RoGA: Towards Generalizable Deepfake Detection through Robust Gradient Alignment

Lingyu Qiu[1,2], Ke Jiang[1,2], Xiaoyang Tan[1,2†]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
{qiulingyu,ke_jiang,x.tan}@nuaa.edu.cn

Abstract—Recent advancements in domain generalization for deepfake detection have attracted significant attention, with previous methods often incorporating additional modules to prevent overfitting to domain-specific patterns. However, such regularization can hinder the optimization of the empirical risk minimization (ERM) objective, ultimately degrading model performance. In this paper, we propose a novel learning objective that aligns generalization gradient updates with ERM gradient updates. The key innovation is the application of perturbations to model parameters, aligning the ascending points across domains, which specifically enhances the robustness of deepfake detection models to domain shifts. This approach effectively preserves domain-invariant features while managing domain-specific characteristics, without introducing additional regularization. Experimental results on multiple challenging deepfake detection datasets demonstrate that our gradient alignment strategy outperforms state-of-the-art domain generalization techniques, confirming the efficacy of our method. The code is available at https://github.com/Lynn0925/RoGA .

Index Terms—Deepfake Detection, Face Forgery Detection

I. Introduction

Highly realistic forgery face images generated by deep-learning methods, such as Deepfake [1], pose a huge threat to social security [2]. Therefore, numerous detection methods [3]–[5] have emerged as a defense technology with the help of deep networks to safeguard over the past time.

Although these previous deepfake detection methods have explored the traces left by forgery from various aspects, they often greatly suffer from poor generalization when exposed to unseen samples at the test stage, especially those generated by unknown manipulation methods, i.e., samples from different data domains. Such limitation significantly hinders the practical application of current deepfake detection methods. To address this issue, recent works [6] introduce prior assumptions over the domain, such as statistics of latent features, and extract the domain-invariant features to enhance the generalization over cross-domain manipulations through techniques like style-mixture [7]. However, such assumptions are often too strong and hard to satisfy in practice [8], which hinders these methods' generalization when deploying practically. Besides, additional regularization is often harmful to the convergence to the optimal solution of the empirical risk minimization (ERM) frameworks. These two points motivate
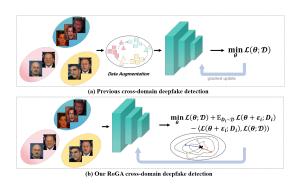

Fig. 1. Comparison among previous cross-domain deepfake detection based on domain adaption such as data/feature augmentation. Our approach in (b) aligns the perturbed gradients, enhancing robustness and generalization.

us to develop a general and regularization-free method to deal with cross-domain deepfake detection.

In this work, we empirically attribute the challenges of cross-domain deepfake detection to the overfitting of neural network parameters to specific domains, resulting in difficulties in cross-domain generalization during the testing phase. This causes neural networks, as illustrated in (a) of Fig.1, learned by traditional methods to focus more on domain-specific features while neglecting broader patterns. Consequently, the learned models would perform well in specific domains, such as certain domains within the dataset, but exhibit poor performance in other domains, particularly when encountering data from unseen domains during testing.

To address this issue, we introduce a novel method in the domain generalization scenario of general face forgery detection, which we term as **Robust Gradient Alignment** (RoGA) for deepfake detection. The key innovation of RoGA is to spontaneously apply perturbations to the model parameters and align the ascending points on each domain during the gradient updates, as is illustrated in (b) of Fig.1. This unique approach guides deepfake detection models to be robust to deal with domain shift, as it can effectively preserve the domain-invariant features while appropriately handling domain-specific characteristics. To be specific, this method attaches certain degrees of perturbation onto the learned models' parameters during the gradient descent process in the parameter space. This perturbation can help the model converge to a flatter local minimum, hence enjoying better robustness. In other words,

---

[†] Corresponding author.

this approach makes it more challenging for the learned model parameters to get trapped in the local optimal solution corresponding to a specific domain, thereby alleviating the overfitting phenomenon of the learned model to that particular domain.

