[[1]](#footnote-1)

BAMMFR: A Brain-Inspired Approach for Privacy-Preserving Face Recognition

***Abstract***

**Purpose—**This paper aims to tackle concerns regarding personal facial privacy and the management of massive video data in massive robotic vision application scenarios. The study aims to mimic the cognitive process of the brain, enabling robots to forgo the need for extensive video storage while leveraging minimal memory (abstract semantic features) to restore face recognition. Additionally, it aids in safeguarding facial privacy amidst the proliferation of camera applications by addressing the risk of raw image exposure.

**Design/methodology/approach—**The paper introduces BAMMFR (Brain-like Associative Memory Mechanism Face Recognition), an algorithm that leverages a CNN-based face recognition approach to accurately detect human faces and employs mosaic encryption to secure the original video. Subsequently, a spatial-temporal index encoding method based on GAN feature extraction and generation is implemented to create a decryption key function tailored for individuals with identical identities, mirroring the perceptual and associative memory mechanisms of the human brain.

**Findings—**Our research demonstrates that BAMMFR exhibits significant potential in person re-identification and personal privacy protection. Experimental validation on a self-built dataset confirms the feasibility of the algorithm for face encryption and decryption. Comparative experiments and analysis validate the importance of high-dimensional abstract semantic features in face association re-recognition.

**Originality/value—**Overall, the paper offers valuable insights and solutions to the progression of machine vision application and the assurance of privacy in the explosion era of robotic intelligence.

***Index Terms*—**Artificial Intelligence, Face recognition, Privacy protection, Brain-like associative memory mechanism**.**

# I. INTRODUCTION

The rapid proliferation of robots is projected to reach millions of applications worldwide in the foreseeable future. However, this surge in robotic deployment brings forth significant challenges, particularly concerning the management of vast amounts of video data generated by their integrated cameras[1]. With a frame rate of 30 frames per second and an average image size of 5MB, a single camera can quickly accumulate a staggering 12,656.25Gb of data per day, overwhelming traditional data centers and necessitating regular data overwriting [2]. Moreover, the redundancy inherent in massive camera video data further compounds the issue, often leading to the loss of crucial information and hindering effective video-based information retrieval [3]. Additionally, the transmission of such massive video data consumes substantial communication bandwidth, resulting in elevated communication costs and impeding the widespread adoption of cameras for collaborative purposes [4]. Concurrently, the use of robotics' cameras raises pertinent concerns regarding the potential leakage of residents' biometric privacy, necessitating ethical and regulatory scrutiny.

In response to these challenges, scholars across various disciplines are focusing their research efforts on developing strategies to safeguard camera functionality while addressing the aforementioned issues. In this paper, we adopt a humanoid cognitive perspective to explore new models for large-scale camera urban applications. Drawing inspiration from human perception and memory mechanisms, we aim to develop innovative approaches to handle massive amounts of video data effectively. As humans, we are adept at perceiving and processing visual information throughout our lives, relying on both our eyes to form lasting memories of people and objects encountered. However, our ability to recall precise image details is limited, often relying on high-dimensional semantic abstraction to facilitate coarse-grained picture recall. Similarly, our recognition of familiar faces is not solely based on specific facial features but rather on general impressions derived from higher-dimensional semantic information. Furthermore, the abstract semantics stored in our brain play a crucial role in blurring human recognition, allowing us to recognize acquaintances even from partially obscured facial images. While the theoretical explanation of the humanoid perceptual memory mechanism remains challenging, the association between low-dimensional fine-grained information and higher-dimensional coarse-grained information holds both theoretical significance and practical value, particularly in the compression and decryption of massive video data.

In this paper, we propose an autonomous face intelligent encryption and decryption algorithm inspired by the humanoid association memory mechanism. Our contributions are outlined as follows:

1. Introduction of an autonomous face intelligent encryption/decryption algorithm based on abstract analysis and modeling of the facial future perception-memory-association mechanism for privacy-preserving face re-recognition.
2. Development of an AI- driven face recognition system featuring a sophisticated encryption and decryption algorithm inspired by human associative memory. Through rigorous testing on a dataset, we demonstrate the algorithm's high effectiveness in selectively protecting facial identity information in videos.
3. Analysis of matching methods and loss function selection, affirming the efficacy of utilizing cosine similarity as the matching criterion and high-dimensional features as the matching basis, thus underscoring the importance of high-dimensional abstract semantic features in the process of face association re-recognition.

The structure of the subsequent sections is as follows: Section II presents a review of related work, including face recognition in video data, face encryption and decryption algorithms, and humanoid memory cognition. Section III models and abstracts the process of human face recognition, memory, and associative re-recognition, proposing an algorithmic framework for face encryption and decryption using human-like cognitive mechanisms. Section IV provides a detailed account of the initial validation of the proposed encryption and decryption algorithm on the dataset. Finally, Section V concludes the paper and outlines future directions for research.

# II. RELATED WORKS

## A. Face recognition of computer vision

In the perspective of recent advances in the field of AI-driven face recognition for video data, the tracking of human faces in video data has become widely adopted in urban security and community management. A lot of scholars are committed to the research of computer vision techniques with promising accuracies and efficiencies for face recognition and object detection [5-7]. The face recognition methods mainly include 1) traditional methods, which rely on hand-crafted feature extraction techniques and a pre-trained classifier along with fusion, and 2) deep learning methods, which automatically learn features and classifiers together by utilizing enormous quantities of data [8-10]. With the development of deep learning technology, the application boundaries of face recognition will gradually expand. The majority of face recognition in video data today is "closed-set," which only recognizes the identities of previously registered individuals. However, the term "open-set" has gained popularity due to the disparities between the source and target domains. These disparities make it less effective to transfer face recognition systems from controlled environments to uncontrolled scenes. To handle this, Suandi *et al*. [11,12], proposed fuzzy ARTMAP neural networks to solve the open-set single-sample face recognition problem. They also developed an automatic pose normalization technique that does not require model fitting or human intervention. These advancements greatly improve the performance of open-set single-sample face recognition methods in surveillance environments. The "open-set" face recognition is prone to increasing the degree of human privacy exposure in the ubiquitous city surveillance network.

The low resolution of urban monitoring pictures and the difficulty of feature extraction from small faces are being addressed. Even though cameras are typically positioned at a distance from the objects they capture, resulting in low-resolution face images, extensive research has been carried out to identify reliable recognition features in low-quality video frames. Zhao *et al*. [13] took an end-to-end approach to match high-resolution (HR) images with low-resolution (LR) images in the videos. Singh *et al*. [14] improved the number of descriptors in the image and mitigated the effects of noise based on super-resolution faces. Dharrao *et al*. [15] used the Viola-Jones algorithm to detect the facial region in the sequential frames of the video. They enhanced the quality of the facial region by applying a super-resolution scheme based on bicubic interpolation. In addition, multi-resolution convolutional neural networks (MRCNN) and anti-aliasing techniques were adopted to solve the low-resolution problems [16].

The development trend of face recognition technologies shows that the challenge of protecting citizens' facial privacy under ubiquitous cameras is becoming increasingly serious. Exploring a new paradigm for large-scale camera urban applications from the perspective of humanoid cognition is meaningful. This can be achieved by performing face-reduction encryption on the recognized video images.

