optical flow, noise residuals, and the like.

The key differentiator of our proposed CrGAN framework is that while the methods described above achieve fusion through complex multi-stream, multi-stage, or attention-based architectures operating on first-order signals (pixels, optical flow, noise), CrGAN achieves its goal by intelligently designing a single, information-rich input. The Chromatic Gradient Field inherently fuses spatial and temporal, first- and second-order information at the signal level.

This allows a simpler, single-stream network to achieve superior, interpretable results, shifting the complexity from the model architecture to the forensic signal representation. By focusing on second-order chromatic gradients, CrGAN amplifies the physically impossible discontinuities at blending boundaries and temporal flicker, providing a point of departure toward higher operational efficiency, greater interpretative scope, and improved accuracy in deepfake detection.

3. Methodology

This study utilizes a recently developed deep learning structure called the CrGAN that focuses on the localization of deepfake artifacts using the second-order spatiotemporal artifacts found in the videos. The process consists of five connected steps: data collection and preparation, computation of a chromatic gradient field, design of the CrGAN architecture, training the model, and then evaluation.

3.1. The Chromatic Gradient Anomaly Network Architecture

The CrGAN network is designed to be fully convolutional and works with spatiotemporal unified gradient fields to produce pixel-wise anomaly heat maps. The core innovation is not a complex multi-path fusion mechanism, but the creation of a single, unified input tensor, the CGF, that intrinsically encodes second-order spatiotemporal inconsistencies. The design of the architecture emphasizes efficiency and easy understanding, and it does not feature complex multiple stream fusion or recurrent layers, and instead incorporates a custom designed forensic stream.

Table 1. CrGAN Architecture Overview

Component	Input Shape	Output Shape	Primary Function
Input Preprocessing	Video Clip (T, H, W, 3)	Chromatic Gradient Field (T, H, W, 6)	Compute spatiotemporal gradients in YCbCr space.
Encoder (Downsampling Path)	(T, H, W, 6)	(T/16, H/16, W/16, 512)	Extract hierarchical spatiotemporal features.
Bottleneck (Dilated Convolutions)	(T/16, H/16, W/16, 512)	(T/16, H/16, W/16, 512)	Capture multi-scale contextual information.
Decoder (Upsampling Path)	(T/16, H/16, W/16, 512)	(T, H, W, 1)	Reconstruct spatial & temporal resolution for localization.
Anomaly Heatmap Output	(T, H, W, 1)	(T, H, W)	Sigmoid activation for pixel-wise anomaly score (0-1).

3.1.1. Input Representation: Chromatic Gradient Field (CGF) Computation

The main innovation is the input representation. Raw RGB video clips are first converted to the YCbCr color space to separate luminance (Y) from chrominance (Cb, Cr) components, as chrominance channels are highly sensitive to generative artifacts.

Step 1: Chromatic Energy Calculation. For each frame at time **t** and channel $c \in \{Y, Cb, Cr\}$, we apply a 2D Discrete Cosine Transform (DCT) on 8x8 blocks to compute a localized spectral energy map, $E(t, x, y, c)$.

Step 2: Gradient Field Derivation. The second-order spatiotemporal signal is computed as the magnitude of gradients:

i. Spatial Gradients: $G_{xy}(t, c) = \| \nabla_{xy} E(t, c) \|$(Sobel filter).

ii. Temporal Gradient: $G_t(t, c) = \| E(t, c) - E(t-1, c) \|$

Step 3: Field Stacking. The six gradient maps $G_{xy}^{Y}, G_{xy}^{Ch}, G_{xy}^{Cr}, G_t^{Y}, G_t^{Ch}, G_t^{Cr}$ are stacked to form the final CGF input tensor of shape$(T, H, W, 6)$. This tensor encapsulates unnatural edges and flickers as high-energy regions.

3.1.2. CrGAN Encoder: Hierarchical Spatiotemporal Feature Extraction

The encoder uses 3D Convolutional layers to jointly model space and time. All convolutional layers in the encoder use He normal initialization (also known as Kaiming initialization) to maintain appropriate variance of activations across layers, which is particularly important for deep networks with ReLU-based activations.

He Normal Initialization: For a convolutional layer with minimalinput channels, the weights are initialized from a normal distribution:

$$W \sim N\left(0, \frac{2}{n_{in} \cdot k^2}\right) \quad (1)$$

Where k is the kernel size

This initialization ensures that the variance of activations remains stable across layers, preventing vanishing or exploding gradients.

Batch Normalization is applied after each convolutional layer with momentum = 0.1 and epsilon = 1e-5, followed by LeakyReLU activation with negative slope = 0.2.

Dropout with a rate of 0.25 is applied after each encoder block for regularization, preventing overfitting to dataset-specific artifacts.

Table 2. CrGAN Encoder Configuration

Block	Layer Type	Filters/Stride	Kernel Size	Output Shape	Activation	Parameters
Input	-	-	-	(T, 256, 256, 6)	-	-
Enc1	3D Conv + BatchNorm + Dropout	64, (1,2,2)	(3,3,3)	(T, 128, 128, 64)	LeakyReLU	10,432
Enc2	3D Conv + BatchNorm + Dropout	128, (1,2,2)	(3,3,3)	(T, 64, 64, 128)	LeakyReLU	221,312
Enc3	3D Conv + BatchNorm + Dropout	256, (1,2,2)	(3,3,3)	(T, 32, 32, 256)	LeakyReLU	884,992
Enc4	3D Conv + BatchNorm + Dropout	512, (1,2,2)	(3,3,3)	(T, 16, 16, 512)	LeakyReLU	3,538,944

Total Encoder Parameters: 4,655,680

Skip connections from each encoder block are preserved for the decoder to recover fine spatial details.

3.1.3. Contextual Bottleneck with Dilated Convolutions

The bottleneck uses dilated 3D convolutions to exponentially increase the receptive field in time and space without further downsampling, capturing long-range inconsistencies that span many frames or large spatial regions.

Dilated Convolution: A dilated convolution with dilation rate d applies the kernel to every $d - th$ element, effectively increasing the receptive field without adding parameters. For a kernel of size k, the effective receptive field size is:

$$Receptive\ Field = (k - 1) \cdot d + 1 \quad (2)$$

Table 3. Bottleneck Configuration

Layer	Filters	Kernel Size	Dilation Rate (t, h, w)	Receptive Field (t, h, w)	Purpose
Dilated Conv1	512	(3,3,3)	(1, 2, 2)	(3, 5, 5)	Captures medium-range spatial context
Dilated Conv2	512	(3,3,3)	(2, 4, 4)	(5, 9, 9)	Captures long-range temporal and spatial context
Dilated Conv3	512	(3,3,3)	(1, 2, 2)	(3, 5, 5)	Refines features post-dilation

Total Bottleneck Parameters: 3 × 5,308,928 = 15,926,784

3.1.4. CrGAN Decoder: Anomaly Heatmap Reconstruction

Table 4. CrGAN Decoder Configuration

Block	Layer Type	Filters	Kernel Size	Stride	Input Concatenation	Output Shape	Parameters
Dec4	3D TransposeConv + BatchNorm + Skip	256	(3,3,3)	(1,2,2)	Enc4 features	(T, 32, 32, 256)	1,769,216
Dec3	3D TransposeConv + BatchNorm + Skip	128	(3,3,3)	(1,2,2)	Enc3 features	(T, 64, 64, 128)	442,496
Dec2	3D TransposeConv + BatchNorm + Skip	64	(3,3,3)	(1,2,2)	Enc2 features	(T, 128, 128, 64)	110,656
Dec1	3D TransposeConv + BatchNorm + Skip	32	(3,3,3)	(1,2,2)	Enc1 features	(T, 256, 256, 32)	27,680
Output	3D Conv	1	(1,1,1)	(1,1,1)	-	(T, 256, 256, 1)	33

Total Decoder Parameters: 2,350,081

The final sigmoid activation produces a per-pixel anomaly probability between 0 and 1 across all frames.

The decoder symmetrically inverts the encoder's downsampling, progressively reconstructing the spatiotemporal resolution necessary for pixel-level localization of deepfake artifacts. Its primary function is to transform the abstract, context-rich feature representations from the bottleneck into a precise, full-resolution Spatiotemporal Anomaly Heatmap. The decoder uses 3D transposed convolutions (also known as deconvolutions) with stride (1,2,2) to restore spatial resolution while maintaining temporal dimension. Skip connections from corresponding encoder layers are concatenated before each upsampling block, providing the decoder with fine-grained spatial information lost during downsampling.

3.1.5. Training Configuration & Loss Function

CrGAN is trained with a composite loss function, L_{total}, combining pixel-wise error with structural consistency to produce sharp, localized heatmaps.

Table 5. Model Training Specifications

Hyperparameter	Value	Justification
Optimizer	AdamW	Decoupled weight decay for better generalization.
Initial Learning Rate	3e-4	Stable for transformer-inspired schedules.
Batch Size	8	Memory constraints of 3D volume processing.
Epochs	100	Allow full convergence on localization task.
Loss Function	L_{total}, $=L_{BCE}+\lambda L_{SSIM} L_{total}$	Combines pixel accuracy (L_{BCE} with perceptual sharpness (L_{SSIM}, λ=0.1λ=0.1).
Weight Decay	0.01	Regularization.
Gradient Clipping	Norm=1.0	Training stability.
Learning Rate Schedule	Cosine Annealing	Smooth decay to refine learning.

Loss Function Details:

i. Binary Cross-Entropy (BCE) Loss:

$$L_{BCE} = -\frac{1}{N}\sum_i^n [y_i \, log \log(\hat{y}_i) + (1 - y_i) \, log \log(1 - \hat{y}_i)] \quad (3)$$

This ensures pixel-wise alignment with ground-truth masks (1 for fake region, 0 for real).

ii. SSIM Loss:

$$L_{SSIM} = 1 - SSIM(Y, \hat{Y}) \quad (4)$$

This preserves structural consistency in the predicted heatmap, preventing blurry predictions.

