DADM: Dual Alignment of Domain and Modality for Face Anti-spoofing

Supplementary Material

7. Proofs of Donsker-Varadhan Representation Theorem

We provide this section for helping understand mutual information maximization formula in Sec. 3.1. We typically need to estimate a lower bound of mutual information and then continuously raise this lower bound to achieve the goal. Among them, Donsker-Varadhan theorem [17] is a commonly used estimation of the lower bound of mutual information. Belghazi et al. [5] converts it into the dual representation:

$$\mathrm{I}_{\mathrm{MINE}} = \mathbb{E}_{p(x,y)}[f(x,y)] - \log(\mathbb{E}_{p(x)p(y)}[e^{(f(x,y))}]). \tag{20}$$

Donsker-Varadhan representation theorem [17]. *The KL-divergence possesses the following dual representation supremum:*

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{Q}) = \sup_{T:\,\Omega\to\mathbb{R},\,T\in\mathcal{F}} \mathbb{E}_{\mathrm{P}}[T] - \log(\mathbb{E}_{\mathrm{Q}}[e^{T}]), \tag{21}$$

where the supremum is taken over all functions T such that the two expectations ar finite. $\mathcal{F}$ be any class of functions $T : \Omega \to \mathbb{R}$ satisfying the integrability constrains of the theorem.

For a given function T, consider the Gibbs distribution G defined by $\mathrm{dG} = \frac{1}{Z}e^{T}\mathrm{dQ}$, where $Z = \mathbb{E}_{\mathrm{Q}}[e^{T}]$. By construction

$$\mathbb{E}_{\mathrm{P}}[T] - \log(\mathbb{E}_{\mathrm{Q}}[e^{T}]) = \mathbb{E}_{\mathrm{P}}[\log\frac{\mathrm{dG}}{\mathrm{dQ}}], \tag{22}$$

as $T = \log[Z\frac{\mathrm{dG}}{\mathrm{dQ}}] = \log Z + \log\frac{\mathrm{dG}}{\mathrm{dQ}} = \log(\mathbb{E}_{\mathrm{Q}}[e^{T}]) + \log\frac{\mathrm{dG}}{\mathrm{dQ}}$. Let Δ be the gap, and combining Eqn. 22:

$$\Delta = \mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{Q}) - \big(\mathbb{E}_{\mathrm{P}}[T] - \log(\mathbb{E}_{\mathrm{Q}}[e^{T}])\big), \tag{23}$$

$$\Delta = \mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{Q}) - \mathbb{E}_{\mathrm{P}}[\log\frac{\mathrm{dG}}{\mathrm{dQ}}], \tag{24}$$

$$\Delta = \mathbb{E}_{\mathrm{P}}[\log\frac{\mathrm{dP}}{\mathrm{dQ}} - \log\frac{\mathrm{dG}}{\mathrm{dQ}}] = \mathbb{E}_{\mathrm{P}}[\log\frac{\mathrm{dP}}{\mathrm{dG}}] = \mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{G}), \tag{25}$$

we can easily draw the conclusion that $\Delta \geq 0$, because KL-divergence $\mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{G})$ is always positive, i.e., $\mathrm{D}_{\mathrm{KL}}(\mathrm{P}||\mathrm{Q}) \geq \mathbb{E}_{\mathrm{P}}[T] - \log(\mathbb{E}_{\mathrm{Q}}[e^{T}])$. The proof is completed.

Since mutual information can be written in the form of the KL-divergence between the joint distribution and the product of the marginal distribution, such a lower bound can also be obtained for mutual information. The idea of [5] is to choose $\mathcal{F}$ to be the family of functions parametrized by a deep neural network with parameters $\theta \in \Theta$, so there exists:

$$\mathrm{I}(X;Y) \geq \mathrm{I}_{\Theta}(X;Y), \tag{26}$$

where $\mathrm{I}_{\Theta}(X;Y)$ is defined as:

$$\mathrm{I}_{\Theta}(X;Y) = \sup_{\theta\in\Theta}\mathbb{E}_{\mathrm{P}_{XY}}[T_{\theta}] - \log(\mathbb{E}_{\mathrm{P}_{X}\mathrm{P}_{Y}}[e^{T_{\theta}}]). \tag{27}$$

In code implementation, we estimate the expectations in Eq. 27 using empirical samples from P_{XY} and $\mathrm{P}_{X}\mathrm{P}_{Y}$ (i.e., by shuffling the samples from the joint distribution along the batch axis). Ultimately, the objective function can be optimized through gradient descent and back propagation. The common approach is to use an independent neural network to process the features of two modalities X and Y. Instead, we employ the average of the mutual information tokens mentioned in Eqn. 12, 15, where the MI tokens represent the summarization of fused features. It can be seen as a special case of the score function $f(x,y)$.