## B. Face encryption and decryption algorithm

The problem of privacy leakage has aroused widespread concern. Face recognition in video data has become ubiquitous in daily life, but it is difficult to strike a balance between intelligent vision applications and personal privacy protection. In addition to improving relevant laws and regulations to govern the acquisition, storage, and use of videos, corresponding technical measures are needed to safeguard personal privacy. The cryptography-based scheme for face privacy protection selectively encrypts the face region in the video that shows the identity. This encrypted data can be decrypted to recover the original video in case of future legitimate demand. How to integrate the autonomous face intelligent encryption and decryption algorithm of humanoid association memory mechanism into the AI face recognition algorithm is an urgent breakthrough direction.

Most of the existing face encryption schemes are homomorphic-based [17-21]. There are three different types of homomorphic encryption schemes: (1) partially homomorphic encryption, (2) somewhat homomorphic encryption, and (3) fully homomorphic encryption (FHE). Tamiya *et al*.[17] proposed a successful homomorphic encryption-based face template protection scheme by computing the squared Euclidean distance between facial features with a single homomorphic multiplication method. Román *et al*.[18] suggested using the Kyber and Saber public key encryption (PKE) algorithms, along with homomorphic encryption (HE) in facial recognition systems. This approach aims to achieve smaller protected templates and key sizes, as well as faster execution times, compared to other HE schemes that utilize lattices. The use of fully homomorphic encryption algorithms provides a higher level of privacy authentication for the queried face. Huang *et al*. [19] offered a successful, privacy-preserving face verification method based on a corrupted circuit and fully homomorphic encryption. Some researchers used CKKS fully homomorphic encryption to encrypt the normalized facial feature vector [20,21].

Due to the low computational efficiency of using homomorphic encryption, other researchers have attempted to identify lightweight algorithms for encrypting facial data. Tan *et al*. [22] proposed a novel approach to implementing video-based ring-learning (ring-LWE) cryptography for face encryption and decryption on a graphics processing unit (GPU). Duong-Ngoc *et al*. [23] proposed a novel method for comprehensively protecting facial images extracted from videos based on NewHope cryptography for post-quantum cryptosystems. This method significantly reduces the time required for encryption and decryption. Zhao *et al*. [24] proposed and implemented a simple and efficient speckle-based optical cryptosystem to encrypt face images using seemingly random optical speckles at the speed of light. They achieved this by training a cryptographic neural network to decrypt face images from random speckles. A fast block scrambling method was used to scramble the detected faces [25,26]. In addition, an encryption technique using face biometrics to generate random phase masks has been proposed [27]. A THM (Tent-Henon Map) chaotic encryption of faces was proposed, combining with the properties of tent chaos and Henon chaos [28]. Liu *et al*. [29] proposed an RGB image encryption algorithm based on DNA encoding and a chaos map. Wu *et al*. [30] proposed a GAN-based method to encrypt facial features using Wasserstein Generative Adversarial Network Encryption (WGAN-E). Ashiba *et al*. [31] used a graph theory-based graph first decomposition mask (GFH) coding algorithm. There is still room for improvement in terms of computational communication efficiency and privacy-preserving effects. Active perception of key privacy features for target encryption, based on humanoid cognitive mechanism, provides a preliminary exploration in this direction.

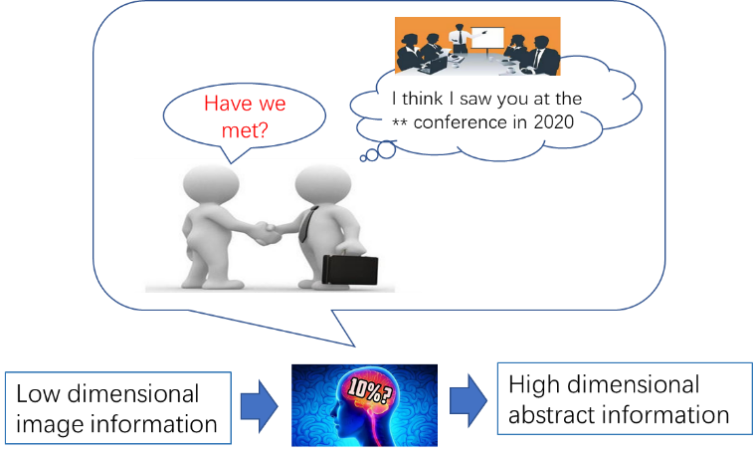
## C. Humanoid associative memory cognition

The human brain is a typical encryption and decryption processing device with low energy consumption and high efficiency. The brain can store learned concepts in memory and recall them when it sees partial or broken patterns. Franklin *et al*. [32] proposed a structured event memory model (SEM) of event cognition, illustrating human abilities in event segmentation, memory, and generalization. SEM can be extended to a high-dimensional input space to produce humanoid event segmentation for natural video data. It also demonstrates various memory phenomena. Sun *et al*. [33] proposed a new model humanoid visual cognitive and language-memory network for visual dialog (HVLM) to simulate global and local dual-view cognition in the human visual system to comprehensively understand images. Inspired by humanoid perception and memory, we explored a new model for face privacy protection in urban large-scale camera monitoring. The research of this algorithm is of great significance to the research of video information compression and storage, character recognition, and personal privacy protection.

# III. PROPOSED APPROACH

## A. Problem description.

The process of human face perception and identity recognition based on fuzzy impression memory association is highly complex. Each of us sees many faces in daily life scenarios, however, not all the information about faces is remembered. As shown in Fig. 1, when people meet with each other unintentionally, their minds will unconsciously recall that they have seen a similar face at a specific time, place, or event. Moreover, they can recall memories of more detailed scenes and clearer features. The process can be simplified, as the human brain perceives the concrete face image information seen by the eyes to extract high-dimensional abstract semantic features. The high-dimensional abstract semantic features are retrieved and matched with the high-dimensional semantic information indexed in memory combining person, event, time, and place. This allows for the clear reproduction of the past feature-blurred memory scene in combination with the current perceived face image.

****

**Fig. 1. Case: Humanoid abstract associations triggered by perceptual features.**

The general expression of the Humanoid Association is as follows.

where *I is the image, Ai* is the set of low-dimensional full-dimensional information about the *ith* person’s face perceived by the brain in the first stage, is the set of high-dimensional abstract semantic features of *ith* person’s face formed by the brain in the mind based on, is the encrypted dataset of , is the decryption set partially from and , is the set of low-dimensional full-dimensional information about the face perceived by the brain in the second stage, and is the set of high-dimensional abstract semantic features formed by the brain in the mind based on . The algorithm for solving the above expression is shown in Table 1.

**TABLE 1**

**ALGORITHM: ENCRYPTION AND DECRYPTION**

|  |
| --- |
| Input: *, Bi, A'i,*  Output: *A'ip*  For *Ai* in Brain Do encryption Key matching  Using cerebral neural network for high dimensional semantic abstraction  *f1(Ai) →AiP,* where*AiP* ⊊*Ai, i*=*1, 2,* ……  For *Bi* in Brain Do decryption  *f1(A'i) →B'i*  Find matching key *B'i* to *Bi*， where is the maximum face similarity  *P*(*A'ip | Aip*)= *f2*(*B'i* ∩ *Bi*)→1  For *AiP* in Brain Do encryption  *f3*(*AiP, Bi*) →*A'ip*  return *A'ip* |

This paper combines the above humanoid perceptual associative memory algorithm with face encryption and decryption requirements to solve the following problems.

