Robust Deepfake Detection and Resilient Adversarial Image Reconstruction with Reduced Features Set

Muhammad Irfan[1], Myung J. Lee [1], Daiki Nobayashi[2]
[1] CUNY - The City College of New York, USA
mirfan001@citymail.cuny.edu, mlee@ccny.cuny.edu
[2] Kyushu Institute of Technology, Japan
nova@ecs.kyutech.ac.jp

Abstract. The rapid advancements in deepfake technology pose significant challenges in detecting manipulated media. This research introduces a feature extraction and selection method to address this threat. An optimally integrated pre-trained model is introduced to extract features from face images, composed of three CNN models —DenseNet-121, EfficientNet-B0, and ResNet-18 —and fine-tuned on the Celeb-DF (V2) dataset. These features are stacked for diverse representations, and a novel Assimilation-Elimination (ASEL) selection algorithm is used to minimize redundancies. The selected features are then fed into a KNN classifier to determine if a image is real or manipulated. Experiments on Celeb-DF (V2) achieve an AUC score of 97.77%, confirming the model's robustness. Additionally, reduced feature sets from real images are transmitted over a noisy communication channel, optimizing storage and bandwidth needs. Afterwards, the recovered bits are fed into the proposed error-resilient Feature-Driven Adversarial Image Reconstruction (FDAIR) model at the receiver, achieving image reconstruction comparable to state-of-the-art methods.

1 Introduction

In the context of remote forensic analysis, the integrity of digital evidence, particularly videos and images, is crucial due to the sophistication of deepfake technology. Ensuring data authenticity during transmission from local servers to forensic labs is critical to prevent tampering. Advances in technology have led to increasingly sophisticated methods for creating fake media content, posing significant challenges for forgery detection. Prior studies have explored various techniques for deepfake detection. Auto-encoders trained on real face images have been used to obscure Generative Adversarial Network (GAN) fingerprints in synthetic images. [1] focused on eliminating visible signs of manipulation to make deepfakes blend seamlessly with authentic imagery. Face-swapping techniques further complicate detection efforts, with methods utilizing disparities between facial regions [2] and targeting specific areas like eyebrows [3]. To address these challenges, advanced methods have been proposed, such as Convolutional Neural Network (CNN) model based on domain-invariant representation learning [4] and methods to identify common GAN features to enhance detection generalization [5]. An attention-based approach focusing on specific facial regions to reduce the search space for identifying manipulated artifacts [6].

Transformers have been integrated to enhance generalization [7], while vision transformer models have been used for feature extraction [8]. Techniques like filter and feature ranking methods [9] and feature-level analysis further improve detection performance, highlighting the need for innovative approaches to counter deepfakes. In image generation and restoration, approaches such as generative priors and diffusion models have been discussed. PULSE proposed StyleGAN inversion for image upsampling [10] but struggles with fidelity at small downsampling factors. BRGM frames GAN inversion as Bayesian inference, with L-BRGM further optimizing quality [11]. Despite these advancements, forgery detection and image reconstruction face challenges like noise interference, limited model applicability, and high computational demands, emphasizing the need for more effective methods.

This paper presents a novel deepfake detection and image reconstruction framework using features from diverse CNNs, stacked as shown in Fig. 1. At the sender, a feature selection strategy avoids overfitting, reduces computational cost, and improves generalization. Selected features are fed into a K-nearest neighbors (KNN) classifier to identify real and fake images. Detected fake images are discarded, while real content is transmitted over a lossy communication channel, whether it is wireless or wired. Upon reception, bits are recovered, and features are processed by the GAN-based Feature-Driven Adversarial Image Reconstruction (FDAIR) model for image restoration.

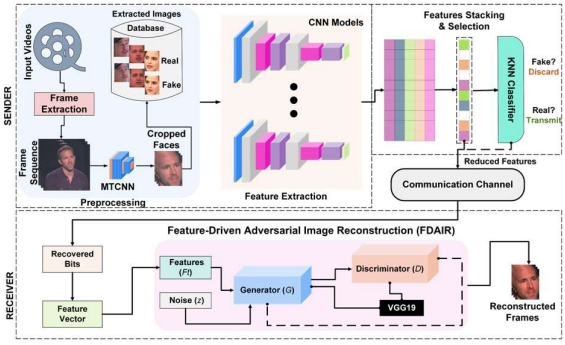

Fig. 1. Workflow of a proposed system model.

