UNMASKING PRIVACY: EVALUATING THE REVERSIBILITY OF VOTER IMAGE ANONYMIZATION IN E-VOTING SYSTEMS

UDC 004.056.55
DOI: 10.56246/18294480-2025.18-08

MASTOYAN KAREN
PhD student, IIAP NAS RA
GSU lecturer
e-mail: kmastoyan@gmail.com

Electronic voting (e-voting) systems have emerged as a promising solution for modernizing elections, offering convenience, accessibility, and potential cost savings. However, such systems face significant challenges, particularly regarding privacy, security, and verifiability. Among the critical areas of concern is the protection of voter identity, especially in internet-based voting (i-voting) systems where sensitive information could be vulnerable to re-identification attacks. In prior research, an anonymization model using pixel shuffling was proposed to distort voter images and safeguard privacy. This study investigates whether these distorted images can be reversed, using techniques including brute-force computation, Convolutional Neural Networks (CNNs), and autoencoders, to assess the robustness of the anonymization. Experimental results demonstrate that the randomization introduced by pixel shuffling effectively prevents brute-force reconstruction due to its astronomical complexity. Further attempts to reconstruct original images with CNNs and autoencoders revealed the limitations of deep learning models in restoring heavily randomized images, as the spatial coherence essential for accurate reconstruction was lost. These findings underscore the strength of pixel shuffling as a privacy-preserving method.

Keywords: *E-voting systems, voter privacy, image anonymization, face recognition, data security, pixel distortion, privacy protection, internet voting*

In recent years, electronic voting (e-voting) systems have gained attention for their potential to streamline elections, improve accessibility, and reduce costs. While e-voting offers significant advantages over traditional voting methods, it also presents unique challenges, especially concerning voter privacy, security, and the verifiability of

election results. Among the various types of e-voting systems, internet-based voting (i-voting) stands out due to its convenience, allowing citizens to cast ballots remotely. However, this remote accessibility also opens doors to complex security threats, necessitating advanced measures to protect voter information.

In the previous study[1], we explored an e-voting model that incorporated face recognition and image distortion techniques to ensure voter authentication while safeguarding voter privacy. A key component of this model was the anonymization of voter images through pixel distortion, a process designed to prevent any potential linkage between voters and their cast ballots. While this approach aimed to enhance privacy, questions remain about the effectiveness and reliability of such distortion methods in preventing image re-identification. Specifically, an adversary might attempt to reconstruct the original voter image from the distorted version, potentially undermining the privacy guarantees essential to a secure e-voting system.

This study seeks to address these concerns by investigating whether distorted voter images in e-voting systems can be restored and, if so, under what conditions. Through a series of experiments, we will explore different computational techniques, including statistical analysis and machine learning methods, to determine the feasibility of reconstructing original images from anonymized data. By assessing the robustness of these anonymization techniques, this study aims to contribute valuable insights into the resilience of privacy safeguards within e-voting systems, highlighting both strengths and areas for potential improvement.

In this study, we hypothesize that while image anonymization techniques, such as pixel distortion and randomization are designed to protect voter privacy within e-voting systems, there may be potential vulnerabilities that could allow for partial or full restoration of the original images. Specifically, we posit that with advanced computational techniques, including machine learning and statistical analysis, it may be possible to reconstruct a recognizable version of a voter's face from the distorted image. This hypothesis aims to assess the robustness of current anonymization methods, testing their effectiveness in truly safeguarding voter identities against potential reidentification attacks.

In this e-voting model, the image distortion process for anonymizing voter faces begins by converting the original image to grayscale. Reducing the image to grayscale simplifies it by collapsing it into a single channel of light intensity values, which decreases data complexity and processing load, especially when working with large numbers of images. Once the image is converted to grayscale, it is transformed into a numerical array, where each element represents the intensity value of a pixel.

At this stage, pixel shuffling is applied to the grayscale image. This process involves randomly rearranging the pixel values in the array, effectively disrupting the spatial coherence that gives the original image its recognizable structure. The result is a highly distorted image that appears as noise, devoid of any facial features or identifiable information. Without knowing the specific pattern of pixel rearrangement, it would be nearly impossible to reconstruct the original face from the shuffled array. After the shuffling, the randomized array is then converted back to an image format, creating the anonymized grayscale version of the voter's face. This distorted image is subsequently used within the voting system to ensure that sensitive facial details are concealed.