1) To model the humanoid cognitive mechanism, the high-dimensional abstract memory and compressed perception process *f1* function need to be solved. We propose an artificial intelligence algorithm for solving and to identify and locate faces in videos, extract high-dimensional semantic features while encrypting video faces with reduced resolution.

2) The algorithm models the storage and matching of high-dimensional semantic features using human-like associative memory. Its associative matching function *f2* is based on and through the recall-triggered indexing mechanism.

3) Drawing on the humanoid perception-triggered recall mechanism, the associative recall of high-dimensional semantic features and low-resolution video is modeled to solve *f3* for indexing location as well as high-resolution decryption.

## B. AI Methodologies

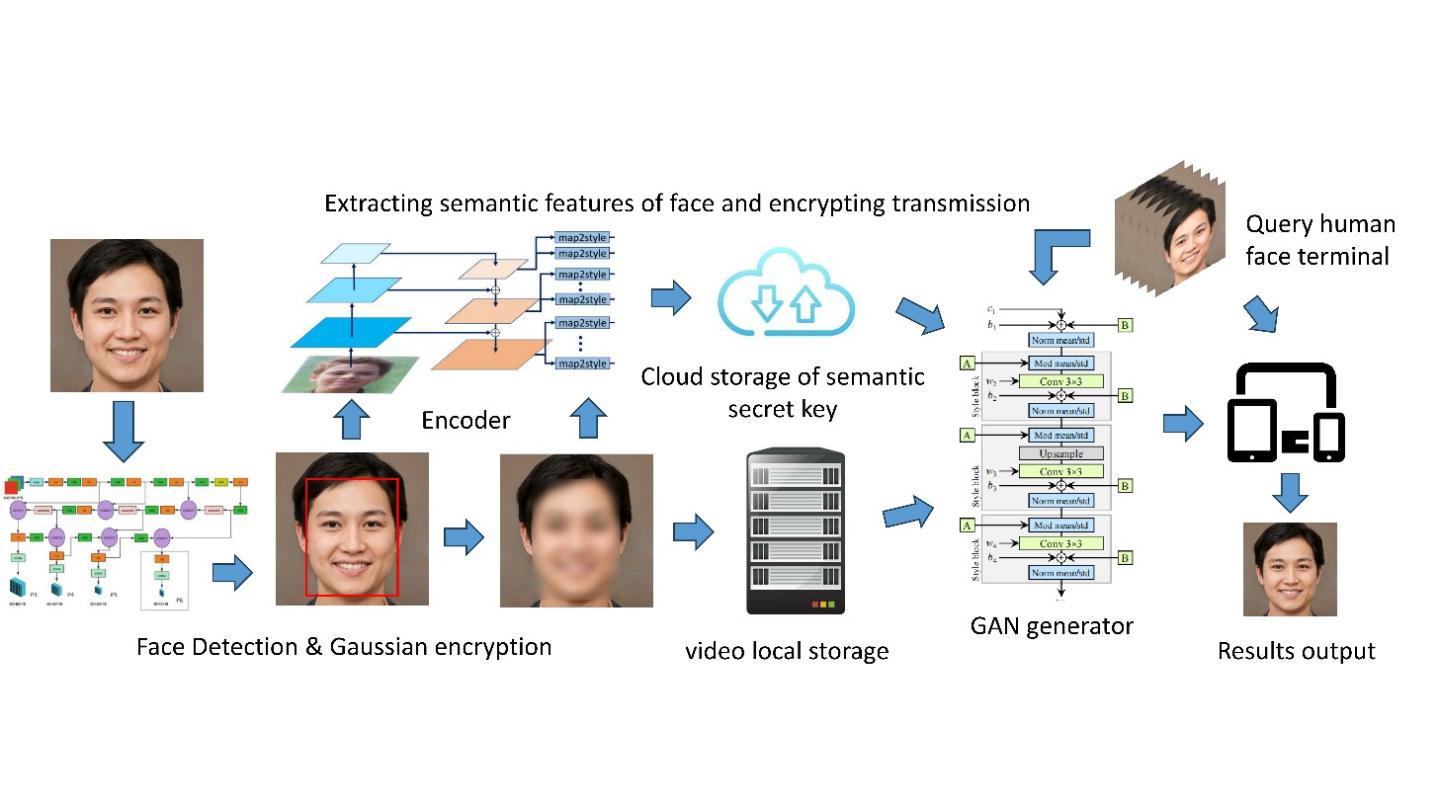
Inspired by humanoid perception, compressed memory, and associative recall, we propose an algorithmic framework that can be used to encrypt/decrypt faces as shown in Fig. 2. In the initial stage of video data production, a convolutional neural network combined with mosaic image processing technology for face recognition and encryption algorithm is employed to encrypt the pictures in the original video stream. Concurrently, a GAN Generator is utilized to extract both high and low-dimensional semantic information from the facial features. This extracted information forms a feature key, establishing an index association with each face in the video. These associations are stored in the corresponding key pool of the video dataset. During the decryption process, the features extracted from newly specified faces are used to search the key pool. Upon finding matching features, the corresponding encrypted faces are decrypted using the GAN generator, resulting in clear faces.

**1) Encryption method**

For the video frame input *V*, the YOLO5-face deep learning model *ф* is used to achieve the recognition and localization of faces by the cameras at the edge end, to obtain *A1* = ф(*V*), YOLO5-face is chosen because the model targets the face recognition segmentation needs, adds landmark branches in YOLOV5, and improves the accuracy of face detection and localization by regressing the wing loss function through five facial key points. The wing-loss function and the overall loss function are as follows.

After completing face target detection, the face in the recognition frame is subjected to Gaussian blurring, i.e. a Gaussian convolution budget is applied to the face image with the probability density distribution function shown below.

After the above face localization + Gaussian blurring, a video will be obtained in which part of the face is displayed by encrypted blurring and the rest is displayed normally, this video corresponds to in Eq. above and is stored locally. The solving process of *f1(Ai)* is thus completed.



**Fig. 2. BAMMFR** **Pipeline based on CNN and GAN-Based Encryption and Decryption Algorithms.**

**2) Key storage and matching**

This method no longer stores the original video. Instead, it chooses to terminal store the encrypted video while uploading the high-dimensional abstract semantics to the cloud for subsequent processing and analysis. For example, in the case of face retrieval service, the terminal storage of encrypted faces results in a significant loss of face feature information. As a result, the video cannot be directly retrieved for review. Instead, it needs to be indexed for the high-dimensional semantic B for query service, which is similar to the human perceptual memory.

Inspired by the pixel2style2pixel (pSp) framework [34] that solves a variety of image-to-image conversion tasks, we use the pSp encoder, which directly encodes a given image into the desired latent code. This allows us to extract high-dimensional abstract semantic features . The model expression is as follows:

Taking the previous original video as input, (\*) denotes the latent code of obtaining to get the abstract semantic feature . denotes the pSp encoder. The potential vector obtained from is summed with the average potential vector in the network model to obtain the final potential vector. This step usually helps to balance the quality and diversity of the generated images. In practical applications, customized mathematical conversion operations can be introduced after the above formula to generate customized conversion keys, to ensure the personalization and security of the keys of each monitoring cloud platform.