Although our method shares similarities with the recent SAM [9] approach - both apply perturbations to model parameters to find a flatter minimum - there are key differences between our basic idea and SAM: (1) SAM does not consider the gradient alignment between different domains, which can influence the direction of the perturbation gradient during the descent step. (2) To our knowledge, parameter-robust methods like SAM have yet to be applied in the context of deepfake detection.

To further address the aforementioned issues, we introduce a novel objective in domain generalization scenario in deepfake detection which is designed to guide the detectors towards an optimal flat minimum towards robust domain shift. The objective can be divided into two parts: 1) the objective should aim for an optimal low minima under perturbation during the training process. 2) Considering the perturb of different domain's alignment, the gradient updates for each domain should be aligned with each other. This dual objective enables our model to learn a more stable local minimum and more robust domain shifts over different face forgery datasets. All in all, the main contributions of our work are concluded as follows:

- We provide a novel perspective on cross-domain deepfake detection. Rather than designing distinct representation learning approaches to extract domain-invariant features, as is common in previous work, we enhance generalization and domain transferability by guiding the model toward the optimal flat and robust minimum by adding perturb during gradient updates.
- We propose a new training objective: During gradient updates at the ascending point, we keep the generalization gradients of each domain consistent with the empirical risk minimum gradient update, which is conducive to domain generalization in deepfake detection.
- We demonstrate that our approach outperforms well-established baselines in performance across a range of popular deepfake detection evaluation protocol settings.

II. RELATED WORK

In deepfake detection, domain generalization methods based on invariant representation learning [10] have been widely used. A common method is through data augmentation [11]–[13], such as SLADD [14] dynamically synthesizing data through adversarial methods, while SBI [15] swapping with the identity of the same person. On the other hand, there are also studies that solve the new problem of general face forgery detection through meta-learning [16], [17] and multi-task settings [18]. RECCE [3] learns realistic face representations by incorporating self-supervised modules. Previous work, which often uses auxiliary modules [5], [19] to remove domain-

specific features, struggles to generalize to unseen domains due to uncertainties in achieving flat loss regions during training.

As for the optimization strategy of domain gradient consistency, ConfR [20] introduces a novel learning objective to reduce gradient conflicts between domains and extract common features across forgeries, but it overlooks the robust generalization to unknown domains. In contrast, our approach leverages the inherent sharpness of a model within specific domains by aligning these models. In summary, our goal is to construct a more robust and universally generalizable model for deepfake detection.

III. PRELIMINARIES

A face forgery detection problem could be formulated as a binary classification problem. To be specific, the observed data (x, y) are assumed to be sampled from a fixed but unknown joint distribution $P(X, Y)$. To model this relationship, we assume that there exists a neural network $f : X \rightarrow Y$, whose weight parameters are represented by θ. Its purpose is to learn a set of parameters $\hat{\theta}$ through the source dataset D to determine whether the input x is *real* or *fake*. Let $\mathcal{L}$ be some loss function, then the standard learning objective of empirical risk minimization can be defined as follows:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim P(X,Y)}[\mathcal{L}(f(x; \theta), y)] \qquad (1)$$

To generalize this formulation to the cross-domain setting, suppose there exists a domain distribution $P(\mathcal{D})$, then we can sample a set of domain d on it, i.e., $D \sim P(\mathcal{D})$.

According to such understanding of domain information, the data distribution could be divided by different domains, i.e., $P(X, Y) = \mathbb{E}_{D \sim P(\mathcal{D})} P(X, Y|D)$. Traditional methods [17], [18], [20] formulate the objective of cross-domain face forgery detection in an expected way, i.e.,

$$\min_{\theta} \mathbb{E}_{D \sim P(\mathcal{D})} \mathbb{E}_{(x,y) \sim P(X,Y|D)}[\mathcal{L}(f(x; \theta), y)] \qquad (2)$$

The ERM objective function optimizes average performance across all domains but may lead to overfitting on shortcut features and convergence to a sharp minimum, impairing generalization to other domains. Inspired by SAM [9], we introduce perturbations during the parameter update process to guide the model toward a flatter minimum, enhancing its robustness, as demonstrated in previous studies.

