Appendix for "Improving Fairness in Deepfake Detection"

This Appendix provides proof of the proposed methods, additional experimental details and results. Specifically, Sections A, B, and C provide proof and details of the proposed methods. Section D provides details of our experiment, including parameter setting and source code. Section E provides further analysis of additional experimental results, including optimization by different metric (in E.1), effect of choosing different hyperparameters (in E.2), performance on Cross-domain Dataset (in E.3), convergence analysis of the proposed methods (in E.4), more comparison results (in E.5 and E.6), and details on datasets and results of each subgroup (in E.7 and E.8).

A. Proofs

A.1. Proof of Proposition 1

Proof. For any m, denote $Z = (X, Y)$ and $\mathcal{D}_m$ as a set that contains samples from m-th group, then $\mathbb{P}(Z) = \pi_m \mathbb{P}(Z|\mathcal{D}_m) + (1-\pi_m)\mathbb{P}(Z|\overline{\mathcal{D}_m})$, where $\overline{\mathcal{D}_m}$ contains samples are not in $\mathcal{D}_m$. Let $\mathbb{Q}(Z) = \mathbb{P}(Z|\mathcal{D}_m)$ and $\mathbb{Q}'(Z) = \frac{\pi_m - \alpha}{1-\alpha}\mathbb{P}(Z|\mathcal{D}_m) + \frac{1-\pi_m}{1-\alpha}\mathbb{P}(Z|\overline{\mathcal{D}_m})$. Then $\mathbb{P}(Z) = \alpha\mathbb{Q}(Z) + (1-\alpha)\mathbb{Q}'(Z)$, which implies that

$$\alpha\mathbb{E}_{\mathbb{Q}(Z)}[\ell(\theta; Z) - \lambda] = \mathbb{E}_{\alpha\mathbb{Q}(Z)}[\ell(\theta; Z) - \lambda] \leq \mathbb{E}_{\alpha\mathbb{Q}(Z)}[[\ell(\theta; Z) - \lambda]_+] \leq \mathbb{E}_{\mathbb{P}(Z)}[[\ell(\theta; Z) - \lambda]_+].$$

The last inequality holds because $\alpha \leq \min_{m=1,\ldots,K} \pi_m$ and $\alpha \in (0, 1)$, which means $\mathbb{Q}'(Z) \geq 0$ and therefore $\mathbb{P}(Z) \geq \alpha\mathbb{Q}(Z)$. From the above inequations, we obtain

$$\begin{aligned}
&\alpha\mathbb{E}_{\mathbb{Q}(Z)}[\ell(\theta; Z) - \lambda] \leq \mathbb{E}_{\mathbb{P}(Z)}[[\ell(\theta; Z) - \lambda]_+] \\
\Rightarrow\ &\mathbb{E}_{\mathbb{Q}(Z)}[\ell(\theta; Z) - \lambda] \leq \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}(Z)}[[\ell(\theta; Z) - \lambda]_+] \\
\Rightarrow\ &\mathbb{E}_{\mathbb{Q}(Z)}[\ell(\theta; Z)] \leq \lambda + \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}(Z)}[[\ell(\theta; Z) - \lambda]_+] = \text{CVaR}_\alpha(\theta)
\end{aligned}$$

In Section 3.1, we have already defined $\mathbb{P}_m$, which is just $\mathbb{Q}(Z)$. Therefore, we have $\mathcal{R}_{\max}(\theta) \leq \text{CVaR}_\alpha(\theta)$.

□

A.2. Proof of Theorem 1

Proof. 1) We first prove that $\frac{1}{k}\sum_{i=1}^{k} \ell_{[i]} = \min_{\lambda\in\mathbb{R}} \lambda + \frac{1}{k}\sum_{i=1}^{q}[\ell_i - \lambda]_+$.

$\Rightarrow$: Suppose $\overline{\ell} := \{\ell_1, \ldots \ell_q\}$. We know $\sum_{i=1}^{k} \ell_{[i]}$ is the solution of

$$\max_{\mathbf{p}} \mathbf{p}^\top \overline{\ell},\ \text{s.t. } \mathbf{p}^\top \mathbf{1} = k, \mathbf{0} \leq \mathbf{p} \leq \mathbf{1}.$$

