[[1]](#footnote-1)

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

Zhongpan Zhu, Qiwei Du, 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 the human-like memory mechanism for parsing and modelling, and combined with specific AI techniques such as YOLO and GAN 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*—Enter keywords or phrases in alphabetical order, separated by commas. For a list of suggested keywords, send a blank e-mail to** [keywords@ieee.org](mailto:keywords@ieee.org) **or visit** <http://www.ieee.org/organizations/pubs/ani_prod/keywrd98.txt>

# I. INTRODUCTION

S

urveillance 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 trust information extraction in surveillance data has aroused many researchers' interest and motivates the analysis of images from numerous IoT visual sensors [1-3]. However, the massive deployment of visual sensors leads to a number of challenges: firstly, the sheer volume of camera video images leads to a data disaster. With 30 frames per second and 5mb per image, a single camera generates a data storage requirement of 12,656.25Gb in a day, while IHS research indicates that there will be over 1 billion surveillance cameras worldwide after 2021. These video stores take up a lot of hardware resources and no data centre can afford the daily growth of video data and has to be covered periodically [4]. Secondly the redundancy of information in massive camera video data leads to key information overwriting and difficulties in retrieving information based on video []. In addition, massive video transmission occupies a large amount of communication bandwidth, with high communication costs, making it difficult to achieve collaborative use of a large range of cameras for mega-city governance. With the development of AI technology, surveillance cameras have led to the leakage of residents' biometric privacy, raising ethical and regulatory concerns. How to safeguard the function of surveillance cameras while ameliorating the above challenges has become a direction of research for a wide range of scholars.

(类人认知角度)

In this paper, we take a human-like cognitive perspective to carry out based theory 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 experience and when and where they happened. 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. Our memories of familiar faces of friends or family members are also often not remembered through images of faces, nor are they even identified by specific features such as single eyelids and double eyelids, but rather translated into a general impression of high-dimensional semantic information. In addition, the high-dimensional abstract semantics in our human brain memory plays an important role in blurring human decryption recognition, where humans can recognise the identity of people we are familiar with in blurred or partially obscured images of faces, whereas strangers have difficulty in doing so. This is a result of the long biological evolution of humans to cope with the challenges of massive amounts of AV data, and it is difficult to have a theoretical explanation of the mechanism by which human low-dimensional fine-grained information is correlated to higher-dimensional coarse-grained information. There are even fewer attempts to use the mechanism to implement video processing for massive surveillance cameras. This paper presents a preliminary exploration in this direction, proposing the study of face degradation encryption in video images and then combining high-dimensional semantic information with identity recognition for autonomous decryption, and achieving high approximation face recovery.

# II. RELATED WORK.

## A. face recognition of video surveillance

With the influence of economic globalization and the accelerating pace of urbanization, urban population density has risen, the flow of people and vehicles has increased, and the layout of urban buildings and infrastructure has become more and more complex. This has led to urban construction of traffic, social security and other urban management issues. In recent years, the diversity and complexity of the security situation is gradually increasing, and the means of crime are becoming more and more hidden, which has posed new challenges to the security management of the city. In the security management, the surveillance cameras distributed in all corners of the city play an indispensable role. And in the process of dealing with security problems, face recognition in surveillance videos is a very important task. The traditional way of exclusion is to identify and capture useful information by watching surveillance records through human eyes, but the current human resources are far from being able to meet the actual demand of the development rate. With the development of artificial intelligence technology, AI face recognition technology combined with surveillance camera technology has become an important means to obtain effective information and improve analysis efficiency for mega-city governance. Compared with ordinary videos, surveillance camera videos have low resolution, complex environment, light interference, small facial size and variable orientation[xxx], which makes face recognition in video surveillance a more challenging task.

