JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

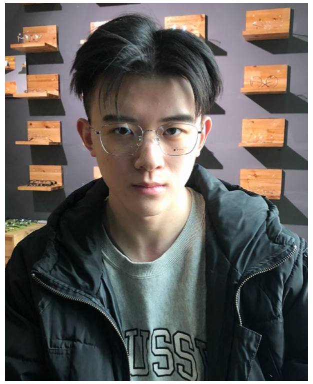
东方理工高等研究院
EASTERN INSTITUTE FOR ADVANCED STUDY

Discrete Point-wise Attack Is Not Enough: Generalized Manifold Adversarial Attack for Face Recognition

Paper Tag: THU-AM-390

Project Page: https://github.com/tokaka22/GMAA


Qian Li


Yuxiao Hu


Ye Liu


Dongxiao Zhang


Xin Jin

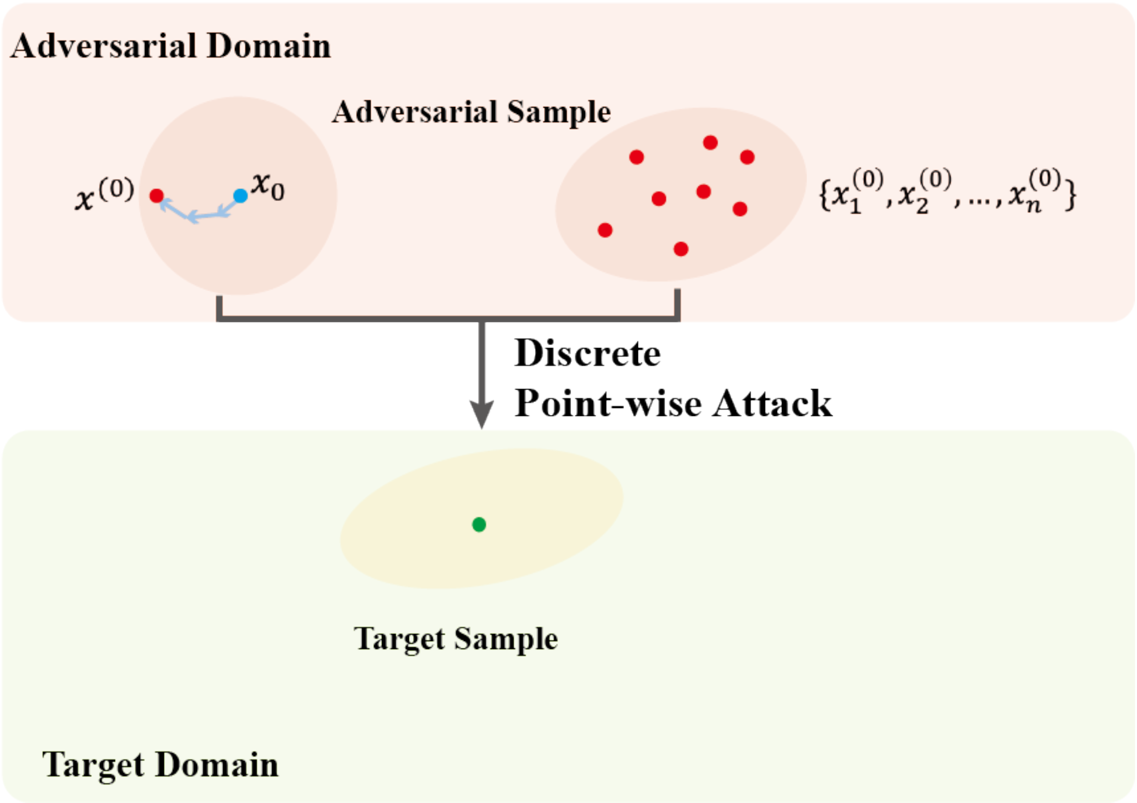

Yuntian Chen

Eastern Institute for Advanced Study, Ningbo Zhejiang, China

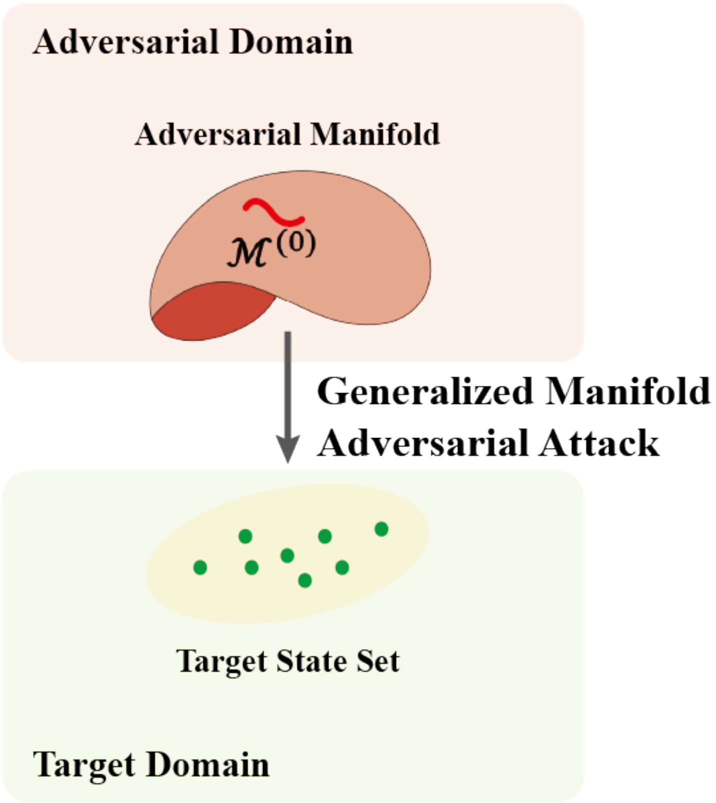
Overview

- We propose a new adversarial attack paradigm **GMAA**.

Discrete Point-wise Attack

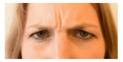


Generalized Manifold Adversarial Attack

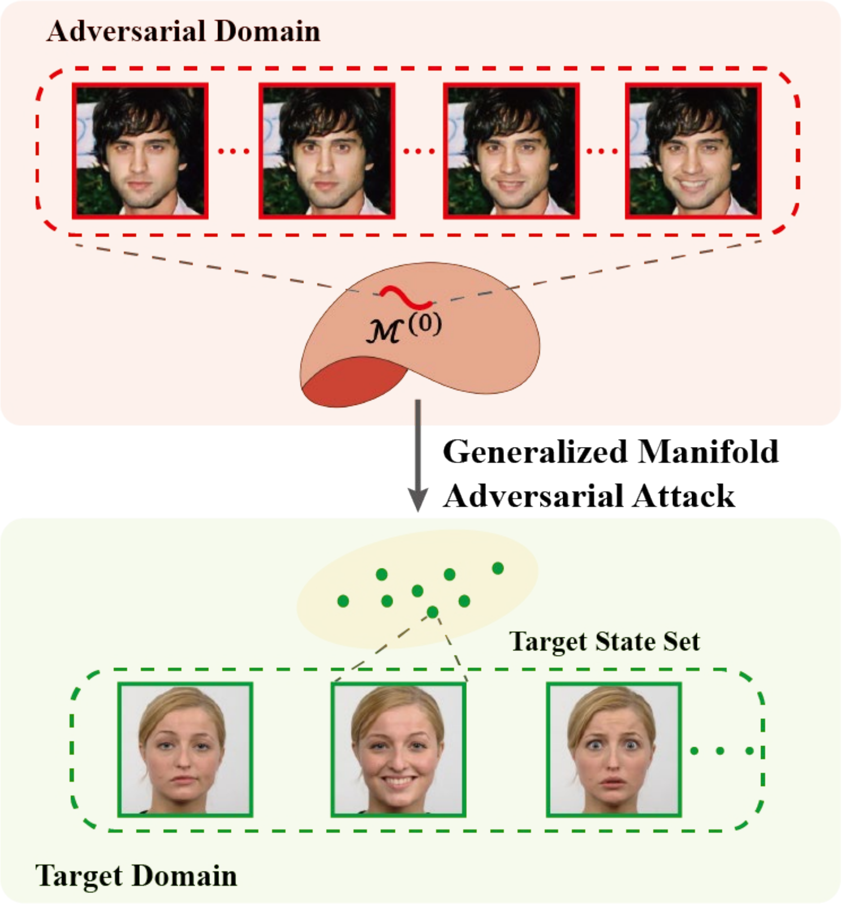


Overview

- We propose a new adversarial attack paradigm **GMAA**.
- We instantiate GMAA in the **face expression state space**.