8. Supplementary Experimental Results

Empirical studies on hyper-parameters. In Tab. I, we conduct empirical studies on the λ coefficients of different loss terms. We select appropriate λ values within the interval of (0,1) to find the relatively optimal combination. The final combination obtained is $(\lambda_{\mathrm{mi}}, \lambda_{\mathrm{angle}}) = (0.1, 0.3)$. In Tab. J, we carry out empirical studies on temperature coefficients τ_l and τ_s. We attempt various values for τ_l and τ_s within the interval of (0,1) to determine the relatively optimal combination. Meanwhile, based on the experience from some previous studies [27, 79], we assume that live and spoof samples exhibit an asymmetry distribution with different degree pf relaxation in the hyper-feature space. Therefore, when conducting our attempts, we prefer to impose a more compact feature distribution to the live samples, while allowing the spoof samples to have a looser feature distribution. The optimal combination we have found is $(\tau_l, \tau_s) = (1.0, 0.85)$.

Convergence speed of CDC-Adapter and vanilla convolutional Adapter. In face anti-spoofing field, there are lots of works [9, 82, 84], apply central difference convolution operator for live/spoof representation capture. The CDC [84] operator combines both intensity-level semantic infor-

Table I. Empirical studies on λ coefficients.

λ_{mi}	λ_{angle}	HTER (%) ↓	AUC (%) ↑
0.3	0.5	15.54	90.54
0.2	0.5	14.98	90.63
0.1	0.5	14.33	91.65
0.1	0.4	14.02	91.94
0.0	0.4	14.52	92.11
0.0	0.3	14.31	92.05
0.1	0.3	**13.63**	**92.96**

Table J. Empirical results on temperature coefficient τ_l and τ_s.

τ_l	τ_s	HTER (%) ↓	AUC (%) ↑
1.0	0.5	15.80	90.77
1.0	0.6	15.25	90.68
1.0	0.7	14.74	91.05
1.0	0.8	14.28	91.93
1.0	0.9	13.91	92.30
0.9	0.8	14.09	91.92
0.95	0.8	13.85	91.98
1.0	0.85	**13.63**	**92.96**

mation and gradient-level messages:

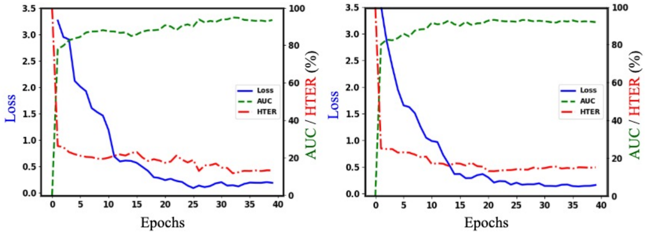
$$y(p_0) = \theta \cdot \underbrace{\sum_{p_n \in \mathcal{P}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}}$$
$$(1-\theta) \cdot \underbrace{\sum_{p_n \in \mathcal{P}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}}, \tag{28}$$

where p_0 is current location on input feature map while p_n enumerates the locations in $\mathcal{P}$ (pixel neighborhood), $w(p_n)$ are the weights of convolutional kernel corresponding to the location p_n, hyper-parameter $\theta \in [0, 1]$ tradeoffs the importance between intensity and gradient information.

In Fig. 7, we compare the convergence speed of Adapters based on vanilla convolution and CDC. Combining the results from Tab. G, they demonstrate that the convergence speed of the vanilla convolution (which tends to stabilize after about 20 epochs) is faster then CDC (which tends to stabilize after about 30 epochs), and the performance of CDC is better. This phenomenon indicates that vanilla convolutional Adapter may have higher risk of overfitting compared to CDC-Adapter, CDC-Adapter is more robust for our backbone's fine-tuning.

9. Algorithm

The multi-modal PG-IRM algorithm is shown as below. Additionally, our input contains three modalities and includes a constraint term of angle margin in the total loss, therefore, the algorithm is designed with dual alignment of hyperplanes and angles. Our DADM optimization pipeline:



(a) CDC-Adapter

(b) Vanilla Conv-Adapter

Figure 7. Convergence speed of different convolutional Adapter. (a) CDC (Central Difference Convolution)-Adapter. (b) Vanilla Convolutional-Adapter.