Although this model utilizes pixel shuffling to achieve image distortion, this method could easily be replaced by alternative techniques, such as pixel block randomization, Gaussian blurring, or pixel intensity modification, depending on specific security requirements. However, for this study, pixel shuffling was chosen to facilitate further experiments aimed at assessing the feasibility of reconstructing original images and testing the limits of image anonymization in e-voting.

For a robust testing approach to assess the reversibility of distorted voter images, we can consider a combination of computational and machine learning techniques, along with a suitable dataset. Here's a detailed plan for both the testing methods and the data sources:

Testing Approach

Brute-Force and Combinatorial Analysis: Start with basic computational methods, like brute-force pixel rearrangement or combinatorial techniques, to attempt reconstructing small images. This method might be feasible on images with fewer pixels, providing insight into the level of distortion needed to prevent reversibility for images of varying resolutions. This would give a baseline for how computationally intense such a reversion is[2].

Machine Learning Approaches: Deep Learning with Convolutional Neural Networks (CNNs): Train a CNN model on paired datasets of original and shuffled images to recognize patterns that could suggest plausible reconstructions. This could involve creating a supervised model that learns to infer the likely locations of shuffled pixels in an attempt to restore spatial coherence in the image.

Autoencoders for Reconstruction: Use autoencoders, which are effective at reconstructing compressed or corrupted images. In this case, you can experiment with a denoising autoencoder by training it on images that have been distorted in a similar manner, then testing if the model can learn any patterns that could reconstruct the original face[3,4].

Generative Adversarial Networks (GANs): GANs could be explored to generate approximations of the original faces by learning from a dataset of both shuffled and original images. The generator could attempt to reconstruct original facial features from shuffled images, while the discriminator would evaluate the authenticity of these reconstructions[5].

Statistical Analysis with Entropy Measures: Compare the entropy levels of distorted images with their original counterparts to see if there is any detectable pattern in the randomness. This approach could provide insights into whether certain shuffling methods leave recognizable statistical traces that could make reconstruction easier.

Dataset Selection

Face Image Databases:

- LFW (Labeled Faces in the Wild): LFW is a well-known dataset with over 13,000 images of faces from diverse individuals. It is suitable for research on face recognition and would provide a range of faces to test shuffling and reconstruction techniques[6].
- CelebA (CelebFaces Attributes Dataset): With over 200,000 celebrity face images, CelebA offers a large, diverse set of faces. This dataset includes multiple angles and facial expressions, which is beneficial for testing distortion and reconstruction across varied images[7].
- FERET Database: This dataset was created for facial recognition research and includes images of individuals taken at different times, with various facial expressions, and under varying lighting conditions, making it ideal for comprehensive distortion testing[8].

Alternatively, we could generate our own custom dataset of grayscale images using a smaller, representative sample. By applying controlled shuffling patterns to this dataset, we would not only be able to test the effectiveness of different distortion techniques but also examine the potential for reversing these distortions. This

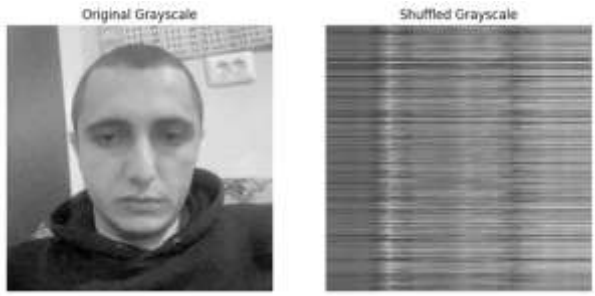

Figure 1 Original and shuffled grayscale images

controlled environment allows us to experiment with both the application of specific distortion patterns and methods aimed at recovering the original images, whether through statistical analysis or AI-based reconstruction techniques. For further experiments, we captured an image using a computer webcam with an initial resolution of 1816x1816 pixels(Figure 1). To optimize computational efficiency, we downscaled the image to a resolution of 64x64 pixels. We chose to use a picture of my own face in these tests to enhance interpretability in the recovery process, as it allows us to more clearly assess how accurately any reconstructed images resemble the original.

