Enhanced Deepfake Detection Using Frequency Domain Upsampling

Mamadou Dian Bah[a] and Mohamed Dahmane[b]
Computer Research Institute of Montreal (CRIM), Canada

Keywords: DeepFake Detection, Media Forensics, Fourier Upsampling.

Abstract: Recent advances in deep learning have dramatically reshaped the image processing landscape, specifically targeting the key challenge of detecting deepfakes in digital media. This study investigates the integration of Fourier or frequency domain upsampling techniques with deep learning models for the purpose of detecting deepfakes. Using the FF++ dataset as a benchmark for evaluating deepfake detection, our research rigorously evaluates the effectiveness of models such as Xception. This evaluation includes an in-depth exploration of various upsampling methods as well as the combination of spatial and frequency domain upsampling. The results reveal clear disparities in performance between the different models and techniques, and our experiments highlight the profound impact of the various upsampling and downsampling approaches on the accuracy of the resulting classification. Remarkably, combining the Xception model with upsampling and downsampling techniques increases detection accuracy by over 2%, while maintaining the constant input size inherent in the Xception architecture.

1 INTRODUCTION

The remarkable progress in artificial neural network (ANN)-based technologies and particularly in generative AI has played a crucial role in manipulating multimedia content. These advancements have made it increasingly feasible to generate highly realistic synthetic images, surpassing previous capabilities. In recent years, generative models have witnessed remarkable advancements in their ability to generate human-like natural language (Brown et al., 2020), high-quality synthetic images (Karras et al., 2020), and diverse human speech and music. These models find utility in various domains, such as image generation from text prompts and feature representation learning.

As a result, numerous captivating applications have emerged in the field of entertainment and education. Through all this capability, the term DeepFake has emerged and refers to multimedia content generated or altered by artificial intelligence models (Gomes et al., 2020; Lattas et al., 2020). Noteworthy examples include FaceApp (FaceApp, 2023), a popular application that leverages an autoencoder-decoder architecture to seamlessly swap faces between two images. DeepFaceLab (DeepfakeVFX.com, 2023), an enhanced framework specifically designed for face-swapping, pushes the boundaries of DeepFake technology. Another notable application is Face Swapping GAN (FSGAN) (Nirkin et al., 2019), an improved version of DeepFakes that employs Generative Adversarial Networks (GANs) to refine segmentation masks, resulting in remarkably higher quality output videos. Additionally, several other tools are employed for DeepFake content generation. DiscoFaceGAN (Deng et al., 2020), based on StyleGAN structure (Karras et al., 2019), generates synthetic face images of virtual individuals with distinct characteristics, while FaceShifter enables high-fidelity face swapping. The recent emergence of a new generation of models, the so–called Denoising Diffusion Models (DDMs), has raised great concern for the spread of fake data, as they proved capable of generating even more realistic and convincing fakes than their predecessors, Generative Adversarial Networks. Models like Stable Diffusion (Rombach et al., 2022) and DALL-E 2 (Ramesh et al., 2022) are some of the best image generators available and are renowned state-of-the-art Diffusion Models (DMs) that excel in text-to-image translation.

Given the significant threats posed by DeepFakes, such as the spread of misinformation, damage to reputation, and invasion of privacy, it is crucial to develop advanced technologies for detecting DeepFake

[a] https://orcid.org/0000-0002-0733-8587
[b] https://orcid.org/0000-0002-2670-1433

Bah, M. and Dahmane, M.
Enhanced Deepfake Detection Using Frequency Domain Upsampling.
DOI: 10.5220/0012473700003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 796-803
ISBN: 978-989-758-679-8; ISSN: 2184-4321

content (Karasavva and Noorbhai, 2021; Firc et al., 2023). Humans often struggle to identify manipulated videos, particularly in terms of spatial aspects, and lack the ability to pinpoint the specific locations and techniques used for manipulation. This highlights the necessity for high-quality algorithms capable of detecting manipulated videos on a large scale.

