A Systematic Literature Review on Datasets for Deepfake Images in Smart Cities

ANAND M K

242CS008

National Institute of Technology Karnataka, Surathkal

Abstract

The management of deepfake image datasets becomes essential for developing smart city environments. This review investigates modern datasets for deepfake image development and recognition while exploring their applications within urban domains including security monitoring and fraud prevention and misinformation management. Deepfake technologies provide valuable benefits to specific fields yet they become dangerous to security and privacy when used improperly.

This review examines deepfake dataset creation methods alongside image generation technologies and preprocessing techniques and detection methods utilized by researchers. The analysis of prominent datasets DFDC, DeeperForensics, FaceForensics++, DeepfakeTIMIT, and UADFV examines their data acquisition approaches together with annotation practices and authenticity and diversity verification methods. This review examines deepfake generative techniques which use Generative Adversarial Networks (GANs) and Autoencoders to evaluate their effects on dataset quality and realism.

This section examines preprocessing methods including data augmentation together with feature extraction and artifact analysis to determine their impact on detection accuracy. The review examines deepfake detection methods based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and hybrid models to demonstrate their ability to detect manipulated images. The research community encounters three main obstacles which include dataset biases and generalization problems and the requirement for real-time processing in smart city applications.

This review presents an extensive analysis of deepfake image dataset availability and detection methods while outlining essential research needs for the future. The research insights serve as a foundation for scientists to create stronger and more scalable countermeasures against deepfake technology within smart cities.

1. Introduction

The fast progress of artificial intelligence and deep learning technologies enabled deepfake techniques to spread widely for creating highly believable synthetic images and videos. These technologies deliver many beneficial uses across entertainment and education and accessibility but simultaneously create major security risks and privacy breaches and trust breakdowns in smart city systems. Deepfake-generated images create security risks through identity theft and misinformation spread and unauthorized system access because effective detection tools remain a critical necessity.

Smart cities use IoT, AI and big data analytics with advanced technologies to improve their urban infrastructure and enhance life quality. The implementation of deepfake technologies within smart city systems creates new security risks which threaten essential services including surveillance monitoring and authentication protocols and public information dissemination. Surveillance systems with manipulated images create false identification problems which disrupt law enforcement operations while damaging public confidence in smart city programs.

The development of deepfake detection methods by researchers uses machine learning algorithms together with robust datasets to tackle these challenges. High-quality deepfake image datasets serve as fundamental components for training and evaluating detection models. Research efforts receive support from five datasets including DFDC, DeeperForensics, FaceForensics++ and DeepfakeTIMIT and UADFV. The detection algorithms experience direct performance impacts from differences in dataset creation methodologies and data preprocessing approaches and annotation strategies.

This systematic review examines deepfake image datasets through an analysis of their creation methods alongside deepfake generation technologies and preprocessing methods and detection approaches and researcher challenges. The review points out current research gaps while recommending future directions to improve both the reliability and scalability of deepfake detection systems for smart cities.

2. Method

2.1. Literature Search Procedure

In order to carry out a systematic literature review on deepfake image datasets for smart city applications, the first step is to conduct a comprehensive literature search. This process involves using various academic databases, scientific journals, repositories, and other reliable sources of information. A thorough search is conducted using relevant keywords covering a specific period to ensure that up-to-date and relevant studies are included in the review.

2.2. Research Problems

After conducting the literature search, studies are selected based on their relevance to the review objectives. The focus is on studies within fields such as smart city security, surveillance systems, deepfake detection, and dataset creation methodologies. Studies are considered based on experimental designs, surveys, and observational studies related to deepfake technology.

2.3. Search Strategy

Once relevant studies are identified, the next step involves a detailed evaluation of the literature. The evaluation process includes a critical analysis of each selected publication based on factors such as study design, dataset creation techniques, preprocessing methods, and the effectiveness of detection approaches.

The digital database search for articles was conducted using platforms such as ScienceDirect and IEEE Xplore. The search string was developed following these steps:

1. Identify search terms from the research objectives.
2. Derive relevant search terms from the formulated research questions.
3. Extract search terms from the title, abstract, and keywords of the studies.
4. Identify synonyms, alternative spellings, and related terms for comprehensive coverage.
5. Construct a sophisticated search query using Boolean operators such as AND and OR.

Here are the search strings used: *"deepfake detection" AND "image dataset" AND "smart city surveillance" OR "deepfake fraud prevention."*

2.4. Selection of Studies

Following the search strategy, relevant data are extracted from each selected study to obtain key insights. The data extraction process involves collecting essential information such as authors, year of publication, research design, dataset generation methods, preprocessing techniques, and detection algorithms used. The extracted data are systematically organized to facilitate further analysis.