Algorithm 1 The optimization pipeline of **DADM**

```
Input: Source Data $S = \{x_i^{\text{RGB}}, x_i^{\text{D}}, x_i^{\text{I}}, y_j, e_i\}_i^N$,Target
    Data $T = \{x_j^{\text{RGB}}, x_j^{\text{D}}, x_j^{\text{I}}, y_j\}_j^M$ , neural network $\phi(\cdot)$,
    classifiers $\beta_{e_1}, \beta_{e_2}, \cdots, \beta_{\mathcal{E}}$, learning rate $\gamma$, alignment
    parameter $\alpha$, alignment starting epoch $T_\alpha$.
Output: $\phi(\cdot)$, mean$(\beta_{e_1}, \beta_{e_2}, \cdots, \beta_{\mathcal{E}})$
for t in 0, 1, $\cdots$, T do
    Data Prep: Sampling a mini-batch $B$ samples, $X_s = \{x_i^{\text{RGB}}, x_i^{\text{D}}, x_i^{\text{I}}, y_j, e_i\}_i^B$
    Forward: Obtain multi-modal features and scores, $[f_i^{\text{RGB}}, f_i^{\text{D}}, f_i^{\text{I}}]_{e_i} = \phi^t([x_i^{\text{RGB}}, x_i^{\text{D}}, x_i^{\text{I}}]_{e_i}), \hat{y}_{e_i} = \beta_{e_i}^t[f_i^{\text{RGB}}, f_i^{\text{D}}, f_i^{\text{I}}]_{e_i}$
    Backward: Compute $\mathcal{L}_{\text{total}}$, update $\phi^{t+1} = \phi^t - \gamma\nabla_{\phi^t}\mathcal{L}_{\text{total}}$
    for $e \in \mathcal{E}$ do
        $\tilde{\beta}_e^{t+1} = \beta_e^t - \gamma\nabla_{\beta_e^t}\mathcal{L}_{\text{total}}$
        select $\beta_{\bar{e}}^t$ with $\bar{e} = \underset{e' \in \mathcal{E} \setminus e}{\text{argmax}} ||\tilde{\beta}_e^{t+1} - \beta_{e'}^t||_2$
        $\alpha' = 1 - \mathbf{1}_{1>T_\alpha}(1-\alpha)$
        $\beta_e^{t+1} = \alpha'\tilde{\beta}_e^{t+1} + (1-\alpha')\beta_{\bar{e}}^t$
    end for
    $\bar{\beta}^{t+1} = $ mean$(\beta_{e_1}^{t+1}, \beta_{e_2}^{t+1}, \cdots, \beta_{\mathcal{E}}^{t+1})$
    Evaluate: Test $\phi^{t+1}(\cdot)$, $\bar{\beta}^{t+1}$ on $T$
    if performance better then
        update $\phi^*(\cdot) = \phi^{t+1}(\cdot), \beta^* = \bar{\beta}^{t+1}$
    end if
end for
Return $\phi^*(\cdot)$, $\beta^*$
```

10. Proofs of the Necessity of Domain Alignment and Angle Alignment

Invariant Risk Minimization (IRM) is a challenging bi-level optimization problem that is hard to solve. Thanks to the efforts of Sun et. al [69], they propose the Projected Gradient Optimization for IRM (PG-IRM) which is an equivalent objective to IRM, with strict proof, and it is easier to optimize. The brief proof process is as follows:

Theorem 1. Projected Gradient Optimization IRM objective is equivalent to IRM objective. *For all $\alpha \in (0, 1)$,* *the IRM objective is equivalent to the following objective:*

$$\begin{aligned}&\min_{\phi,\beta_{e_1},\cdots,\beta_{\mathcal{E}}} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} R^e(\phi,\beta_e),\\ &s.t.\,\forall e\in\mathcal{E},\,\exists\beta_e\in\Omega_e(\phi),\beta_e\in\Upsilon_\alpha(\beta_e)\,,\end{aligned}\tag{29}$$

where the parametric constrained set for each environment is simplified as $\Omega_e(\phi) = \underset{\beta}{\mathrm{argmin}} R^e(\phi,\beta)$*, and the* α*-adjacency set is defined as:*

$$\begin{aligned}\Upsilon_\alpha(\beta_e) =&\{v\,|\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||v-\beta_{e'}||_2\\ &\leq\alpha\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||\beta_e-\beta_{e'}||_2\}.\end{aligned}\tag{30}$$

Proofs 1.

