Supplementary Materials

This supplementary material offers a detailed examination of the methodologies and results that underpin the experiments in our study. It is designed to provide comprehensive information to validate the findings and ensure reproducibility and transparency.

The content is organized as follows:

- **Section 6:** A comprehensive and detailed description of the experimental setup and methodology.
- **Section 7:** A formal presentation of the proposed algorithm, including its pseudocode representation.
- **Section 8:** In-depth evaluation of model performance under various transfer attack scenarios.
- **Section 9:** Detailed results from adversarial training using synthetic data for robustness assessment.
- **Section 10:** Extensive exploration of techniques for generating foreground-background attention maps.
- **Section 11:** Thorough examination of the model's robustness under varying levels of attack intensity.

6. Experimentation Details

In this part, we provide a comprehensive overview of the experimental setup.

6.1. Training Setup

6.1.1. CIFAR-10 and CIFAR-100

For the CIFAR-10 and CIFAR-100 [9] datasets, we employ the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay factor of 5×10^{-4}. The initial learning rate is set to 0.1, and models undergo 100 epochs of training with the learning rate reduced by a factor of 0.1 at the 80th and 90th epochs.

For adversarial training on these datasets, we create adversarial examples through a 10-iteration attack where the maximum ℓ_∞-norm of the adversarial perturbation is limited to $\epsilon = 8/255$, using a step size of $\alpha = 2/255$. To ensure reliability and fairness, we align the inverse adversarial example generation process with the UIAT settings [5], employing the same loss function and perturbation constrained by $\epsilon = 4/255$.

6.1.2. ImageNet-1K

For the ImageNet-1K dataset, we strictly follow the training protocol established in [14] to ensure fair comparison with existing methods. Specifically, we implement a 2-iteration PGD over 50 epochs, and set $\epsilon = 4/255$ and $\alpha = 1/255$ for adversarial perturbations. This adherence to established benchmarking parameters enables direct and meaningful comparison with state-of-the-art approaches evaluated on this challenging large-scale dataset.

6.1.3. Common Settings

The hyperparameters λ_1 and λ_2 are consistently set to 1.0 across all datasets. To maintain experimental fairness, all comparative methods undergo identical training strategies with the same number of epochs and optimizer configurations. For Grad-CAM to quantify the spurious correlation bias, we use a pre-trained ResNet-18 model under simple training, introducing minimal additional training cost.

All experiments were conducted on a single NVIDIA Tesla A100. Notably, our method does not employ label smoothing techniques to avoid potential conflicts with efforts to mitigate spurious correlation bias, as label smoothing can blur class boundaries and undermine the precision required for robust decision-making. Additionally, unlike UIAT, our approach does not utilize momentum terms to stabilize the generation process of inverse adversarial examples, instead focusing solely on the intrinsic properties and underlying dynamics of our method. Comprehensive training details are available in the supplementary material.

6.2. Evaluation Setup

Our evaluation consists of two main components: robustness performance and robust generalization. To evaluate robustness performance, we use PGD [12], C&W [2], and AutoAttack (AA) [3] within the ℓ_∞-norm. AutoAttack includes several attack methods such as APGD-DLR [3], APGD-CE [3], FAB [4], and Square [1]. Adversarial attacks are generated using a step size of $\alpha = 2/255$ and a specified maximum ℓ_∞-norm. Note that "Clean" indicates natural examples unaffected by adversarial perturbations.

To evaluate robust generalization, we introduce the concept of the robust generalization gap [21] denoted as the "Robust Gap" quantifying the difference in robust performance between training and test sets under adversarial attacks. A smaller robust gap indicates improved robust generalization, reflecting reduced vulnerability to robust overfitting in the adversarial training model.

7. Algorithm Details

Algorithm 1 outlines the complete workflow of our proposed *Debiased High-Confidence Adversarial Training* (DHAT) framework, which systematically mitigates spurious correlations in adversarial training through two primary components: Debiased High-Confidence Logit Regularization (DHLR) and Foreground Logit Orthogonal Enhancement (FLOE).