This study investigates the integration of Fourier or frequency domain oversampling techniques with deep learning models such as FourierUp (Zhou et al., 2022) for the purpose of detecting deepfakes. Our paper makes the following contributions:

- Illustration of the enhancement in results by employing spatial or frequency domain upsampling at the onset of the deep CNN model.
- Demonstration of improved model accuracy by initially upsampling the input image with Deep Fourier Up-Sampling, followed by downsampling in the spatial domain, while minimally impacting the architecture of the deep CNN.

The rest of the paper is organized as follows. Section 2 provides a literature review of commonly used algorithms for deepfake detection. Section 3 presents details on the proposed approach. Section 4 provides the results and discussions. Finally, Section 5 concludes with implications, limitations, and suggestions for future research.

2 RELATED WORKS

Deep learning techniques have emerged as the dominant approach for DeepFake detection, as evidenced by a comprehensive analysis of 122 studies conducted by Rana et al. (Rana et al., 2022). Approximately 77% of these studies employed deep learning models, specifically CNNs (Tariq et al., 2018) and Recurrent Neural Networks (RNNs) mostly used for fake videos detection. These models have shown great promise in effectively detecting DeepFake content.

In (Afchar et al., 2018) MeSoNet is a CNN architecture which is used to detect Face2Face and deepfakes manipulations. XceptionNet which uses depthwise separable convolutional layers with residual connections (Chollet, 2017) has given the best result in (Rossler et al., 2019) on Faceforensics++ (FF++) dataset.

Although the use of deep learning different approaches are used based on the knowledge of human face. The authors in (Haliassos et al., 2021) propose a method of detecting high-level semantic anomalies in mouth motion, leveraging the hypothesis that most video generators display a degree of high-level semantic irregularities near the mouth. In (Zhao et al., 2021a; Zhao et al., 2021b), fine-grained classification is applied to distinguish subtle differences in visual appearance and patterns. The authors of Fake-Buster (Hubens et al., 2021) addressed the issue of detecting face modification in video sequences using recent facial manipulation techniques. In (Ismail et al., 2021), the YOLO face detector is used to extract the face area from video frames, while the InceptionResNetV2 CNN is utilized to extract features from the detected face region.

However detecting DeepFakes in videos solely based on counterfeit images can be difficult due to the temporal features of videos and variation in frame resolution. In (Ranjan et al., 2020), the CNN-LSTM combo is used to identify and classify the videos as fake or real. FSSPOTTER (Chen et al., 2020), for instance, uses spatial and temporal clues to detect swapped faces in videos. These features are fed into the XGBoost, which works as a recognizer on the top level of the CNN network. Physiological signals are also used for DeepFake detection. DFT-MF (Elhassan et al., 2022) is a deepfake detection model that uses deep learning to detect deepfake videos by isolating, analyzing, and verifying lip/mouth movement. Eye blinking based signal detection is also used to determine if a video is real or a DeepFake generated. Deep Vision detects DeepFake videos by focusing on eye blink patterns. FakeCatcher (Ciftci et al., 2020) is a method that addresses the challenge of detecting DeepFakes by exploiting the fact that biological signals obtained from facial regions are not positionally and temporally well-preserved in the content of portrait videos.

3 METHOD

In the model shown in Figure 1 according to *Up* or *Up and Down* we explore the way to apply it in different configurations. The deep CNN used is Xception, since XceptionNet which uses depth-wise separable convolutional layers with residual connections (Chollet, 2017) has given the best result in (Rossler et al., 2019) in Faceforensics++ dataset. Here we decide to add one more module to its input in the aim to improve the deepfakes detection. This added module involves two potential approaches: solely performing upsampling (Up) or employing a sequence that combines upsampling and downsampling (Up and Down) on the input image. We explore different techniques of upsampling.

