DPV-VPP: A Dual-Layer Video Privacy Protection Model Design Combining Differential Privacy and Variational Autoencoder-Based Face Replacement

Junfeng Wu
Academy of Fine Arts, Weifang University, Weifang 261061, China
E-mail: wujunfeng9809@163.com

Keywords: communication privacy, differential privacy, face replacement, video call, encoder-decoder

Received: May 22, 2025

To tackle the risk of visual content privacy leaks during video calls, the study proposes a two-layer protection method combining differential privacy with variational autoencoder-based face replacement. The first layer uses a 3D convolutional structure based on optical flow to extract temporal features. It also applies a block-level cropping perturbation to sensitive areas, ensuring frame consistency and effective privacy masking. In the second layer, a variational autoencoder is uses to replace faces, achieving natural transitions via semantic generation and boundary fusion. Experiments on the Celeb-DF dataset show the method achieves a 96.9% privacy protection success rate, 3.7% false negative rate, and 96.8% misdirection success rate against attacks. In simulated platform attack tests, the protection success rates against cross-site scripting injection and forged request attacks were 99.2% and 98.9%, respectively. In 95.1% of the test video frames, the system processing rate reached 30 frames per second, with a minimum CPU usage of 0.9% during processing. The results indicate that the method ensures visual privacy security while maintaining good real-time performance and deployment adaptability.

Povzetek: Razvita je DPV-VPP dvoslojna zaščita videoklicev: optični tok + 3D konvolucije z blokovnimi DP-motnjami ter VAE zamenjava obraza. Na Celeb-DF doseže dobre rezultate, je nizka poraba, visoka odpornost na XSS/CSRF.

1 Introduction

With the rapid development of mobile communication and Internet technologies, various network attacks are also evolving. Traditional privacy protection methods can no longer resist these advanced attacks, putting video call content at risk of being monitored, stolen, or tampered with [1–2]. Conventional privacy protection systems show limitations when facing these upgraded threats [3]. Therefore, there is an urgent need for a privacy protection method tailored to video calls that can counter new forms of network attacks. Differential Privacy (DP) protects private data by adding random noise that distorts the original data and prevents attackers from inferring sensitive information [4]. Encoder-decoder frameworks can encrypt data by converting its structure [5]. Based on this, this paper designs a visual content privacy protection algorithm using the disturbance capability of DP and the optical flow estimation technique. A face replacement method is also designed using a variational autoencoder to protect sensitive information. Finally, the disturbance algorithm and face replacement method are integrated into a visual content data privacy protection model for video calls named DP and Variational Visual Privacy Protection (DPV-VPP). This model provides dual-layer protection for both sensitive and global data, enhancing overall privacy protection. The study aims to construct a dual-layer visual privacy protection model that integrates DP and VAE to protect sensitive information in video call scenarios. The goal is to ensure the privacy and security of key areas of video data while balancing the system's real-time processing capabilities and computing resource consumption, thereby improving the model's practicality and adaptability in complex communication environments.

2 Related works

DP has been a reliable method for protecting privacy and promoting data sharing. Many researchers domestically and internationally have conducted extensive studies on DP. For example, in response to the issue where attackers use different classifiers to bypass defenses, leading to poor protection performance, Zhang et al. put forward a statistical privacy method by using the statistical analysis capability of DP. They validated its effectiveness in improving privacy protection success through experiments and evaluations [6]. Zhang's team also applied DP to optimize the performance of federated learning, aiming to provide strong privacy protection for users and to overcome the limitations of traditional methods. Their final experimental results confirmed the effectiveness of this method [7]. Encoder-decoder models can re-encode various types of data to support later processing. Because of their advantages in data handling, many scholars have studied encoder-decoder frameworks

in different applications. For instance, García's team applied encoder-decoder models to medical language processing. They used the framework to translate global languages into Spanish and solved the gap in Spanish electronic health records. Clinical tests across 17 datasets demonstrated the feasibility of this approach [8]. To address diagnostic errors caused by blurry chest X-ray images, Ukwuoma et al. integrated convolutional neural networks with transformer encoders and proposed a learning model with strong feature extraction capabilities. Their experimental results confirmed the model's high accuracy and outstanding classification performance [9].

Faced with growing concerns over privacy leaks, many researchers have worked on methods to achieve high success rates in privacy protection. For example, to reduce privacy exposure caused by recommendation systems, Chen et al. put forward a privacy-preserving federated collaborative filtering scheme. Simulation results showed the scheme achieved high accuracy while also reducing communication overhead [10]. Liang et al. designed a personal data protection method using consortium blockchain. By combining blockchain with distributed private cluster storage, they encrypted and protected private data. Simulation tests confirmed the strong practicality of this method [11]. To solve the problem of communication data leakage between patients and hospitals caused by centralized artificial intelligence training, Ali et al. discussed how federated learning could be used to address the issue. They also explored strategies for protecting private data in future smart healthcare systems [12]. In order to avoid the negative impact of privacy leaks on data sharing among network users, Li's team built a secure data sharing scheme for the Internet of Things based on blockchain. Simulation results demonstrated that the scheme was both secure and efficient [13]. Facing potential privacy leaks at the edge of 6G networks, Mao et al. analyzed the strengths and weaknesses of various countermeasures. Their findings offered useful guidance for future research on privacy protection in 6G communication systems and supported the development of safer 6G networks [14]. Larriba et al. addressed the issue of low trust in electronic voting systems by proposing the introduction of political parties as active partners in elections and using blockchain technology to build a voting system that is open and auditable by third parties, thereby enhancing the credibility of the voting system [15].

In summary, current privacy protection methods in various fields can defend against some types of attacks but still show limited performance and poor generalization against more advanced threats. DP offers a way to disrupt data and achieve global protection. Face replacement based on variational autoencoder can replace sensitive information, thereby safeguarding key content. Most current research focuses on privacy protection for structured data or static images, lacking dynamic protection mechanisms for visually sensitive information such as faces in video sequences. Therefore, the proposed DPV-VPP model simultaneously applies data perturbation and face replacement to protect video communication data. This dual-layer approach is expected to enhance user communication security across various scenarios. Table 1 summarises the details of the comparison between the existing methods and the proposed method.

3 Construction of visual content data privacy protection model for video calls

3.1 Design of visual content privacy protection algorithm based on DP

Facing the privacy protection of video content, most methods convert video into a set of images, thereby reducing the problem to image-level privacy protection. However, since video data are continuous, adjacent frames often share high similarity [16]. Attackers may exploit complementary information from these frames to restore video content. To address this issue, this study introduces an optical flow estimation algorithm. The optical flow estimation identifies the position of perturbation noise based on pixel motion to ensure consistency between adjacent processed frames and preventing attackers from inferring video content [17-18]. The structure of the optical flow estimation algorithm is shown in Figure 1.