In summary, the contributions of this paper are presented as follows:

- Developed a detection workflow model using features from CNNs (DenseNet-121, EfficientNet-B0, ResNet-18) to enhance detection accuracy through increased features diversity.
- Proposed an Assimilation-Elimination (ASEL) feature selection algorithm to mitigate overfitting, lower computational costs, and optimize bandwidth (BW) efficiency. Experiments show ASEL outperforms state-of-the-art methods in classification accuracy and Area Under Curve (AUC) score.
- Proposed a FDAIR model to reconstruct images from reduced noisy feature sets. Our experimental evaluations of Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Normalized Root Mean Square Error (NRMSE) demonstrate its effectiveness in image reconstruction, indicating resilience against limited and noisy features.

The rest of the paper is organized as follows. Section 2 presents a technical approach for our proposed methods, while Section 3 discusses the experimental results. Finally, the paper is concluded in Section 4.

2 Technical Approach

This section discusses two key areas of our framework: feature selection and deepfake detection at sender, and a FDAIR model to reconstruct the images at the receiver.

2.1 Deepfake Detection

The proposed deepfake detection method uses the ASEL algorithm to select relevant features from stacked CNN outputs, achieving a near-optimal subset. A KNN classifier is then used for deepfake classification.

2.1.1 Deep Feature Extraction

For effective feature extraction, our approach is to integrate multiple CCN models. To determine the optimal number of component models, the accuracy gain is evaluated each time of adding one component model. The CCN models for the evaluation include. DenseNet-121 [12], EfficientNet-B0 [13], ResNet-18 [14], and GoogLeNet [15]. As shown in Fig. 2, the accuracy improvement Δ by adding a model diminishes rapidly when adding fourth model at only 0.77%, which justifies our choice of three models. As shown in Fig. 1, flattened feature maps were obtained by applying global average pooling (GAP) to the top layer of each CNN model, resulting in dimensions of 1024 (DenseNet-121), 1280 (EfficientNet-B0), and 2048 (ResNet-18). This compact feature representation facilitated subsequent feature combination and analysis for our deepfake detection algorithm. Using these diverse CNN models aimed to capture superior features, enhancing overall performance.

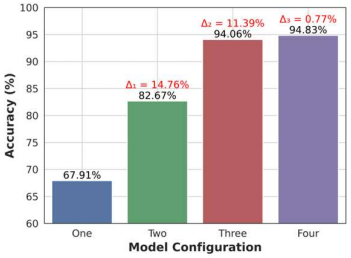

Fig. 2. Models' complexity-accuracy.

2.1.2 Feature Rankings

The features selection process is indicated in Fig. 3 where the stacked features are given to three different filtering methods: ReliefF [16], Mutual Information (MI) [17], and Minimum Redundancy Maximum Relevance (mRMR) [18] to produce a final rank vector. **ReliefF** identifies relevant features by computing weighted distances between dataset instances, distinguishing same-class from different-class instances. Features with larger differences among same-class neighbors and smaller differences with different-class neighbors receive higher ReliefF scores, indicating their importance. **MI** evaluates feature relevance by computing mutual information scores between

features and the target variable. It starts with dimensionality reduction, computes marginal probabilities, applies a binning strategy for discretization. **mRMR** identifies informative features by evaluating relevance to the target and redundancy with other features. It prioritizes high-relevance, non-redundant features. All methods generate normalized scores for feature selection and are combined to compute a $score(X_i)$, weighted by ω_1= 0.5, ω_2= 0.25, and ω_3= 0.25 respectively as indicated in Eq. (1). These weights determine the importance of each method while ranking.

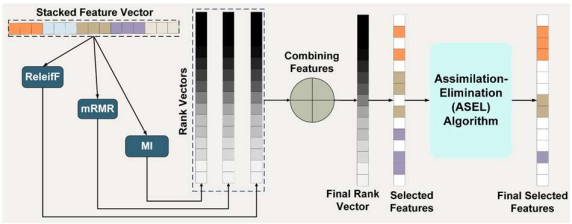

Fig. 3. Feature ranking and selection.

$$score(X_i) = \omega_1 \times ReliefF^{scores} + \omega_2 \times mRMR^{scores} + \omega_3 \times MI^{scores} \quad (1)$$

where, $ReliefF^{scores}$, $mRMR^{scores}$, and MI^{scores} represent the importance score assigned to feature X_i by respective method. These combined feature scores are then used to rank the features. Eq. (1) can be represented by standardizing the scores obtained from each feature selection method before combining them, ensuring that each method contributes proportionally to the final score as depicted in Eq. (1a).