After extracting the high-dimensional abstract semantic feature from the original video using the above encoder, the temporal sequence of the video frames, along with the latitude, longitude, and pixel coordinates of the edge camera itself, are fused as the symbolic bit encoding of the high-dimensional abstract semantic feature This symbolic bit encoding is then associated with the mapping relationship between the high-dimensional abstract semantic feature and the encrypted face image frames, making it easier to decrypt the above video in subsequent steps. These together serve as the key for decrypting the set and are stored in the cloud key pool.

The high-dimensional abstract semantic features learned by the above model are often not interpretable, so it is necessary to conduct an in-depth study of the abstract semantics and to investigate how to match face retrieval by both. At present, in the field of face recognition, the technology of matching face features by deep neural networks to determine the identity of faces is relatively mature. However, using abstract semantics and as the identity key to determine the identity of faces poses a greater challenge.

The identity recognition of an input face is the process of searching for the most similar high-dimensional abstract feature, which is paired with one’s identity information, in the key pool using the feature extracted from the input face. When a face needs to be decrypted, the same model is first used to extract features from the face, and then certain dimensions are taken to match with the features in the human cloud keystore, and the accuracy of the matching is calculated. Thus, the recall-triggered call function *f2* (*B'i*∩*Bi*) is implemented.

**3) Decryption process**

The previous section implements the original video encryption and describes the process of key storage and query matching. This section discusses how to output the decrypted video by the encrypted video , high-dimensional abstract semantics , and the facein the second video that matches the identity of a person in the first video. This process is equivalent to associative memory, where we can associate images from the past with the current image, and the blurred features of a face can often be made clear again.

To this end, we build an open-set face re-recognition and decryption model based on the styleGAN generator [35]. Firstly, through the method described in the previous section, the high-dimensional abstract semantics is extracted from . Then, the similarity is calculated between and all the high-dimensional abstract semanticsin the key pool corresponding to the encrypted video , and the with the highest similarity is taken. If the similarity is lower than a certain threshold, the face is judged to be strange and further decryption is rejected; if the similarity exceeds a certain threshold, the high-dimensional abstract semantics is added to the generator with as input as a constraint, and the decrypted video is output. Then the solution of the decryption *f3*(*AiP, Bi*) in the human-like memory mechanism is obtained.

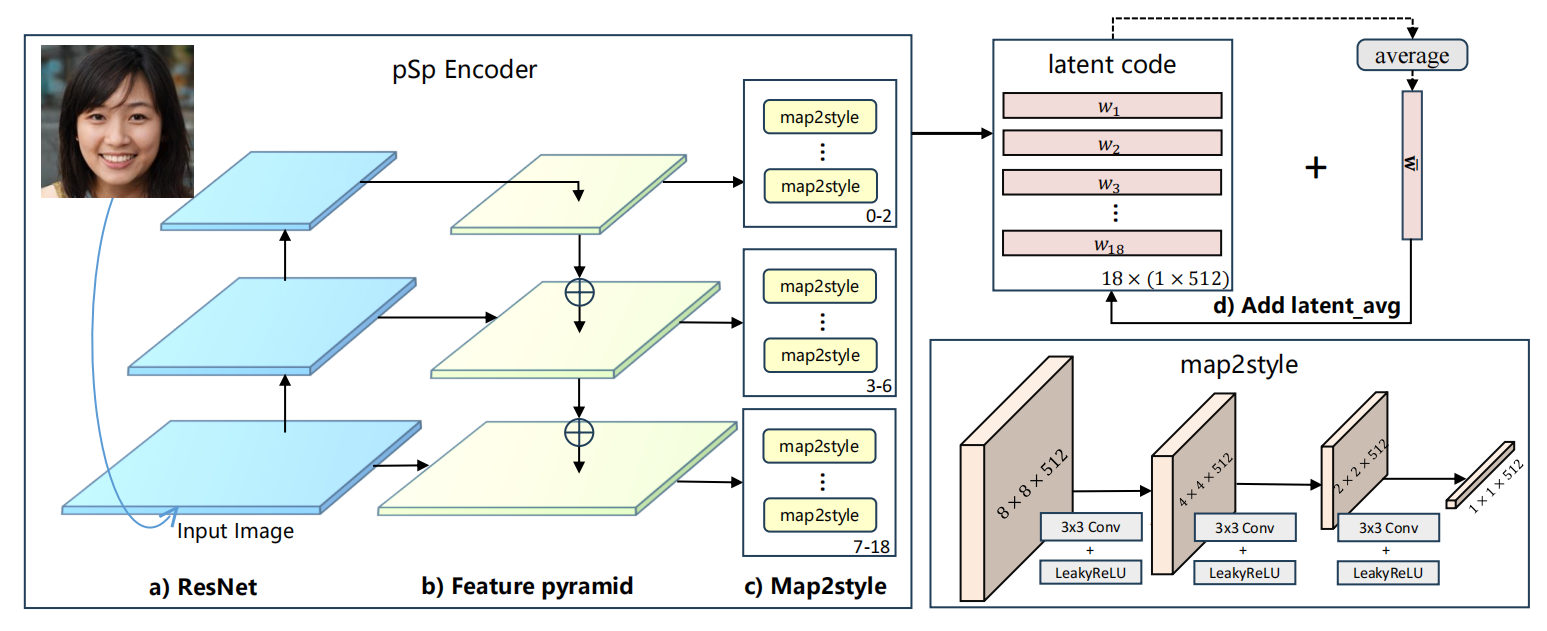
# IV. EXPERIMENTS AND RESULTS

## A. Dataset

The dataset used in the experiments is generated\_yellow-stylegan2[[2]](#footnote-2), which is a purely random (no filtering) dataset of yellow faces generated by a face generator made based on StyleGAN2, and contains 10,000 images of yellow faces with different genders, ages, and head pose variations. During the training and encryption of faces, we take this dataset as input and use YOLO5-face-based Gaussian encryption to obtain a dataset that consists of Gaussian encrypted images paired with the original ones. When decryption is needed, we use the Age Filter[[3]](#footnote-3) in AIlab, an artificial intelligence cloud platform created by Wondershare, to perform aging processing on 47 images from the generated\_yellow-stylegan2 dataset. The majority of the images are set to 50 years old using the default setting in the platform. This setting yields improved aging processing results, ensuring that the facial features remain relatively recognizable while producing more natural and plausible outcomes. The aging of the image is not so extreme and unrealistic that it loses its recognizability, which ensures the realism and credibility of the generated results. The aged image is used as the new input image, and the aging process is employed to simulate the natural physiological changes of the original face after a certain period in a real decryption restoration scenario. This allows for a more realistic simulation of the entire process of face encryption and decryption in the video.

## B. Training and Encryption Process

We used the pSp encoder to accomplish the training and extraction of potential codes for the above dataset. The framework depicted in Fig. 3 utilizes a standard feature pyramid architecture based on ResNet and mapping networks for feature extraction. ResNet, short for Residual Network, is a deep learning architecture that introduces residual connections to alleviate the vanishing gradient problem, enabling the training of very deep neural networks. The mapping networks named map2style are consists of several fully connected layers. This network transforms the latent code into a style vector which contains information about various aspects of the image's appearance, such as facial features, colors, and textures. Each element of the style vector corresponds to a different attribute of the image. We employ different mapping networks of varying sizes to extract learned styles from the corresponding feature mappings for each of the 18 target styles. Notably, small feature mappings generate features for 0-2 styles, while medium feature mappings handle 3-6 styles, and large feature mappings are responsible for 7-18 styles. This approach facilitates the realization of high-dimensional abstract memory and compressed perception processes akin to human cognitive mechanisms. By generating high-dimensional feature information through smaller feature mappings, we can effectively acquire and store feature information at different levels.

