


Generative AI
Semester Project Presentation

Dimas Christos

Electrical and Computer Engineering
Technical University of Crete

July 1, 2025


Contents

Introduction

Introduction

About this presentation

This presentation is a semester project for the course *Generative AI* and aims to explain the ***Adv-CPG: A Customized Portrait Generation Framework with Facial Adversarial Attacks*** paper, a novel approach to generating personalized portraits while protecting user privacy, published by Junying Wang, Hongyuan Zhang, and Yuan Yuan in 2025.

Key features:

- Customized Portrait Generation (CPG) creates personalized, high-fidelity portraits form a facial image and text prompts.
- Current CPG methods lack effective protection against malicious face recognition (FR) systems.
- Adv-CPG integrates facial adversarial attacks into CPG to protect user privacy.
- It generates visually consistent portraits that deceive FR systems while allowing fine-grained customization.
- Key innovation: Progressive two-layer privacy protection combined with multi-modal image customization.

Concept 1: Customized Portrait Generation (CPG)

What is CPG?

Definition

Customized Portrait Generation, or **CPG**, is a cutting-edge AI technique that creates personalized portrait images based on a user's photo and a descriptive text prompt.

What is CPG?

Definition

Customized Portrait Generation, or **CPG**, is a cutting-edge AI technique that creates personalized portrait images based on a user's photo and a descriptive text prompt.

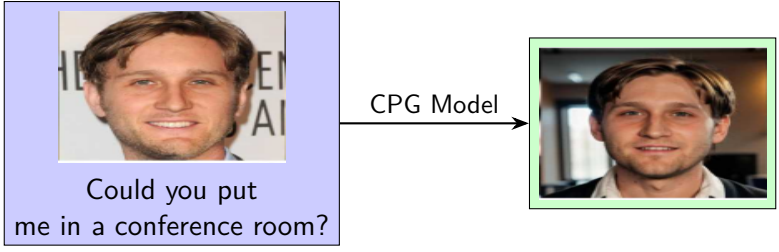


Figure: Customized Portrait Generation: From prompt to result. (Images from the paper)

CPG Types

Depending on whether or not fine-tuning is performed during testing, mainstream CPG methods can be divided into two categories:

CPG Types

Depending on whether or not fine-tuning is performed during testing, mainstream CPG methods can be divided into two categories:

- **Early methods**: require resource-intensive backpropagation, with each fine- tuning taking tens of minutes or hours to complete.

CPG Types

Depending on whether or not fine-tuning is performed during testing, mainstream CPG methods can be divided into two categories:

- **Early methods**: require resource-intensive backpropagation, with each fine- tuning taking tens of minutes or hours to complete.
- **Recent methods**: do not require fine-tuning, allowing for real-time generation of customized portraits.

CPG Types

Depending on whether or not fine-tuning is performed during testing, mainstream CPG methods can be divided into two categories:

- **Early methods**: require resource-intensive backpropagation, with each fine- tuning taking tens of minutes or hours to complete.
- **Recent methods**: do not require fine-tuning, allowing for real-time generation of customized portraits.

More specifically, the recent methods of CPG enable large-scale, high-fidelity, and flexible portrait generation for social media, art, and entertainment.

The Problem: A Privacy Risk

Problem Definition

Even though the exciting potential of CPG methods, they also pose significant privacy risks. Specifically, the high-fidelity portraits generated by CPG can be collected by malicious *face recognition (FR)* systems. Some of the most common risks include:

Problem Definition

Even though the exciting potential of CPG methods, they also pose significant privacy risks. Specifically, the high-fidelity portraits generated by CPG can be collected by malicious *face recognition (FR)* systems. Some of the most common risks include:

- These images can be used for unauthorized tracking, monitoring and profiling.

Problem Definition

Even though the exciting potential of CPG methods, they also pose significant privacy risks. Specifically, the high-fidelity portraits generated by CPG can be collected by malicious *face recognition (FR)* systems. Some of the most common risks include:

- These images can be used for unauthorized tracking, monitoring and profiling.
- Unlike a password, facial data cannot be changed if compromised.

Problem Definition

Even though the exciting potential of CPG methods, they also pose significant privacy risks. Specifically, the high-fidelity portraits generated by CPG can be collected by malicious *face recognition (FR)* systems. Some of the most common risks include:

- These images can be used for unauthorized tracking, monitoring and profiling.
- Unlike a password, facial data cannot be changed if compromised.
- Data breaches involving facial data can lead to identity theft, stalking, and harassment.