Table 1: Comparison of different methods.

Method	Mechanism	Target domain	Methodological limitations
Zhang et al. [6]	DP	Traffic data protection	Applies noise only to static traffic packets; lacks modeling of continuous data streams
Zhang et al. [7]	DP + federated learning	Federated learning	Focused on parameter perturbation; not applicable to multimodal or visual content
García et al. [8]	Encoder-decoder architecture	Medical language processing	Designed for text vector transformation; not transferable to image/video scenarios
Ukwuoma et al. [9]	CNN + Transformer	Medical image recognition	Operates on static images; lacks temporal modeling for video content
Chen et al. [10]	Federated collaborative filtering	Recommender systems	No support for visual input; unsuitable for image/video privacy protection
Liang et al. [11]	Blockchain + distributed storage	Personal data encryption	Focuses on data encryption and storage; lacks content disturbance or replacement mechanisms
Ali et al. [12]	Federated learning	Smart healthcare communication	Emphasizes secure parameter updates; neglects facial privacy in video frames
Li et al. [13]	Blockchain	IoT data sharing	No design for visual content protection; limited applicability to video-based scenarios
Mao et al. [14]	Survey of Security Strategies	6G Edge Communication	Provides strategic overview without concrete algorithmic implementation

Larriba et al. [15]	Blockchain + Multi-party voting	Electronic Voting Privacy	Highly application-specific; lacks generalizability to video communication privacy
This paper	DP + VAE-based face replacement	Video call privacy protection	-

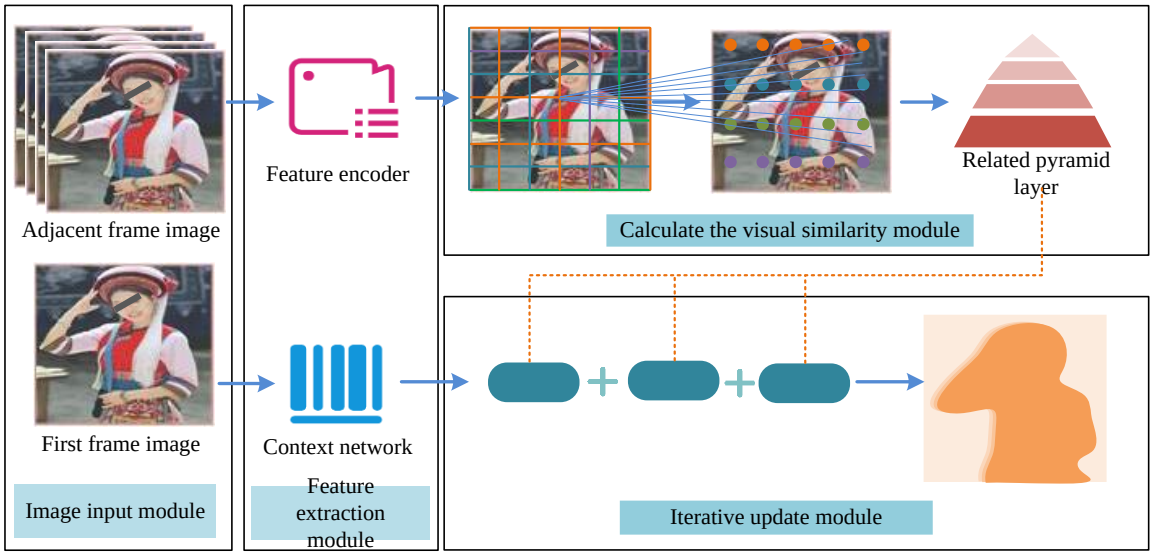


Figure 1: Structure diagram of optical flow estimation algorithm.

As shown in Figure 1, after receiving adjacent video frames from the image input module, the algorithm extracts feature through a feature encoder and a context network. Then, the image similarity calculation module divides the images into multiple regions and assigns relevance factors to each one. The relevance of other regions is determined based on the position of pixel values. The relationship between two adjacent frames is calculated as shown in Equation (1).

$$\begin{cases} C_{AB} = \lambda_k \cdot region(A) \otimes region(J(A)) \\ C_V = \underset{\substack{A=A_1A_2\cdots A_n \\ B=B_1B_2\cdots B_n}}{\cup} C_{AB} \end{cases} \tag{1}$$

In Equation (1), A and B represent region indexes, $J(\cdot)$ indicates the mapping between regions, λ_k is an adaptive relevance factor within the range [0, 1], $region(\cdot)$ and $\otimes$ refer to the dot product between image regions and feature maps, and C_V represents the computed correlation. Multiple convolutional kernels then extract four-dimensional relational features. This preserves high resolution and enables the computation of subtle motions. The optical flow sequence is updated iteratively to complete the estimation, as shown in Equation (2).

$$f_{k+1} = f_k + \Delta f_k \tag{2}$$

In Equation (2), the current optical flow f_{k+1} is updated by adding f_k to obtain Δf_k, and after k iterations, the sequence $\{f_1, f_2, \cdots, f_n\}$ is formed. For current optical flow displacement, three parallel small convolutional kernels are used, as shown in Equation (3).

$$\begin{cases} F_{flow}^{i} = \mathrm{Re}LU(Conv_{3\times3}(f_k)), i = 1,2,3 \\ F_{export} = cat(F_{flow}^{1}, F_{flow}^{2}, F_{flow}^{3}) \end{cases} \tag{3}$$

In Equation (3), $F_{flow}^{i}(i=1,2,3)$ denotes the extracted features, i is the number of kernels, $cat(\cdot)$ represents the concatenation of F_{flow}^{1}, F_{flow}^{2}, and F_{flow}^{3}, and F_{export} is the final result. This process ensures accurate optical flow estimation. Effective video content feature extraction requires both spatial and temporal features. 3D convolution captures features across both dimensions [19]. Therefore, this study uses 3D convolution to extract temporal features and enhance feature completeness. However, after applying optical flow and 3D convolution, the feature maps become large and increase the burden on mobile devices. Based on the difference between feature maps and video frames, this study adjusts the perturbation degree: no processing for minor differences and stronger perturbation for significant differences. A convolutional kernel feature analysis is added to 3D convolution to classify feature maps, as shown in Figure 2.

