Supplementary File

Toward Improving Robustness and Accuracy in Unsupervised Domain Adaption

Anonymous submission

Contributions

- We introduce a novel self-training method called Consistent Attention Mapping with Self Pseudo Label Refinement (CAM+SPLR), designed to enhance both the robustness and accuracy of UDA models. This method leverages adversarial target data generated from pseudo labels to strengthen robustness through a self-training paradigm in two step gradient descent process. Simultaneously, it encourages consistency between the attention maps of clean examples and their adversarial counterparts, while progressively refining the pseudo labels.
- We propose the Consistent Attention Mapping (CAM) method to prevent the model from concentrating on less informative regions that may be influenced by adversarial perturbations or noisy pseudo-labels. During training of the TargetNet model, CAM ensures that attention maps remain consistent between clean target data processed by the frozen Anchor model and their adversarial counterparts processed by the TargetNet model. By focusing on semantically relevant key areas, CAM enhances the learning of more discriminative features.
- We further introduce the Self Pseudo Label Refinement (SPLR) method to prevent the model from overfitting caused by inevitable noisy pseudo labels. To achieve this, we progressively refine the pseudo labels during the training of the TargetNet model by incorporating feedback from the labeled source data . This refinement occurs during the second step of the gradient descent process, ensuring that the model remains accurate and resilient as it updates.
- We achieve improvement in robustness and gain in standard accuracy across multiple datasets (OfficeHome (Wang et al. 2021), PACS (Li et al. 2017), VisDA (Peng et al. 2017)) compared to state-of-the-art methods (DART (Wang et al. 2024), SRoUDA (Zhu et al. 2023), and ARTUDA (Yang et al. 2021a)). Specifically, we observed remarkable average robustness gain at $\epsilon = 2/255$ of 5.2%, 4.9%, and 10.2% on the OfficeHome, VisDA, and PACS datasets, respectively. Additionally, average accuracy improved by 0.9%, 8.1%, and 6% over the UDA baseline DANN (Ganin et al. 2016) on the OfficeHome, VisDA, and PACS datasets, respectively.

CAM+SPLR algorithm

The training steps of the proposed method, as defined in Algorithm 1.

Related Work

Unsupervised Domain Adaption

Unsupervised domain adaption is a scenario in which the rich knowledge of the labelled source domain is transferred to the unlabelled target domain to perform different machine learning tasks. There exists various approaches (Ganin et al. 2016; Long et al. 2018, 2017; Saito et al. 2018; Sun and Saenko 2016; Xu et al. 2019; Choi et al. 2019; Yang et al. 2021b), which minimize the distributional differences between the source domain and target domain to perform UDA. DANN (Ganin et al. 2016) is one of the important approaches of UDA that utilize the GAN (Goodfellow et al. 2020) method to learn the domain invariant feature between the source domain and target domain via discriminator and minimize the distributional gaps. CDAN (Long et al. 2018) extends the concept of DANN by adding class conditions to learn more discriminative domain invariant features. Similarly, the authors in (Long et al. 2017), proposed a UDA approach namely, JAN that learns features by aligning the joint distribution of domain-specific layers of source and target domains. Next, the MCD approach (Saito et al. 2018) attempts to align source and target distributions by utilizing task-specific decision boundaries. The authors in (Sun and Saenko 2016) develop a technique, CORAL, to lean nonlinear transformation that aligns correlations of activation layers in deep neural networks. Apart from only minimizing the domain gaps, the prior approaches (Xu et al. 2019; Choi et al. 2019; Yang et al. 2021b) emphasised the use of pseudo labels, generated using the source model, for training the model in the target domain to reduce the domain divergence. These UDA approaches primarily focus on improving the feature alignment from the labelled source domain to the unlabelled target domain to perform specific machine learning tasks; however, they do not consider improving the adversarial robustness of the models.

Adversarial Robustness

Deep neural networks find applications in various domains, including autonomous driving vehicles, recognition sys-