We apply Lagrangian to this equation and get

$$\mathcal{L} = -\mathbf{p}^\top \overline{\ell} - \mathbf{v}^\top \mathbf{p} + \mathbf{u}^\top(\mathbf{p} - 1) + \lambda(\mathbf{p}^\top \mathbf{1} - k)$$

where $\mathbf{u} \geq \mathbf{0}$, $\mathbf{v} \geq \mathbf{0}$ and $\lambda \in \mathbb{R}$ are Lagrangian multipliers. Taking its derivative with respect to $\mathbf{p}$ and set it to 0, we have $\mathbf{v} = \mathbf{u} - \overline{\ell} + \lambda\mathbf{1}$. Substituting it back into the Lagrangian, we get

$$\min_{\mathbf{u},\lambda} \mathbf{u}^\top \mathbf{1} + k\lambda,\ \text{s.t. } \mathbf{u} \geq \mathbf{0}, \mathbf{u} + \lambda\mathbf{1} - \overline{\ell} \geq 0.$$

This means

$$\sum_{i=1}^{k} \ell_{[i]} = \min_{\lambda} k\lambda + \sum_{i=1}^{q}[\ell_i - \lambda]_+.$$

Therefore,

$$\frac{1}{k}\sum_{i=1}^{k} \ell_{[i]} = \min_{\lambda} \lambda + \frac{1}{k}\sum_{i=1}^{q}[\ell_i - \lambda]_+. \tag{A.1}$$

$\Leftarrow$: Denote $\overline{\mathcal{L}} := \lambda + \frac{1}{k}\sum_{i=1}^{q}[\ell_i - \lambda]_+$. Since $\overline{\mathcal{L}}$ is a convex function with respect to λ, we can set the $\partial_\lambda \overline{\mathcal{L}} = 0$ to get the optimal value of λ^*. Thus, we have $\partial_\lambda \overline{\mathcal{L}} = 1 - \frac{1}{k}\sum_{i=1}^{q} \mathbb{1}_{[\ell_i \geq \lambda^*]} = 0$, then $\lambda^* = \ell_{[k]}$ can be an optimal value. Taking $\lambda^* = \ell_{[k]}$ into $\overline{\mathcal{L}}$, we obtain $\overline{\mathcal{L}} = \frac{1}{k}\sum_{i=1}^{k} \ell_{[i]}$.

Based on the above analysis, we get $\frac{1}{k}\sum_{i=1}^{k} \ell_{[i]} = \min_{\lambda\in\mathbb{R}}\{\lambda + \frac{1}{k}\sum_{i=1}^{q}[\ell_i - \lambda]_+\}$.

2) Using the above result, we can directly replace $\mathcal{L}_g(\theta) = \frac{1}{k_g}\sum_{j=1}^{k_g} \ell^g_{[j]}(\theta)$ from (8) with $\mathcal{L}_g(\theta) = \min_{\lambda_g\in\mathbb{R}}\{\lambda_g + \frac{1}{\alpha_g n_g}\sum_{i\in\mathcal{I}_g}[\ell(\theta; X_i, Y_i) - \lambda_g]_+\}$. This is also shown in (9).

□

B. Pseudocode of the DAW-FDD

Algorithm 2: DAW-FDD

Input: A training dataset $\mathcal{S}$ with demographic variable G, A set of subgroups $\mathcal{G}$, α, α_g, max_iterations, num_batch, η
Output: A fair deepfake detection model with parameters θ^*

Initialization: θ_0, $l = 0$
for $e = 1$ *to* max_iterations **do**
  for $b = 1$ *to* num_batch **do**
    Sample a mini-batch $\mathcal{S}_b$ from $\mathcal{S}$
    Compute $\ell(\theta_l; X_i, Y_i), \forall (X_i, Y_i) \in \mathcal{S}_b$
    For each $g \in \{1, ..., |\mathcal{G}|\}$, set λ_g^* to be the value of λ_g that minimizes $\mathcal{L}_g(\theta, \lambda_g)$ as given in (9b). This minimization is solved using binary search.
    Set $L_g(\theta) \leftarrow L_g(\theta, \lambda_g^*)$ using (9b), $\forall g$
    Using binary search to find λ that minimizes (9a)
    Set $\theta_{l+1} \leftarrow \theta_l - \eta \partial_\theta \mathcal{L}_{\text{DAW-FDD}}(\theta_l, \lambda)$
    $l \leftarrow l + 1$
  end
end
return $\theta^* \leftarrow \theta_l$

C. Explicit Forms of (sub) gradients

From equation (9), we have

$$\min_{\theta\in\Theta,\lambda\in\mathbb{R}} \mathcal{L}_{\textit{DAW-FDD}}(\theta, \lambda) := \lambda + \frac{1}{\alpha|\mathcal{G}|}\sum_{g\in\mathcal{G}}[\mathcal{L}_g(\theta) - \lambda]_+,$$

$$\textit{s.t. } \mathcal{L}_g(\theta) = \min_{\lambda_g\in\mathbb{R}} \mathcal{L}_g(\theta, \lambda_g) := \lambda_g + \frac{1}{\alpha_g n_g}\sum_{i\in\mathcal{I}_g}[\ell(\theta; X_i, Y_i) - \lambda_g]_+.$$