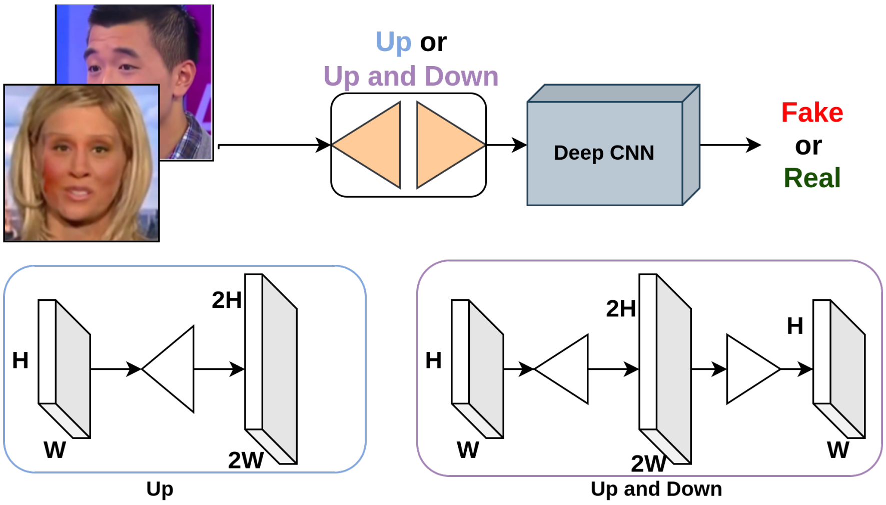

Figure 1: Flowchart of the proposed method: ”Up” denotes the application of solely the upsampling process to the input. ”Up and Down” signifies the sequential application of upsampling followed by downsampling to the input.

3.1 Bilinear Interpolation

Bilinear interpolation, a widely-used technique, offers a balance between computational efficiency and visual fidelity. Image resolution enhancement, or upsampling, is crucial in various domains such as medical imaging, satellite imagery analysis, and computer graphics. The bilinear interpolation method enabled us to estimate pixel values in the upsampled image by blending adjacent pixel intensities, resulting in smoother transitions and reduced aliasing effects compared to other interpolation techniques. This approach retained visual details while enhancing the resolution. In deepfake detection details are often imperceptible for the human eye.

Bilinear interpolation is a method used for estimating the value of a new pixel by considering the weighted average of its neighboring pixels. In the context of image upsampling, bilinear interpolation calculates the value of a new pixel within the enlarged image grid based on the surrounding pixels' values in the original image.

In essence, bilinear interpolation linearly interpolates in both the horizontal and vertical directions, combining contributions from adjacent pixels to estimate the value of the new pixel within the upsampled image.

3.2 Upsampling with Fourier Transform

Based on (Durall et al., 2019), it can be inferred that genuine and counterfeit images exhibit distinguishable spectral differences at higher frequencies. The authors in (Zhou et al., 2022) highlight the strong reliance of spatial up-sampling methods (e.g., interpolation, transposed convolution, and un-pooling) on local pixel attention. Despite the significant role of spatial up-sampling in multi-scale modeling, recent attention has shifted towards up-sampling in the frequency domain (Yang and Soatto, 2020; Zhou et al., 2022).

Although spatial upsampling holds significance in multi-scale modeling domain. In contrast to the spatial domain, the Fourier domain does not share the same scale invariant property and local texture similarity, and hence cannot implement up-sampling using the same techniques as the spatial domain. To address this, we employed techniques proposed in (Zhou et al., 2022) for upsampling images in the Fourier domain, as their approach has been validated in various computer vision tasks, such as object detection, image segmentation, de-raining, dehazing, and guided image super-resolution, demonstrating performance improvements.

We explored two of the 3 proposed Deep Fourier up-sampling variants, including i) periodic padding of magnitude and phase, ii) area up-sampling of magnitude and phase because they were found more robust.

3.2.1 Periodic Padding of Magnitude and Phase

This process involves performing periodic padding twice, both horizontally and vertically, over the phase and amplitude. Subsequently, the inverse Fourier transform is applied to project the interpolated elements back into the spatial domain.