$$score(X_i) = \omega_1 \times z_{ReleifF}(X_i) + \omega_2 \times z_{mRMR}(X_i) + \omega_3 \times z_{MI}(X_i) \quad (1a)$$

here, $z_{ReleifF}(X_i)$, $z_{mRMR}(X_i)$, and $z_{MI}(X_i)$ are the standardized scores obtained from ReliefF, mRMR, and MI methods respectively for feature X_i. The standardized scores $z(X_i)$ for feature X_i can be computed as $z(X_i) = \frac{score(X_i) - \mu}{\sigma}$ that ensures that the scores have a mean of 0 and a standard deviation of 1.

2.1.3 ASEL Based Feature Selection

The conventional filter method for feature selection typically involves selecting top-ranked features based on their relevance scores, often determined experimentally. However, this approach lacks validation beyond experimental results, raising doubts about the true contribution of chosen features to the classification process. Moreover, prioritizing only the highest-ranked features may not always lead to optimal selection, as redundant information may persist. To address this, we propose a novel method that iteratively refines the feature subset to enhance classification accuracy. Initially, the top η% of ranked features are selected, and the subset is iteratively updated by incorporating discarded features and excluding some previously selected ones. In each iteration, α% of the selected features are randomly removed, while β% of the non-selected features are added for comparison with the existing feature subset to search for an optimal feature subset. The values of η, α%, and β% are determined experimentally. The comparison between the generated feature subset and the previous one is evaluated based on a fitness function as indicated in Eq. (2).

$$fitness(X) = \theta \times acc \times (1 - \theta) \times \left(1 - \frac{N_{sel}}{n}\right) - \gamma \times penalty(X) \quad (2)$$

where, θ represents a weight factor, acc represents the accuracy of the selected features, N_{sel} represents the number of selected features, n represents the total number of features, γ is the regularization parameter allows controlling the trade-off between accuracy and feature complexity in the fitness calculation, and $penalty(X)$ is a penalty term that accounts for redundancy of the selected features and can be represented as $penalty(X) = \delta \times redundancy(X)$. It penalized based on redundancy among selected features. For instance, high redundancy will increase the penalty. The δ is a scaling factor set to 0.6 and $redundancy(X)$ measures the redundancy among the selected features which can be computed as the average correlation among selected features. The fitness function, expressed as Eq. (2), balances feature accuracy acc and the length of the selected feature subset (N_{sel}), with the trade-off controlled by weight θ. In this scenario, θ is set to 0.9. Feature accuracy acc is determined using a KNN classifier on the validation dataset.

2.2 Proposed FDAIR Model for Image Reconstruction

The proposed FDAIR model reconstructs images from a reduced feature set recovered at the receiver. The generator produces high-quality images, while the discriminator ensures their realism by evaluating feature authenticity. The training process combines perceptual, feature, adversarial, and feature preservation losses, weighted by hyperparameters λ_1, λ_2, and λ_3, to optimize reconstruction quality and feature fidelity. FDAIR employs adversarial training and a feature space discriminator, effectively learning to reconstruct images without direct access to real images. The generator and discriminator both use a learning rate of 0.0001, a batch size of 32, and train for 100 epochs. The overall objective function given as Eq. (3) involves a generator G and a discriminator D, as seen in GANs. The goal is to optimize G and D such that G generates realistic data, while D discriminates between real and generated data.

$$min_G max_D(V(D,G) + \lambda_1 L_{percep} + \lambda_2 L_{feat} + \lambda_3 L_{preserv}) \quad (3)$$

Here, λ_1, λ_2, and λ_3 are hyperparameters that control the contribution of each loss term to the overall objective.

2.2.1 Generator and Discriminator

As shown in Fig. 4, the novel generator design for high-quality image reconstruction integrates several key modules. It starts by receiving reduced noisy features, then uses a Multi-Scale Feature Extraction module to capture diverse scales. Attentional Residual

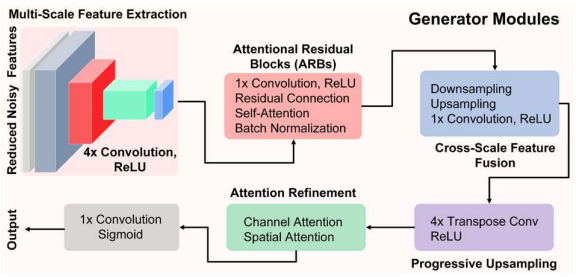

Fig. 4. Proposed generator modules in the FDAIR model.

