International Journal of Electrical, Electronics and Computer Systems
ISSN: 2347-2820
Volume 15 Issue 1s (2026) | Pages: 208-211

A Survey on AI-based Deep Fake Detection for Human Face Images and Videos

Tejas Ahirrao[1], Atharv Bhondave[2], Atharva Bhutkar[3], Siddhi Neharkar[4], Smitha Sapkal[5]

[1,2,3,4,5] Department of Computer Engineering, Genba Sopanrao Moze College of Engineering, Pune
Email: tejasahirrao1@gmail.com, atharvbhondave@gmail.com, atharvabhutkar0@gmail.com, Pune siddhineharkar1111@gmail.com, Pune sapkalss.2025@gmail.com

Peer Review Information

Type: *Article*
Received: *13 February 2026*
Revised: *14 March 2026*
Accepted: *15 April 2026*
Published: *21 May 2026*

Abstract

Technology for making and changing multimedia stuff has come a long way, and now we can create visuals that look really real. DeepFake tech uses these deep learning models to mess with faces or make new ones, and it's so good that telling real from fake is tough sometimes. I think that's part of why it's exciting but also scary. There are good sides to it, like in movies or TV shows where they improve effects, or even video games to make things look better. But people are using it badly too, for spreading else info or pretending to be famous people, which can cause big problems. To fight that, researchers are working on ways to detect DeepFakes, mostly with deep neural networks. Basically, DeepFakes are just media that's been changed or created by training these models to swap or add visual bits, especially faces. This paper looks at different detection methods for images and videos of faces, sorting them by how they detect, what techniques they use, and how well they work in tests. It also covers how DeepFakes are made in the first place, putting them into five main groups. I am not totally sure about all the details there, but it seems important to understand both sides. Datasets for DeepFakes are another thing they review, looking at what's common and how theyve gotten better or more varied lately. One big issue is making detection models that work on new kinds of fakes they haven't seen before, which sounds tricky. Overall, the survey points out challenges in creating and spotting these things, and some open questions that need more work. Hopefully, this helps push forward better deep learning ways to catch DeepFakes in faces and videos, though it might take time to get really reliable.

Keywords: Deep Fake Detection; Artificial Intelligence; Human Face Images; Deep Learning; Face Manipulation Detection; Video Forensics.

How to Cite This Article

Ahirrao, T., Bhondave, A., Bhutkar, A., Neharkar, S., & Sapkal, S. (2026). A Survey on AI-based DeepFake Detection for Human Face Images and Videos. *International Journal of Electrical, Electronics and Computer Systems*, 15(1s), 208-211.

Introduction

Digital media has blown up so much lately, with all these social sites, video apps, and messaging things everywhere. People share images and videos all the time, and they feel like solid proof of whats happening in the world. But verifying if theyre real is getting tougher. I mean, editing photos or splicing videos has been around for years, like with special effects in movies. Now though, deep learning stuff has changed everything, making fake visuals way more realistic and easier for anyone to do without much know how. DeepFakes are probably the scariest part of this. They use these neural networks, like GANs or autoencoders, to mess with faces in videos or pictures. The models train on tons of data and can swap someones face onto another, or make fake expressions, or even create whole new faces from scratch. It seems like the old signs of fakes, you know, weird lighting or jerky movements, are fading away. Theyre so subtle now that it's hard for people to notice without help.

This spread of DeepFakes is worrying a lot of areas, from social media to news, politics, cops, and forensics work. Bad uses include fake speeches to trick voters, or pretending to be someone for scams, or making nonconsensual explicit videos. It even hits security systems like face recognition. All this shakes trust in what we see online, messes with privacy, and could stir up bigger problems in society or elections. And since theyre so simple to make and share fast, the damage can spread quick in big online spaces. Traditional ways to spot fakes arent cutting it anymore as these techniques get better. Old methods use handpicked features or watermarks, but they can't handle how deep networks create stuff that looks totally natural. They just werent built for this adaptive kind of forgery. So theres this pushes for better detection that can keep up.

Researchers are working on automated tools now, using deep learning classifiers or looking at time-based patterns in videos, or even body signals like heartbeats from the footage. Some pull out tiny artifacts to tell real from fake. Results look good on test sets, but many struggle with new tricks, or get fooled by attacks, or lose accuracy after compression. It feels like were still catching up. This survey tries to lay out the generation methods, detection approaches, datasets out there, and how to evaluate them. It goes through whats working and whats not, pointing out gaps like needing systems that generalize better and explain themselves for real use in security or investigations. I think the biggest challenge is making something robust enough for everyday threats, but it's not fully clear how yet.