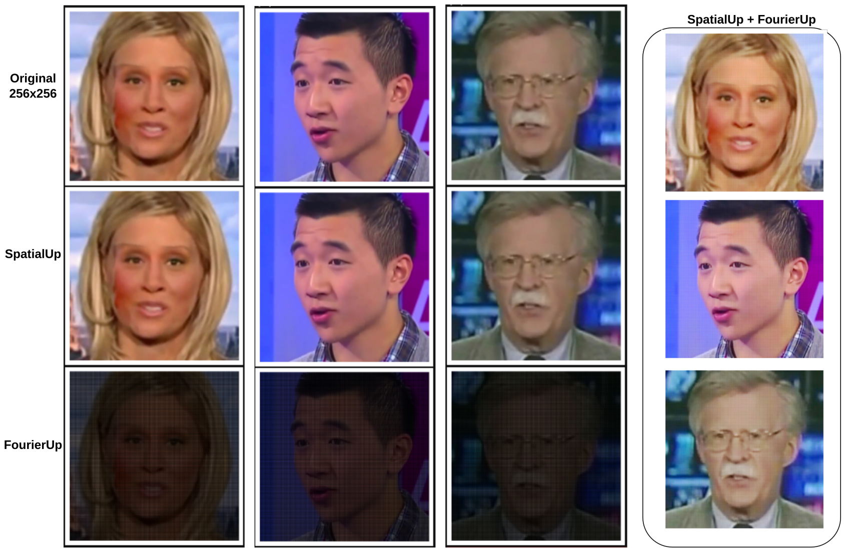

Figure 2: From top to down original input 256, upsampling by Bilinear interpolation (spatial upsampling) and Fouier UP with period padding in frequency domain (Frequency upsampling). The last column represents the sum between spacial and frequential domain upsampling.

3.2.2 Area Up-Sampling of Magnitude and Phase

Area up-sampling begins with area interpolation conducted over the phase and amplitude using a 2 × 2 area interpolation method with identical pixels. Following this, the inverse Fourier transform is employed to project the interpolated elements back into the spatial domain.

```
procedure DFU_PADDING($X$)
    Input: $X$ (input with shape $[N,C,H,W]$)
    Output: $Y$ (transformed output with shape $[N,C,2H,2W]$)
    $A,P \leftarrow$ FFT($X$)    ▷ A and P are the amplitude and phase
    $A_{pep} \leftarrow$ Periodic-Padding($A$)
    $P_{pep} \leftarrow$ Periodic-Padding($P$)
    $A_{pep} \leftarrow$ Convs_1x1($A_{pep}$)
    $P_{pep} \leftarrow$ Convs_1x1($P_{pep}$)
    $Y \leftarrow$ iFFT($A_{pep}, P_{pep}$)    ▷ Inverse Fourier transform
    return $Y$    ▷ Output shape: $[N,C,2H,2W]$
end procedure
```

Algorithm 1: Periodic Padding Upsampling.

For both periodic padding and area interpolation, the phase and amplitude undergo feeding into two independent convolution modules using a 1 × 1 filter after the up-sampling process.

Figures 2 and 3 present examples of images with differents upsampling process.

In this paper, we use the term ”FourierUp” to indicate upsampling in the frequency domain. Algorithms 1 and 2 demonstrate the Fourier upsampling process as proposed in (Zhou et al., 2022). Algorithm 3 illustrates the sequential application of our upsampling and downsampling method to the input data.

4 EXPERIMENT AND RESULTS

To assess the effectiveness of our proposed approach, we conducted extensive experiments on the FF++ dataset (Rossler et al., 2019). The dataset stands out as a large-scale facial manipulation dataset created through state-of-the-art video editing techniques. In particular, it combines classical computer graphics methods, such as Face2Face (Thies et al., 2016) and FaceSwap, with learning-based strategies, including deepfakes and NeuralTextures (Thies et al., 2019). Each of these methods was applied to 1000 high-quality YouTube videos downloaded from YouTube, carefully selected to ensure that the subjects faces were nearly frontal and free from occlusions. These video sequences consist of at least 280 frames each. Ultimately, this dataset comprises over 1.8 million images obtained from 4000 manipulated videos.