IV. ROBUSTNESS GRADIENT ALIGNMENT

In this section, we propose a method named Robustness Gradient Alignment (RoGA). Specifically, we first introduce the expected perturbation optimization on ERM inspired by SAM [9]; the second part acts as a conservative term to ensure that the perturbation in the optimization direction will not differ too much from the original one.

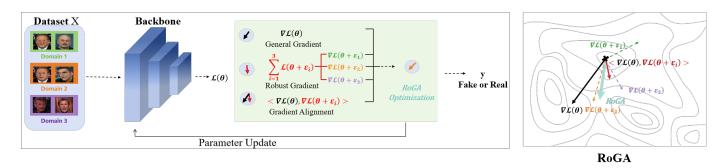Subsequently, we analyze it and propose an implementable algorithm.

Fig. 2. The left part illustrates the framework of RoGA. The right part demonstrates the basic idea of RoGA, where it aligns the gradients with different perturbations during training, hence reducing the risk of overfitting to specific domain patterns.

A. Robustness Gradient Optimization

For a single domain, we use the following SAM [9] objective:

$$\min_{\theta} \max_{\|\epsilon_i\| \leq \rho} \mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i) \quad (3)$$

where ϵ_i denotes some perturbation imposed on the parameter θ for the $i-$th domain. The objective penalizes the model by its 'sharpness' or sensitivity to the small perturbation ϵ_i at θ. Using the Taylor expansion around ϵ_i, we can transform the inner maximization in Eq. 3 into a linearly constrained optimization, yielding the following solution:

$$\epsilon_i^* = arg \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon; \mathcal{D}_i) \approx \rho \frac{\nabla \mathcal{L}(\theta; \mathcal{D}_i)}{\|\nabla \mathcal{L}(\theta; \mathcal{D}_i)\|} \quad (4)$$

With this, we construct the following objective for our cross-domain deepfake detection, which basically seeks to minimize the empirical robust loss function over multiple (K) domains,

$$\min_{\theta} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(\theta + \hat{\epsilon}_i; \mathcal{D}_i) \text{ where } \hat{\epsilon}_i \stackrel{\triangle}{=} \rho \frac{\nabla \mathcal{L}(\theta; \mathcal{D}_i)}{\|\nabla \mathcal{L}(\theta; \mathcal{D}_i)\|} \quad (5)$$

It is noteworthy that there exists an important difference between the above objective function (Eq. 5) and the original (Eq. 2). That is, the original objective was a "strong requirement," which means that the model was expected to simultaneously interpret data from multiple domains. In reality, achieving this goal is quite challenging and is prone to be overfitting. However, our new objective function (Eq. 5) relaxes this requirement. It merely demands that the model learn an optimal shared position in the parameter space, allowing it to interpret data from different domains after a single-step gradient adjustment. This relaxation is more reasonable than the original strong-fitting criterion.

B. Domain Aware Gradient Alignment

To further improve the performance, we consider the issue due to gradient misalignment caused by domain artifacts and data imbalance in multidomain settings. As Fig. 2 shows, domain-specific perturbations (e.g., $\nabla \mathcal{L}(\theta + \epsilon_1; \mathcal{D}_1)$) may significantly deviate from the main gradient direction $\nabla \mathcal{L}(\theta; \mathcal{D})$, compromising optimization stability.

To address this issue, we require that the perturbed gradient $\nabla \mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i)$ for each domain remain aligned with its empirical risk gradient $\nabla \mathcal{L}(\theta; \mathcal{D}_i)$. Finally, we obtain the following Robustness Gradient Alignment (RoGA) optimization loss:

$$\frac{1}{K} \sum_{i=1}^{K} [\mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i) - \alpha \langle \nabla \mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i), \nabla \mathcal{L}(\theta; \mathcal{D}_i) \rangle] \quad (6)$$

where α is the balance coefficient. Note that this is a general objective - as shown later, it can be easily incorporated into any cross-domain fake face detection algorithms. During the optimization process, we decoupled ϵ from the parameters θ by first estimating the ϵ value with the current θ_t and then update the θ_{t+1} value by stochastic gradient decent.