Action Unit	Description	Facial Muscle	Example
1	Inner Brow Raiser	*Frontalis, pars medialis*	
2	Outer Brow Raiser (unilateral, right side)	*Frontalis, pars lateralis*	
4	Brow Lowerer	*Depressor Glabellae, Depressor Supercilli, Currugator*	
5	Upper Lid Raiser	*Levator palpebrae superioris*	
6	Cheek Raiser	*Orbicularis oculi, pars orbitalis*	
7	Lid Tightener	*Orbicularis oculi, pars palpebralis*	

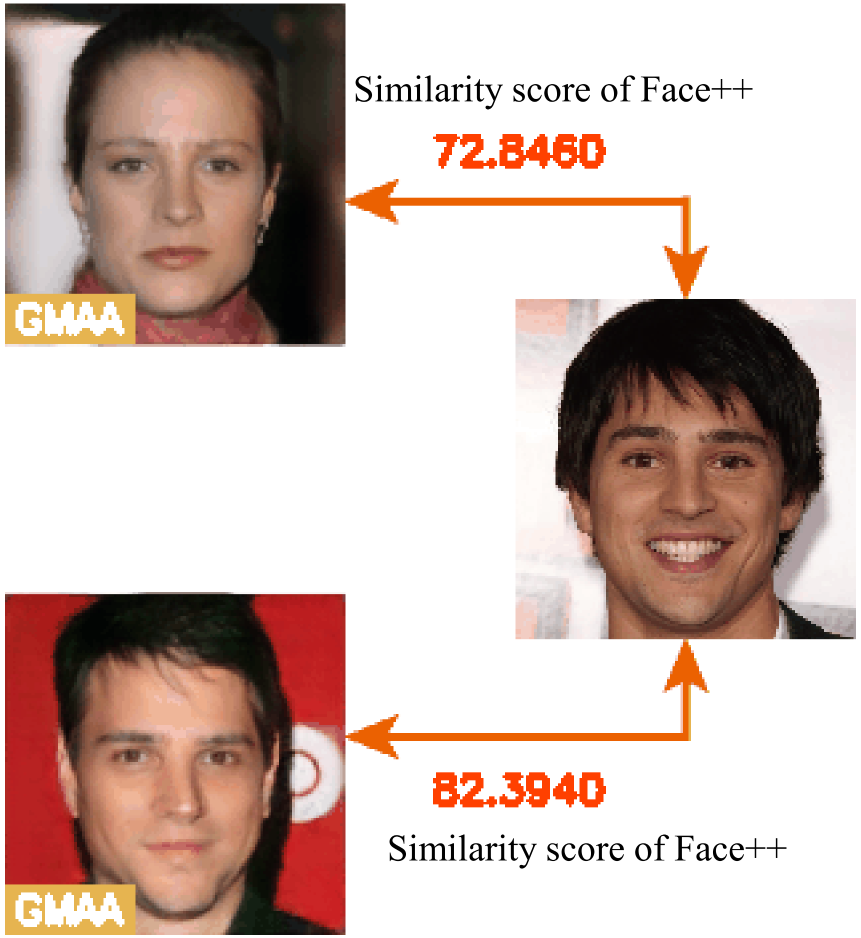
Domain knowledge
AU Vector[1]



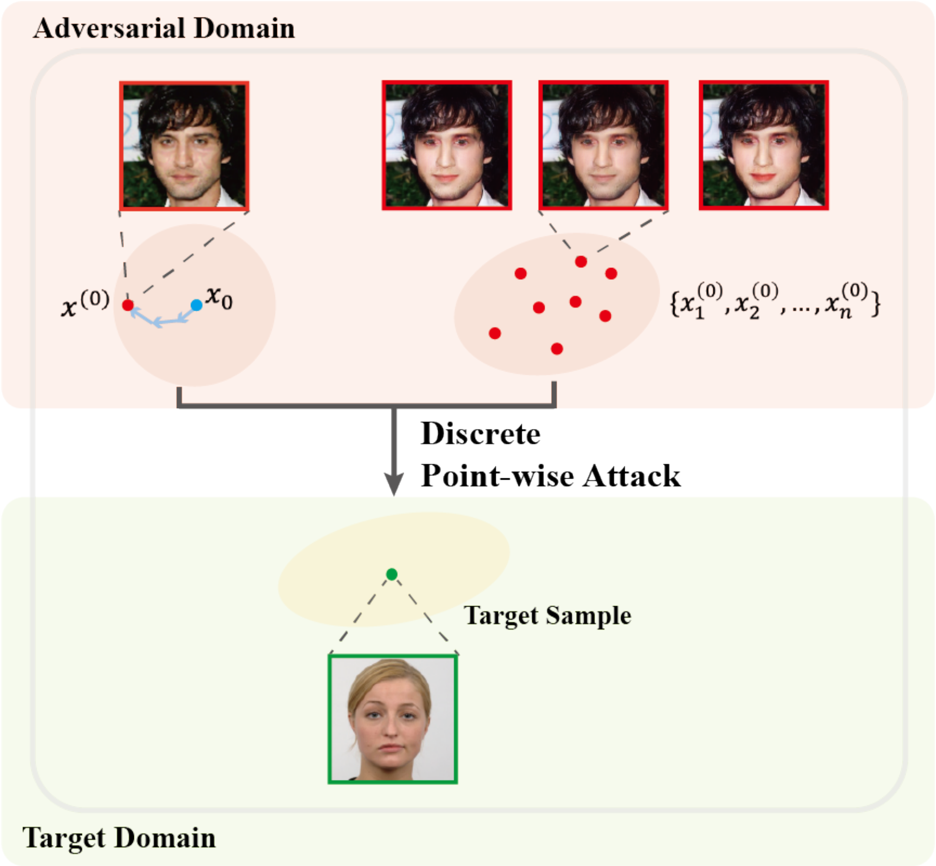
[1] https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/

Overview

- We propose a new adversarial attack paradigm **GMAA**.
- We instantiate GMAA in the **face expression state space**.
- Our method has better attack performance and higher visual quality.

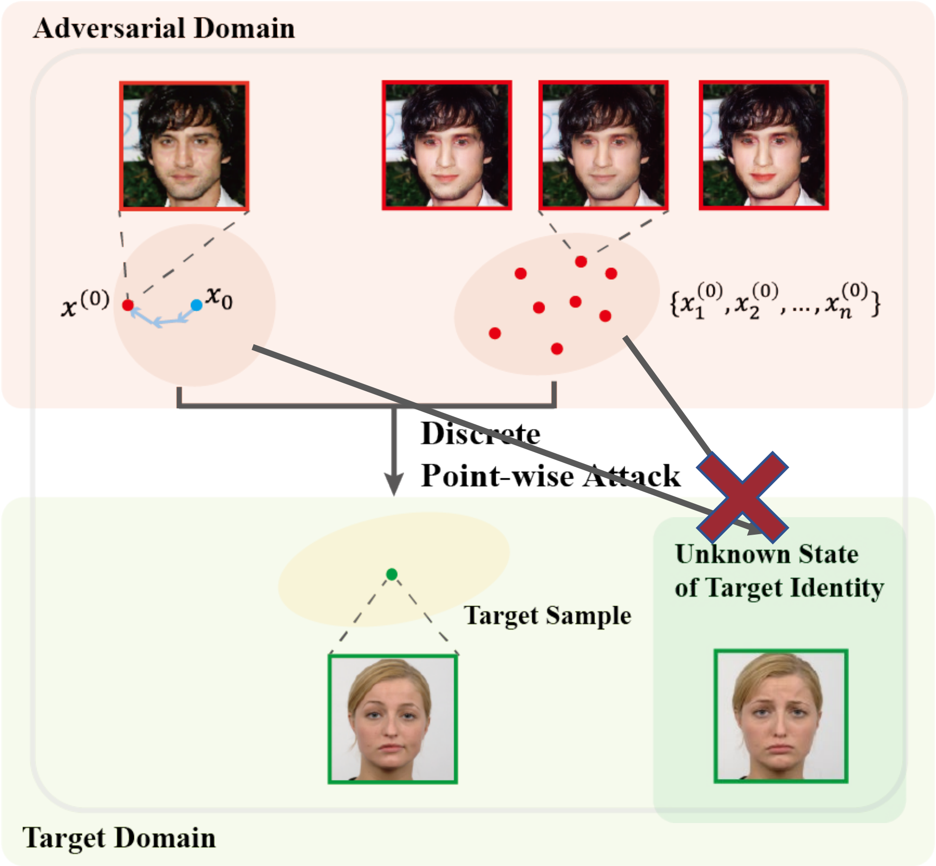


Limitations of previous work

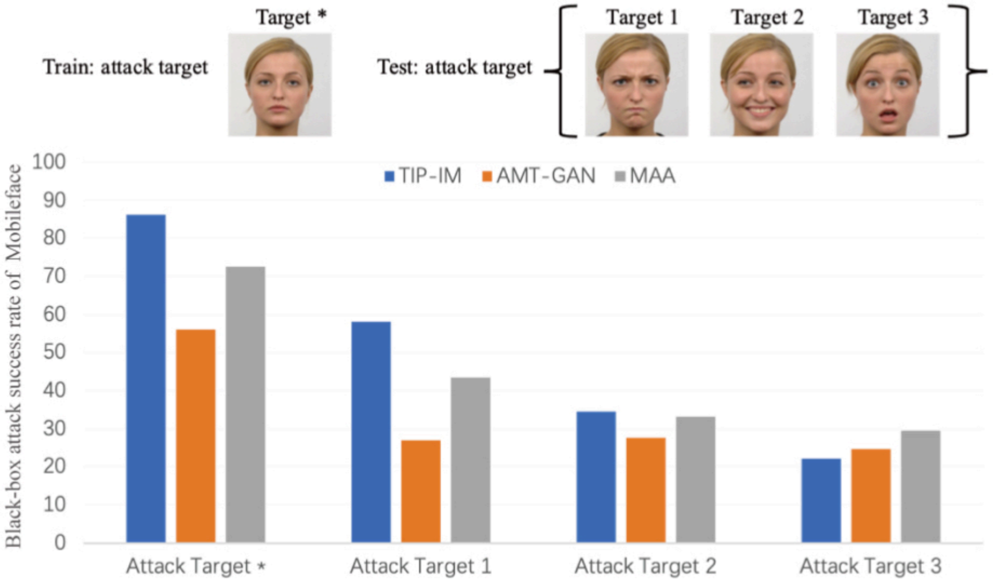


- For the target domain, previous methods tend to attack a single target identity sample.