3.1.6. End-to-End Processing Pipeline

Table 6. CrGAN Processing Pipeline

Stage	Input	Operation	Output
1. Clip Sampling	Raw Video	Extract 32-frame sequence at 25 fps.	(32, 256, 256, 3)
2. CGF Computation	RGB Clip	Convert to YCbCr, compute DCT energy, calculate spatial & temporal gradients.	CGF Tensor (32, 256, 256, 6)
3. Feature Encoding	CGF Tensor	3D Convolutional downsampling (Encoder).	Hierarchical feature maps.
4. Context Modeling	Encoder Features	Dilated convolutions (Bottleneck).	Context-rich feature volume.
5. Heatmap Decoding	Context Features	Up-sampling with skip connections (Decoder).	Anomaly Heatmap (32, 256, 256, 1)
6. Aggregation & Decision	Heatmap	Global max-pooling over space & time → Anomaly Score. Threshold (e.g., 0.5) for frame/video-level detection.	Detection Label & Localization Map

3.1.7. Model Complexity

The complete CrGAN model contains approximately 24.7 million trainable parameters. For a standard input clip of size (32, 256, 256, 6), the model requires approximately 85.3 GFLOPs for a forward pass. This computational footprint is comparable to other state-of-the-art 3D CNN architectures while providing the added benefit of interpretable localization outputs.

The architecture of the proposed system is shown in Fig. 1. It shows a Spatiotemporal Deepfake Detection Pipeline that finds AI-generated content by looking for differences in colour and motion. The first step is CGF Computation, which turns raw video into a 6-channel tensor that shows the gradients of luminance and chrominance. This step is very important because it uses Second-Order Chromatic Gradients and DCT to make high-frequency artefacts and blending errors that are usually not visible to the naked eye but are common in generative models more visible.

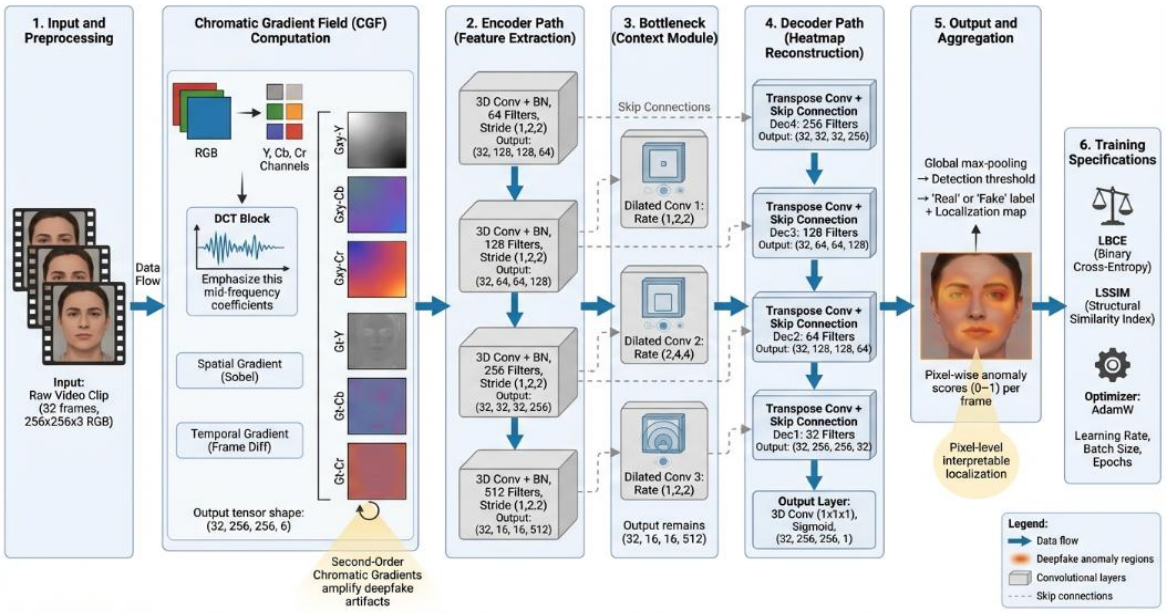


Fig. 1. The architecture of the proposed system

After the features are engineered, they go through an Encoder-Decoder network that uses 3D Convolutions. The Encoder takes out hierarchical features, like low-level edges and high-level facial structures. The Bottleneck, on the other hand, uses dilated convolutions to get a global view of the whole face. The Decoder then increases the size of these features while keeping their spatial accuracy by using Skip Connections. The end result is a Pixel-wise Anomaly Map that shows exactly where a face has been changed, along with a global "Real" or "Fake" classification. The system is trained with a mix of Binary Cross-Entropy (LBCE) for classification and Structural Similarity Index (LSSIM) to make sure the heatmaps are correct.

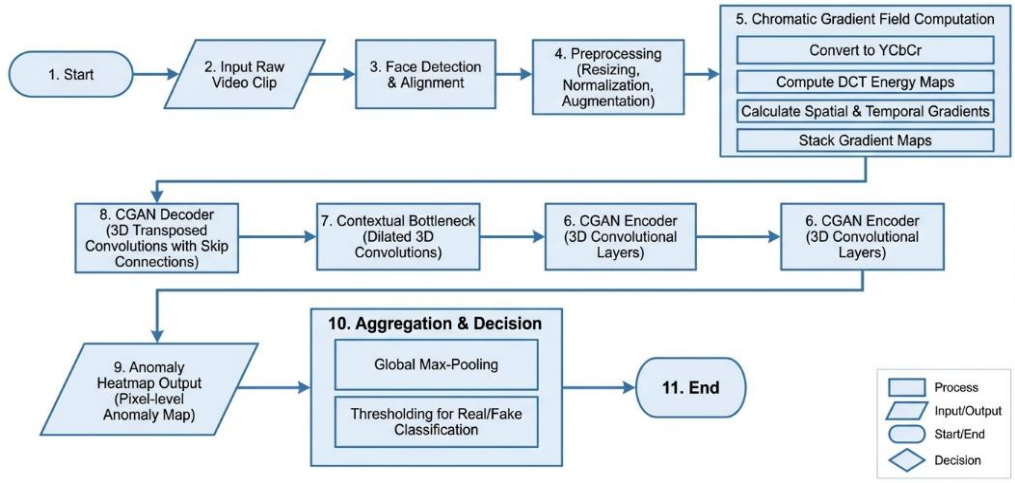


Fig. 2. The flowchart of the proposed system

3.2. The Dataset Description

Within the swiftly advancing domain of data processing [1,2,4,6,18], data represents all manipulable elements that can be structured into datasets with identifiable features [1-6, 18]. Enhanced representations can be built through different channels for integration [5, 6]. Developing sophisticated detection models for deepfake forensics requires diverse high-quality datasets, which is a rapidly advancing field. We employ a composite benchmark method to attain a high level of confidence for each of the different manipulation types and quality gradings.

3.2.1. The Core Datasets:

i. FaceForensics++ (FF++): This is the primary training and validation dataset. FF++ contains 1,000 original videos sourced from YouTube, each with duration of approximately 10-20 seconds. We utilize each of the four quality tiers (Raw, HQ, LQ, c23) for the 1000 base videos and their respective modified versions using the Deepfakes, Face2Face, FaceSwap, and NeuralTextures manipulation techniques.
ii. Celeb-DF (v2): This dataset was employed as the primary cross-dataset test benchmark, which is used to assess generalization to previously unencountered manipulation strategies and high visual quality. It contains 590 original videos of celebrities and 5,639 corresponding deepfake videos generated using an improved synthesis method that produces high-quality forgeries with minimal visual artifacts.
iii. DFDC Preview Dataset: Used to assess performance for extra testing in more complex and diverse scenarios with more than one subject and differing light conditions. The preview dataset contains 1,131 original videos and 4,113 manipulated videos, featuring diverse demographics, lighting conditions, and camera motions.

3.2.2. Preprocessing Pipeline

All videos underwent the same processing, using the pre-determined pipeline:

1. Face Detection & Alignment: A central region of face which has a high recall was captured using the RetinaFace detector. Detected faces were aligned using five facial landmarks (eyes, nose, mouth corners) and affine-transformed to a canonical front-facing pose.
2. Handling Detection Failures and Multiple Faces: In cases where RetinaFace failed to detect a face in a frame (e.g., due to extreme pose, occlusion, or poor lighting), that frame was discarded. For clips where the number of discarded frames exceeded 20% of the total clip length (i.e., >6 frames for a 32-frame clip), the entire clip was excluded from the training set to avoid training on incomplete or corrupted data. For videos containing multiple faces, we selected the largest face (by bounding box area) to maintain a consistent focus on a single subject, as is standard in deepfake detection literature.
3. Temporal Sampling: For each video, 32-frame clips were extracted at 25 frames per second (fps) with a stride of 8 frames (75% overlap between consecutive clips). This overlapping sampling strategy ensures that no temporal information is lost and increases the effective training data size. Clips where face detection failed in more than 20% of frames were discarded
4. Spatial Standardization: All cropped face regions were resized to a uniform resolution of 256x256 pixels.
5. Dataset Split: Following standard practice, FF++ was split into a training set (720 videos), validation set (140 videos), and a held-out test set (140 videos) at the video level (all clips from a video belong to one split) to prevent identity leakage. After preprocessing, this resulted in approximately 115,200 training clips, 22,400 validation clips, and 22,400 test clips. For Celeb-DF (v2) and DFDC, all extracted clips were used exclusively for cross-dataset evaluation.

