Deepfake Detection Using VOLO D5 Model

Kuldeep Chaudhary, and Prof. Pratik Narang
Birla Institute of Technology And Science - Pilani
Department of Computer Science and Information Systems
h20240184@pilani.bits-pilani.ac.in, pratik.narang@pilani.bits-pilani.ac.in

1 INTRODUCTION

Advances in generative models enable photorealistic face manipulation, undermining trust in digital media. This work studies deepfake detection using the Vision Outlooker (VOLO) architecture (L. Yuan and Yan, 2022), whose outlook attention couples fine-grained local cues with global context. We fine-tune VOLO-D5 on a curated forensic corpus and evaluate accuracy, AUC, and true detection rates at low false-positive levels. Results, ablations, and considerations demonstrate a robust baseline for screening applications.

2 LITERATURE REVIEW

The detection of deepfakes, both in images and videos, has become a critical area of research due to the progressing advancements in manipulation techniques. A range of methods has been introduced, addressing a variety of different challenges in deepfake detection, ranging from identifying low-quality forgeries to detecting manipulations generated by unseen methods.

2.1 Image Deepfake Detection

Early works in deepfake detection were focused on using Convolutional Neural Networks (CNNs) to identify semantic distortions. To address this, *Attention-based Deepfake Detection Distiller* (Woo, 2022) makes use of frequency attention distillation and multi-view attention to detect low-quality compressed deepfakes. Additionally, *On the Detection of Digital Face Manipulation* (H. Dang and Jain, 2020) proposes a novel approach using attention-based layers to not only detect manipulated faces but also localize the modified regions by generating attention maps. This technique outperforms existing methods by accurately detecting the altered facial regions.

2.2 Generalization Challenges and Augmentation Techniques

One of the key challenges in deepfake detection is ensuring that models can generalize well across various deepfake generation methods. The paper *Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection* (L. Chen and Wang, 2022) tackles this by creating augmented forgeries from a wide range of forgery configurations. This approach helps the model stay sensitive to different types of manipulations. By using adversarial training, the method dynamically generates the most challenging examples for the detector, which enables it to outperform other techniques. This approach avoids overfitting to specific datasets, making it more robust and adaptable in real-world scenarios.

2.3 Face and Context Discrepancies for Detection

Another promising approach involves leveraging the discrepancies between manipulated face regions and their surrounding context. *DeepFake Detection Based on Discrepancies Between Faces and Their Context* (Y. Nirkin and Hassner, 2021) proposes using two separate identity recognition networks: one for recognizing the face region and another for recognizing the context. The method detects identity-to-context discrepancies, which are robust even in the absence of traditional artifacts, and it performs well on multiple benchmarks, including *FaceForensics++* (A. Rossler and Nie0̈0dfner, 2019) and *Celeb-DF-v2* (Y. Li and Lyu, 2020).

2.4 Temporal Patterns in Video Deepfakes

Detecting temporal inconsistencies plays an important role in video deepfake detection. *Delving into Sequential Patches for Deepfake Detection* (J. Guan and Zhao, 2022) introduces a Local and Temporal-aware Transformer-based Deepfake Detection (LTTD) framework that models the temporal consistencies within sequences of local patches. The method focuses on detecting local forgery cues and achieves state-of-the-art robustness, especially for deep-fakes generated with previously unseen techniques. The use of Local Sequence Transformers (LST) allows the model to capture both short- and long-term temporal patterns, providing superior generalization.

2.5 Large-Scale Datasets for Deepfake Detection

The development of large-scale datasets has also been vital in the advancement of deepfake detection technology. *The DeepFake Detection Challenge (DFDC) Dataset* (B. Dolhansky and Ferrer, 2020) introduces the largest publicly available face-swap video dataset, featuring over 100,000 clips from 3,426 paid actors, and it has been used to benchmark video deepfake detection methods. The DFDC dataset enables training models that generalize well to *in-the-wild* deepfakes. The *Celeb-DF* dataset (Y. Li and Lyu, 2019), another deepfake video dataset, includes 5,639 high-quality deepfake videos generated from YouTube clips of celebrities. It is particularly challenging as it features high-quality manipulations that more closely resemble deepfakes circulated online, making it a tough benchmark for existing methods. The *Celeb-DF v2* dataset (Y. Li and Lyu, 2020) further improves on this by reducing visual artifacts and employing an advanced synthesis method, resulting in even more realistic deepfakes. This makes it an even more valuable resource for developing and testing robust detection models. In contrast, the *Diverse Fake Face Dataset (DFFD)*, introduced in *On the Detection of Digital Face Manipulation* (H. Dang and Jain, 2020), provides a deepfake image dataset. The paper collects 781,727 real images and 1,872,007 fake images, from which a subset of 58,703 real images and 240,336 fake images was randomly selected to make the dataset size manageable and to balance the subcategories, providing a comprehensive dataset that aids in training and evaluating detection models for various types of face manipulations.