Problem Definition

Even though the exciting potential of CPG methods, they also pose significant privacy risks. Specifically, the high-fidelity portraits generated by CPG can be collected by malicious *face recognition (FR)* systems. Some of the most common risks include:

- These images can be used for unauthorized tracking, monitoring and profiling.
- Unlike a password, facial data cannot be changed if compromised.
- Data breaches involving facial data can lead to identity theft, stalking, and harassment.

The main reason why such things can happen is because current CPG methods lack built-in security mechanisms to protect users.

Concept 2: Adversarial Attacks

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

There are two types of adversarial attacks on face recognition:

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

There are two types of adversarial attacks on face recognition:

- Noise-based approaches exploit adding λ_θ boundary perturbations directly in the pixel space, which struggles with black-box transferability and stealthiness.

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

There are two types of adversarial attacks on face recognition:

- Noise-based approaches exploit adding λ_θ boundary perturbations directly in the pixel space, which struggles with black-box transferability and stealthiness.
- Recent unrestricted methods concentrate on learning perturbations in the semantic space, further improving transferability and stealthiness

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

1. For face recognition (FR), an adversarial attack subtly alters an image so an FR system identifies it as a different person (a "target identity") or fails to identify it at all.

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

1. For face recognition (FR), an adversarial attack subtly alters an image so an FR system identifies it as a different person (a "target identity") or fails to identify it at all.
2. These alterations are often imperceptible to the human eye, maintaining "stealthiness."

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

1. For face recognition (FR), an adversarial attack subtly alters an image so an FR system identifies it as a different person (a "target identity") or fails to identify it at all.
2. These alterations are often imperceptible to the human eye, maintaining "stealthiness."
3. Previous methods either add noise (less stealthy) or modify attributes like makeup (limited customization).

Adversarial Attacks

Definition

Adversarial attacks are techniques used to manipulate machine learning models by introducing small, imperceptible perturbations to input data, causing the model to make incorrect predictions or classifications.

Idea: Use adversarial attacks to protect privacy

The Goal of Adv-CPG: The Best of Both Worlds

The Goal of Adv-CPG: The Best of Both Worlds

What is Adv-CPG?

Adv-CPG is a novel approach that combines the strengths of CPG and adversarial attacks to generate personalized portraits while protecting user privacy.

The Goal of Adv-CPG: The Best of Both Worlds

What is Adv-CPG?

Adv-CPG is a novel approach that combines the strengths of CPG and adversarial attacks to generate personalized portraits while protecting user privacy.

It aims to generate a single image that is simultaneously:

- A high-fidelity, fine-grained, and personalized portrait.

The Goal of Adv-CPG: The Best of Both Worlds

What is Adv-CPG?

Adv-CPG is a novel approach that combines the strengths of CPG and adversarial attacks to generate personalized portraits while protecting user privacy.

It aims to generate a single image that is simultaneously:

- A high-fidelity, fine-grained, and personalized portrait.
- A stealthy and effective adversarial attack that deceives Face Recognition (FR) systems.

The Goal of Adv-CPG: The Best of Both Worlds

What is Adv-CPG?

Adv-CPG is a novel approach that combines the strengths of CPG and adversarial attacks to generate personalized portraits while protecting user privacy.

It aims to generate a single image that is simultaneously:

- A high-fidelity, fine-grained, and personalized portrait.
- A stealthy and effective adversarial attack that deceives Face Recognition (FR) systems.

Goal: Give users the creative freedom of CPG without sacrificing their facial privacy.

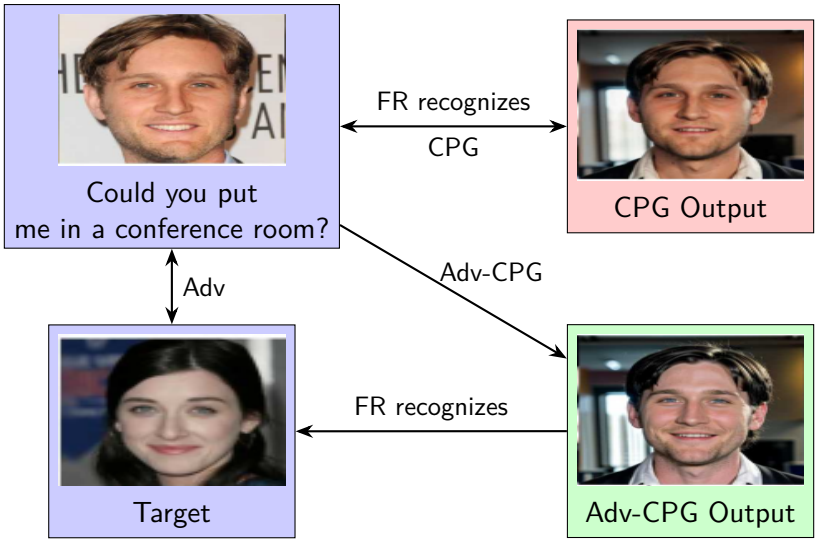


Figure: Adv-CPG: Combining CPG and Adversarial Attacks. (Images from the paper)

How It Works: The Adv-CPG Pipeline

How It Works: The Adv-CPG Pipeline

Now that we established the "what" and "why" of Adv-CPG, it's time to dive into the "how".