**Fig. 3. The process of extracting latent code**. **The psp Encoder uses a ResNet-based feature pyramid and three mapping networks to extract 18 target styles. These styles are combined with the average potential vector of the network model to produce the final potential vector.**

The training process is as follows: We extract the original face from the training set and train it using the algorithmic process shown in Fig 3. In this process, we gradually extract multi-dimensional face features from low to high dimensions. These extracted features are integrated and encrypted to form a key paired with the identity ID of the processed face. These encrypted keys are then added to the face key pool and stored in a file named existing\_faces.pkl.

## C. Decryption Process

In the decryption phase, we use the same coding model to encode the newly captured face to extract its unique features. These extracted features will be matched with the keys stored in the face feature key pool according to certain dimensions or matching criteria. By measuring the similarity of the face features, we can determine whether a face has appeared in the cloud face key pool. In our experiments, we can train multiple times and define an acceptable minimum similarity threshold to determine if the face is recognized as having appeared.

If the face features are brand new, i.e., no matching key can be found in the key pool, the system will refuse to decrypt them. This security measure ensures that we only decrypt faces that have already been registered, thereby protecting data privacy.

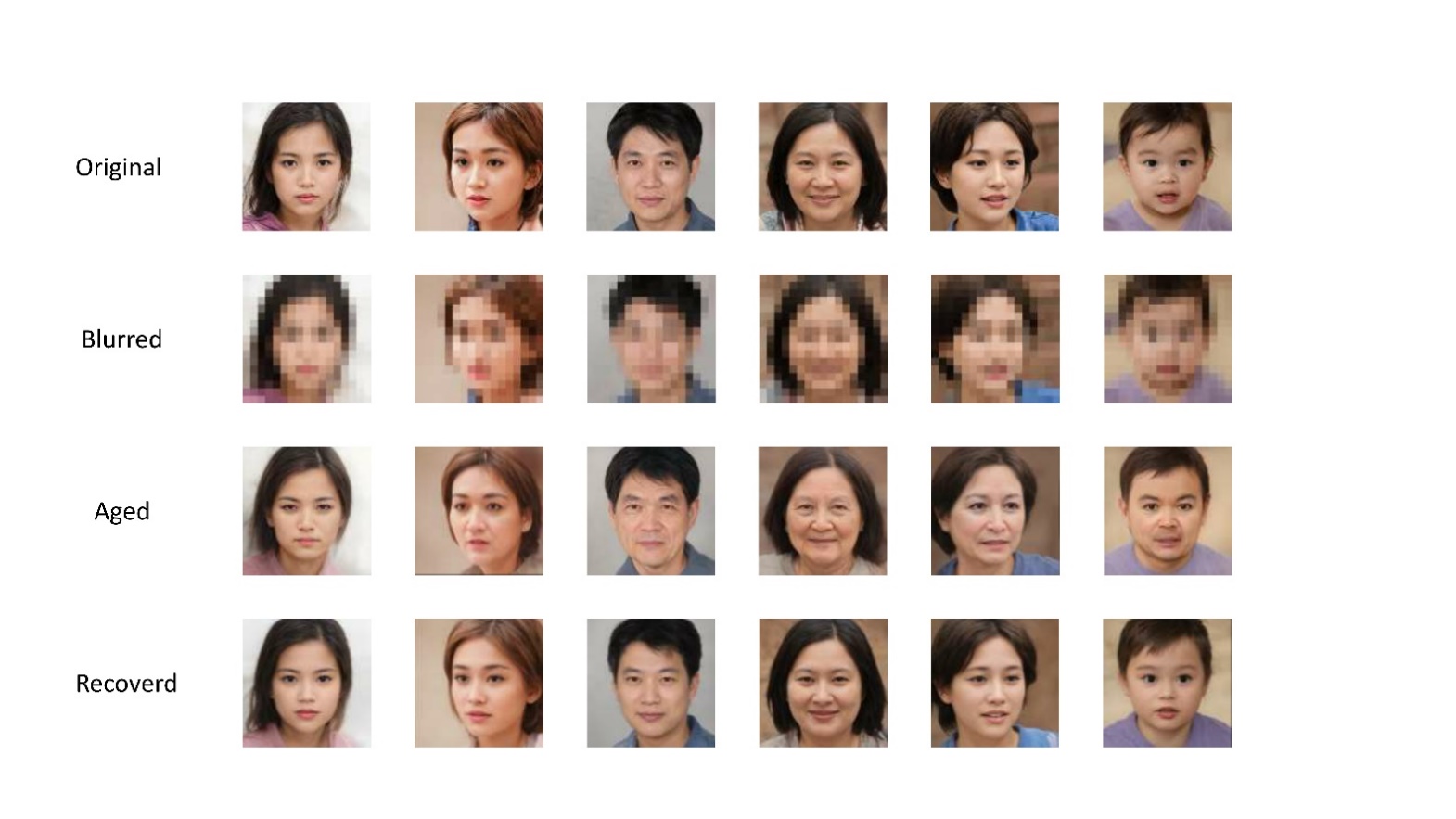
If the face feature can find a matching key in the key pool, we will confirm its identity ID and the key is accompanied by a spatiotemporal index that provides information about the face region in the video. Since the locally stored video is the face region after Gaussian blurring, there is no way to directly decrypt the face region in this video. Therefore, it is necessary to generate the decrypted face using GAN to proceed with the face recovery process. The key is fed into the generator of styleGAN, which will use the specified key to generate a restored face image, which is then applied to the video to realize the decryption. Further, styleGAN utilizes unsupervised image super-resolution methods to transform low-resolution images into high-quality, high-resolution images. This process reshapes the intricate details of the image, such as skin color, eyes, lips, etc., so that the decrypted face is restored to its original appearance.

With the above process, we have implemented a secure decryption method that allows decryption of only registered faces while protecting personal privacy and maintaining image quality.

## D. Results Discussion

**a) Encryption and decryption effects**

The effect of face encryption and decryption is shown in Fig. 4. The original image is denoted by "Original", which presents the unprocessed face . The encrypted image is represented as "Blurred", in which the face is Gaussian blurred so that it is difficult to recognize the identity of the person, while also protecting their privacy. On the other hand, the image after the



**Fig. 4.** **The effect of face encryption and decryption**

aging process is called "Aged", which represents the newly captured face . After has gone through the coding model to extract features and match with the cloud key pool, the decryption of the encrypted face image is realized under the guidance of the matched face feature key. The results show that the decryption process produces the "Recovered" image which presents the face restored to its original state. This process iscapable of restoring previously encrypted blurred images to regain clear and recognizable features. Moreover, the face feature key has high robustness, even if the young people’s face turns old, it still has the ability to decrypt, and support the recovery of the young appearance of the people. It shows that the algorithm has the similar ability of human brain’s perception associative memory