2.6 Architectural Attribution in Deepfakes

Some methods focus on identifying the specific architectures used to create deepfakes. In *Deepfake Network Architecture Attribution* (T. Yang and Li, 2022), the authors introduce DNA-Det, a technique that classifies deepfake images based on the architecture of the GAN used, even when the models are fine-tuned or retrained with different settings. The method relies on consistent architectural fingerprints, which makes it a reliable way to attribute deepfakes across different generation techniques.

Summary: These methods cover various approaches, from enhancing model generalization and using new datasets to detecting identity manipulation and analyzing temporal patterns in video deepfakes. The focus on high-quality datasets and innovative training is advancing deepfake detection, improving its effectiveness in real-world applications.

3 DESIGN AND IMPLEMENTATION

3.1 Dataset Description

The Diverse Fake Face Dataset (H. Dang and Jain, 2020) is a comprehensive resource for detecting digitally manipulated face images. The dataset introduced in the paper *On the Detection of Digital Face Manipulation* (H. Dang and Jain, 2020) consists of a subset of 58,703 real images and 240,336 fake images. In the original paper, the subset was split into 50% for

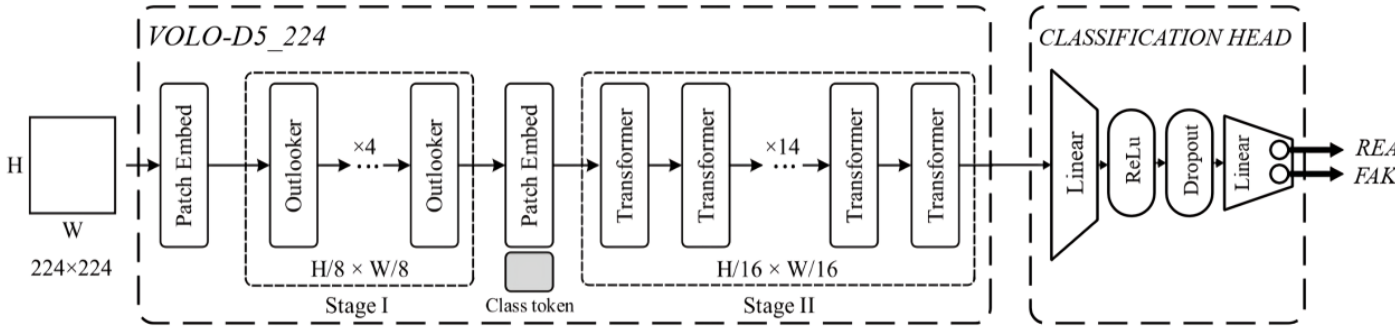

Figure 1: Model architecture

training, 5% for validation, and 45% for testing, providing a balanced and diverse representation suitable for comprehensive evaluation. These images include manipulations from various categories such as identity swaps, expression changes, attribute edits, and entirely synthetic faces.

The version of the DFFD dataset available for this study was smaller, but maintained the core structure and diversity outlined in the original work. The splits for training, validation and testing were predefined, ensuring consistency in evaluation protocols. Specifically:

- Training set:
  - Fake: 62,574 images
  - Real: 10,000 images
- Validation set:
  - Fake: 6,996 images
  - Real: 999 images
- Test set:
  - Fake: 66,862 images
  - Real: 9,000 images

This predefined split is nearly proportional to the original dataset structure and ensures comprehensive coverage of both real and fake samples. Each image was resized to a resolution of 224 × 224, a standard preprocessing step for compatibility with the VOLO-D5 architecture.

3.2 VOLO Architecture for Deepfake Detection

The Vision Outlooker (VOLO) architecture (L. Yuan and Yan, 2022) represents a significant advance in visual recognition tasks by incorporating outlook attention. Unlike traditional self-attention mechanisms, outlook attention aggregates fine-level features in a sliding window manner, ensuring:

- Efficient encoding of local structural information (e.g., edges, textures).
- Reduced computational overhead compared to global self-attention.