To ensure a fair comparison, in our experiment we adopted the evaluation protocol defined in (Bonettini et al., 2021). We used similar splits, selecting 720 videos for training, 140 for validation and 140 for test from the pool of original sequences taken from YouTube. The corresponding fake videos are assigned to the same split. We primarily focus on the

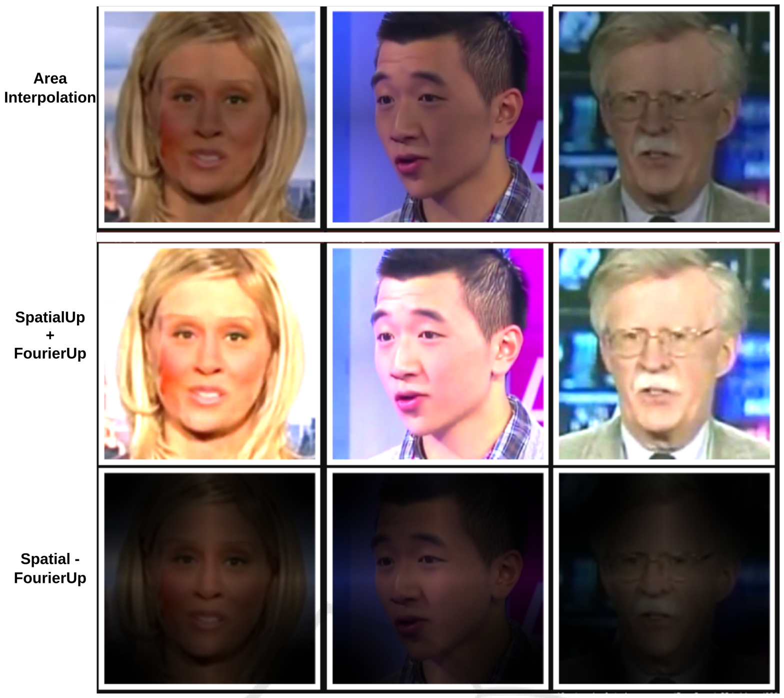

Figure 3: Example of upsampling using Area upsampling of magnitude and phase. The last row shows the difference between upsampling with bilinear interpolation and upsampling using Area up-sampling of magnitude and phase.

Table 1: Performance Evaluation on FF++ Dataset. FourierUp and SpatialUp use Xception. P-Padding and Area-Inter means respectively periodic padding and area interpolation. AUC is in percentage (%).

Up/Down	Model	Up	Down	Size	AUC
-	Xception*	-	-	256x256	93.69
-	Xception (Bonettini et al., 2021)	-	-	256x256	95.05
FourierUpDown (ours)	Xception*	P-Padding	Bilinear	256x256	95.82
SpatialUp (ours)	Xception	Bilinear	-	512x512	**97.28**
FourierUp (ours)	Xception	P-Padding	-	512x512	96.99
FourierUp (ours)	Xception	A-Inter	-	512x512	94.93

```
procedure DFU_AREAINTERPOLATION($X$)
    Input: $X$ (input with shape $[N, C, H, W]$)
    Output: $Y$ (transformed output with shape $[N, C, 2H, 2W]$)
    $A, P \leftarrow$ FFT($X$)    ▷ A and P are the amplitude and phase
    $A_{aip} \leftarrow$ Area-Interpolation($A$)
    $P_{aip} \leftarrow$ Area-Interpolation($P$)
    $A_{aip} \leftarrow$ Convs_1x1($A_{aip}$)
    $P_{aip} \leftarrow$ Convs_1x1($P_{aip}$)
    $Y \leftarrow$ iFFT($A_{aip}, P_{aip}$)    ▷ Inverse Fourier transform
    $Y \leftarrow$ Area-Cropping($Y$)
    $Y \leftarrow$ Resize($Y$)
    return $Y$    ▷ Output shape: $[N, C, 2H, 2W]$
end procedure
```