To efficiently compute attention maps, we employ Grad-CAM (Line 11) due to its computational efficiency. However, our framework is compatible with alternative saliency-based methods, such as SAM [8], which can provide enhanced performance at a higher computational cost. The extraction of background features (Lines 13-14) is performed

Algorithm 1 Debiased High-Confidence Adversarial Training (DHAT)

Require: Training dataset $(X, Y) = \{(x_i, y_i)\}_{i=1}^{N}$; Network parameters θ; Perturbation budget ϵ; Hyperparameters λ_1, λ_2; Attention threshold ω.
Ensure: Robust model parameters θ^*

1: Initialize network parameters θ
2: **for** each epoch **do**
3: **for** each mini-batch (x, y) **do**
4: // Generate adversarial examples
5: $\hat{x} \leftarrow \arg\max_{||x'-x||_p \leq \epsilon} \mathcal{L}_{AT}(f_\theta(x'), y)$
6: $\hat{z} \leftarrow f_\theta(\hat{x})$
7: // Generate inverse adversarial examples
8: $\check{x} \leftarrow \arg\min_{||x'-x||_p \leq \epsilon} \mathcal{L}_{Inv}(f_\theta(x'), y)$
9: $\check{z} \leftarrow f_\theta(\check{x})$
10: // Compute attention maps using a selected method $\mathcal{A} \in \{$Grad-CAM, Integrared-Grad, SAM, *etc.*$\}$
11: $M \leftarrow \mathcal{A}(x)$ {Different attention map computation techniques}
12: // Extract background features from inverse adversarial examples
13: $[\check{x}_{(B)}]_{(i,j)} \leftarrow \mathbb{I}_{(M_{i,j}<\omega)} \cdot \check{x}_{(i,j)}$
14: $\check{z}_{(B)} \leftarrow f_\theta(\check{x}_{(B)})$
15: // Compute debiased high-confidence logits
16: $\check{z}^* \leftarrow \check{z} - \check{z}_{(B)}$
17: // Compute DHLR loss
18: $\mathcal{L}_{DHLR} \leftarrow \mathcal{L}_{KL}(\phi(\check{z}^*)||\phi(\hat{z}))$
19: // Compute FLOE loss
20: $\mathcal{L}_{FLOE} \leftarrow -|\check{z} - \frac{\check{z} \cdot \check{z}_{(B)}}{|\check{z}_{(B)}|^2} \cdot \check{z}_{(B)}|_p$
21: // Compute total loss
22: $\mathcal{L}_{DHAT} \leftarrow \mathcal{L}_{AT}(\hat{z}, y) + \lambda_1 \cdot \mathcal{L}_{DHLR} + \lambda_2 \cdot \mathcal{L}_{FLOE}$
23: // Update network parameters
24: $\theta^* \leftarrow \theta - \nabla_\theta \mathcal{L}_{DHAT}$
25: **end for**
26: **end for**
27: **return** θ^*

using an adaptive threshold ω, ensuring the identification of non-discriminative regions that contribute to spurious correlations. The overall training objective (Line 23) integrates standard adversarial training with our proposed debiasing regularization terms, weighted by λ_1 and λ_2.

Despite incorporating additional minimal computational steps, the DHAT framework introduces only a marginal increase in resource consumption compared to standard adversarial training while significantly enhancing model robustness. Crucially, DHAT achieves these improvements without compromising clean-data accuracy, making it a highly practical and scalable solution for real-world adversarial defense applications.

8. Performance Under Transfer Attack

In this part, we evaluate our proposed model's performance under transfer attacks and compare it with the UIAT [5]. Transfer attacks evaluate a model's robustness by testing its performance against adversarial examples generated from different source models, which simulate real-world conditions where attackers may use varied strategies. This testing is essential to ensure that a model's adversarial defenses generalize effectively beyond its training environment and are robust against diverse attack methods. By evaluating the performance under transfer attacks, we ensure that our model is robust not only to attacks generated by its own architecture but also to those from different models, providing a more comprehensive measure of robustness.

Table 8 presents the robust accuracy of various models under transfer attacks with an ℓ_∞-norm perturbation of $\epsilon = 8$. The table compares the performance of the UIAT method with that of our proposed model across different target models and transfer attack types.