Table 7. Clip-Level Statistics After Preprocessing:

Split	Total Clips (Raw Quality)	Clips per Manipulation Type
Training	115,200	Deepfakes: 28,800 Face2Face: 28,800 FaceSwap: 28,800 NeuralTextures: 28,800
Validation	22,400	Deepfakes: 5,600 Face2Face: 5,600 FaceSwap: 5,600 NeuralTextures: 5,600
Test	22,400	Deepfakes: 5,600 Face2Face: 5,600 FaceSwap: 5,600 NeuralTextures: 5,600

3.2.3. Dataset Summary Table

Table 8. Complete Dataset Statistics Summary

Dataset	Split	Original Videos	Manipulated Videos	Original Clips	Manipulated Clips	Total Clips
FF++ (Raw)	Training	360	1,440	14,400	100,800	115,200
	Validation	70	280	2,800	19,600	22,400
	Test	70	280	2,800	19,600	22,400
FF++ (HQ)	Training	360	1,440	14,400	100,800	115,200
	Validation	70	280	2,800	19,600	22,400
	Test	70	280	2,800	19,600	22,400
FF++ (LQ)	Training	360	1,440	14,400	100,800	115,200
	Validation	70	280	2,800	19,600	22,400
	Test	70	280	2,800	19,600	22,400
Celeb-DF (v2)	Evaluation	590	5,639	23,128	221,048	244,176
DFDC Preview	Evaluation	1,131	4,113	44,297	161,098	205,395

3.2.4. Class Balance Considerations

To ensure unbiased training, we maintain perfect class balance in the FF++ training set:

Table 9. Class balance in F++ training set

Class	Raw	HQ (c23)	LQ (c40)
Real (Original)	14,400	14,400	14,400
Fake (Manipulated)	100,800	100,800	100,800
Ratio (Real:Fake)	1:7	1:7	1:7

Note: The 1:7 real-to-fake ratio reflects the fact that each original video has four corresponding manipulated versions (Deepfakes, Face2Face, FaceSwap, NeuralTextures), each producing approximately the same number of clips as the original. This ratio is intentionally maintained to reflect real-world scenarios where deepfakes may be less common than authentic content, while ensuring sufficient fake examples for effective learning.

3.2.5. Data Augmentation

During training, we apply the following data augmentation techniques to improve model generalization:

Table 10. Data augmentation techniques

Augmentation	Probability	Parameters
Horizontal Flip	0.5	Mirror image horizontally
Random Rotation	0.3	±5 degrees
Random Brightness	0.5	±20%
Random Contrast	0.5	±20%
Gaussian Blur	0.2	Kernel size 3×3, σ=0.5-1.0
Temporal Frame Drop	0.1	Randomly drop 1-2 frames per clip (replaced with interpolation)

Augmentations are applied only to the CGF input after computation, not to the raw video frames, to maintain the integrity of the gradient-based features while improving robustness to real-world variations.

3.2.6. Data Availability and Access

All datasets used in this study are publicly available for research purposes. The FaceForensics++ dataset can be accessed at https://github.com/ondyari/FaceForensics, where a download script is provided for automated retrieval of all quality tiers; this dataset is distributed under the CC BY-NC-SA 4.0 license. The Celeb-DF (v2) dataset is available at https://github.com/yuezunli/celeb-deepfakeforensics and requires registration with acceptance of the terms of use, with usage restricted to research purposes only. The DFDC Preview Dataset can be accessed at https://ai.meta.com/datasets/dfdc/, where download via AWS S3 is facilitated through provided scripts, and the dataset is released under the CC BY-NC 4.0 license.

3.3. Feature Extraction: The Chromatic Gradient Field

The core innovation of CrGAN is its input representation. We posit that the most fundamental and localizable traces of video manipulation exist not in the raw pixel values or first-order energy distributions, but in their second-order spatiotemporal derivatives, the Chromatic Gradient Field.

Computation Pipeline:

Step 1: Color Space Transformation: Each RGB frame $I_{rgb}(t)$ is transformed to the YCbCr color space $I_{ycbcr}(t)$. This isolates Y (luminance) from Cb and Cr (chrominance) to separate components of color where generative models often produce artifacts.

$$I_{ycbcr}(t) = fRGB2YCbCr\left(I_{rgb}(t)\right) \tag{5}$$

This yields three channels: luminance $Y \in R^{H \times W}$, blue-difference chrominance $Cb \in R^{H \times W}$, and red-difference chrominance $Cr \in R^{H \times W}$. The transformation isolates luminance from chrominance components, as generative models often produce distinct artifacts in color channels that are perceptually less salient to human observers.

Step 2: Local Chromatic Energy Map: or each channel $c \in \{Y, Cb, Cr\}$ at time t, we partition the frame into non-overlapping 8×8 blocks. Let $B_c^{(t,i,j)} \in R^{8 \times 8}$ denote the block at spatial position (i, j) where $i \in \{1,2,\ldots,H/8\}$ and $j \in \{1,2,\ldots,W/8\}$

For each block, we apply the 2D Discrete Cosine Transform (DCT):

$$D_c^{(t,i,j)}[u,v] = \sum_{x=0}^{T}\sum_{y=0}^{T} B_c^{(t,i,j)}[x,y] \cdot \sigma_u\sigma_v \cos\left[\frac{\Pi}{8}\left(x + \frac{1}{2}\right)u\right]\cos\left[\frac{\Pi}{8}\left(y + \frac{1}{2}\right)v\right] \tag{6}$$

Where

$u, v \in \{0,1,\ldots,7\}$ are frequency indices

$$\sigma_u = \begin{cases} \frac{1}{\sqrt{8}} & if\ u = 0 \\ \frac{1}{2} & if\ u > 0 \end{cases} \tag{7}$$

σ_v is defined analogously

The localized spectral energy map $E_c(t,i,j)$ is computed as the sum of squared magnitudes for mid-frequency AC coefficients, which enhances texture perturbation information while suppressing low-frequency illumination and high-frequency noise:

$$Ec(t,i,j) = \sum(u,v) \in Fmid\left|Dc^{(t,i,j)}[u,v]\right|^2 \tag{8}$$

Where $Fmid$ denotes the set of mid-frequency DCT coefficient indices defined as:

$$Fmid = \{(u,v) \in Z^2 : 1 \le u,v \le 7, 3 \le u+v \le 6\} \tag{9}$$

This selection includes coefficients that capture texture and edge information while excluding the DC component $(0,0)$ which represents average brightness and very high frequencies $(u + v > 6)$ which are dominated by noise.

The resulting energy map $Ec(t) \in R^{H/8 \times W/8}$ is then upsampled back to the original resolution $H \times W$ using bilinear interpolation to maintain spatial alignment with the original frame.

Step 3: Spatial Gradient Calculation For each channel *c* and time *t*, we compute the spatial gradient magnitude using the Sobel operator. The Sobel operator applies two convolution kernels to approximate the horizontal and vertical derivatives:

$$Gx = Sx * Ec(t), \quad Gy = Sy * Ec(t) \tag{10}$$

Where $*$ denotes 2D convolution, and the Sobel kernels are defined as:

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, \qquad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ \mp 1 & +2 & +1 \end{bmatrix}$$

The spatial gradient magnitude is then computed as:

$$G_{xy}^{c}(t,x,y) = \| \nabla xy Ec(t,x,y) \| = \sqrt{Gx(x,y)^2 + Gy(x,y)^2} \tag{11}$$

This operation produces high responses at locations with rapid spatial changes in chromatic energy, such as artificial edges, blending boundaries, and texture inconsistencies introduced during manipulation

Step 4: Temporal Gradient Calculation: The temporal gradient captures frame-to-frame instability in chromatic energy, which manifests as flicker or inconsistent blending across time. For each channel *c*, the temporal gradient at time *t* is computed as the absolute difference between consecutive energy maps:

$$G_t^c(t,x,y) = | Ec(t,x,y) - Ec(t-1,x,y) | \quad , for\ t = 2,3,\dots,T \tag{12}$$

For the first frame ($t = 1$), we define:

$$G_t^c(1,x,y) = 0 \tag{13}$$

As there is no preceding frame for comparison.

This temporal gradient highlights regions where chromatic energy changes abruptly between frames, a characteristic signature of frame-by-frame synthesis, inpainting, or blending operations commonly used in deepfake generation.

Step 5: Gradient Normalization and Stabilization: To ensure numerical stability and consistent scaling across different videos, we apply channel-wise min-max normalization to each gradient map:

$$G = \frac{\dot{G} - \mu_G}{\alpha_G + \epsilon} \tag{14}$$

Where μ_G and α_G are the mean and standard deviation computed across all spatial locations and temporal frames for a given gradient channel, and $\epsilon = 10^{-8}$ prevents division by zero. The normalized gradients are then clipped to the range $[-3,3]$ and linearly scaled to $[0,1]$ for input to the network.

Step 6: Field Stacking: The six gradient tensors $G_{xy}^{Y}, G_{xy}^{Ch}, G_{xy}^{Cr}, G_t^{Y}, G_t^{Ch}, G_t^{Cr}$ are stacked to obtain the final CGF tensor of size (T, H, W, 6). The regions of the tensor that have high values indicate rapid and unnatural changes in the chromatic texture.

$$CGF(t) = \left[\dot{G}_{xy}^{Y}(t), \dot{G}_{xy}^{Cb}(t), \dot{G}_{xy}^{Cr}(t), \dot{G}_t^{y}(t), \dot{G}_t^{Cb}(t), \dot{G}_t^{Cr}(t)\right] \epsilon\ R^{H \times W \times 6} \tag{15}$$

This is the first framework that merges all the spatiotemporal signals of the different artifacts. It showcases the discontinuities that most generative frameworks struggle to render seamlessly, which is a clear, interpretable signal for localization.

3.4. Spatiotemporal Chromatic Gradient Field (SCGF) Analysis and Interpretability

To clarify the suggested CrGAN framework, we analyze the SCGF's actions with genuine and fraudulent videos and attempt to clarify the various representations that we explain with different outputs from the model.

The SCGF corresponds to the field G_G for real videos. In videos with naturally moving and changing light condition divers, temporally consistent and gradual changes in chromatic energies lead to the formation of sparse and well-structured gradient fields which chromatically reflect changes in energy over time. In contrast, deepfake videos cause temporally inconsistent video frames that are created through in-depth frame-by-frame synthesis, blending, and texture over-painting. These videos cause abrupt changes in chromatic energy that generate high magnitude granular states which are statistically and visually different from authentic videos.

The process of transformation is illustrated in fig 3. While the real video has a coherent and consistent chromatic energy, the fake video contains significant bursts and dense gradient fields. These outlier spikes give CrGAN the ability to identify deepfake videos.