3. Research Results

3.1. DeepFake Detection Challenge (DFDC) Dataset

The DeepFake Detection Challenge (DFDC) dataset stands as the largest and most diverse collection of deepfake videos that researchers use for deepfake detection work. The dataset emerged from a Facebook-

organized competition to support research into detection models that identify manipulated media. The dataset contains over 100,000 total video clips, sourced from 3,426 paid actors, that include real videos alongside deepfake videos, which were created through multiple production methods using multiple actors. The extensive diversity and large size of this dataset provide researchers with a powerful tool to develop deepfake detection technologies for smart city surveillance systems [1].

The DeepFake Detection Challenge (DFDC) dataset includes multiple deepfake generation methods to create diverse and challenging manipulated content. One such method is the Deepfake Autoencoder (DFAE), which utilizes a convolutional autoencoder with a shared encoder and two separately trained decoders, each dedicated to one identity in the swap. Another approach, the MM/NN face swap method, employs a custom frame-based morphable-mask model that computes facial landmarks in both target and source images, morphing the source pixels to match the target landmarks and using blending techniques and spherical harmonics for illumination adjustments. The Neural Talking Heads (NTH) method generates realistic talking head videos using a meta-learning approach with two stages: meta-parameter learning and fine-tuning with limited training images. Additionally, the FSGAN method leverages generative adversarial networks (GANs) for face swapping and reenactment, ensuring pose and expression variations are accurately maintained by incorporating adversarial loss and Poisson blending techniques. These varied methodologies enhance the dataset's diversity, making it an effective resource for developing robust deepfake detection models [1].

After inference, all methods generated a 256x256 cropped image of the face, but some methods did not capture details like hair or background. To address this, the face was re-blended onto the original full-resolution frame using Poisson blending, with a mask that extended to the forehead region to avoid artifacts such as "double eyebrows." Augmentations were applied to the videos in both the public Kaggle test set and the final evaluation set. These included two types: (1) Distractors, which overlay objects like images, shapes, and text, and (2) Augmenters, which apply geometric, color, and framerate transformations. Approximately 70% of the videos were augmented with effects like Gaussian blurring, contrast adjustment, flipping, and noise addition, with grayscale conversion being used only in the final evaluation set [1].

In the research [2], the proposed convolutional-transformer architectures, Efficient ViT and Convolutional Cross ViT, are applied to deepfake detection on the DFDC dataset. Faces are pre-extracted from video frames using the state-of-the-art MTCNN detector, and the models are trained to classify the faces as real or manipulated. During inference, the models process each face individually, and the final classification is derived by aggregating the results across time and multiple faces in the video. These methods effectively address the challenges presented by the DFDC dataset, which includes diverse deepfake generation techniques and variations in video content.

3.2. FaceForensics++ Dataset

The FaceForensics++ dataset serves as a popular resource for researchers who study facial manipulation detection. The dataset serves two purposes: it establishes standardized evaluation criteria for detection methods while offering extensive research material through its large-scale dataset. The proposed automated benchmark system for facial manipulation detection comes with a public dataset that contains hidden test data for unbiased evaluation[3].

The FaceForensics++ dataset includes a variety of manipulation methods based on both classical computer graphics and deep learning approaches. FaceSwap is a graphics-based method that extracts and transfers the face region using sparse facial landmarks, fitting a 3D template model with blendshapes. DeepFakes, a deep learning-based method, refers to face replacement techniques that use neural networks, notably implemented in tools like FakeApp and faceswap. Face2Face utilizes facial reenactment to transfer expressions from a source video to a target video while preserving identity, and NeuralTextures applies a neural texture-based rendering approach to achieve photorealistic facial reenactments through adversarial learning. These diverse manipulation methods contribute to the dataset's complexity and make it valuable for training and evaluating deepfake detection models[3].

The FaceForensics++ dataset implements postprocessing to improve realism through the simulation of video quality degradation that occurs on social media platforms and video-sharing websites. The widespread use of H.264 codec for encoding videos means that raw uncompressed videos are almost never found online. The dataset undergoes this step to match real-world compression artifact conditions which affect detection performance. Two levels of compression are applied: The dataset contains two compression levels: high-quality (HQ) with a quantization parameter of 23 for minimal visual degradation and low-quality (LQ) with a quantization parameter of 40 that produces noticeable quality loss.