Algorithm 1: Proposed Method Algorithm.

```
Input: Source and target domain datasets: $\mathcal{D}^s = \{x_s^i, y_s^i\}_{i=1}^n$ and $\mathcal{D}^t = \{x_t^j\}_{j=1}^m$, Pre-trained UDA model $F_p$, Anchor
    model $F_a$, TargetNet model $F_t$ , Batch size $B$, Learning rate $lr$, Training epoch $epoch_{max}$, Hyperparameters;
Output: Adversarial trained target model $F_t$;
Pre-training UDA model $F_p$ using: $\min\left(\mathcal{L}_{CE}(F_s(x_s), y_s) + \omega\mathcal{L}_{dd}(x_s, x_t)\right)$;
Initialize $F_t$ and $F_a$ by copying parameters from $F_p$;
Anchor model $F_a$ is frozen in training process;
for $i = 1$ to $epoch_{max}$ do
    Sampling a random mini-batch $B$ from $\mathcal{D}^t$ and $\mathcal{D}^s$;
    Compute the hard pseudo label $y_t$ for unlabeled target data $\mathcal{D}^t$ using $F_t$;
    Generate the target adversarial image $x_t$ using pseudo labels $y_t$;
    Train $F_t$ using Adversarial Target data $\hat{x}_t, y_t$;
    Compute attention map from $F_a$ and $F_t$ using $x_t$ and $\hat{x}_t$ respectively;
    Compute the loss $L_1(\theta_M)$ gradient with the target pseudo labels;
    Update the model $F_t$ by $\theta_M{}' = \theta_M - \eta_1 \cdot \nabla L_1(\theta_M)$;
    Compute the new loss $L_2(\theta_M{}')$ and gradient with the target pseudo labels and labelled source data $(x_s, y_s)$;
    Update the model $F_t$ by $\theta_M{}'' = \theta'_M - \eta_2 \cdot \nabla L_2(\theta'_M)$;
return TargetNet model $F_t$;
```

tems, and security-related applications (malware, intrusion, spam detections *etc.*) Despite their higher applicability in threat detection and classification, deep networks are vulnerable to attacks. It introduces the challenges for training adversarially robust neural networks, enhancing reliability while dealing with maliciously manipulated inputs. Such adversarial attacks can readily perturb the trained weight of the constructed classifier due to the high memorization capability of the deep networks (Arpit et al. 2017). One of the popular attacks, known as Fast Gradient Sign Methods (FGSM) (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014), creates the adversarial example by a one-step gradient ascent across the model loss surface. Similarly, Projected Gradient Descent (PGD) (Madry et al. 2017) is a multi-step or iterative perturbation generation method for adversarial examples, a classical and effective method to generate the perturbations. Other attacks, such as Moment Iterative FGSM (MI-FGSM) (Dong et al. 2018) and Multiplicative adversarial examples(multiadv) (Lo and Patel 2021) are more frequent in neural networks. Similarly, the PGT-AT (Madry et al. 2017) method employs max-min optimization to generate the adversarial examples and train the model with these examples only.

Adversarial Robustness of Unsupervised Domain adaption

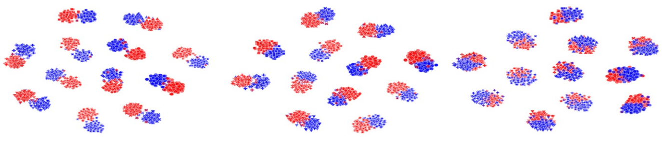
Different methods have been proposed in the existing literature to enhance the robustness of deep learning models. However, only a few methods are proposed to enhance the robustness in unsupervised domain adaption (Awais et al. 2021; Wang et al. 2024; Lo and Patel 2022; Zhu et al. 2023; Yang et al. 2021a). The authors in (Awais et al. 2021) utilize the adversarial pre-trained Imagenet model to improve the robustness of the unsupervised domain adaption. Though the mechanism provides some degree of robustness, but assumes a pre-trained model is somewhere impractical in real scenarios. Similarly, ARTUDA (Lo and Patel 2022) also propose a self-supervised method to achieve the robustness of unsupervised domain adaption in white-box attacks. It uses an additional regularizer and UDA model losses to minimize the distance between target and adversarial logits. Further, ARTUDA utilizes self-supervised signals to generate adversarial examples. SRoUDA (Zhu et al. 2023) introduces a meta-learning-based adversarial training method to improve the robustness. It utilizes a pre-trained UDA model to generate pseudo labels for target domains to generate adversarial images. Afterwards, adversarial training is performed on the target model to enhance robustness by fine-tuning the pseudo-label predictor. Finally, DART (Wang et al. 2024) also used pseudo labels to generate the adversarial example and re-train the UDA model by utilizing the source domain and adversarial target domain by considering the joint loss along with classifier and discriminator loss.