We can get

$$\partial_\theta \mathcal{L}_{\text{DAW-FDD}}(\theta, \lambda) = \frac{1}{\alpha|\mathcal{G}|}\sum_{g\in\mathcal{G}}\left[\left(\frac{1}{\alpha_g n_g}\sum_{i\in\mathcal{I}_g}\partial_\theta \ell(\theta; X_i, Y_i)\cdot \mathbb{1}_{[\ell(\theta;X_i,Y_i)>\lambda_i]}\right)\cdot \mathbb{1}_{[\mathcal{L}_g(\theta)>\lambda]}\right]$$

D. Additional Experimental Details

D.1. α and α_g Settings on Each Dataset

We tune α and α_g on the following hyperparameter grid: 0.1, 0.3, 0.5, 0.7, 0.9. We provide a reference for setting α and α_g to reproduce our experimental results in Table D.1.

Parameter	Xception				ResNet-50	EfficientNet-B3	DSP-FWA	RECCE
	FF++	Celeb-DF	DFD	DFDC	FF++	FF++	FF++	FF++
α in DAG-FDD	0.5	0.5	0.3	0.7	0.5	0.5	0.5	0.5
α, α_g in DAW-FDD	0.5, 0.9	0.5,0.7	0.7, 0.9	0.5, 0.7	0.5, 0.9	0.5, 0.9	0.7, 0.9	0.5, 0.9

Table D.1. *Hyperparameter settings of DAG-FDD and DAW-FDD.*

D.2. Trade-off Parameters for $Cons.$ EFPR and $Cons.$ EO

For $Cons.$ EFPR and $Cons.$ EO baselines, we tune the trade-off hyperparameters on the following grid: 0.5, 0.6, 0.7, 0.8, 0.9. Finally, we use 0.6 for both methods since this hyperparameter can return the best performance.

E. Additional Experimental Results

E.1. Optimization by Metric F_{EO}

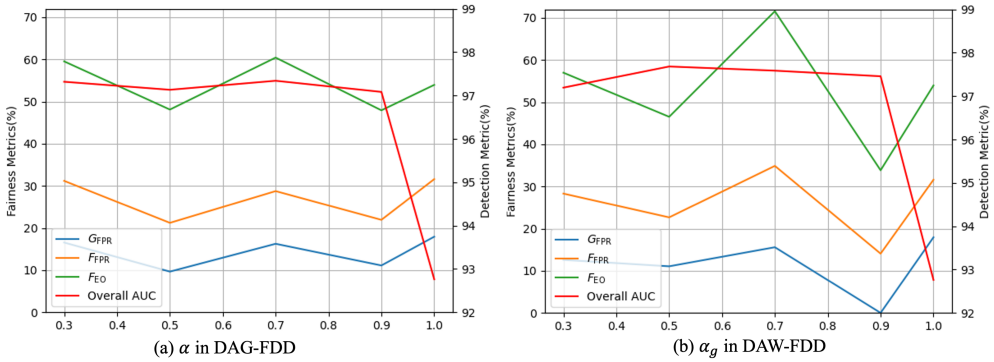
We employ F_{EO} as an index to tune the hyperparameter and report the results in Tabel E.1. The results illustrate that optimizing hyperparameters using F_{EO} can improve TPR and F_{EO} (compared with results in Table 4), which demonstrates that our method can generalize to different metric.

Methods	Fairness Metrics (%) ↓				Detection Metrics (%)			
	Intersection				Overall			
	G_{AUC}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Original	8.53	11.81	15.66	39.95	97.17	13.01	**95.83**	**94.05**
DAG-FDD (Ours)	**5.86**	7.04	8.23	**29.65**	98.50	5.06	93.48	93.78
DAW-FDD (Ours)	6.67	**2.96**	**3.96**	30.52	**98.81**	**2.78**	91.99	93.04

Table E.1. *Test set results of Xception on the Celeb-DF dataset, optimized by F_{EO} metric.*

E.2. Effect of α and α_g

Fig. E.1 shows the fairness metrics and performance metric AUC to different α and α_g values in DAG-FDD and DAW-FDD methods, respectively, when using Xception as backbone in FF++ dataset. Experiment result in Fig. E.1 (a) demonstrates that the model achieves the best fairness performance when setting α as 0.5 in DAG-FDD and also keeps fair AUC score. In FAW-FDD, we set α as 0.5 selected from the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ based on the best fairness performance first. Secondly, we searched for the optimal value of α_g in the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ while keeping α fixed at its optimal value. Fig. E.1 (b) shows that the proposed DAW-FDD performs best when α_g is set to 0.9 when α is fixed on 0.5.