Limitations of previous work

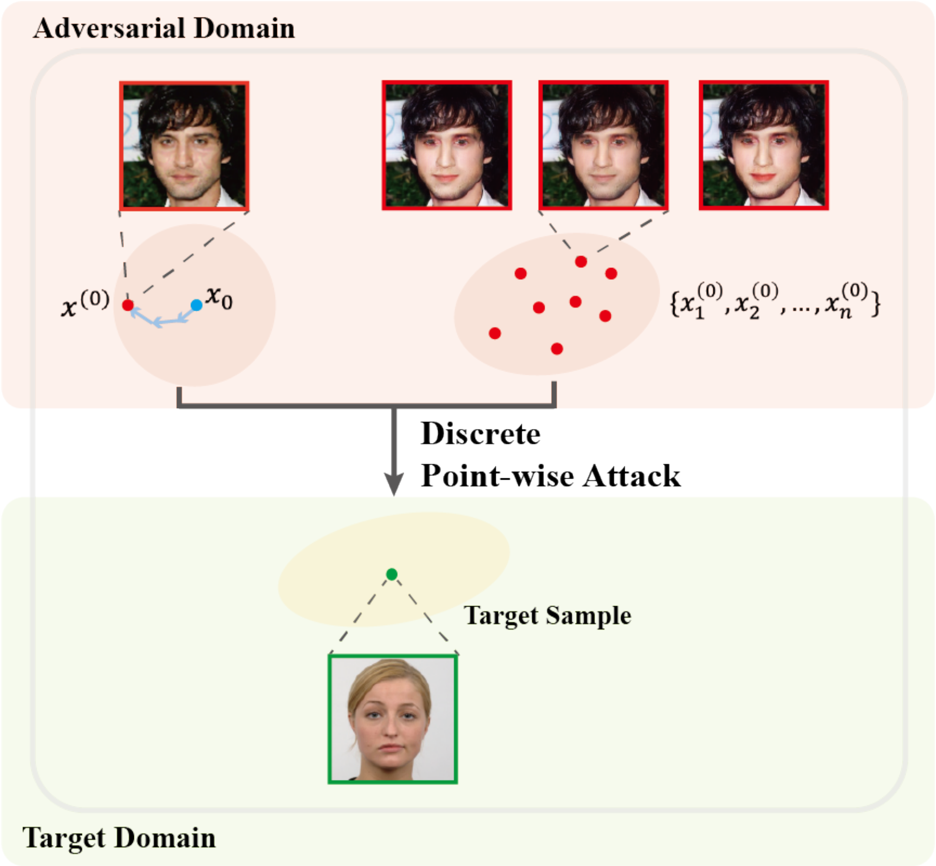


- For the target domain, previous methods tend to attack a single target identity sample.



Poor generalization on unknow state target images !

Limitations of previous work



- For the target domain, previous methods tend to attack a single target identity sample.

Poor generalization on unknow state target images !

Generate highly generalizable adversarial examples !

- For the adversarial domain, many methods searching for **discrete adversarial examples** in a hypersphere of the clean sample.

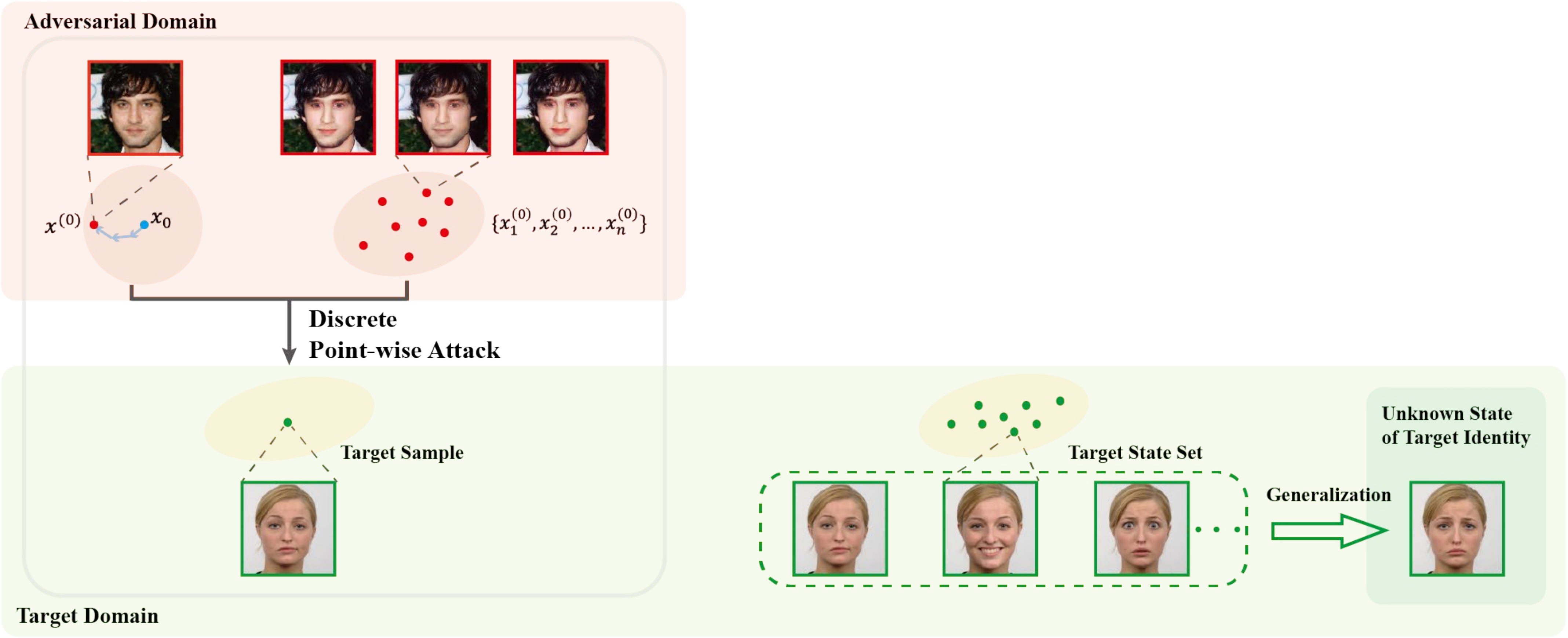
Ignore the continuity of the adversarial domain !

Find a continuous adversarial manifold instead of discrete adversarial examples!

Existing works are not strong enough both in target domain and adversarial domain.

