IMPROVING HYPERSPECTRAL ADVERSARIAL ROBUSTNESS UNDER MULTIPLE ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic segmentation models classifying hyperspectral images (HSI) are vulnerable to adversarial examples. Traditional approaches to adversarial robustness focus on training or retraining a single network on attacked data, however, in the presence of multiple attacks these approaches decrease in performance compared to networks trained individually on each attack. To combat this issue we propose an *Adversarial Discriminator Ensemble Network* (ADE-Net) which focuses on attack type detection and adversarial robustness under a unified model to preserve per data-type weight optimally while robustifiying the overall network. In the proposed method, a discriminator network is used to separate data by attack type into their specific attack-expert ensemble network.

1 INTRODUCTION

Current semantic segmentation models are vulnerable to the addition of imperceptible perturbations to the input data Biggio et al. (2013). These perturbations are well-crafted attacks that when added in small amounts to a sample, drastically fool the model and decrease accuracy Szegedy et al. (2013b). To combat this challenge and make the model robust toward attacks, adversarial training is performed on the model. To perform adversarial training on a model, one generates the adversarial examples and then either continually trains the trained model, or mixes the attack examples with non-attacked examples and trains a new model from scratch Goodfellow et al. (2014b); Szegedy et al. (2013c). In this approach, the model performs worse on each individual data attack type due to the updated weights, but better overall. Another approach is to detect the attacked data to avoid it during classification via adversarial detection. In this case, the attacks are often ignored, despite having information that might be valuable to classify Metzen et al. (2017). In this work, we fuse the strengths of these two approaches. To detect and defend against adversarial attacks in one model, a novel Adversarial Discriminator Ensemble Network (ADE-Net) model is proposed. While most models focus on network architecture and robustness, ADE-Net can use any semantic segmentation network while classifying attack type and class type in one unified model.

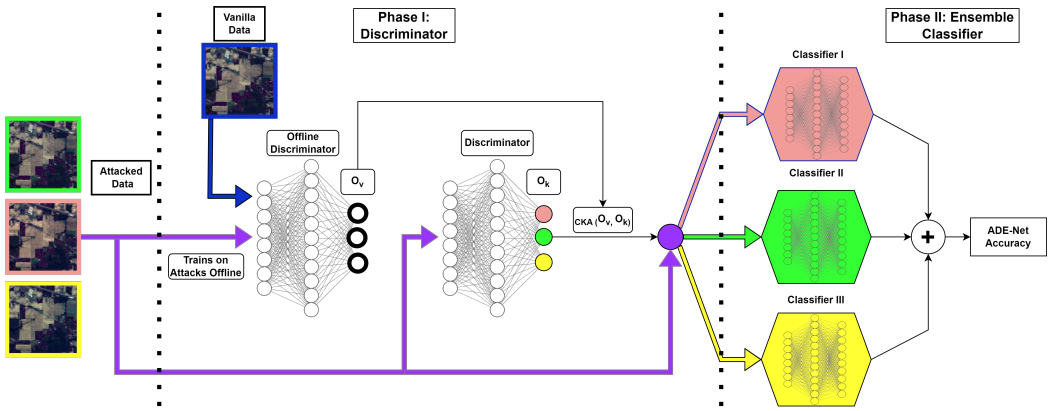

Figure 1: Overview of ADE-Net model structure: **Phase I** contains the Discriminator Network and **Phase II** contains the Ensemble Networks.