The IRM objective is the following constrained optimization problem:

$$\begin{aligned}&\min_{\phi,\beta^*}\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}R^e(\phi,\beta^*),\\ &s.t.\,\beta^*\in\underset{\beta}{\mathrm{argmin}}R^e(\phi,\beta),\,\forall e\in\mathcal{E},\end{aligned}\tag{31}$$

where ϕ represents a neural network, β denotes the hyperplane for classification, $\mathcal{E} = \{e_1, e_2, \cdots, e_{|\mathcal{E}|}\}$ represents the entire environment, e is one of the sub-environments, and $f(x;\beta,\phi)$ is the function processing x via ϕ,β and obtaining y. The risk function $R^e(\phi,\beta)$, based on the loss function $\mathcal{L}(\cdot,\cdot)$, for a given environment e, is defined as:

$$R^e(\phi,\beta)=\mathbb{E}_{(x,y)\sim e}[\mathcal{L}(f(x;\beta,\phi),y)].\tag{32}$$

The constrain $\beta^* = \underset{\beta}{\mathrm{argmin}}R^e(\phi,\beta)$, $\forall e\in\mathcal{E}$, means that the β^* is the optimal linear classifier for all $e\in\mathcal{E}$, which is equivalent to $\beta^*\in\bigcap_{e\in\mathcal{E}}\Omega_e(\phi)$, and equivalent to:

$$\forall e\in\mathcal{E},\,\exists\beta_e\in\Omega_e(\phi),\beta^*=\beta_e\,.\tag{33}$$

This indicates that for all $e\in\mathcal{E}$, there is a hyperplane in the optimal set $\Omega_e(\phi)$ that also lies in the intersection of other environments' optimal set ($\bigcap_{e'\in\mathcal{E}\backslash e}\Omega_{e'}(\phi)$), i.e.:

$$\forall e\in\mathcal{E},\,\exists\beta_e\in\Omega_e(\phi),\beta_e\in\bigcap_{e'\in\mathcal{E}\backslash e}\Omega_{e'}(\phi)\,.\tag{34}$$

Sun et. al [69] relax the constrain to:

$$\beta_e\in\bigcap_{e'\in\mathcal{E}\backslash e}\Omega_{e'}(\phi)\rightarrow\max_{e'\in\mathcal{E}\backslash e}||\beta_e-\Omega_{e'}(\phi)||_2\leq\epsilon\,,\tag{35}$$

due to one key challenge for constrain 34 is that there is a no guarantee that is non-empty for a feature extractor ϕ and β_e. Then they define the l_2 distance between a vector β and a set Ω as: $||\beta-\Omega||_2=\min_{e'\in\mathcal{E}\backslash e}||\beta-v||_2$. Practically, ϵ can be set to be any variable converging to 0 during the optimization stage. Without losing the generality, they change the constraint to the following form:

$$\forall e\in\mathcal{E},\,\exists\beta_e\in\Omega_e(\phi),\tag{36}$$

$$\begin{aligned}&\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||\beta_e-\beta_{e'}||_2\leq\\ &\alpha\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||\beta_e-\beta_{e'}||_2,\end{aligned}\tag{37}$$

where $\alpha\in(0,1)$. Note that constraint 36 will be satisfied only when $\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||\beta_e-\beta_{e'}||_2=0$. Therefore constraint 34 is equivalent to constraint 36.

Let's define $\Upsilon_\alpha(\beta_e)$:

$$\begin{aligned}\Upsilon_\alpha(\beta_e) =&\{v\,|\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||v-\beta_{e'}||_2\\ &\leq\alpha\max_{e'\in\mathcal{E}\backslash e}\min_{\beta_{e'}\in\Omega_{e'}(\phi)}||\beta_e-\beta_{e'}||_2\},\end{aligned}\tag{38}$$

then the constraint 33 can be simplified to:

$$s.t.\,\forall e\in\mathcal{E},\,\exists\beta_e\in\Omega_e(\phi),\beta_e\in\Upsilon_\alpha(\beta_e)\,.\tag{39}$$

Proofs 1 Completed.

Above Theorem 1 ensures that the PG-IRM's optimization objective being equivalent to the IRM's optimization objective.

Why we need dual alignment of hyperplane and angle margin? In uni-modality scenarios, misalignment has always been a critical concern, as it relates to whether domain-invariant representations have been truly learned. MMDG [42] found that directly incorporating multi-modality into DG-FAS can result in performance degradation, indicating the significance impact from domain and modality misalignment. Therefore, dual alignment of modality and domain is crucial.

Theorem 2. Misalignment of angle margin for modality features leads to severe shift and difficult convergence of the optimal classification hyperplane β^* in PG-IRM. *For misaligned angle margin among modalities features in varies domains* $[f_0^e,...,f_i^e,...,f_{\mathcal{M}}^e]_{\mathcal{M}}\in\mathcal{E}$*, where* $[f_0^e,...,f_i^e,...,f_{\mathcal{M}}^e]=\phi([x_0^e,...,x_i^e,...,x_{\mathcal{M}}^e])\in\mathbb{R}^{\mathcal{D}\times\mathcal{M}}$*,* x_i^e *represents single modality input* i *from environment* e*. The the optimal classification hyperplane* β^* *will severely shift.*

Notation declarations.