In designing our experiments, we aim to systematically test various methods for reversing image distortion in a controlled environment. This experiment involves multiple attempts using different computational and machine learning techniques to assess the feasibility and quality of image recovery. Here is a detailed breakdown of each method and its specific approach:

1. Brute-Force and Combinatorial Analysis

In this initial attempt, we apply brute-force methods and combinatorial analysis to reverse the pixel shuffling on a downscaled 64x64 image. Brute-force reconstruction involves generating all possible permutations of pixel positions in search of the original configuration, which could theoretically reveal the original face. The factorial growth of possible pixel arrangements, even for 64x64 images, limits brute-force feasibility, particularly when high levels of shuffling have been applied. This attempt helps establish the impracticality of brute-force for larger images, setting a baseline for the limitations of purely computational approaches. Let us reduce the scale of the picture one more time to 16x16 for simplicity of calculations.

To brute-force a 16x16 grayscale image, let us break down the calculations involved and determine the resources needed. A 16x16 grayscale image has 256 pixels, each with a grayscale value from 0 to 255, representing 256 possible values for each pixel. Steps to Calculate the Computational Resources:

Total Possible Combinations:

- Each pixel has 256 possible values (from 0 to 255).
- Since there are 256 pixels, the total number of combinations would be:

$$256^{256} = 2^{2048}$$

- This is an astronomical number, making brute-forcing extremely computationally intensive.

Storage Requirements:

- Storing all possible combinations would be infeasible. If each combination (image) were stored as a 256-byte structure (1 byte per pixel), the storage needed would be:

$$256x2^{2048}\ bytes$$

- This is vastly beyond any storage system available today.

Computational Power:

- Brute-forcing even 1% of these combinations would require a processing capability far beyond today's supercomputers.
- A modern GPU can process a few teraflops (trillions of floating-point operations per second), but brute-forcing 16x16 grayscale images would require exaflops or higher to approach reasonable time frames.
  A high-end GPU can process about 10^{12} operations per second.
  If each "operation" is generating one unique 16x16 image and comparing it.
  Estimated time for brute-forcing:

$$\frac{2^{2048}}{10^{12} images\ per\ second}$$

This would still take longer than the age of the universe. Given the extreme complexity, brute-forcing a 16x16 grayscale image by checking all possible combinations is infeasible. Alternative methods, such as heuristic approaches or machine learning techniques, would be more practical for solving specific image-related problems or optimizations.

2. Deep Learning with Convolutional Neural Networks (CNNs)

For our next approach, we employ a deep learning model—specifically a Convolutional Neural Network (CNN)—to learn patterns in pixel arrangements that could facilitate reconstruction. The model is trained on paired images of original and distorted (shuffled) faces to predict the original pixel positions.

The CNN model is provided with a dataset of original and distorted image pairs, including my own face image, to learn a mapping between distorted and original configurations. During training, the CNN attempts to identify and approximate the correct pixel locations to recover the facial structure in a coherent way.

We use metrics such as Mean Squared Error (MSE) and Structural Similarity Index (SSIM) to compare reconstructed images with their original versions. These metrics help evaluate the quality and accuracy of each reconstruction attempt.

Advantages and Limitations: CNNs have shown promise in reconstructing structured distortions, especially where patterns can be learned. However, their effectiveness depends on the training data and the model's ability to generalize across different shuffling patterns. For training purposes LFW database was used.

The visualization of the training and validation loss (figure 2) suggests potential issues in the model training process. From the outset, we observe that the training loss

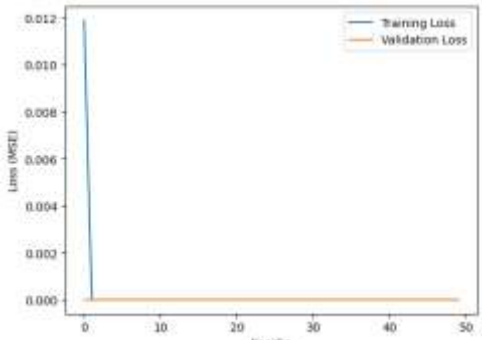

Figure 2 Visualization of the training

drops sharply to near zero, reaching a stable point almost immediately, while the validation loss remains similarly flat and close to zero throughout the entire training period.