Adv-CPG is built upon a pre-trained diffusion model (SDXL) and features three core modules:

- **Multi-Modal Image Customizer (MMIC)**: Responsible for fine-grained portrait generation and customization using text prompts and image features.

How It Works: The Adv-CPG Pipeline

Now that we established the "what" and "why" of Adv-CPG, it's time to dive into the "how".

Adv-CPG is built upon a pre-trained diffusion model (SDXL) and features three core modules:

- **Multi-Modal Image Customizer (MMIC)**: Responsible for fine-grained portrait generation and customization using text prompts and image features.
- **ID Encryptor (En 1)**: Injects the target identity directly into the diffusion model's identity-dependent layers for global privacy protection.

How It Works: The Adv-CPG Pipeline

Now that we established the "what" and "why" of Adv-CPG, it's time to dive into the "how".

Adv-CPG is built upon a pre-trained diffusion model (SDXL) and features three core modules:

- **Multi-Modal Image Customizer (MMIC)**: Responsible for fine-grained portrait generation and customization using text prompts and image features.
- **ID Encryptor (En 1)**: Injects the target identity directly into the diffusion model's identity-dependent layers for global privacy protection.
- **Encryption Enhancer (En 2)**: Adds additional identity guidance during denoising steps for enhanced local protection.

Generation Process

The generator process is divided into two steps, based on MMIC's operation:

- **Stage 1: Progressive Facial Privacy Protection**: *En 1* and *En 2* primary work with the initial text prompt to embed the adversarial target identity and introduce the background.

Generation Process

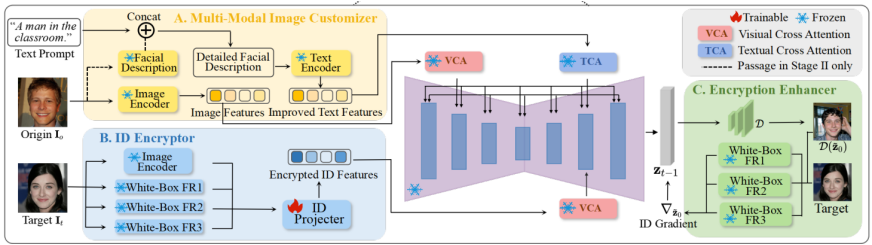
The generator process is divided into two steps, based on MMIC's operation:

- **Stage 1: Progressive Facial Privacy Protection**: *En 1* and *En 2* primary work with the initial text prompt to embed the adversarial target identity and introduce the background.
- **Stage 2: Fine-Grained Custommized Portrait Generation**: *MMIC* bacomes active, using detailed facial descriptions to refine the portrait, while *En 1* and *En 2* continue to maintain privacy protection.

Generation Process

The generator process is divided into two steps, based on MMIC's operation:

- **Stage 1: Progressive Facial Privacy Protection**: *En 1* and *En 2* primary work with the initial text prompt to embed the adversarial target identity and introduce the background.
- **Stage 2: Fine-Grained Custommized Portrait Generation**: *MMIC* bacomes active, using detailed facial descriptions to refine the portrait, while *En 1* and *En 2* continue to maintain privacy protection.



How It's Trained and Generated

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:

2. Inference Phase:

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)

2. Inference Phase:

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.

2. Inference Phase:

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.
   - Only parameters of the ID projector in En 1 will be optimized, all other are frozen.

2. Inference Phase:

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.
   - Only parameters of the ID projector in En 1 will be optimized, all other are frozen.
   - Loss function minimizes the difference between predicted and true noise in the diffusion denoising process, conditioned on encrypted ID, image, and text features.
2. Inference Phase:

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.
   - Only parameters of the ID projector in En 1 will be optimized, all other are frozen.
   - Loss function minimizes the difference between predicted and true noise in the diffusion denoising process, conditioned on encrypted ID, image, and text features.
2. Inference Phase:
   - All of the modules are involved (MMIC, En 1, and En 2).

