AUEditNet: Dual-Branch Facial Action Unit Intensity Manipulation with Implicit Disentanglement

Shiwei Jin [1], Zhen Wang [2], Lei Wang [2], Peng Liu [2], Ning Bi [2], Truong Nguyen [1]
[1] ECE Dept. UC San Diego, [2] Qualcomm Technologies, Inc.
{sjin, tqn001}@eng.ucsd.edu, {zhewang, wlei, peli, nbi}@qti.qualcomm.com

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

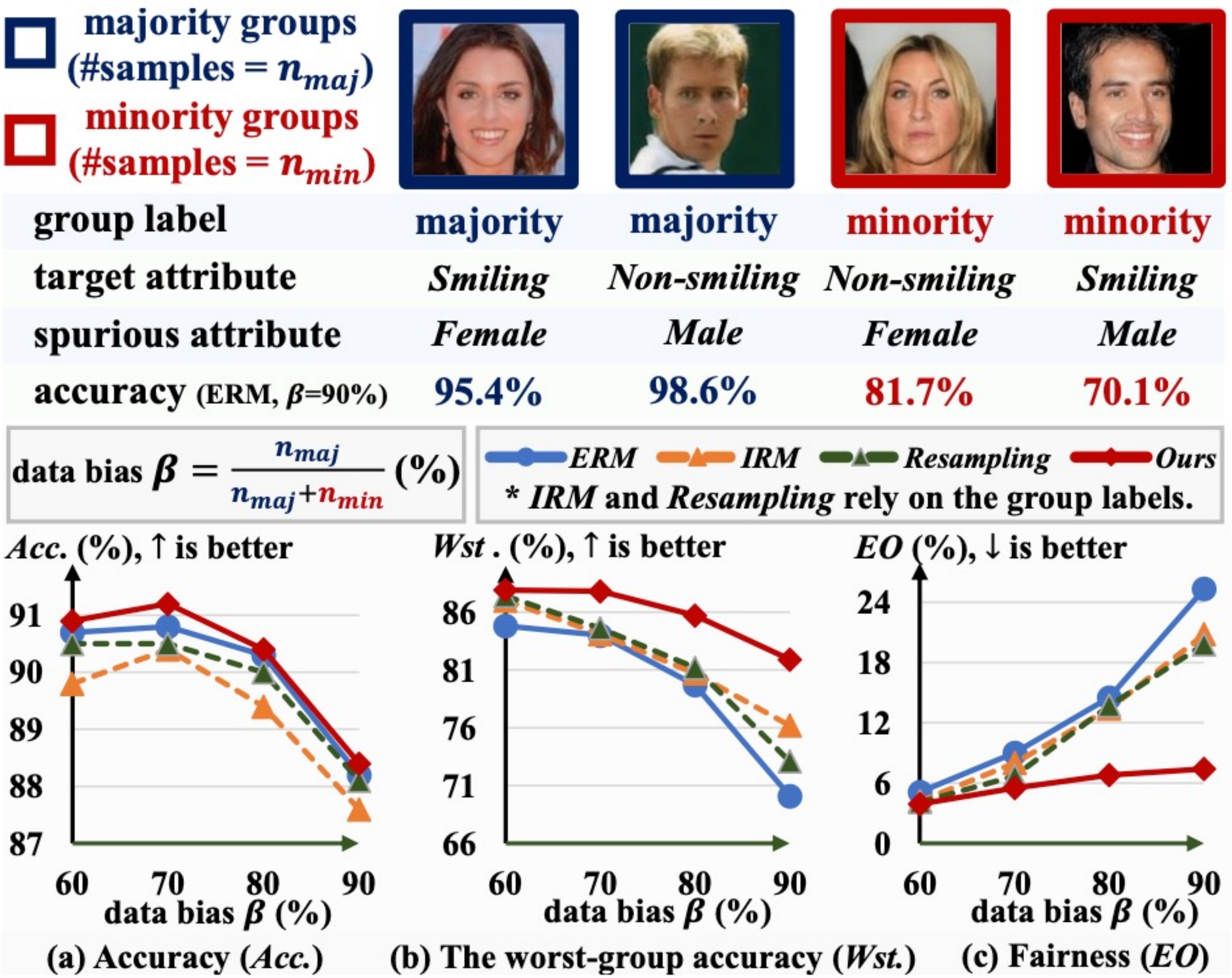
Introduction

- Unfairness is largely attributed to bias in data, where some spurious attributes (e.g., Male) statistically correlate with the target attribute (e.g., Smiling)