Despite our attempts to train the model, it became evident that it was impossible to reconstruct the original images from the shuffled versions with this

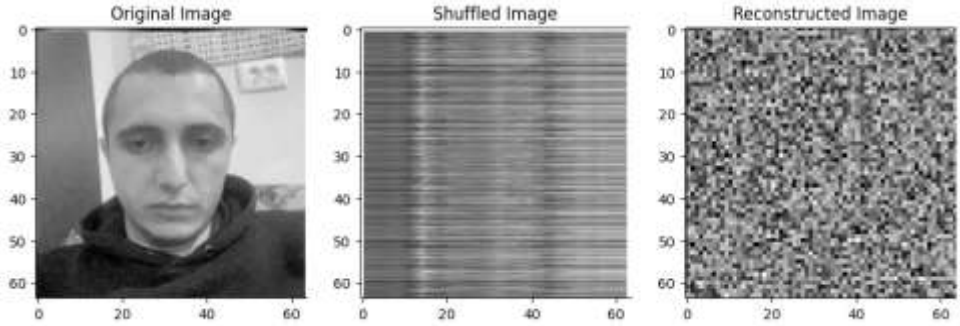

Figure 3 Original, shuffled and reconstructed images using CNN

approach(figure 3). The inability to achieve reconstruction likely stems from the high level of randomness introduced by pixel shuffling, which disrupts the spatial coherence necessary for CNNs to learn meaningful relationships between pixels. Without this spatial information, the model lacks the structure needed to reverse the shuffling process, rendering it ineffective at generating a coherent image reconstruction. This outcome suggests that highly randomized distortions, such as full pixel shuffling, present a formidable challenge for deep learning models and highlight

the limitations of current CNN architectures for tasks that require precise spatial reordering.

3. Autoencoders for Reconstruction

Autoencoders, particularly denoising autoencoders, are another deep learning-based approach that can be useful for reconstruction. We explore the potential of an autoencoder to "denoise" the shuffled images by training it to recognize the original pixel patterns. We train the autoencoder with pairs of original and shuffled images, using the original image as the "clean" target. The autoencoder, which compresses the image and then reconstructs it, learns to approximate the mapping back to the original structure. The network attempts to decode or "denoise" the shuffled image back to a recognizable face. We measure the quality of the output based on visual accuracy and MSE to determine if the autoencoder effectively restores facial features. Autoencoders excel at learning efficient representations, especially for compressed data, but may struggle with high levels of random distortion. This experiment highlights their strengths in controlled, structured distortions but may expose limitations in generalized recovery. For training purposes LFW database was used.

Autoencoders training loss plot shows similar issues as the previous one. Both the training and validation losses drop to near-zero right after the first epoch and remain constant throughout the rest of the training. This pattern indicates that the autoencoder model is likely not learning to perform meaningful reconstruction. Here are possible explanations:

- The training loss dropping immediately to near zero suggests that the model is finding a trivial solution, such as simply outputting a constant value or memorizing an unintended pattern. This can happen if the model is not receiving the correct input-target pairs or if there is a fundamental issue with the data preparation.
- If the shuffling applied is highly randomized, the autoencoder might lack the necessary complexity to reconstruct spatial information from completely unstructured data. The model could converge on a near-zero error without actually learning to reconstruct the original images properly, especially if the shuffling is beyond its reconstructive capability.

Another option is to adjust the learning rate or model architecture (e.g., adding more layers or increasing the depth of the encoder-decoder) to see if this changes the training dynamics which we have tried but there weren't any obvious differences. The current setup might be too challenging for the autoencoder due to complete randomness in shuffling, making it impossible for the model to reconstruct the original images. If the shuffling is highly randomized, even more complex

architectures may struggle, as random shuffling disrupts the spatial information that autoencoders rely on.

These 3 approaches - brute-force, CNNs, autoencoders allow us to evaluate the effectiveness and challenges of each method for reversing image distortion. By combining these methods, we gain a comprehensive view of both the feasibility and limitations of reconstruction attempts. Each experiment contributes uniquely to understanding the robustness of image anonymization in e-voting systems, particularly regarding the reversibility of applied distortions.

Conclusion

In this study, we evaluated the feasibility of reversing pixel-shuffled voter images in e-voting systems through a series of computational and deep learning approaches, including brute-force methods, Convolutional Neural Networks (CNNs), and autoencoders. Our experiments demonstrated that brute-force reconstruction is impractical due to the exponential complexity of pixel arrangement possibilities, making this approach computationally infeasible. CNNs and autoencoders, while promising for structured tasks, struggled with the high level of randomness introduced by pixel shuffling. Both models showed rapid convergence with negligible loss improvement, indicating that they failed to learn meaningful representations for image reconstruction. The inability of CNNs and autoencoders to restore the spatial coherence disrupted by pixel shuffling suggests that such distortions effectively prevent re-identification attempts through these methods.