VOLO's fine-grained feature encoding makes it particularly effective for tasks like deepfake detection, where subtle artifacts (e.g., blending inconsistencies, edge distortions) must be identified. Among VOLO variants, we selected VOLO-D5 because it is a heavier model that generates better results, providing more sophisticated feature extraction capabilities for detecting complex deepfake characteristics.

3.2.1 Why VOLO for Deepfake Detection?

Deepfake images often introduce subtle manipulations, such as:

- Seam artifacts at boundaries of blended regions.
- Irregularities in high-frequency textures.

VOLO's design specifically addresses these challenges by:

1. Encoding both global dependencies (via transformer layers) and local structural details (via outlook attention).
2. Providing robustness to small datasets, a common limitation in deepfake detection, through efficient tokenization and aggregation mechanisms.

By leveraging its two-stage architecture (Outlooker blocks followed by Transformer layers), VOLO-D5 ensures comprehensive feature extraction essential for identifying both real and fake images.

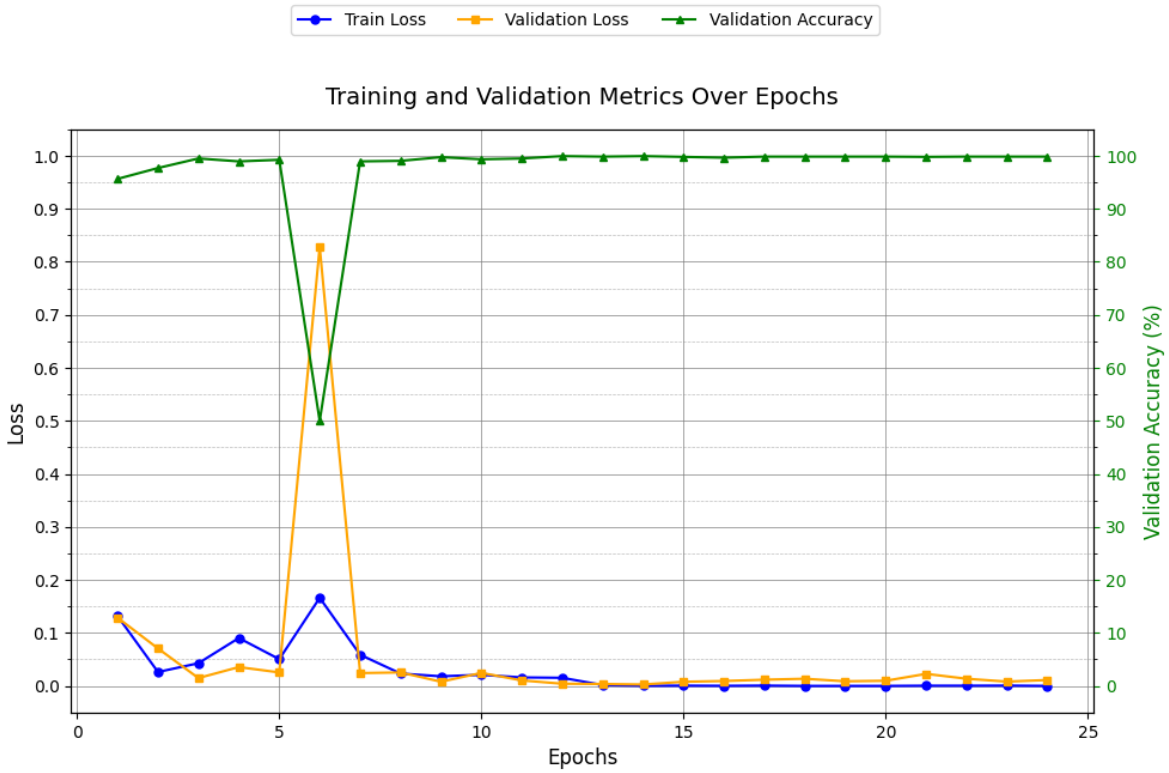

Figure 2: Training loss, validation loss, and validation accuracy over all epochs.