2 PROPOSED METHOD: ADE-NET

Notation: Consider vanilla data as $\{x_i^v, y_i^v\}_{i=1}^{n_v}$ with size n_v. Define attack label $k = 1, \ldots, K$ and consider k-th attack data $\{x_i^k, y_i^k\}_{i=1}^{n_k}$ with size n_k. Each attacked point has two labels: class label and attack label, $\{x_i^k, y_i^k, c_i = k\}_{i=1}^{n_k}$. For a classifier F, consider a discriminator F_D that focuses on separating input based on attack type, and ensemble classifiers F_k, which denotes the k-the classifier expert in attack type k. For each ensemble network, ω_k^* is the optimal weight of F_k, and ω_D^* is the optimal weight of F_D on all K attack data. Let $O_v \in \mathcal{R}^{m \times n_v}$ be the logits of passing vanilla on an offline trained $F_D^{offline}$ on attacks, and $O_k \in \mathcal{R}^{m \times n_k}$ be the logits of F_D on single attack k.

Methodology: For ADE-Net, our overall model structure consists of two phases: 1) Discriminator Phase I: Adds adversarial detection to ADE-Net via attack type classification and CKA Kornblith et al. (2019) to separate mixed attack data. 2) Ensemble Phase II: Constituted by a number of networks, equal to n_k, that focus on regular class label classification. Each network is an expert at a particular attack determined by the output label from the discriminator. Figure 1 presents a visual overview of ADE-Net. We propose in ADE-Net that the ensemble classifiers are trained collaboratively with the discriminator focusing on class and attack classification respectively:

$$\omega_D^*, \omega_k^* = \arg\min_{\omega_D, \omega_k} \sum_{k=1}^{K} \alpha_k \mathcal{L}_{\omega_k}(F_k(\mathbf{x}), y) + \sum_{k=1}^{K} \lambda_k \, CKA(O_v, O_k) + \beta \, \mathcal{L}_{\omega_D}(F_D(\mathbf{x}), c) \quad (1)$$

where the first α_k is for Phase II and the remaining λ_k and β terms are for Phase I. Extended mathematical formalization of the loss function can be found in the Appendix in Section C. The overview of our ADE-Net algorithm is shown in Algorithm 1.

# Attack	All = 1.0	λ_k=β=0.1 α_k=10.0	λ_k=β=10.0 α_k=0.1	λ_k=β=1.0 α_k Attack Dependent	λ_k=0 (No CKA) β=α_k=1.0	Baseline
2	**I: 95.29 ± 0.01** **II: 80.37 ± 0.01**	I: 95.19 ± 0.04 II: 80.00 ± 0.01	I: 95.34 ± 0.03 II: 80.11 ± 0.02	I: 95.21 ± 0.04 II: 80.30 ± 0.01	I: 95.00 ± 0.05 II: 79.67 ± 0.02	**78.78 ± 0.08**
3	I: 63.97 ± 0.03 II: 81.57 ± 0.02	I: 64.12 ± 0.04 II: 81.71 ± 0.02	I: 63.97 ± 0.04 II: 81.62 ± 0.03	**I: 63.99 ± 0.03** **II: 81.72 ± 0.02**	I: 63.50 ± 0.04 II: 81.49 ± 0.04	**80.11 ± 0.13**
4	I: 48.03 ± 0.04 II: 82.23 ± 0.12	I: 47.98 ± 0.10 II: 82.06 ± 0.11	I: 47.98 ± 0.09 II: 82.1 ± 0.12	**I: 48.05 ± 0.05** **II: 82.51 ± 0.09**	I: 47.51 ± 0.10 II: 81.66 ± 0.14	**81.37 ± 0.02**
5	I: 50.30 ± 0.03 II: 83.27 ± 0.15	I: 50.40 ± 0.02 II: 83.27 ± 0.14	I: 50.45 ± 0.03 II: 83.17 ± 0.16	**I: 50.30 ± 0.03** **II: 83.55 ± 0.13**	I: 50.40 ± 0.04 II: 83.10 ± 0.16	**82.65 ± 0.22**

Table 1: Experimental results of ADE-Net with Indian Pines dataset. Best performing results for each row are in bold, worst-performing is in red. The α_k hyperparameter is for the ensemble class cross-entropy, λ_k is for the discriminator CKA and β is for the discriminator attack cross-entropy.

3 EXPERIMENTS

For our experiments, we use the HSI dataset Indian Pines (IP). Two additional datasets, Kennedy Space Center (KSC) Amigo (2020), and Houston Labate et al. (2019) are available in Appendix Section D. Two averaged overall accuracies (OA) terms will be reported: Phase I and Phase II accuracy as shown in Figure 1. Phase I Discriminator accuracy captures attack-label classification while Phase II Ensemble accuracy captures class-label accuracy.

Evaluation: We showcase different hyperparameter values to find the experimentally optimal values for better ADE-Net accuracy. We also explore the role of CKA by removing it in certain experiments. Results are found in Table 1 for IP. For all numbers of attacks, ADE-Net outperforms the baseline. The discriminator performs worse when more attacks are added, however, with the addition of vanilla data in our five attack experiments, we can see the discriminator performing better due to the difference between attacked and non-attacked data. The highest Phase II accuracy is achieved in our attack dependent α_k tests, showing that giving more attention to more difficult attacks increases the overall accuracy of ADE-Net. The No CKA ($\lambda_k = 0$) tests perform worse than all the others, showing the positive effect CKA has on ADE-Net.

Conclusion: We analyzed the effect that multiple adversarial attacks have on HSI semantic segmentation. An approach was developed leveraging attack-type detection and robustness in one unified network: ADE-Net. Though a Phase I discriminator attack-type classifier leveraging the similarity measure CKA, and a Phase II attack-expert ensemble network, ADE-Net outperformed the single model baseline for all datasets with all attack combinations.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

José Manuel Amigo. Hyperspectral and multispectral imaging: Setting the scene. In *Data Handling in Science and Technology*, volume 32, pp. 3–16. Elsevier, 2020.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Demetrio Labate, Mohamadkazem Safaripoorfatide, Nikolaos Karantzas, Saurabh Prasad, and Farideh Foroozandeh Shahraki. Structured receptive field networks and applications to hyperspectral image classification. pp. 23, 09 2019. doi: 10.1117/12.2527712.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pp. 6640–6650. PMLR, 2020.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

Sungjune Park, Hong Joo Lee, and Yong Man Ro. Adversarially robust hyperspectral image classification via random spectral sampling and spectral shape encoding. *IEEE Access*, 9:66791–66804, 2021.

Jiahao Qi, Zhiqiang Gong, Xingyue Liu, Kangcheng Bin, Chen Chen, Yongqian Li, Wei Xue, Yu Zhang, and Ping Zhong. Masked spatial-spectral autoencoders are excellent hyperspectral defenders. *arXiv preprint arXiv:2207.07803*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013a.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013b.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013c.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Yonghao Xu, Bo Du, and Liangpei Zhang. Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification. *IEEE Transactions on Image Processing*, 30:8671–8685, 2021a.

Yonghao Xu, Bo Du, and Liangpei Zhang. Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification. *IEEE Transactions on Image Processing*, 30:8671–8685, 2021b.

A Related Works

In Szegedy et al. (2013a) the authors showed that adding a small well-crafted perturbation to a sample can be catastrophic to a network's performance. These adversarially attacked images often look the same to a human observer but make the network vulnerable. The study of adversarial examples and robustness in HSI semantic segmentation is a new entry into the field with Xu et al. (2021a) being the first to introduce the idea of adversarial attack-defense in the HSI domain. They created a self-attention context network (SACNet) to better defend against these attacks. The authors in Xu et al. (2021b) use their Masked Spatial-Spectral Autoencoder (MSSA) which consists of masked sequence attention learning, dynamic graph embedding, and self-supervised reconstruction. Rather than focusing on semantic segmentation via network architecture, some works like Park et al. (2021); Qi et al. (2022) try using rich spectral information to robustify the entire process. In Park et al. (2021), they propose a spectral sampling and shape encoding to increase adversarial robustness as a preprocessing step to traditional per-pixel classification via random sampling. However, all these approaches focus on a single attack at a time and do not explore network robustness in the presence of multiple attacks. Recently, AutoAttack Croce & Hein (2020) extends PGD to create an aggregate attack and achieves lower robust networks. A union of attacks similar to AutoAttack is created in Maini et al. (2020) by using a generalized PGD-based procedure to incorporate multiple perturbation models into a single attack which also leads to drastically worse performance than individual perturbations. Many existing works use these aggregate attacks now to test performance, but still, test one aggregate attack at a time.