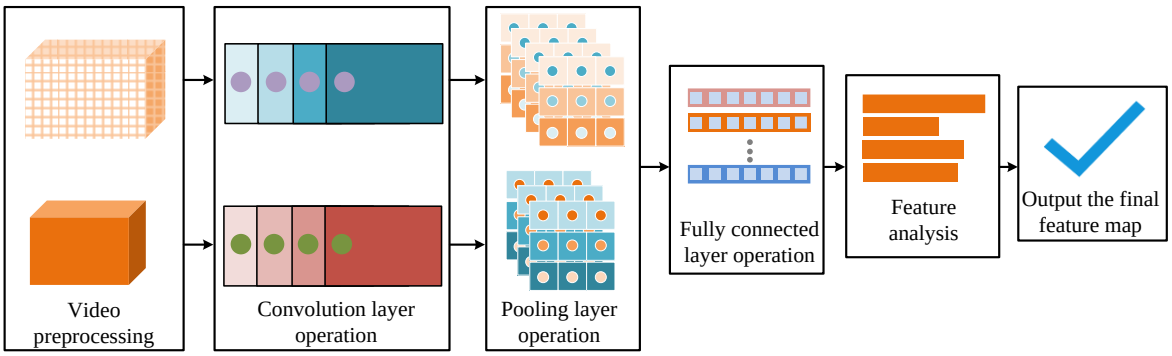


Figure 2: Schematic diagram of the feature analysis 3D convolution operation process.

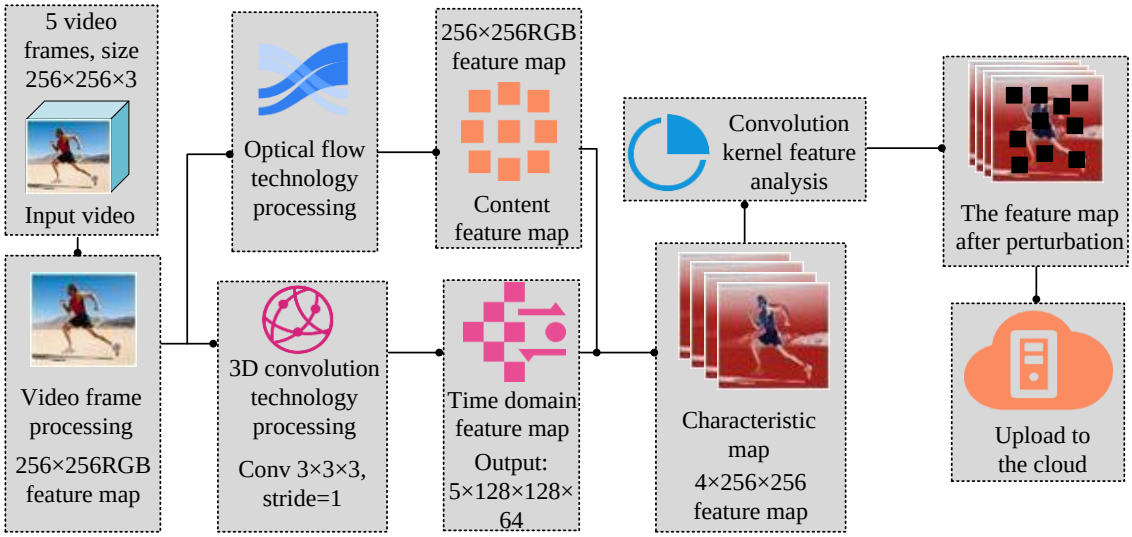


Figure 3: DP-VPM mechanism structure diagram.

Figure 2 shows that the 3D convolution first normalizes the video frames in size and pixel values. The normalized frames are passed through convolutional layers to capture spatial-temporal relationships. Then, 3D max pooling is used to reduce the dimensions of the feature maps. A fully connected layer converts the map into a 1D vector. Finally, similarity with the original video determines the perturbation level. Structural similarity is used to measure this, considering brightness, contrast, and structure. A higher score indicates higher similarity. The brightness similarity function is shown in Equation (4).

$$l(f,g)=\frac{2\mu_f\mu_g+C_1}{\mu_f^2+\mu_g^2+C_1} \tag{4}$$

In Equation (4), μ_f and μ_g represent the brightness means of the feature and original images. The contrast similarity function is shown in Equation (5).

$$c(f,g)=\frac{2\sigma_f\sigma_g+C_2}{\sigma_f^2+\sigma_g^2+C_2} \tag{5}$$

In Equation (5), σ_f and σ_g are the standard deviations of the feature and original images. The structural similarity function is shown in Equation (6).

$$s(f,g)=\frac{\sigma_{fg}+C_3}{\sigma_f\sigma_g+C_3} \tag{6}$$

In Equation (6), $s(f,g)$ is the calculated similarity value. C_1, C_2, and C_3 in Equations (4), (5), and (6) are constants used to prevent zero denominators. The final similarity score is calculated as shown in Equation (7).

$$SSIM(f,g)=l(f,g)\cdot c(f,g)\cdot s(f,g) \tag{7}$$

In Equation (7), $SSIM(f,g)$ is the comprehensive similarity score. To clearly distinguish between "minor differences" and "significant differences" in feature maps, the study classified each region based on the structural similarity index. The specific classification criteria are as follows: if the structural similarity index is greater than 0.85, the region is classified as having minor difference region and mild perturbation is applied. When the score is <0.65, it is classified as a significant difference region and strong perturbation is applied. Traditional perturbation methods based on DP suffer from low data usability and insufficient protection [20]. To address this, this study combines pixel-level noise with a block-level mosaic approach to propose a novel block-cutting perturbation mechanism. It first generates a matrix to determine the center of the region to be perturbed, converting pixel-wise noise into block-wise perturbation. Then, the pooling operation in the mosaic is replaced by pixel zeroing. Combining this with video feature extraction forms a DP-based visual content protection mechanism named DP-VPM. Its structure is shown in Figure 3.

As shown in Figure 3, the DP-VPM mechanism follows these steps: first, convert the target video into image frames. Second, extract spatial and temporal features via optical flow and 3D convolution modules. Third, analyze feature maps using kernel-based classification to assign appropriate perturbation levels. Fourth, apply block-cutting perturbation to finalize protection. The perturbation mechanism consists of three closely connected stages: optical flow estimation, structural similarity analysis, and block-level perturbation execution. The optical flow estimation algorithm analyzes pixel motion between adjacent video frames and generates a map of motion intensity. Subsequently, the structural similarity index analysis is introduced within the candidate regions to measure the visual consistency between each region and the original frame. Based on the values of the structural similarity index, the perturbation intensity is categorized into different levels. Regions with high similarity do not require perturbation, those with moderate similarity are subjected to light perturbation, and regions with low similarity receive strong perturbation. This adaptive approach allows for precise control of perturbation levels based on visual similarity. Finally, the system executes the block-based perturbation mechanism according to the positions identified by optical flow and the intensity levels determined by the structural similarity index. For light perturbation, pixel-level Gaussian noise is applied, while strong perturbation involves setting entire pixel blocks to zero. The study sets the encoder and context network in the optical flow estimation module to include four layers of convolution operations, with a 3×3 kernel size, a stride of 1, padding of 1, and channel numbers of 64, 128, 128, and 256 for each layer, respectively. In the structural similarity assessment, the brightness, contrast, and structural constants are 0.01, 0.03, and 0.015, respectively, and the structural similarity threshold is 0.85. In the DP-VPM module, the disturbance noise is sampled from a Gaussian distribution $N(0,\sigma^2)$, and the variance σ^2 is dynamically adjusted according to the scene. In low-motion scenes, σ^2 is 0.04, and in high-motion scenes, it is 0.08. This noise is initially applied at the pixel level. It is then transformed into block-level zero-value masks using a regional occlusion mechanism, which enhances privacy protection while preserving visual continuity.

