Text-to-Image Synthesis for Domain Generalization in Face Anti-Spoofing

Supplementary Material

A. Text prompt analysis

A.1. Text prompts

Tables A, B and C detail the textual prompts related to facial attributes, lighting conditions, and spoofing attacks used in our study. In subsequent sections we explore the effects of textual prompts with various spoofing attacks, aiming to understand their impact more comprehensively. Following this, we delve deeper into the significance of facial attributes and lighting conditions, building on the preliminary analysis presented in Table 7 from the main paper.

A.2. Spoof attacks

In Table D, we assess the significance of incorporating diverse spoof attack textual prompts. We compare the performance of our SpoofFusion approach, which employs textual guidance for facial attributes, lighting conditions, and spoof attacks, against a variant of our method without spoof attack prompts, labeled as *SpoofFusion w/o Spoof Attacks*. The findings underscore the crucial role of spoof attack textual guidance, particularly given that the training sets have only replay and print attacks, whereas the test datasets encompass a broader array of spoofing techniques.

In Table E, we assess the role of textual guidance in training solely with synthetic images. We observe that training with a complete set of synthetic images yields an impressive average HTER of 2.10, outperforming the results of training with only base images. Conversely, omitting textual guidance results in a significantly higher average HTER of 4.92. This finding underscores the critical value of textual guidance in enhancing model performance.

A.3. Facial attributes

We analyze the facial attributes in Table C as textual guidance for generation of the synthetic images. First, we categorize the attributes based on whether they affect human identification or not. According to [22], there are controllable and uncontrollable attributes for de-identification.

In Table G, we examine the influence of facial attributes as text guidance. Our analysis reveals that the controllable facial attributes enhance performance better compared to the uncontrollable ones. This can be attributed to the fact that controllable attributes tend to yield a greater diversity in facial identities. Furthermore, we observe that a combination of both controllable and uncontrollable attributes surpasses the performance of using either category in isolation. This finding underscores the key insight: greater diversity in images, augmented by varied text prompts, leads to improved overall performance.

A.4. Lightning conditions

To assess the effect of different lighting conditions, as detailed in Table B, we conduct a comparative performance analysis presented in Table H. Among the three conditions—Dim, Bright, and One Side—Dim lighting yielded the best average Half Total Error Rate (HTER). This could be attributed to the limited number of facial images under dim lighting in the base dataset, making the synthesis of such images particularly beneficial. Although Bright and One Side lighting also contributed to performance improvements, a pattern emerges, similar to the findings in Table G: integrating all lighting conditions into a single training regimen proved more effective than individual condition training. Thus, both in facial attributes and lighting conditions, the creation of diverse images using a variety of text prompts emerges as a crucial factor for enhancing Face Anti-Spoofing performance.

B. Failure Analysis

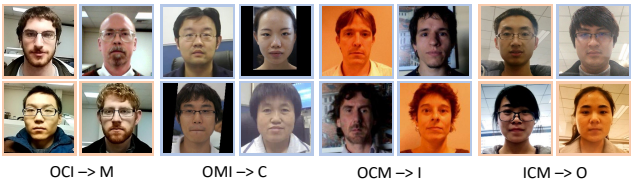

Figure A. Mis-classified examples. Blue boxes show real faces mis-classified as spoof and orange boxes show the reverse.

The majority of errors in OCI $\rightarrow$ M and ICM $\rightarrow$ O involved misidentifying spoof faces as real. Conversely, in the other scenarios, real faces were often misclassified as spoof, largely due to black padding and varying lighting conditions.

Attack type	Description
Replay Attack	A cropped face shown through a digital screen to simulate a replay attack
Full Mask	A cropped face wearing a realistic full-face mask for a disguise
Transparent Mask	A cropped face with a barely visible transparent mask altering its features
Paper Mask	A cropped face covered by a paper mask with a printed face on it
Silicone Head	A cropped image of a silicone head mimicking human features for deception
Mannequin	A cropped face of a mannequin, styled to appear almost human
Print Attack	A cropped face holding a printed photo of another face in front of it

Table A. Spoof attack text guidance

Condition	Description
Overhead Lighting	A face lit from above
Side Lighting	A face lit from sideways
Dim Lighting	A face in dim light
Bright Lighting	A face under bright light

Table B. Facial Lighting Conditions

Facial Features			
arched eyebrows	mouth slightly open	straight hair	bags under eyes
mustache	wavy hair	high cheekbones	big lips
no beard	heavy makeup	big nose	oval face
pale skin	smiling	blond hair	pointy nose
eyeglasses	brown hair	receding hairline	sideburns
bushy eyebrows	rosy cheeks	double chin	attractive
chubby	bald	young	

Table C. Facial Features from CelebA dataset [66]

Method	MCIO → CS			MCIO → SM			MCIO → IW			Avg.
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
FLIP-MCL (ICCV' 23)	4.69	98.86	89.14	13.57	93.72	35.41	15.51	91.78	60.36	11.26
SpoofFusion w/o Spoof Attacks	4.33	98.95	88.89	11.65	94.91	**37.07**	13.21	93.52	**62.49**	9.73
SpoofFusion (Ours)	**3.60**	**99.24**	**91.62**	**9.47**	**96.08**	31.34	**11.97**	**94.32**	62.13	**8.34**

Table D. Evaluation of cross-domain performance: Collaborative training on CASIA (**C**), Idiap Replay (**I**), MSU-MFSD (**M**), and Oulu-NPU (**O**) databases with evaluation on CelebA-Spoof [66] (**CS**), SiW-Mv2 [17] (**SM**), and Insightface Wild [55] (**IW**).

Prompt	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
All Syn	2.62	99.85	0.67	99.94	2.00	99.81	3.11	99.23	2.10
Syn w/o Text	4.29	98.80	1.33	99.91	9.50	95.87	4.55	98.86	4.92

Table E. Evaluating the impact of textual guidance on training exclusively with synthetic images: A comparison between training using all synthetic images versus synthetic images without textual prompts.

Controllable		Uncontrollable
arched eyebrows	mouth slightly open	bald
bags under eyes	mustache	receding hairline
high cheekbones	big lips	straight hair
no beard	heavy makeup	wavy hair
big nose	oval face	
pale skin	smiling	
blond hair	pointy nose	
eyeglasses	brown hair	
sideburns	bushy eyebrows	
rosy cheeks	double chin	
attractive	chubby	
young		

Table F. Controllable and Uncontrollable Facial Attributes for De-identification [22]

Prompt	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
Controllable	3.10	**99.19**	0.67	99.61	2.85	99.58	**1.53**	**99.81**	2.03
Uncontrollable	2.86	98.29	0.78	99.83	3.90	99.41	1.88	99.65	2.35
Combined	**2.86**	98.20	**0.67**	**99.85**	**2.45**	**99.63**	1.86	99.73	**1.96**

Table G. Impact of facial attributes as text guidance. Controllable and Uncontrollable facial attributes for de-identification as grouped in Table F

Prompt	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
Dim	2.86	98.53	**0.56**	**99.99**	2.50	99.56	2.03	99.71	1.99
Bright	2.86	99.02	0.67	99.70	3.00	99.57	**1.86**	**99.69**	2.09
One side	3.10	99.07	0.67	99.92	3.10	99.30	1.89	99.78	2.19
Combined	**2.62**	**99.57**	0.67	99.55	**1.90**	**99.79**	2.02	99.73	**1.80**

Table H. Impact of light conditions as text guidance. Dim is ‘a face in dim light’, Bright is ‘A face under bright light’, One side is lighting from one side, which contains ‘A face lit from above’ and ‘A face lit from sideways’ .