V. EXPERIMENTAL RESULT

A. Experiment Settings

Datasets. To evaluate the generalization of our method, we conduct experiments on widely used deepfake datasets: FaceForensics++ (FF++) [21], Deepfake Detection Challenge (DFDC) [22], CelebDF (v1, v2) [23], DFDCP [24], and UADFV [25]. FF++ includes four manipulation methods—DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTexture (NT)—with two compression levels: low quality (c23) and high quality (c40).

Training Details. ResNet34 [26] serves as the baseline architecture for feature learning, with parameters initialized via ImageNet pre-training [27]. The model is optimized using SGD [28] as the base optimizer with a learning rate of 0.005. The hyperparameter α is set as 0.0002, while $\rho = 0.1$.

Evaluation Metrics. For a comprehensive evaluation, we evaluate the performance of the detector at image-level and report Area Under Curve (AUC) and Accuracy (ACC) as primary metrics. Average Precision (AP) and Equal Error Rate (EER) are also included for additional insights. We ensure a fair comparison by adhering to the experimental settings defined by DeepfakeBench [29].

B. Domain Generalization Evaluation

Cross-Datasets Evaluation To assess the generalization capacity of our method, we follow a standard cross-dataset evaluation protocol. The models are trained on FF++(c23) [21] and tested on unseen datasets, including Celeb-DF [23]

TABLE I
COMPARATIVE PERFORMANCE FOR VARIOUS DOMAIN-ADAPTION-BASED METHODS UNDER CROSS-DATASET EVALUATION USING AUC(%) VALUES. WE
DIVIDE THE METHODS INTO FOUR CATEGORIES, "NAIVE" REPRESENTS WITHOUT USE OF DOMAIN GENERALIZATION, AND THE OTHERS REPRESENT THE
CATEGORIES OF DOMAIN GENERALIZATION. THEN, WE BOLD THE **TOP.1** METHODS WITH THE BEST CROSS-DOMAIN PERFORMANCE.

Method	Detector	Backbone	CDF-v1	CDF-v2	UADFV	DFDC	DFDCP
Naive	Meso4 [30]	MesoNet	0.736	0.609	0.715	0.556	0.599
Naive	CNN-Aug [26]	ResNet	0.742	0.703	0.874	0.636	0.617
Naive	Xception [31]	Xception	0.779	0.737	0.937	0.708	0.737
Naive	EfficientB4 [32]	EfficientNet	0.791	0.749	0.947	0.696	0.728
Representation Learning	SPSL [5]	Xception	0.815	0.765	0.942	0.704	0.741
Representation Learning	CORE [10]	Xception	0.780	0.743	0.941	0.705	0.734
Representation Learning	Recce [3]	Designed	0.768	0.732	0.945	0.713	0.734
Data Augmentation	SRM [11]	Xception	0.793	0.755	0.942	0.700	0.741
Data Augmentation	UCF [12]	Xception	0.779	0.753	0.953	0.719	0.759
Data Augmentation	AdaForensics [13]	ResNet	0.869	0.793	0.955	0.747	0.779
Learning Strategy	MLDG [17]	Xception	0.641	-	-	0.682	-
Learning Strategy	Multi-task [18]	Xception	0.609	-	-	0.682	-
Learning Strategy	LTW [16]	Xception	0.609	-	-	0.690	-
Learning Strategy	ConfR [20]	EfficientNet	0.873	-	-	**0.803**	**0.828**
Ours(RoGA)	ResNet34	ResNet	**0.877**	**0.858**	**0.959**	0.751	0.753

and DFDC [22], serving as target domains to evaluate domain robustness. Table I shows that our method outperforms state-of-the-art approaches, achieving the highest AUC of 0.877 on Celeb-DF and 0.751 on DFDC. The results highlight the efficacy of our multi-source gradient alignment strategy, which enhances domain-agnostic feature learning without introducing additional complexity. Compared to domain adaptation and data augmentation-based methods, our approach demonstrates superior robustness across unseen datasets.