3.2 Construction of Visual Content Privacy Protection Model for Video Calls

Although the VPM mechanism protects overall video content, attackers may still recover critical details like faces, compromising privacy [21]. Therefore, this study builds a face replacement model using the Variational Autoencoder (VAE), which models latent features of video frames probabilistically. The structure is shown in Figure 4.

As shown in Figure 4, the VAE consists of an encoder and decoder. The encoder analyzes the input image to obtain the probability values of facial features. The decoder reconstructs the encoded values into a new arrangement for output. To enhance the expressive capability of facial information reconstruction, the VAE designed by the research institute adopts a symmetric structure, with the encoder and decoder each consisting of four convolutional layers and two fully connected layers. The input image size is 256×256×3, which is compressed into a latent variable vector with a dimension of 128 after encoding. The decoder reconstructs the image from this latent space. All convolutional layers use a 3×3 convolutional kernel with a stride of 2, padding of 1, and the ReLU activation function. During training, the total loss function of the VAE consists of two parts, with a balancing coefficient of 0.1 for the loss weights. The first is the pixel-level mean squared error loss between the input image and the reconstructed image, and the second is the Kullback-Leibler divergence between the encoder output distribution and the standard normal distribution. Additionally, the model is trained for 120 epochs using the Adam optimizer with a learning rate of 0.0002 and a batch size of 32. During reconstruction, latent vectors are obtained from inputs as described in Equation (8).

$$Z \sim q_{\varphi}(z|x) \tag{8}$$

In Equation (8), Z is the latent vector, φ denotes encoder weights, x is the data sample, and Z is the output latent vector. Then, the image set is reconstructed from the latent vector, as shown in Equation (9).

$$X' \sim p_{\theta}(z)p_{\theta}(x'|z) \tag{9}$$

In Equation (9), X' is the reconstructed image, θ denotes decoder weights, and x' is a data sample. The encoder's posterior distribution is expected to be $p(z|x)$, which is approximated using $p(z|x)$ instead of $q(z|x)$. To minimize the difference between q and p, the KL divergence between q and p is minimized, as shown in Equation (10).

$$\min KL(q(z|x)\|p(z|x)) \tag{10}$$

Equation (10) ensures consistency between q and p, allowing accurate inference of complex distributions. Using VAE, it is possible to generate realistic fake faces that resemble the original and capture multiple angles. Based on this, the study proposes a face replacement method named VA-FR, shown in Figure 5.

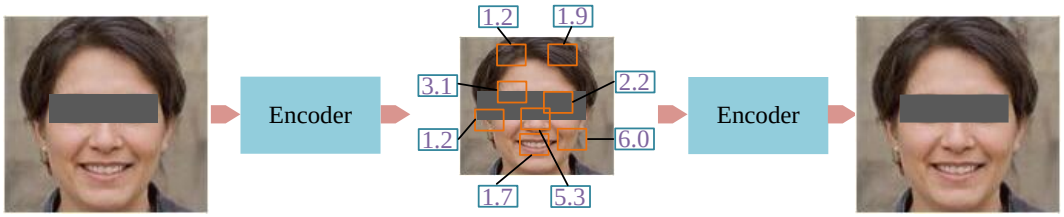


Figure 4: Schematic diagram of the VAE.

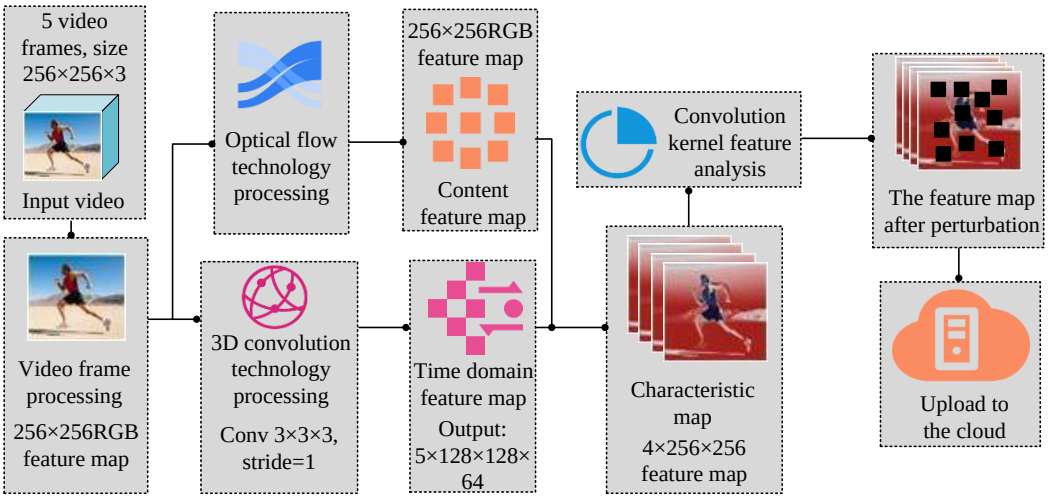


Figure 5: Flowchart of VA-FR face replacement method.

As shown in Figure 5, the first stage of face detection uses the RetinaFace model based on the ResNet-50 backbone network. After detection, the faces are geometrically aligned using a five-point affine transformation to standardize the face scale and improve the accuracy of subsequent reconstruction. In the face segmentation stage, a supervised semantic segmentation model based on the U-Net architecture is used to extract the foreground face region. In the experiment, the frame discard rate due to segmentation failure was approximately 2.1%, mainly concentrated in overexposed or blurred frames. Therefore, in terms of feature modeling, the VAE structure is symmetrically composed of an encoder and a decoder, each containing four convolutional layers and two fully connected layers, with the latent variable dimension set to 128. In the final output stage, the system uses a Poisson fusion algorithm to perform edge smoothing and lighting adjustment on the replaced face, and completes frame rate synchronization to ensure the naturalness and continuity of the generated video in terms of visual perception. The content smoothing is described in Equation (11).

