Adversarial Level of Face Images Generated by Prompt-based Image Coding in Face Recognition System

Yurika Fujinami
School of FSE, Waseda University
Tokyo, Japan
ynami73.37imany@fuji.waseda.jp

Hiroshi Watanabe
School of FSE, Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—In this study, we investigate the adversarial level of face images generated by a prompt-based image coding. The adversarial level is the criterion by which the image produced by diffusion is judged to be consistent with the unprocessed original image. The Prompt-based image coding is designed to combine semantic compression and faithful image representation. The quality of the coded image can be controlled by adjusting the amount of edge information. Face recognition systems rely on patches of faces formed by vectors created from feature points with significant edge contributions. It is therefore worth investigating how much facial edge information should be retained in prompt-based image coding to fool face recognition systems. Experimental results show high possibility of falsification when coded images are fed into the face recognition model.**

Keywords—face recognition, prompt-based image coding, diffusion model, stable diffusion, canny edge detector

I. Introduction

Research and development of image generation methods based on diffusion models have been progressing rapidly. Text-to-image or prompt-to-image conversion is the main process. In particular, many models for generating face images have been studied, and variations of high-quality face images have been obtained. Especially, prompt-based image coding schemes that extract prompts, edges, and color information from images have excellent reproducibility of face images.

CNN-based face recognition system has achieved high recognition accuracy. It has been suggested to be applied to various applications as a social infrastructure. It should not be vulnerable to external attacks in the practical application stage [1],[2]. Therefore, an important issue to consider is what properties the face images created by image generation methods based on diffusion models have in a face recognition system. In this study, we examine the degree of similarity by which face images obtained by prompt-based image encoding are judged true or false when input to face image recognition system.

II. Related Work

Prompt-based image coding is a novel technique based on a diffusion model [3],[4]. It combines prompt extraction, edge, and color detection, and particular embedding optimization in the diffusion model for image compression. This innovation offers applications of pseudo-face generation generated from an input face image. Generated facial characteristics are strongly influenced by the model employed in the diffusion model. Furthermore, the shapes of facial parts are diversely generated when the constrained edge information is not enough. The amount of detected edge can be controlled by two thresholds of canny edge detection.

DeepFace is a state-of-the-art face recognition framework [5],[6]. It shows that human beings have 97.53% accuracy on facial recognition tasks. The facial attribute analysis covers age, gender, emotion and race. The face recognition models represent facial images as multi-dimensional vectors created by feature points such as start and end points of eyes, nose, and mouth. Textures may not be meaningful for a face recognition. Conversely, it can be said that facial edge information plays an important role in recognition.

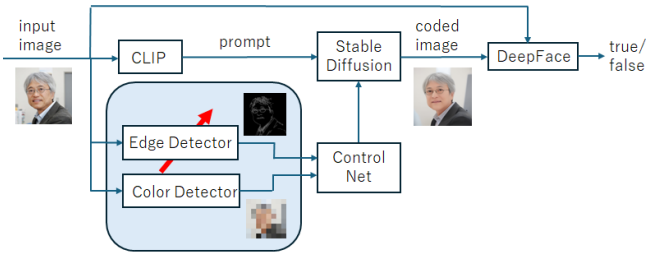


Fig. 1. Processing diagram using Stable Diffusion and DeepFace.

III. Edge Control in Stable Diffusion

Prompt-based image coding uses the Canny edge detector for its edge extraction. The Canny edge detector extracts edges after applying the Sobel operator, sets a first threshold, extracts edges above the threshold as major edges. Next, edges above the second threshold connected to the major edges are retained. Therefore, the larger the first threshold value, the smaller the number of edges to be extracted. When Canny edges are used in the ControlNet control model for Stable Diffusion, the edges constrain the boundaries of the generated image and preserve its shape. On the other hand, texture is generated by the diffusion model according to the conditions specified by the prompt and hyperparameter. Therefore, the number of edges is controlled by setting a threshold value. The color information and texture generation process are fixed.

We assume that a face image encoded based on a prompt is misidentified as a false face image in a face image recognition system. The fidelity of the encoded image from the original image changes depending on the number of edges. The face images with varying edge values are judged by DeepFace, which calculates the face recognition result as the degree of difference and judges it to be true if it is below a threshold value. DeepFace uses a constant threshold to make a decision, although it shows different evaluation values depending on the encoding process of the face image. We avoid the same threshold values as usual for true/false judgements and evaluate the degree of difference as a numerical value. The process is shown in Fig.1.

IV. Experiment

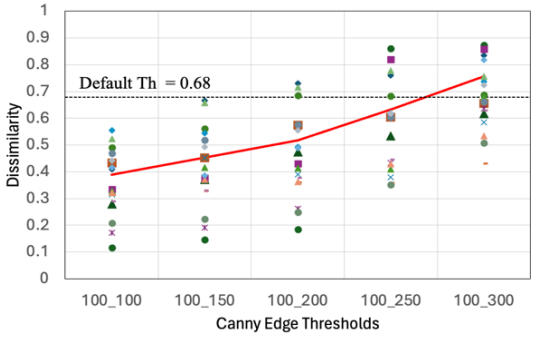
We use the state-of-the-art DeepFace model for our face image recognition system in our experiments. Three datasets of 20 images each are used. (A) Male and female faces dataset

979-8-3503-5506-2/24/$31.00 ©2024 IEEE

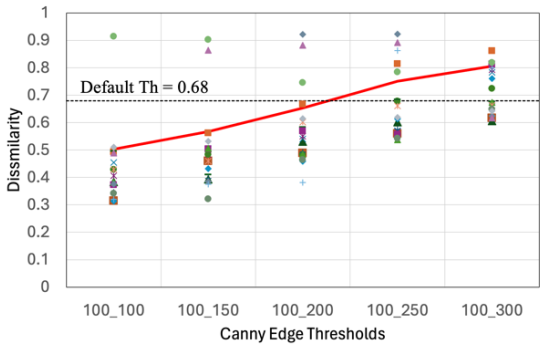
[7], (B) DeepFace dataset [5], and (C) Face Research Lab London set [8].