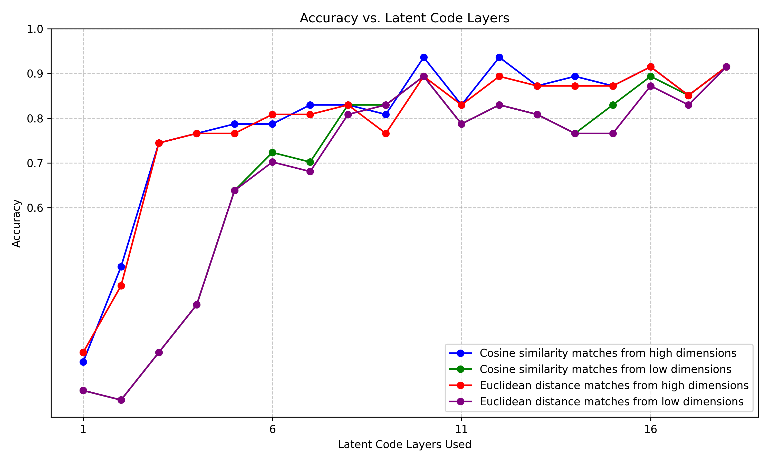
**b) Feature Matching Methods**

In our experiments, we used the same model for extracting features from 47 photos that had been processed with face aging. Then, we selected certain dimensions from these extracted features to match with the features in the pre-constructed face feature key pool and calculated the matching accuracy. This matching accuracy can be calculated as follows.

where TP denotes the number of faces that were successfully restored by matching correctly, and N denotes the number of faces in the test set, i.e., the number of new input images that have been processed by face aging = 47.

To fully investigate the impact of different matching methods, we used different combinations of dimensionality selection in our experiments. These combinations cover the matching methods from high to low dimensions. Specifically, the high-dimensional matching starts from dimension 1, followed by 1-2, 1-3, all the way up to 1-18; while the low-dimensional matching starts from dimension 18 and decreases in steps of 1, followed by 17-18, 16-18, and so on.

In the matching process, we used cosine similarity and Euler distance as matching criteria. For cosine similarity, we chose the maximum similarity as the matching result; and for Euler distance, we chose the minimum distance as the matching result. The matching accuracies of different matching methods are listed in Table 2 and Table 3. Table 2 represents the matching accuracies of two matching criteria, cosine similarity, and Euler distance, for matching starting from high dimensions. Table 3 represents the matching accuracies of the two matching criteria, starting from the low dimension. The comparisons of the four matching methods are shown in Fig. 5, from which it can be seen that the matching accuracies of the four methods basically show an increasing trend as more layers of latent code are used. "Cosine similarity matches from high dimensions" basically achieves the highest precision, and the highest precision is 0.9362. However, the highest precision does not appear at latent code layers used=18, which suggests that the matching may start from high dimensions, and as more layers of latent code are used, more low dimensional information is introduced into the feature matching, which may lead to worse matching results.



**Fig. 5.** **The Accuracy of the four matching methods.**

This observation provokes deeper thinking, revealing the diversity of feature extractors of different dimensions in capturing data. Different dimensions may play an important role in focusing on different aspects and details of the data. Specifically, feature extractors with low dimensionality may be more focused on capturing localized features, such as small

**TABLE 2**

|  |  |  |
| --- | --- | --- |
| MATCHING FROM HIGH DIMENSIONS. | | |
| latent code layers used | **Accuracy** | |
| Cosine similarity | Euclidean distance |
| 1-1 | 0.2553 | 0.2766 |
| 1-2 | 0.4681 | 0.4255 |
| 1-3 | 0.7447 | 0.7447 |
| 1-4 | 0.7660 | 0.7660 |
| 1-5 | 0.7872 | 0.7660 |
| 1-6 | 0.7872 | 0.8085 |
| 1-7 | 0.8298 | 0.8085 |
| 1-8 | 0.8298 | 0.8298 |
| 1-9 | 0.8085 | 0.7660 |
| 1-10 | 0.9362 | 0.8936 |
| 1-11 | 0.8298 | 0.8298 |
| 1-12 | 0.9362 | 0.8936 |
| 1-13 | 0.8723 | 0.8723 |
| 1-14 | 0.8936 | 0.8723 |
| 1-15 | 0.8723 | 0.8723 |
| 1-16 | 0.9149 | 0.9149 |
| 1-17 | 0.8511 | 0.8511 |
| 1-18 | 0.9149 | 0.9149 |

**TABLE 3**

**MATCHING FROM LOW DIMENSIONS.**

|  |  |  |
| --- | --- | --- |
|  | | |
| latent code layers used | **Accuracy** | |
| Cosine similarity | Euclidean distance |
| 18-18 | 0.1915 | 0.1915 |
| 18-17 | 0.1702 | 0.1702 |
| 18-16 | 0.2766 | 0.2766 |
| 18-15 | 0.3830 | 0.3830 |
| 18-14 | 0.6383 | 0.6383 |
| 18-13 | 0.7234 | 0.7021 |
| 18-12 | 0.7021 | 0.6809 |
| 18-11 | 0.8298 | 0.8085 |
| 18-10 | 0.8298 | 0.8298 |
| 18-9 | 0.8936 | 0.8936 |
| 18-8 | 0.7872 | 0.7872 |
| 18-7 | 0.8298 | 0.8298 |
| 18-6 | 0.8085 | 0.8085 |
| 18-5 | 0.7660 | 0.7660 |
| 18-4 | 0.8298 | 0.7660 |
| 18-3 | 0.8936 | 0.8723 |
| 18-2 | 0.8511 | 0.8298 |
| 18-1 | 0.9149 | 0.9149 |

variations and details in the data. In contrast, high-dimensional feature extractors are more inclined to capture global features that involve the overall structure and abstract concepts of the data.

This view is particularly notable in applications in the field of face recognition. The increased reliance on high-dimensional

abstract information for face recognition tasks may echo our human face recognition process. In the human mind, face recognition involves not only local features such as eyes and mouth but also overall facial structure and high-level semantic features. This observation actually provides strong support for face association re-recognition, suggesting that humans pay more attention to high-dimensional semantic features in face recognition.

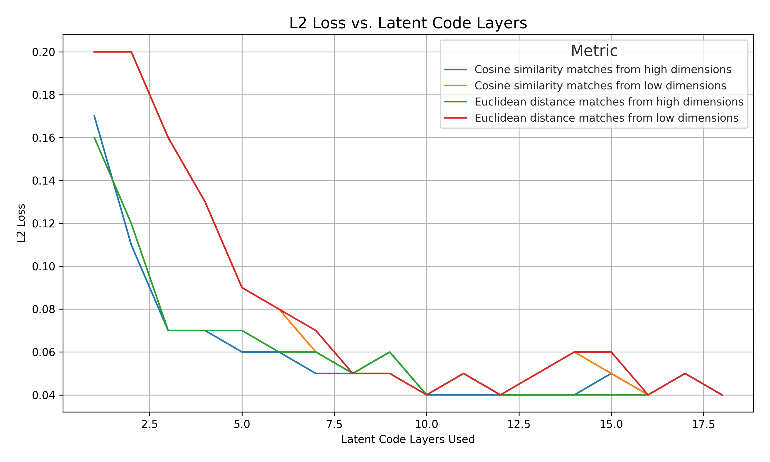
When solving matching problems, it is important to acknowledge the impact of differences in dimensionality on feature representation. Properly balancing the integration of low-dimensional and high-dimensional information, as well as understanding how to effectively fuse them, will enhance the matching effect. Taken together, this observation provides useful insight into understanding the diversity of feature extraction and the role of different dimensional information in the matching task.