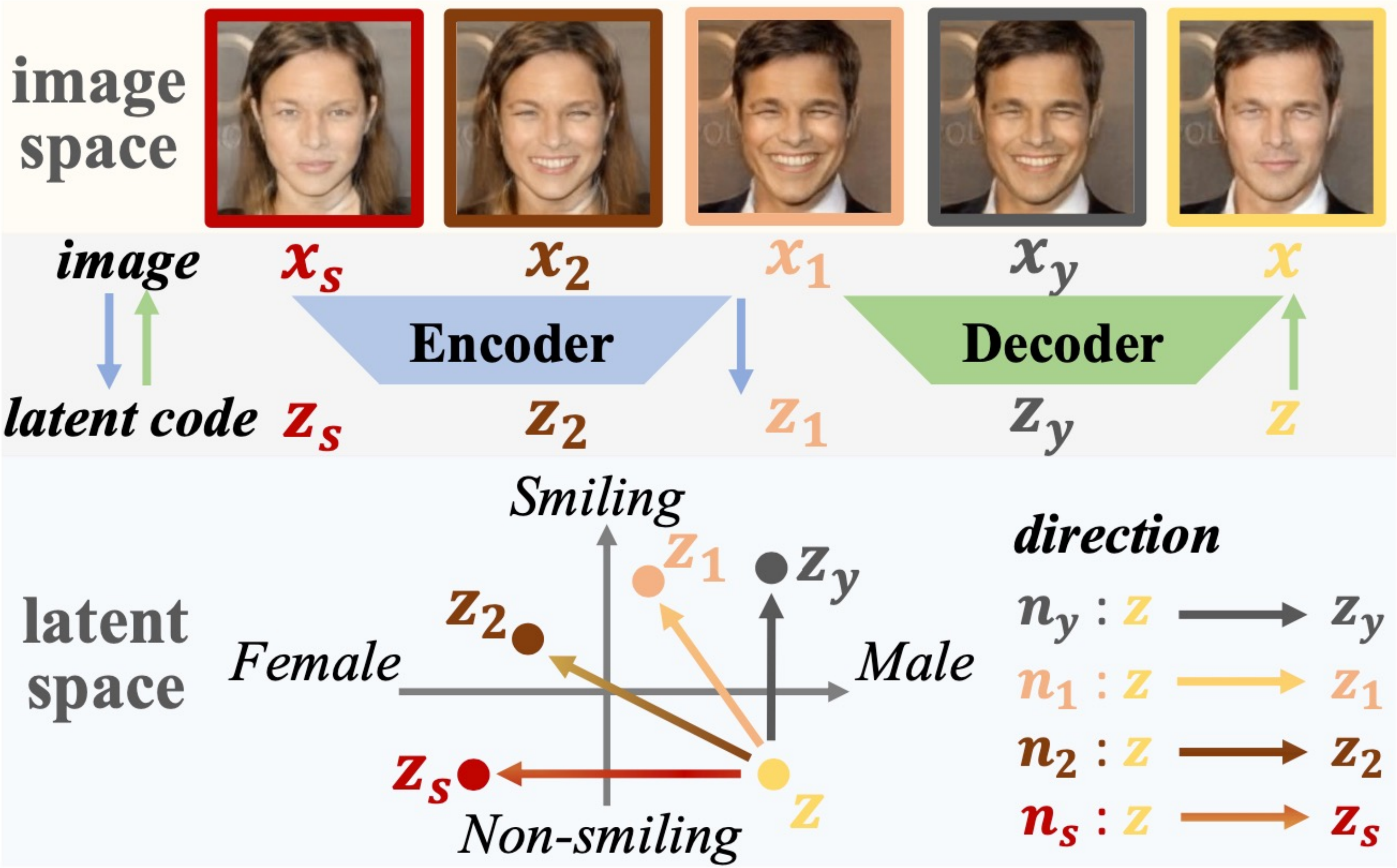
- Proposes a novel, generation-based two-stage framework to train a fair FAC model on biased data without additional annotation.

- Train a fair FAC model by fostering model invariance to these augmentation. Extensive experiments on three common datasets demonstrate the effectiveness.