$$\min_{f} \iint_{\Omega} |\nabla f - v|^2 \tag{11}$$

In Equation (11), Ω is the foreground region of the synthesized image, and f is the pixel representation function in the merged image Ω. To improve the natural transition of synthetic images at the boundaries, a boundary smoothing mechanism was introduced in this study. First, a Poisson mixture algorithm is used to gradientally blend the foreground region with the background image, ensuring consistency in brightness and texture at the boundary. Specifically, the pixel values in boundary region $\partial\Omega$ are not directly taken from values outside the image but are adjusted based on the solution of the Laplace operator in the Poisson equation, thereby constructing a smooth transition of pixel value distribution in the boundary region. This process is illustrated in Equation (12).

$$\begin{cases} \Delta f^* = \Delta f_s & \text{in} \quad \Omega \\ f^* = f_t & \text{on} \quad \partial\Omega \end{cases} \tag{12}$$

In Equation (12), $\partial\Omega$ is the boundary of Ω, and f^* is the pixel representation function outside the boundary. f_t represents the gradient information of the source image. f_s represents the boundary value of the target background image; Δ represents the Laplace operator. The VA-FR method replaces key facial information in video content to further protect privacy. Finally, this study combines the VA-FR and VPM mechanisms to build the DPV-VPP model. Its structure is shown in Figure 6.

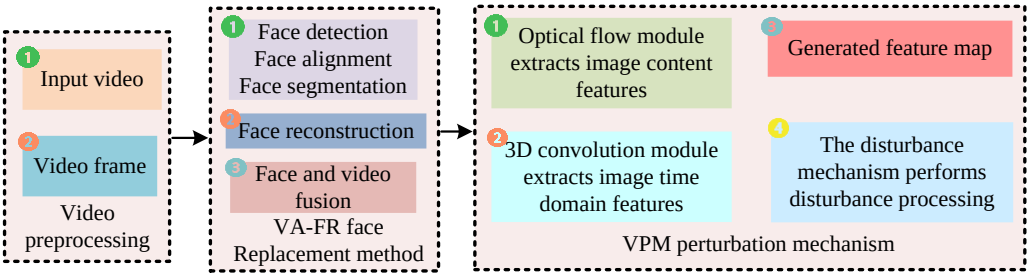


Figure 6: DPV-VPP model structure diagram.

As shown in Figure 6, the DPV-VPP model preprocesses the target video into frames. Then, the VA-FR mechanism replaces sensitive face information. Finally, the VPM mechanism perturbs the video frames, achieving two-layer protection.

4 Performance evaluation of the DPV-VPP model

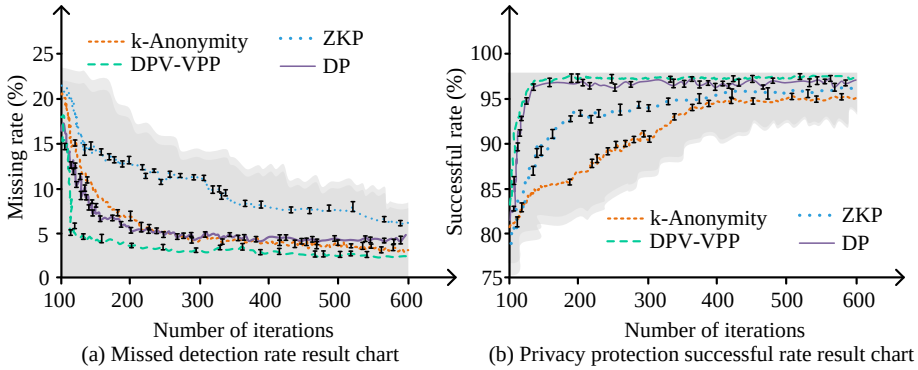
4.1 Experimental environment and training results of the DPV-VPP model

After the construction of the DPV-VPP model, in order to evaluate its performance in protecting privacy in video calls, the study introduced three domain-related models—K-Anonymity, DP, and Zero-Knowledge Proof (ZKP)—as comparison models. A Huawei Mate 30 device with a Kirin 990 chip was used as the local mobile terminal, while the cloud server was equipped with an Intel Xeon E5-2682 V4 CPU and an NVIDIA Tesla P4 GPU. The programming language used was Python, and the operating system was Windows 10. The Celebrity Deepfake Detection (Celeb-DF) dataset was used as the experimental dataset to provide test samples. Details of the Celeb-DF dataset are shown in Table 2.

As shown in Table 2, the Celeb-DF dataset suffers from a significant class imbalance problem, with the number of face-swapped videos (5639) far exceeding that of real videos (590), resulting in a ratio of real to fake samples of approximately 1:10. This severe imbalance may cause the model to favour identifying fake samples during training while neglecting its ability to distinguish real samples, leading to certain generalisation errors in real-world applications. To mitigate this bias, the study introduced a category weight adjustment mechanism and a stratified sampling strategy during training, and also adjusted the loss function with category weights. Based on the above experimental environment and the Celeb-DF dataset, the study first conducted experiments on the missed detection rate and privacy protection success rate for the four privacy protection models. The missed detection rate referred to the proportion of non-sensitive information that was incorrectly protected, while the privacy protection success rate referred to the proportion of sensitive information that was correctly identified and protected. The results represent the average value of five independent experiments run under the same data set and parameter configuration, with error bars representing the 95% confidence interval. The results are shown in Figure 7.

Table 2: Celeb-DF dataset details.

Parameter	Details
The number of themes	590
The number of deepfake videos	5639
Average duration	13s
Standard frame rate	30 frames per second
Video format	MPEG4.0



(a) Missed detection rate result chart (b) Privacy protection successful rate result chart

Figure 7: Results of missed detection rate and privacy protection success rate.

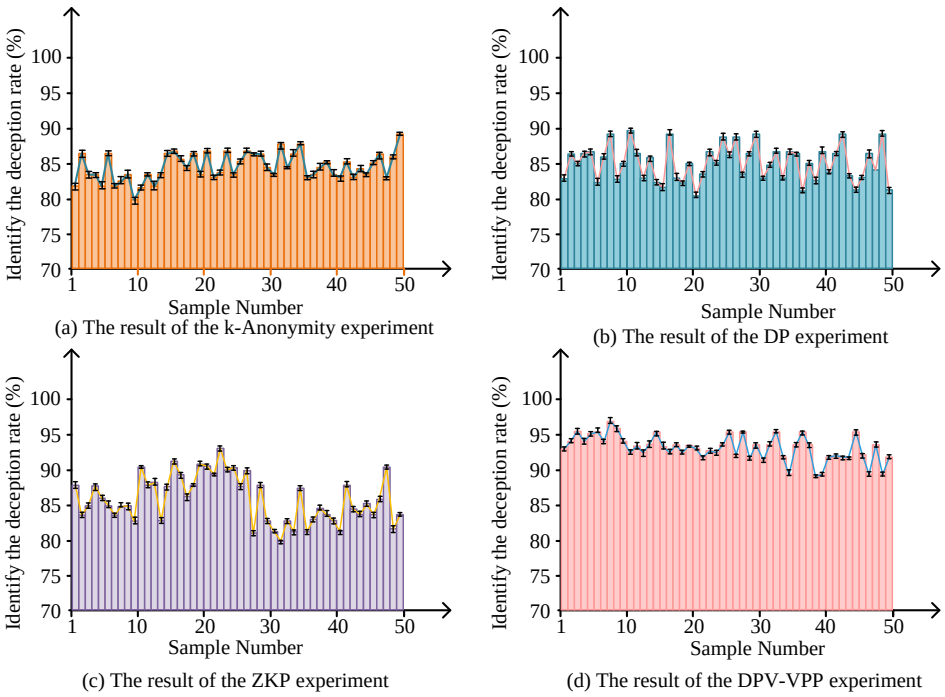


Figure 8: Successful deception rate experimental results.