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.
   - Only parameters of the ID projector in En 1 will be optimized, all other are frozen.
   - Loss function minimizes the difference between predicted and true noise in the diffusion denoising process, conditioned on encrypted ID, image, and text features.
2. Inference Phase:
   - All of the modules are involved (MMIC, En 1, and En 2).
   - Uses the two-stage generation process described earlier (see slide "Generation Process p*18*").

Training and Generation Process

In order to achieve the model's goals, two phases were introduce and it is really important to analyze them:

1. Training Phase:
   - Adv-CPG is built on the pre-trained SDXL model (SOTA T2I diffusion model)
   - Only MMIC and En 1 particate.
   - Only parameters of the ID projector in En 1 will be optimized, all other are frozen.
   - Loss function minimizes the difference between predicted and true noise in the diffusion denoising process, conditioned on encrypted ID, image, and text features.
2. Inference Phase:
   - All of the modules are involved (MMIC, En 1, and En 2).
   - Uses the two-stage generation process described earlier (see slide "Generation Process p*18*").
   - Switches from original text features to augmented text features after a certain diffusion timestep to balance privacy and customization.

Results (Part 1): It's an Effective Attack

Results (Part 1): It's an Effective Attack

Adversarial Attack Effectiveness

Adv-CPG successfully generates portraits that:

- Maintain high visual fidelity and customization.

Results (Part 1): It's an Effective Attack

Adversarial Attack Effectiveness

Adv-CPG successfully generates portraits that:

- Maintain high visual fidelity and customization.
- Achieves state-of-the-art attack success rates (ASR) against multiple face recognition (FR) models in black-box settings.

Results (Part 1): It's an Effective Attack

Adversarial Attack Effectiveness

Adv-CPG successfully generates portraits that:

- Maintain high visual fidelity and customization.
- Achieves state-of-the-art attack success rates (ASR) against multiple face recognition (FR) models in black-box settings.
- Provide a balance between privacy protection and personalization.

Results (Part 1): It's an Effective Attack

Adversarial Attack Effectiveness

Adv-CPG successfully generates portraits that:

- Maintain high visual fidelity and customization.
- Achieves state-of-the-art attack success rates (ASR) against multiple face recognition (FR) models in black-box settings.
- Provide a balance between privacy protection and personalization.

The following slides will showcase the effectiveness of Adv-CPG in generating adversarial portraits.

Face Verification

Method	Dataset ASR% on FR Model ↑	FFHQ IR152	FFHQ IRSE50	FFHQ FaceNet	FFHQ MF	CelebA-HQ IR152	CelebA-HQ IRSE50	CelebA-HQ FaceNet	CelebA-HQ MF	Average
Clean	—	3.14	4.73	1.49	8.96	4.64	5.74	1.06	13.27	5.38
Noise-Based	FGSM (ICLR'15) [9]	9.86	48.53	4.23	51.48	12.07	45.81	1.35	53.02	28.29
	MI-FGSM (CVPR'18) [6]	46.31	69.24	20.62	69.26	46.53	70.57	27.09	58.94	51.07
	PGD (ICLR'18) [33]	31.64	75.38	18.59	63.12	41.87	63.23	19.62	57.34	46.35
	TI-DIM (CVPR'19) [7]	43.57	65.89	14.72	53.31	35.13	62.38	13.68	52.84	42.69
	TIP-IM (ICCV'21) [54]	46.25	67.38	59.82	52.03	41.26	57.29	39.07	49.56	51.58
Makeup-Based	Adv-Hat (ICPR'21) [25]	13.77	15.36	5.26	9.83	5.04	16.88	4.91	12.64	10.46
	Adv-Makeup (IJCAI'21) [56]	10.03	25.57	1.08	20.38	12.68	19.95	1.37	22.11	14.15
	AMT-GAN (CVPR'22) [17]	11.52	56.02	9.74	41.32	12.09	53.26	4.87	47.95	29.60
	Clip2Protect (CVPR'23) [43]	52.12	86.53	45.01	76.29	47.63	80.96	42.57	73.64	63.09
	GIFT (ACMMM'24) [27]	69.72	87.64	54.49	82.93	73.84	83.72	56.48	86.37	74.40
	DFPP (arXiv'24) [44]	54.25	90.63	52.13	80.09	46.38	80.59	45.37	72.13	69.20
	DiffAM (CVPR'24) [47]	67.24	90.30	**64.96**	89.56	65.09	**89.66**	62.99	84.51	76.79
Facial Semantic Invariant	DiffProtect (arXiv'23) [31]	57.62	60.14	49.38	67.52	58.64	79.34	24.69	75.91	59.16
	DPG (arXiv'24) [60]	34.87	76.82	36.57	69.03	42.89	62.47	35.83	66.42	53.11
	SD4Privacy (ICME'24) [1]	51.31	79.94	43.57	71.55	66.89	79.96	53.49	74.58	65.16
	Adv-Diffusion(AAAI'24) [29]	50.93	81.76	30.84	67.52	52.84	81.67	34.95	70.78	58.91
	P3-Mask (ECCV'25) [3]	73.18	85.35	57.92	70.69	73.47	83.40	60.24	69.64	71.74
Portraits	Adv-CPG (Ours)	**75.26**↑2.1	**91.03**↑0.4	63.84↓1.1	**89.94**↑0.4	**76.96**↑3.1	88.72↓0.9	**63.50**↑0.5	**87.95**↑1.6	**79.65**↑2.9