Introduction



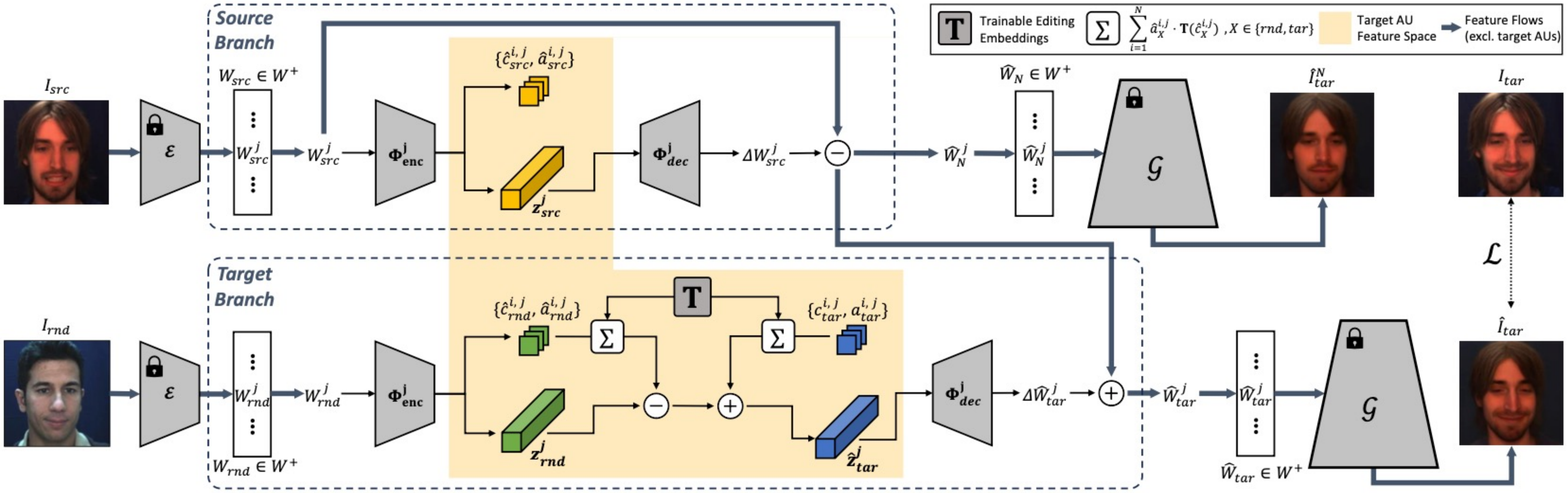
Introduction



Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework

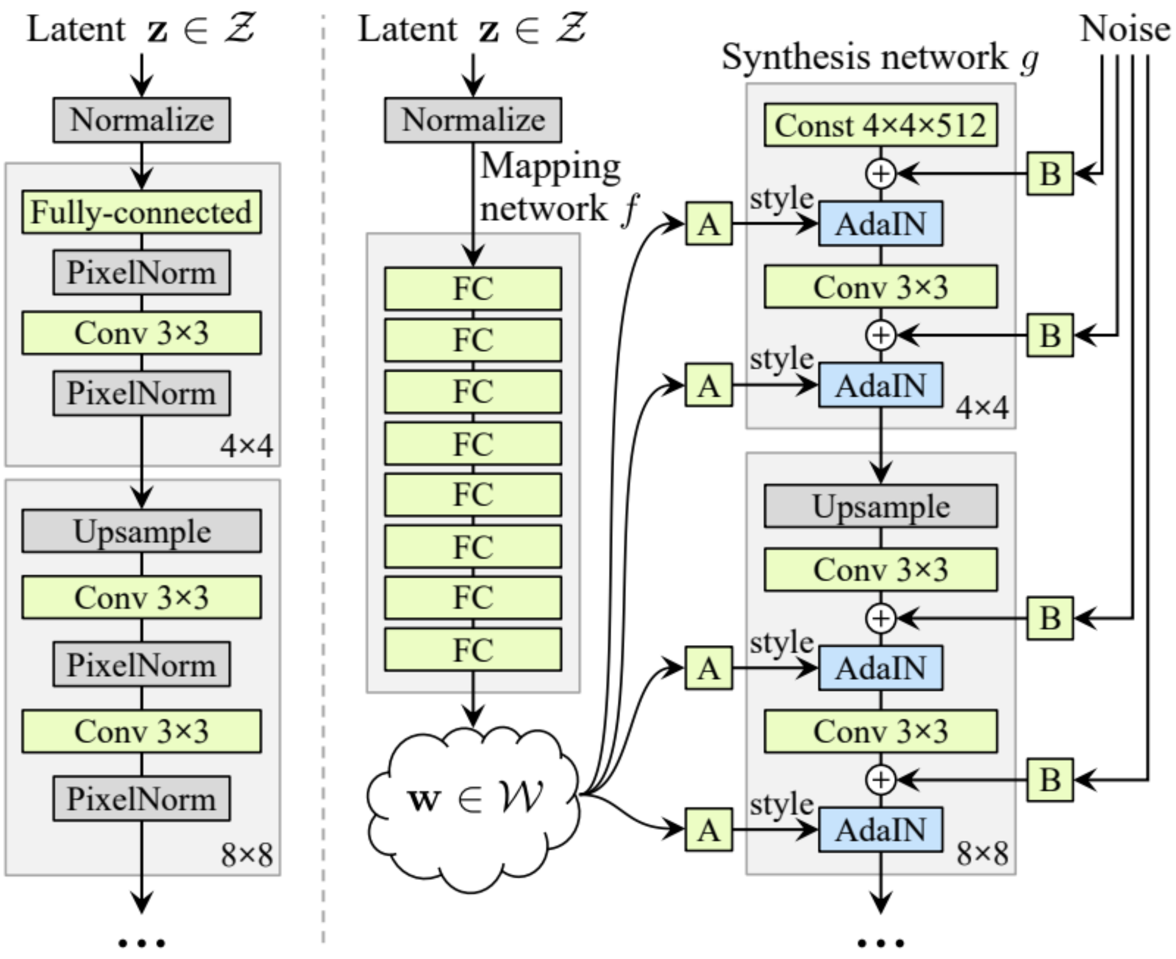


Source Branch
Target Branch
Trainable Editing Embeddings
$\sum_{i=1}^{N} \hat{a}_X^{i,j} \cdot \mathbf{T}(\hat{c}_X^{i,j})$, $X \in \{rnd, tar\}$
Target AU Feature Space
Feature Flows (excl. target AUs)
I_{src}
I_{rnd}
$\mathcal{E}$
$W_{src} \in W^+$
$W_{rnd} \in W^+$
W^j_{src}
W^j_{rnd}
Φ^j_{enc}
Φ^j_{dec}
$\{\hat{c}^{i,j}_{src}, \hat{a}^{i,j}_{src}\}$
$\{\hat{c}^{i,j}_{rnd}, \hat{a}^{i,j}_{rnd}\}$
$\{c^{i,j}_{tar}, a^{i,j}_{tar}\}$
z^j_{src}
z^j_{rnd}
$\hat{z}^j_{tar}$
ΔW^j_{src}
$\Delta \hat{W}^j_{tar}$
$\hat{W}^j_N$
$\hat{W}_N \in W^+$
$\hat{W}^j_{tar}$
$\hat{W}_{tar} \in W^+$
$\mathcal{G}$
$\hat{I}^N_{tar}$
I_{tar}
$\hat{I}_{tar}$
$\mathcal{L}$

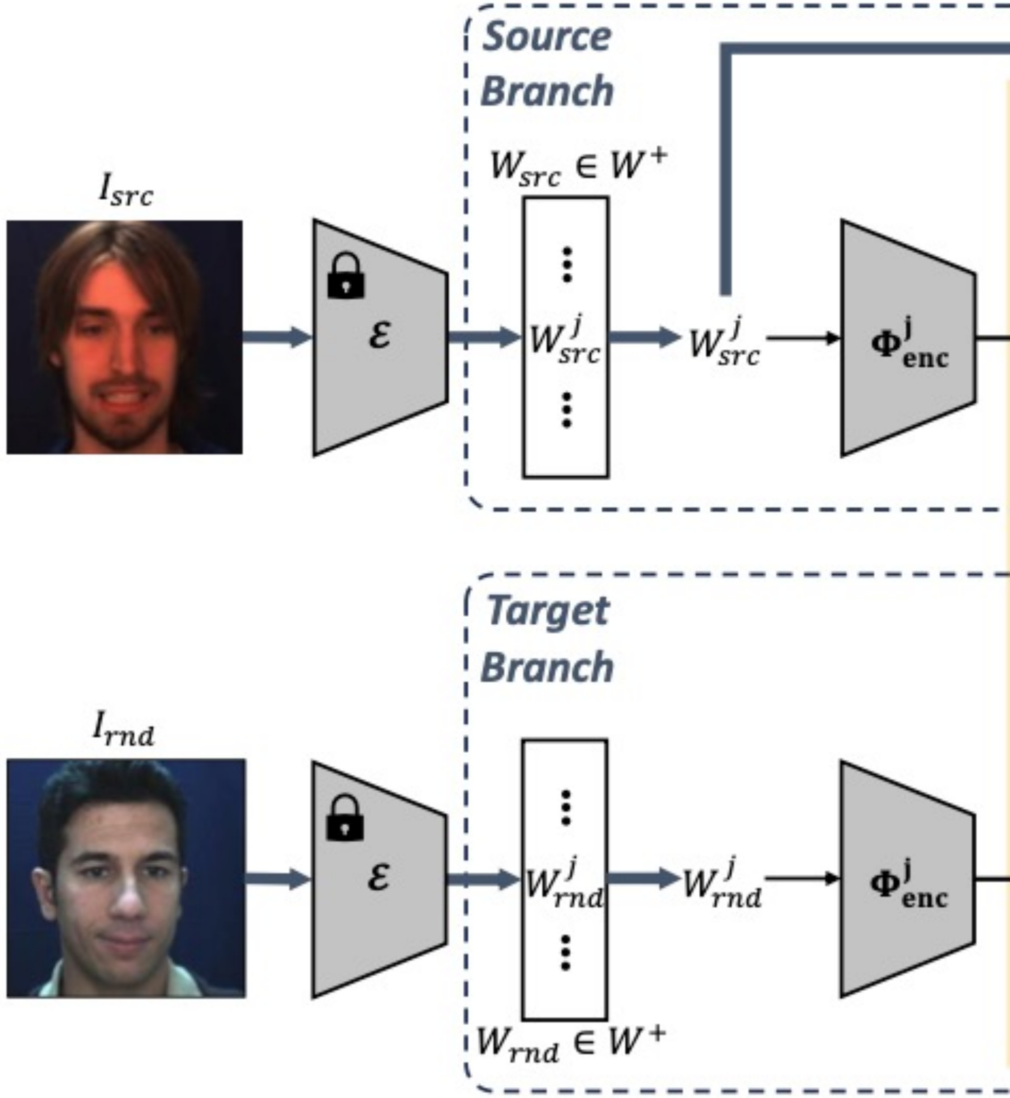
Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