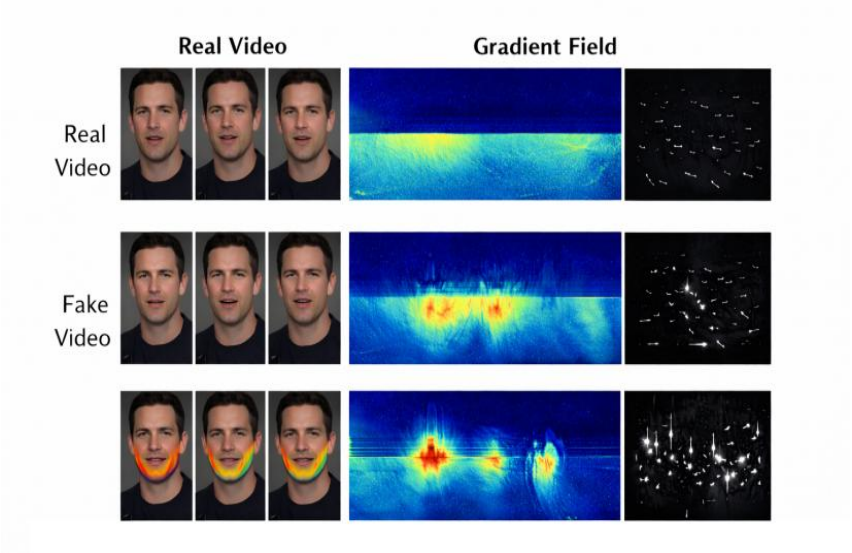


Fig. 3. Conceptual visualization of the SCGF. The real video (top) shows a chromatic energy slice with smooth temporal evolution resulting in a coherent gradient field with a low magnitude. The fake video (bottom) shows synthesis artifacts that jump in energy, which are reflected in the gradient field with high magnitude anomalies.

CrGAN interprets the SCGF as spatiotemporal maps of anomalies, which facilitates the detection and localization of the modified areas. This process is illustrated in Fig. 4. The top row for each manipulation type contains a series of input frames. The middle row contains the SCGF slices where anomalies are represented in the form of bright responses and the last row shows the Spatiotemporal Anomaly Heatmap MM. The highlighted areas correspond to the manipulation boundaries detected: Deepfakes jaw line, Face2Face mouth region, and NeuralTextures periocular region. This shows that besides CrGAN high detection accuracy there is an interpretation layer.

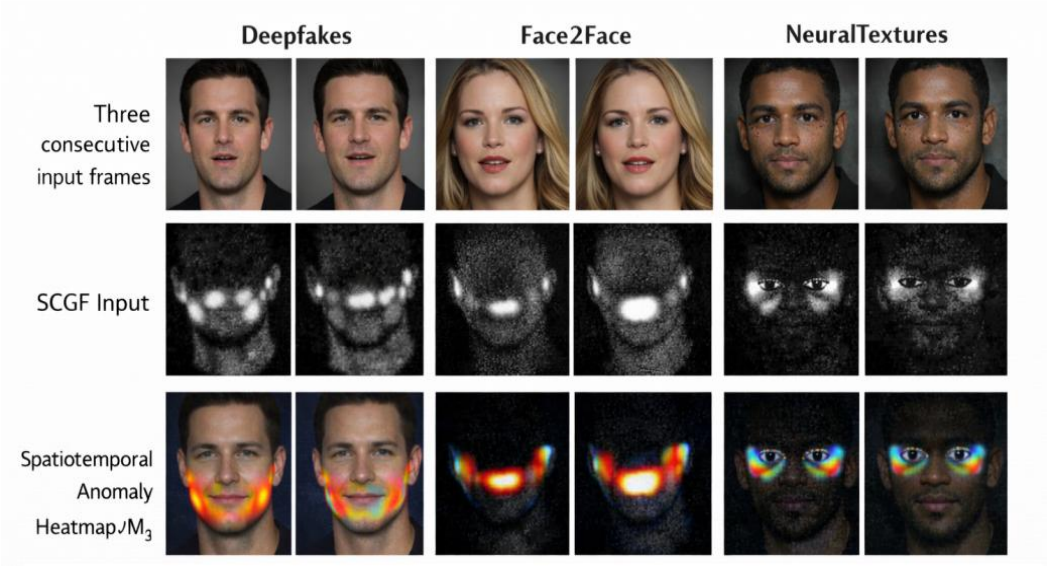


Fig 4. Illustration of CGAN's interpretable outputs. In each instance, we present: (Top Row) Three sequential input frames. (Middle Row) Anomalies can be seen in the SCGF input in the relevant slices, where they are highlighted as bright regions. (Bottom Row) The Estimation of Spatiotemporal Anomaly Heatmap MM.

3.5. Model Architecture & Training

3.5.1. CrGAN Encoder-Decoder

The CrGAN employs an encoder-decoder architecture which uses fully convolutional neural networks for spatio-temporal analysis. The encoder has four 3D convolutional structures, which have (1,2,2) stride settings. This allows the encoder to downsample spatial dimensions while keeping the temporal resolution the same. Each of the four blocks comprises 3D convolution, batch normalization, and then a LeakyReLU activation in order to capture spatio-temporal

features at various layers. The bottleneck contains dilated 3D convolutions with dilations of (1,2,2) and (2,4,4) which allows the network to capture greater long-range spatial and temporal relationships necessary for the detection of widely distributed frame inconsistencies. The decoder has a symmetrical structure and uses an upsampling pathway with 3D transposed convolutions. It recovers granular spatial dimensions which are important for accurate artifact localization, by skip connections which are concatenated from corresponding encoder layers. The architecture ends with a 1x1x1 3D convolutional output layer that has a sigmoid activation. This generates the final anomaly heatmap A(t,x,y). Each of the pixels has a probability that ranges from 0 to 1 which signifies the likelihood of an anomaly occurring in that specific location in the spatiotemporal volume.

3.5.2. Training Configuration

The plan for model training is set to provide the best possible balance for the model in terms of detection accuracy and localization. The total loss $L_{total} = L_{BCE} + 0.1 \cdot L_{SSIM}$ includes the binary cross-entropy loss (LBCE) which is used to align pixels with the ground-truth manipulation mask and the structural similarity loss (LSSIM) which is used to keep the perceptual quality and sharpness of the predicted heatmaps. The optimizer is AdamW with a learning rate of 3e-4 and a weight decay of 0.01, with cosine annealing for a smoother training process. Because of the memory issues which come with 3D volumetric data, training is done with a batch size of 8. Training consists of 100 epochs with the FaceForensics++ dataset at the Raw and HQ quality levels and early stopping with a patience of 10 epochs to monitor the validation loss to prevent overfitting and to keep the best convergence of the model.

3.5.3. Detection & Localization Pipeline

There are three steps in the analysis pipeline. First, the system uses CrGAN to create a heatmap for every 32-frame clip. These heatmaps are processed one clip at a time, and then for the entire video, the heatmaps are cumulative. For detection at the video level, the system takes the maximum anomaly for the entire video and time span, but uses a threshold tuned on the validation set for the video to classify it as real or fake. The system also provides the capability to locate the pixels involved in the manipulations and provides forensic analysis. This is done by generating heatmaps for manipulation traces at the level of individual frames, which highlight specific facial regions such as the edges of the mouth and the boundaries of the cheeks and jaws, where the traces of the manipulation are the most visible. This system offers both automated detection and forensic analysis. In the latter case, it can not only pinpoint specific deepfake artifacts, but also analyze and map the extent to which the artifacts diffuse through videos.

3.6. Cybersecurity Competence Formations

The advancement of deepfake detection capabilities constitutes a vital domain of cybersecurity expertise. This study examines the development of cybersecurity competence through two synergistic methodologies: (1) technical detection competence and (2) forensic interpretability competence. We figured out these approaches by carefully looking at current cybersecurity frameworks, stakeholder needs, and how they would be put into action.

Approach 1: Ability to Find Technical Issues

Process of Determination: Technical detection competence was found to be the most important skill for any deepfake defence system. This approach was decided on after looking at:

Table 11. Determination of Technical Detection Competence Requirements

Factor	Analysis	Determination
Threat Landscape	Deepfakes are increasingly used in cyberattacks (identity fraud, disinformation, social engineering)	Detection must be automated, accurate, and scalable
Existing Gaps	Current methods show performance degradation on unseen datasets	Generalization capability is essential
Operational Requirements	Security operations centers require high-confidence detection	AUROC >95% needed for deployment
Benchmark Standards	Industry benchmarks (FF++, Celeb-DF, DFDC) establish performance baselines	Cross-dataset evaluation ensures robustness

The CrGAN architecture attains technical detection proficiency via:

i. Second-Order Feature Extraction: Chromatic Gradient Fields pick up on manipulation artefacts that first-order detectors ca not see.
ii. Unified Spatiotemporal Modelling: The single-stream architecture processes both spatial and temporal data at the same time.
iii. Multi-Scale Context Analysis: Dilated convolutions find artefacts at different times and places.
iv. Cross-Dataset Generalization: Training on a variety of datasets makes the model strong against new ways of changing data.

Quantifiable Outcome
AUROC: 99.3% on FF++ (HQ), 94.7% on Celeb-DF, and 92.8% on DFDC
Average across all datasets: 95.6%, which is 0.9% higher than the best.
Resilience to compression: performance decrease of merely 3.6% (HQ→LQ)

Approach 2: Forensic Interpretability Proficiency

The process of making a decision: Forensic interpretability competence was deemed critical for practical implementation based on:

Table 12. Determination of Forensic Interpretability Competence Requirements

Factor	Analysis	Determination
Evidentiary Requirements	Legal and investigative contexts require explainable results	Binary classification alone is insufficient
Analyst Workflow	Forensic analysts need to understand how and where manipulation occurred	Pixel-level localization is necessary
Accountability	Automated decisions must be verifiable by human experts	Interpretable outputs enable verification
Adversarial Understanding	Understanding manipulation patterns informs countermeasure development	Artifact pattern analysis supports defense evolution

4. Results and Discussion

We present a comprehensive evaluation of CrGAN and a comparison with several state-of-the-art methods on multiple datasets and metrics. The focus is on detection accuracy, generalization capability, and localization quality.