8.1. Robustness Performance

Our model consistently outperforms UIAT across all transfer adversarial attacks and target models. For example, when the WRN28-10 model is trained with our proposed DHAT and subjected to AA attacks generated from ResNet-50, VGG16, and Inc-V3 source models, the defense success rate improves by 1.21%, 1.14%, and 1.44% compared to UIAT, respectively. This demonstrates the model's robustness, highlighting its effectiveness not only in specific attack scenarios but also across various model architectures.

8.2. Robustness Across Various Attack Types

Our proposed model DHAT demonstrates enhanced robustness, particularly against more challenging attacks such as PGD-50 and C&W. For instance, when using the ResNet-18 as the source model, our method achieves improvements of up to 1.88% and 2.39% under PGD-50 and C&W attacks from WRN28-10, respectively, compared to UIAT. This highlights the model's superior capability to withstand adversarial perturbations.

8.3. Generalization Across Various Source Models

Our model exhibits strong performance across a range of source models, including VGG-16, WRN28-10, and ResNet-18. This indicates that the improved robustness of our model is not limited to specific attacks generated from source architectures but generalizes effectively across different adversarial settings.

9. Performance with Generated Data

We employ Diffusion Denoising Probabilistic Models (DDPM) [6] to generate an additional 50K samples for

Table 8. Transfer attack accuracy (%) in the single-model transfer scenario. The **number** in bold indicates the best accuracy.

Attack ($\epsilon = 8$)	Performance of UIAT / DHAT								
	Target: VGG-16			Target: WRN28-10			Target: ResNet-18		
	⇒ ResNet-50	⇒ Inc-V3	⇒ WRN28-10	⇒ ResNet-50	⇒ VGG16	⇒ Inc-v3	⇒ WRN28-10	⇒ VGG16	⇒ Inc-v3
FGSM	63.42/**64.83**	63.44/**64.51**	63.73/**64.88**	77.11/**78.73**	78.08/**79.27**	77.84/**78.69**	73.01/**74.80**	74.00/**75.03**	72.43/**73.18**
PGD-10	53.09/**54.46**	52.81/**53.96**	54.17/**55.36**	61.93/**62.87**	64.73/**65.70**	61.40/**62.56**	58.56/**60.30**	59.91/**60.89**	57.82/**58.75**
PGD-20	52.73/**54.25**	52.63/**53.70**	53.80/**55.29**	61.52/**62.68**	64.60/**65.65**	61.19/**62.38**	58.13/**60.04**	59.80/**60.78**	57.48/**58.42**
PGD-50	52.68/**54.25**	52.62/**53.67**	53.77/**55.27**	61.51/**62.65**	64.56/**65.68**	61.17/**62.24**	58.15/**60.03**	59.78/**60.83**	57.45/**58.46**
C&W	51.90/**53.44**	51.38/**52.61**	55.31/**56.35**	61.05/**62.02**	63.32/**64.40**	60.76/**61.98**	57.73/**60.12**	57.44/**58.74**	57.71/**59.25**
AA	56.31/**57.94**	56.02/**56.98**	59.65/**60.77**	67.18/**68.39**	71.86/**73.00**	64.39/**65.83**	62.36/**63.19**	65.19/**66.07**	60.69/**61.94**

Table 9. Comparison of robustness (%) and robust generalization gap (%) for models trained on generated data. The **bolded numbers** indicate the best performance.

CIFAR-10	ResNet-18			WRN28-10		
Method	Clean↑	AA↑	Robust Gap↓	Clean↑	AA↑	Robust Gap↓
MART [17]	83.45	49.45	2.91	84.26	51.95	8.73
AWP [20]	83.78	50.79	2.03	84.10	53.29	6.11
FSR [7]	83.19	50.53	2.35	83.88	53.03	7.05
CFA [19]	84.97	50.85	2.98	85.81	53.35	8.94
UIAT [5]	85.10	52.09	3.04	86.73	54.59	9.41
SGLR [10]	86.35	51.27	2.75	87.72	53.77	11.25
DHAT (Ours)	**87.62**	**54.42**	**0.83**	**88.94**	**56.92**	**2.48**
CIFAR-100	ResNet-18			WRN28-10		
Method	Clean↑	AA↑	Robust Gap↓	Clean↑	AA↑	Robust Gap↓
MART [17]	54.73	27.70	2.88	55.87	30.20	8.86
AWP [20]	57.55	29.33	2.37	59.71	31.83	7.13
FSR [7]	58.10	28.94	2.47	59.03	30.44	7.40
CFA [19]	60.13	28.85	3.01	61.56	29.61	9.13
UIAT [5]	59.92	28.48	4.47	60.24	30.98	14.32
SGLR [10]	61.25	29.10	4.15	62.39	30.15	17.05
DHAT (Ours)	**63.14**	**32.21**	**0.98**	**64.80**	**34.71**	**2.96**