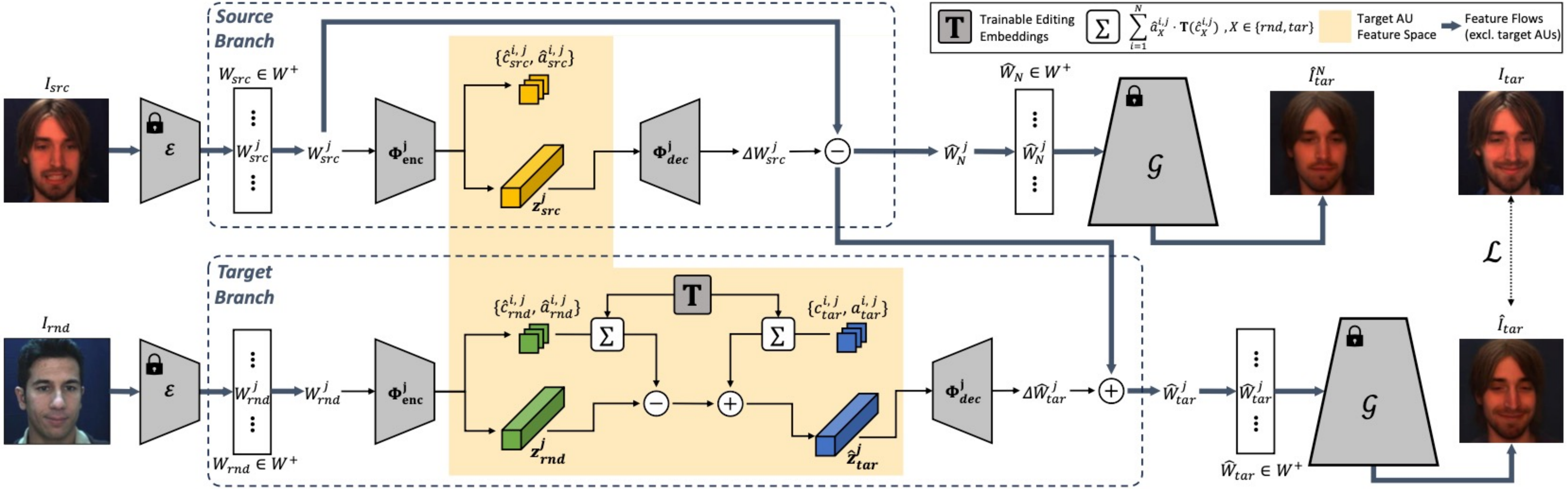
Feature Space for Target AUs



(a) Traditional (b) Style-based generator



Feature Space for Target AUs

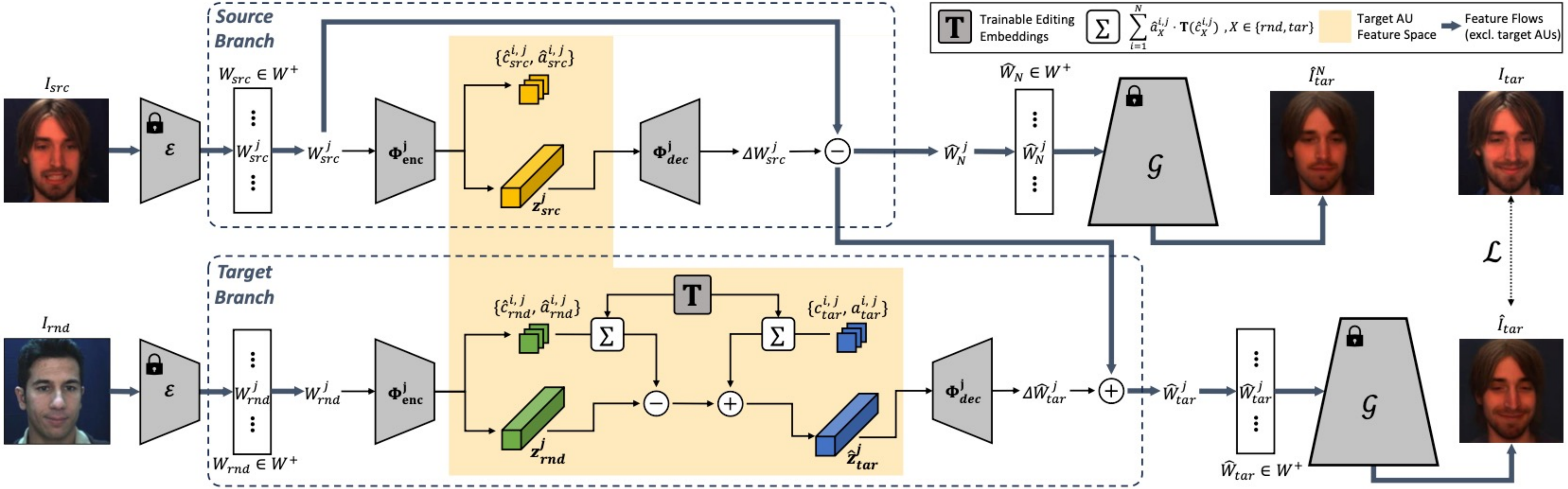


$$\Phi^{j}_{enc}(W^{j}_{src}) = \left\{\hat{c}^{i,j}_{src}, \hat{a}^{i,j}_{src}, z^{j}_{src}\right\}, i \in [1, N], \qquad (1)$$

$$z^{j}_{N} = z^{j}_{src} - \sum_{i=1}^{N} \hat{a}^{i,j}_{src} \cdot \mathbf{T}(\hat{c}^{i,j}_{src})$$