Literature Review

Early work in multimedia forensics mostly focused on catching basic image and video edits—things like copy-move forgery, splicing, resizing, or weird compression traces. Back then, researchers leaned on handcrafted features. They'd dig into pixel stats, frequency analysis, camera sensor noise, and telltale compression marks. Classic signal processing tricks like DCT analysis, PRNU, and block-wise correlation were the tools of choice. Honestly, these methods worked pretty well for old-school Photoshop jobs. But they assumed a world before DeepFakes—before AI started creating super-realistic fakes that don't leave the same obvious fingerprints.

Now, DeepFake tech has changed the game. Deep neural networks crank out fake content by learning patterns from huge datasets. The result? DeepFakes look seamless, with smooth textures and hardly any of the low-level artifacts traditional forensics relied on. When you throw in extra editing—compression, smoothing, resizing—the old methods struggle even more. They just can't spot the difference between a real image and a DeepFake anymore.

Because of this, researchers have shifted to DeepFake-specific detection techniques. CNN-based models dominate here. They're great at picking up on tiny texture inconsistencies, color screw-ups, weird blending around faces, and those hard-to-spot GAN fingerprints. Networks like XceptionNet, EfficientNet, ResNet, and Inception get trained and fine-tuned for this job, and they usually nail it on benchmarks. Some teams even mix in features from steganalysis to make their models tougher against compression and noise.

But images are only part of the story. When it comes to video, just looking at single frames

isn't enough. DeepFake generators are getting so good that frame-by-frame, everything looks legit. So, researchers started using temporal modeling—basically, looking at how things change over time. RNNs, LSTMs, GRUs, and 3D CNNs dig into sequences of frames to catch glitches like flickering, jitter, stiff or unnatural motion, or faces that just don't move like real ones. The idea is to spot those moments where AI messes up the flow.

On top of that, people have started looking for clues in human behavior and physiology. Think odd eye blinks, impossible head turns, gaze that doesn't match, lip-sync fails, or even heart rate signals hidden in tiny skin color changes. These features are tough for AI to fake and add another layer of security.

Still, even with all this progress, DeepFake detectors hit a wall when it comes to generalizing. They learn from one set of fakes, but when you test them on new DeepFake methods or real-world stuff pulled off the internet, their accuracy drops—sometimes a lot. It's a big disconnect between lab results and what actually works in the wild. Plus, most deep learning models are black boxes; they don't explain their decisions, which makes them hard to trust and easy targets for adversarial attacks. So, the field's facing a real challenge: building DeepFake detectors that are not just accurate, but also robust, explainable, and ready for anything that comes next.

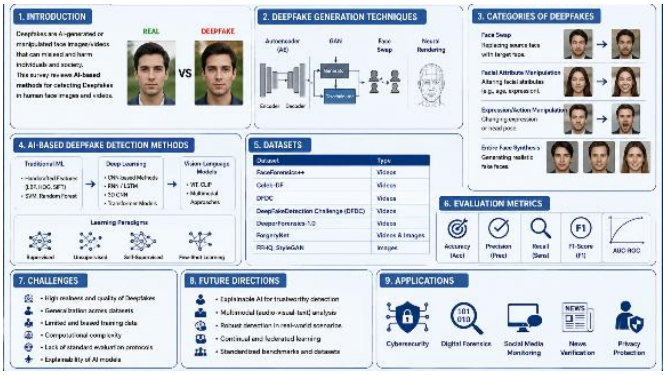

Fig 1. *Survey Framework of AI-Based Deepfake Detection for Human Face Images and Videos*

Problem Gap & Motivation:

DeepFake detection has come a long way, but there are still some big gaps researchers haven't closed yet. One of the main problems is that most models lean too heavily on the datasets they're trained on. Usually, these datasets are pretty limited—they only cover a handful of manipulation methods. So, when a model runs into a DeepFake technique it's never seen before, its performance drops a lot. There are also the issue of adversarial attacks and all the basic stuff people due to media, like compressing videos, resizing, adding blur, or injecting noise. These tweaks are common when people share videos online, and they can throw off even the best detectors.