Algorithm 2: Area Interpolation Upsampling.

```
procedure UPDOWN($X$)
    Input: $X$ (input tensor)
    Output: $xn$ (output tensor)
    $x1 \leftarrow x$
    $x2 \leftarrow$ SpatialUP($x1$)
    $x3 \leftarrow$ Up($x1$)    ▷ Upsampling with Up method
    $xm \leftarrow x2 + x3$
    $xn \leftarrow$ Convs_1x1($xm$)    ▷ Applying convolution
    $xn \leftarrow$ Down($xn$)    ▷ Down sampling with Down method
    $xn \leftarrow$ Convs_1x1($xn$)    ▷ Applying convolution again
    return $xn$    ▷ Final output
end procedure
```

Algorithm 3: Up and Down.

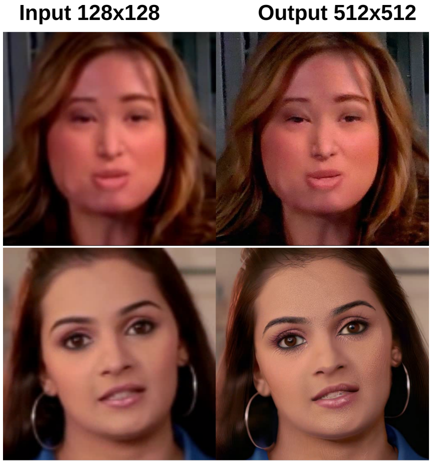

Figure 4: Upsampling using stable-diffusion-x4-upscaler (Rombach et al., 2022) with 50 DDIM steps. First row without prompt and second row with prompt: *best quality and high resolution of a human face*.

subject face region for analysis. We use the BlazeFace extractor for preprocessing, extracting the best-confidence face from each frame. Our network input image shape is 256x256. During training and validation, we enhance the model robustness with data augmentation, including downscaling, flipping, brightness, contrast, hue, saturation adjustments, noise addition, and JPEG compression. A total of 230302 frames were extracted from the dataset for training, 26879 frames were set aside for validation and 26879 for testing purposes. During experiments, we utilized the Adam optimizer with an initial learning rate of 0.0001 and binary cross-entropy loss. Batch size of 32 is used and the model is validated every 500 iteration on 6000 sample randomly selected from the validation set. Models are trained about 100k iterations. To evaluate the performance of our proposed method, we conducted various tests. We compared our results using pretrained Xception and perform cross-data validation. During the tests, we evaluated two Xception models: the one labeled simply as Xception* refers to a model fine-tuned directly from the deep-learning library ”timm”, while the Xception model is the one obtained from the GitHub repository of (Bonettini et al., 2021).

The three approaches we tested include upsampling with bilinear interpolation in spatial domain (SpatialUP), upsampling with two versions of FourierUp with padding and area interpolation of magnitude and phase.

Table 1 represents the performance evaluation on the FF++ Dataset, specifically assessing different models and techniques for upsampling and downsampling images. The table presents details regarding the models used, their associated techniques (such as FourierUp and SpatialUp), the methodology for upsampling and downsampling (e.g., Padding, Bilinear), the resulting image size (e.g., 256x256, 512x512), and the corresponding Area Under the Curve (AUC) values. Each method effectiveness is measured based on the AUC metric, reflecting their capability in distinguishing features within the dataset. The performance assessment on the FF++ dataset, showcasing the effectiveness of various models and techniques for image enhancement and manipulation detection. The Xception model achieves an AUC of 95.05 on a 256x256 resolution, surpassing Xception* model by a significant margin, which scored 93.69. However, the method employing upsampling followed by downsampling outperforms both Xception* models with the same resolution achieving an AUC of 95.82. Among the proposed approaches, the Fourier-based techniques, specifically the method employing Periodic Padding (*FourierUp*), demonstrates notable success, yielding an AUC of 96.99 on 512x512 input images. This technique stands out as a robust method for effectively capturing manipulated features. Moreover, our SpatialUp approach, employing Xception and Bilinear interpolation, achieves a remarkable AUC of 97.28 on a resolution of 512x512. Although the FourierUp method using periodic padding is slightly inferior to the SpatialUp method, both perform significantly better than the other models.