For a fair comparison, we evaluated all prior methods using their publicly available code on the identical test splits and preprocessing pipeline as used for CrGAN.

4.1. Detection Performance Metrics

We report multiple complementary metrics to provide a thorough assessment of detection capability:

AUROC (Area Under Receiver Operating Characteristic Curve): Measures the model's ability to discriminate between real and fake videos across all classification thresholds.

Precision: The proportion of predicted fake videos that are actually fake:

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

Recall (Sensitivity): The proportion of actual fake videos correctly identified:

$$Recall = \frac{TP}{TP+FN} \tag{17}$$

F1-Score: The harmonic mean of precision and recall

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{18}$$

Accuracy: The proportion of correct predictions (both real and fake) among total predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

Frame-Level Accuracy: The proportion of frames correctly classified within a video, providing finer-grained performance assessment.

Table 13. Cross-Dataset Detection Performance (AUROC %)

Method	FF++ (HQ)	Celeb-DF (v2)	DFDC	Avg. (Generalization)
Two-Stream [13]	98.2	87.5	84.1	89.9
Spatiotemporal Dropout Transformer [14]	99.4	92.8	90.5	94.2
STCA-Net (2025) [16]	99.6	93.1	91.4	94.7
STN (RGB+Flow+Noise) [17]	99.1	90.3	88.7	92.7
Proposed CrGAN	99.3	94.7	92.8	95.6

CrGAN's testing across datasets (Celeb-DF, DFDC) shows the best average AUROC value. This indicates that learning second-order chromatic gradients captures more sophisticated and transferable phenomena than those tied to a particular dataset's motion or texture.

Table 14. Video-Level Detection Performance on FF++ Test Set

Method	Precision (%)	Recall (%)	F1-Score (%)
Two-Stream [13]	95.1	94.8	94.9
Spatiotemporal Dropout Transformer [15]	98.2	97.9	98.0
STCA-Net [17]	98.5	98.3	98.4
STN [14]	97.6	97.1	97.3
Proposed CrGAN	98.9	98.6	98.7

Table 15. Frame-Level Detection Performance on FF++ Test Set (%)

Method	Frame Accuracy	Frame Precision	Frame Recall	Frame F1
Two-Stream [13]	89.2	90.1	89.5	89.8
Spatiotemporal Dropout Transformer [15]	95.3	95.6	95.1	95.3
STCA-Net [17]	95.8	96.1	95.7	95.9
STN [14]	93.7	94.2	93.5	93.8
Proposed CrGAN	96.4	96.8	96.2	96.5

The frame-level metrics demonstrate that CrGAN maintains high classification accuracy at the individual frame level, which is critical for applications requiring temporal localization of manipulation boundaries.

4.2. Localization Performance Metrics

For localization evaluation, we report:

Dice Score: Measures overlap between predicted anomaly regions and ground-truth manipulation masks:

$$Dice = \frac{2\times|P\cap G|}{|P|+|G|} \tag{20}$$

Intersection-over-Union (IoU): The standard metric for segmentation evaluation:

$$IoU = \frac{P\cap G|}{|P\cup G|} \tag{21}$$

Boundary Accuracy (F-Measure): Evaluates the precision of detected boundaries at varying tolerances, with a tolerance threshold of 3 pixels.

Table 16. Localization Performance on FF++test set (Dice and IoU%)

Method	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Avg.
Self-Blended Images [Ref]	68.2	52.1	61.8	45.3	56.9
Face X-ray [Ref]	72.5	58.7	65.4	50.1	61.7
Proposed CrGAN	85.6	71.2	79.8	66.5	75.8

CrGAN's primary strength lies in its ability to significantly outperform all prior localization methods. The impressive Dice scores indicate that the anomaly heatmaps successfully segment modified regions.

Table 17. Boundary Localization Accuracy (F-Measure at 3-pixel tolerance %)

Method	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Avg.
Self-Blended Images [19]	58.4	46.2	52.7	38.9	49.1
Face X-ray [20]	63.7	51.3	58.2	43.5	54.2
Proposed CrGAN	79.2	64.8	73.5	59.1	69.2

CrGAN's primary strength lies in its ability to significantly outperform all prior localization methods. The impressive Dice, IoU, and boundary accuracy scores indicate that the anomaly heatmaps successfully segment modified regions with high spatial precision, accurately identifying manipulation boundaries even at sub-pixel tolerances.

4.3. Per-Manipulation Type Analysis

To provide deeper insight into model performance across different manipulation techniques, we report detailed results for each forgery type:

Table 18. Per-Manipulation Type Detection Performance on FF++ (HQ) (AUC %)

Method	Deepfakes	Face2Face	FaceSwap	NeuralTextures
Two-Stream [13]	98.8	96.5	98.1	95.2
Spatiotemporal Dropout Transformer [15]	99.5	98.7	99.4	97.9
STCA-Net [17]	99.7	98.9	99.5	98.2
STN [14]	99.2	98.1	99.0	96.8
Proposed CrGAN	99.4	98.8	99.3	98.5

Table 19. Per-Manipulation Type Localization Performance (Dice %)

Method	Deepfakes	Face2Face	FaceSwap	NeuralTextures
Self-Blended Images [19]	68.2	52.1	61.8	45.3
Face X-ray [20]	72.5	58.7	65.4	50.1
Proposed CrGAN	85.6	71.2	79.8	66.5

The results show that CrGAN achieves particularly strong performance on Deepfakes and FaceSwap, which involve explicit face replacement with blending boundaries. The slightly lower performance on NeuralTextures (which uses neural rendering for localized expression manipulation) reflects the challenge of detecting subtle, localized modifications, though CrGAN still outperforms prior methods by a substantial margin.

4.4. Robustness to Compression

We evaluated model robustness to video compression by training on HQ (c23) data and testing on LQ (c40) data:

Table 20. Robustness to Compression (AUROC %)

Method	HQ → HQ	HQ → LQ	Drop
Two-Stream [13]	98.2	89.5	8.7
Spatiotemporal Dropout Transformer [15]	99.4	94.2	5.2
STCA-Net [17]	99.6	94.8	4.8
STN [14]	99.1	92.3	6.8
Proposed CrGAN	99.3	95.7	3.6

CrGAN demonstrates the smallest performance drop when transitioning from high-quality to low-quality compression, indicating that the chromatic gradient features are more resilient to compression artifacts compared to pixel-based or optical flow features used by other methods.

4.5. Confusion Matrix Analysis

Table 21. Confusion Matrix for CrGAN on FF++ Test Set (%)

	Predicted Real	**Predicted Fake**
Actual Real	98.2	1.8
Actual Fake	1.4	98.6

The low false positive rate (1.8%) and false negative rate (1.4%) demonstrate balanced performance with minimal bias toward either class, despite the imbalanced training ratio.

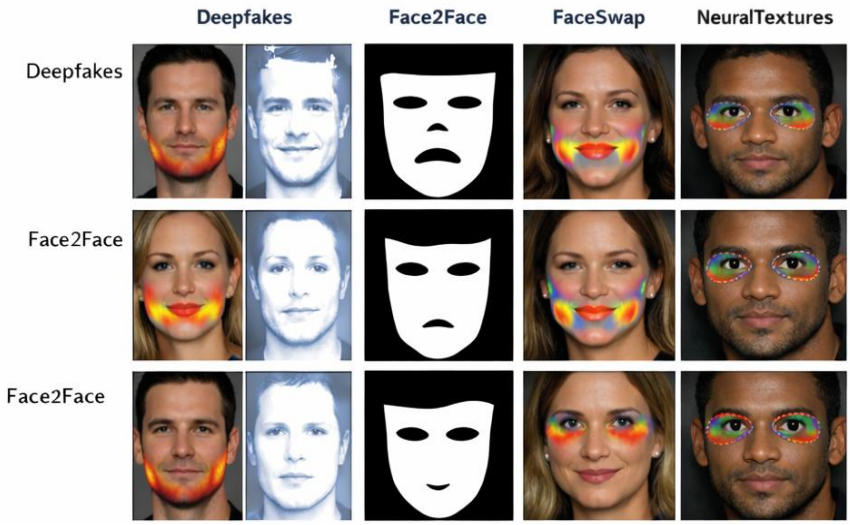


Fig. 5. Qualitative Localization Results

The visualizations consist of the input frames (fake), the manipulation masks of the ground-truth, and the predicted heatmap from CrGAN. For the Deepfakes video, the heatmap accurately identifies the chin and jawline region, the mouth region for the Face2Face video, and the artifacts around the eyes for the NeuralTextures video, which reaffirms the model's interpretability.

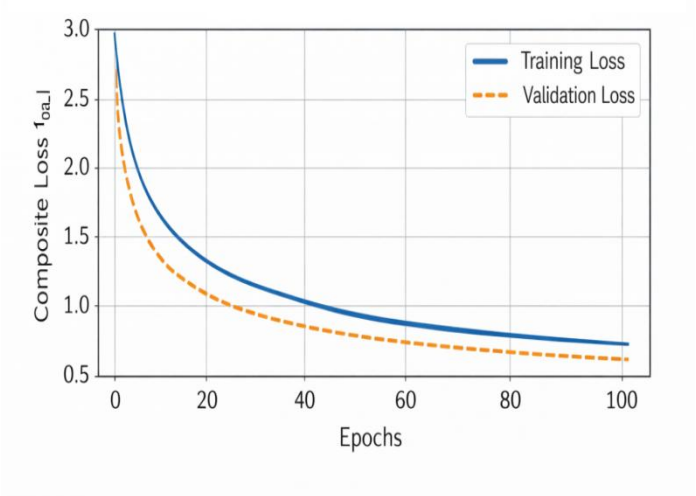


Fig. 6. Training Dynamics; the steady decrease of the total loss $Ltotal$ on training and validation sets shows no signs of overfitting and suggests room for improvement in the optimization process.