Generalized Manifold Adversarial Attack

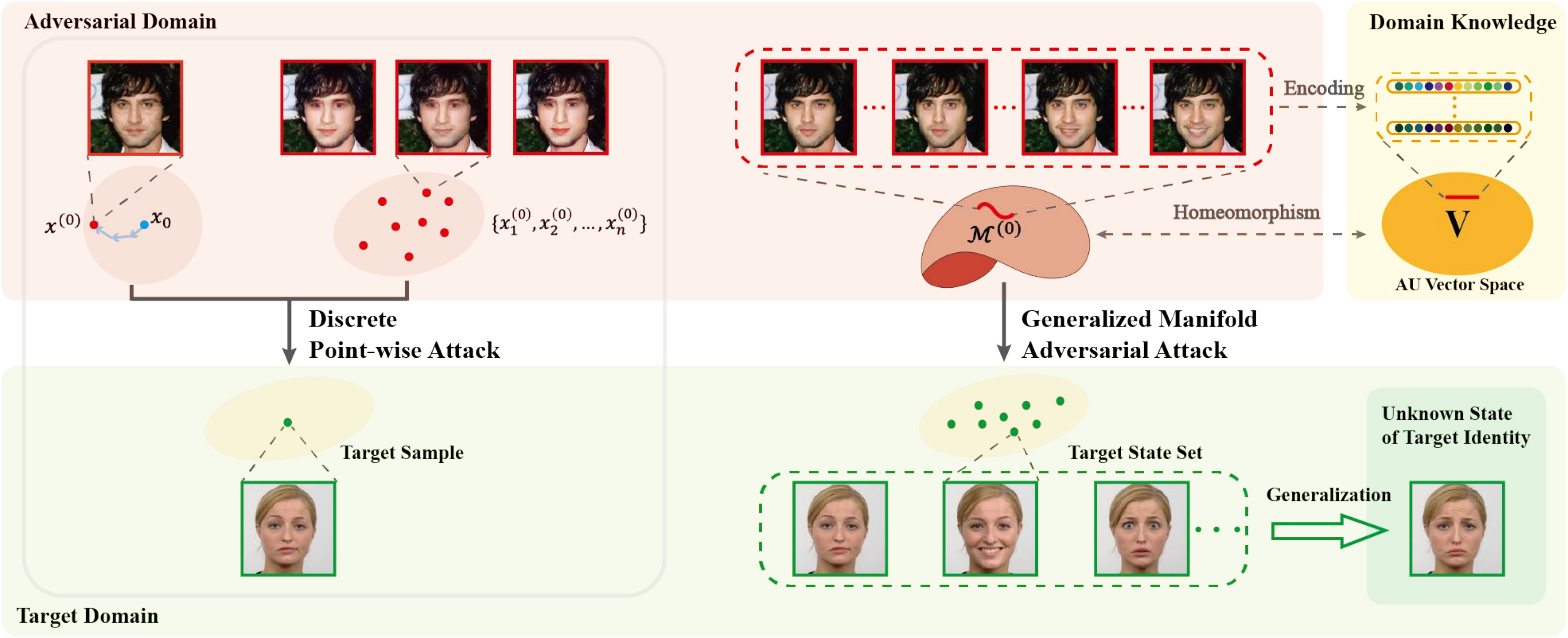
Core idea



- Expand the target domain **from one to many** to encourage a good generalization.

GMAA

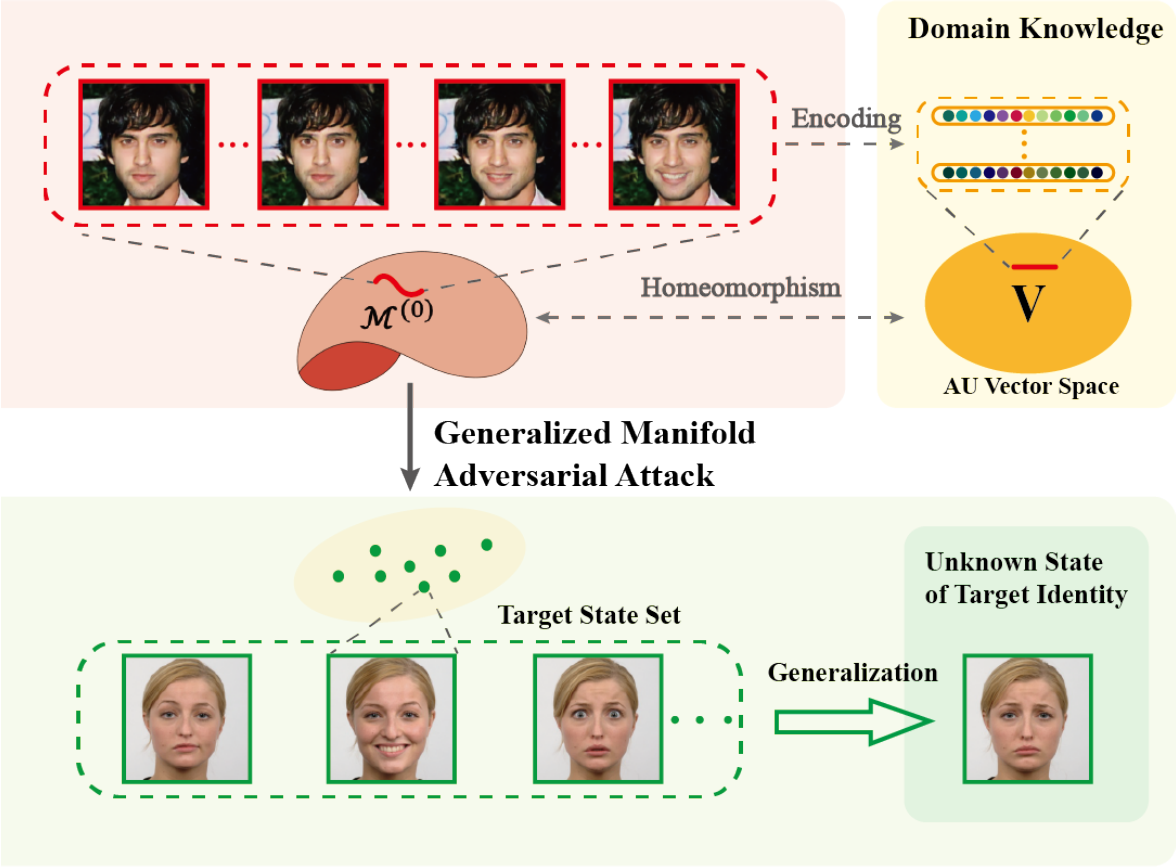
Core idea



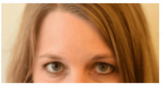
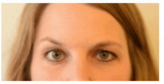
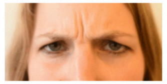
GMAA

- Expand the target domain **from one to many** to encourage a good generalization.
- Expand the adversarial domain **from discrete points to manifold** to strengthen the attack effect.

Core idea



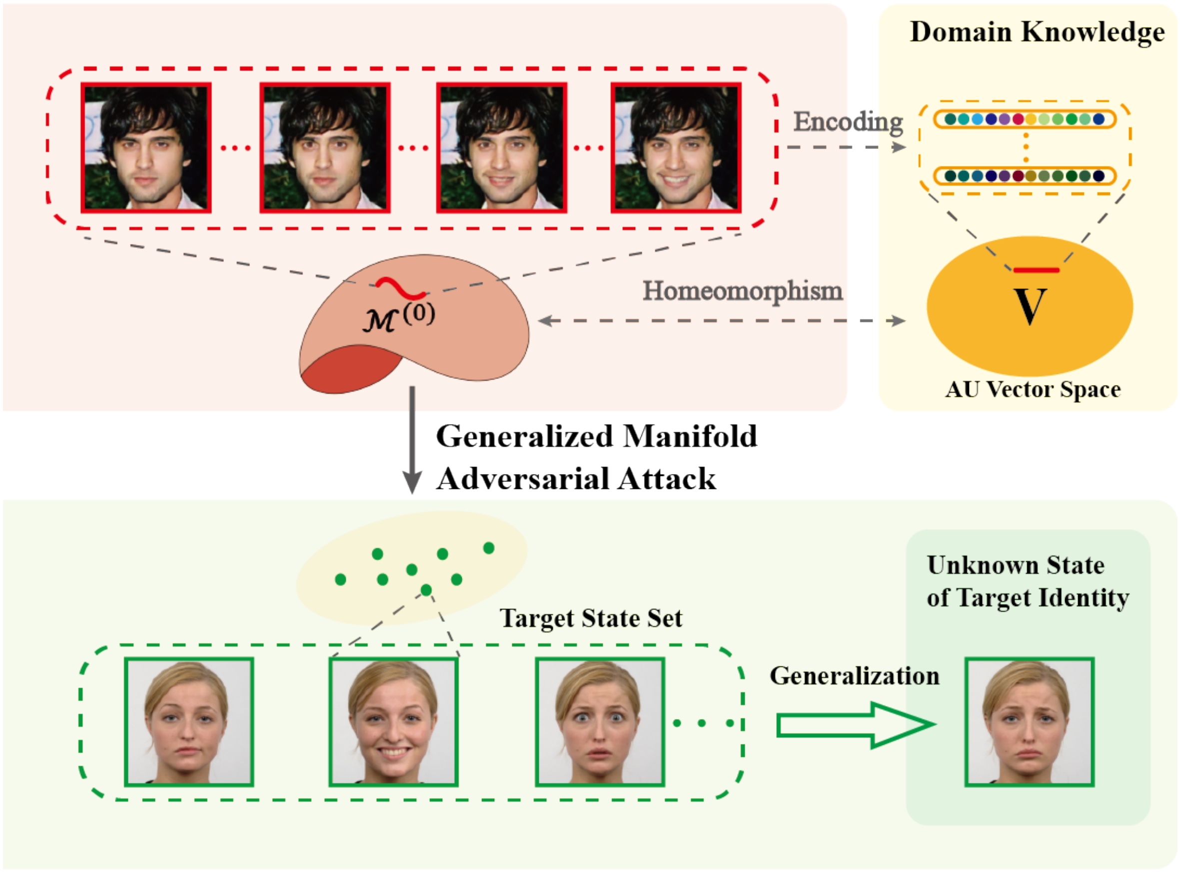
Facial Action Coding System

Action Unit	Description	Facial Muscle	Example
1	Inner Brow Raiser	*Frontalis, pars medialis*	
2	Outer Brow Raiser (unilateral, right side)	*Frontalis, pars lateralis*	
4	Brow Lowerer	*Depressor Glabellae, Depressor Supercilli, Currugator*	
5	Upper Lid Raiser	*Levator palpebrae superioris*	
6	Cheek Raiser	*Orbicularis oculi, pars orbitalis*	
7	Lid Tightener	*Orbicularis oculi, pars palpebralis*	

GMAA

- Expand the target domain **from one to many** to encourage a good generalization.
- Expand the adversarial domain **from discrete points to manifold** to strengthen the attack effect.