3.3 Model Architecture and Implementation

The architecture of the deepfake detection model as shown in Figure 1 based on VOLO-D5 includes the following key components and design strategies:

- **VOLO-D5_224 Architecture:** A vision transformer model with 224-pixel input resolution, designed to capture complex visual features through innovative attention mechanisms and pretrained on large-scale image classification datasets.
- **Backbone Network:** Utilizes transfer learning with a pretrained VOLO-D5 model, extracting high-level feature representations by removing the original classification head and allowing custom adaptation for deepfake detection.
- **Classification Head:** A custom multi-layer neural network comprising:
  - Linear dimension reduction to 512 features
  - ReLU activation for non-linearity
  - Dropout Regularization Layer: Applied with a probability of 0.5 to mitigate overfitting during training
  - Final linear layer for binary classification
- **Optimizer:** The model was trained using the Adam optimizer, chosen for its efficiency and adaptive learning rate capabilities.
- **Gradient Scaler:** Used for mixed precision training, improving computational efficiency while maintaining numerical stability through `torch.amp` techniques.
- **Loss Function:** CrossEntropyLoss was employed to distinguish between real and fake images, providing an effective metric for binary classification.

3.4 Implementation Details

- Framework: The model was implemented in PyTorch.
- Hardware: Training was conducted on an NVIDIA GeForce RTX 2080 Ti GPU core with a batch size of 16.
- Hyperparameters:
  - Learning Rate: 1e-4 for the first 12 epochs; reinitialised to 1e-5 on resuming training for the remaining epochs.
  - Epochs: 24

By combining VOLO's robust architecture with domain-specific enhancements, our model achieves high accuracy in detecting manipulated images within the DFFD subset. The results are discussed in detail in the subsequent section.

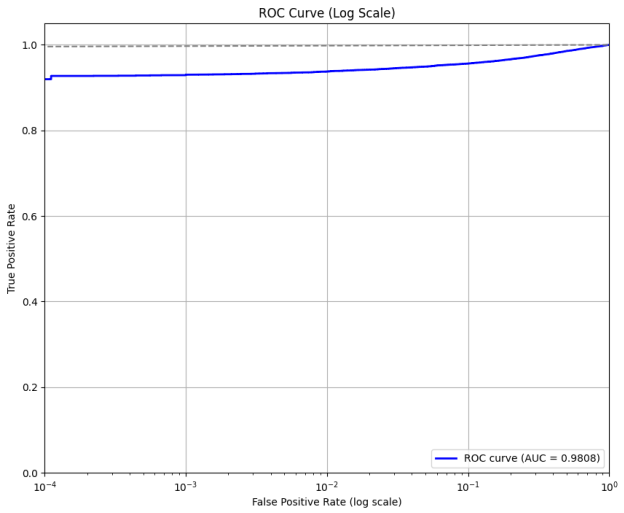

Figure 3: ROC curve at Epoch 14.

4 EXPERIMENTAL RESULTS

4.1 Training Setup

The VOLO-D5 model was trained over 24 epochs, divided into two distinct phases. In the first phase, the model was trained for 12 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. In the second phase, training resumed for an additional 12 epochs, with the optimizer reinitialized and the learning rate reduced to 1×10^{-5}. This reduction aimed to fine-tune the model and improve its convergence during the later stages of training.

The training leveraged the entire dataset, encompassing both real and manipulated images, to ensure the model captured the full spectrum of image manipulations. The Adam optimizer was particularly effective due to its adaptive learning rate, which is beneficial for handling imbalanced datasets.

4.2 Validation Strategy

To ensure that the validation accuracy reflected the model's true performance, a balanced subset of the validation dataset was used. This subset contained an equal number of real and fake images, which prevented inflated accuracy metrics due to dataset imbalance.

4.3 Performance Trends and Key Insights

The model's learning dynamics, as shown in Figure 2, highlight its rapid convergence and strong classification performance. Early in training, the VOLO-D5 demonstrated its ability to distinguish between real and manipulated images, with validation accuracy rising from 95% to nearly 99.9% within a few epochs.

Notably, there was a sharp drop in validation accuracy to 50% at epoch 6. This occurred because the model became stuck in a sub-optimal minima, where it predominantly predicted all images as fake. The presence of the dropout layer may have also contributed to this instability by temporarily disrupting the model's ability to generalize. However, the model quickly recovered in subsequent epochs, regaining its ability to classify real and fake images with high accuracy.

By epoch 14, the VOLO-D5 achieved its peak validation accuracy of 99.95%. The training loss steadily decreased and asymptotically approached zero, while validation accuracy stabilized at 99.85% during later epochs. These trends demonstrate the model's capacity to achieve consistent and accurate classification of manipulated images, even when trained on a relatively small dataset.