(a) α in DAG-FDD (b) α_g in DAW-FDD

Figure E.1. *Parameters of DAG-FDD and DAW-FDD on FF++ dataset with Xception.*

E.3. Performance on Cross-domain Dataset

We further evaluate the performance of our methods using Xception on cross-domain dataset. The models are trained on FF++ dataset and tested on DFDC dataset. The results are presented in Table E.2.

Methods	Fairness Metrics (%) ↓				Detection Metrics (%)			
	Intersection				Overall			
	G_{AUC}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Original	33.76	17.19	30.70	122.51	58.81	59.54	71.60	51.57
DAG-FDD (Ours)	25.42	24.16	49.27	117.19	56.32	35.29	47.06	58.41
DAW-FDD (Ours)	26.96	21.50	45.34	119.32	59.95	43.70	60.69	57.87

Table E.2. *Cross-domain Performance. Models are trained on FF++ and tested on DFDC.*

E.4. Convergence of the Proposed Loss Functions

We also show the training loss convergence of our methods when applying to Xception on FF++ dataset in Fig. E.2. The results show that our methods can converge within reasonable epochs.

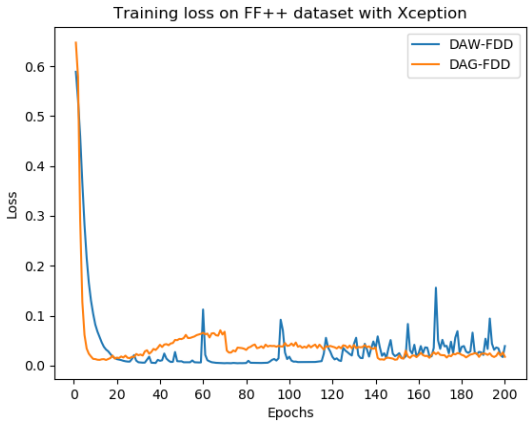


Figure E.2. *Training loss convergence.*

Methods	Require Demo-graphics	Fairness Metrics (%) ↓									Detection Metrics (%)			
		Gender			Race			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Original	–	0.87	0.87	**3.14**	18.81	27.65	30.07	30.26	67.38	80.34	98.05	21.20	**98.21**	94.74
DRO_{χ^2} [30]	×	1.46	1.46	4.17	13.87	20.05	23.15	23.63	42.14	57.29	98.32	15.65	97.28	94.97
DAG-FDD (Ours)	×	0.55	0.55	3.71	12.68	17.41	20.33	15.40	36.17	54.24	98.33	12.01	96.80	**95.23**
Naive [16]	✓	9.48	9.48	13.05	18.26	20.86	22.27	28.74	73.59	89.87	93.64	27.57	95.96	91.76
FRM [22]	✓	2.15	2.15	5.48	8.50	10.00	13.75	14.88	30.59	**49.86**	98.06	15.21	97.05	94.86
Group DRO [59]	✓	0.74	0.74	3.71	12.08	16.26	20.01	15.17	32.95	51.08	98.22	11.75	96.59	95.10
Cons. EFPR [60]	✓	6.13	6.13	11.15	10.71	15.00	19.46	13.67	38.48	63.80	97.17	14.72	96.29	94.32
DAW-FDD (Ours)	✓	**0.25**	**0.25**	4.75	**6.99**	**7.96**	**11.95**	**13.54**	**23.44**	52.95	**98.35**	**8.15**	94.59	94.10

Table E.3. *Comparison results with different fairness solutions using RECCE Deepfake detector on FF++ testing set across Gender, Race, and Intersection groups. The best results are shown in* ***Bold****. ↑ means higher is better and ↓ means lower is better. Gray highlights mean our methods outperform the baselines in the group (i.e., DAG-FDD vs. Original/*DRO_{χ^2}*, DAW-FDD vs. Original/Naive/FRM/Group DRO/Cons. EFPR).*

E.5. Comparison on SOTA Deepfake Detector

Since the RECCE model achieves SOTA detection performance on several datasets, we apply our methods and baselines based on the RECCE detector and show the results in Table E.3. The results demonstrate the adaptability and efficiency of our methods.