Core idea



Definition 1. *Let $\boldsymbol{x}_0 \in \mathbb{R}^{3\times H\times W}$, then $\mathcal{M}^0 = G(\boldsymbol{x}_0;\boldsymbol{\theta})$ is a continuous adversarial space if and only if*
(1) $\mathcal{M}^0$ is a subspace of $\mathbb{R}^{3\times H\times W}$.
(2) $\forall \boldsymbol{x}_i^0 \in \mathcal{M}$, $\boldsymbol{x}_i^0$ is an adversarial version of $\boldsymbol{x}_0$.

Theorem 1. *$\mathcal{M}^0$ generated by G_0 is a continuous adversarial manifold, where $G_0 : V \to \mathcal{M}$ is a map when fixed the input $\boldsymbol{x}_0$ in G.*

Remark 1. *Since the $\mathcal{M}^0$ generated by G_0 is a continuous adversarial manifold when fixed the $\boldsymbol{x}_0$, then we can assert over the sample space Ω, the adversarial examples space generated by G constitutes an adversarial fiber bundle.*

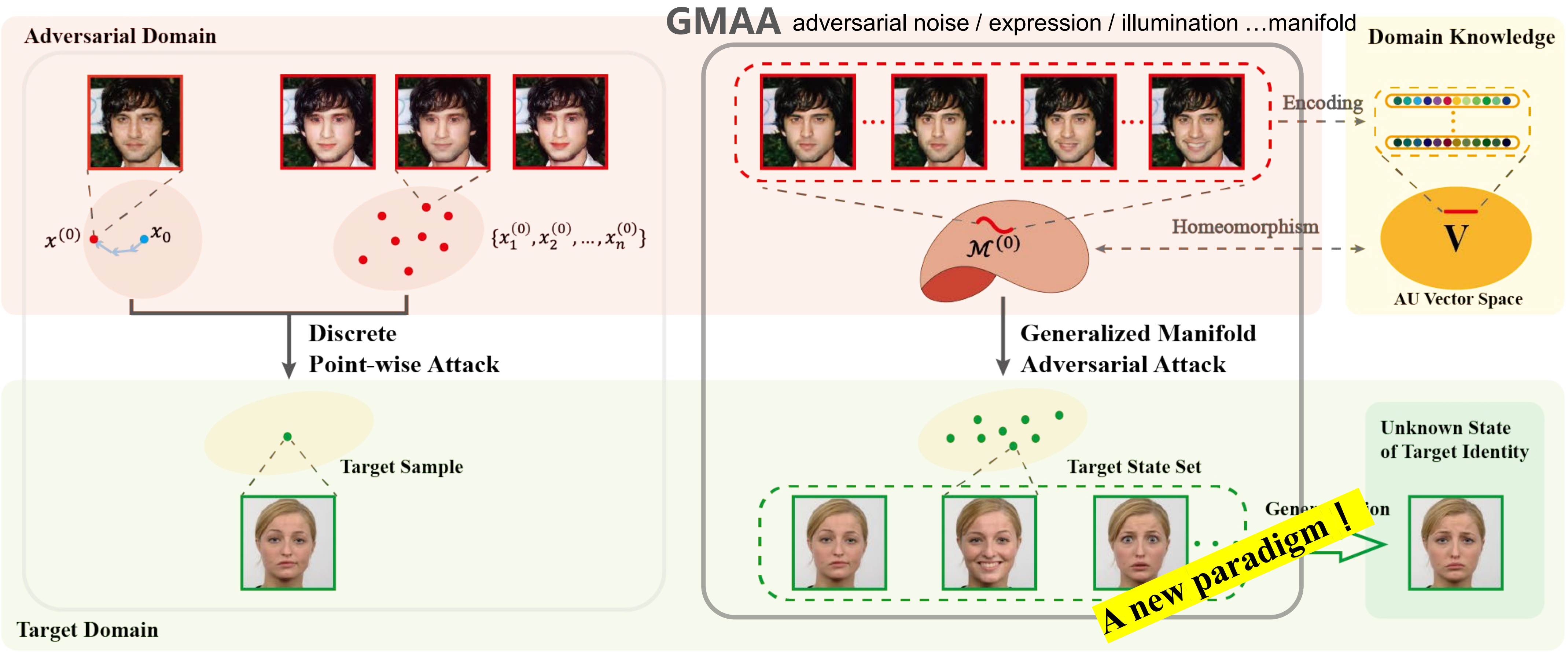
Definition 2. *$\mathcal{M}^0$ generated by $\boldsymbol{x}_0 \in \mathbb{R}^{3\times H\times W}$ is a semantic continuous adversarial space if and only if*
(1) $\mathcal{M}^0$ is a continuous adversarial space.
(2) $\forall \boldsymbol{x}_1^0, \boldsymbol{x}_2^0 \in \mathcal{M}^0$, if $\boldsymbol{x}_1^0$ is close to $\boldsymbol{x}_2^0$ on the $\mathcal{M}^0$, then $\boldsymbol{x}_1^0$ and $\boldsymbol{x}_2^0$ satisfy the semantic consistency.

Theorem 2. *$\mathcal{M}^0$ generated by G_0 is a semantic continuous adversarial manifold, where $G_0 : V \to \mathcal{M}$ is a map when fixed the input $\boldsymbol{x}_0$ in G.*

GMAA

- Expand the target domain **from one to many** to encourage a good generalization.
- Expand the adversarial domain **from discrete points to manifold** to strengthen the attack effect.

Core idea

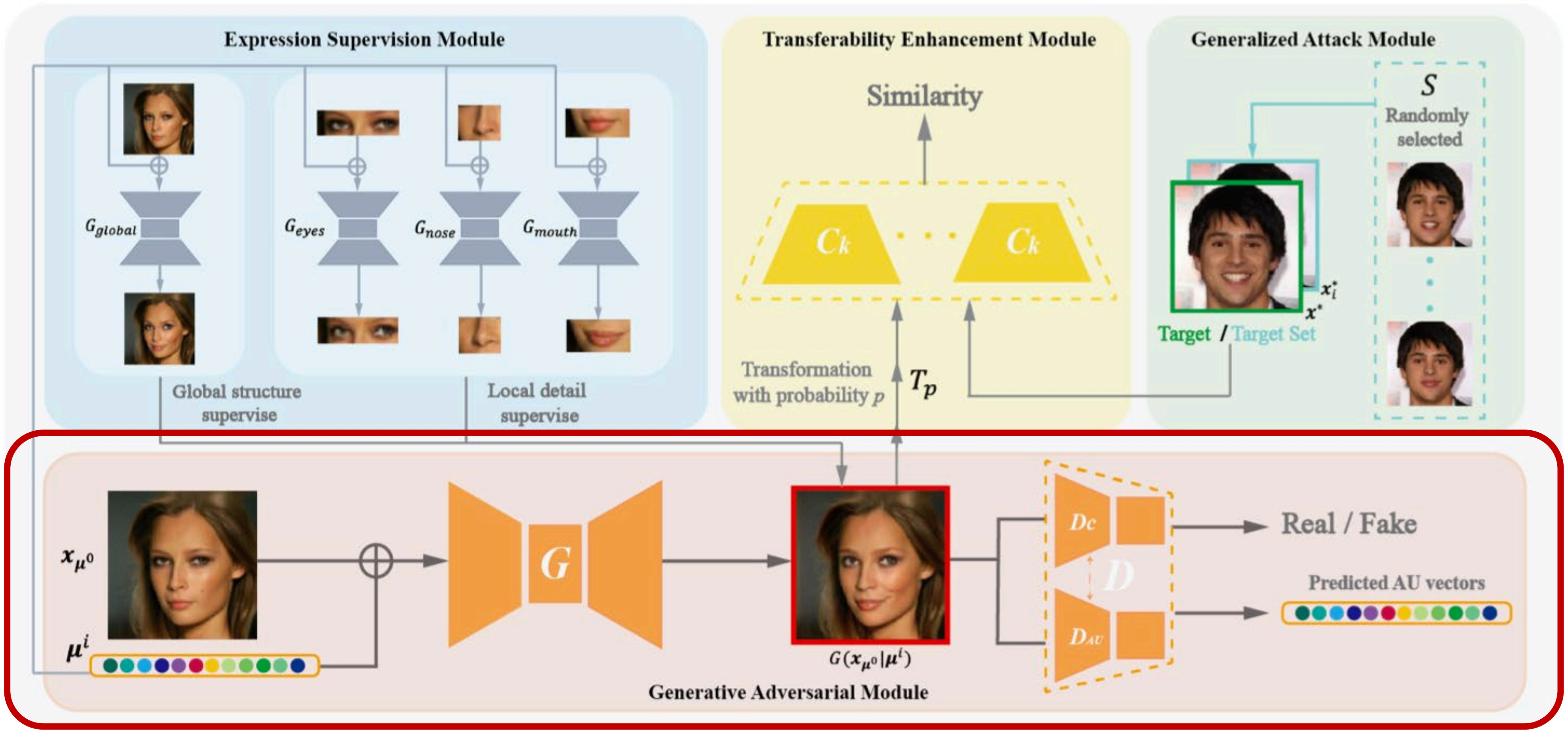


GMAA

- Expand the target domain **from one to many** to encourage a good generalization.
- Expand the adversarial domain **from discrete points to manifold** to strengthen the attack effect.