4.4 Evaluation Metrics

To comprehensively assess the performance of the models, the following evaluation metrics were employed:

1. **Area Under the ROC Curve (AUC):** The AUC measures the ability of a model to distinguish between real and fake images across varying thresholds. A higher AUC value indicates better discriminatory power, with a value of 1 representing perfect classification.
2. **True Detection Rate (TDR):** The TDR represents the proportion of correctly identified fake images at specific false positive rates (FPR). Two strict thresholds were considered in this study:
   - TDR@0.01% FPR: Detection rate when the FPR is 0.01%.
   - TDR@0.1% FPR: Detection rate when the FPR is 0.1%.
3. **Percentage of Binary Classification Accuracy (PBCA):** The PBCA measures the overall accuracy of the model in correctly classifying both real and fake images. It provides a straightforward measure of performance based on the binary nature of the task.

By analyzing these metrics, the robustness and practical applicability of each model can be effectively compared.

Table 1: Comparison of Backbone Models: Performance Metrics

Backbone Models	AUC (%)	TDR@0.01%	TDR@0.1%	PBCA (%)
Ours				
VOLO-D5	98.08	91.97	92.91	96.49
From (H. Dang and Jain, 2020)				
Xception+Reg	99.64	83.83	90.78	88.44
VGG+MAM	99.67	75.89	87.25	86.74
VGG+Reg	99.46	44.16	61.97	91.29

4.5 Comparative Analysis of Results

The DFFD dataset utilized in this study was a smaller version of the original but still preserved the essential diversity of real and fake manipulations required for effective evaluation. The dataset was used without any data augmentation. Furthermore, all images were resized to a resolution of 224×224 to reduce computational load. Despite these constraints, the VOLO-D5 model exhibited strong performance, effectively identifying subtle artifacts indicative of deepfake manipulations.

The comparative analysis presented in Table 1 includes our VOLO-D5-based model and the results reported in the paper *On the Detection of Digital Face Manipulation* (H. Dang and Jain, 2020). In our study, the VOLO-D5 model achieved its peak performance at Epoch 14, demonstrating its ability to effectively generalize even under resource-constrained settings.

At Epoch 14, our VOLO-D5 model achieved an AUC of 98.08%, while maintaining a PBCA of 96.49%. Particularly impressive is its TDR@0.01% of 91.97%, which highlights its robustness in detecting deepfake images at very low false positive rates.

The models from the paper (H. Dang and Jain, 2020)—Xception+Reg, VGG+MAM, and VGG+Reg—demonstrated slightly higher AUC values but fell short in stringent detection scenarios, such as TDR@0.01%. This indicates that while these models excelled in overall performance, their robustness under stricter thresholds was limited.

These results emphasize the strong balance achieved by our VOLO-D5-based models between high detection accuracy and robustness at low false positive rates. The results are particularly significant considering the model was trained on a smaller version of the DFFD dataset, without any augmentation, and at a lower resolution of 224×224. This highlights the VOLO-D5 architecture's ability to capture subtle deepfake artifacts effectively, positioning it as a promising solution for deepfake detection tasks in real-world settings.

5 CONCLUSIONS

Please note that ONLY the files required to compile your paper should be submitted. Previous versions or examples MUST be removed from the compilation directory before submission.

We hope you find the information in this template useful in the preparation of your submission.

ACKNOWLEDGEMENTS

If any, should be placed before the references section without numbering. To do so please use the following command: *section*{*ACKNOWLEDGEMENTS*}

REFERENCES

A. Rossler, D. Cozzolino, L. V. C. R. J. T. and Nie0̈0dfner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11.

B. Dolhansky, J. Bitton, B. P. J. L. R. H. M. W. and Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

H. Dang, F. Liu, J. S. X. L. and Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790.

J. Guan, H. Zhou, Z. H. E. D. J. W. C. Q. and Zhao, Y. (2022). Delving into sequential patches for deepfake detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 4517–4530.

L. Chen, Y. Zhang, Y. S. L. L. and Wang, J. (2022). Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719.

L. Yuan, Q. Hou, Z. J. J. F. and Yan, S. (2022). Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6575–6586.

T. Yang, Z. Huang, J. C. L. L. and Li, X. (2022). Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4662–4670.

Woo, S. (2022). Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 122–130.

Y. Li, X. Yang, P. S. H. Q. and Lyu, S. (2019). Celeb-df (v2): A new dataset for deepfake forensics.

Y. Li, X. Yang, P. S. H. Q. and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216.

Y. Nirkin, L. Wolf, Y. K. and Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121.

APPENDIX

If any, the appendix should appear directly after the references without numbering, and not on a new page. To do so please use the following command: $\backslash section*\{APPENDIX\}$