**c) Multiple loss functions**

In face encryption and decryption, the fidelity and similarity of decrypted restored faces play a crucial role in determining the quality of decryption results. In order to pursue high-quality face reconstruction and restoration, we use several methods to calculate the similarity of face restoration. These methods not only consider the pixel-level fidelity of the image, but also deeply analyze the facial features and attributes in the image to more accurately assess the degree of similarity between the restoration result and the original reference, and provide direction and reference for the selection of the training loss function.

First, we introduce the pixel-level L2 loss function, calculated as follows.

Where denotes the input face image, denotes the entire face encryption and decryption algorithm framework, and denotes the entire face reconstructed by restoring after the encryption and decryption process.

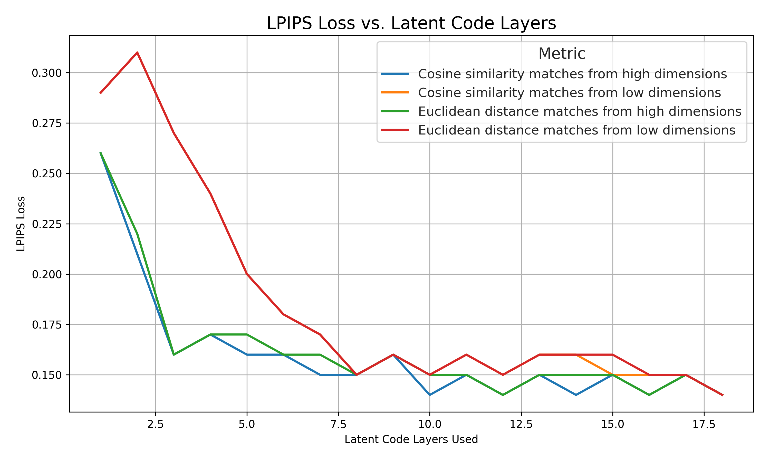


**Fig. 6.** **Trends in the L2 loss function in the test set**

Using this loss function, we can measure the pixel-level difference between the decrypted facial image and the restored facial image, thus providing a quantifiable metric for evaluating the fidelity of the reconstruction process. The evaluation results of restoring 47 aged images using the above loss function are shown in Fig. 6. Among the four different matching methods, the cosine similarity matching utilizing high dimensionality produces the lowest loss value. This indicates that using this matching method ultimately results in the highest facial similarity during the restoration process. It is also worth noting that the use of a relatively small amount of high-dimensional information has resulted in a significant reduction in the loss values to a minimum point. However, the pixel-level L2 loss function may not be able to capture subtle differences in facial features that are important for visual similarity.

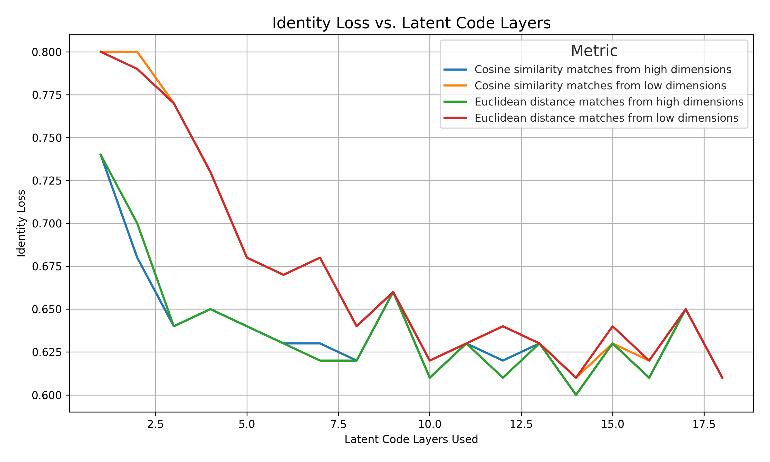
To measure the similarity of face restoration more comprehensively and ensure perceptual similarity between the restored face image and the original image, the Loss Perceptual Image Patch Similarity (LPIPS) loss is introduced. Specifically, this loss evaluates the similarity between the original face image and the decrypted and reconstructed face image by comparing their distances in the feature representation space. The mathematical expression for this loss function is:

where denotes the perceptual feature extractor.



**Fig. 7.** **Trends in the LPIPS loss function in the test set**

By using the LPIPS loss function, we are able to ensure that the restored face image remains perceptually similar to the original image, thus providing for the overall quality of the decryption process. This approach not only quantitatively evaluates the perceptual similarity of the images, but also optimizes the decryption results to better align with the criteria of perceived image quality by the human eye. The evaluation results of 47 aged images restored using the above loss function are shown in Fig. 7. it can be observed that the four different matching methods during the restoration process using the LPIPS loss function gradually reduce the loss as the number of latent code layers used increases, similar to Fig 6.



**Fig. 8.** **Trends in the Identity loss function in the test set**

Finally, in order to ensure that both the input and output of the encryption and decryption process are face images, we introduce an identity loss function related to face recognition, which ensures the identity consistency of the images by comparing the similarity between the input image and the output image in terms of face recognition.

Specifically, the identity loss is computed as follows:

Where denotes the pre-trained ArcFace network and the input images are subjected to crop and resize operations to unify them to a size of 112 \* 112 for matching with the network. denotes the cosine similarity between the output image and the original image is calculated.

The results of the evaluation are shown in Fig. 8. It can be observed from the figure that when using the identity loss, the loss decreases gradually as the number of latent code layers used increases, which is consistent across all four different matching methods. However, the loss function converges slightly slower compared to the other two loss functions. This implies that introducing identity loss increases the complexity of model training but can lead to better fidelity and similarity.

In this section, we describe in detail the various loss functions we employ and their design principles. By embedding these loss functions in the face encryption and decryption process, we realize a comprehensive assessment of the quality of decrypted restored faces. Experimental results show that the combination of the above multiple loss functions can effectively measure the fidelity and similarity of the restoration results, thus providing a solid foundation for high-quality face reconstruction restoration.

# V. Conclusion

## In this paper, we propose an algorithm for encrypting and decrypting video abstractions based on a cognitive mechanism that associates high and low-dimensional information. The aim is to address problems related to personal privacy and the management of massive video data in robotic vision application scenarios. When producing the video data, a face recognition and face encryption algorithm based on CNN and Mosaic image processing technology is used to encrypt the image in the video stream. Simultaneously, the encoder based on GAN generator is utilized to extract the high and low-dimensional semantic information of the face to form a feature key and establish an index association with the face in the video, which is stored in the corresponding key pool of the video. When decrypting, the features extracted from the specified new faces are used to search in the key pool, and the matched faces are recovered to former faces using the GAN. We initially validate the feasibility of this video encryption and decryption algorithm on a self-built dataset. At the same time, we conduct comparative experiments and analysis of multiple matching methods and loss function selection to inversely verify the importance of high-dimensional abstract semantic features in the process of face association re-recognition. Results show that the face feature key exhibits considerable robustness, enabling decryption and restoration of individuals' youthful appearance despite aging, reflecting the associative memory capability akin to the human brain. Our research has important implications in terms of how to strike a balance between safeguarding privacy and advancing machine vision research. In the future, this algorithm is promising to play an important role in privacy protection and big data storage, face re-identification, and other fields.

Acknowledgement

This work was co-supported by the National Key Research and Development Program of China (No. 2021YFE0193900 and 2020AAA0108901), the National Natural Science Foundation of China (No. 52002286, 51975415 and 62088101), the Fundamental Research Funds for the Central Universities (22120220642).