For $f_i^e(k)$, f represents the feature, the superscript e denotes f^e comes from environment e, the subscript i denotes the i-th modality feature, $f(k)$ indicates the k-th element of f. Specially, f^e (without subscript) denotes final fusion feature from environment e, $f^e(k)$ indicates the k-th element of f^e.

Proofs 2.

For $[f_0^e,...,f_i^e,...,f_{\mathcal{M}}^e]=\phi([x_0^e,...,x_i^e,...,x_{\mathcal{M}}^e])\in\mathbb{R}^{\mathcal{D}\times\mathcal{M}}$, where $f_i^e\in\mathbb{R}^{\mathcal{D}\times 1}$ and $\mathcal{M}$ is the number of modalities, we construct a modality matrix for environment e:

$$\mathrm{F}^e=[f_0^e,...,f_i^e,...,f_{\mathcal{M}}^e]\in\mathbb{R}^{\mathcal{D}\times\mathcal{M}},\tag{40}$$

the final fusion feature f^e is obtained via a linear projecting $\mathrm{P} \in \mathbb{R}^{\mathcal{M}\times 1}$:

$$f^e = \mathrm{F}^e\mathrm{P} = \sum_i^{\mathcal{M}} p_i f_i^e \in \mathbb{R}^{\mathcal{D}\times 1}. \tag{41}$$

Intra-domain case.

The intra-domain co-variance matrix of f^e is as follows (for the simplicity, we omit the superscript e):

$$\begin{aligned}\mathbb{E}[f] &= \sum_i^{\mathcal{M}} p(i)\mathbb{E}[f_i],\\ \mathbb{D}[f] &= \mathbb{E}[(f-\mathbb{E}[f])(f-\mathbb{E}[f])^{\mathbf{T}}]\\ &= \mathbb{E}[ff^{\mathbf{T}}] - \mathbb{E}[f]\mathbb{E}[f]^{\mathbf{T}}\\ &= \mathbb{E}[\mathrm{F}\mathrm{P}\mathrm{P}^{\mathbf{T}}\mathrm{F}^{\mathbf{T}}] - \mathbb{E}[f]\mathbb{E}[f]^{\mathbf{T}},\end{aligned} \tag{42}$$

where $p(i)$ denotes the i-th element of P.

Assuming that the modality features f_i^e have been normalized before classification by the classifier β, i.e., $\mathbb{E}[f_i] = 0, ||f_i|| = 1$, thus Eqn. 42 can be rewritten as:

$$\begin{aligned}\mathbb{E}[f] &= 0,\\ \mathbb{D}[f] &= \mathbb{E}[\mathrm{F}\mathrm{P}\mathrm{P}^{\mathbf{T}}\mathrm{F}^{\mathbf{T}}],\end{aligned} \tag{43}$$

and k-th diagonal elements of co-variance matrix $\mathbb{D}[f]$, which represents the $\mathbb{D}[f(k)]$:

$$\begin{aligned}\mathbb{D}[f(k)] &= p(k)^2\mathbb{E}[< f_k, f_k >]\\ &= p(k)^2\mathbb{E}[||f_k|| \cdot ||f_k||\mathrm{cos}\theta_k]\\ &= p(k)^2\mathbb{E}[\mathrm{cos}\theta_{kk}],\end{aligned} \tag{44}$$

where $<,>$ indicates the inner product, $p(k)$ denotes the k-th element of P.

Since we consider that the distribution of angles θ_{kk} (θ) without intervention generally does not approach a constant distribution, in order to maintain generality, we suppose that θ follows a Gaussian distribution with μ and variance σ, $N(\mu, \sigma)$:

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{exp}(-\frac{(\theta-\mu)^2}{2\sigma^2}). \tag{45}$$

Then we can calculate the value of $\mathbb{E}[\mathrm{cos}\theta]$:

$$\mathbb{E}[\mathrm{cos}(\theta)] = \int_{-\infty}^{\infty} \mathrm{cos}(\theta)\cdot\frac{1}{\sigma\sqrt{2\pi}}\mathrm{exp}(-\frac{(\theta-\mu)^2}{2\sigma^2})\mathrm{d}\theta, \tag{46}$$

$$\mathrm{cos}(\theta) = \frac{\mathrm{exp}(-i\theta) + \mathrm{exp}(i\theta)}{2}, \tag{47}$$

$$\mathbb{E}[\mathrm{cos}(\theta)] = \frac{1}{2}(\mathbb{E}[\mathrm{exp}(-i\theta)] + \mathbb{E}[\mathrm{exp}(i\theta)]). \tag{48}$$

For Gaussian distribution $N(\mu, \sigma)$, its characteristic function is $\Phi(t) = \mathbb{E}[\mathrm{exp}(-it\theta)] = \mathrm{exp}(i\mu t - \frac{\sigma^2 t^2}{2})$.