As shown in Figure 7(a), the missed detection rates of the four models decreased steadily as the number of training iterations increased. When the number of iterations reached 389, 267, 214, and 145 respectively, the missed detection rates of ZKP, K-Anonymity, DP, and DPV-VPP stabilized at 7.4%, 5.2%, 5.8%, and 3.7%. These results showed that DPV-VPP achieved a lower missed detection rate compared to the other three models, indicating better performance in identifying sensitive information. As illustrated in Figure 7(b), after training, the privacy protection success rates of ZKP, K-Anonymity, and DP stabilized at 94.3%, 93.9%, and 96.1% respectively. The DPV-VPP model achieved a success rate of 96.9%, which was higher than the other three models. The results in Figure 7(a) and Figure 7(b) demonstrate that DPV-VPP achieved favorable performance in terms of both missed detection rate and privacy protection success rate, providing reliable data support for subsequent experiments. Next, the study conducted a comparison experiment on the deception success rate of the four models. The deception success rate referred to the probability that a privacy protection model successfully misled and deceived attackers, causing them to analyze or attack incorrect information. The study tested 50 samples using the ZKP, K-Anonymity, DP, and DPV-VPP models. The experimental results are shown in Figure 8.

Figure 8 presents the deception success rates of the four models. As shown in Figure 8(a), the K-Anonymity model reached a highest deception success rate of 87.5%. According to Figures 8(b) and 8(c), the highest deception success rates of the DP and ZKP models were 88.1% and 91.2%, respectively. Figure 8(d) shows that the DPV-VPP model achieved the highest deception success rate of 96.8%, surpassing the other three models. These results indicated that the DPV-VPP model successfully disturbed the original call data and effectively misled attackers, thereby reducing the attack success rate.

4.2 Practical performance evaluation of the DPV-VPP model

After verifying the training performance of the DPV-VPP model, the study further evaluated its practical applicability. The experimental environment and dataset remained consistent with the training experiments. The study first conducted experiments on the number of video frames processed per second for the four models. This metric measured whether the privacy protection model affected the smoothness of terminal usage. If the number of frames processed per second exceeded 30 fps, it indicated that the model did not cause noticeable latency. The results are presented in Figure 9.

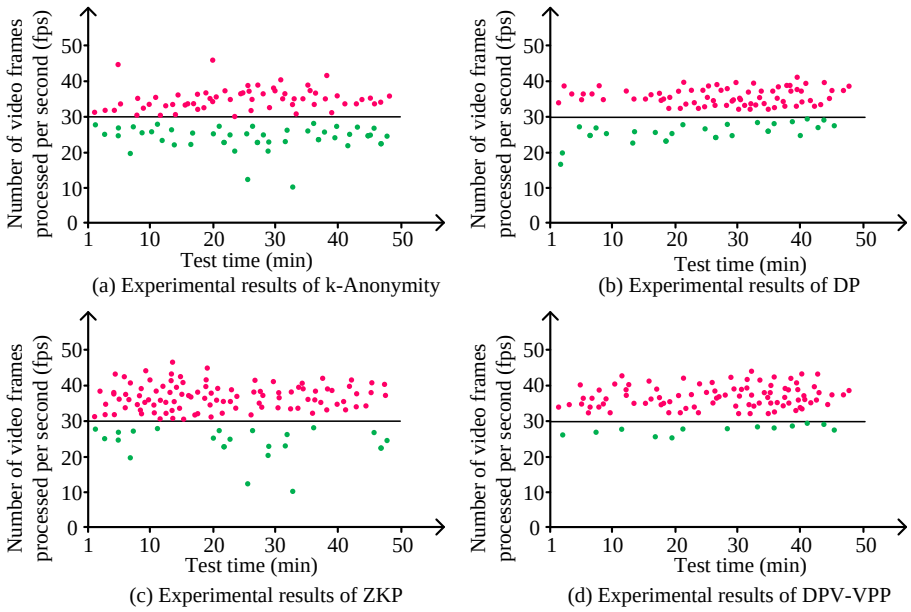


Figure 9: Experimental results of processing video frames per second.

In Figure 9, the average frame rate of the DPV-VPP model is approximately 34.0 fps, the DP model is approximately 32.0 fps, the ZKP model is approximately 30.5 fps, and the K-Anonymity model is approximately 29.0 fps. Based on the calculation of intra-sample variance, the 95% Confidence Intervals (CI) for the frame rates of all models are controlled within ±1.5-2.2 fps, indicating that the models exhibit strong real-time stability. As shown in Figure 9(a), the K-Anonymity model processed more than 30 frames per second in 75.1% of the samples, which was relatively low and could affect the smooth performance of the terminal. According to Figures 9(b) and 9(c), the proportions for the DP model were 84.4% and 91.1%, indicating a certain degree of impact on performance. Figure 9(d) shows that the DPV-VPP model achieved processing speeds above 30 frames per second in 95.1% of the cases, suggesting minimal impact on device smoothness. These results demonstrated the practicality and reliability of the DPV-VPP model in real-world applications. Subsequently, the study conducted experiments on the attack protection success rate of the four models. Two types of attacks were simulated: Cross-Site Scripting (XSS) and Cross-Site Request Forgery (CSRF), representing different levels of attack intensity. Although XSS and CSRF attacks typically target the platform logic layer, in actual video call systems, attackers can bypass video desensitisation modules by forging application programming interface requests or injecting scripts, thereby submitting unprotected raw image frames and causing user privacy leaks. The study deployed the DPV-VPP module in the front-end video capture process, using structural perturbation and face replacement mechanisms to ensure that even if the interface is tampered with, the system cannot access the original visual content. Therefore, XSS/CSRF attack simulations were introduced to verify the proposed method's indirect protective capabilities against potential visual content leakage attacks. In the simulated XSS and CSRF attack experiments, the study used a black-box attack method to test the protection capabilities of different visual privacy protection models. Attackers could not access model parameters and were only able to submit video frames embedded with attack payloads via standard HTTP interfaces. The results are shown in Figure 10.

