[[1]](#footnote-1)

Face Privacy Protection and Self-decryption Method Based on Humanoid Association Mechanism

Kaijing Ma, Qiwei Du, Zhongpan Zhu\*, Bin He\*, *Member, IEEE*, Zhipeng Wang, Gang Li

***Abstract*—** **With the large-scale industry application of artificial intelligence and video surveillance, massive video data storage and personal privacy issues are highlighted, which restrict the application expansion. From the perspective of humanoid memory mechanism, we propose a video abstraction encryption and decryption algorithm based on high and low dimensional information association cognitive mechanism, which uses face recognition algorithm to locate human faces and encrypt the original video with mosaic encryption method, and then perform spatio-temporal index encoding, and further use abstract face feature memory to match and decode with the same identity person to construct a de-mosaic decryption key function. The core innovation is a humanoid memory mechanism for parsing and modeling, which is combined with specific AI techniques such as YOLO5-face, pSp encoder and styleGAN generator for initial experimental validation in face encryption and decryption. The algorithm research will be important in the research of video information compression and storage, person re-identification and personal privacy protection.**

***Index Terms*—Face detection, Privacy protection, Humanoid cognition, Association cechanism, GAN**

# I. INTRODUCTION

The surveillance cameras distributed in all corners of the city play an indispensable role in the city security management. Surveillance video/image data collection and analysis based on AI and IOT technologies has been an important technical grip for different scenes in smart city development.

The extraction of trust information from surveillance data has attracted the interest of many researchers and has led to the analysis of images from numerous IoT vision sensors [1-3]. However, the large-scale deployment of vision sensors leads to a number of challenges: 1) First, the huge number of camera video images leads to a data disaster. At 30 frames per second and 5MB per image, a single camera generates a data storage requirement of 12,656.25Gb a day, while IHS research indicates that there will be over one billion surveillance cameras worldwide in future. These video stores take up a large amount of hardware resources, and no data center can withstand the daily growth of video data, which must be overwritten on a regular basis [2]. 2) Secondly, information redundancy in massive camera video data leads to key information being overwritten and video-based information retrieval being difficult [3]. 3) In addition, massive video transmission takes up a large amount of communication bandwidth, and communication costs are high, making it difficult to achieve widespread cameras for collaborative use to achieve mega-city governance [4]. 4) Meanwhile surveillance cameras have led to the leakage of residents' biometric privacy, raising ethical and regulatory concerns. How to safeguard the functionality of surveillance cameras while improving the above challenges has become a research direction for a wide range of scholars.

In this paper, we take a humanoid cognitive perspective to carry out theoretical research for exploring new models of large-scale camera urban applications. We humans, from infants to the elderly, perceive a large amount of picture information with both eyes over decades and can have long-term clear memories of the people and things we experienced. However, we are often unable to reproduce all of the image information that occurred, but rather combine it with high-dimensional semantic abstraction to achieve coarse-grained picture recall. We also tend to remember familiar faces not through detailed facial features such as single or double eyelids, but into general impressions of higher-dimensional semantic information. In addition, the high-dimensional abstract semantics in our human brain memory plays an important role in blurring human decryption recognition. Humans can recognize acquaintances through blurred or partially blurred facial images, but not strangers. The process of humanoid perceptual memory mechanism to handle the massive amounts of video data is difficult to have a theoretical explanation. But the association between low-dimensional fine-grained information and higher-dimensional coarse-grained information for humanoid perceptual data compression and decryption has theoretical significance and practical value, which are worth using for processing massive surveillance vidio data. In this paper, we try to propose an autonomous face degradation encryption and decryption algorithm based on the above humanoid association memory mechanism.

The main contributions are summarized as follows.

1. For the first time, the face facial perception-memory-association re-recognition mechanism is applied to massive surveillance video processing, the mechanism is abstractly analyzed and modeled, and an autonomous face degradation encryption/decryption algorithm process based on the above mechanism is proposed.
2. The autonomous face degradation encryption and decryption algorithm modeled after the human associative memory mechanism is integrated into the AI face recognition algorithm, and the algorithm for selective encryption and decryption of the face region showing identity in the video is preliminarily validated on the dataset.
3. We analyze the matching method and loss function selection, and prove that the feature matching effect will be better by using the cosine similarity (taking the highest similarity) as the matching criterion and selecting some high-dimensional features as the matching basis, which inversely verifies the importance of the high-dimensional abstract semantic features in the process of face association re-recognition.