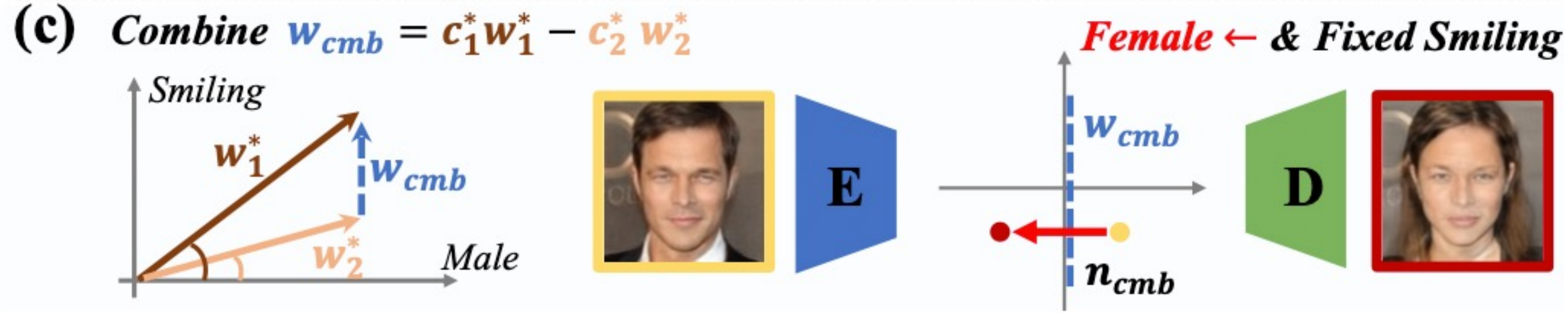
$$\hat{z}^{j}_{tar} = z^{j}_{rnd} - \sum_{i=1}^{N} \hat{a}^{i,j}_{rnd} \cdot \mathbf{T}(\hat{c}^{i,j}_{rnd}) + \sum_{i=1}^{N} a^{i,j}_{tar} \cdot \mathbf{T}(c^{i,j}_{tar})$$

Source Latent Vectors Editing

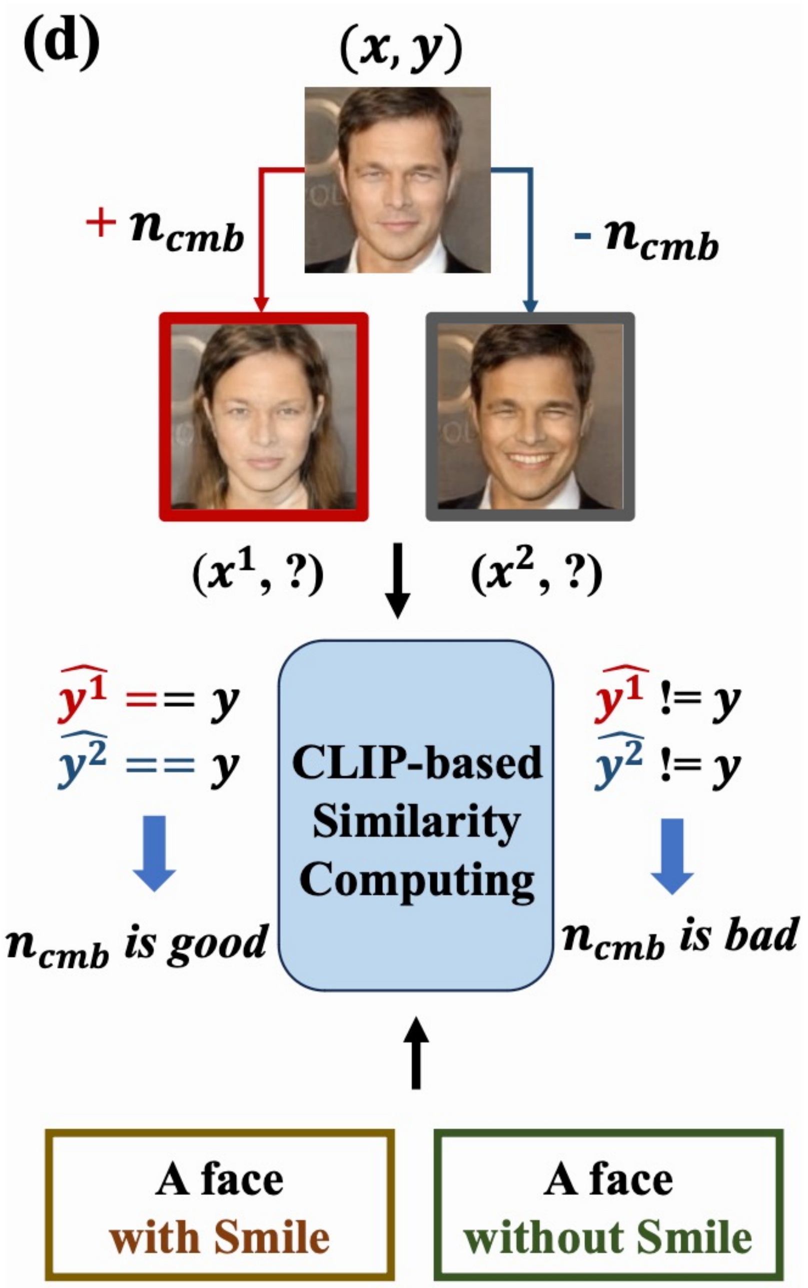


$$
\begin{cases} \hat{W}_N^j = W_{src}^j - \Delta W_{src}^j, \\ \hat{W}_{tar}^j = \hat{W}_N^j + \Delta \hat{W}_{tar}^j, \end{cases} \qquad (2)
$$