The characteristic function when t takes 1 and -1 is:

$$\mathbb{E}[\mathrm{exp}(i\theta)] = \mathrm{exp}(i\mu - \frac{\sigma^2}{2}), \tag{49}$$

$$\mathbb{E}[\mathrm{exp}(-i\theta)] = \mathrm{exp}(-i\mu - \frac{\sigma^2}{2}). \tag{50}$$

Substitute Eqn. 49, 50 into Eqn. 48:

$$\begin{aligned}\mathbb{E}[\mathrm{cos}(\theta)] &= \frac{1}{2}(\mathrm{exp}(i\mu - \frac{\sigma^2}{2}) + \mathrm{exp}(-i\mu - \frac{\sigma^2}{2})),\\ \mathbb{E}[\mathrm{cos}(\theta)] &= \frac{1}{2}\mathrm{exp}(-\frac{\sigma^2}{2})\cdot 2\mathrm{cos}(\mu) = \mathrm{exp}(-\frac{\sigma^2}{2})\mathrm{cos}(\mu).\end{aligned} \tag{51}$$

Thus, increasing the variance (σ) of θ will leads to a decrease in the value of $\mathbb{D}[f(k)]$:

$$\mathbb{D}[f(k)] = p(k)^2\mathrm{exp}(-\frac{\sigma^2}{2})\mathrm{cos}(\mu). \tag{52}$$

This result indicates that when the angle margins θ between modalities exhibit a significant disturbance, the $\mathbb{D}[f(k)]$ will decrease.

Inter-domain case.

The inter-domain co-variance matrix between f^{e_1} and f^{e_2} is as follows:

$$\begin{aligned}\mathbb{E}[f^{e_1}] = \sum_i^{\mathcal{M}} p(i)\mathbb{E}[f_i^{e_1}],\ &\mathbb{E}[f^{e_2}] = \sum_i^{\mathcal{M}} p(i)\mathbb{E}[f_i^{e_2}],\\ \mathbb{C}[f^{e_1}, f^{e_2}] &= \mathbb{E}[(f^{e_1} - \mathbb{E}[f^{e_1}])(f^{e_2} - \mathbb{E}[f^{e_2}])^{\mathbf{T}}]\\ &= \mathbb{E}[f^{e_1}f^{e_2\mathbf{T}}] - \mathbb{E}[f^{e_1}]\mathbb{E}[f^{e_2}]^{\mathbf{T}}\\ &= \mathbb{E}[\mathrm{F}^{e_1}\mathrm{P}^{e_1}\mathrm{P}^{e_2\mathbf{T}}\mathrm{F}^{e_2\mathbf{T}}] - \mathbb{E}[f^{e_1}]\mathbb{E}[f^{e_2}]^{\mathbf{T}}.\end{aligned} \tag{53}$$

Please note that f^{e_1} and f^{e_2} exhibit the same liveness label.