The FaceForensics++ dataset employs various detection methods to identify manipulated facial images, categorized into handcrafted and learned feature-based approaches. Handcrafted detection relies on steganalysis features, where co-occurrences of pixel patterns are extracted from high-pass filtered images and used to train a linear Support Vector Machine (SVM)[4]. The method demonstrates high accuracy when processing raw images yet it faces challenges when analyzing compressed video content. In contrast, learned feature-based detection leverages deep learning techniques, including convolutional neural networks (CNNs) trained on the dataset. Several architectures are evaluated, such as models that adapt handcrafted steganalysis features into CNN frameworks, networks designed to suppress high-level content, and well-known architectures like MesoInception-4[5] and XceptionNet[6]. Furthermore, detection methods are also tested against GAN-based manipulations, with results indicating that neural texture-based forgeries present greater chal-

lenges due to their diverse artifact patterns compared to other methods with more consistent post-processing artifacts.

3.3. DeeperForensics-1.0 Dataset

DeeperForensics-1.0 is the largest face forgery detection dataset, with 60,000 videos and 17.6 million frames, designed to improve detection robustness through diverse real-world perturbations. Fake videos are generated using a novel face-swapping framework, producing superior quality validated by user studies. Its hidden test set includes highly deceptive videos, making it a vital resource for advancing forgery detection in smart city applications[7].

The DeeperForensics-1.0 dataset emphasizes source face videos to improve face swapping robustness under diverse conditions. It includes 50,000 high-resolution videos (1920×1080) with 12.6 million frames, featuring 100 actors of varied genders, ages (20–45), skin tones (white, black, yellow, brown), and nationalities (26 countries). The data collection took place in a professional indoor setting under nine different lighting conditions and seven camera perspectives to achieve high-quality data with diverse variations. The dataset provides superior diversity and scale compared to existing datasets which positions it as a fundamental benchmark for face forgery detection research [7]. .

To tackle low visual quality problems of previous works, three key requirements are considered for a high-fidelity face swapping method: 1) scalability to generate high-quality videos, 2) addressing face style mismatch issues, and 3) ensuring temporal continuity of generated videos. DeepFake Variational Auto-Encoder (DF-VAE) is a learning-based face swapping framework that includes a structure extraction module, a disentangled module, and a fusion module. It disentangles structure (expression and pose) from appearance (texture, skin color) and introduces masked adaptive instance normalization (MAdaIN) to address style mismatches, focusing on face areas while preserving the original background [7]. DF-VAE generates 1,000 raw manipulated videos by swapping faces from 100 identities onto 1,000 target YouTube videos. The method's scalability and multimodality reduce model training and data generation time by a factor of five compared to traditional Deepfake methods, without compromising quality. To ensure diversity, the dataset includes various distortions such as color saturation changes, Gaussian blur, and video compression—applied at different intensity levels. These perturbations simulate real-world video conditions and improve the robustness of the dataset for face forgery detection research.

Deepfake detection methods applied to the DeeperForensics datasets include both image-level and video-level techniques. For image-level detection, methods like Xception-Net are used, while video-based methods such as C3D[8] and TSN[9] capture spatiotemporal features to detect forgeries across frames. Additionally, face detection methods like MTCNN and RetinaFace are employed to accurately localize and track faces,

enhancing the robustness of the detection models in identifying manipulated videos.

3.4. DF-TIMIT Dataset

The DF-TIMIT dataset is a collection of Deepfake videos created by applying generative adversarial networks (GANs) to videos from the VidTIMIT database. This dataset was developed to address the increasing challenges posed by Deepfake videos, which involve the manipulation of faces in video content. The DF-TIMIT dataset consists of 640 videos in total, with 320 videos generated with low visual quality and 320 videos with high visual quality[10].

The DF-TIMIT dataset is created by selecting 16 subject pairs from the VidTIMIT database, where each pair consists of subjects with similar visual features, such as mustaches or hairstyles. Using a GAN-based face-swapping algorithm, videos are generated by swapping faces between subjects in each pair. Two versions of the dataset are created: low quality (LQ) and high quality (HQ). For the LQ model, face regions are generated with a size of 64x64, and videos are trained using 200 frames extracted at 4 fps. For the HQ model, a larger image size of 128x128 is used with 400 frames extracted at 8 fps. Different blending techniques are applied in each model, with the LQ model using a CNN-based face segmentation algorithm for blending, while the HQ model employs facial landmark alignment and histogram normalization for better face integration[10].