First, five images with different quality are generated from the original image using Stable Diffusion. At this stage, the color and contour are specified by ControlNet so that something similar to the original image are generated. When generating the images, Canny Edge detector is used for edge extraction and the value of threshold is changed.

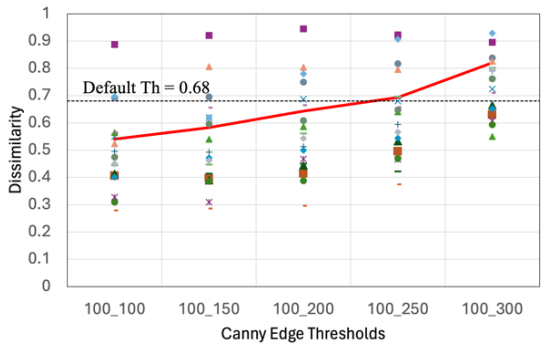
The similarity between the generated image and the original image is measured by using DeepFace. Authentication is done with VGG-Face, detection with opencv, and distance with cosine. Dissimilarity is expressed as (*1.0 - cosine similarity)*, and with the value closer to 0 indicating that images are more similar. The distribution of dissimilarity value is shown in Fig.2. They tend to decrease as the number of edges increase in any dataset. Red lines indicate the average values of dissimilarity at each set of thresholds. Some dissimilarity values are close to 0.9, where many edges exist in the background or areas outside of faces.



(A) Male and female faces dataset [7].



(B) DeepFace [5].



(C) Face Research Lab London set [8].

Fig. 2. Dissimilarities of coded images for each data set.

TABLE I TRUE RATE BY DEEPFACE FOR DIFFERENT EDGE IMAGES

Dataset	True rate (%)				
	100_100	100-150	100_200	100_250	100_300
A	1.00	1.00	0.85	0.75	0.55
B	0.95	0.90	0.80	0.60	0.50
C	0.85	0.85	0.75	0.70	0.35

(a) original (b) 100_100 Dissimilarity: 0.43 (c) 100_200 Dissimilarity: 0.60 (d) 100_300 Dissimilarity: 0.66

Fig. 3. Examples of original and coded images with Canny thresholds. All coded images are authorized for research purposes.

The true rates for five levels of encoded images with DeepFace at a threshold of 0.68 are shown in Table I. Images in dataset (A) at thresholds 100_100 and 100_150 reach 100% true, indicating that the coded faces are recognized as original.

The coded images in Fig. 3 illustrates the transition from natural to artificial faces at thresholds 100_100, 100_200, and 100_300. Lower thresholds are crucial to avoid falsification with DeepFace. However, if prompt-based image coding is treated as regular lossy coding, the images can be used for personal identification.

Higher accuracy increases the risk of the system mistaking fake images for the actual person. It is vital to determine the amount of edge data reduction needed for errors to occur, thereby enhancing security.

V. CONCLUSION

This paper demonstrates the adversarial potential of prompt-based image coding. Experiments show that DeepFace recognizes faces as real unless the number of edges is significantly reduced. Therefore, DeepFace should not be used with the default dissimilarity value if the generated image is to be rejected.

REFERENCES

[1] F. Vakhshiteh, A. Nickabadi, R. Ramchandra, "Adversarial Attacks against Face Recognition: A Comprehensive Study," arXiv: 2007.11709v3, Feb. 2021.

[2] Y. Xu, K. Raja, R. Ramchandra, C. Busch, "Adversarial Attacks on Face Recognition Systems," Springer, Handbook of Digital Face Manipulation and Detection, pp.139-161, Jan. 2022.

[3] H. Watanabe, L. Jin, T. Hayami, T. Chujoh, T. Aono, Y. Yasugi, S. Hong, Z. Fan, and T. Ikai, "Prompt-based Image Coding with Edge Information," 2023 Picture Coding Symposium, Image Media Symposium (PCSJ/IMPS2023), P1-12, Nov. 2023

[4] H. Watanabe, L Jin, T. Hayami, T. Chujoh, Y. Yasugi, S. Hong, Z. Fan, T. Ikai, "Prompt-based Image Coding with Edge and Color Information," IEICE General Conference, D-11A-27, Mar. 2024.

[5] S. Serengil and A. Özpınar, "LightFace: A Hybrid DeepFace Recognition Framework," IEEE Innovations in Intelligent Systems and Application Conference (ASYU) 2020, pp.1-5, Oct 2020.

[6] S. Serengil and A. Özpınar, "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules," Journal of information Technologirs, Vol.17, No.2, pp.95-107, Mar. 2024.

[7] Ashwin Gupta, Male and femel faces dataset, https://www.kaggle.com/datasets/ashwingupta3012/male-and-female-faces-dataset

[8] Lisa DeBruine, Benedict Jones, Face Research Lab London set, https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666