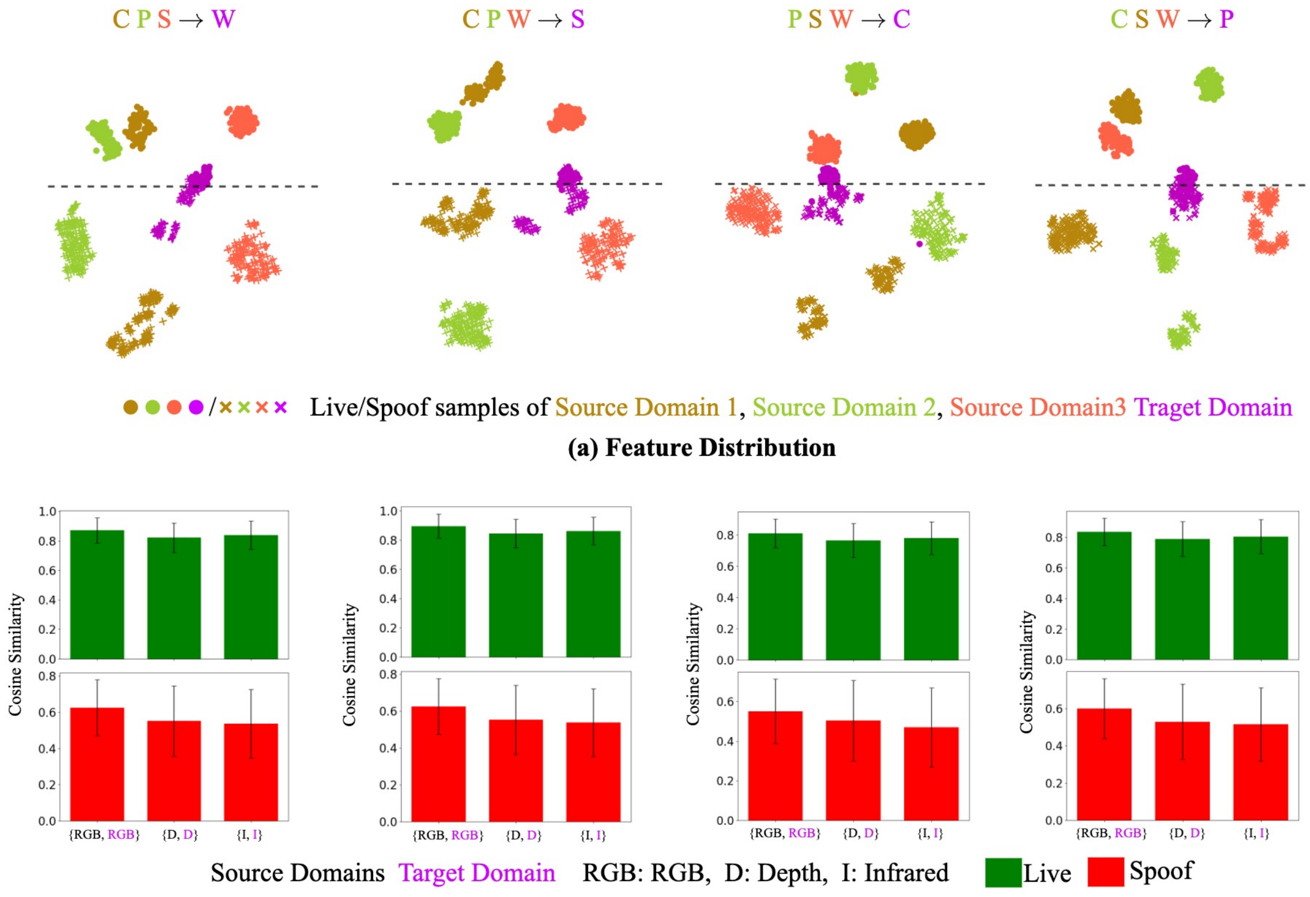
Assuming that the modality features f_i^e have been normalized before classification by the classifier β, i.e., $\mathbb{E}[f_i^e] = 0, ||f_i^e|| = 1$, thus Eqn. 53 can be rewritten as:

$$\begin{aligned}\mathbb{E}[f^{e_1}] = 0,\ &\mathbb{E}[f^{e_2}] = 0,\\ \mathbb{C}[f^{e_1}, f^{e_2}] &= \mathbb{E}[\mathrm{F}^{e_1}\mathrm{P}^{e_1}\mathrm{P}^{e_2\mathbf{T}}\mathrm{F}^{e_2\mathbf{T}}],\end{aligned} \tag{54}$$

the k-th diagonal elements of $\mathbb{C}[f^{e_1}, f^{e_2}]$, which represents the co-variance between $f^{e_1}(k)$ and $f^{e_2}(k)$:

$$\begin{aligned}\mathbb{C}[f^{e_1}(k), f^{e_2}(k)] &= p(k)^2\mathbb{E}[< f_k^{e_1}, f_k^{e_2} >]\\ &= p(k)^2\mathbb{E}[||f_k^{e_1}|| \cdot ||f_k^{e_2}||\mathrm{cos}\theta_{kk}]\\ &= p(k)^2\mathbb{E}[\mathrm{cos}\theta_{kk}].\end{aligned} \tag{55}$$

Similarly, we can also conclude that increasing the variance (σ) of θ will also lead to a decrease in the value of $\mathbb{C}[f^{e_1}(k), f^{e_2}(k)]$ according to **Intra-domain case**.



(a) Feature Distribution

(b) Angle Margin

Figure 8. Illustration of dual alignment of domain and modality for four sub-protocols. (a) feature distribution of source and target domains, the dotted line represents the decision hyperplane in 2D space. (b) Mean and Std. of cosine similarity among modalities in the source and target domains.