We compared our method with a diffusion-based upsampling technique that quadruples image resolution using prompts (Rombach et al., 2022). Testing with and without prompts, we observed that using prompts resulted in more alterations during upsampling, complicating deepfake detection (Figure 4). Though image quality improved qualitatively, deepfake elements were either added, removed or reduced. Due to long inference times, results on FF++ test images are not reported, but the validation results consistently demonstrated an AUC of around 90, making deepfake detection challenging post-upsampling with this diffusion method.

Furthermore, we performed cross-dataset evaluation on Celeb-DF (V2) (Yuezun Li and Lyu, 2020), an extensive dataset designed to mimic the visual quality of online videos. Unlike its predecessor, Celeb-DF (V1), which contained only 795 deepfake videos, this updated version includes 590 original videos from YouTube, spanning various ethnicities. Additionally, Celeb-DF (V2) encompasses 5639 corresponding deepfake videos, making it a valuable resource for evaluation and analysis. For testing purposes, 16,565 frames were selected from a subset of 518 designated as test videos.

Table 2: Performance Evaluation on Celebdf Dataset. FourierUp and SpatialUp use Xception with separable convolution. P-Padding and Area-Inter means respectively periodic padding and area interpolation. AUC is in percentage (%).

Up/Down	Model	Up	Down	Size	AUC
-	Xception*	-	-	256x256	74.30
-	Xception (Bonettini et al., 2021)	-	-	256x256	76.54
FourierUpDown (ours)	Xception*	P-Padding	Bilinear	256x256	75.56
SpatialUp (ours)	Xception	Bilinear	-	512x512	77.61
FourierUp (ours)	Xception	P-Padding	-	512x512	**78.57**
FourierUp (ours)	Xception	Area-Inter	-	512x512	74.52

The performance evaluation presented in Table 2 demonstrates the robustness and generalizability of the models and techniques when applied to a different dataset, particularly the Celebdf dataset, serving as a cross-dataset evaluation. The results obtained from training on FF++ and testing on Celebdf underscore the efficacy of the models in detecting image manipulations across different datasets. Xception model achieves significant AUC improvements in both datasets, scoring 95.05 on FF++ and 76.54 on Celebdf. Moreover, in our proposed methods, the Fourier-based techniques specifically FourierUp employing Periodic Padding (*FourierUp*) demonstrate promising results not only on the FF++ dataset (AUC of 96.99) but also on the Celebdf dataset (AUC of 78.57). This consistent performance underscores the reliability of Fourier-based methods in detecting manipulated features across diverse datasets. Similarly, the SpatialUp approach, maintains its effectiveness across datasets, displaying AUC scores of 97.28 on FF++ and 77.61 on Celebdf.

Overall, the consistent performance trends observed across the FF++ and Celebdf datasets validate the models ability to generalize and detect image manipulations effectively, affirming their reliability and efficacy in cross-dataset evaluation scenarios.

Furthermore, upon comparing our results with the state-of-the-art approaches, we observe that for the same data split, the reported AUC values of the different methods in (Bonettini et al., 2021) do not exceed 95.