both CIFAR-10 and CIFAR-100 datasets following the [18]. This synthetic data is then used to augment the training of both our method and all baseline models. The performance comparisons, as shown in Table 9, illustrate the impact of additional data on robustness and generalization.

9.1. Robustness Performance

The results in Table 9 demonstrate that our method outperforms baseline methods in terms of robustness when trained with the additional synthetic data. This improvement highlights the effectiveness of our approach in leveraging extra data to enhance model robustness, particularly under adversarial conditions.

9.2. Generalization Performance

We observe that, compared to the results in Table 1, most models trained with the additional data exhibit a reduced Robust Gap, indicating improved generalization. However, both UIAT and SGLR exhibit limited improvements in robust generalization. These methods rely on spurious correlations during training, which hinders their generalization performance even with the added data. Although UIAT and SGLR benefit from enhanced robustness due to increased data diversity, their robust generalization remains suboptimal, likely due to their dependence on non-essential features. This finding underscores the unique advantage of our method in achieving both robust accuracy and generalization with enriched datasets.

Table 10. Comparison of robustness (%) and robust generalization gap (%) for using various attention map generation techniques using WRN28-10 on the CIFAR-10.

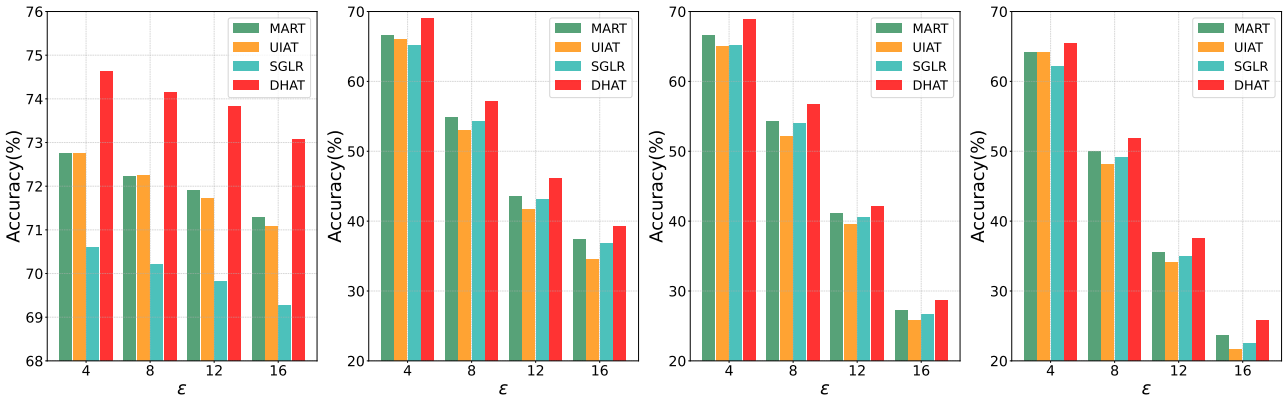
Method	Clean↑	PGD-10↑	C&W↑	AA↑	Robust Gap↓
-	82.94	58.66	54.11	52.17	7.92
Grad-CAM [13]	83.95	60.49	55.27	53.10	3.51
Integrated-Grad [15]	83.97	60.35	55.18	53.04	3.68
SOLO [16]	84.26	61.60	56.74	54.93	3.09
SAM [8]	85.65	62.44	58.10	56.38	2.46

10. Impact of Foreground-Background Recognition Techniques

In this part, we investigate the influence of different attention map generation techniques on the performance of Debiased High-Confidence Logit Regularization (DHAT), as seen in Table 10. Our exploration extends beyond the initially employed Grad-CAM method to encompass a variety of attention map generation approaches, thereby providing a comprehensive analysis of their effects on model robustness and generalization.