Robustness Evaluation in Feature Space

This section utilizes t-SNE visualization to evaluate the robustness of the proposed method in the feature space of adversarial and clean examples of the target data. In feature space, adding a small perturbation in the clean image results in large changes in the feature space; thus, the model predicts the wrong class corresponding to the perturb images. To study this, we evaluate our method in feature space for the VisDA dataset ($Real \rightarrow Syn$) to determine the robustness of the TargetNet model. We choose features from the last layer of the feature extractor (*i.e.,* ResNet-50) for clean and adversarial examples of the target data. From Figure 1, While UDA+AT reveals a significant distribution gap between clean and adversarial data, and the SRoUDA method reduces this gap to some extent, our proposed method (CAM+SPLR) effectively align the clean and adversarial examples within the target data.

Table 1: An illustration of comparison on PGD 20 attack at $\epsilon = 2/255$ using VisDA dataset.

Source→Target	**Syn→ Re**		**Re→Syn**		**Avg Accuracy**	
Method	Clean	PGD	Clean	PGD	Clean	PGD
DANN(Ganin et al. 2016)	67.5	0.3	78.5	0.5	73.0	0.4
UDA+AT	69.6	58.3	85.7	82.0	77.6	70.1
UDA+Trades(Zhang et al. 2019)	68.1	57.9	85.1	81.5	76.6	69.7
UDA+Mart(Wang et al. 2019)	64.8	58.1	82.1	83.9	73.4	71.0
ARTUDA (Yang et al. 2021a)	45.2	32.5	72.5	62.6	58.8	47.5
SRoUDA(Zhu et al. 2023)	48.2	33.4	81.2	72.9	64.7	53.1
DART(Wang et al. 2024)	69.5	58.0	87.3	85.3	78.4	71.6
Ours	**72.8**	**65.9**	**89.5**	**87.1**	**81.1**	**76.5**



a) – UDA+AT b) – SRoUDA c) – Ours

Figure 1: The t-SNE visualization of extracted features from model trained with UDA+AT, SRoUDA, and Ours) on the **Real→ Synthetic** source target domain, respectively. The blue symbols represent clean target data, while the red symbols denote the adversarial examples of the target data. Our method demonstrates a remarkable overlap between clean and adversarial examples in the feature space, as shown in (c).

Comparison results on PGD20 at $\epsilon = 2/255$

Table 1 displays the results for $\epsilon = 2/255$ using various methods on the VisDA dataset.

Additional Visual results

Figure 2 show the attention maps defense against White-box attacks on the VisDA dataset.

References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A Closer Look at Memorization in Deep Networks. In *International conference on machine learning*, 233–242. PMLR.

Awais, M.; Zhou, F.; Xu, H.; Hong, L.; Luo, P.; Bae, S.-H.; and Li, Z. 2021. Adversarial robustness for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8568–8577.

Choi, J.; Jeong, M.; Kim, T.; and Kim, C. 2019. Pseudo-Labeling Curriculum for Unsupervised Domain Adaptation. arXiv:1908.00262.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59): 1–35.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5542–5550.

Lo, S.-Y.; and Patel, V. 2022. Exploring adversarially robust training for unsupervised domain adaptation. In *Proceedings of the Asian Conference on Computer Vision*, 4093–4109.

Lo, S.-Y.; and Patel, V. M. 2021. Multav: Multiplicative adversarial videos. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6.

Long, M.; CAO, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional Adversarial Domain Adaptation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31, 1–11.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep Transfer Learning with Joint Adaptation Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2208–2217.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