A lot of current methods only look at spatial or temporal features, not both, which means they miss parts of how DeepFakes actually get made. On top of that, there just aren't enough big, diverse, well-annotated real-world datasets for training and testing. All these challenges point to one thing: there's a real need for DeepFake detectors that are tough, flexible, and can actually explain their decisions—something that can keep up as DeepFake tech keeps changing and move smoothly into the real world.

Proposed direction & Scope:

DeepFake detection research needs to break out of its own little bubble. Right now, most models are good at catching fakes only when the test data looks a lot like the data they trained on. But real-world DeepFakes are a mess—they're made with all sorts of tools, new architectures pop up constantly, and post-processing tricks keep getting sneakier. So, if we want to actually stop DeepFakes outside the lab, we need detection systems that can handle the unknown. That's the real challenge. One way forward is to design deep learning models that don't just focus on one thing. Instead, they should pull information from all over: spatial, temporal, and even physiological cues. Spatial features spot weird textures, blending mistakes, or telltale GAN artifacts in

single frames. Temporal features watch how faces move from frame to frame—does the smile look natural, does the blink rate make sense, are the expressions smooth? Then you've got physiological and behavioral signals: stuff like eye blinks, lip sync, head movement, gaze direction, or heart rate. Generative models still struggle to fake all of those at once. If you combine these clues in one system, you get a detector that's a lot harder to fool.

There's another big issue: labeled data. Most current detectors rely on huge, carefully annotated datasets, but labeling DeepFakes takes forever and costs a lot. Plus, there's always some new DeepFake method right around the corner, so your dataset is never truly up to date. That's where self-supervised and unsupervised learning come in. These methods dig out useful patterns from unlabeled or weakly labeled data. They don't wait for someone to tag every frame as real or fake—they learn from the data itself and spot stuff that just doesn't look natural. That makes them much better at catching new types of DeepFakes, especially in the wild.

Generalization is still tough, though. Video quality changes, compression artifacts, weird lighting, different cameras, and cultural differences all mess with detection accuracy. To deal with this, research needs to take domain adaptation and cross-dataset evaluation seriously. Domain adaptation helps models adjust to new settings, while cross-dataset tests show how well a detector works outside its comfort zone. If we want these systems to actually work on real platforms, we can't skip this step.

And DeepFakes aren't just about faces anymore. Now we have audio-visual DeepFakes—fake voices that match fake faces—plus text-to-video synthesis and full-on multimodal DeepFakes that mix video, audio, and text into totally convincing fakes. Fighting these requires detectors that can handle all these different signals at once, not just look for a weird eyebrow in a video clip.

Finally, there's the trust problem. Most deep learning detectors are black boxes. That's a problem for investigators, lawyers, or anyone who needs to explain in court why a video is fake. Detectors need to show their work—highlight the manipulated spots, flag strange timing, or point out physiological oddities. And speed matters. Social media, livestreams, surveillance—these places need answers fast, not in a few hours. Real-time or near real-time detection isn't just nice to have; it's essential if we want to stop DeepFakes before they spread.

Challenges & Open Issues:

DeepFake detection keeps turning into a cat-and-mouse game. As soon as detection tools catch up, new DeepFake generators come out that look even more convincing—fewer obvious glitches, smoother videos, all that. So, the old trick of looking for telltale artifacts doesn't work as well anymore. And let's be honest, the datasets out there aren't great. They're often limited, not diverse, and don't capture everything you'll see in the wild, which means detection models can miss a lot when faced with something new. It gets trickier. Some people deliberately tweak DeepFakes to slip past detectors, throwing in adversarial attacks. On top of that, you've got to spot tiny edits or just parts of a video that have been changed. Working with blurry, low-res, or super-compressed clips? That's a headache too. Plus, there's the big question of privacy and ethics when collecting data for training. Detecting unknown manipulation methods, building systems that actually scale up, and getting all this to work fast enough for real-time use—these are tough problems. But tackling them matters. Without solid DeepFake detection, it's just too easy for fake content to slip through and mess with the trust we have in digital media.

References:

1. Goodfellow et al., "Generative Adversarial Nets," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2014, pp. 2672–2680.
2. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), Hong Kong, Dec. 2018, pp. 1–7.
3. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 46–52.
4. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1–11.
5. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1–6.
6. Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE Int. Conf. Image Processing (ICIP), Athens, Greece, 2018, pp. 251–255.
7. B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020.
8. Y. Li et al., "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3207–3216.
9. L. Jiang et al., "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 2889–2898.
10. Malik et al., "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18778, 2022.