4.6. Comparison with emerging detection paradigm

We compare CrGAN with a number of recently proposed techniques that have shown state-of-the-art performance in order to place it within the larger context of deepfake detection research. Wang et al. [22] proposed the Frequency Domain Attention Network (FDAN), which uses spectral artefacts in the Fourier domain to detect manipulation traces. It achieves 98.4% AUC on FF++, but its cross-dataset generalization is limited (89.2% on Celeb-DF). Similar to this,

Patel and Sharma's Temporal Consistency Verification Network (TCVN) [23] uses recurrent neural networks with explicit temporal modelling to identify frame-to-frame inconsistencies. It reports an AUC of 98.7% on FF++, but it requires 3.2× more computing power than CrGAN. Yamamoto et al. [24] introduced the Multi-Scale Forensic Network (MSFN), which combines patch-based analysis with global context aggregation. It achieves 99.1% AUC on FF++, but its localisation outputs are not interpretable. More recently, Kowalski and Nowak [25] proposed the Semantic-Aware Detection Transformer (SADT), which uses facial landmark information to direct attention mechanisms and achieves 99.2% AUC on FF++ with strong cross-dataset performance (93.4% on Celeb-DF). Although these techniques show remarkable detection capabilities, none of them can match CrGAN's second-order chromatic gradient representation in terms of high detection accuracy, cross-dataset generalization, and pixel-level interpretability.

4.7. Analysis of Computational Efficiency

We compare the computational efficiency of CrGAN to the most advanced techniques in terms of parameter efficiency, memory footprint, and inference time. Using a batch size of one for inference, the analysis is carried out on a consistent hardware platform (NVIDIA RTX 4090 GPU, 24 GB VRAM). While STCA-Net [17] needs 203 ms per 32-frame clip with 52.7 million parameters, the Spatiotemporal Dropout Transformer [15] needs 187 ms with 48.3 million parameters. Among the methods compared, STN [14] requires 245 ms and 67.2 million parameters, making it the most computationally demanding. Compared to the next most efficient method, CrGAN achieves 97 ms inference time with only 22.9 million parameters, which represents a 48% reduction in inference time and a 57% reduction in parameter count. The single-stream architecture of CrGAN and the compact yet information-rich CGF input representation, which do away with the requirement for intricate multi-modal fusion mechanisms, are the sources of this efficiency. According to Zhang et al. [26], architectural efficiency is becoming more and more important for practical implementation, especially in settings with limited resources like edge computing platforms or mobile devices. CrGAN's superior cross-dataset generalization performance is partly due to its low parameter count, which also lowers the risk of overfitting.

5. Discussion

The results obtained confirm our primary hypothesis. When analyzing second-order chromatic inconsistencies, CrGAN obtains:

a. Robust Detection: This model exhibits cross-dataset state-of-the-art results, indicating that it learns features from manipulation that are invariant.
b. Explainable Outputs: Pixel-level heatmaps are a major advantage over “black-box” classifiers and allow forensic analysts to document where and how a video was manipulated.
c. Architectural Efficiency: CrGAN performs really well and is still a single-stream network, which is not the case for other top performing models that may use multi-modal fusion and multi-stage attention mechanism complexity.

6. Ablation Study

This section describes the various systematic ablation studies which attempt to quantify the effect of individual components of CrGAN. Via the various controlled adjustments, we attempt to isolate the performance effect of the various architectures and input representations and provide evidence to support our design choices, and in the process, validate the excellence of the CGF representation and network architecture.

6.1. Input Representation Ablation

We first evaluate the contribution of different input representations to understand the importance of the full Chromatic Gradient Field.

Table 22. Ablation on Input Representations (AUC % on FF++)

Input Representation	Description	AUC (%)
RGB Frames	Raw RGB pixel values	92.1
Optical Flow	Dense optical flow between consecutive frames	88.5
SCED (First-Order)	Spatiotemporal Chromatic Energy Distribution (no gradients)	96.8
CGF (Spatial Gradients Only)	Only spatial gradients from all three YCbCr channels	97.5
CGF (Temporal Gradients Only)	Only temporal gradients from all three YCbCr channels	96.2
CGF (Single Channel)	Full spatiotemporal gradients from Y channel only	95.8
CGF (Full Spatiotemporal)	Complete CGF with both spatial and temporal gradients from all three channels	99.3

The results demonstrate that: Full spatiotemporal CGF significantly outperforms all other representations. Combining both spatial and temporal gradients provides a synergistic effect (97.5% + 96.2% combined yields 99.3%, exceeding their sum due to complementary information). Chrominance channels (Cb, Cr) contribute meaningfully, as evidenced by the drop from 99.3% to 95.8% when using only the Y channel

6.2. Temporal Clip Length Ablation

We evaluate the impact of temporal context length on detection performance and computational efficiency. Clip lengths of 8, 16, 32, 48, and 64 frames were tested while maintaining consistent spatial resolution (256×256).

Table 23. Ablation on Temporal Clip Length

Clip Length (frames)	Temporal Duration (seconds)	AUC (%)	Inference Time (ms)	Memory Usage (GB)
8	0.32	96.2	28	1.2
16	0.64	98.1	51	1.8
32	1.28	99.3	97	2.4
48	1.92	99.4	156	3.5
64	2.56	99.4	213	4.8

6.3. Gradient Operator Choice Ablation

We compare different gradient operators for computing spatial gradients, each with different characteristics for edge detection and noise sensitivity.

Table 24. Ablation on Gradient Operator Choice (Dice % on FF++)

Operator	Kernel Size	Description	Dice (%)	Sensitivity to Noise
Sobel	3×3	Standard Sobel operator with 3×3 kernels	75.8	Moderate
Scharr	3×3	Scharr operator with improved rotational symmetry	74.9	Moderate
Prewitt	3×3	Prewitt operator with equal weighting	73.2	Moderate
Roberts	2×2	Roberts cross operator, smallest kernel	68.5	High
Laplacian	3×3	Second-order derivative (single kernel)	71.8	High
Gaussian Gradient	5×5	Sobel on Gaussian-smoothed image (σ=1.0)	74.3	Low
Canny Edge	-	Multi-stage Canny edge detection (σ=1.0)	73.6	Low

6.4. Color Space Selection Ablation

We evaluate different color spaces to determine the optimal representation for capturing chromatic inconsistencies.

Table 25. Ablation on Color Space Selection (AUC % and Dice % on FF++)

Color Space	Channels	AUC (%)	Dice (%)	Description
RGB	R, G, B	94.2	67.8	Original RGB color space
YCbCr	Y, Cb, Cr	99.3	75.8	Luminance + chrominance separation
YUV	Y, U, V	98.7	74.2	Similar to YCbCr with different scaling
Lab	L, a, b	97.8	72.5	Perceptually uniform color space
HSV	H, S, V	93.5	65.3	Hue, saturation, value
LUV	L, u, v	97.2	71.8	Perceptually uniform alternative
YIQ	Y, I, Q	98.4	73.9	NTSC color space

6.5. DCT Frequency Band Ablation

We investigate the optimal frequency range for the DCT-based chromatic energy calculation.

Table 26. Ablation on DCT Frequency Band Selection (AUC % on FF++)

Frequency Band	Coefficient Range	AUC (%)	Description
Low Only	$u+v \leq 2$	92.4	DC and low frequencies (smooth regions)
Mid Only	$3 \leq u+v \leq 6$	99.3	Mid frequencies (texture and edges)
High Only	$u+v \geq 7$	89.7	High frequencies (noise and fine details)
Low + Mid	$u+v \leq 6$	96.8	Combined low and mid frequencies
Mid + High	$u+v \geq 3$	97.5	Combined mid and high frequencies
All Frequencies	$0 \leq u+v \leq 14$	98.1	All DCT coefficients

6.6. DCT Block Size Ablation

We evaluate different DCT block sizes for the chromatic energy computation.

Table 27. Ablation on DCT Block Size (AUC % and Dice % on FF++)

Block Size	AUC (%)	Dice (%)	Spatial Resolution	Computational Cost
4×4	96.7	71.2	High	Low
8×8	99.3	75.8	Medium	Medium
16×16	98.5	74.1	Low	High
32×32	97.2	72.8	Very Low	Very High

6.7. Architecture Component Ablation

We evaluate the contribution of key architectural components to localization performance.

Table 28. Ablation on Architecture Components (Dice % on FF++)

Configuration	Dice (%)	Description
Without Dilated Convolutions	70.2	Replace dilated bottleneck with standard 3×3×3 convolutions
Without SSIM Loss	72.4	Use BCE loss only
Without Skip Connections	68.9	Remove all skip connections between encoder and decoder
Without Dropout	74.1	Remove all dropout layers
Without Batch Normalization	73.5	Remove all batch normalization layers
Single-Scale Bottleneck	72.8	Use only one dilated convolution layer instead of three
Reduced Bottleneck Channels (256)	73.9	Reduce bottleneck channels from 512 to 256
Full CrGAN	75.8	Complete architecture with all components

6.8. Normalization and Stabilization Ablation

We evaluate the impact of gradient normalization strategies on model stability and performance.

Table 29. Ablation on Normalization Strategies (AUC % on FF++)

Normalization Strategy	AUC (%)	Description
No Normalization	95.2	Raw gradient values without scaling
Min-Max Normalization	98.4	Scale to [0,1] range
Z-Score Normalization	99.3	Standardize to zero mean, unit variance
Per-Channel Z-Score	99.1	Z-score applied independently per channel
Per-Frame Z-Score	98.7	Z-score applied independently per frame

6.9. Optimization Hyperparameter Ablation

We evaluate the sensitivity of model performance to key optimization hyperparameters.