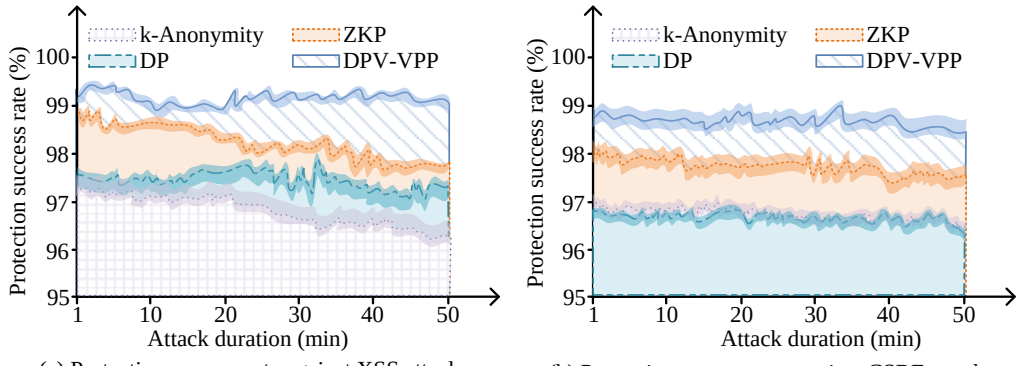
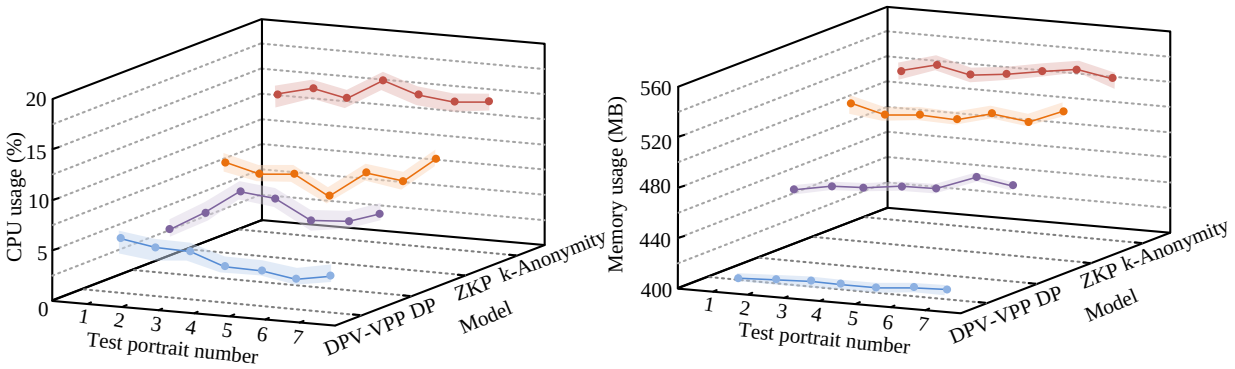


Figure 10: Protection success rate against XSS attacks and CSRF attacks.



(a) CPU usage experimental results

(b) Memory usage experimental results

Figure 11: CPU and memory usage experimental results.

As shown in Figure 10(a), after 50 minutes of XSS attacks, the protection success rates of ZKP, K-Anonymity, DP, and DPV-VPP were 97.8%, 96.5%, 97.4%, and 99.2%, respectively. The DPV-VPP model achieved the highest success rate among the four. According to Figure 10(b), after 50 minutes of CSRF attacks, the protection success rates of ZKP, K-Anonymity, DP, and DPV-VPP were 97.9%, 96.8%, 96.6%, and 98.9%, respectively. Again, DPV-VPP outperformed the other models. These results indicated that DPV-VPP consistently provided effective protection under attacks of varying complexity, demonstrating strong generalization capabilities. Finally, to further verify the practical applicability of the DPV-VPP model, the study evaluated CPU and memory usage while each of the four models processed seven test samples. The results are shown in Figure 11.

As shown in Figure 11(a), the K-Anonymity model exhibited high CPU usage, exceeding 10% in every test sample. For the ZKP and DP models, the highest CPU usage rates were 7.8% and 4.9%, while the lowest were 4.8% and 2.5%, respectively. In contrast, the DPV-VPP model achieved significantly better performance, with a maximum CPU usage of only 3.8% and a minimum of 0.9%. As shown in Figure 11(b), the DPV-VPP model consistently maintained memory usage below 405 MB across all test samples, significantly outperforming other models and demonstrating better resource stability and deployment adaptability. These results demonstrated that the DPV-VPP model did not interfere with normal call operations, further validating its excellent performance in practical scenarios.

5 Discussion

Compared with traditional differential privacy methods, DPV-VPP integrates two layers of protection mechanisms into its structural design. Compared with References [7] and [10], DPV-VPP combines a dynamic perturbation algorithm based on optical flow estimation and structural similarity analysis to adaptively adjust the perturbation intensity, effectively addressing privacy-sensitive areas of varying degrees in videos. Additionally, the VA-FR face replacement strategy based on VAE achieves deep semantic replacement and smooth boundary fusion in the target face region, addressing the limitations of traditional occlusion or blurring methods in terms of visual deceptiveness. Experiments show that DPV-VPP outperforms existing ZKP models, K-Anonymity processing methods, and GAN-based disguise generation techniques, particularly in terms of false detection rate (3.7%) and deception success rate (96.8%).

Furthermore, in terms of system resource control, the DPV-VPP model also demonstrates excellent real-time processing capabilities and terminal adaptability. In 95.1% of video frames, the frame rate exceeds 30fps, meeting the smoothness requirements for video call applications. In video tests on seven samples, CPU usage dropped as low as 0.9%, and memory usage remained under 406MB, with resource overhead significantly better than the multi-stage convolution-based face blurring processing methods proposed in References [11] and [14]. In terms of platform security testing, DPV-VPP achieved interception rates of 99.2% and 98.9% in evaluations against XSS and CSRF forgery attacks, respectively.

However, running dynamic perturbation and VAE replacement in parallel causes slightly higher memory usage when processing high-resolution videos. Future research will explore lightweight network architectures or model pruning optimisation strategies. Additionally, XSS and CSRF attacks primarily target platform interfaces and transmission processes. The protection provided by this method is primarily manifested in the irreversibility of content after front-end data perturbation and face replacement, representing an 'indirect protective effect' rather than a core design objective of the method itself. Therefore, the experiments in this section serve primarily as a reference for usability and compatibility verification in a system integration context. Future research will further focus on the portability and resource adaptation capabilities of DPV-VPP on mobile devices and edge computing platforms to enhance its engineering practicality.