E.6. Results on DF-Platter Dataset

We apply our methods and baseline to the Xception network on a recent Deepfake dataset with demographic annotations, namely DF-Platter, to further illustrate the effectiveness of our methods. We mainly consider Gender (Male and Female) and Age (Young Adult, Adult, Old) attributes based on the official annotations. In addition to the single attribute fairness, we also consider the combined attributes (Intersection) group, including Male-Young Adult (M-Y), Male-Adult (M-A), Male-Old (M-O), Female-Young Adult (F-Y), Female-Adult (F-A), and Female-Old (F-O). We train and evaluate our methods on a subset of the DF-Platter dataset consisting of real and FSGAN-generated data from Set A with C23 compression, and use Dlib [61] for face extraction and alignment. The cropped faces are resized to 380×380 for training and testing. Training/validation/test datasets are divided following the official split, without identity overlapping. Experiment results shown in Table E.4 demonstrate that our methods outperform baseline for most metrics.

E.7. Dataset Details

We show the total number of train/val/test samples of each dataset and the attributes included in our experiment in Table E.5. Specifically, the number of training samples within each subgroup for four datasets is shown in Table E.6.

Methods	Fairness Metrics (%) ↓									Detection Metrics (%)			
	Gender			Age			Intersection			Overall			
	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Original	3.70	3.70	3.92	3.43	3.90	5.03	4.96	11.94	14.56	99.93	2.80	99.82	98.54
DAG-FDD (Ours)	3.05	3.05	3.18	3.40	3.29	4.06	4.72	10.35	12.03	**99.97**	2.42	**99.91**	98.77
DAW-FDD (Ours)	**1.95**	**1.95**	**2.13**	**1.97**	**2.17**	**2.96**	**3.27**	**6.81**	**8.81**	**99.97**	**1.75**	99.82	**99.05**

Table E.4. *Comparison results with different fairness solutions using the Xception detector on DF-Platter testing set across Gender, Age, and Intersection groups. The best results are shown in* ***Bold****. ↑ means higher is better and ↓ means lower is better. Gray highlights mean our methods outperform the Original baseline.*

Dataset	# Samples	Sensitive Attributes
FF++	126,956	Gender (Male, Female), Race (White, Black, Asian, Others)
Celeb-DF	143,273	Gender (Male, Female), Race (White, Black, Others)
DFD	40,246	Gender (Male, Female), Race (White, Black, Others)
DFDC	117,065	Gender (Male, Female), Race (White, Black, Asian, Others)

Table E.5. *Sample number and attributes in each dataset.*

Datasets	Gender		Race				Intersection							
	M	F	A	B	W	O	M-A	M-B	M-W	M-O	F-A	F-B	F-W	F-O
FF++	33549	42590	10488	2579	56724	6348	2475	1468	31281	4163	8013	1111	31281	2185
Celeb-DF	87344	6251	-	630	86583	6382	-	600	81194	5550	-	30	5389	832
DFD	16607	7227	-	8121	11911	3802	-	6482	7784	2341	-	1639	4127	1461
DFDC	37911	33567	4059	18909	40257	8253	2144	9603	21755	4409	1915	9306	18502	3844

Table E.6. *Number of training samples of each group in the FF++, Celeb-DF, DFD and DFDC datasets. "-" means group does not exist in the dataset.*

E.8. Detailed Results

Detailed test results of each subgroup on four datasets based on four models are presented in this section. Table E.7 provides comprehensive metrics of each subgroup on the four datasets, while Table E.8 displays details of the four models. These findings align with the results reported in Tables 2, 4, 5 and Figures 1, 2 of the submitted manuscript.