Blocks (ARBs) enhance focus and stability, followed by Cross-Scale Feature Fusion Modules (CSFFMs) that refine feature representation through downsampling and upsampling. The Progressive Upsampling Module (PUM) increases resolution and reduces artifacts, while the Attention Refinement Module (ARM) emphasizes important regions. Finally, the output module generates the image, ensuring superior feature fidelity.

The discriminator evaluates the realism of reconstructed images using only their feature representations, without access to input features or real images. This approach ensures high-quality reconstruction. Employing a U-Net architecture designed in [19], the discriminator captures both local and global contextual information. The U-Net's encoder-decoder structure extracts hierarchical feature representations, identifying detailed patterns and distinctions crucial for differentiating real from generated features. Despite not accessing original images, the discriminator uses adversarial, feature, and preservation loss functions to guide reconstruction. The adversarial loss ensures the generator produces realistic feature representations, while feature and preservation losses maintain alignment and essential attributes during reconstruction.

2.2.2 Loss Functions

Adversarial Loss: It represents the adversarial component of the GAN.

$$min_G max_D\left(E_{F_t}[\log D(I|F_t)] + E_{F_{tz}}[\log(1 - D(G(F_t, z)|F_t))]\right) \quad (3a)$$

where F_t represent recovered set of reduced features at the receiver, I is the real image associated with features F_t, $G(F_t, z)$ represents generated image from generator G given input features F_t and random noise z. The discriminator D tries to maximize this loss term by correctly classifying real images as real $\log D(I|F_t)$ and generated images as fake $\log(1 - D(G(F_t, z)|F_t))$. The generator minimizes this loss term to fool the discriminator.

Feature Loss: It ensures the generated image's features match the input features.

$$L_{feat} = \|F_t - G(F_t, z)\|_2^2 \quad (3b)$$

Perceptual Loss: It measures the perceptual difference between the synthesized image I_{syn} and the reconstructed image $\hat{I}$.

$$L_{percep} = \frac{1}{N}\sum_{i=1}^{N}\left\|\varphi_i(I_{syn}) - \varphi_i(\hat{I})\right\|_2^2 \quad (3c)$$

The φ_i represents a feature extraction function at layer i, and N represents the number of layers used to extract features.

Preservation Loss: This loss ensures that certain attributes are preserved in the generated image:

$$L_{preserv} = \frac{1}{M}\sum_{j=1}^{M}\left\|\psi_j(F_t) - \psi_j(G(F_t, z))\right\|_2^2 \quad (3d)$$

The ψ_j function extracts specific attributes from the image, and M represents the number of attributes being preserved.

3 Results and Discussions

In this section, we evaluated our proposed deepfake detection method's robustness by analysing its performance under various challenges present in Celeb-DF (V2) dataset as "CeDF" [20]. It was divided into training, validation, and test sets. We used equidistant frames for training and the first I-frame for validation and testing. Faces were cropped using the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm, and only the face image with the highest confidence score was retained. We employed three CNNs with batch size of 32 for 100 epochs, using the Adam optimizer. Accuracy was computed with a KNN classifier with n_neighbors set to 5. Using Area Under Curve (AUC) score and test accuracy as primary metrics, we evaluated its effectiveness and generalization capability. These metrics revealed how well the method distinguished between real and fake content and performed across varied conditions in the dataset. Also, we comprehensively evaluated our FDAIR model's effectiveness by comparing it with other state-of-the-art approaches such as SRGAN [21], ESRGAN [22], and A-ESRGAN [19]. To ensure objectivity, we utilized three standard evaluation metrics: PSNR, SSIM, and NRMSE, particularly focusing on image reconstruction.

3.1 Evaluation for Deepfake Detection

Our experiments determined the optimal parameters for feature selection: η (30-50), α (5-10), and β (5-10). As indicated in the Table 1., the best performance was achieved with η=30, α=10, and β=10.

Table 1. Ablation study on the hyperparameters of the proposed ASEL.