B Centered Kernel Alignment (CKA)

In ADE-Net, we use a network-layer similarity measure called Centered Kernel Alignment (CKA) Kornblith et al. (2019) as follows:

$$CKA(O_v, O_k) = \frac{\left\|(O_v)^T O_k\right\|_F^2}{\left\|(O_v)^T O_v\right\|_F^2 \left\|(O_k)^T O_k\right\|_F^2}, \tag{2}$$

For this particular use of CKA, we are using vanilla logits O_v from a discriminator trained offline on attacks to determine the dissimilarity between different attacks. O_k is calculated during training where k is the assigned attack type for each sample during training by the discriminator. CKA was chosen for its intuition over other similarity measures to aid our attack separation. Adversarial attacks have a prevalent effect on a network in deeper layers, therefore, once CKA calculates the relative dissimilarity between the logits of vanilla data and attacked data, we have an intuitive measure to use attack information in the logit space rather than just the feature space.

C ADE-NET EXTENDED FORMALIZATION

Discriminator Phase I: The aim of the discriminator is to add adversarial detection to ADE-Net via attack type classification and CKA to separate mixed attack data. For the discriminator, we use a categorical cross-entropy loss function as $\omega_D^* = \arg\min_{\omega_D} \mathcal{L}_{\omega_D}(F_D(\mathbf{x}))$. To increase attack classification accuracy further, we use the similarity measure CKA from 2 to incorporate more information about attacks into the loss function. This term is the sum of CKAs between logit-layer vanilla data output on a trained discriminator, and the logit-layer output of each attack present during training. The optimal ω_D^* is learned by solving the optimization problem: $\arg\min_{\omega_D} \beta \ \mathcal{L}_{\omega_D}(F_D(\mathbf{x})) + \sum_{k=1}^{K} \lambda_k \ CKA(O_v, O_k)$, where λ_k and β are hyperparameters and $\mathcal{L}_{\omega_D}$ is the cross-entropy function on attack labels while also incorporating the similarity of the attack compared to non-attack data in deeper layers of a trained network.

Ensemble Phase II: The ensemble phase is constituted by a number of networks, equal to n_k, that focus on regular class label classification. Each network is an expert at a particular attack determined by the output label from the discriminator. We use an ensemble categorical cross-entropy loss function $\omega_k^* = \arg\min_{\omega_k} \sum_{k=1}^{K} \alpha_k \mathcal{L}_{\omega_k}(F_k(\mathbf{x}), y)$, where α_k, $k = 1, \ldots, K$ are hyperparameters and allow for individual networks to get more attention than others.