Figure: Face Verification Attack Success Rate (ASR) of Adv-CPG compared to other adversarial attacks. (Image from the paper Table 3)

Face Verification

Method	Dataset	FFHQ				CelebA-HQ				Average
	ASR% on FR Model ↑	IR152	IRSE50	FaceNet	MF	IR152	IRSE50	FaceNet	MF	
Clean	——	3.14	4.73	1.49	8.96	4.64	5.74	1.06	13.27	5.38
Noise-Based	FGSM (ICLR'15) [9]	9.86	48.53	4.23	51.48	12.07	45.81	1.35	53.02	28.29
	MI-FGSM (CVPR'18) [6]	46.31	69.24	20.62	69.26	46.53	70.57	27.09	58.94	51.07
	PGD (ICLR'18) [33]	31.64	75.38	18.59	63.12	41.87	63.23	19.62	57.34	46.35
	TI-DIM (CVPR'19) [7]	43.57	65.89	14.72	53.31	35.13	62.38	13.68	52.84	42.69
	TIP-IM (ICCV'21) [54]	46.25	67.38	59.82	52.03	41.26	57.29	39.07	49.56	51.58
Makeup-Based	Adv-Hat (ICPR'21) [25]	13.77	15.36	5.26	9.83	5.04	16.88	4.91	12.64	10.46
	Adv-Makeup (IJCAI'21) [56]	10.03	25.57	1.08	20.38	12.68	19.95	1.37	22.11	14.15
	AMT-GAN (CVPR'22) [17]	11.52	56.02	9.74	41.32	12.09	53.26	4.87	47.95	29.60
	Clip2Protect (CVPR'23) [43]	52.12	86.53	45.01	76.29	47.63	80.96	42.57	73.64	63.09
	GIFT (ACMMM'24) [27]	69.72	87.64	54.49	82.93	73.84	83.72	56.48	86.37	74.40
	DFPP (arXiv'24) [44]	54.25	90.63	52.13	80.09	46.38	80.59	45.37	72.13	69.20
	DiffAM (CVPR'24) [47]	67.24	90.30	**64.96**	89.56	65.09	**89.66**	62.99	84.51	76.79
Facial Semantic Invariant	DiffProtect (arXiv'23) [31]	57.62	60.14	49.38	67.52	58.64	79.34	24.69	75.91	59.16
	DPG (arXiv'24) [60]	34.87	76.82	36.57	69.03	42.89	62.47	35.83	66.42	53.11
	SD4Privacy (ICME'24) [1]	51.31	79.94	43.57	71.55	66.89	79.96	53.49	74.58	65.16
	Adv-Diffusion(AAAI'24) [29]	50.93	81.76	30.84	67.52	52.84	81.67	34.95	70.78	58.91
	P3-Mask (ECCV'25) [3]	73.18	85.35	57.92	70.69	73.47	83.40	60.24	69.64	71.74
Portraits	Adv-CPG (Ours)	**75.26**↑2.1	**91.03**↑0.4	63.84↓1.1	**89.94**↑0.4	**76.96**↑3.1	88.72↓0.9	**63.50**↑0.5	**87.95**↑1.6	**79.65**↑2.9

Average ASR is 28.1% higher than the best noise-based attacks and 2.86% higher than the best unrestricted attacks.