Discrete Point-wise Attack ⇨ Generalized Manifold Adversarial Attack

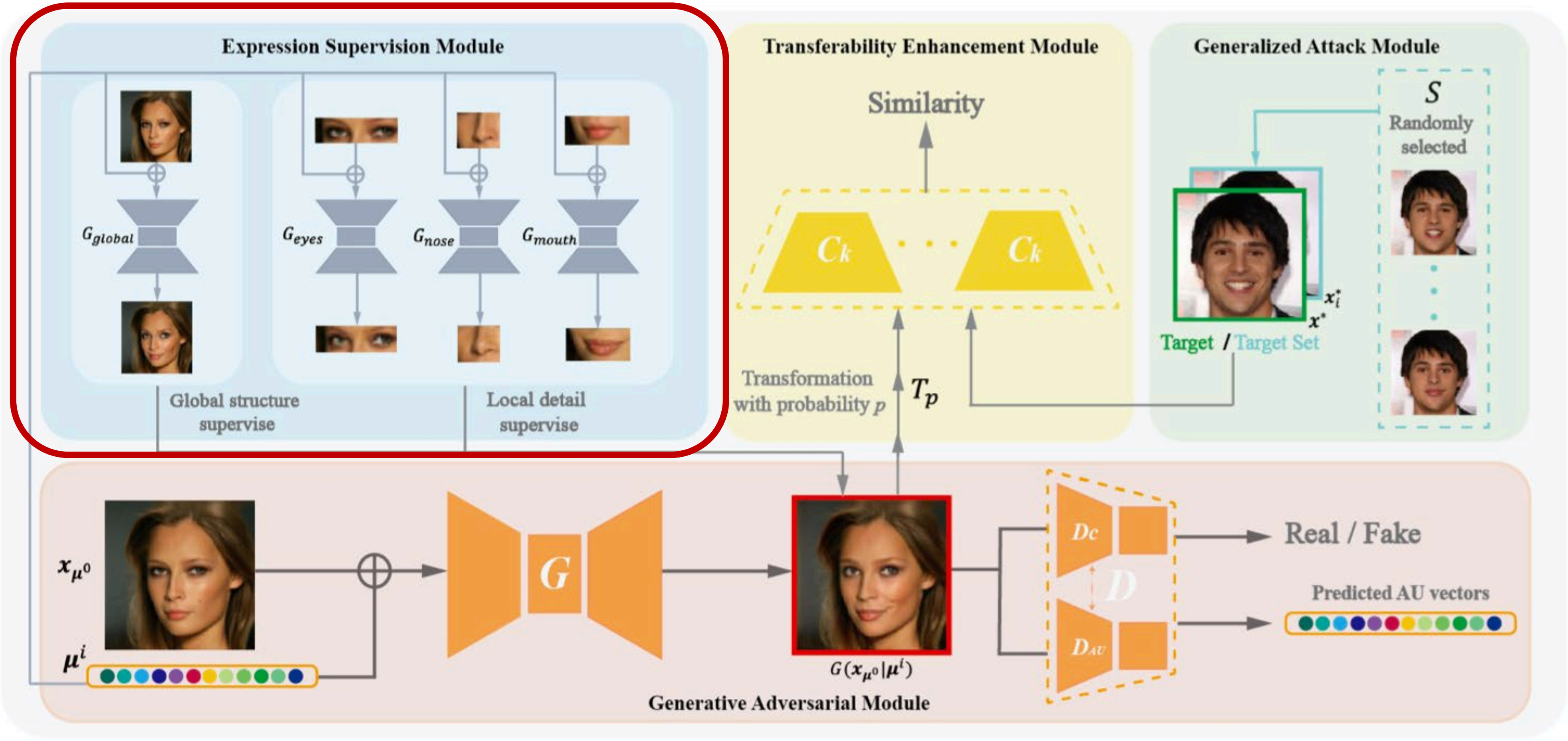
Method



Generative adversarial module

- The generator G produces adversarial example wearing the expression matching to the supplied AU label.
- The discriminator D_c learns to distinguish real images from generated images.
- The AU predictor D_{AU} learns the AU coding rules by real images and their AU labels.

Method

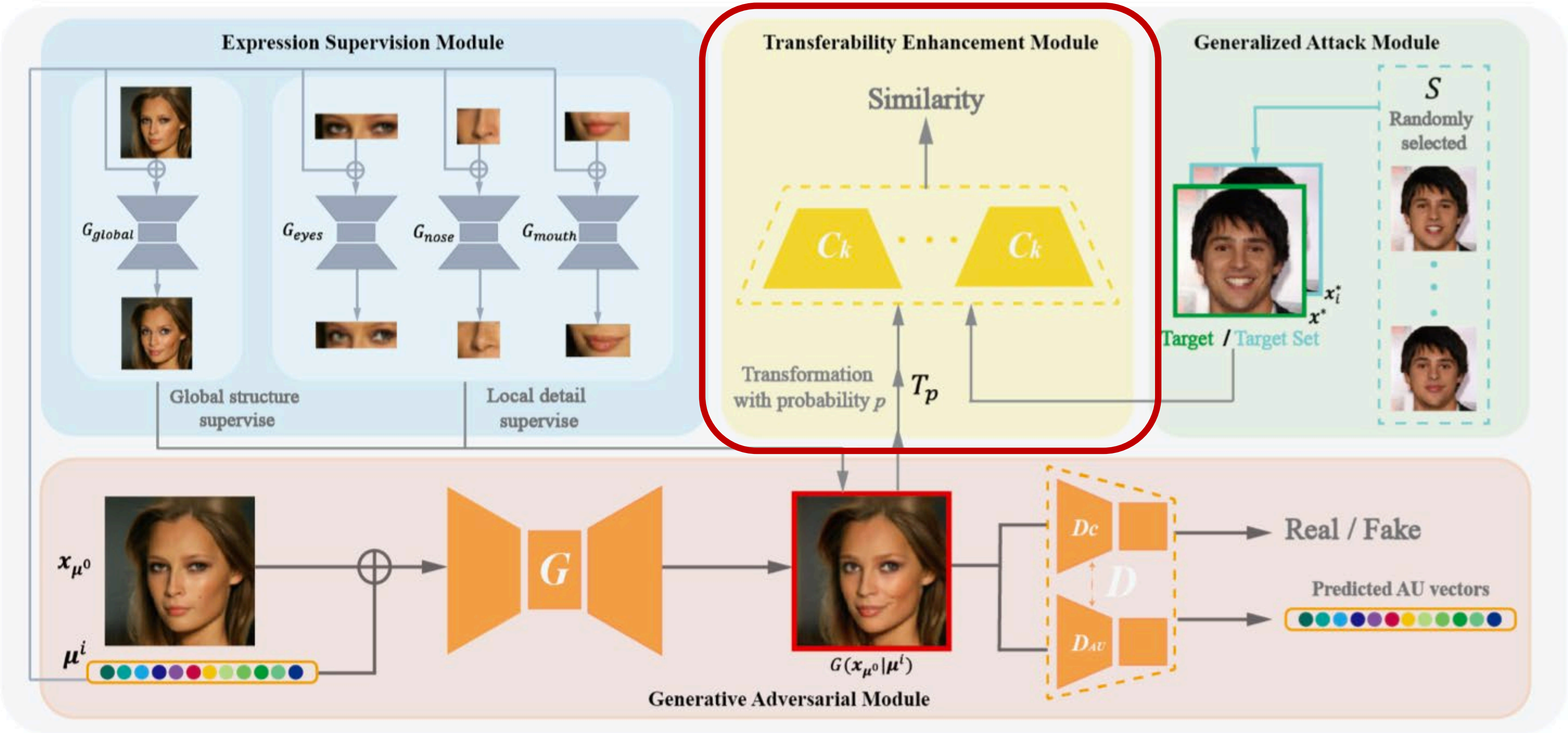


Expression supervision module

- Four pre-trained expression supervision networks protect the visual identity and guide G in expression editing.
- The global branch focuses on structural features of the face, whereas the local branch protects important facial details.
- Each generator has the network structure similar to [2].

[2] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Al- berto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.