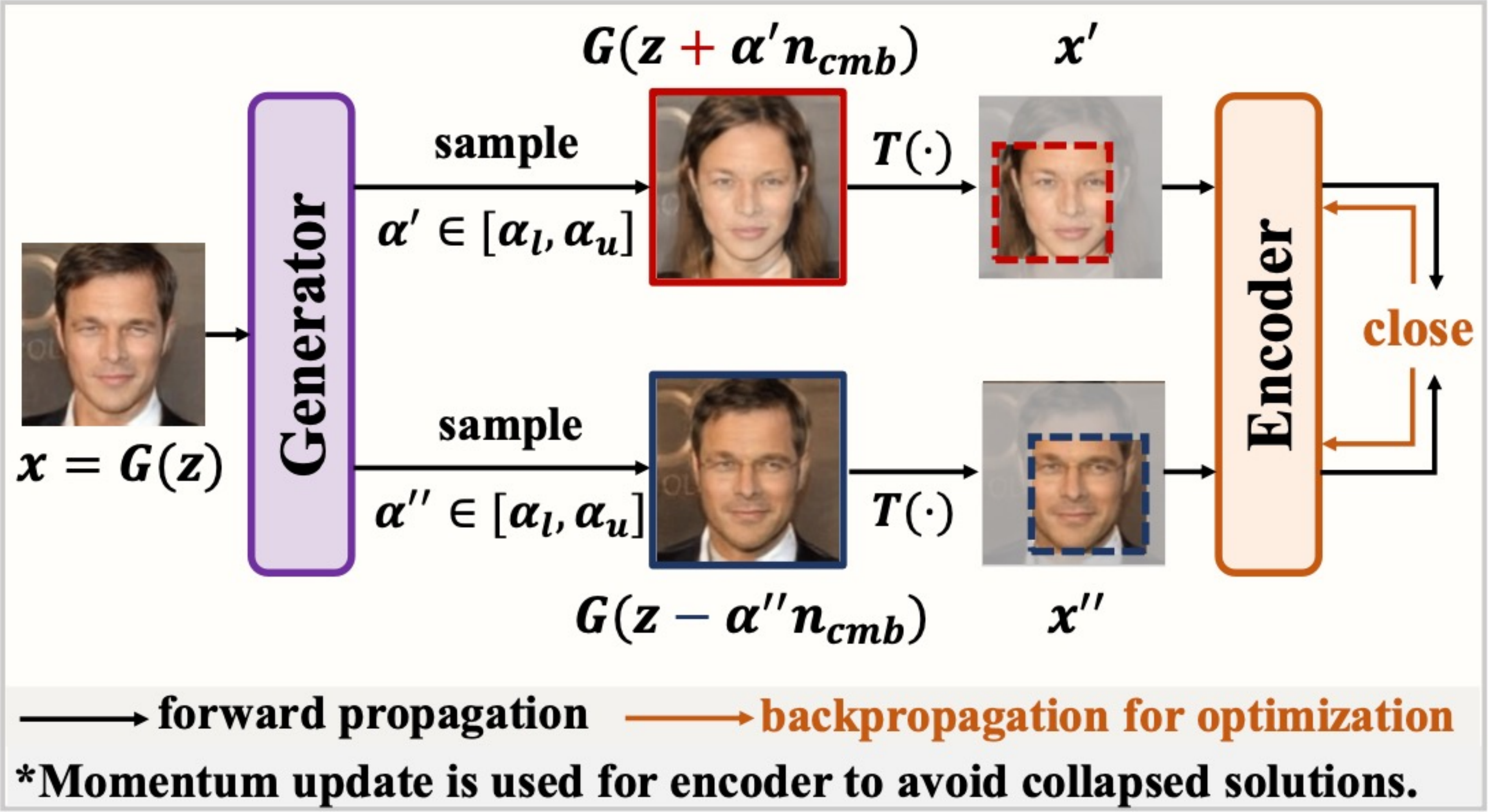
Bias Detection via Generative Modeling



- Using different regularization strengths, we can obtain two different biased semantic directions of target attribute
- Combine these two biased directions by some appropriate combination coefficients



Bias Mitigation via Generative Augmentation



Outline

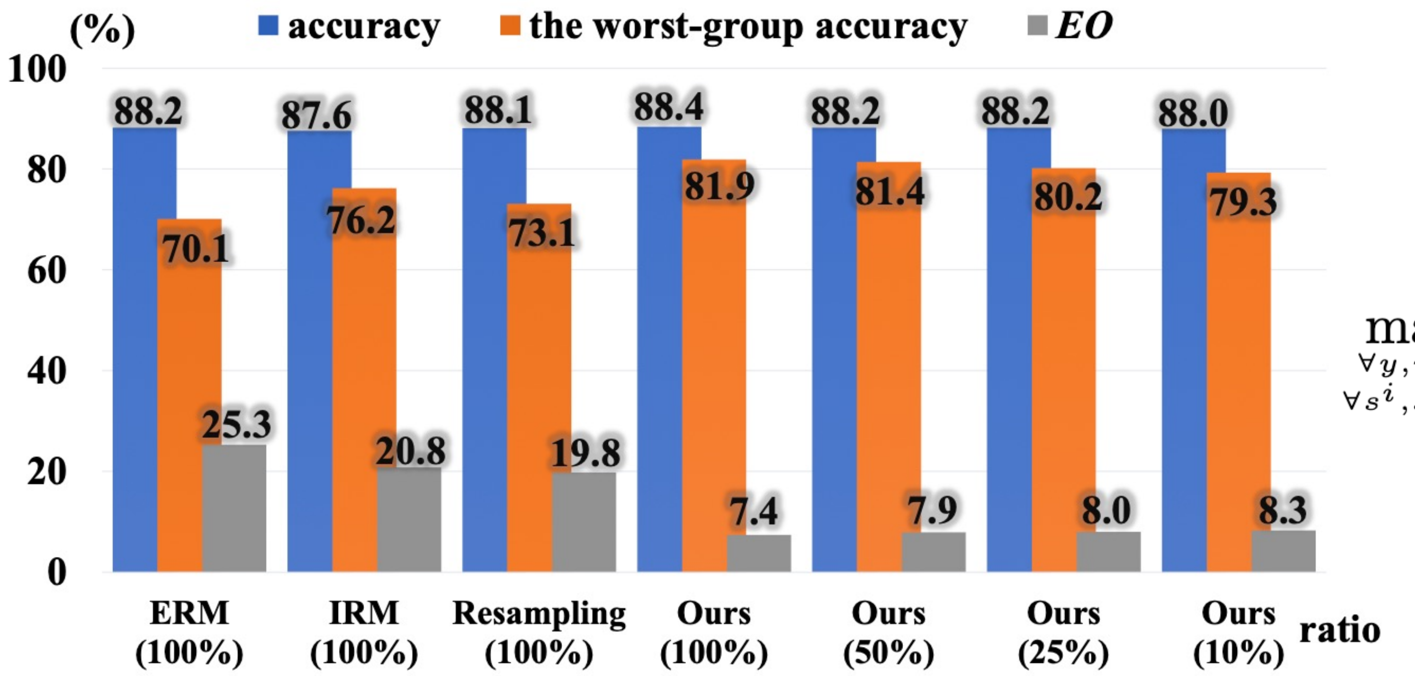
- Introduction
- Framework
- Method
- Experiment
- Conclusion