6 Conclusion

Facing the continuous evolution of cyberattacks, traditional privacy protection methods have become increasingly ineffective in safeguarding users' call privacy. Therefore, this study put forward a dual-layer

privacy protection model, DPV-VPP, by combining a perturbation mechanism based on DP with face replacement using a VAE. Experimental results showed that DPV-VPP not only provided reliable privacy protection but also offered strong practical performance, meeting the demands of modern communication for privacy protection.

Although this study validated the privacy protection performance and practical applicability of the DPV-VPP model, there are still certain limitations. The study has not yet been deployed and validated in a real remote presentation system architecture platform, and there is a lack of testing of generalisation capabilities under complex facial expressions or lighting conditions. In the future, we will expand the adaptability of multi-person interaction scenarios, enhance adversarial robustness, and strengthen lightweight deployment capabilities.

References

[1] Daniele Scarpi, Gabriele Pizzi, and Shashi Matta. Digital technologies and privacy: State of the art and research directions. Psychology & Marketing, 39(9):1687-1697, 2022. https://doi.org/10.1002/mar.21692

[2] Md ABU IMRAN Mallick, and Rishab Nath. Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. World Scientific News, 190(1):1-69, 02024.

[3] Ranjan Chaudhuri, Sheshadri Chatterjee, and Demetris Vrontis. Antecedents of privacy concerns and online information disclosure: Moderating role of government regulation. EuroMed Journal of Business, 18(3):467-486, 2023. https://doi.org/10.1108/emjb-11-2021-0181

[4] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 84(1):3-37, 2022. https://doi.org/10.48550/arXiv.1905.02383

[5] Jin H. Bae, Ruolin Liu, Eugenia Roberts, Erica Nguyen, Shervin Tabrizi, Justin Rhoades, Timothy Blewett, Kan Xiong, Gregory Gydush, Douglas Shea, Zhenyi An, Sahil Patel, Ju Cheng, Sainetra Sridhar, Mei Hong Liu, Emilie Lassen, Anne-Bine Skytte, Marta Grońska-Pęski, Jonathan E. Shoag, Gilad D. Evrony, Heather A. Parsons, Erica L. Mayer, G. Mike Makrigiorgos, Todd R. Golub, and Viktor A. Adalsteinsson. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. Nature Genetics, 55(5):871-879, 2023. https://doi.org/10.1038/s41588-023-01376-0

[6] Xiaokuan Zhang, Jihun Hamm, Michael K. Reiter, and Yinqian Zhang. Defeating traffic analysis via differential privacy: A case study on streaming traffic. International Journal of Information Security, 21(3):689-706, 2022. https://doi.org/10.1007/s10207-021-00574-3

[7] Lefeng Zhang, Tianqing Zhu, Ping Xiong, Wanlei Zhou, and Philip S. Yu. A robust game-theoretical federated learning framework with joint differential privacy. IEEE Transactions on Knowledge and Data Engineering, 35(4):3333-3346, 2022. https://doi.org/10.1109/TKDE.2021.3140131

[8] Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. A comparative analysis of Spanish Clinical encoder-based models on NER and classification tasks. Journal of the American Medical Informatics Association, 31(9):2137-2146, 2024. https://doi.org/10.1093/jamia/ocae054

[9] Chiagoziem C Ukwuoma, Zhiguang Qin, Md Belal Bin Heyat, Faijan Akhtar, Olusola Bamisile, Abdullah Y Muaad, Daniel Addo, and Mugahed A Al-Antari. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. Journal of Advanced Research, 48:191-211, 2023. https://doi.org/10.1016/j.jare.2022.08.021

[10] Yucheng Chen, Chenyuan Feng, and Daquan Feng. Privacy-preserving hierarchical federated recommendation systems. IEEE Communications Letters, 27(5):1312-1316, 2023. https://doi.org/10.1109/LCOMM.2023.3245101

[11] Wei Liang, Yang Yang, Ce Yang, Yonghua Hu, Songyou Xie, Kuan-Ching Li, and Jiannong Cao. PDPChain: A consortium blockchain-based privacy protection scheme for personal data. IEEE Transactions on Reliability, 72(2):586-598, 2022. https://doi.org/10.1109/tr.2022.3190932

[12] Mansoor Ali, Faisal Naeem, Muhammad Tariq, and Georges Kaddoum. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. IEEE Journal of Biomedical and Health Informatics, 27(2):778-789, 2022. https://doi.org/10.1109/JBHI.2022.3181823

[13] Tian Li, Huaqun Wang, Debiao He, and Jia Yu. Blockchain-based privacy-preserving and rewarding private data sharing for IoT. IEEE Internet of Things Journal, 9(16):15138-15149, 2022. https://doi.org/10.1109/JIOT.2022.3147925

[14] Bomin Mao, Jiajia Liu, Yingying Wu, and Nei Kato. Security and privacy on 6G network edge: A survey. IEEE Communications Surveys & Tutorials, 25(2):1095-1127, 2023. https://doi.org/10.1109/COMST.2023.3244674

[15] Antonio M. Larriba, Aleix Cerdà i Cucó, José M. Sempere, and Damián López. Distributed trust, a blockchain election scheme. Informatica, 32(2):321-355, 2021. https://doi.org/10.15388/20-INFOR440

[16] Mohammad Amin Satvati, Mehrdad Lakestani, Hossein Jabbari Khamnei, and Tofigh Allahviranloo. Deblurring medical images using a new grünwald-letnikov fractional mask. Informatica, 35(4):817-836, 2024. https://doi.org/10.15388/24-INFOR573

[17] Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. Deep multi-modal network based automated depression severity estimation. IEEE transactions on affective computing, 14(3):2153-2167, 2022. https://doi.org/10.1109/TAFFC.2022.3179478

[18] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. Advances in Neural Information Processing Systems, 35:3502-3516, 2022. https://doi.org/10.48550/arXiv.2210.10716

[19] Hüseyin Firat, Mehmet Emin Asker, Mehmet İlyas Bayindir, and Davut Hanbay. 3D residual spatial-spectral convolution network for hyperspectral remote sensing image classification. Neural Computing and Applications, 35(6):4479-4497, 2023. https://doi.org/10.1007/s00521-022-07933-8

[20] Fabian Bach. Differential privacy and noisy confidentiality concepts for European population statistics. Journal of Survey Statistics and Methodology, 10(3):642-687, 2022. https://doi.org/10.48550/arXiv.2012.09775

[21] Punam Kumari, and Bhaskar Mondal. An encryption scheme based on grain stream cipher and chaos for privacy protection of image data on IoT network. Wireless Personal Communications, 130(3):2261-2280, 2023. https://doi.org/10.1007/s11277-023-10382-8