Overall, the findings affirm the robustness of pixel shuffling as a method for anonymizing images in e-voting systems, offering substantial resistance to reconstruction attacks. This study contributes valuable insights into the limits of current machine learning techniques in handling fully randomized distortions, reinforcing the importance of carefully designed privacy safeguards in digital voting environments. Future research could explore alternative anonymization methods or hybrid approaches that combine shuffling with additional distortions to further enhance voter privacy. Additionally, future experiments could investigate the use of more complex or deeper neural networks, as well as various machine learning models and architectural configurations, to assess if further advancements in model depth or architecture could improve the ability to reconstruct voter images.

References

1. Haroutunian M. E., Margaryan A. S., Mastoyan K. A., New Approach for Online Voting Ensuring Privacy and Verifiability, ISSN 0361-7688, Programming and Computer Software, 2024, Vol. 50, Suppl. 1, pp. S60–S68
2. Reneta P. Barneva, Valentin E. Brimkov, Jake K. Aggarwal, **Combinatorial Image Analysis**, 2012/
3. Simonyan, K., Zisserman, A., *Very Deep Convolutional Networks for Large-Scale Image Recognition,* 2014.
4. Zhao H., Gallo Frosio I., Kautz, J., *Loss Functions for Image Restoration with Neural Networks,* 2017, IEEE Transactions on Computational Imaging.
5. Wang T. C., Liu M. Y., Zhu J. Y., Tao A., Kautz J., Catanzaro, B., *High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs,* 2018, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8798–8807.
6. Huang G. B., Ramesh M., Berg T., Learned-Miller E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report, 2007, 07-49.
7. Liu Z., Luo P., Wang X., Tang X., *Deep Learning Face Attributes in the Wild.* Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, 3730-3738.
8. Phillips P. J., Moon H., Rizvi S. A., Rauss P. J., *The FERET Evaluation Methodology for Face-Recognition Algorithms,* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000 22(10), 1090-1104.

ԳԱՂՏՆԻՈՒԹՅԱՆ ԲԱՑԱՀԱՅՏՈՒՄ: ԷԼԵԿՏՐՈՆԱՅԻՆ ՔՎԵԱՐԿՈՒԹՅԱՆ ՀԱՄԱԿԱՐԳԵՐՈՒՄ ԸՆՏՐՈՂԻ ՊԱՏԿԵՐԻ ԱՆԱՆՈՒՆԱՑՄԱՆ ՀԵՏԱԴԱՐՁԵԼԻՈՒԹՅԱՆ ԳՆԱՀԱՏՈՒՄԸ

ՄԱՍՏՈՅԱՆ ԿԱՐԵՆ
ՀՀ ԳԱԱ ԻԱՊԻ ասպիրանտ,
ԳՊՀ դասախոս
Էլփոստ՝ kmastoyan@gmail.com

Այս հոդվածում քննվել է էլեկտրոնային քվեարկության համակարգերում պիքսելներով խառնված ընտրողների պատկերների հետադարձման հնարավորությունը՝ օգտագործելով հաշվողական և խորը ուսուցման մի շարք մոտեցումներ, ներառյալ broot force, կոնվոլյուցիոն (փաթույթային) նեյրոնային ցանցերը (CNN) և ավտոկոդավորիչները: Մեր փորձերը ցույց են տալիս, որ broot force վերակառուցումն անիրագործելի է պիքսելների տեղադրման հնարավորությունների էքսպոնենցիալ բարդության պատճառով՝ այս մոտեցումը հաշվողականորեն անիրագործելի դարձնելով: CNN-ները և autoencoder-ները, չնայած խոստումնալից են կառուցվածքային առաջադրանքների համար, սակայն նույնպես թույլ են պատահականության բարձր մակարդակի դեմ, որը ներկայացվել է պիքսելների խառնման արդյունքում: Երկու մոդելներն էլ ցույց տվեցին արագ կոնվերգենցիա՝ կորստի փոքր բարելավմամբ, որը ցույց է տալիս, որ նրանք չեն կարողացել սովորել պատկերի վերակառուցման իմաստալից ներկայացումներ: CNN-ների և ավտոկոդավորիչների անկարողությունը՝ վերականգնելու տարածական համահունչությունը, որը խաթարվել է պիքսելների խառնման հետևանքով, վկայում է այն մասին, որ նման աղավաղումները արդյունավետորեն կանխում են պատկերի վերականգման փորձերը՝ օգտագործելով այս մեթոդները: Ընդհանուր առմամբ արդյունքները հաստատում են պիքսելների խառնման կայունությունը՝ որպես պատկերի անանունացման մեթոդ էլեկտրոնային քվեարկության համակարգերում՝ զգալի կայունություն առաջարկելով վերականգման հարձակումներին: Այս ուսումնասիրությունը արժեքավոր պատկերացումներ է տալիս մեքենայական ուսուցման ընթացիկ մեթոդների սահմանափակումների վերաբերյալ՝ լիովին պատահականացված դեպքերը կարգավորելու համար՝ ամրապնդելով թվային քվեարկության միջավայրում մշակված գաղտնիության պաշտպանության կարևորությունը: Հետագա հետազոտություններում կարող են ուսումնասիրվել անանունացման այլընտրանքային մեթոդները կամ հիբրիդային մոտեցումները, որոնք համատեղում են խառնաշփոթը լրացուցիչ աղավաղումների հետ՝ ընտրողների գաղտնիության հետագա բարելավման համար: Բացի դրանից՝