Methods of face recognition in video surveillance can be broadly classified into two main categories: close-set and open-set[xxx]. Close-set method performs a classification task and it can only recognize the faces that appeared in the training set. The recognition process is to calculate the similarity between the detected faces and known faces one by one, and then select the highest similarity as the output result.[xxx] When a strange face appears, it will result in incorrect classifications. Therefore, this method can generally only be used when there are few people in the video and their identities have all been confirmed, and its universality and scalability are not strong. On the contrary, open-set method can not only recognize faces that have appeared, but also recognize strange faces. If the detected face is judged to be known, the recognition system should output the identity of it. Otherwise, the detected face should be simply ignored or be added into the dataset and become a new category.[xxx] Due to its ability to properly handle unknown faces, this method is more valuable in large scale practical application scenarios. (后面写一些其他人的具体方法)

.

## B. Face encryption and decryption algorithm

To be defined

## C.Brain-like memory cognition

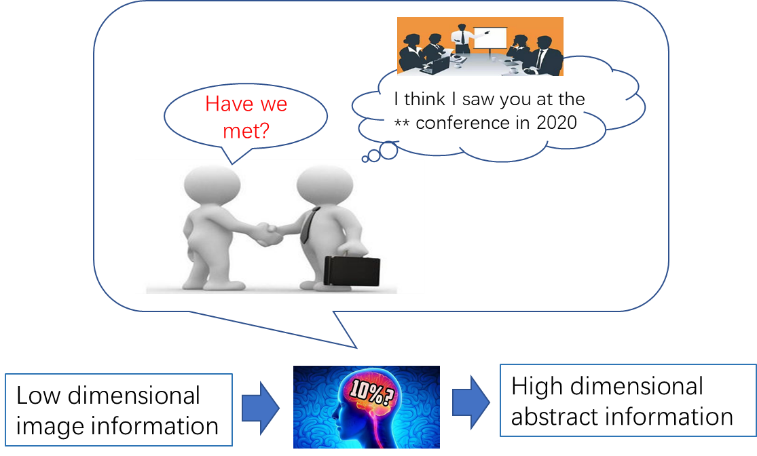
To be defined

# III. PROPOSED APPROACH

## We describe the human brain face perception-memory-association re-identification process through a typical scenario and outline the scientific problem to be solved in this paper. An algorithmic model for solving the problem and a framework for the solution process are then proposed.

## A. Problem description.

The process by which humans perceive faces and recognise their identity based on fuzzy impression memory associations is highly complex. Each of us sees many faces in life scenarios, such as in crowded crowds, at an academic conference we attend, at a dinner party where we meet, etc. However, not all information about faces is generally remembered. However, when meeting new people face-to-face, humans tend to look for déjà vu impressions in brain sessions. Let's start with a typical scenario to describe this process. As shown in Figure 1, we meet a lot of people at academic conferences but not everyone interacts with each other. When the two of you meet again sometime in the future, your mind will unconsciously recall that you have seen such a face at a certain time, place and event, and in combination with the real person in front of you, you seem to recall the scene and get a clearer picture of his 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: Humanlike abstract association based on perceptual feature trigger.**

The general expression of the model is as follows.

(1)

where A1 is the set of low-dimensional full-dimensional information about the face currently perceived by the brain, B1 is the set of high-dimensional abstract semantic features formed by the brain in the mind based on A1, A2p is the set of partially low-dimensional residual information from fuzzy memory in the brain, A2 is the set of low-dimensional full-dimensional information about the face currently perceived by the brain, A2p is a subset of A2, and B2 is the set of high-dimensional abstract semantic features formed by the brain in the mind based on A2 . The algorithm for solving the above expression is as follows.

|  |
| --- |
| **Algorithm: perceives- retrieves-associates-recalls** |
| Input: A1, A2p, B2  Output: A2  For A1 in Brain Do  Using cerebral neural network for high dimensional semantic abstraction  f1(Ai) →AiP, Bi, where A1P ⊊A1, i=1, 2, ……  For B1 in Brain Do  Find B2 in B1，  Meet the maximum face similarity feature  P(A2p| A1p)= f2(B2 ∩ B1 )→1  For A2p in Brain Do  f3(A1, A2p, B2) → A2  return A2 |

## This paper combines the above human-like perceptual associative memory algorithm with the face encryption and decryption requirements of surveillance video to solve the following problems.

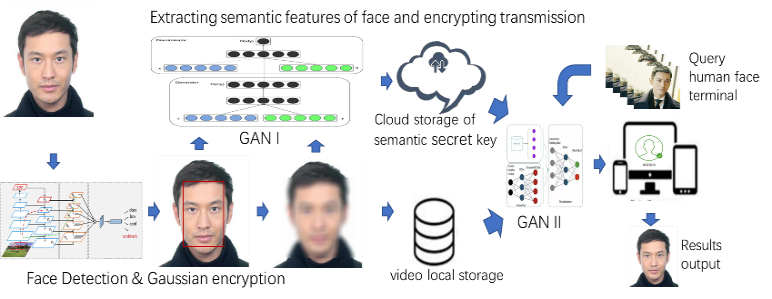
## 1) Drawing on the human-like cognitive mechanism, we model and solve the high-dimensional abstract memory and compressed perception process f1 function, and propose an artificial intelligence algorithm for solving Aip and Bi to identify and locate faces in videos, extract high-dimensional semantic features while encrypting video faces with reduced resolution.

## 2) Drawing on human-like 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 B1 and B2.

## 3) Drawing on the human-like perception-triggered recall mechanism, the associative recall of high-dimensional semantic features A1 and B2 low-resolution video is modelled to solve f3 for indexing location as well as high-resolution decryption.

## B. AI Methodologies

Inspired by human-like 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 figure below.

****

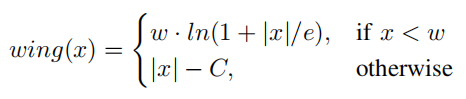
**Fig. 2. Framework of personal self-decryption.**

1. Encryption method

**问题：人脸定位的部分**

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.





After completing face target detection, the faces in the recognized frames are encrypted. We use the pixel2style2pixel (pSp) model to encrypt and encode the faces localized by YOLO5-face.The model expression for encrypting the face to obtain the high dimensional abstract semantics B1 is as follows:

)

Taking the previous original video A1 as input, L(\*) denotes the latent code of obtaining A1 to get the abstract semantic feature B1.E(\*) denotes the Encoder of the pSp model.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. Further, the abstract semantic feature B1 is used as the key to decode A1p and stored in the cloud keystore. The solving process of f1(Ai) is thus completed.

1. 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.

In order to equip the AI system with human-like high-dimensional abstract computation capability, the model in 1） is utilized to extract the high-dimensional abstract semantic feature B from the original video, and fuses the time sequence of the video frames, the latitude and longitude of the edge camera itself, and the pixel coordinates as the sign bit encoding of the high-dimensional abstract semantic feature, and associates the mapping relationship between the high-dimensional abstract semantic feature B and encrypted face image frames with the abovementioned sign bit encoding to facilitate the subsequent video decryption work. During the training process, the abstract semantic feature B1 of the video face dataset is extracted as the key for decoding A1p and stored in the cloud keystore.

The high-dimensional abstract semantic features learned by the above model are often not interpretable, so in-depth research on abstract semantics B1 and B2 is needed to study how to match face retrieval through abstract semantics B1 and B2. 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 B1 and B2 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 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 A1 by the encrypted video A1p, high-dimensional abstract semantics B1 and the face A2 in 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 generators. Firstly, through the method described in the previous section, the high-dimensional abstract semantics B2 is extracted from A2. Then, the similarity is calculated between B2 and all the high-dimensional abstract semantics B1i in the key pool corresponding to the encrypted video A1p, and the B1i 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 B1i is added to the generator with A1p as input as a constraint, and the decrypted video A1 is output.

# IV. EXPERIMENTS AND RESULTS

## Dataset

数据集部分暂时不知道怎么改，使用的是generated\_yellow-stylegan2

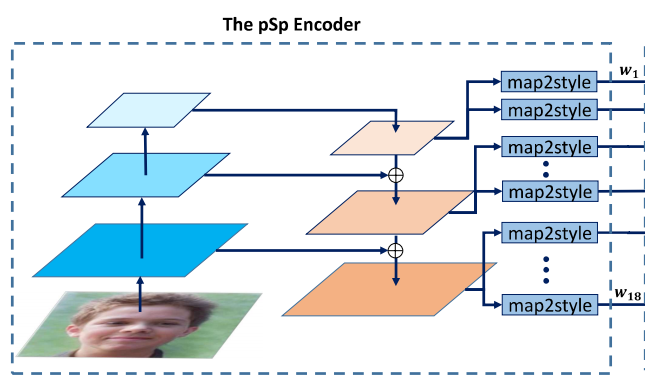
We acquired face images based on temporal head pose changes of experimental subjects of different genders and ages, and obtained a dataset with Gaussian encryption paired with the original images using a method based on YOLO5-face with Gaussian encryption. 1000 images were acquired for each person, for a total of 20 people with a total of 20,000 images, in order to exclude other background features in the images from the subsequent GAN network high-dimensional semantic abstraction In order to exclude other background features in the images from the subsequent GAN network high-dimensional semantic abstraction, all faces were collected in the same background for the face collection process. The figure below shows a portion of the extracted paired dataset. This data encryption process also validates the feasibility of the video encryption method.



## GANI Training and Encryption Process

（模型图的问题）

We use the pixel2style2pixel (pSp) model to complete the training of the above dataset,which uses the standard feature pyramid on top of ResNet to extract the feature mapping.For each of the 18 target styles, a small mapping net is trained to extract the learned styles from the corresponding feature maps.0-2 styles are generated from small feature mappings, 3-6 styles are generated from medium feature mappings,and 7-18 styles are generated from large feature mappings. Corresponding to the high-dimensional abstract memory and compressed perception processes of the human brain in humanoid cognitive mechanisms, high-dimensional feature information is generated from smaller feature mappings.



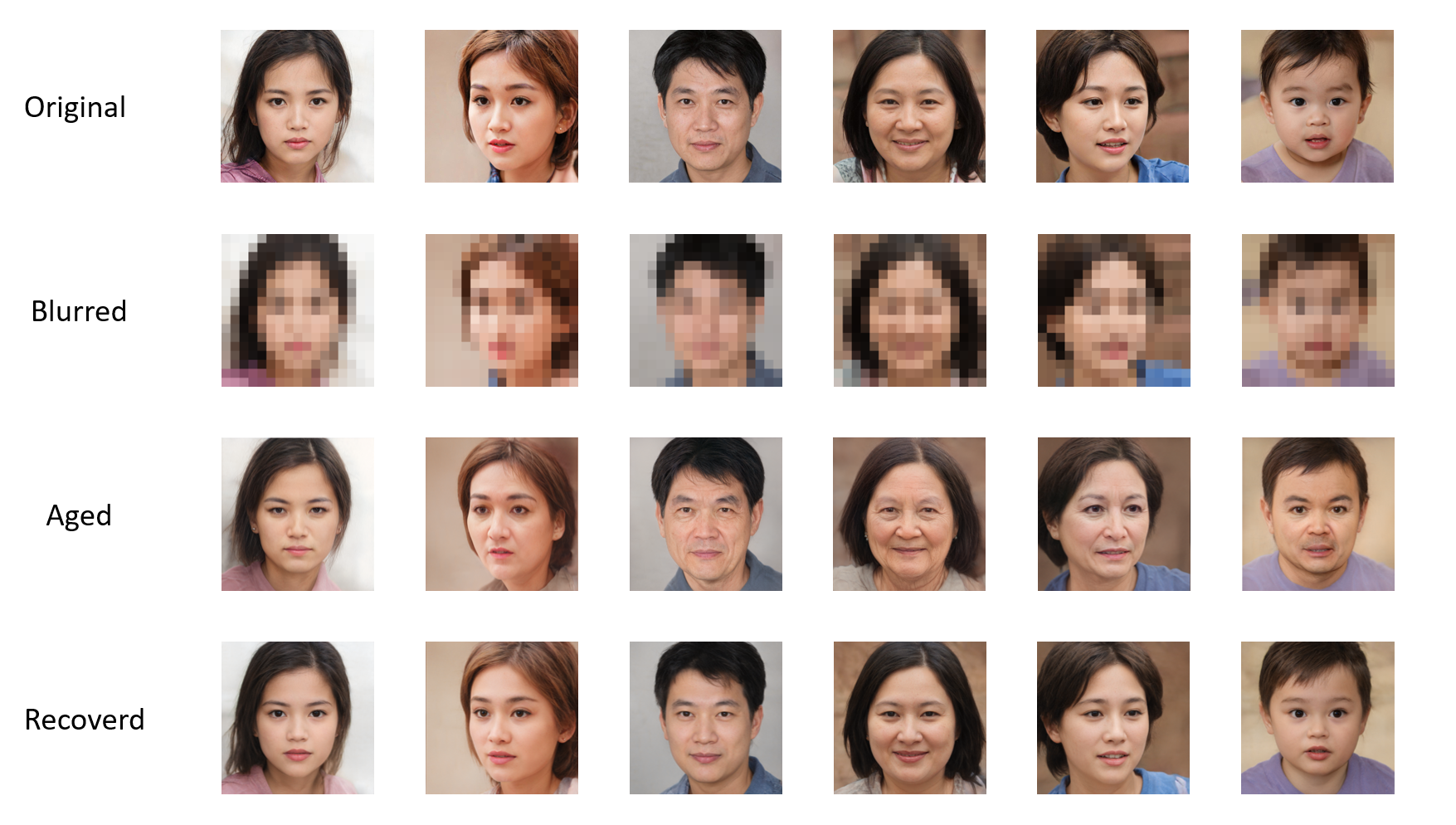
We first extract the original face P1 from the training set, use psp encoder to train P1, extract multi-dimensional face features from low-dimensional to high-dimensional, and integrate and encrypt this information to form a key, which is bound to the processed face identity ID, added to the face key pool, and stored in existing\_faces.pkl.

## C. GANI Decryption Process

For decryption, a newly captured face is input. The newly captured face P2 is encoded using the same coding model to extract features, and the extracted features are taken in certain dimensions and matched with the features in the face feature key pool. Based on the similarity of the face features determine whether it has appeared in the dataset. If it has not appeared, the decryption is rejected. If it has appeared, then confirm its identity ID and input it into the generator of styleGAN with the previously stored feature key, decrypt the face with the mosaic of the specified ID, and then output the decrypted face. styleGAN uses an unsupervised image super-resolution method to transform low-resolution images into high-quality, high-resolution images, thus reproducing the detailed features of the image such as skin color, eyes, lips, and so on.

## D. Experimental evaluation and detailed analysis

a) Encryption and decryption effects

The encryption and decryption effects are shown below.Original denotes the original face P1.Blurred denotes the encrypted face.Aged denotes the photo of Original after face aging treatment, most of the photos are set to be aged to 50 years old as the newly captured face P2 in this experiment, and Recoverd denotes the face after decryption under the guidance of the previously stored face feature key. 

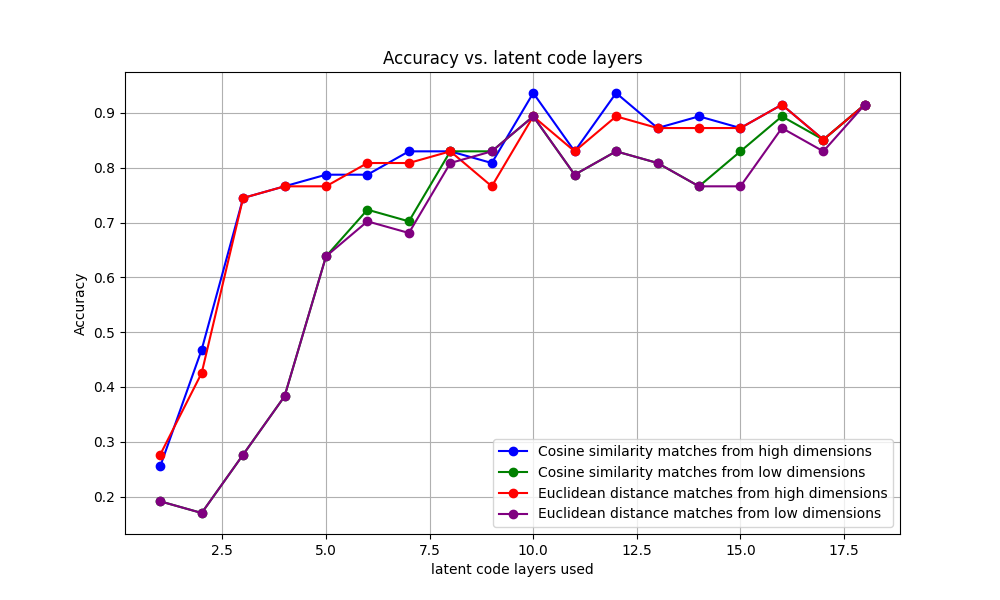
b) Feature Matching Methods

In the experiment, there are 47 photos that have been processed by face aging, and after extracting the features with the same model, certain dimensions are taken to match with the features in the face feature key pool, and the accuracy of matching is calculated.

Different dimension selections and different matching criteria are chosen in the experiment to fully explore the impact of different matching methods. The dimension selection combinations used are matching from high dimension:1, 1-2, 1-3, ... , 1-18; and from lower dimensions: 18, 17-18, 16-18, ... ,1-18.The matching criteria used are cosine similarity (taking the highest similarity) and Euler distance (taking the smallest distance). The matching accuracies of different matching methods are shown in the following table.

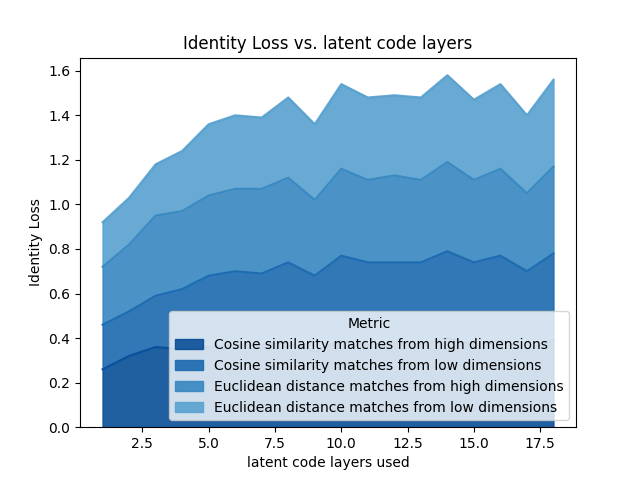
|  |  |  |
| --- | --- | --- |
| Table 1. Matching from high dimensions | | |
| latent code layers used | Accuracy | |
| Cosine similarity | Euclidean distance |
| 1-1 | 0.2553 | 0.2766 |
| 1-2 | 0.4681 | 0.4255 |
| 1-3 | 0.7447 | 0.7447 |
| 1-4 | 0.7660 | 0.7660 |
| 1-5 | 0.7872 | 0.7660 |
| 1-6 | 0.7872 | 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 |

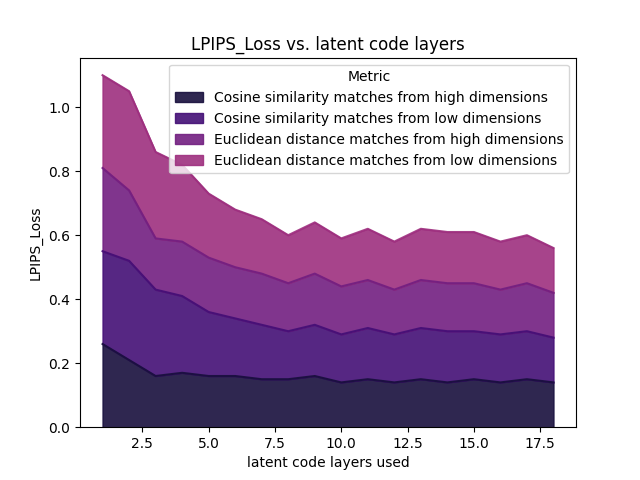
|  |  |  |
| --- | --- | --- |
| Table 2. Matching from low dimensions | | |
| latent code layers used | Accuracy | |
| Cosine similarity | Euclidean distance |
| 18-18 | 0.1915 | 0.1915 |
| 18-17 | 0.1702 | 0.1702 |
| 18-16 | 0.2766 | 0.2766 |
| 18-15 | 0.3830 | 0.3830 |
| 18-14 | 0.6383 | 0.6383 |
| 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 |

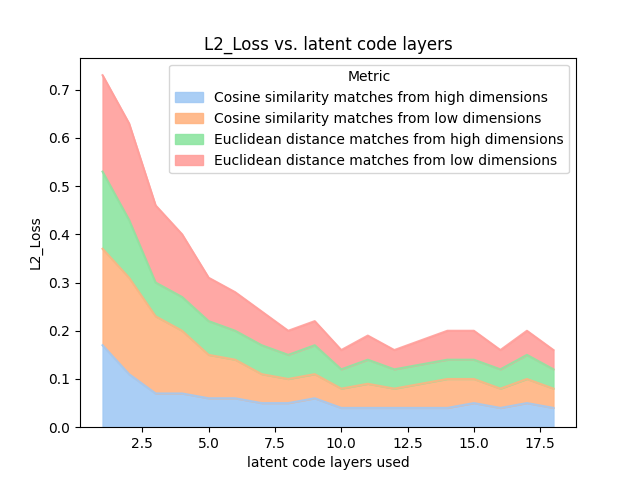


From the above figure, it can be seen that selecting the cosine similarity (taking the highest similarity) as the matching criterion and selecting some high-dimensional features as the basis of matching is more effective.

c) Multiple loss functions







# 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, the YOLO-based face recognition and face encryption algorithm is used to encrypt the video, while the high and low dimensional semantic information of the face is extracted to form a feature key, and index association is established with the face in the video and stored in the key pool corresponding to 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 the GAN model. We initially validate the feasibility of this video encryption and decryption algorithm on a self-built dataset. 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.

Appendix

Appendixes, if needed, appear before the acknowledgment.

Acknowledgment

The preferred here.

# 

# IEEE Guidelines and Policies

A full.

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, doi: 10.1109/JIOT.2019.2963701.
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, doi: 10.1109/JIOT.2018.2872606.
3. O. Styles, T. Guha and V. Sanchez, "Multi-Camera Trajectory Forecasting with Trajectory Tensors," IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3107958.
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, doi: 10.1109/JIOT.2020.3005727.

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

.

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

**Third C. Author, Jr.** (Member, IEEE), photograph and biography not available at the time of publication.

1. This work was supported in part by the U.S. Department of Commerce under Grant BS123456. *(Corresponding author: Bin. He).* Here you may also indicate if authors contributed equally or if there are co-first authors.

   The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, First A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

   Second B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

   Third C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba 305-0047, Japan (e-mail: author@nrim.go.jp).

   Mentions of supplemental materials and animal/human rights statements can be included here.

   Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org [↑](#footnote-ref-1)