Parameters			Performance Score (%)		
η	α	β	Validation Accuracy	Feature Length	Objective Score
30	05	05	97.76	482	95.13
30	05	10	98.27	557	94.97
30	10	05	98.59	373	96.49
30	**10**	**10**	**98.83**	**339**	**97.03**
40	05	05	96.56	1046	90.53
40	05	10	96.79	1199	89.59
40	10	05	94.30	1161	88.03
40	10	10	97.01	1176	89.17

The confusion matrix for the proposed method is shown in Fig. 5(a), reveals that only 11 images are misclassified when trained and tested on the CeDF dataset. To evaluate the randomness impact, the algorithm ran for 1000 iterations. Fitness scores, shown in Fig. 5(b), saturated after about 240 iterations with regularization term set to 0, suggesting near-optimal features were found. Higher values of regularization term improved the fitness score more slowly, indicating a balance between regularization and performance. In Fig. 5(c), the reduced feature set (339 features, 10.8 kbits) optimizes transmission BW and storage usage, requiring only 16.5%, 26.3%, and

Table 2. Performance of individual and combined models.

Model	# Features	Accuracy (%)	AUC (%)
DenseNet-121	1024	92.48	90.79
EfficientNet-B0	1280	92.85	91.49
ResNet-18	2048	95.18	93.13
Stacked Features	4352	95.73	97.01
Reduced Features	**339**	**97.53**	**97.77**

33% of the storage compared to ResNet-18 (2048 features, 65.5 kbits), EfficientNet-B0 (1280 features, 41.0 kbits), and DenseNet-121 (1024 features, 32.8 kbits), respectively. This reduction is quite significant for bandwidth-constraint networks.

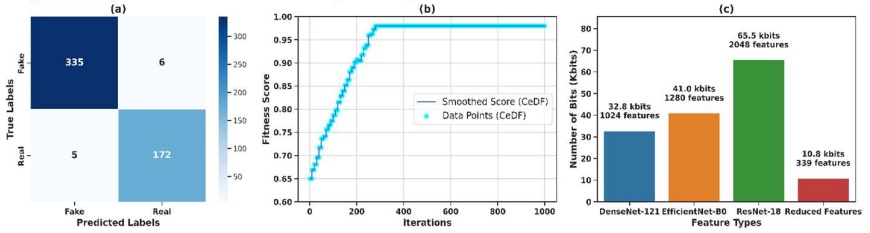

Fig. 5. Performance evaluation. (a) Confusion matrix. (b) Fitness scores. (c) Comparing storage and BW gain among feature types obtained from different models.

We also evaluate the overall effectiveness of the architecture and its components when test artifacts are known during model training. We assess three CNN models both individually and in combination. Additionally, we evaluate the proposed method's performance, which includes a feature selection process applied to the combined features. The test accuracy and AUC scores for the CeDF dataset are summarized in Table 2. Results indicate that the model using combined features outperforms the individual models with improved performance of 1.75%. We evaluated our deepfake detection method using test accuracy and AUC score as performance metrics. As reported in Table 3, our method outperforms the other state-of-the-art techniques, demonstrating superior performance with only 339 features, achieving a test accuracy of 97.53% and an AUC score of 97.77%.

Table 3. Comparison with other state-of-the-art methods.

Experiment	Method	# Features	Accuracy (%)	AUC (%)
CeDF Dataset	[20]	2048	95.37	98.88
	[23]	1024	65.64	65.33
	[24]	300	68.33	78.04
	Proposed	**339**	**97.53**	**97.77**

3.2 Evaluation for Proposed Error-resilient FDAIR Model

Fig. 6 compares the performance of image reconstruction methods: SRGAN, ESRGAN, A-ESRGAN, and a proposed method, using SSIM, PSNR, and NRMSE. The average results are shown in Table 4 indicating that the proposed method excels in all metrics, achieving the highest SSIM and PSNR and lowest NRMSE when comparing with other state-of-the-art methods. For instance, the proposed method improves SSIM by

Table 4. Evaluating proposed FDAIR model with other state-of-the-art methods.

Method	SSIM	PSNR (dB)	NRMSE
SRGAN	0.657	23.72	0.072
ESRGAN	0.770	28.04	0.044
A-ESRGAN	0.858	31.93	0.028
Proposed	**0.876**	**32.89**	**0.025**

33.47%, PSNR by 38.65%, and reduces NRMSE by 65.19% compared to SRGAN. These results highlight the proposed method's superior image reconstruction quality, visual fidelity, and noise reduction ability.

Fig. 6. Performance evaluation of proposed FDAIR model.

Fig. 7. The FDAIR's resilience against noisy features.