Face Verification

Method	Dataset ASR% on FR Model ↑	FFHQ IR152	 IRSE50	 FaceNet	 MF	CelebA-HQ IR152	 IRSE50	 FaceNet	 MF	Average
Clean	—	3.14	4.73	1.49	8.96	4.64	5.74	1.06	13.27	5.38
Noise-Based	FGSM (ICLR'15) [9]	9.86	48.53	4.23	51.48	12.07	45.81	1.35	53.02	28.29
	MI-FGSM (CVPR'18) [6]	46.31	69.24	20.62	69.26	46.53	70.57	27.09	58.94	51.07
	PGD (ICLR'18) [33]	31.64	75.38	18.59	63.12	41.87	63.23	19.62	57.34	46.35
	TI-DIM (CVPR'19) [7]	43.57	65.89	14.72	53.31	35.13	62.38	13.68	52.84	42.69
	TIP-IM (ICCV'21) [54]	46.25	67.38	59.82	52.03	41.26	57.29	39.07	49.56	51.58
Makeup-Based	Adv-Hat (ICPR'21) [25]	13.77	15.36	5.26	9.83	5.04	16.88	4.91	12.64	10.46
	Adv-Makeup (IJCAI'21) [56]	10.03	25.57	1.08	20.38	12.68	19.95	1.37	22.11	14.15
	AMT-GAN (CVPR'22) [17]	11.52	56.02	9.74	41.32	12.09	53.26	4.87	47.95	29.60
	Clip2Protect (CVPR'23) [43]	52.12	86.53	45.01	76.29	47.63	80.96	42.57	73.64	63.09
	GIFT (ACMMM'24) [27]	69.72	87.64	54.49	82.93	73.84	83.72	56.48	86.37	74.40
	DFPP (arXiv'24) [44]	54.25	90.63	52.13	80.09	46.38	80.59	45.37	72.13	69.20
	DiffAM (CVPR'24) [47]	67.24	90.30	**64.96**	89.56	65.09	**89.66**	62.99	84.51	76.79
Facial Semantic Invariant	DiffProtect (arXiv'23) [31]	57.62	60.14	49.38	67.52	58.64	79.34	24.69	75.91	59.16
	DPG (arXiv'24) [60]	34.87	76.82	36.57	69.03	42.89	62.47	35.83	66.42	53.11
	SD4Privacy (ICME'24) [1]	51.31	79.94	43.57	71.55	66.89	79.96	53.49	74.58	65.16
	Adv-Diffusion(AAAI'24) [29]	50.93	81.76	30.84	67.52	52.84	81.67	34.95	70.78	58.91
	P3-Mask (ECCV'25) [3]	73.18	85.35	57.92	70.69	73.47	83.40	60.24	69.64	71.74
Portraits	Adv-CPG (Ours)	**75.26**↑2.1	**91.03**↑0.4	63.84↓1.1	**89.94**↑0.4	**76.96**↑3.1	88.72↓0.9	**63.50**↑0.5	**87.95**↑1.6	**79.65**↑2.9

Demonstrates strong transferability across diverse FR models including IR152, IRSE50, FaceNet, and MobileFace.

Face Verification

Method	Dataset	FFHQ				CelebA-HQ				Average
	ASR% on FR Model ↑	IR152	IRSE50	FaceNet	MF	IR152	IRSE50	FaceNet	MF	
Clean	——	3.14	4.73	1.49	8.96	4.64	5.74	1.06	13.27	5.38
Noise-Based	FGSM (ICLR'15) [9]	9.86	48.53	4.23	51.48	12.07	45.81	1.35	53.02	28.29
	MI-FGSM (CVPR'18) [6]	46.31	69.24	20.62	69.26	46.53	70.57	27.09	58.94	51.07
	PGD (ICLR'18) [33]	31.64	75.38	18.59	63.12	41.87	63.23	19.62	57.34	46.35
	TI-DIM (CVPR'19) [7]	43.57	65.89	14.72	53.31	35.13	62.38	13.68	52.84	42.69
	TIP-IM (ICCV'21) [54]	46.25	67.38	59.82	52.03	41.26	57.29	39.07	49.56	51.58
Makeup-Based	Adv-Hat (ICPR'21) [25]	13.77	15.36	5.26	9.83	5.04	16.88	4.91	12.64	10.46
	Adv-Makeup (IJCAI'21) [56]	10.03	25.57	1.08	20.38	12.68	19.95	1.37	22.11	14.15
	AMT-GAN (CVPR'22) [17]	11.52	56.02	9.74	41.32	12.09	53.26	4.87	47.95	29.60
	Clip2Protect (CVPR'23) [43]	52.12	86.53	45.01	76.29	47.63	80.96	42.57	73.64	63.09
	GIFT (ACMMM'24) [27]	69.72	87.64	54.49	82.93	73.84	83.72	56.48	86.37	74.40
	DFPP (arXiv'24) [44]	54.25	90.63	52.13	80.09	46.38	80.59	45.37	72.13	69.20
	DiffAM (CVPR'24) [47]	67.24	90.30	**64.96**	89.56	65.09	**89.66**	62.99	84.51	76.79
Facial Semantic Invariant	DiffProtect (arXiv'23) [31]	57.62	60.14	49.38	67.52	58.64	79.34	24.69	75.91	59.16
	DPG (arXiv'24) [60]	34.87	76.82	36.57	69.03	42.89	62.47	35.83	66.42	53.11
	SD4Privacy (ICME'24) [1]	51.31	79.94	43.57	71.55	66.89	79.96	53.49	74.58	65.16
	Adv-Diffusion(AAAI'24) [29]	50.93	81.76	30.84	67.52	52.84	81.67	34.95	70.78	58.91
	P3-Mask (ECCV'25) [3]	73.18	85.35	57.92	70.69	73.47	83.40	60.24	69.64	71.74
Portraits	Adv-CPG (Ours)	**75.26**↑2.1	**91.03**↑0.4	63.84↓1.1	**89.94**↑0.4	**76.96**↑3.1	88.72↓0.9	**63.50**↑0.5	**87.95**↑1.6	**79.65**↑2.9