5 CONCLUSION

In conclusion, our study underscores the pivotal role of upsampling and the combination of upsampling and downsampling techniques, both in spatial and frequency domains, in augmenting the performance of deep CNN models for deepfake detection. By demonstrating the substantial improvements achieved through these methods while minimizing alterations to the deep CNN architecture, our research sheds light on a promising avenue for enhancing model accuracy in detecting deepfakes within digital media. In our future research, we aim to explore alternative image sampling techniques based on generative models. Additionally, we have planned to assess the effectiveness of our method on datasets featuring deepfakes generated by diffusion techniques.

ACKNOWLEDGEMENTS

The work was partially supported by a NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grant under grant agreement No RGPIN-2020-05171.
We gratefully acknowledge the support of the Computer Research Institute of Montreal (CRIM), the Ministère de l'Économie et de l'Innovation (MEI) of Quebec, and The Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7.

Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., and Tubaro, S. (2021). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5012–5019.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Chen, P., Liu, J., Liang, T., Zhou, G., Gao, H., Dai, J., and Han, J. (2020). Fsspotter: Spotting face-swapped video by spatial and temporal clues. In *2020 IEEE*

International Conference on Multimedia and Expo (ICME), pages 1–6.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Ciftci, U. A., Demir, I., and Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

DeepfakeVFX.com (2023). DeepFaceLab - DeepfakeVFX.com — deepfakevfx.com. https://www.deepfakevfx.com/downloads/deepfacelab/. [Accessed 21-Jun-2023].

Deng, Y., Yang, J., Chen, D., Wen, F., and Tong, X. (2020). Disentangled and controllable face image generation via 3d imitative-contrastive learning.

Durall, R., Keuper, M., Pfreundt, F.-J., and Keuper, J. (2019). Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*.

Elhassan, A., Al-Fawa'reh, M., Jafar, M. T., Ababneh, M., and Jafar, S. T. (2022). Dft-mf: Enhanced deepfake detection using mouth movement and transfer learning. *SoftwareX*, 19:101115.

FaceApp (2023). FaceApp: Face Editor — faceapp.com. https://www.faceapp.com/. [Accessed 21-Jun-2023].

Firc, A., Malinka, K., and Hanáček, P. (2023). Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*.

Gomes, T. L., Martins, R., Ferreira, J., and Nascimento, E. R. (2020). Do as i do: Transferring human motion and appearance between monocular videos with spatial and temporal constraints.

Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049.

Hubens, N., Mancas, M., Gosselin, B., Preda, M., and Zaharia, T. (2021). Fake-buster: A lightweight solution for deepfake detection. In *Applications of Digital Image Processing XLIV*, volume 11842, pages 146–154. SPIE.

Ismail, A., Elpeltagy, M., S. Zaki, M., and Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using xgboost. *Sensors*, 21(16):5413.

Karasavva, V. and Noorbhai, A. (2021). The real threat of deepfake pornography: A review of canadian policy. *Cyberpsychology, Behavior, and Social Networking*, 24(3):203–209. PMID: 33760666.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan.

Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., and Zafeiriou, S. (2020). Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild".

Nirkin, Y., Keller, Y., and Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*.

Ranjan, P., Patil, S., and Kazi, F. (2020). Improved generalizability of deep-fakes detection using transfer learning based cnn framework. In *2020 3rd international conference on information and computer technologies (ICICT)*, pages 86–90. IEEE.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Tariq, S., Lee, S., Kim, H., Shin, Y., and Woo, S. S. (2018). Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security*, pages 81–87.

Thies, J., Zollhöfer, M., and Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Yang, Y. and Soatto, S. (2020). FDA: fourier domain adaptation for semantic segmentation. *CoRR*, abs/2004.05498.

Yuezun Li, Xin Yang, P. S. H. Q. and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021a). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194.

Zhao, L., Zhang, M., Ding, H., and Cui, X. (2021b). Mff-net: deepfake detection network based on multi-feature fusion. *Entropy*, 23(12):1692.

Zhou, M., Yu, H., Huang, J., Zhao, F., Gu, J., Loy, C. C., Meng, D., and Li, C. (2022). Deep fourier upsampling. *arXiv preprint arXiv:2210.05171*.