Method

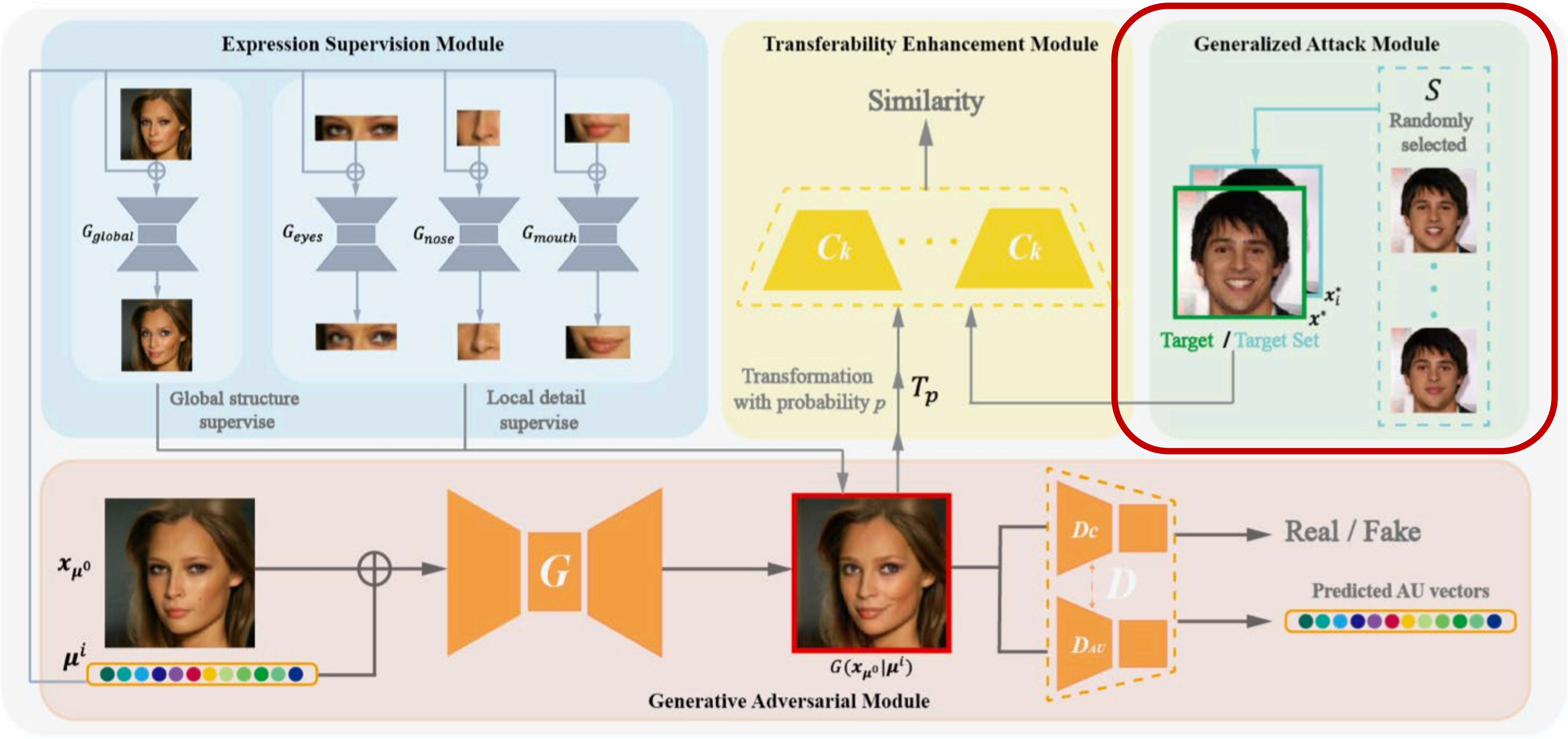


Transferability enhancement module

- To improve the transferability of adversarial examples and the black-box attack success rate, we introduce the transferability enhancement module from [3].
- All baselines are equipped with this module for a fair comparison.

[3] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15014–15023, 2022.

Method



Generalized attack module

- This module intends to raise the attack success rate on the unseen face belonging to the target identity.
- It is a generic module, which can be introduced into other adversarial attack approaches.
- Manifold Adversarial Attack (MAA) means the method without this module, just expand adversarial domain from point to manifold.
- When the model is coupled with this module, we call it G-(method name).

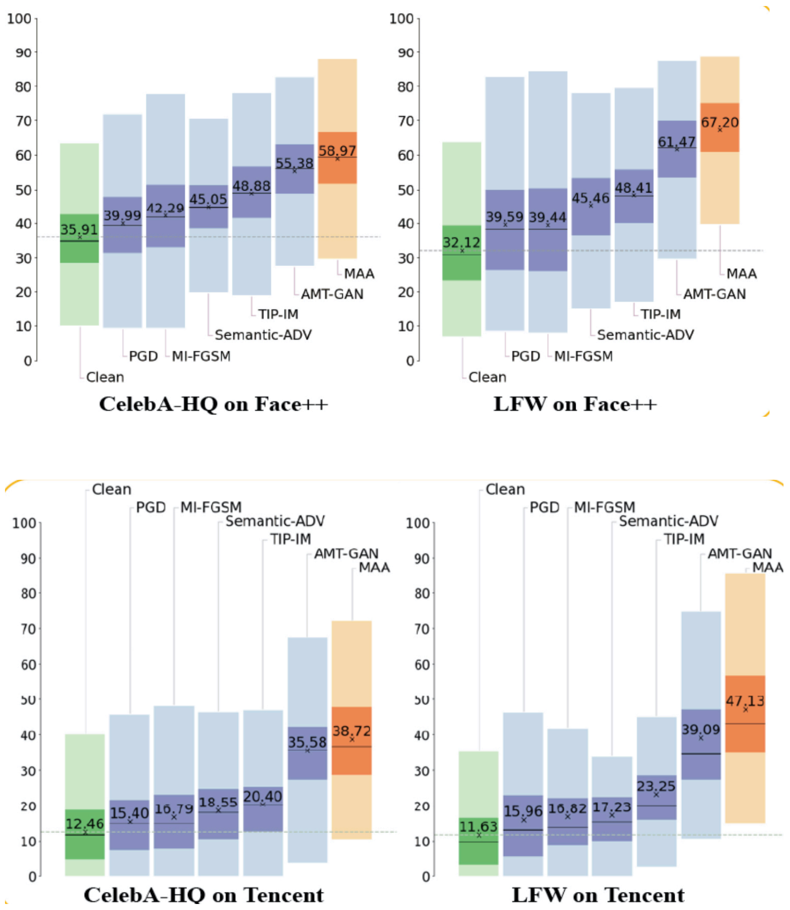
Experiment —— Part 1

- **Black-box attack success rate**

	CelebA-HQ			
	IRSE50	IR152	Facenet	Mobileface
Clean	3.68	3.08	1.31	8.43
PGD [23]	24.20	13.37	5.86	28.72
MI-FGSM [7]	38.90	20.76	9.25	40.48
SemanticAdv [26]	26.53	10.24	7.80	55.32
TIP-IM [34]	44.20	16.09	14.46	**65.36**
AMT-GAN [16]	51.06	15.63	11.63	33.27
MAA	**60.40**	**29.43**	**18.91**	56.13

	LFW			
	IRSE50	IR152	Facenet	Mobileface
Clean	3.20	0.06	0.04	5.00
PGD [23]	31.30	10.20	7.40	33.50
MI-FGSM [7]	38.20	14.20	7.60	39.40
SemanticAdv [26]	33.60	10.40	8.80	37.40
TIP-IM [34]	32.80	15.20	13.00	**79.00**
AMT-GAN [16]	40.72	25.23	13.89	35.67
MAA	**55.80**	**29.20**	**18.00**	60.80