Outperforms 17 benchmark methods on FFHQ and CelebA-HQ datasets.

Face Identification

Method	FR Model	IR152		IRSE50		FaceNet		MobileFace		Average	
	ASR% ↑	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
Noise-based	MI-FGSM (CVPR'18)	3.0	13.2	3.8	12.6	8.4	22.0	7.8	20.6	5.75	17.10
	TI-DIM (CVPR'19)	8.2	14.2	4.2	13.8	17.8	32.2	21.4	43.8	12.90	26.00
	TIP-IM (ICCV'21)	9.8	29.8	7.4	27.8	25.8	55.8	31.8	52.6	18.70	41.50
Unrestricted	Clip2Protect (CVPR'23)	17.8	51.8	10.6	35.2	27.2	54.4	39.0	67.8	23.65	52.30
	GIFT (ACMMM'24)	21.2	**57.2**	**34.6**	49.4	33.2	65.6	41.2	67.6	32.55	59.95
	DFPP (arXiv'24)	14.8	41.4	13.6	37.2	25.2	53.8	36.6	60.4	22.55	48.20
	SD4Privacy (ICME'24)	15.6	26.8	23.4	41.2	33.6	53.8	31.8	49.8	26.10	42.90
	Adv-CPG (Ours)	**24.4**↑3.2	56.4↓0.8	33.8↓0.8	**51.2**↑1.8	**36.6**↑3.0	**67.4**↑1.8	**43.4**↑2.2	**70.4**↑2.6	**34.55**↑2.0	**61.35**↑1.4

Figure: Face Identification Attack Success Rate (ASR) of Adv-CPG compared to other adversarial attacks. (Image from the paper Table 3)

In most cases, the results of the proposed Adv-CPG are superior to recent approaches under Rank-1 and Rank-5 settings.

Comparison on Image Quality

Method (Metric)	PGD	TIP-IM	Adv-Makeup	AMT-GAN	CLIP2Protect	DiffAM	SD4Privacy	Adv-Diffusion	**Adv-CPG**
FID ↓	78.92	38.7325	4.2282	34.5703	26.1272	26.1015	26.4078	22.5751	26.0758
PSNR ↑	27.96	33.2106	34.5152	19.5048	19.3542	20.5260	26.8864	28.8501	29.9996
SSIM↑	0.8573	0.9236	0.9850	0.7924	0.6025	0.8861	0.8092	0.8046	0.8978

Figure: Comparison of image quality metrics for Adv-CPG and other methods. (Image from the paper Table 4)

- In contrast to the noise-based method TIP-IM, the adversarial examples of Adv-CPG are more natural, with no obvious noise patterns.

Comparison on Image Quality

Method (Metric)	PGD	TIP-IM	Adv-Makeup	AMT-GAN	CLIP2Protect	DiffAM	SD4Privacy	Adv-Diffusion	**Adv-CPG**
FID ↓	78.92	38.7325	4.2282	34.5703	26.1272	26.1015	26.4078	22.5751	26.0758
PSNR ↑	27.96	33.2106	34.5152	19.5048	19.3542	20.5260	26.8864	28.8501	29.9996
SSIM↑	0.8573	0.9236	0.9850	0.7924	0.6025	0.8861	0.8092	0.8046	0.8978

Figure: Comparison of image quality metrics for Adv-CPG and other methods. (Image from the paper Table 4)

- In contrast to the noise-based method TIP-IM, the adversarial examples of Adv-CPG are more natural, with no obvious noise patterns.
- In terms of unrestricted makeup-based approaches, CLIP2Protect generates reasonable and high-quality makeup, and DiffAM achieves outstanding performance in terms of precision and refinement