Comparison of AU intensity manipulation

Source | e4e Inversion | AUEditNet (−) | AUEditNet (+) | ReDirTrans (−) | ReDirTrans (+) | DeltaEdit (+Δ) | Target

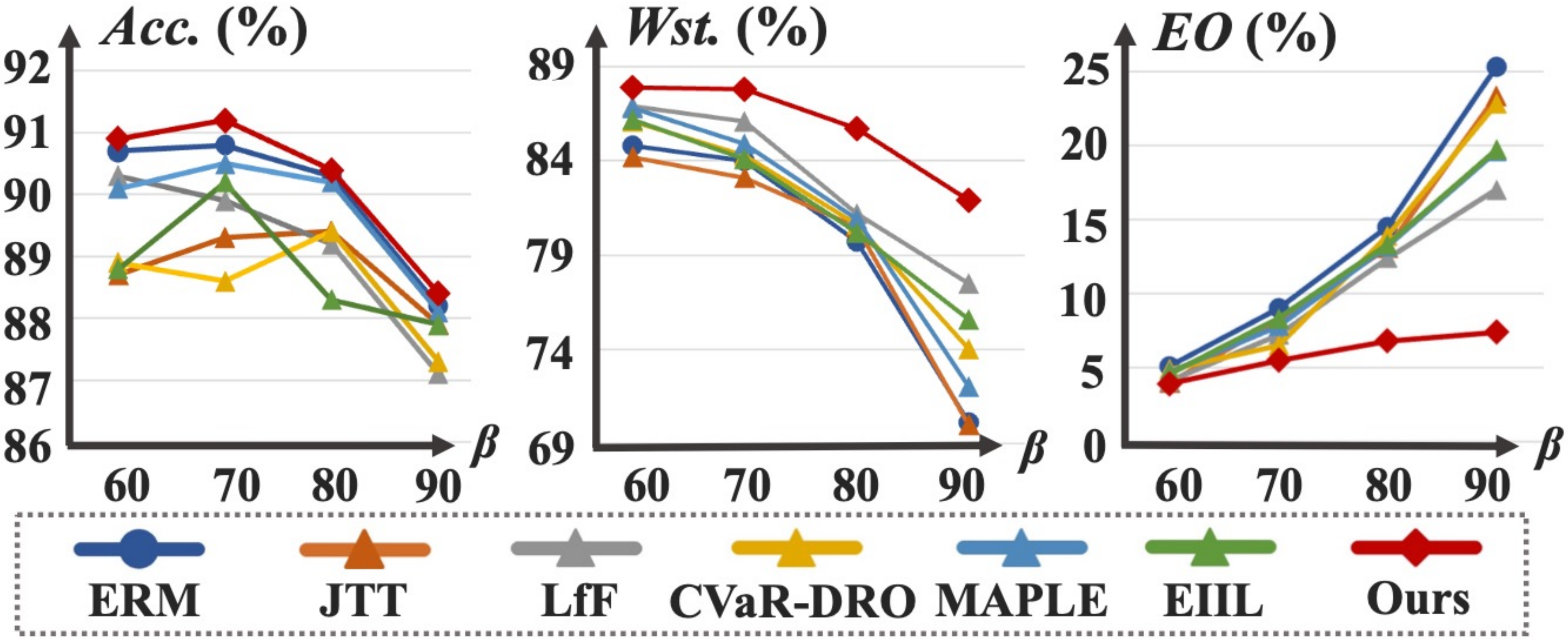
Classification results on facial datasets

Method	T=*s* / S=*m*			T=*s* / S=*y*			T=*b* / S=*m*			T=*a* / S=*y*			T=*m* / S=*y*			T=*y* / S=*m*			T=*b*&*a*&*r* / S=*m*			T=*s* / S=*m*&*y*			T=*g* / S=*e*		
	Acc.	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*	*Acc.*	*Wst.*	*EO*
ERM [23]	88.2	70.1	25.3	88.3	71.5	15.6	84.2	73.3	17.1	82.8	70.1	19.4	97.2	92.8	5.4	77.7	42.0	52.0	90.6	69.3	24.1	87.3	60.4	33.8	91.4	83.5	12.2
CVaR DRO [38]	87.3	74.0	22.8	87.0	76.1	13.9	84.0	73.9	15.5	81.4	71.8	15.2	96.5	93.0	5.3	75.4	42.3	48.8	90.0	71.8	22.0	86.3	64.0	28.4	90.6	84.5	11.9
EIIL [8]	87.9	75.6	19.7	87.9	72.5	13.3	84.1	73.9	15.7	81.9	73.3	14.4	96.2	93.3	4.9	77.5	45.6	39.2	90.4	71.5	22.0	86.4	60.8	19.7	89.2	84.3	8.3
LfF [52]	87.1	77.5	17.0	85.3	72.9	14.3	84.0	74.0	15.1	82.4	72.5	14.2	97.1	92.9	5.1	77.4	44.2	43.6	89.8	70.8	20.5	85.0	62.5	26.6	86.7	84.6	11.1
JTT [42]	88.0	74.8	19.4	87.6	73.3	14.2	83.9	74.1	16.7	81.1	71.1	16.6	97.0	92.4	5.8	76.3	43.6	47.7	88.3	69.1	23.3	87.3	61.0	31.0	90.5	85.0	10.4
MAPLE [85]	88.1	72.0	19.6	88.1	73.6	13.6	83.7	73.9	14.7	82.4	74.7	13.8	97.1	92.9	4.8	76.3	46.2	43.5	89.9	72.8	18.6	86.0	64.8	31.2	89.4	85.3	9.4
DiGA (ours)	**88.4**	**81.9**	**7.4**	**89.1**	**78.5**	**9.5**	**84.5**	**74.5**	**13.5**	**83.6**	**78.6**	**10.8**	**97.4**	**94.8**	**4.3**	**80.0**	**51.3**	**33.3**	**90.7**	**79.7**	**15.8**	**88.4**	**75.8**	**15.6**	**92.7**	**89.0**	**6.8**