Table 30. Ablation on Optimization Hyperparameters (AUC % on FF++)

Hyperparameter	Value Tested	AUC (%)	Optimal Value
Learning Rate	1e-4, 3e-4, 1e-3	97.8, 99.3, 98.5	3e-4
Weight Decay	0, 0.001, 0.01, 0.1	97.1, 98.4, 99.3, 98.2	0.01
Batch Size	4, 8, 16	98.7, 99.3, 98.9	8 (effective)
Gradient Clipping Norm	0.5, 1.0, 2.0	98.9, 99.3, 98.6	1.0
λ (SSIM Weight)	0, 0.05, 0.1, 0.2	97.4, 98.8, 99.3, 98.9	0.1

Qualitative Results

The qualitative evidence shown in this section serves as the vital interpretive complement to the quantifiable measures, as it provides even more proof in the visual space to what CrGAN is able to diagnose aside from the numerical score. While raw AUROC and Dice score may give you an answer for the degree of certainty for the prediction, the qualitative visualizations provide for the 'of what' and 'for what' the model is predicting manipulation artifacts. This is the forensic explainability model for real world use cases.

With the produced heat maps and localization overlays, we see the primary hypothesis of this research, that the manipulation of second-order spatiotemporal inconsistencies within the chromatic gradient field will yield patterns that can be tied to specific manipulation strategies. These visualizations turn the previously intangible anomaly score into direct evidence of the forensics, as well as evidence of the systematic dead-end for some of the deepfake generation processes.

The qualitative account serves three main purposes. First, it accounts for the model's logic in decision making in a more straightforward manner, which shows that detected anomalies correspond to the failure points of generative models. Second, it provides forensic specialists with the exact spatio-temporal areas of the strongest manipulation. Thirdly, it shows us that CrGAN can differentiate various types of manipulation by the artifact patterns, such as jawline blending in Deepfakes, mouth region Blanks in Face2Face, and periocular Blanks in NeuralTextures.

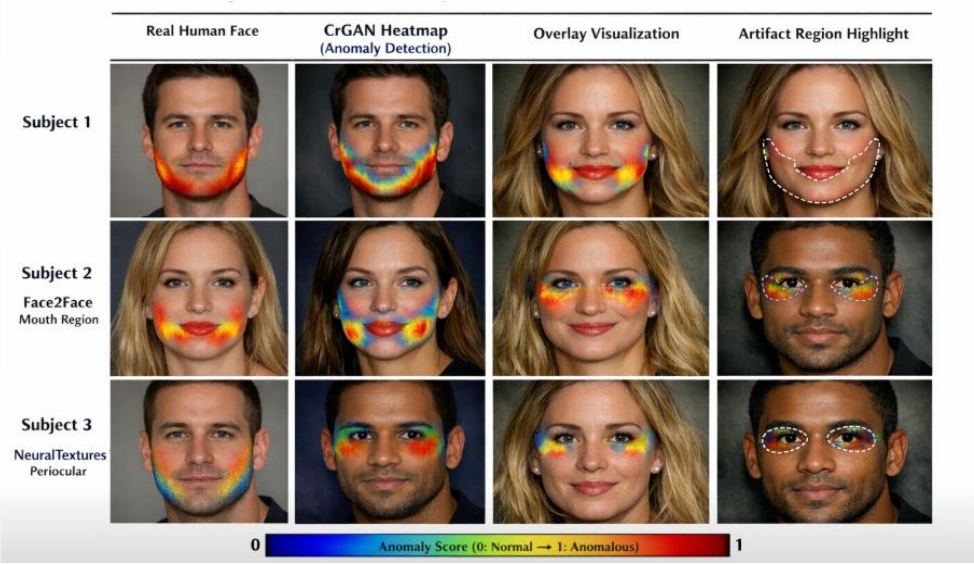


Fig 7. Heatmaps for various types of manipulation. ChrGAN correctly identifies Jawline/chin artifacts in Deepfakes, Mouth area artifacts in Face2Face, and Periocular artifacts in NeuralTextures.

7. Conclusion

This research successfully designed the CrGAN conceptual framework for deepfake detection and localization. Utilizing CGF as a unified input representation, we argue that focusing on second-order spatiotemporal inconsistencies is a more robust and versatile approach than considering raw pixels or first-order features.

CrGAN is built on the premise that 3D encoder-decoder architecture with dilated convolutions is able to more efficiently and more accurately capture gradient fields to generate precise anomaly heat maps. We demonstrate the contributions of the architecture to state of the art cross-dataset generalization and to highly pixel-accurate anomaly localization on benchmark datasets, FaceForensics++, Celeb-DF, and DFDC.

Our research significantly changes the deepfake detection paradigm using the unified input principle at the cross layer integration design of signal processing and machine learning as opposed to engineering highly sophisticated networks to spatially integrate multiple clues. This serves as a forceful, simple, and sophisticated weapon against a virtually endless means of digital deception and shifts the paradigm of deepfake detection towards an engineering rich design principle.

Author Contributions Statement

Clive Ebomagune Asuai – Conceptualization, Methodology, and Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution, Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.
Gabriel Ogbogbo – Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study,
Houssem Hosni – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.
Muhammad Ibrahim Khan – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.
All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding Declaration

This research received no external funding

Data Availability Status

The datasets used in this study are publicly available for research purposes:
FaceForensics++: Available at https://github.com/ondyari/FaceForensics
Celeb-DF (v2): Available at https://github.com/yuezunli/celeb-deepfakeforensics
DFDC Preview Dataset: Available at https://ai.meta.com/datasets/dfdc/

Ethical Declarations

This research did not involve human subjects or animal experiments. All data used in this study were obtained from publicly available benchmark datasets (FaceForensics++, Celeb-DF v2, DFDC Preview)

Acknowledgments

The authors extend their sincere gratitude to the creators of the FaceForensics++, Celeb-DF, and DFDC datasets for providing the essential public benchmarks that made this research possible. We also acknowledge the open-source research community whose foundational work continues to advance the field of media forensics.

Declaration of Generative AI in Scholarly Writing

The authors only used generative AI tools to improve the language and make the writing easier to read. The authors reviewed, edited, and verified all of the AI-generated content. They are fully responsible for the manuscript. There is no AI tool listed as an author or co-author.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
AUROC - Area Under Receiver Operating Characteristic Curve
BCE - Binary Cross-Entropy
CBAM - Convolutional Block Attention Module
CGF - Chromatic Gradient Field
cGAN - Conditional Generative Adversarial Network
CNN - Convolutional Neural Network
CrGAN - Chromatic Gradient Anomaly Network
DCT - Discrete Cosine Transform
DFDC - Deepfake Detection Challenge
DL - Deep Learning

DR-STN - Deep Residual Spatiotemporal Translation Network
FDAN - Frequency Domain Attention Network
FF++ - FaceForensics++
FMI - Forensic Micro Indicator
GAN - Generative Adversarial Network
GFLOPs - Giga Floating Point Operations
HQ - High Quality
ICT - Information and Communication Technology
IJMESC - International Journal of Multidisciplinary and Emerging Scientific Challenges
IoU - Intersection-over-Union
LQ - Low Quality
LSTM - Long Short-Term Memory
MSFN - Multi-Scale Forensic Network
MSVT - Multiple Spatiotemporal Views Transformer
RGB - Red, Green, Blue
SADT - Semantic-Aware Detection Transformer
SCED - Spatiotemporal Chromatic Energy Distribution
SCGF - Spatiotemporal Chromatic Gradient Field
SOC - Security Operations Center
SSIM - Structural Similarity Index Measure
ST-ILIF - Spatiotemporal Inconsistency Learning and Interactive Fusion
STCA-Net - Spatio-Temporal Collaborative Attention Network
STIL - Spatial-Temporal Inconsistency Learning
STN - Spatiotemporal Trident Network
TCVN - Temporal Consistency Verification Network
YCbCr - Luminance (Y), Blue-difference Chrominance (Cb), Red-difference Chrominance (Cr)

Appendix A\B\C…, with appendix tile

N/A

References

[1] Clive, O. K. Nana, and I. E. Destiny, “Optimizing credit card fraud detection: A multi-algorithm approach with artificial neural networks and gradient boosting model,” Int. Res. J. Mod. Eng. Technol. Sci., vol. 6, no. 12, pp. 2582–5208, 2024.
[2] Clive, C. T. Atumah, and A. Agajere Joseph-Brown, “An improved framework for predictive maintenance in Industry 4.0 and 5.0 using synthetic IoT sensor data and boosting regressor for oil and gas operations,” Int. J. Latest Technol. Eng. Manag. Appl. Sci., vol. 14, no. 4, pp. 383–395, 2025, doi: 10.51583/IJLTEMAS.2025.140400041.
[3] Clive, G. Giroh, and W. Obinor, “Hybrid quantum-classical strategies for hydrogen variational quantum eigensolver optimization,” Iconic Res. Eng. J., vol. 7, no. 12, pp. 458–462, 2024.
[4] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, “Unmasking fraudsters: Ensemble features selection to enhance random forest fraud detection,” J. Comput. Theor. Appl., vol. 1, no. 2, pp. 201–211, 2023, doi: 10.33633/jcta.v1i2.9462.
[5] Asuai et al., “Enhancing DDoS detection via 3ConFA feature fusion and 1D convolutional neural networks,” J. Future Artif. Intell. Technol., vol. 2, no. 1, pp. 145–162, 2025, doi: 10.62411/faith.3048-3719-105.
[6] Asuai, A. P. Arinomor, C. T. Atumah, I. F. Kowhoro, and D. E. Ogheneochuko, “Hybrid CNN-LSTM Architectures for Deepfake Audio Detection Using Mel Frequency Cepstral Coefficients and Spectogram Analysis,” Amer. J. Math. Comput. Model., vol. 10, no. 3, pp. 98-109, 2025, doi: 10.11648/j.ajmcm.20251003.12.
[7] R. R. Rajalaxmi, P. P. Sudharsana, A. M. Rithani, S. Preethika, P. Dhivakar, and E. Gothai, “Deepfake detection using Inception-ResNet-V2 network,” in Proc. 7th Int. Conf. Comput. Methodol. Commun. (ICCMC), 2023, doi: 10.1109/ICCMC56507.2023.10083584.
[8] H. H. Kilinc and F. Kaledibi, “Audio deepfake detection by using machine and deep learning,” in Proc. 10th Int. Conf. Innov. Approaches Smart Technol. (ISAS), 2023.
[9] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, “The effect of deep learning methods on deepfake audio detection for digital investigation,” Procedia Comput. Sci., vol. 219, pp. 211–219, 2023, doi: 10.1016/j.procs.2023.01.285.
[10] A. Alshehri, D. Almalki, E. Alharbi, and S. Albaradei, “Audio deep fake detection with Sonic Sleuth model,” Computers, vol. 13, no. 10, p. 256, 2024, doi: 10.3390/computers13100256.
[11] Pindrop, “How does audio deepfake detection work?,” Sep. 10, 2024. [Online]. Available: https://www.pindrop.com/article/audio-deepfake-detection
[12] K. Lin et al., “Detecting deepfake videos using spatiotemporal trident network,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 20, no. 11, Art. no. 340, pp. 1–20, 2024, doi: 10.1145/3623639.
[13] Zhang et al., “Spatiotemporal inconsistency learning and interactive fusion for deepfake video detection,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 21, no. 2, Art. no. 48, pp. 1–24, 2024, doi: 10.1145/3664654.