Comparison on Image Quality

Method (Metric)	PGD	TIP-IM	Adv-Makeup	AMT-GAN	CLIP2Protect	DiffAM	SD4Privacy	Adv-Diffusion	**Adv-CPG**
FID ↓	78.92	38.7325	4.2282	34.5703	26.1272	26.1015	26.4078	22.5751	26.0758
PSNR ↑	27.96	33.2106	34.5152	19.5048	19.3542	20.5260	26.8864	28.8501	29.9996
SSIM↑	0.8573	0.9236	0.9850	0.7924	0.6025	0.8861	0.8092	0.8046	0.8978

Figure: Comparison of image quality metrics for Adv-CPG and other methods. (Image from the paper Table 4)

1. Among all quantitative evaluations, Adv-Makeup performs the best. The reason is that Adv-makeup only generates eye shadow, not full-face makeup, which makes minimal modifications to the image. Also, Adv-Makeup had a really low ASR.
2. In summary, Adv-CPG obtains comparatively low FID scores and high PSNR and SSIM scores, indicating that the images generated by Adv-CPG are relatively natural.

Results (Part 2): It works on Real-World Systems

Commercial APIs Testing

For benchmarking, recent methods are chosen, including 4 unrestricted makeup- based methods and 5 unrestricted facial semantical invariant methods.

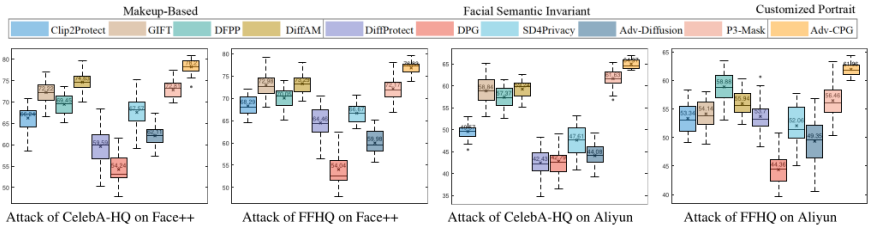


Figure: The confidence scores (↑) returned from commercial APIs, Face++ and Aliyun. (Images from the paper)

The results demonstrate that the average confidence scores of Adv-CPG on each API are about 77.5 and 63.0, respectively, which is superior to current SOTA methods, and the attack effectiveness is comparatively stable.

Commercial APIs Testing

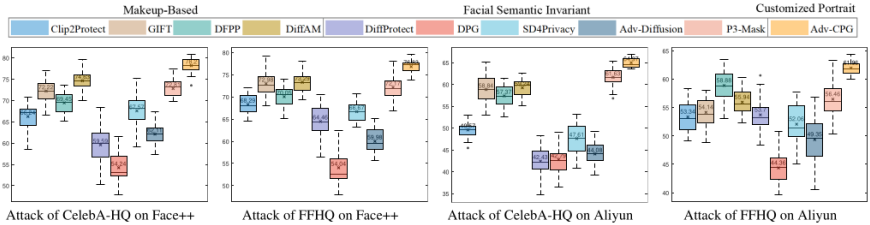


Figure: The confidence scores (↑) returned from commercial APIs, Face++ and Aliyun. (Images from the paper)

Adv-CPG has higher and more stable confidence scores than state-of-the-art noise-based and makeup-based facial privacy protection methods.

Conclusions

Conclusions

Adv-CPG Overview

- Adv-CPG is the first framework to integrate facial adversarial attacks into customized portrait generation.

Conclusions

Adv-CPG Overview

- Adv-CPG is the first framework to integrate facial adversarial attacks into customized portrait generation.
- It achieves progressive two-layer facial privacy protection by injecting a target identity and adding identity guidance.

Conclusions

Adv-CPG Overview

- Adv-CPG is the first framework to integrate facial adversarial attacks into customized portrait generation.
- It achieves progressive two-layer facial privacy protection by injecting a target identity and adding identity guidance.
- The multi-modal image customizer enables fine-grained, high-fidelity, and personalized portrait generation.

Conclusions

Experimental results show Adv-CPG outperforms state-of-the-art methods in:

- Robust black-box attacks against diverse face recognition models and commercial APIs.
- Maintaining high image quality and customization.

Future work could explore the application of this facial privacy protection paradigm to other facial manipulation tasks, such as face editing and face swapping, to fully protect facial privacy in the age of artificial intelligence.

References


Junying Wang, Hongyuan Zhang, and Yuan Yuan. *Adv-CPG: A Customized Portrait Generation Framework with Facial Adversarial Attacks*.
https://arxiv.org/pdf/2503.08269

Thank you!

cdimas@tuc.gr