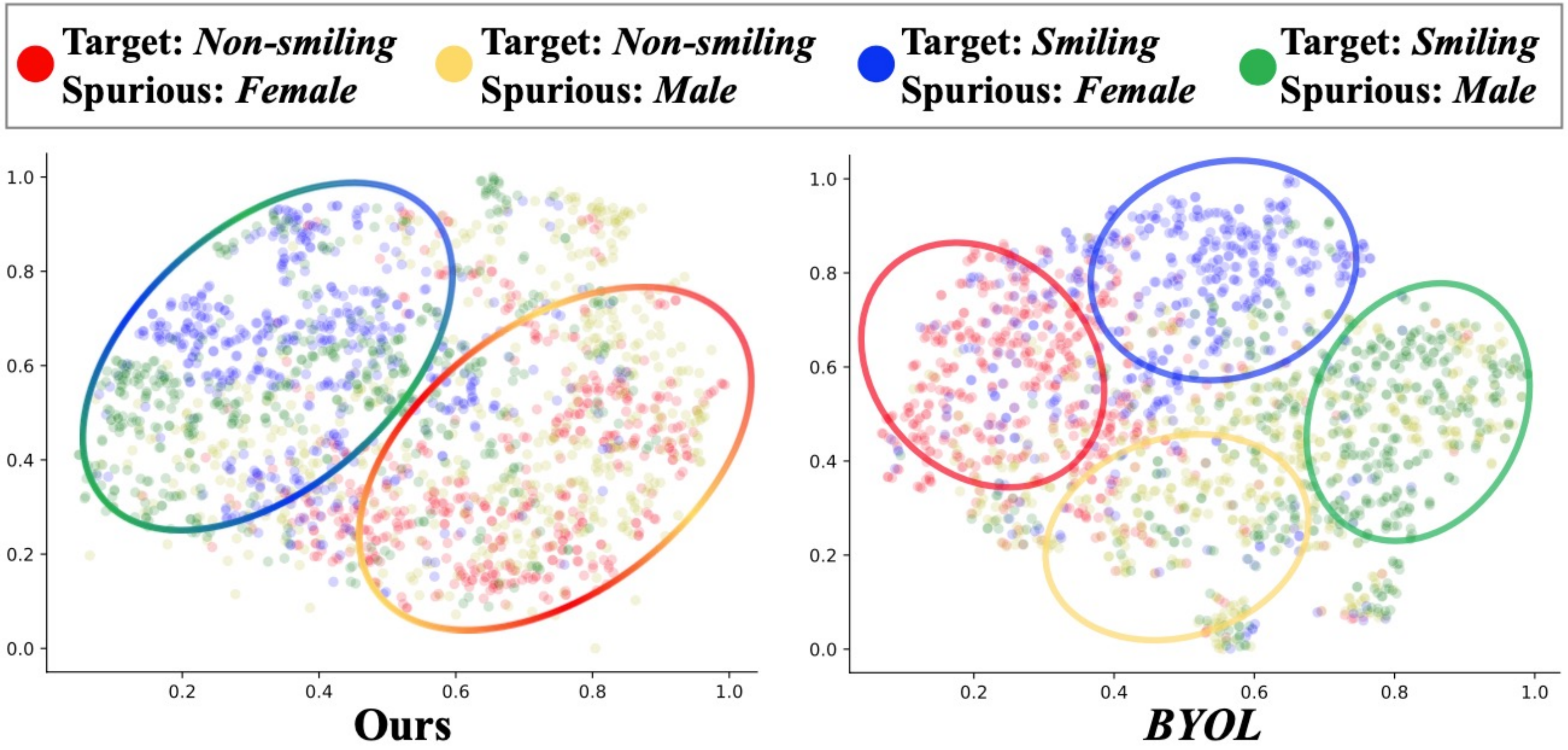


$$\max_{\substack{\forall y,\hat{y}\in\mathcal{Y}\\ \forall s^i,s^j\in\mathcal{S}}} \left|P_{s^i}(\hat{Y}=\hat{y}\mid Y=y) - P_{s^j}(\hat{Y}=\hat{y}\mid Y=y)\right|,$$

Classification results on CelebA dataset under different of data bias



Ablation Studies



- T-SNE visualization for the learned representations on CelebA

Ablation Studies

	Acc.	*Wst.*	*EO*
λ_1=2e-4 λ_2=5e+3	88.3	82.7	9.3
λ_1=1e-4 λ_2=1e+4	88.4	81.9	7.4
λ_1=2e-5 λ_2=5e+4	88.8	85.1	4.8
λ_1=1e-6 λ_2=1e+6	88.8	82.3	7.4

	T=*s* / S=*c*		
Method	*Acc.*	*Wst.*	*EO*
ERM	87.5	67.8	26.1
CVaR DRO	86.6	72.9	22.1
EIIL	86.2	71.3	22.5
LfF	86.9	75.5	19.4
JTT	87.3	72.9	20.1
MAPLE	87.4	73.7	23.8
DiGA (ours)	**88.4**	**81.1**	**7.8**

- Ablation studies of regularization strength $\lambda 1$, $\lambda 2$ on CelebA
- Classification results on non-facial dataset Dogs and Cats

Results on the Cross-Domain Benchmark.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP-ResNet-50	16.11	87.26	55.89	40.37	25.79	62.77	74.82	82.97	60.85	59.48	56.63
CoOp	15.12	86.53	55.32	37.29	26.20	61.55	75.59	87.00	58.15	59.05	56.18
CoCoOp	14.61	87.38	56.22	38.53	28.73	65.57	76.20	**88.39**	59.61	57.10	57.23
TPT	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT	17.60	86.89	**60.71**	40.72	41.04	63.53	**79.21**	83.40	**62.72**	62.67	59.85
TDA (Ours)	**17.61**	**89.70**	57.78	**43.74**	**42.11**	**68.74**	77.75	86.18	62.53	**64.18**	**61.03**
CLIP-ViT-B/16	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
CoCoOp	22.29	93.79	64.90	45.45	39.23	70.85	83.97	**90.46**	66.89	68.44	64.63
TPT	24.78	94.16	66.87	**47.75**	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT	**25.60**	92.49	67.01	47.00	43.13	70.10	**87.23**	88.22	65.74	62.67	65.47
TDA (Ours)	23.91	**94.24**	**67.28**	47.40	**58.00**	**71.42**	86.14	88.63	**67.62**	**70.66**	**67.53**

- Comprehensive evaluation of the model's adaptability during test time across various class spaces

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Proposed a **generation-based** two-stage framework to train a fair FAC model on **biased data without additional annotations**.

- In the first stage, **detecting the spurious attributes** via generative models. This method enhances interpretability by explicitly representing the spurious attributes in the image space.

- In the second stage, for each image, first **edit its spurious attributes**, where the editing degree follows a **uniform distribution**. Then training a fair **FAC model by promoting its invariance to these augmentation**.