Algorithm 1: ADE-Net model

Input: Data $\{x_i, y_i\}_{i=1}^N$. Set K, E: # of attacks and epochs. Learning rates η_D and η_j, $j = 1, \ldots, K$.
Output: Overall Test Accuracy
Attack data $\{x_i, y_i\}_{i=1}^N$ with k attack types to get $\{x_i^k, y_i^k, c_i = k\}_{i=1}^{n_k}$ and vanilla data $\{x_i^v, y_i^v\}_{i=1}^N$.
Train Discriminator F_D on attacked training data offline and pass vanilla and logits O_v. Shuffle attack data $\{x_i^k, y_i^k, c_i = k\}_{i=1}^{n_k}$ (including vanilla data).
for $e = 1, \ldots, E$ **do**
  for $b = bach_1, \ldots, bach_B$ **do**
    Input data into F_D on the attack type and compute $CKA(O_v, O_k)$. Update F_D's weights as $\omega_D \leftarrow \omega_D - \eta_D \nabla_{\omega_D}(\beta \ \mathcal{L}_{\omega_D}(F_D(\mathbf{x})) + \sum_{k=1}^{K} \lambda_k \ CKA(O_v, O_k))$
    Discriminate attacks based on the ω_D.
    for $j = 1, \ldots, k$ **do**
      Input $\{x_i^j, y_i^j, c_i = j\}_{i=1}^{n_j}$ discriminated by F_D into F_j. Update classifier F_j weights as $\omega_j \leftarrow \omega_j - \eta_j \nabla_{\omega_j} \sum_{k=1}^{K} \alpha_k \mathcal{L}_{\omega_k}(F_k(\mathbf{x}), y)$
    end
  end
end
Report Acc

D EXPERIMENTAL PARAMETERS & EXTENDED EXPERIMENTS

In our main paper, we used the Indian Pines dataset. For extended experiments, we include the Kennedy Space Center (KSC) and Houston datasets. Experimental results for these extra datasets can be found in Table 2 for KSC and Table 3 for Houston. To better visualize the impact of CKA on ADE-Net, Figure 2 is shown to denote how CKA can effectively separate attacks in the logit layer during training.

For all of our experiments, each dataset is reduced to 30 bands using Principal Component Analysis (PCA). We use U-Net Ronneberger et al. (2015) for the discriminator and all ensemble networks. We use Adam optimizer with a learning rate of 0.001 for all networks under three trials. We use a batch size of 256 and train ADE-Net for 100 epochs. For attacks, we use the Fast Gradient Signed Method (FGSM) Goodfellow et al. (2014a), Carlini and Wagner (CW) Carlini & Wagner (2017), Projected Gradient Decent (PGD) Madry et al. (2017), and Iterative Fast Gradient Signed Method

(I-FGSM) Xie et al. (2019). In our experiments each attack is added on at a time starting at, FGSM, then CW, PGD, I-FGSM, and finally non-attacked vanilla data. All attacks were generated using an ϵ = 0.1. In the experiment where $\alpha_k = X_k$, as shown in Tables 1, 2 and 3, values are weighted by the relative difficulty for classification on that attack as: α_{FGSM} = 1.4, α_{CW} = 2.3, α_{PGD} = 1.7, α_{I-FGSM} = 1.3, and for vanilla $\alpha_{Vanilla}$ = 1.0.

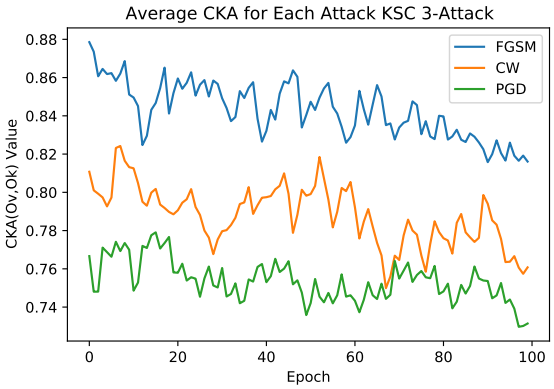

Figure 2: Comparison of attacks' CKA values calculated from the average of each batch for each FGSM, CW, and PGD attack during ADE-Net training for the KSC dataset. Note that for each epoch the three attacks are consistently separated.