10.1. Exploration of Various Attention Map

The primary approach detailed in the main text utilizes Grad-CAM [13], a widely adopted method in weakly supervised object segmentation [11], to extract foreground and background feature maps efficiently. Grad-CAM offers a balance between computational efficiency and effectiveness, making it suitable for large-scale evaluations. However, to understand the broader applicability and potential improvements, we incorporated additional attention map generation techniques, including Integrated Gradients [15],



(a) FGSM Attack (b) PGD-10 Attack (c) PGD-20 Attack (d) C&W Attack

Figure 6. Comparisons with varying ϵ values using ResNet-18 on the CIFAR-10. The x-axis represents the ϵ value, while y-axis represents the robust accuracy (%).

SOLO [16], and SAM [8].

10.2. Evaluation of Advanced Attention Map Models

Incorporating more sophisticated attention map generation models, such as SAM, demonstrated enhanced robustness and reduced robust generalization gaps compared to simpler methods like Grad-CAM. These advanced models provide finer and more accurate segmentation of foreground and background features, which in turn leads to better alignment of high-confidence logits under adversarial conditions. However, this improvement in performance comes at the cost of increased computational overhead. More complex models require greater processing power and longer training times, which may be a limiting factor in resource-constrained environments.

10.3. Trade-Off Between Performance and Computational Efficiency

While the adoption of advanced attention map techniques can yield superior performance metrics, researchers must consider the trade-offs involved. Enhanced models may offer marginal gains in robustness and generalization, but the additional computational resources and time required may not always justify these benefits, especially in applications where real-time processing is essential. Therefore, the choice of attention map generation technique should be informed by the specific requirements and constraints of the deployment scenario.

11. Robustness under Various Magnitude Attacks

In this part, we evaluate the robustness of our proposed model under various levels of adversarial perturbations, quantified by different ϵ values. We compare our model with two baseline methods, MART [17] and UIAT [5], and the state-of-the-art SGLR method [10] across various attack types, including FGSM, PGD-10, PGD-20, and C&W. We aim to demonstrate the superior generalization and robustness of our model under progressively challenging adversarial conditions. Figure 6 shows the accuracy of different methods under various ϵ values for each attack type.

The comparative analysis of our model with MART, UIAT, and SGLR reveals several key insights:

11.1. Robustness Across Various Attack Types

Our model consistently outperforms the baselines and SGLR across all attack types and ϵ values. This consistent superiority underscores our model's robust generalization ability, allowing it to maintain high accuracy even under more challenging adversarial attack conditions.

11.2. Generalization Against Various Attacks

The results suggest that the defensive mechanisms in our model are highly effective in mitigating the impact of various adversarial attacks (*i.e.,* FGSM, PGD-10, PGD-20, C&W). This robustness is especially evident in scenarios with stronger attacks and higher ϵ values, where our model shows superior performance compared to existing methods.

11.3. Practical Applicability

The enhanced robustness of our model across different ϵ values and attack types demonstrates its practical applicability in real-world scenarios, where adversarial perturbations can vary in strength and sophistication.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient

black-box adversarial attack via random search. In *ECCV*, 2020. 1

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *IEEE Symp. Security Privacy*, pages 39–57, 2017. 1

[3] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 1

[4] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020. 1

[5] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving aversarial training. In *CVPR*, 2023. 1, 2, 3, 4

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[7] Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration for adversarial robustness. In *CVPR*, 2023. 3

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 3, 4

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *Toronto, ON, Canada*, 2009. 1

[10] Zhuorong Li, Daiwei Yu, Lina Wei, Canghong Jin, Yun Zhang, and Sixian Chan. Soften to defend: Towards adversarial robustness via self-guided label refinement. In *CVPR*, 2024. 3, 4

[11] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *CVPR*, 2022. 3

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1

[13] 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 *ICCV*, 2017. 3

[14] Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. In *NeurIPS*, 2023. 1

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 3

[16] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 3, 4

[17] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020. 3, 4

[18] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *ICML*, 2023. 3

[19] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *CVPR*, 2023. 3

[20] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 3

[21] Shenglin Yin, Kelu Yao, Sheng Shi, Yangzhou Du, and Zhen Xiao. Again: Adversarial training with attribution span enlargement and hybrid feature fusion. In *CVPR*, 2023. 1