Datasets	Methods	Metric (%)	Gender		Race				Intersection							
			M	F	A	B	W	O	M-A	M-B	M-W	M-O	F-A	F-B	F-W	F-O
FF++	Original	AUC	92.42	93.30	89.33	94.44	92.93	97.01	88.09	95.21	92.47	95.43	90.33	93.42	93.53	99.40
		FPR	19.86	23.95	32.67	24.29	20.10	19.58	25.63	21.74	19.01	18.79	36.72	26.27	21.08	20.53
		TPR	91.84	96.80	94.92	95.66	94.09	96.07	89.13	95.69	91.70	93.43	97.96	95.63	96.35	99.86
		ACC	89.80	93.01	89.55	92.17	91.57	93.49	86.12	93.02	89.83	91.55	91.38	91.30	93.20	96.20
	DAG-FDD (Ours)	AUC	96.59	97.65	96.74	96.76	97.08	98.76	93.20	99.44	96.55	98.34	98.24	94.19	97.60	99.31
		FPR	8.67	10.30	13.65	8.57	9.21	5.42	14.29	9.78	8.06	6.08	13.29	7.63	10.26	4.64
		TPR	91.93	96.51	94.62	95.25	94.28	93.63	88.16	98.43	92.06	91.11	98.02	91.88	96.39	97.25
		ACC	91.82	95.26	93.01	94.58	93.66	93.79	87.66	97.18	92.04	91.55	95.86	91.97	95.19	96.91
	DAW-FDD (Ours)	AUC	96.91	98.05	96.39	97.92	97.54	98.23	94.63	97.81	97.07	97.24	97.35	98.23	98.11	99.22
		FPR	11.29	11.61	12.58	13.33	11.18	10.84	9.24	15.22	11.20	12.71	14.49	11.86	11.17	8.61
		TPR	93.48	97.15	94.40	96.36	95.47	95.77	88.91	96.08	93.92	93.13	97.29	96.67	96.94	99.57
		ACC	92.65	95.55	93.04	94.67	94.29	94.68	89.29	94.35	93.02	92.23	95.05	94.98	95.48	98.10
Celeb-DF	Original	AUC	87.83	98.04	-	91.47	97.40	99.91	-	91.47	-	-	-	-	98.00	100
		FPR	16.47	11.55	-	11.55	13.31	10.00	-	11.55	-	-	-	-	11.81	0.00
		TPR	79.62	96.74	-	79.62	96.74	100	-	79.62	-	-	-	-	96.74	100
		ACC	81.90	95.44	-	82.88	94.89	91.67	-	82.88	-	-	-	-	95.42	100
	DAG-FDD (Ours)	AUC	91.61	98.53	-	92.56	98.28	99.98	-	92.56	-	-	-	-	98.51	100
		FPR	3.84	1.82	-	2.54	2.43	1.33	-	2.54	-	-	-	-	1.86	0.00
		TPR	73.43	88.18	-	73.43	88.16	100	-	73.43	-	-	-	-	88.16	100
		ACC	86.68	89.75	-	82.31	89.90	98.89	-	82.31	-	-	-	-	89.71	100
	DAW-FDD (Ours)	AUC	88.72	98.93	-	91.52	98.38	100	-	91.52	-	-	-	-	98.91	100
		FPR	4.78	0.97	-	3.80	1.90	0.33	-	3.80	-	-	-	-	0.99	0.00
		TPR	70.22	85.33	-	70.22	85.31	100	-	70.22	-	-	-	-	85.31	100
		ACC	84.79	87.49	-	79.81	87.67	99.44	-	79.81	-	-	-	-	87.43	100
DFD	Original	AUC	92.41	93.34	-	95.27	92.12	-	-	94.12	90.85	-	-	98.39	93.10	-
		FPR	23.44	26.39	-	19.48	26.83	-	-	19.65	26.78	-	-	18.18	26.86	-
		TPR	94.57	97.14	-	96.32	95.95	-	-	94.33	94.41	-	-	100	97.25	-
		ACC	88.36	89.68	-	88.37	88.48	-	-	86.22	88.86	-	-	95.48	88.20	-
	DAG-FDD (Ours)	AUC	92.68	93.93	-	94.93	92.89	-	-	93.64	92.26	-	-	98.51	93.58	-
		FPR	26.53	29.44	-	23.51	29.59	-	-	23.75	28.98	-	-	21.59	29.89	-
		TPR	95.26	97.14	-	97.11	96.13	-	-	95.75	94.96	-	-	99.62	97.13	-
		ACC	87.75	88.72	-	86.73	87.70	-	-	84.44	88.69	-	-	94.35	86.99	-
	DAW-FDD (Ours)	AUC	92.38	93.77	-	94.55	92.68	-	-	93.23	91.93	-	-	98.47	93.42	-
		FPR	27.01	28.41	-	25.97	28.34	-	-	26.54	27.43	-	-	21.59	28.79	-
		TPR	94.97	96.71	-	96.84	95.86	-	-	95.55	94.64	-	-	99.25	96.90	-
		ACC	87.39	88.75	-	85.36	87.93	-	-	82.74	88.86	-	-	94.07	87.26	-
DFDC	Original	AUC	91.19	93.41	79.27	94.69	92.24	89.33	66.96	92.61	92.67	86.82	99.77	95.50	91.54	94.58
		FPR	8.04	6.40	9.30	5.28	7.72	8.67	20.96	4.98	6.61	13.02	0.80	5.57	9.09	0.99
		TPR	74.69	77.41	56.68	81.36	76.45	68.15	44.44	71.80	77.80	67.57	93.33	85.14	75.45	68.61
		ACC	86.90	86.83	87.31	90.66	85.72	84.00	73.36	90.34	87.91	82.22	98.94	90.90	83.53	86.44
	DAG-FDD (Ours)	AUC	90.70	94.22	82.44	95.79	92.00	89.73	69.71	94.49	91.18	87.02	99.63	96.29	92.22	95.60
		FPR	7.22	5.91	6.81	3.87	7.60	8.47	15.28	3.49	6.54	12.54	0.64	4.24	8.91	1.28
		TPR	71.97	76.04	52.50	80.92	74.74	62.38	42.22	70.41	74.85	64.36	83.33	85.06	74.67	60.77
		ACC	86.69	86.54	89.14	91.51	85.09	82.32	77.74	91.25	86.92	81.80	98.63	91.70	83.25	83.03
	DAW-FDD (Ours)	AUC	93.30	96.24	88.66	98.23	93.64	93.69	77.89	96.73	93.24	91.43	100	98.97	93.87	97.72
		FPR	5.06	3.34	4.88	1.95	5.43	4.31	11.35	3.04	4.99	6.27	0.16	0.91	5.97	0.85
		TPR	74.11	75.77	54.17	82.59	74.74	65.15	40.00	74.56	76.92	65.35	96.67	85.76	73.14	64.99
		ACC	88.84	87.93	91.05	93.36	86.36	86.04	80.66	92.45	88.65	86.77	99.70	94.03	84.06	85.03