- **Attack performance on commercial API**



CelebA-HQ on Face++ **LFW on Face++**

CelebA-HQ on Tencent **LFW on Tencent**

- **Visual quality**

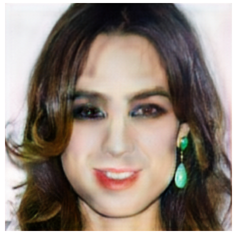
Ours

Target image

Attack Success Rate:100%

Original

TIP-IM
ICCV21



AMT-GAN
CVPR22

Experiment —— Part 2

- **Ablation studies of generalized attack module**

  ——Attack real state set

Targe set $S = \{$  $\cdots \}$

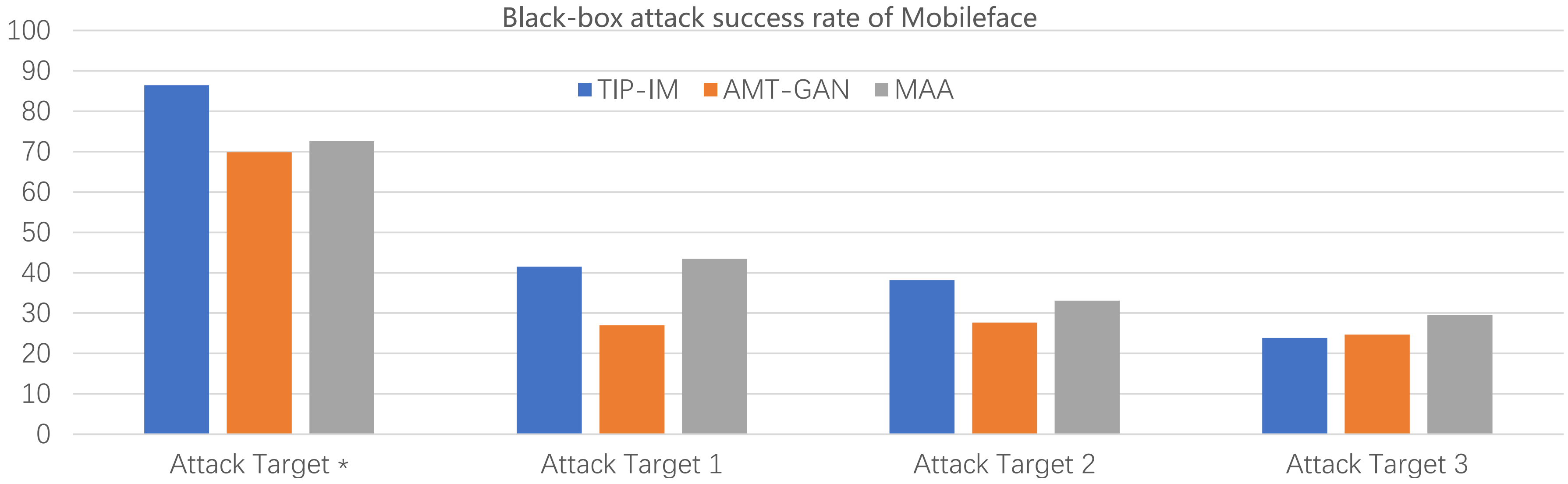
1 2 3

Targe * 

Case 0

Train: attack target *

Test: attack target *\1\2\3



Experiment —— Part 2

- **Ablation studies of generalized attack module**

  ——Attack real state set

Targe set $S = \{$  1 2 3 $\cdots \}$

Targe * 

Case 0

Train: attack target *

Test: attack target *\1\2\3

Case 1

Train: attack target $S/\{1\}$

Test : attack target 1

Case 2

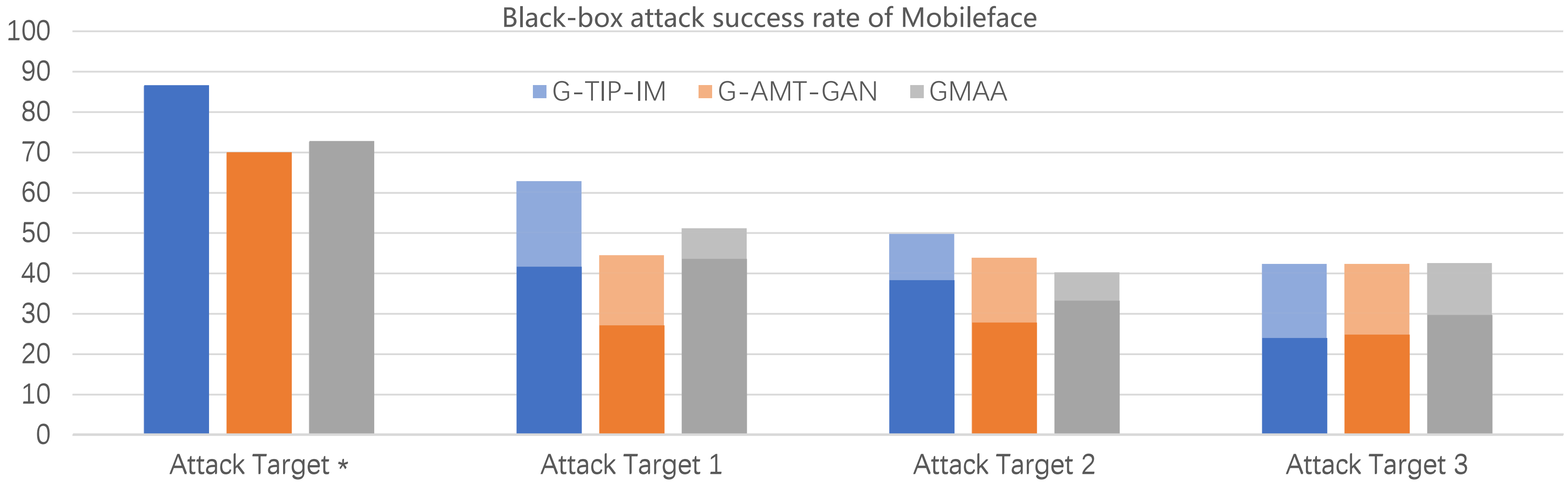
Train: attack target $S/\{2\}$

Test: attack target 2

Case 3

Train: attack target $S/\{3\}$

Test: attack target 3



Experiment —— Part 2

- **Ablation studies of generalized attack module**

  ——Attack real state set

Targe set $S = \{$


1


2


3

$\cdots \}$

Targe *

Case 0

Train: attack target *

Test: attack target *\1\2\3

Case 1

Train: attack target $S/\{1\}$

Test : attack target 1

Case 2

Train: attack target $S/\{2\}$

Test: attack target 2

Case 3

Train: attack target $S/\{3\}$

Test: attack target 3

	Target*		**Target 1**		**Target 2**		**Target 3**	
	Facenet	Mobileface	Facenet	Mobileface	Facenet	Mobileface	Facenet	Mobileface
TIP-IM [34] / **G-TIP-IM**	17.68	86.33	4.54 / **7.62**	58.03 / **70.93**	10.75 / **20.42**	34.42 / **49.20**	11.93 / **19.41**	22.21 / **42.43**
AMT-GAN [16] / **G-AMT-GAN**	16.12	55.95	8.22 / **13.23**	26.99 / **47.14**	9.78 / **17.12**	27.67 / **43.93**	10.91 / **16.16**	24.69 / **42.37**
MAA / **GMAA**	25.22	72.62	11.43 / **17.84**	43.44 / **67.50**	13.30 / **21.71**	33.08 / **41.24**	12.64 / **19.15**	29.56 / **47.21**

Experiment —— Part 2

- **Ablation studies of generalized attack module**

  ——Attack synthesized state set

Generated targe set $S = \{$

$\cdots \}$

	Facenet	Mobileface
TIP-IM [34] / **G-TIP-IM**	5.80 / **9.50**	17.20 / **23.5**
AMT-GAN [16] / **G-AMT-GAN**	4.04 / **8.27**	9.82 / **12.45**
MAA / **GMAA**	6.60 / **10.60**	13.50 / **21.60**

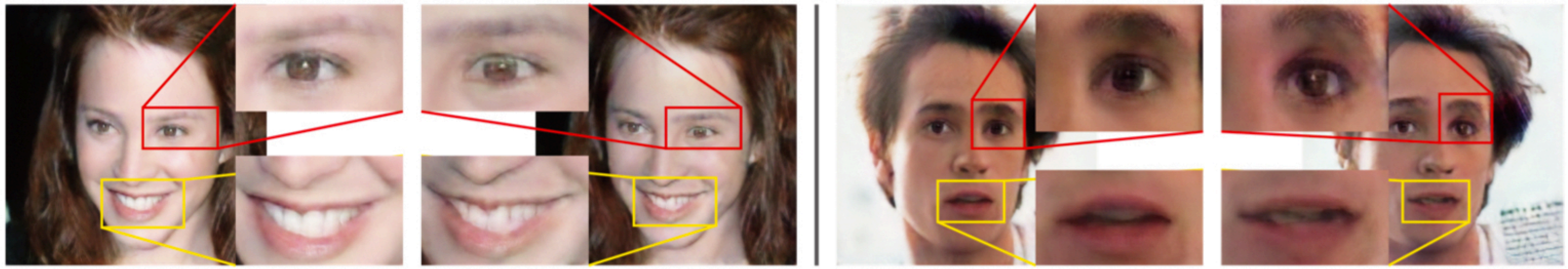
Experiment —— Part 3

- **Other Ablation studies**

—— Ablation studies of D_{AU}

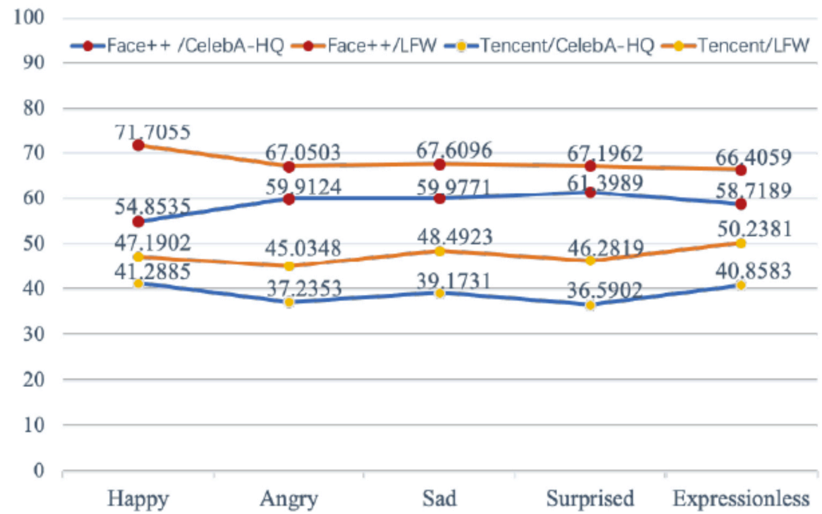
	Without D_{AU}	Without local editors	Original
MSE	0.5549	0.6283	**0.3582**

—— Ablation studies of local editors



with local editors **without local editors** **with local editors** **without local editors**

—— Ablation studies of different expressions



Thanks!

This work was supported and partially funded by the National Natural Science Foundation of China (Grant No. 62106116). This work was also supported in part by ZJNSFC under Grant LQ23F010008. We also would like to thank Han Fang of NUS for the valuable discussion.