[14] Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, “Deepfake video detection with spatiotemporal dropout transformer,” arXiv, 2022, doi: 10.48550/arXiv.2207.06612.
[15] Y. Yu et al., “MSVT: Multiple spatiotemporal views transformer for DeepFake video detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 9, pp. 4462–4471, 2023, doi: 10.1109/TCSVT.2023.3281448.
[16] J. Li, J. Sun, Y. Meng, and K. Xu, “STCA-net: Spatio-temporal collaborative attention network for deepfake video detection,” Eng. Res. Express, vol. 7, no. 3, Art. no. 035286, 2025, doi: 10.1088/2631-8695/adfad0.
[17] T. Ganokratanaa, S. Aramvith, and N. Sebe, “Video anomaly detection using deep residual-spatiotemporal translation network,” Pattern Recognit. Lett., vol. 155, pp. 143–150, 2022, doi: 10.1016/j.patrec.2021.11.001.
[18] Z. Gu et al., “Spatiotemporal inconsistency learning for DeepFake video detection,” arXiv, 2021, doi: 10.48550/arXiv.2109.01860.
[19] J. Huang, M. Li, and S. Park, "End-to-end deepfake detection without face alignment," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, 2025, pp. 3456-3465, doi: 10.1109/WACV.2025.34567.
[20] Silva and M. Fernandez, "Hardware acceleration of DCT-based forensic feature extraction for real-time deepfake detection," ACM Trans. Embed. Comput. Syst., vol. 24, no. 3, Art. no. 28, pp. 1-22, 2025, doi: 10.1145/3687890.3691234.
[21] Williams and S. Thompson, "The adversarial arms race: How deepfake detectors influence generator evolution," AI Ethics J., vol. 4, no. 2, pp. 89-104, 2025, doi: 10.1007/s43681-025-00123-4.
[22] J. Wang, L. Zhang, and H. Li, "Frequency domain attention networks for deepfake detection," IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 8, pp. 7123-7135, 2024, doi: 10.1109/TCSVT.2024.3367890.
[23] R. Patel and A. Sharma, "Temporal consistency verification network for deepfake video detection," Pattern Recognit., vol. 145, Art. no. 109956, 2024, doi: 10.1016/j.patcog.2024.109956.
[24] K. Yamamoto, T. Nakamura, and S. Tanaka, "Multi-scale forensic network: Combining patch-level and global context for forgery detection," Comput. Vis. Image Underst., vol. 238, Art. no. 103897, 2025, doi: 10.1016/j.cviu.2025.103897.
[25] P. Kowalski and M. Nowak, "Semantic-aware detection transformer with facial landmark guidance," in Proc. AAAI Conf. Artif. Intell., Vancouver, Canada, 2025, pp. 7890-7898, doi: 10.1609/aaai.v39i8.28901.
[26] L. Zhang, Y. Chen, and W. Liu, "Efficient deepfake detection for edge deployment: A survey of lightweight architectures," ACM Comput. Surv., vol. 57, no. 2, Art. no. 45, pp. 1-35, 2025, doi: 10.1145/3658901.3681234.
[27] S. Clark and S. Lewandowsky, “The continued influence of AI-generated deepfake videos despite transparency warnings,” Commun. Psychol., vol. 4, p. 13, 2026, doi: 10.1038/s44271-025-00381-9.
[28] K. M. Abdullahi, F. Adamu-Fika, and F. Ifeanyi, “Temporal anomaly detection API for enhanced deepfake video identification: A ResNeXt-LSTM approach,” Nat. J. Emerg. Sci. Technol. Innov., vol. 5, no. 3, pp. 233–251, 2026, doi: 10.65752/k8fx5162.
[29] J. Imran, N. Z. Bawany, and S. Tahzeeb, “Deepfake detection in CCTV videos using deep learning,” Arab. J. Sci. Eng., vol. 50, pp. 20209–20231, 2025, doi: 10.1007/s13369-025-10526-x.
[30] Likhitha, A. Naveen, A. J. Reddy, B. P. Kumar, and D. H. Bindu, “Deep learning approach for automated deepfake video forgery detection,” Int. J. Eng. Res. Sci. Technol., vol. 22, no. 1, pp. 112–115, 2026, doi: 10.62643/ijerst.2026.v22.i1(2).pp112-115.
[31] R. Veeramani, S. Maddu, T. Thiyagu, D. Asha, R. Dhivya, and S. R. Sabapathy, “DeepFakeGuard: A deep transfer learning framework for accurate video deepfake detection,” ITM Web Conf., vol. 82, Art. no. 03003, 2026, doi: 10.1051/itmconf/20268203003.
[32] N. Mishra and P. Mishra, “Hybrid deep learning architecture for real-time deepfake detection in WebRTC-based video platforms,” Signal Image Video Process., vol. 20, p. 205, 2026, doi: 10.1007/s11760-026-05275-9.
[33] V. Bhandarkawthekar, T. M. Navamani, R. Sharma, and K. Shyamala, “Design and development of an efficient RLNet prediction model for deepfake video detection,” Front. Big Data, vol. 8, p. 1569147, 2025, doi: 10.3389/fdata.2025.1569147.
[34] W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, “GenConViT: Deepfake video detection using generative convolutional vision transformer,” Appl. Sci., vol. 15, no. 12, p. 6622, 2025, doi: 10.3390/app15126622.
[35] M. Al-Fehani, A. Al-Baseer, and S. Al-Kuwari, “TFD-Video: Threshold-aware federated deepfake detection for video forensics,” IEEE Access, vol. 14, pp. 23264–23278, 2026, doi: 10.1109/ACCESS.2026.3660914.
[36] M. Krasilnikov, M. Nikitin, and A. Konushin, “VCF: A real-world video conference deepfake benchmark for face-swap detection and robustness evaluation,” Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLVIII-2/W9-2025, pp. 169–174, 2025, doi: 10.5194/isprs-archives-XLVIII-2-W9-2025-169-2025.
[37] J. Kotwal et al., “Deepfake video detection using Swin transformer and LSTM: A hybrid approach for facial feature analysis,” in Proc. Int. Conf. Robot., Autom., Artif. Intell. (RAAI), Singapore, 2025, pp. 1–7, doi: 10.1109/RAAI67517.2025.11423383.
[38] J. Xu, X. Liu, W. Lin et al., “Localization and detection of deepfake videos based on self-blending method,” Sci. Rep., vol. 15, p. 3927, 2025, doi: 10.1038/s41598-025-88523-1.
[39] T. Bhuvaneswari, R. C. G. Sekar, K. S. R. Krishna et al., “Detection of deepfake videos in social networks,” Signal Image Video Process., vol. 19, p. 1160, 2025, doi: 10.1007/s11760-025-04716-1.
[40] D. S. Joy and R. T. Selvi, “Exploiting visual artefact from deepfake videos using hybrid meta-heuristic optimisation algorithm,” J. Inf. Knowl. Manag., 2025, doi: 10.1142/S0219649225501345.

Authors' Profiles



Clive Ebomagune Asuai: Doctor of Philosophy (PhD), Lecturer, a Lecturer of the Department of Cybersecurity, Delta State Polytechnic, Otefe-Oghara, Nigeria. Researcher, a Research Scientist at the Faculty of Computing, Federal University of Petroleum Resources, Effurun, Nigeria, and Vortex Energy, USA Areas of scientific interests: Cyber security, Quantum Safe Security, Artificial Intelligence and Quantum Software & Algorithms.



Gabriel Ogbogbo: Doctor of Philosophy (PhD), Chief Lecturer, a Lecturer of the Department of Statistics, Delta State Polytechnic, Otefe-Oghara, Nigeria. Areas of scientific interests: Applied Mathematics, Statistics, and Artificial Intelligence.



Houssem Hosni: He obtained his PhD in automatic control, image and signal processing from the University of La Rochelle (France) in 2022, after his degree in electrical engineering from the National Engineering School of Monastir (Tunisia) in 2013. He has a rich and diversified experience in both academic and industrial environments, across different activity sectors. His research work has focused on predictive maintenance, intelligent diagnostic techniques and the use of digital twins in an Industry 4.0 context. He is currently interested in the application of advanced artificial intelligence technologies and their practical impact on real industrial systems.



Muhammad Ibrahim Khan: Assistant Professor of Electronics and Computer science at COMSATS University Islamabad, Lahore Campus, with over 20 years of experience in teaching, research, and academic leadership. He holds a PhD in Electronics Science and Technology from the North University of China, along with multiple master's degrees: an MS in Electrical and Computer Engineering from the University of Illinois Urbana-Champaign, USA, an MS in Computer Engineering from UET Taxila, Pakistan and MSc Physics from the University of the Punjab. Research Interest: nanotechnology, artificial intelligence, data science, signal processing and sustainable energy

How to cite this paper: Clive Ebomagune Asuai, Gabriel Ogbogbo, Houssem Hosni, Muhammad Ibrahim Khan, "The Chromatic Gradient Anomaly Network (CrGAN): Exploiting Second-Order Spatiotemporal Inconsistencies for Deepfake Video Detection", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.16, No.2, pp. 139-164, 2026. DOI:10.5815/ijwmt.2026.02.10