Muli-source Cross-Manipulation Evaluation To verify the effectiveness of the proposed Domain-Aware Robustness Gradient Alignment(RoGA), we conducted experiments on the multi-source forgery method of FF++ [21]. Specifically, models are trained on three forgery methods while incorporating domain labels and tested on the remaining forgery method as the target domain. This setup simulates diverse manipulation transformations originating from identical source images, thus providing an evaluation of cross-manipulation generalization. Table II reports the experimental results consistently outperform competitive baselines, particularly achieving remarkable AUC of 98.08% for GID-FS and 90.08% for GID-DF, which are **11.78%** and **3.98%** higher than the second detector, respectively. These results demonstrate the effectiveness of gradient alignment in capturing subtle manipulative variations, significantly boosting the model's generalization to unseen forgery methods. In summary, our approach introduces a robust optimization strategy, advancing cross-domain deepfake detection with superior performance across diverse datasets and forgery techniques.

C. Ablation Study

1) Effects of proposed objectives: Next, we conducted experiments on the FF++(c23) benchmark and selected ResNet34 [26] as the backbone to investigate the effectiveness of the two objectives of **RoGA**. Specifically, (a) denotes the baseline optimizer (SGD [28]), (b) incorporates the robust gradient term $\frac{1}{K}\sum_{i=1}^{K}\mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i)$, and (c) adds the domain-aware align-

TABLE II
MULTI-SOURCE EVALUATION RESULTS ON ACC(%)/AUC (%)."GID-DF"
MEANS TRAINING ON THE FF++ EXCLUDED DF WHILE TESTING IN DF.

Methods	GID-DF	GID-F2F	GID-FS	GID-NT
MLDG [17]	67.2/73.1	58.1/61.7	58.1/61.7	56.9/60.7
LTW [16]	69.1/75.6	65.7/72.4	62.5/68.1	58.5/60.8
DisGRL [33]	77.3/86.1	75.8/84.3	76.9/86.3	**66.3/72.8**
Ours(RoGA)	**80.13/90.08**	**76.34/86.54**	**93.41/98.08**	59.43/70.52

ment term $\langle \nabla\mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i), \nabla\mathcal{L}(\theta; \mathcal{D}_i)\rangle$, with (d) representing the complete method, it achieve an optimal AUC performance of **99.30%**, surpassing individual gains of 0.27% and 0.10%.

TABLE III
ABLATION STUDY: ANALYSIS OF THE TWO OBJECTIVES WHILE TRAINING
ON THE FF++(C23) AND TESTING ON THE DEEPFAKES.

	AUC ↑	ACC ↑	Loss ↓	EER ↓
(a) Base-optimizer	97.65	90.00	0.32	7.83
(b) + $\mathcal{L}(\theta + \hat{\epsilon}_i; \mathcal{D}_i)$	99.03	93.39	0.18	4.79
(c) + $\langle \nabla\mathcal{L}(\theta + \epsilon_i; \mathcal{D}_i), \nabla\mathcal{L}(\theta; \mathcal{D}_i)\rangle$	99.20	91.61	0.22	5.38
(d) RoGA	99.30	94.25	0.15	4.13

2) Effects of the optimizer selection: To assess the efficacy of our proposed optimizer, we conduct a comparative analysis with various optimization techniques across both cross-manipulation and cross-dataset settings. As shown in Table IV, sharpness-aware optimizers like SAM [9] and ASAM [34] outperform traditional ERM-based optimizers such as Adam [35] and SGD [28]. Our RoGA optimizer, which integrates perturbation regularization and domain-specific gradient alignment, consistently delivers superior performance with a 1.47% ACC improvement over Adam [35] in-domain and a notable 6.1% gain in cross-domain tests.