As depicted in Fig. 7, our proposed FDAIR model demonstrates strong resilience to noise, maintaining consistent performance across different SNR levels. Key metrics—SSIM (average 0.872, standard deviation 0.0015), PSNR (average 32.64 dB, standard deviation 0.0861), and NRMSE (average 0.025, standard deviation 0.0026)—show minimal variation. This indicates that the model effectively preserves image quality and structural similarity despite increasing noise, showcasing its robustness in noisy environments.

4 Conclusion

This research presents a robust approach to deepfake detection and adversarial image reconstruction, utilizing deep feature extraction and selection. Leveraging pre-trained CNN models—DenseNet-121, EfficientNet-B0, and ResNet-18—fine-tuned on the Celeb-DF (V2) dataset, the study achieves high accuracy in distinguishing real from manipulated images. The novel ASEL feature selection algorithm effectively reduces redundant features, enhancing KNN classifier efficiency, resulting in an impressive AUC score of 97.77%, outperforming existing approaches. Moreover, the method optimizes BW by transmitting reduced feature sets over noisy wireless channels. At the receiver, the proposed FDAIR model significantly improves image quality, evident in a 33.47% SSIM enhancement, 38.65% PSNR improvement, and 65.19% NRMSE reduction compared to state-of-the-art SRGAN method. Also, the FDAIR model shows strong resilience to noise, maintaining consistent performance across SNR levels. This study enhances deepfake detection and ensures image quality in adverse conditions, offering value for applications like remote forensic analysis. Future work aims to balance accuracy and computational complexity, potentially exploring knowledge distillation approaches.

Acknowledgments. This work is supported by NSF PAWR COSMOS (#1827923) and NSF IRNC COSMIC (#2029295) grants.

References

1. C. Liu, H. Chen, T. Zhu, J. Zhang and W. Zhou, "Making DeepFakes More Spurious: Evading Deep Face Forgery Detection via Trace Removal Attack", 2023.
2. Y. Nirkin, L. Wolf, Y. Keller and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context", 2022.
3. H. M. Nguyen and R. Derakhshani, "Eyebrow Recognition for Identifying Deepfake Videos", 2020.
4. W. Yuanlu, W. Yan, L. Caiyu, H. Guoqiang, "Learning domain-invariant representation for generalizing face forgery detection", 2023.
5. Z. Yan, Y. Zhang, Y. Fan and B. Wu, "UCF: Uncovering Common Features for Generalizable Deepfake Detection", 2023.
6. C. Beijing, L. Tianmu, D. Weiping, "Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM", 2022.
7. W. Tianyi, C. Harry, P. Kam, Chow, N. Liqiang, "Deep Convolutional Pooling Transformer for Deepfake Detection", 2023.
8. Y. Heo, W. Yeo, B. Kim, "DeepFake detection algorithm based on improved vision transformer", 2023.
9. G. Kushal, B. Shemim, S. Aritra, A. Sukdev, G. Manosij, K. Munish, S. Ram, "Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data", 2020.
10. M. Sachit, D. Alexandru, H. Shijia, R. Ravi, R. Cynthia, "Pulse: Self supervised photo upsampling via latent space exploration of generative models", 2020.
11. C. Arthur, M. Subhadip, S. Carola-Bibiane, "Stylegan-induced data-driven regularization ¨for inverse problems", 2022.
12. G. Huang, Z. Liu, L. Van Der Maaten and K. Weinberger, "Densely Connected Convolutional Networks", 2017.
13. M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", 2019.
14. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", 2016.
15. N. Kasim, N. Rahman, Z. Ibrahim, N. Mangshor, "Celebrity face recognition using deep learning", 2018.
16. N. Aggarwal, U. Shukla, G.J. Saxena, "Mean based relief: An improved feature selection method based on ReliefF", 2023.
17. B. Mario, M. Alberto, P. Matteo, T. Andrea, R. Marcello, "Feature Selection via Mutual Information: New Theoretical Insights", 2019.
18. G. Wang, F. Lauri and A. H. E. Hassani, "Feature Selection by mRMR Method for Heart Disease Diagnosis," in IEEE Access, vol. 10, pp. 100786-100796, 2022.
19. Z. Wei, Y. Huang, Y. Chen, C. Zheng, and J. Gao, "A-esrgan: Training real-world blind super-resolution with attention u-net discriminators", 2023.
20. Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics", 2020.
21. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network", 2017.
22. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change, "Esrgan: Enhanced super-resolution generative adversarial networks", 2018.
23. D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network", 2018.
24. G. Zhiqing, Y. Gaobo, C. Jiyou, S. Xingming, "Fake face detection via adaptive manipulation traces extraction network", 2021.