For Deepfake detection on the DF-TIMIT dataset, several detection methods were applied, including both audio-visual and image-based systems. The lip-syncing detection system, which analyzes the synchronization between lip movements and audio, was tested on the dataset. This system extracts audio features using Mel-frequency cepstral coefficients (MFCCs)[11] and visual features based on distances between mouth landmarks. These features are then processed using Principal Component Analysis (PCA) and classified with a Long Short-Term Memory (LSTM) network to differentiate between genuine and tampered videos. Additionally, image-based systems such as Pixels+PCA+Linear Discriminant Analysis[13] (PCA+LDA), Image Quality Measures (IQM)[12]+PCA+LDA, and IQM+Support Vector Machine (SVM) were employed. These systems utilize raw image data or image quality measures (IQM) to identify inconsistencies in the Deepfake videos. The IQM+SVM system demonstrated promising results, though high-quality videos posed a greater challenge due to the advanced face-swapping techniques used in the dataset.

3.5. UADFV Dataset

The UADFV (Unconstrained Audio-Visual Deepfake Video) dataset comprises 49 real videos and their corresponding 49 Deepfake videos, as utilized in prior research. Each video in this dataset has an average duration of approximately 11.14 seconds and a resolution of 294×500 pixels. Additionally, the dataset is

complemented by a subset from the DARPA MediFor GAN Image/Video Challenge, which includes 241 real images and 252 Deepfake images. This dataset is a foundational resource for studying and benchmarking Deepfake detection methods[14].

The UADFV dataset utilizes a method involving 3D head pose inconsistencies to detect DeepFake content. This approach trains SVM classifiers using features derived from differences in head poses estimated from the entire set of facial landmarks and those in the central face regions. Facial landmarks are extracted using the DLib software package, and head poses are calculated with standard 3D facial landmark models from OpenFace2. Rotation matrices and translation vectors are computed for both the central and full face regions, and the differences are flattened into feature vectors. These standardized feature vectors serve as input for classification, effectively identifying discrepancies introduced by DeepFake manipulations[14].

4. Conclusion

This systematic review highlights the evolution of DeepFake datasets and their implications for smart cities. Early datasets like UADFV and DF-TIMIT provided small-scale resources for initial detection efforts, while second-generation datasets like FaceForensics++ introduced higher quality and ethical considerations but remained limited in diversity. Third-generation datasets, such as DeeperForensics-1.0 and the DFDC, advanced scale, diversity, and ethical compliance, supporting more robust detection systems.

Table 1 presents a quantitative comparison of these datasets, illustrating their differences in size, scope, and subject diversity. These datasets are essential for developing reliable DeepFake detection methods, which are crucial for preserving the integrity of digital media in smart cities. Continued innovation in dataset creation is necessary to address the growing challenges posed by DeepFake technologies.

Table 1: Quantitative comparison of various Deepfake datasets

Dataset	**Unique fake videos**	**Total videos**	**Total subjects**
DF-TIMIT	640	960	43
UADFV	49	49	49
FF++ DF	4,000	5,000	?
DeeperForensics-1.0	1,000	60,000	100
DFDC	104,500	128,154	960

References

[1] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer, *The DeepFake Detection Challenge (DFDC) Dataset*, 2020.

[2] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi, *Combining EfficientNet and Vision Transformers for Video Deepfake Detection*, 2022.

[3] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Niessner, *FaceForensics++: Learning to Detect Manipulated Facial Images*, 2019.

[4] J. Fridrich and J. Kodovsky, *Rich Models for Steganalysis of Digital Images*, IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, June 2012.

[5] Darius Afchar, Vincent Nozick, Junichi Yamagishi, Isao Echizen, *MesoNet: A Compact Facial Video Forgery Detection Network*, 2018.

[6] Francois Chollet, *Xception: Deep Learning with Depthwise Separable Convolutions*, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[7] L. Jiang, R. Li, W. Wu, C. Qian and C. C. Loy, *DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection*, 2020

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, *Learning spatiotemporal features with 3D convolutional networks*, ICCV, 2015.

[9] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, *Temporal segment networks: Towards good practices for deep action recognition*, CVPR, 2016.

[10] Pavel Korshunov and Sébastien Marcel, *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*, 2018.

[11] N. Le and J.-M. Odobez, *Learning multimodal temporal representation for dubbing detection in broadcast media*, in *Proceedings of the 2016 ACM on Multimedia Conference*

[12] J. Galbally and S. Marcel, *Face anti-spoofing based on general image quality assessment*, in *2014 22nd International Conference on Pattern Recognition* Aug 2014

[13] D. Wen, H. Han, and A. K. Jain, *Face spoof detection with image distortion analysis*, *IEEE Transactions on Information Forensics and Security*, April 2015.

[14] Xin Yang, Yuezun Li, and Siwei Lyu, *Exposing Deep Fakes Using Inconsistent Head Poses*,2018