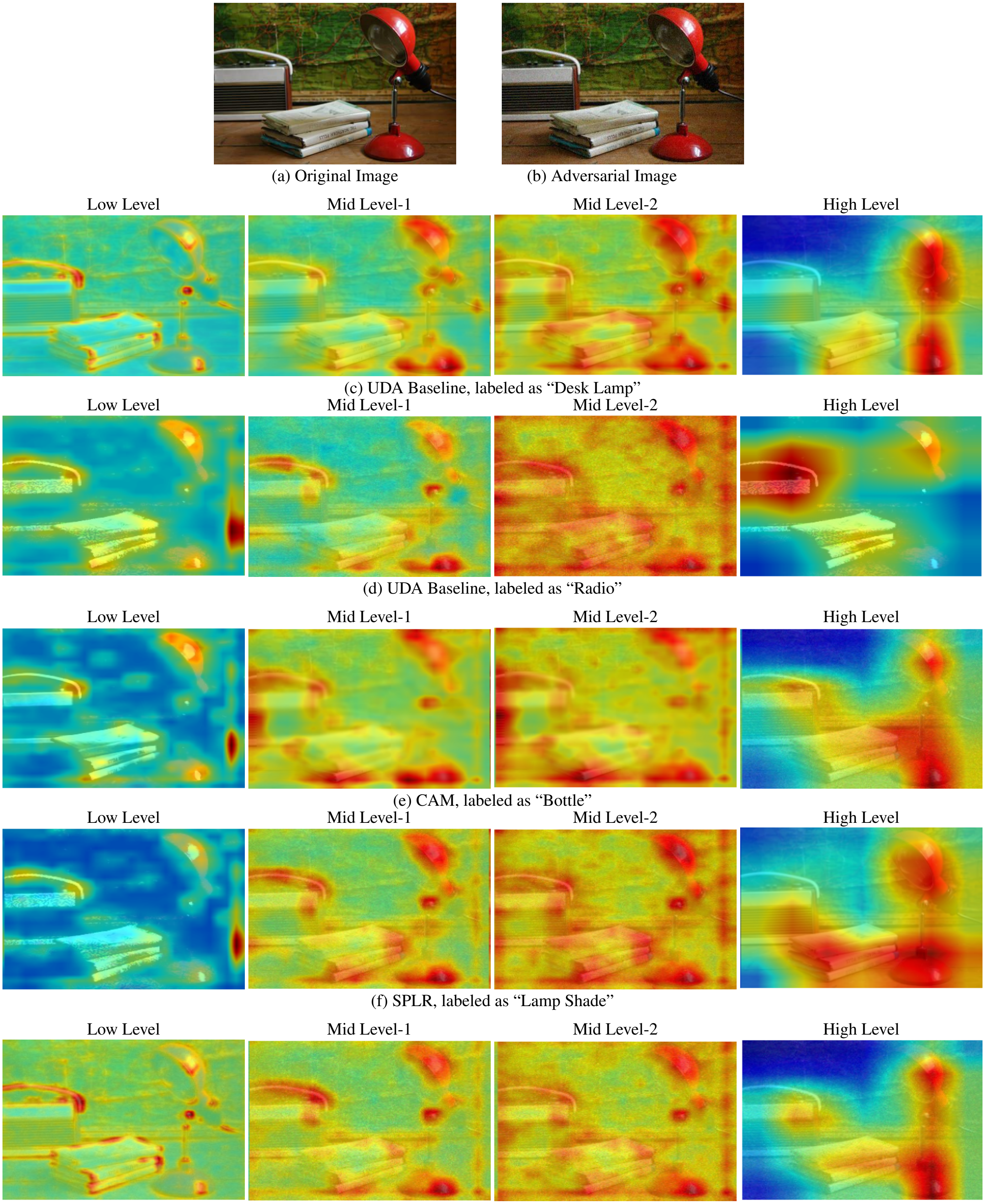
Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, B.; and Saenko, K. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops*, 443–450. Cham: Springer International Publishing.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wang, J.; Tian, K.; Ding, D.; Yang, G.; and Li, X. 2021. Unsupervised Domain Expansion for Visual Categorization. *ACM Transactions on Multimedia Computing Communications and Applications (TOMM)*. In press.



(a) Original Image

(b) Adversarial Image

(c) UDA Baseline, labeled as "Desk Lamp"

(d) UDA Baseline, labeled as "Radio"

(e) CAM, labeled as "Bottle"

(f) SPLR, labeled as "Lamp Shade"

(g) CAM+SPLR, labeled as "Desk Lamp"

Figure 2: Attention maps for defense against White-box PGD20 attacks ($\epsilon = 8/255$) on VisDA Dataset. (a) Clean image, and (b) the corresponding adversarial image. (c) and (d) are attention maps of Clean and Adversarial images, and ((e) and (f) display the attention maps for CAM and SPLR method individually applied to the adversarial images, while (g) attention map for the combined CAM+SPLR method.

Wang, Y.; Hazimeh, H.; Ponomareva, N.; Kurakin, A.; Hammoud, I.; and Arora, R. 2024. DART: A Principled Approach to Adversarially Robust Unsupervised Domain Adaptation. *arXiv preprint arXiv:2402.11120*.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 1–14.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1426–1435.

Yang, J.; Li, C.; An, W.; Ma, H.; Guo, Y.; Rong, Y.; Zhao, P.; and Huang, J. 2021a. Exploring robustness of unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9194–9203.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2021b. ST3D: Self-Training for Unsupervised Domain Adaptation on 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10368–10378.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhu, W.; Yin, J.-L.; Chen, B.-H.; and Liu, X. 2023. SRoUDA: meta self-training for robust unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3852–3860.