Table E.7. *Detailed test set results of each group in Xception on the FF++, Celeb-DF, DFD and DFDC datasets. '-' means not applicable.*

Models	Methods	Metric (%)	Gender		Race				Intersection							
			M	F	A	B	W	O	M-A	M-B	M-W	M-O	F-A	F-B	F-W	F-O
ResNet-50	Original	AUC	93.54	95.15	92.19	96.38	94.51	96.10	87.06	97.67	94.22	94.04	94.75	96.16	94.92	98.23
		FPR	24.57	27.15	33.28	27.62	24.97	20.48	35.29	20.65	23.16	24.86	32.13	33.05	26.61	15.23
		TPR	94.24	98.30	96.89	97.68	96.17	96.49	93.65	96.67	94.17	94.04	98.59	98.75	98.06	100
		ACC	90.96	93.65	91.01	93.25	92.42	93.69	87.75	94.02	91.15	91.12	92.76	92.48	93.60	97.27
	DAG-FDD (Ours)	AUC	93.50	95.56	92.40	95.21	94.69	96.58	89.36	97.17	93.78	94.82	93.92	94.82	95.70	98.82
		FPR	21.33	23.54	27.76	26.67	21.32	21.69	27.31	17.39	20.20	25.41	28.02	33.90	22.34	17.22
		TPR	93.16	97.32	96.48	95.96	95.03	96.01	93.86	95.88	92.83	93.64	97.85	96.04	97.10	99.42
		ACC	90.64	93.51	91.76	92.00	92.12	93.09	89.55	93.85	90.56	90.69	92.94	90.13	93.59	96.43
	DAW-FDD (Ours)	AUC	92.78	94.78	91.78	95.79	93.84	95.50	88.59	96.75	93.17	93.23	93.29	95.93	94.80	97.81
		FPR	21.52	25.31	29.29	23.81	22.55	22.29	29.83	19.57	19.76	27.07	28.99	27.12	25.07	16.56
		TPR	90.29	96.72	94.66	95.66	93.20	95.00	90.74	94.90	89.75	91.62	96.72	96.46	96.46	99.86
		ACC	88.23	92.69	90.00	92.25	90.40	92.15	86.55	92.69	88.09	88.73	91.84	91.81	92.57	96.91
EfficientNet-B3	Original	AUC	94.72	97.07	94.56	98.80	95.96	96.85	90.78	98.64	94.87	96.37	96.59	99.00	97.03	98.29
		FPR	19.19	21.17	25.61	19.05	19.65	16.57	24.79	18.48	19.13	12.71	26.09	19.49	20.11	21.19
		TPR	96.07	98.25	97.33	99.19	97.19	96.07	93.00	99.02	96.56	93.74	99.60	99.38	97.78	99.42
		ACC	93.42	94.70	92.86	96.00	94.20	93.99	89.37	96.35	93.83	92.74	94.73	95.65	94.55	95.72
	DAG-FDD (Ours)	AUC	97.01	97.46	96.06	99.45	97.27	97.73	94.67	99.74	97.08	97.26	96.68	99.19	97.51	98.81
		FPR	8.14	8.61	10.43	0.95	8.46	8.43	9.66	0.00	8.37	8.29	10.87	1.70	8.55	8.61
		TPR	90.32	95.20	93.77	94.34	92.50	94.05	86.55	94.71	90.48	90.40	97.57	93.96	94.40	99.28
		ACC	90.59	94.51	92.95	95.17	92.33	93.64	87.32	95.52	90.68	90.61	95.97	94.82	93.87	97.86
	DAW-FDD (Ours)	AUC	95.96	96.68	95.80	98.09	96.22	97.79	95.09	97.45	95.83	97.20	95.92	98.69	96.68	98.88
		FPR	8.24	8.20	9.51	9.05	8.16	5.72	8.82	11.96	8.24	5.53	9.90	6.78	8.09	5.96
		TPR	88.55	94.05	93.36	95.15	90.68	92.98	86.44	93.73	88.34	89.50	97.00	96.67	92.90	97.97