# Attack	All = 1.0	λ_k=β=0.1 α_k=10.0	λ_k=β=10.0 α_k=0.1	λ_k=β=1.0 α_k Attack Dependent	λ_k=0 (No CKA) β=α_k=1.0	Baseline
2	I: 100.0 ± 0.0 II 87.45 ± 0.39	I: 100.0 ± 0.0 II: 86.87 ± 0.38	I: 100.0 ± 0.0 II: 86.94 ± 0.29	**I: 99.98 ± 0.01** **II: 87.63 ± 0.24**	I: 99.54 ± 0.02 II: 86.01 ± 0.36	**83.59 ± 0.40**
3	**I: 98.3 ± 0.04** **II: 88.07 ± 0.04**	I: 97.90 ± 0.06 II: 87.39 ± 0.09	I: 97.73 ± 0.03 II: 88.00 ± 0.05	I: 97.70 ± 0.04 II: 87.20 ± 0.08	I: 97.45 ± 0.05 II: 87.11 ± 0.10	**84.69 ± 0.32**
4	I: 73.30 ± 0.10 II 88.18 ± 0.12	**I: 74.04 ± 0.08** **II: 88.65 ± 0.08**	I: 74.21 ± 0.10 II: 88.22 ± 0.09	I: 73.39 ± 0.12 II: 88.50 ± 0.10	I: 73.44 ± 0.10 II: 88.00 ± 0.12	**86.12 ± 0.14**
5	I: 78.57 ± 0.01 II 88.73 ± 0.11	I: 78.66 ± 0.03 II: 88.25 ± 0.12	I: 79.07 ± 0.02 II: 88.66 ± 0.13	**I: 78.68 ± 0.04** **II: 88.75 ± 0.09**	I: 78.01 ± 0.04 II: 87.89 ± 0.13	**86.46 ± 0.17**

Table 2: Experimental results of ADE-Net with Kennedy Space Center dataset. Best performing results for each attack combination are in bold and the worst-performing is in red. The α_k hyperparameter is for the ensemble class cross-entropy, λ_k is for the discriminator CKA and β is for the discriminator attack cross-entropy.

# Attack	All = 1.0	λ_k=β=0.1 α_k=10.0	λ_k=β=10.0 α_k=0.1	λ_k=β=1.0 α_k Attack Dependent	λ_k=0 (No CKA) β=α_k=1.0	Baseline
2	I: 98.61 ± 0.29 II: 95.84 ± 0.59	**I: 98.87 ± 0.22** **II: 96.24 ± 0.54**	I: 98.18 ± 0.19 II: 95.28 ± 0.61	I: 98.16 ± 0.21 II: 95.38 ± 0.49	I: 98.11 ± 0.22 II: 95.22 ± 0.53	**93.94 ± 0.12**
3	**I: 67.05 ± 0.05** **II: 96.32 ± 0.49**	I: 66.81 ± 0.06 II: 95.77 ± 0.35	I: 67.24 ± 0.05 II: 95.70 ± 0.44	I: 67.20 ± 0.05 II: 95.94 ± 0.48	I: 67.03 ± 0.07 II: 95.89 ± 0.42	**95.15 ± 0.05**
4	I: 50.51 ± 0.06 II: 96.02 ± 0.16	I: 50.47 ± 0.06 II: 97.21 ± 0.09	I: 50.40 ± 0.05 II: 96.10 ± 0.11	**I: 50.52 ± 0.04** **II: 97.24 ± 0.08**	I: 50.39 ± 0.06 II: 95.89 ± 0.14	**95.48 ± 0.04**
5	I: 55.66 ± 0.03 II: 96.45 ± 0.05	I: 55.70 ± 0.04 II: 96.57 ± 0.04	I: 55.45 ± 0.03 II 96.35 ± 0.05	I: 55.76 ± 0.03 II: 96.42 ± 0.04	**I: 54.97 ± 0.03** **II: 96.63 ± 0.04**	**96.08 ± 0.16**

Table 3: Experimental results of ADE-Net with the Houston dataset. Best performing results for each attack combination are in bold and the worst-performing is in red. The α_k hyperparameter is for the ensemble class cross-entropy, λ_k is for the discriminator CKA and β is for the discriminator attack cross-entropy.