Data availability

"generated\_yellow-stylegan2" dataset. Available at: <https://github.com/a312863063/generators-with-stylegan2>.

Age Filter from the Wondershare AI platform AIlab. Link to the aging process website: <https://ailab.wondershare.com/tools/aging-filter.html>

References

1. Robinson N, Tidd B, Campbell D, et al. Robotic vision for human-robot interaction and collaboration: A survey and systematic review[J]. ACM Transactions on Human-Robot Interaction, 2023, 12(1): 1-66.
2. Ş. Kolozali et al., "Observing the Pulse of a City: A Smart City Framework for Real-Time Discovery, Federation, and Aggregation of Data Streams," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2651-2668, April 2019.
3. O. Styles, T. Guha and V. Sanchez, "Multi-Camera Trajectory Forecasting with Trajectory Tensors," IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, 44(11): 8482-8491.
4. C. W. Chen, "Internet of Video Things: Next-Generation IoT With Visual Sensors," IEEE Internet of Things Journal, vol. 7, no. 8, pp. 6676-6685, Aug. 2020.
5. Zou, Zhengxia, et al. "Object detection in 20 years: A survey." arXiv preprint arXiv:1905.05055 (2019).
6. W. N. I. Al-Obaydy and S. A. Suandi, "Open-set face recognition in video surveillance: a survey," in *InECCE2019*: Springer, 2020, pp. 425-436.
7. A. H. Ahmad et al., "Real time face recognition of video surveillance system using haar cascade classifier," vol. 21, no. 3, pp. 1389-1399, 2021.
8. C. Shirley, N. Ram Mohan, B. J. M. S. Chitra, and S. Processing, "Gravitational search-based optimal deep neural network for occluded face recognition system in videos," vol. 32, no. 1, pp. 189-215, 2021.
9. Z. Lei, X. Zhang, S. Yang, Z. Ren, and O. F. J. E. I. S. Akindipe, "RFR-DLVT: a hybrid method for real-time face recognition using deep learning and visual tracking," vol. 14, no. 9-10, pp. 1379-1393, 2020.
10. M. Liu, J. Liu, P. Zhang, and Q. J. I. A. Li, "PA-GAN: A patch-attention based aggregation network for face recognition in surveillance," vol. 8, pp. 152780-152789, 2020.
11. W. N. I. Al-Obaydy, S. A. J. N. C. Suandi, and Applications, "Open-set single-sample face recognition in video surveillance using fuzzy ARTMAP," vol. 32, no. 5, pp. 1405-1412, 2020.
12. W. N. I. Al-Obaydy, S. A. J. M. T. Suandi, and Applications, "Automatic pose normalization for open-set single-sample face recognition in video surveillance," vol. 79, no. 3, pp. 2897-2915, 2020.
13. X. Zhao, Y. Chen, E. Blasch, L. Zhang, and G. Chen, "Face recognition in low-resolution surveillance video streams," in Sensors and Systems for Space Applications XII, 2019, vol. 11017, pp. 147-159: SPIE.
14. N. Singh, S. S. Rathore, S. J. M. T. Kumar, and Applications, "Towards a super-resolution based approach for improved face recognition in low resolution environment," pp. 1-33, 2022.
15. D. S. Dharrao, N. J. J. I. J. o. C. I. Uke, and Applications, "Fractional Krill–Lion algorithm based actor critic neural network for face recognition in real time surveillance videos," vol. 18, no. 02, p. 1950011, 2019.
16. M. J. I. J. o. A. C. S. Imandito and Applications, "Face Recognition on Low-Resolution Image using Multi Resolution Convolution Neural Network and Antialiasing Method," vol. 10, no. 12, 2019.
17. H. Tamiya, T. Isshiki, K. Mori, S. Obana, and T. Ohki, "Improved Post-quantum-secure Face Template Protection System Based on Packed Homomorphic Encryption," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1-5: IEEE.
18. R. Román, R. Arjona, P. López-González, and I. Baturone, "A Quantum-Resistant Face Template Protection Scheme using Kyber and Saber Public Key Encryption Algorithms," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2022, pp. 1-5: IEEE.
19. H. Huang, L. J. J. o. I. S. Wang, and Applications, "Efficient privacy-preserving face verification scheme," vol. 63, p. 103055, 2021.
20. Y. Yang, Q. Zhang, W. Gao, C. Fan, Q. Shu, and H. J. W. P. C. Yun, "Design on Face Recognition System with Privacy Preservation Based on Homomorphic Encryption," vol. 123, no. 4, pp. 3737-3754, 2022.
21. L. Jiasen, W. X. An, C. Bowei, T. Zheng, Z. J. I. J. o. M. C. Kaiyang, and M. Communications, "Outsourced Secure Face Recognition Based on CKKS Homomorphic Encryption in Cloud Computing," vol. 12, no. 3, pp. 27-43, 2021.
22. T. N. Tan, Y. Hyun, J. Kim, D. Choi, and H. Lee, "Ring-LWE based face encryption and decryption system on a GPU," in 2019 International SoC Design Conference (ISOCC), 2019, pp. 15-16: IEEE.
23. P. Duong-Ngoc, T. N. Tan, and H. J. I. A. Lee, "Efficient NewHope cryptography based facial security system on a GPU," vol. 8, pp. 108158-108168, 2020.
24. Q. Zhao et al., "Speckle-based optical cryptosystem and its application for human face recognition via deep learning," 2022.
25. K. Nakai, M. Kuribayashi, and N. Funabiki, "A Study of Privacy Protection of Photos Taken by a Wide-angle Surveillance Camera," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021, pp. 1865-1871: IEEE.
26. K. M. Hosny, M. A. Zaki, H. M. Hamza, M. M. Fouda, and N. A. J. I. A. Lashin, "Privacy Protection in Surveillance Videos Using Block Scrambling-Based Encryption and DCNN-Based Face Detection," vol. 10, pp. 106750-106769, 2022.
27. Y. Shen, C. Tang, M. Xu, Z. J. O. Lei, and L. Technology, "Optical selective encryption based on the FRFCM algorithm and face biometric for the medical image," vol. 138, p. 106911, 2021.
28. Z. Liu, J. Li, and J. J. M. B. E. Liu, "Encrypted face recognition algorithm based on Ridgelet-DCT transform and THM chaos," vol. 19, pp. 1373-1387, 2022.
29. Liu, Y. , J. Tang , and T. Xie . "Cryptanalyzing a RGB image encryption algorithm based on DNA encoding and chaos map." Optics & Laser Technology 60(2014):111-115.
30. C. Wu, B. Ju, Y. Wu, N. N. Xiong, and S. J. E. Zhang, "WGAN-E: A generative adversarial networks for facial feature security," vol. 9, no. 3, p. 486, 2020.
31. H. J. M. T. Ashiba and Applications, "Presented cancelable face recognition system using graph theory," pp. 1-22, 2022.
32. N. Franklin, K. A. Norman, C. Ranganath, J. M. Zacks, and S. J. Gershman, "Structured event memory: a neuro-symbolic model of event cognition," 2019.
33. K. Sun, C. Guo, H. Zhang, and Y. Li, "HVLM: Exploring Humanoid Visual Cognition and Language-Memory Network for Visual Dialog," *Information Processing & Management,* vol. 59, no. 5, p. 103008, 2022/09/01/ 2022.
34. Richardson, Elad, et al. "Encoding in style: a stylegan encoder for image-to-image translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
35. Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)