3) Effects of the backbone: We evaluate the flexibility of our method by testing it with various backbones, including Xception [31], EfficientNetB4 [32], and ResNet34 [26], which are commonly used in deepfake detection. Experiments were conducted on FF++(c23), and evaluated on four manipulation

TABLE IV
ABLATION STUDY: ALTERNATIVE APPROACH USING DOMAIN GRADIENT
FOR DOMAIN GENERATION UNDER CROSS-MANIPULATION AND
CROSS-DATASET EVALUATION

Train Test	FS, F2F, NT DF		FF++ UADFV	
Optimizer	AUC ↑	ACC ↑	AUC ↑	ACC ↑
Adam [35]	92.96	83.83	81.63	55.98
SGD [28]	92.27	84.82	76.45	56.18
SAM [9]	93.18	83.39	81.50	50.63
SAGM [36]	93.05	84.46	85.13	59.86
Ours(RoGA)	**93.63**	**85.30**	**95.95**	**62.08**

Fig. 3. The GradCAM visualizations [37] comparing SRM [39] baseline and our **RoGA**, across four forgery types on FF++(c23).

methods from FF++(c23) and cross-domain datasets UADFV [25]. The results presented in Table V (AUC) unequivocally demonstrate that our method enhances performance without adding any additional model parameters, regardless of the backbone type. Specifically, we observe a **7.21%** average performance increase on the Xception [31] framework. This underscores the remarkable versatility of our approach, which exhibits seamless integration capabilities, thereby offering substantial performance gains across a diverse set of well-established detection frameworks.

TABLE V
ABLATION STUDY: THE AUC(%) VALUES OF ALTERNATIVE APPROACH
USING **OURS ROGA** IN DIFFERENT BACKBONE

Backbone	+Ours	DF	F2F	FS	NT	UADFV
Meso4 [30]	✗	75.89	70.38	61.31	66.95	64.49
	✓	79.84	75.57	64.57	67.61	71.78
Xception [31]	✗	91.51	91.86	88.64	82.22	88.27
	✓	99.00	99.08	99.06	96.42	91.55
EffcientNet [32]	✗	90.51	88.05	86.98	81.14	83.40
	✓	97.36	97.57	97.53	94.82	87.87
ResNet34 [26]	✗	97.65	97.62	97.81	96.19	85.31
	✓	99.20	98.85	99.05	96.41	95.95

4) Hyparameter Sensitivity: In our proposed RoGA, the hyperparameters α and ρ are critical for performance. Table VI presents the AUC(%) results across intra- and cross-dataset settings on FF++ [21]. The optimal values, $\alpha = 0.001$ and $\rho = 0.1$, consistently yield superior results under both evaluations.

TABLE VI
HYPERPARAMETERS SENSITIVITY OF α AND ρ UNDER BOTH INTRA- AND
CROSS-DATASET SETTINGS WHERE MODEL IS TRAINED ON FF++(C23)

Hyperparam. α	ρ	DF	F2F	FS	NT	UADFV	CelebDF
0.001	0.05	99.20	98.85	99.05	96.41	91.01	81.09
0.002	0.05	99.02	98.75	99.12	95.88	90.44	80.61
0.001	0.1	99.09	98.87	99.39	96.26	94.23	87.70

D. Analysis

To enhance our understanding of the RoGA method, we use interpretable methods such as Grad-CAM [37] and t-SNE [38] for analysis.

Visualizations of the captured artifacts. GradCAM [37] is employed to visualize the regions most critical for the detector, offering interpretability of the model. As shown
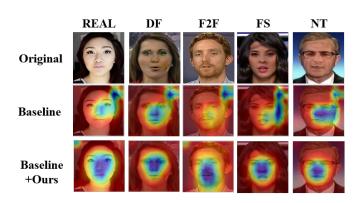
in Figure 3, our RoGA significantly outperforms the SRM baseline [39], which focuses on uniform but limited artifact patterns. In contrast, **RoGA** precisely localizes manipulated facial regions, demonstrating its capability to learn robust and domain-invariant features.