		ACC	89.11	93.64	92.80	94.42	90.89	93.19	87.40	92.86	88.93	90.27	95.69	95.99	92.72	97.27
DSP-FWA	Original	AUC	89.75	93.97	90.13	95.60	91.72	94.16	83.99	96.38	90.10	90.63	93.32	96.51	93.62	98.77
		FPR	28.48	34.37	40.64	31.43	29.58	34.94	38.24	20.65	26.50	37.02	42.03	39.83	32.37	32.45
		TPR	90.08	95.99	95.33	96.77	92.46	94.11	91.82	95.29	89.48	90.30	97.17	98.33	95.28	99.57
		ACC	86.85	90.45	88.33	91.83	88.55	89.31	85.69	92.86	86.70	86.08	89.73	90.80	90.29	93.82
	DAG-FDD (Ours)	AUC	90.19	92.78	89.50	95.43	91.79	91.83	84.10	95.13	90.78	91.04	92.08	96.06	92.90	93.54
		FPR	29.86	34.50	42.64	37.14	30.11	31.63	42.86	41.30	27.25	29.83	42.51	33.90	32.71	33.78
		TPR	91.02	96.15	94.96	98.59	93.14	93.99	88.38	98.63	90.98	89.90	98.42	98.54	95.18	99.86
		ACC	87.39	90.55	87.64	92.33	89.01	89.76	82.01	92.53	87.80	86.85	90.65	92.14	90.15	93.82
	DAW-FDD (Ours)	AUC	88.15	93.54	90.54	94.44	90.63	92.05	85.08	95.23	87.81	91.13	94.30	93.86	93.49	94.07
		FPR	28.81	31.83	34.20	31.43	29.19	34.94	26.05	38.04	27.88	35.91	38.89	26.27	30.37	33.78
		TPR	87.64	95.92	92.73	96.06	91.22	95.18	82.35	95.88	87.10	92.42	98.19	96.25	95.12	99.13
		ACC	84.77	90.85	87.49	91.25	87.60	90.21	80.63	90.70	84.49	88.04	91.16	91.81	90.52	93.22
RECCE	Original	AUC	97.15	98.86	97.44	98.65	98.12	98.51	94.75	99.20	97.31	97.76	98.71	98.40	98.89	99.71
		FPR	21.67	20.80	32.67	28.10	19.26	13.86	39.50	41.30	19.07	11.05	28.74	17.80	19.43	17.22
		TPR	97.03	99.29	98.04	100	98.17	97.80	95.37	100	97.10	96.47	99.43	100	99.18	99.71
		ACC	93.77	95.62	92.06	95.08	95.08	95.88	88.26	93.69	94.29	95.30	94.09	96.49	95.82	96.67
	DAG-FDD (Ours)	AUC	97.71	98.90	97.13	99.65	98.40	99.04	94.02	99.69	97.97	98.09	98.56	99.62	98.85	99.88
		FPR	11.71	12.26	18.87	6.19	11.45	7.83	19.75	4.35	11.08	10.50	18.36	7.63	11.80	4.64
		TPR	95.15	98.31	96.55	98.79	96.79	96.13	92.47	99.22	95.35	94.04	98.70	98.33	98.16	99.13
		ACC	93.95	96.38	93.55	97.92	95.33	95.48	89.97	98.67	94.23	93.34	95.46	97.16	96.36	98.45
	DAW-FDD (Ours)	AUC	97.73	98.98	98.17	98.85	98.27	99.00	95.52	98.24	97.94	98.31	99.54	99.49	98.70	99.89
		FPR	8.29	8.04	7.67	10.00	8.64	3.01	7.56	16.30	8.56	2.76	7.73	5.09	8.72	3.31
		TPR	92.24	96.74	93.70	97.37	94.60	94.29	85.79	98.24	92.81	90.81	97.85	96.46	96.29	99.28
		ACC	92.15	95.86	93.43	96.08	94.02	94.73	87.15	96.01	92.57	91.80	96.79	96.15	95.38	98.81

Table E.8. *Detailed test set results of each group in ResNet-50, EfficientNet-B3, DSP-FWA, and RECCE on the FF++ dataset.*