կարող են ուսումնասիրել ավելի բարդ կամ խորը նեյրոնային ցանցերի, ինչպես նաև մեքենայական ուսուցման տարբեր մոդելների և ճարտարապետական կոնֆիգուրացիաների օգտագործման հնարավորությունները՝ գնահատելու, թե արդյոք մոդելի խորության կամ ճարտարապետության հետագա առաջընթացը կարող է բարելավել ընտրողների պատկերները վերակառուցելու ունակությունը:

Բանալի բառեր՝ *էլեկտրոնային քվեարկության համակարգեր, ընտրողների գաղտնիություն, պատկերների անանունացում, դեմքի ճանաչում, տվյալների անվտանգություն, պիքսելների աղավաղում, գաղտնիության պաշտպանություն, ինտերնետ քվեարկություն:*

РАЗОБЛАЧЕНИЕ КОНФИДЕНЦИАЛЬНОСТИ: ОЦЕНКА ОБРАТИМОСТИ АНОНИМИЗАЦИИ ИЗОБРАЖЕНИЙ ИЗБИРАТЕЛЕЙ В СИСТЕМАХ ЭЛЕКТРОННОГО ГОЛОСОВАНИЯ

МАСТОЯН КАРЕН
Аспирант ИПИА, НАН РА
Преподаватель ГГУ
электронная почта: kmastoyan@gmail.com

Системы электронного голосования (e-voting) стали многообещающим решением для модернизации выборов, обеспечивая удобство, доступность и потенциальную экономию средств. Однако такие системы сталкиваются со значительными проблемами, особенно в отношении конфиденциальности, безопасности и проверяемости. Среди критических проблемных областей — защита личности избирателей, особенно в системах интернет-голосования (i-voting), где конфиденциальная информация может быть уязвима для атак повторной идентификации. В предыдущих исследованиях была предложена модель анонимизации с использованием перетасовки пикселей для искажения изображений избирателей и защиты конфиденциальности. В данной статье рассматривается возможность восстановления таких искаженных изображений с использованием методов, таких как вычисления методом перебора, сверточные нейронные сети (CNN) и автокодировщики, для оценки надежности анонимизации. Экспериментальные результаты показывают, что рандомизация, введенная перетасовкой пикселей, эффективно предотвращает реконструкцию методом перебора из-за ее астрономической сложности. Попытки восстановления исходных изображений с помощью CNN и автокодировщиков выявили

ограничения моделей глубокого обучения при работе с сильно рандомизированными изображениями, поскольку пространственная когерентность, необходимая для точной реконструкции, была утрачена. Эти результаты подчеркивают эффективность перемешивания пикселей как метода сохранения конфиденциальности.

Ключевые слова: *системы электронного голосования, конфиденциальность избирателей, анонимизация изображений, распознавание лиц, безопасность данных, искажение пикселей, защита конфиденциальности, интернет-голосование.*

Հոդվածը ներկայացվել է խմբագրական խորհուրդ 10.01.2025թ.:
Հոդվածը գրախոսվել է 25.01.2025թ.:
Ընդունվել է տպագրության 25.04.2025թ.: