In this supplementary material, we first provide a brief description of the datasets used in our experiment Section (Section A). Next, the proof of Theorem 1 is provided in Section B. In Section C, we will detail the implementation and present ablation studies about the detectors' robustness towards unseen corruptions. Lastly, in Section D, we discuss the limitations of our proposed method and outline our planned future work.

A. Datasets

We describe here four popular benchmark datasets used to evaluate our proposed GAC-FAS:

- **Idiap Replay Attack** (denoted as I) [7]: This dataset includes 1,300 videos captured from 50 clients under two different lighting conditions. It features four types of replayed faces and one type of printed face for spoof attacks.
- **OULU-NPU** (denoted as O) [3]: Comprising high-resolution videos, this dataset contains 3,960 spoof face videos and 990 live face videos captured from six different cameras. It includes two kinds of printed faces and two kinds of replayed faces.
- **CASIA-MFSD** (denoted as C) [76]: Consisting of 50 subjects, each with 12 videos, this dataset features three types of attacks: printed photo, cut photo, and video attacks.
- **MSU-MFSD** (denoted as M) [67]: This dataset includes 280 videos for 35 subjects recorded with different cameras. It encompasses three spoof types: one kind of printed face and two kinds of replayed faces.

Following the pre-processing steps outlined in [57], we utilized MTCNN [73] to detect faces in each frame of the videos.

B. Proof of Theorem 1

Theorem 1 (***Restate***): Suppose that the loss function $\ell(\theta_t) = \ell(f(x;\theta_t), y)$ satisfies the following assumptions. (i) its gradient $g(\theta_t) = \nabla\ell(\theta_t)$ is bounded, i.e., $\|g(\theta_t)\| \leq G, \forall t$. (ii) The stochastic gradient is L-Lipchitz, i.e., $\|g(\theta_t) - g(\theta_t')\| \leq L\|\theta_t - \theta_t'\|, \forall \theta_t, \theta_t'$. Let the learning rate η_t be $\frac{\eta_0}{\sqrt{t}}$, and and let the perturbation be proportional to the learning rate, i.e., $\rho_t = \frac{\rho}{\sqrt{t}}$, and $\gamma_t = \frac{\gamma}{\sqrt{t}}$, we have:

$$\frac{1}{T}\Sigma_{t=1}^{T}\mathbb{E}_{\mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}\mathbb{E}_{(x,y)\sim\mathcal{S}_i}\left[\|\nabla\ell(\theta_t)\|^2\right] \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right), \text{and}$$
$$\frac{1}{T}\Sigma_{t=1}^{T}\mathbb{E}_{\mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}\mathbb{E}_{(x,y)\sim\mathcal{S}_i}\left[\|\nabla\ell(\theta_t^{\text{adv}})\|^2\right] \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right),$$

where $\theta_t^{\text{adv}} = \theta_t + \hat{\epsilon}_t - \gamma_t\delta_t$, $\delta_t = \Sigma_{j=1}^{k}\nabla\ell(f(x_j';\theta_t), y_j')$ with $(x_j', y_j') \sim \mathcal{S}_j$.

For simplicity, we denote the update at step t as:

$$d_t = -\eta_t g(\theta_t) - \eta_t g(\theta_t^{\text{adv}}). \tag{11}$$

By L-smoothess of ℓ and the definition of $d_t = \theta_{t+1} - \theta_t$, we have:

$$\begin{aligned}
&\ell(\theta_{t+1}) - \ell(\theta_t) \leq \langle\nabla\ell(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= \langle\nabla\ell(\theta_t), d_t\rangle + \frac{L}{2}\|d_t\|^2 \\
&= -\eta_t\langle\nabla\ell(\theta_t), g(\theta_t) + g(\theta_t^{\text{adv}})\rangle + \frac{L\eta_t^2}{2}\|g(\theta_t) + g(\theta_t^{\text{adv}})\|^2 \\
&= -\eta_t\langle\nabla\ell(\theta_t), \nabla\ell(\theta_t) + g(\theta_t^{\text{adv}})\rangle + \frac{L\eta_t^2}{2}\|g(\theta_t) + g(\theta_t^{\text{adv}})\|^2 \\
&= -\eta_t\langle\nabla\ell(\theta_t), \nabla\ell(\theta_t) + \nabla\ell(\theta_t) - \nabla\ell(\theta_t) + g(\theta_t^{\text{adv}})\rangle \\
&+ \frac{L\eta_t^2}{2}\|g(\theta_t) + g(\theta_t^{\text{adv}})\|^2 \\
&\leq -2\eta_t\|\nabla\ell(\theta_t)\|^2 - \eta_t\langle\nabla\ell(\theta_t), g(\theta_t^{\text{adv}}) - g(t)\rangle + L\eta_t^2G^2
\end{aligned}$$

Taking expectation on both sides, and let $\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}} = \mathbb{E}_{\mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}\mathbb{E}_{(x,y)\sim\mathcal{S}_i}$, we have:

$$\begin{aligned}
&\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\ell(\theta_{t+1}) - \ell(\theta_t)\right] \leq -2\eta_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \\
&+ \eta_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\langle\nabla\ell(\theta_t), g(t) - g(\theta_t^{\text{adv}})\rangle\right] + L\eta_t^2G^2.
\end{aligned} \tag{12}$$

Here we need to bound the term $\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\langle\nabla\ell(\theta_t), g(t) - g(\theta_t^{\text{adv}})\rangle\right]$. We have:

$$\begin{aligned}
&\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\langle\nabla\ell(\theta_t), g(t) - g(\theta_t^{\text{adv}})\rangle\right] \\
&\leq \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|g(t) - g(\theta_t^{\text{adv}})\|\right] \\
&\leq L\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|\theta_t - \theta_t^{\text{adv}}\|\right] \text{ (assumption (ii))} \\
&= L\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|\hat{\epsilon}_t - \gamma_t\delta_t\|\right] \\
&\leq L\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|\hat{\epsilon}_t\|\right] \\
&+ L\gamma_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|\delta_t\|\right] \\
&\leq L\rho_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|\right] \\
&+ L\gamma_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\| \cdot \|\delta_t\|\right] \ (\hat{\epsilon}_t \leq \rho_t) \\
&\leq L\rho_t\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|\right] + L\gamma_t kG\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i\\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|\right] \\
&\leq L\rho_t G + L\gamma_t kG^2 \text{ (assumption (i)).}
\end{aligned} \tag{13}$$

Replace Equation 13 into Equation 12 we obtain:

$$\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\ell(\theta_{t+1})-\ell(\theta_t)\right] \leq -2\eta_t \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] + L\rho_t G + Lk\gamma G^2 + L\eta_t^2 G^2. \quad (14)$$

Re-arrange the above formula, we have:

$$2\eta_t \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \leq \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\ell(\theta_t)-\ell(\theta_{t+1})\right] + L\rho_t G + Lk\gamma_t G^2 + L\eta_t^2 G^2. \quad (15)$$

Perform telescope sum and taking expectation on each step we have:

$$2\Sigma_{t=1}^T \eta_t \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \leq \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\ell(\theta_0)-\ell(\theta_T)\right] + LG\Sigma_{t=1}^T\rho_t + LkG^2\Sigma_{t=1}^T\gamma_t + LG^2\Sigma_{t=1}^T\eta_t^2. \quad (16)$$

Note that our schedules are $\eta_t \frac{\eta_0}{\sqrt{t}}$ $\rho_t = \frac{\rho}{\sqrt{t}}$, and $\gamma_t = \frac{\gamma}{\sqrt{t}}$ then we have:

$$\begin{aligned}
&\frac{2\eta_0}{\sqrt{T}}\Sigma_{t=1}^T \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \leq \text{LHS(16)} \leq \text{RHS(16)} \\
&\leq \ell(\theta_0) - \ell_{\min} + LG\rho\Sigma_{t=1}^T\frac{1}{\sqrt{t}} + LkG^2\gamma\Sigma_{t=1}^T\frac{1}{\sqrt{t}} \\
&+ LG^2\eta_0^2\Sigma_{t=1}^T\frac{1}{t} \\
&\leq \ell(\theta_0) - \ell_{\min} + LG\rho(2\sqrt{T}-1) + LkG^2\gamma(2\sqrt{T}-1) \\
&+ LG^2\eta_0^2(1+\log(T)). \quad (17)
\end{aligned}$$

Hence,

$$\frac{1}{T}\Sigma_{t=1}^T \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \leq C_0 + \frac{C_1}{\sqrt{T}} + C_2\frac{\log T}{\sqrt{T}} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) \quad (18)$$

where C_0, C_1, C_2 are some constants.

For the second part of the Theorem, we have that :

$$\begin{aligned}
&\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t^{\text{adv}})\|_2^2\right] \\
&= \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t) + \nabla\ell(\theta_t^{\text{adv}}) - \nabla\ell(\theta_t)\|^2\right] \\
&\leq 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \\
&+ 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t^{\text{adv}}) - \nabla\ell(\theta_t)\|^2\right] \\
&\leq 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] + 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|g(\theta_t^{\text{adv}}) - g(\theta_t)\|\right] \\
&\leq 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] \\
&+ 2L^2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\theta_t^{\text{adv}} - \theta_t\|^2\right] \text{ (assumption (ii))} \\
&\leq 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] + 2L^2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\hat{\epsilon}_t - \gamma_t\delta_t\|^2\right] \\
&\leq 2\mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t)\|^2\right] + 2L^2\left(\rho_t^2 + \gamma_t^2 k^2 G^2\right) \quad (19)
\end{aligned}$$

Sum over t and average, then we have:

$$\begin{aligned}
&\frac{1}{T}\Sigma_{t=1}^T \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t^{\text{adv}})\|_2^2\right] \\
&\leq 2\left(C_0 + \frac{C_1}{\sqrt{T}} + C_2\frac{\log T}{\sqrt{T}}\right) \\
&+ 2L^2\frac{1+\log T}{T}(\rho_0^2 + \gamma_0^2 k^2 G^2). \quad (20)
\end{aligned}$$

Therefor,

$$\begin{aligned}
\frac{1}{T}\Sigma_{t=1}^T \mathbb{E}_{\substack{(x,y)\sim\mathcal{S}_i \\ \mathcal{S}_i\sim\boldsymbol{\mathcal{S}}}}\left[\|\nabla\ell(\theta_t^{\text{adv}})\|_2^2\right] &\leq C_3 + \frac{C_4}{\sqrt{T}} + C_5\frac{\log T}{\sqrt{T}} \\
&= \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) \quad (21)
\end{aligned}$$

where C_3, C_4, C_5 are some constants.

C. More Empirical Experiment

Details of our Implementation. We observe in Alg. 1 that random sampling in each iteration does not necessarily include images from all source domains. Specifically, the algorithm functions effectively even if, during certain iterations, the images in a minibatch belong to only 2, or even a single, domain. However, this issue can be mitigated by designing a balanced sampler. Regarding the training process, the hyperparameters are detailed precisely in Table 6.

	lr. step	FC lr. scale	Logit scale	Weight decay	Epochs
ICM → O	40	10	12	1×10^{-4}	150
OMI → C	40	1	16	5×10^{-4}	80
OCM → I	40	10	32	6×10^{-4}	50
OCI → M	5	10	12	6×10^{-4}	20

Table 6. Hyper-parameter settings in our experiment.

Robustness to Unseen Corruptions. In assessing the generalization capabilities of a FAS detector, it is crucial to evaluate its robustness against various types of input corruptions, a topic extensively explored in prior works [24, 32]. Adopting the experimental settings from [24], we examine the detector's performance under six common image corruptions: saturation, contrast, block-wise distortion, white Gaussian noise, blurring, and JPEG compression, each with five levels of severity. In Figure 6, we showcase examples of live and spoof faces affected by *six* types of image corruption techniques [24], each applied with a severity level of 3. While these represent digital corruptions, they are still pertinent for assessing the resilience of spoof face detectors.

We compare our method with 4 baselines: SSAN [66], SSDG [44], SA-FAS [57], and IADG [77]. These comparisons are based on the available official models. The results are demonstrated in Fig. 7. Our method consistently exhibits robustness across varying severity levels, as indicated by its lower HTER performance on average.

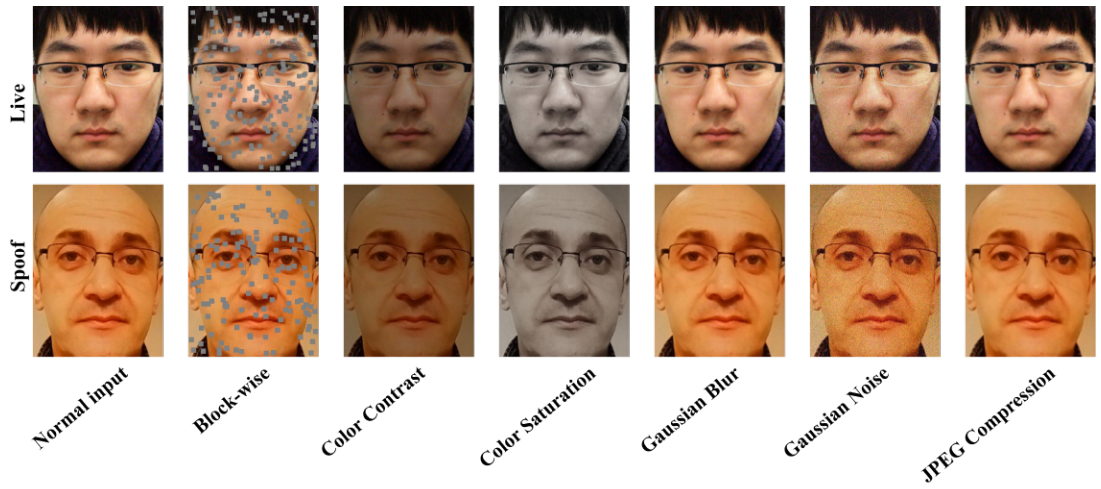

Figure 6. Illustration of six corruption types applied on live and spoof faces in OULU-NPU dataset.

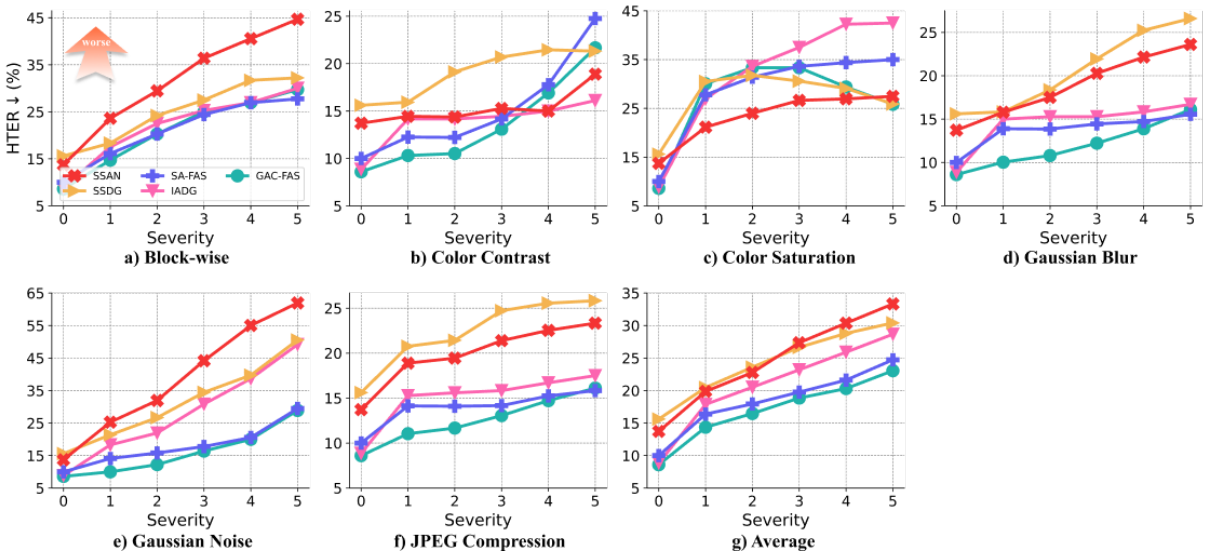

Figure 7. HTER performance (%) of DG spoof detectors under various image corruptions with different severity levels [24]. The experiment are conducted on ICM→O with the corruptions are applied on OULU-NPU dataset.

D. Limitations and Future Works

While our proposed method has achieved SoTA performance across various experiments, we acknowledge two limitations in our work. First, the training dataset requires domain labels to derive ascending points, which may limit its applicability in the in scenarios where training data from multiple sources are combined. Second, although our method maintains comparable computational demands to other methods during the validation phase, GAC-FAS could incur higher computational costs during training when handling a large number of domains as the rising number of ascending points.

In our forthcoming research, we aim to reduce the number of ascending points by exploring similarities across domains. This endeavor includes developing a more efficient regularization approach to gain deeper insights into generalization updates. Notably, our proposed method, which employs a SAM-based optimizer, demonstrates parallels in generating domain-specific gradients with meta-learning techniques [5, 55, 63], albeit in a contrasting direction. While meta-learning methods require additional domain-specific gradient steps and may underperform compared to our approach, the potential synergy of combining ascending vectors from our GAC-FAS with descending vectors from meta-learning promises further enhancements in domain generalization. Our future research will concentrate on investigating these synergistic possibilities.