The rest of the paper is organized as follows. Section II describes the related work of the method, including face recognition of video surveillance, face encryption and decryption algorithm and humanoid memory cognition. Section III models and abstracts the process of human face recognition, memory and associative re-recognition, and proposes an algorithmic process for face encryption and decryption with human-like cognitive mechanisms, which is the core of the paper. In Section IV, the initial validation of the proposed encryption and decryption algorithm on the dataset is presented in detail. Section V presents the conclusions and future work.

# II. RELATED WORK.

## A. Face recognition of video surveillance

In the perspective of recent advances in the field of AI-driven face recognition of video surveillance, the human face object tracing for video surveillance has gained widespread adoption in urban security and community management. A lot of scholars are committed to the research of computer vision technique 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 utilizing enormous quantities of data [8-10]. With the development of deep learning technology, the application boundary of face recognition will be gradually opened. The majority of face recognition in video surveillance today is "closed-set," which only recognizes the identity of previously registered objects. However, "open-set" has gained popularity as a result of the differences between the source and target domains, which make it less effective when transferring face recognition systems from controlled environments to uncontrolled scenes. Suandi *et al*. [11,12], proposed fuzzy ARTMAP neural networks to solve the open-set single-sample face recognition problem and an automatic pose normalization technique without model fitting and human intervention, which greatly improves the performance of open-set single-sample face recognition methods in surveillance environments. The "open-set" face recognition prone to increase the human privacy exposure degree in the ubiquitous city surveillance network.

The low resolution of urban monitoring picture and the difficulty of small face feature extraction are being changed. Even though the surveillance cameras are usually placed far away from the objects and the resolution of the captured face images is low due to distance, extensive research has been carried out for recognizing acceptable recognition features at 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 surveillance videos. Singh *et al*. [14] improved the number of descriptors in the image and mitigates the effects of noise based on super-resolution faces. Dharrao *et al*. [15] used the Viola-Jones algorithm to detect the face part in the video sequential frames and improved the quality of the face part by applying a super-resolution scheme based on bicubic interpolation. In addition, the 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 citizen's face privacy feature under the ubiquitous cameras is more and more serious. How to explore a new paradigm for large-scale camera urban applications from the perspective of humanoid cognition by performing face reduction encryption on the recognized video images are meaningful.

## B. Face encryption and decryption algorithm

The problem of privacy leakage has aroused widespread concern. Face recognition of video surveillance have become ubiquitous in daily lives, but it is difficult to balance between intelligent vision applications and personal privacy protection. In addition to improving relevant laws and regulations to regulate the acquisition, storage and use of videos, corresponding technical measures are needed to protect personal privacy. The cryptography-based face privacy protection scheme selectively encrypts the face region in the video that shows the identity and can be decrypted to recover the original video in case of future legitimate demand. How to integrate the autonomous face degradation encryption and decryption algorithm of humanoid association memory mechanism into 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 to achieve smaller protected template and key sizes and faster execution times than other HE schemes that use 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 normalised facial feature vector [20,21].

Due to the low computational efficiency of using homomorphic encryption, other studies tried to find lightweight algorithms to encrypt faces. Tan *et al*. [22] proposed a novel approach to implement 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 to comprehensively protect facial images extracted from videos based on NewHope cryptography for post-quantum cryptosystems, greatly reducing the time for encryption and decryption. Zhao *et al*. [24] proposed and implemented a simple and efficient speckle-based optical cryptosystem to encrypt face images by seemingly random optical speckles at the speed of light, by training an 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 [27]. A THM (Tent-Henon Map) chaotic encryption of faces was proposed in combined with the properties of tent chaos and Henon chaos [28]. Liu *et al*. [29] proposed a RGB image encryption algorithm based on DNA encoding and chaos map. Wu *et al*. [30] proposed a Generative Adversarial Network (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 are still room for improvements 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.

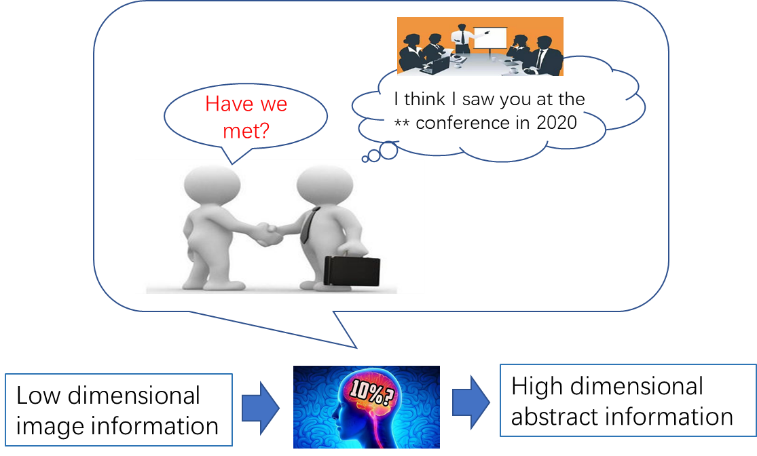
## C.Humanoid memory cognition

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, and illustrates a wide range of 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 of face privacy protection for urban large-scale camera monitoring with. 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 are remembered. As shown in figure 1 for example, when some people meet with each other unintentionally, their mind will unconsciously recall that they have seen such a face at a certain time, place and event. Moreover, they can recall the memory of more detailed scene and clearer features. The process can actually 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, and the past feature-blurred memory scene is clearly reproduced 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* people’s face perceived by the brain in the first stage, is the set of high-dimensional abstract semantic features of *ith* people’s face formed by the brain in the mind based on, is the encrypted data set 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 as follows.

|  |
| --- |
| **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 the face encryption and decryption requirements of surveillance video to solve the following problems.

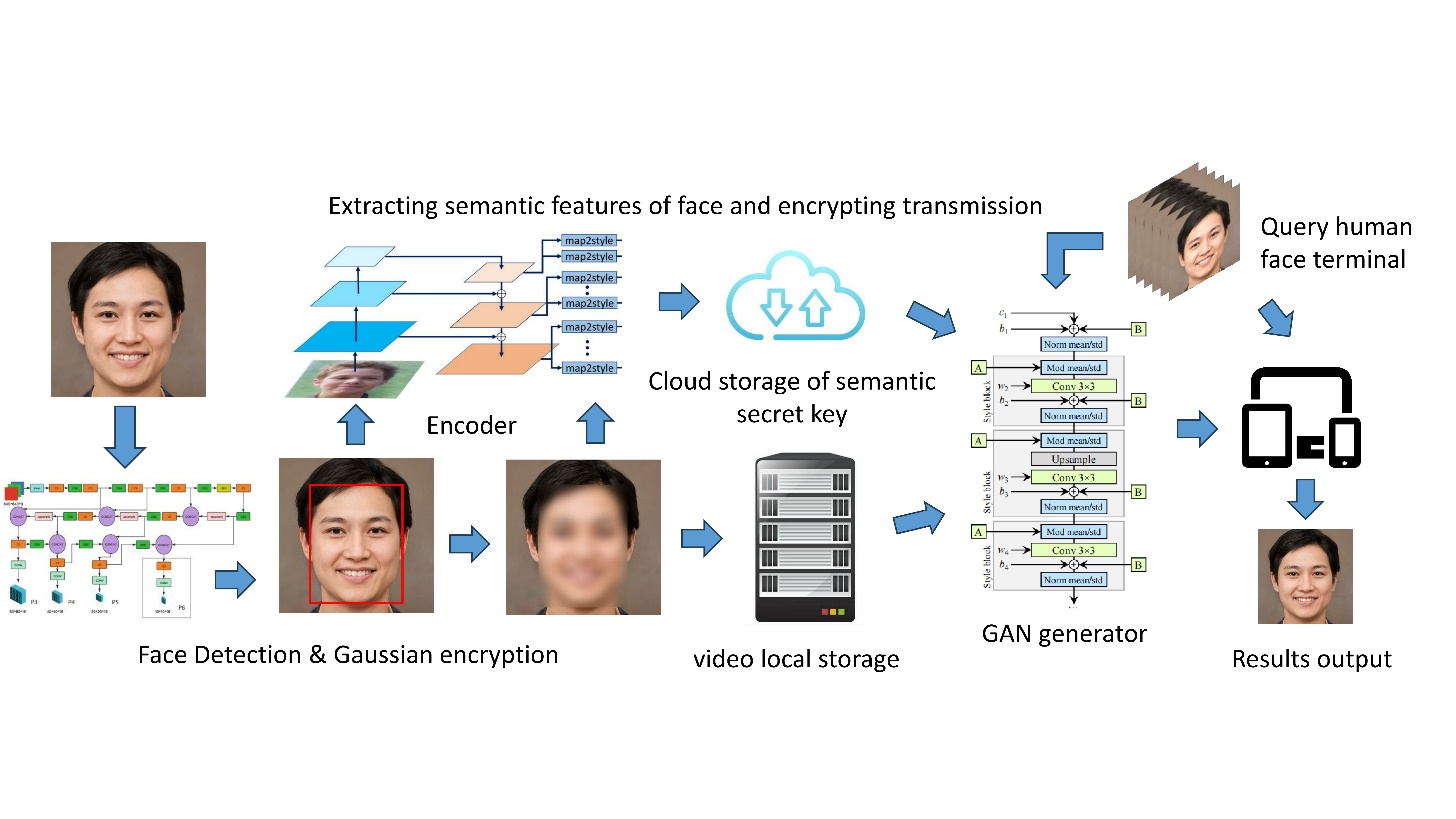
1) To modeling the humanoid cognitive mechanism, the high-dimensional abstract memory and compressed perception process *f1* function need to be solved. and propose an artificial intelligence algorithm for solving and to identify and locate faces in videos, extract high-dimensional semantic features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

2) Drawing on humanoid associative memory mechanism, the algorithm models the memory storage of high-dimensional semantic features and associative matching *f2*, and proposes a recall-triggered matching index mechanism to achieve associative memory matching based on and .

3) Drawing on the humanoid perception-triggered recall mechanism, the associative recall of high-dimensional semantic features and low-resolution video is modelled 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 surveillance video faces as shown in the Fig 2.



**Fig. 2. Framework of personal self-decryption.**

**1) Encryption method**

For the video frame input *V*, the YOLO5-face deep learning model *ф* is used to achieve the recognition and localisation of faces by the surveillance cameras at the edge end, into obtaining *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 localisation 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.

**2) Key storage and matching**

Unlike traditional video surveillance systems, this method no longer stores the original video, but chooses to locally store the encrypted video, while uploading the high-dimensional abstract semantics to the cloud for subsequent processing and analysis. For example, for the face retrieval service of post-surveillance, as the local storage of encrypted faces loses a large amount of face feature information, the video cannot be retrieved for review, but 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 to the desired latent code, 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 . E(\*) 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, so as 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, 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, and the mapping relationship between the high-dimensional abstract semantic feature and the encrypted face image frames is associated with the above symbolic bit encoding to facilitate subsequent video decryption work. These together serve as the key for decrypting the 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, while using abstract semantics and as the identity key to determine the identity of faces becomes more challenging.

The identity recognition of a face is divided into two parts, one part is to determine the identity by matching the specific features of the face, not simply by comparing the specific features; the other part belongs to the impression abstract semantics, just like the general impression given by the face, and the identity is determined by matching the specific features of the 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 the process of key storage and query matching. This section discusses how to output the 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 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.

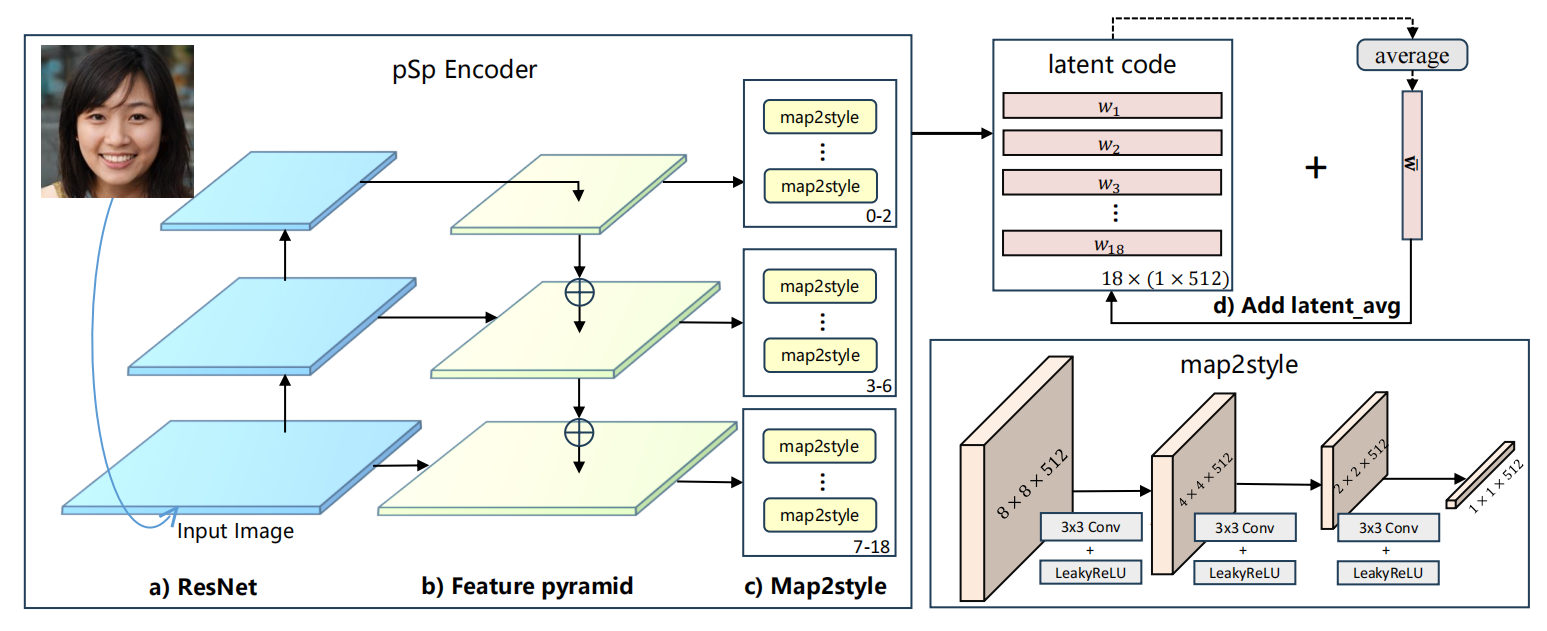
# IV. EXPERIMENTS AND RESULTS

## 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 of different genders, ages, and head pose variations. During training and encryption of faces, we take this dataset as input and use YOLO5-face based Gaussian encryption to obtain a Gaussian encrypted dataset paired with the original images. When decryption is needed, we use the Age Filter[[3]](#footnote-3) in AIlab, an artificial intelligence cloud platform created by Wondershare, to do aging processing on 47 images from the generated\_yellow-stylegan2 dataset, and most of the images adopt the default aging to 50 years old setting in the platform, which can obtain better aging processing results, ensuring that the feature changes of the face are still relatively recognizable, obtaining more natural and reasonable results. 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 used to simulate the normal physiological changes of the original face after a period of time in a real decryption restoration situation, so as to more realistically simulate the whole process of face encryption and decryption of surveillance 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 mode

**Fig. 3. The process of extracting latent code.**

uses a standard feature pyramid based on ResNet to extract feature mappings. For each of the 18 target styles, we trained three mapping networks of different sizes to extract the learned styles from the corresponding feature mappings. Specifically, features for 0-2 styles were generated from small feature mappings, 3-6 styles were generated from medium feature mappings, and 7-18 styles were generated from large feature mappings. In this way, we realize the high-dimensional abstract memory and compressed perception processes in human-like cognitive mechanisms. The high-dimensional feature information is generated by smaller feature mappings, which enables us to acquire and store feature information at different levels.

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 multidimensional face features from low to high dimensions. These extracted feature information are integrated and encrypted to form a key bound to 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 the 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, thus protecting data privacy.

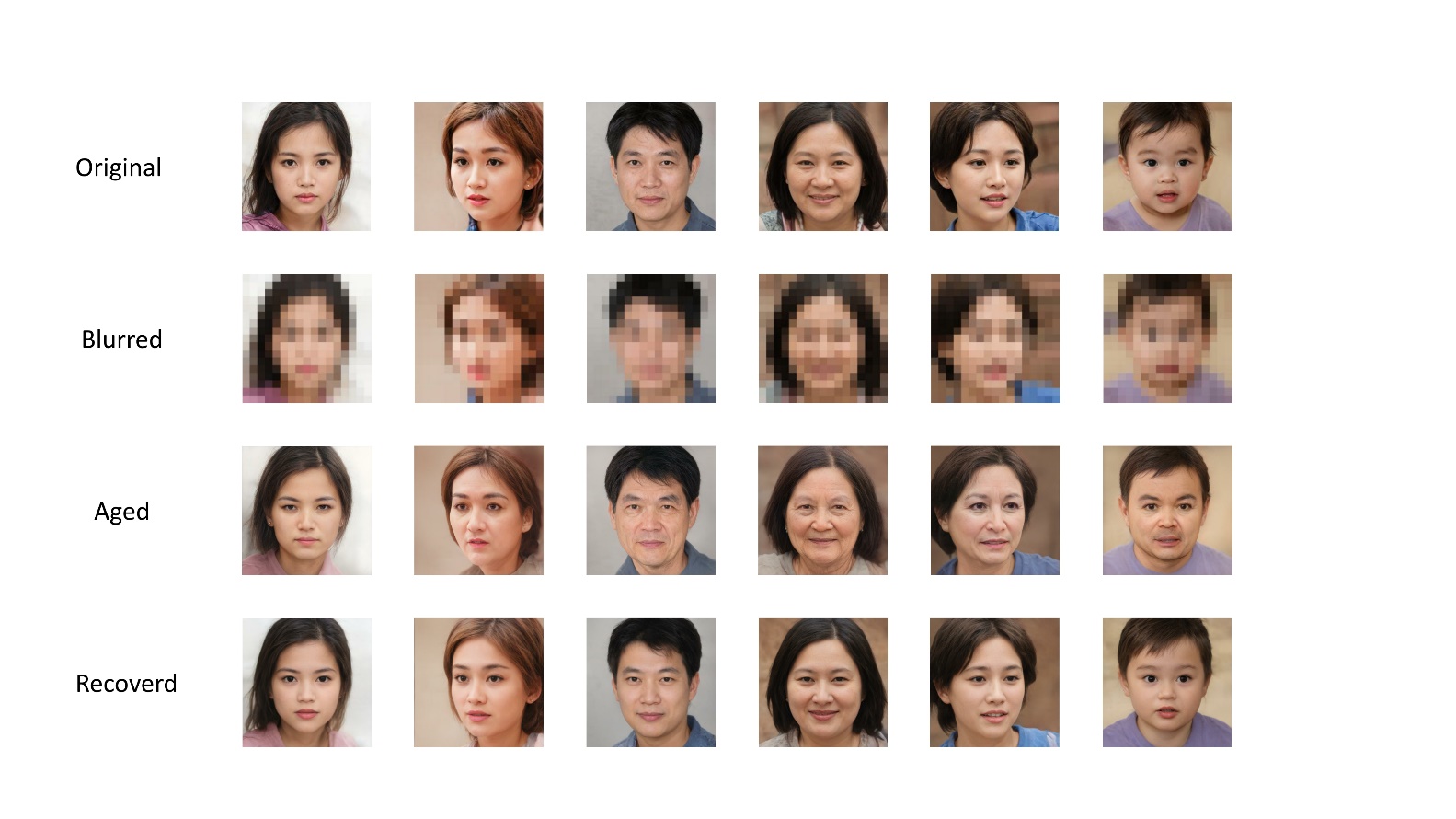
If the face feature can find a matching key in the key pool, we will confirm its identity ID, and the key is accompanied by a spatio-temporal index that will provide information about the face region in the surveillance video. Since the locally stored video is the face region after Gaussian blurring, there is no way to directly decrypt the face region in this video, so it is necessary to generate the decrypted face through GAN, so that the face recovery can be carried out. 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 surveillance 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 detailed features 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 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 by "Blurred", in which the face is Gaussian blurred so that it is difficult to recognize the identity of the face, and at the same time protects the privacy of the individual. On the other hand, the image after the aging process is called "Aged", which represents the newly captured face . After has gone through the coding model to extract features and matched with the cloud key pool, the decryption of the encrypted face image is realized under the guidance of the matched face feature key. This decryption process produces the "Recovered" image which presents the face restored to its original state. This process is capable of restoring previously encrypted blurred images to regain clear and recognizable features.



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

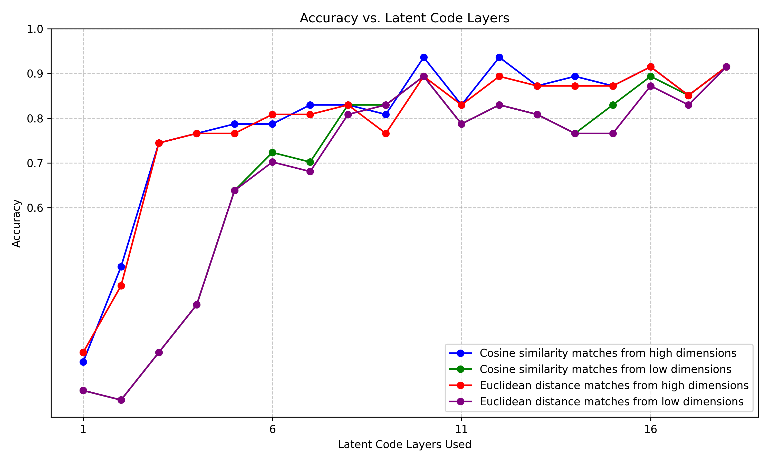
**b) Feature Matching Methods**

In our experiments, we used the same model for extracting features for 47 photos that have 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.

In order 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 I and Table II. Table I represents the matching accuracies of the two matching criteria, cosine similarity and Euler distance, for matching starting from high dimensions. Table II 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 achieve the highest precision, and the highest precision is 0.9362. However, the highest precision does not appear at latent code layers used=18, which This suggests that the matching may start from high dimensions, and as more layers of latent code layers 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 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

**TABLE I**

|  |  |  |
| --- | --- | --- |
| 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 Ⅱ**

**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 |

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.

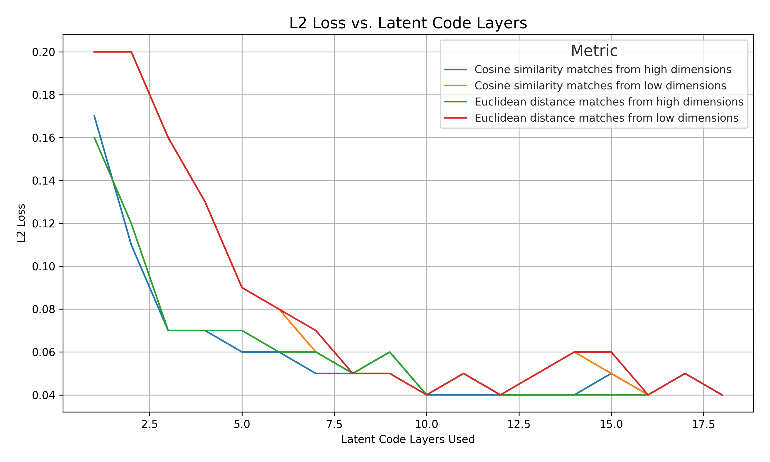
When solving matching problems, we need to recognize the role of dimensionality differences in feature representation. Properly balancing the introduction of low-dimensional and high-dimensional information and how to fuse them will help improve the matching effect. Taken together, this observation provides 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.

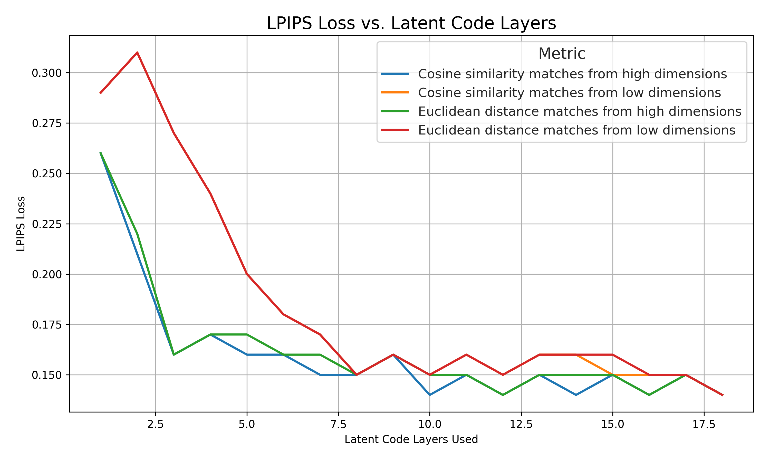


**Fig. 6.** **Trends in the L2 loss function in the test set**

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 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 some subtle facial feature differences that are important for visual similarity.

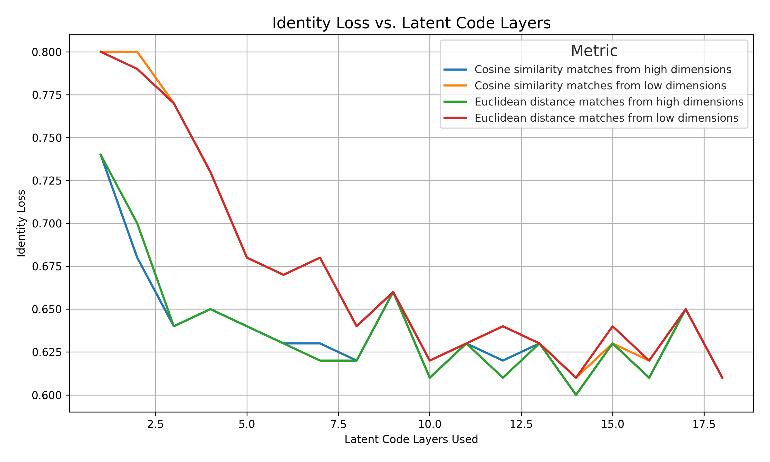
In order to measure the similarity of face restoration more comprehensively and to ensure the perceptual similarity between the restored face image and the original image, the Loss Perceptual Image Patch Similarity (LPIPS) loss is introduced. Specifically, this loss evaluates the similarity between the 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 match the human eye's criteria for perceived image quality. The evaluation results of 47 aged images restored using the above loss function are shown in Fig. 7. From the figure, 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 latent code layers used increase as the same for the 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.

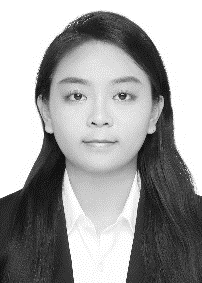
# V. Conclusion

## In this paper, we propose a video abstraction encryption and decryption algorithm based on high and low dimensional information association cognitive mechanism for massive video data disaster and personal privacy problems. When storing the video, a face recognition and face encryption algorithm based on YOLO5-face is used to encrypt the video, and at the same time, pSp encoder 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 reduced to clear faces using styleGAN generator. 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. Our research has important implications in terms of how to strike a balance between privacy protection and 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.

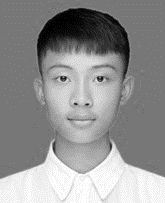
References

1. Z. Gao, C. Xu, H. Zhang, S. Li and V. H. C. de Albuquerque, "Trustful Internet of Surveillance Things Based on Deeply Represented Visual Co-Saliency Detection," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4092-4100, May 2020.
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.

**First A. Author** (Fellow, IEEE)

**** **Kaijing Ma**.

Kaijing Ma is currently a junior undergraduate student majoring in Computer Science and Technology at Tongji University. Her current research interests are computer vision and robotics.

**** **Qiwei Du**.

Qiwei Du is currently a senior undergraduate student majoring in Computer Science and Technology at Tongji University. His current research interest lies in computer vision and machine learning.

**Second B. Author**, photograph and biography not available at the time of publication.

**Zhongpan zhu**.

Zhongpan Zhu is a Research Associate Professor at School of Electronic and Information engineering of Tongji University. He received M.S. degree in Power Machinery and Engineering from School of Automotive Studies of Tongji University in 2012 and the Ph.D. degree in Vehicle Engineering 2019. He was a postdoctoral research scientist at School of electronic and information engineering of Tongji university from 2019 until 2022. He was the chief of automotive software development section in TJARD institute from 2013 until 2016.His current research interests include intelligent robots, intelligent vehicles, and AI algorithms.

**Bin He**

Bin He received the B.S. degree in engineering machinery from Jilin University, Changchun, China, in 1996, and the Ph.D. degree in mechanical and electronic control engineering from Zhejiang University, Hangzhou, China, in 2001.

He held postdoctoral research appointments with The State Key Lab of Fluid Power Transmission and Control, Zhejiang University from 2001 and 2003. He is currently a Professor with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai, China. His current research interests include intelligent robot control, biomimetic microrobots, and wireless networks.

**Zhipeng Wang** .

Zhipeng Wang received the M.S. degree from Zhejiang University, Hangzhou, China, in 2011, and the Ph.D. degree from Tongji University, Shanghai, China, in 2015.

From 2015 to 2018, he held post-doctoral research appointments with the College of Mechanical Engineering, Tongji University. He is currently a Vice Professor with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University. His current research interests include biped robots, mechatronics, and dynamics.

**Gang Li**

Gang Li received the B.S. degree in measuring and control technology and instrumentations from China Jiliang University, Hangzhou, China, in 2010, and the M.S. and Ph.D. degrees in control theory and control engineering from Tongji University, Shanghai, China, in 2013 and 2018, respectively.

He is currently an Assistant Research Fellow with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University. His current research interests include intelligent sensing and wireless sensor networks.

1. This work was co-supported by the National Key Research and Development Pro-gram of China (No. 2021YFE0193900 and 2020AAA0108901), the National Natural Science Foundation of China (No. 52002286, 51975415 and 62088101). *(Corresponding author: Zhongpan Zhu, Bin. He).* Kaijing Ma and Qiwei Du contributed equally as first author.

   Kaijing Ma , Qiwei Du, Zhongpan Zhu, Bin He, Zhipeng Wang, Gang Li are with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and with Frontiers Science Center for Intelligent Autonomous Systems, Shanghai 201210, China (e-mail: 2151400@tongji.edu.cn; 2051500@tongji.edu.cn; 521bergsteiger@tongji.edu.cn; hebin@tongji.edu.cn; zhipeng wang@tongji.edu.cn; lig@tongji.edu.cn;). [↑](#footnote-ref-1)
2. "generated\_yellow-stylegan2" dataset. Available at: https://github.com/a312863063/generators-with-stylegan2. [↑](#footnote-ref-2)
3. Age Filter from the Wondershare AI platform AIlab. Link to the aging process website: https://ailab.wondershare.com/tools/aging-filter.html [↑](#footnote-ref-3)