Visualizations of learned latent space. To assess the learned representations, we visualized the latent space of 5,000 test samples from FF++(c23) using t-SNE [38]. As shown in Figure 4, our RoGA method indicates that our enhanced method (right) has indeed learned a much clearer and more distinct decision boundary compared to the unenhanced baseline (left).
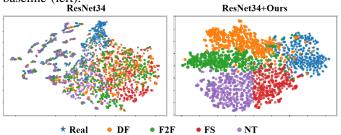

Fig. 4. t-SNE [38] visualization of latent space *w* and *w/o* our **RoGA** when the model is trained on FF++(c23).

VI. CONCLUSION

We propose RoGA (Robustness Gradient Alignment), a novel optimization framework for domain generalization in deepfake detection. RoGA combines gradient perturbation and alignment, with its core innovation being gradient orthogonal alignment in the loss function. This ensures that perturbation and ERM gradients are harmonized, guiding updates towards robust, generalizable minima.

By modeling domain variability via perturbations and resolving inter-domain conflicts through alignment, RoGA effectively enhances cross-domain generalization. Extensive experiments both ablation studies and rigorous comparisons with state-of-the-art methods demonstrate RoGA's adaptability and state-of-the-art performance across diverse architectures, underscoring its significance in advancing optimization strategies for multi-domain deepfake detection.

REFERENCES

[1] Siwei Lyu, "Deepfake detection: Current challenges and next steps," in *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[2] Lingyu Qiu, Ke Jiang, Sinan Liu, and Xiaoyang Tan, "Multi-level distributional discrepancy enhancement for cross domain face forgery detection," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024, pp. 508–522.

[3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.

[4] HuyH. Nguyen, Junichi Yamagishi, and Isao Echizen, "Capsuleforensics: Using capsule networks to detect forged images and videos," *Cornell University - arXiv,Cornell University - arXiv*, Oct 2018.

[5] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[6] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi, "Exploiting style latent flows for generalizing deepfake video detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1133–1143.

[7] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[8] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou, "How does mixup help with robustness and generalization?," *arXiv preprint arXiv:2010.04819*, 2020.

[9] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[10] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao, "Core: Consistent representation learning for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12–21.

[11] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16317–16326.

[12] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22412–22423.

[13] Xiaoke Yang, Haixu Song, Xiangyu Lu, Shao-Lun Huang, and Yueqi Duan, "Adaforensics: Learning a characteristic-aware adaptive deepfake detector," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[14] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18710–18719.

[15] Kaede Shiohara and Toshihiko Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.

[16] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji, "Domain general face forgery detection by learning to weight," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 2638–2646.

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[18] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2019, pp. 1–8.

[19] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020.

[20] Jin Chen, Jiahe Tian, Cai Yu, Xi Wang, Zhaoxing Li, Yesheng Chai, Jiao Dai, and Jizhong Han, "Confr: Conflict resolving for generalizable deepfake detection," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[21] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[22] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) preview dataset," 2019.

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] B Dolhansky, "The dee pfake detection challenge (dfdc) pre view dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[25] Chang M Liy and LYUS InIctuOculi, "Exposingaicreated fakevideosbydetectingeyeblinking," in *2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE*, 2018.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[28] David Pollard, *Convergence of stochastic processes*, Springer Science & Business Media, 2012.

[29] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu, "Deepfakebench: a comprehensive benchmark of deepfake detection," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 4534–4565.

[30] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[31] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[33] Zenan Shi, Haipeng Chen, Long Chen, and Dong Zhang, "Discrepancy-guided reconstruction learning for image forgery detection," *arXiv preprint arXiv:2304.13349*, 2023.

[34] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5905–5914.

[35] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang, "Sharpness-aware gradient matching for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3769–3778.

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[38] Geoffrey E. Hinton and Sam T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, Eds. 2002, pp. 833–840, MIT Press.

[39] Yang He, Ning Yu, Margret Keuper, and Mario Fritz, "Beyond the spectrum: Detecting deepfakes via re-synthesis," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Aug 2021.