The impact of $\mathbb{D}[f(k)]$ on the convergence and shift of β. Before computing the loss function, we need to use a linear classifier β and $\mathrm{softmax}(\cdot)$ projecting final fusion feature f to logits $\hat{y}$, where $z = \beta f$, $\hat{y} = [\frac{\exp(z_p)}{\exp(z_p)+\exp(z_n)}, \frac{\exp(z_n)}{\exp(z_p)+\exp(z_n)}]$, p represents positive sample, n represents negative sample. Considering that using cross-entropy loss:

$$\mathcal{L} = -\mathbb{I}(\mathrm{label}_{\mathrm{GT}})\hat{y}\log\hat{y} \tag{56}$$

where $\hat{y}_p, \hat{y}_n \in (0,1)$, the gradient of $\mathcal{L}$:

$$\begin{aligned}\nabla_{\beta_e^t}\mathcal{L} &= \mathbb{I}(\mathrm{label}_{\mathrm{GT}})\nabla_{\beta_e^t}\hat{y}\,\frac{\partial\mathcal{L}}{\partial\hat{y}} \\ &= -\mathbb{I}(\mathrm{label}_{\mathrm{GT}})\nabla_{\beta_e^t}\hat{y}(\log\hat{y}+1)\end{aligned} \tag{57}$$

then we consider the variance of $z_p = \beta_p f = \sum_k^{\mathcal{D}} w(k)f(k)$ (the same applies to the analysis of $z_n = \beta_n f$), which $\mathbb{D}[z_p] = \sum_k^{\mathcal{D}} w(k)^2\mathbb{D}[f(k)]$, $w(k)$ is the weight of β_p. When the σ of θ_{kk} increases, the $\mathbb{D}[f(k)]$ decreases, so does the $\mathbb{D}[z_p]$. The smaller $\mathbb{D}[z_p]$ and $\mathbb{D}[z_n]$ will lead to the difference between z_p and z_n may be subtle at the start of training (supposing that randomly initialization does not favor either z_p or z_n), resulting in a flatten value of softmax output $\hat{y}$, i.e., the value of logits ($\hat{y}$) tend towards a uniform distribution. And we can easily know that the function $\hat{y}\log\hat{y}$ has its maximum value when the probability of $\hat{y}$ reaches $1/n$ (for two categories, n equals to 2).

At this point, the drastic fluctuation in θ_{kk} will cause the absolute value of gradient $|\nabla_{\beta_e^t}\mathcal{L}|$ to be difficult to converge to a smaller value. According to line 6-9 in Algorithm 1:

$$\begin{aligned}&\beta_e^{t+1} = \alpha'\tilde{\beta}_e^{t+1} + (1-\alpha')\beta_{\bar{e}}^t \rightarrow \bar{e} = \underset{e'\in\mathcal{E}\backslash e}{\mathrm{argmax}}||\tilde{\beta}_e^{t+1} - \beta_{e'}^t||_2, \\ &\beta_e^{t+1} = \alpha'(\beta_e^t - \gamma\nabla_{\beta_e^t}\mathcal{L}_{\mathrm{total}}) + (1-\alpha')\beta_{\bar{e}}^t, \\ &\beta_e^{t+1} = \beta_{\bar{e}}^t + \alpha'(\beta_e^t - \beta_{\bar{e}}^t) - \alpha'\gamma\nabla_{\beta_e^t}\mathcal{L}_{\mathrm{total}},\end{aligned} \tag{58}$$

t starts from 0 to T, the single-step shift will accumulate increasingly, resulting in the optimal classification hyperplane $\bar{\beta}^{T+1} = \mathrm{mean}(\beta_{e_1}^{T+1}, \beta_{e_2}^{T+1}, \cdots, \beta_{\mathcal{E}}^{T+1})$ shift severely.
Proofs 2 Completed.

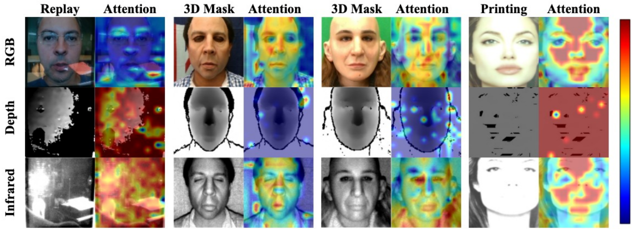


Figure 9. More visualization attention maps on varies attack samples, for example, replay attack, 3D mask, paper printing.

11. Visualization

Comprehensive visualization of dual alignment of domain and modality. Fig. 8 presents the visualizations of dual alignment of hyperplanes and angles for four sub-protocols in Tab. A. we can observe that in the sub-protocols CPS → W and CPW → S, the hyperplane for the live/spoof decision remains consistent across different source domains and is also transferable to unseen target domain. Moreover, the angles between the source domains and the target domain are relatively close to the expected values. In contrast, the other two sub-protocols PSW → C and CSW → P, exhibit slightly poorer illustration. Correspondingly, they also show poorer performance in Tab. A, which might be due to their encountering of a more significant domain shift.

More visualization attention maps. In Fig. 9, For 3D masks, the face region in the depth map is shown with cooler color, indicating its weak influence. For paper-printing attacks, depth information is particularly revealing of spoof cues, thereby warranting higher importance. For video replay attacks, more obvious spoofing traces were observed from infrared and depth maps, so both of them have higher importance than RGB.