SELF-ENSEMBLE VARIANCE REGULARIZATION FOR DOMAIN ADAPTATION

Xinyi Liu[⋆,†] *Tao Dai*[‡] *Shu-Tao Xia*[†,‡] *Yong Jiang*[⋆,†,‡,II]

[⋆]Tsinghua Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China
[†]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[‡] PCL Research Center of Artificial Intelligence, Peng Cheng Laboratory, Shenzhen, China
liuxinyi19@mails.tsinghua.edu.cn, daitao.edu@gmail.com, {xiast,jiangy}@sz.tsinghua.edu.cn

ABSTRACT

Unsupervised domain adaptation (UDA) aims to transfer knowledge from a label-rich source domain to a different yet related fully-unlabeled target domain. Existing approaches utilize self-training scheme to learn discriminative target features and thus enforce class-level distribution alignment implicitly across the source and target domains. However, inherent noise of the pseudo labels due to domain shift could compromise the training process to negatively affect the adapted model performance. In this paper, we propose Self-Ensemble Variance Regularization for Domain Adaptaton (VRDA) method to rectify the learning with pseudo labels. To be specific, we regard the prediction distinction between the student and its self-ensemble teacher model as prediction variance, to regularize target domain prediction bias from pseudo labels. The experimental results reveal that the proposed VRDA achieves the state-of-the-art performance on several standard UDA datasets.

***Index Terms*—** unsupervised domain adaptation, self-ensemble, prediction variance

1. INTRODUCTION

Despite great success of deep learning in computer vision tasks, the success is highly dependent on numerous labels in large datasets and the assumption that training and testing data are drawn from the same distribution. In reality, deep models often suffer performance drops when testing on the new data with different characteristics, due to the presence of *domain shift* [1]. Unsupervised Domain Adaptation (UDA) aims to learn a model from a label-rich source domain that can generalize well to a different yet related unlabeled target domain. In this way, UDA techniques address domain shift problem and thus cause significant interests in both academia and industry.

The intention of mainstream methods in UDA is to learn domain-invariant features between source and target domains. Concretely, most works model the discrepancy between domains to construct domain alignment loss, for joint optimization with source domain task-specific loss. Some works [2, 3] explicitly measure the statistic discrepancy on corresponding activation layers of the source and target network streams, while adversarial-based methods [4] utilize an adversarial objective with respect to a domain discriminator to encourage domain confusion, which can minimize domain discrepency implicitly. However, learning domain-invariant features only considers marginal feature distributions alignment, leading to difficulty in learning discriminative features in target domain. Besides, [5] proves an information theoretic lower bound on the joint error of this type of methods and proposes that better distribution alignment causes increasing error on the target domain when label distributions differ in two domains.

Recently, another line of works based on self-training framework emerges for UDA. Typically, self-training repeats alternative optimization to generate pseudo labels based on model prediction probabilities and train the model with pseudo labels. Compared to domain-invariant features learning methods, self-training based methods can learn discriminative target features via pseudo labeling, yet they can only implicitly align feature discrimination between domains by joint training, without theoretical guarantee. Therefore, reliability of pseudo labels (self-generated supervision) is crucial within self-training methods. Most previous works perform easy-to-hard stragety to select pseudo labels, which means setting thresholds to guarantee the higher signal-to-noise ratio. [6, 7].

Nevertheless, the existence of false-easy (false-pseudo-labeled due to high confidence) samples causes the error accumulation and consequent performance drop in the target domain. To tackle this problem, we get motivation from learning with noisy labels [8] that samples with large prediction distinctions at different phases of model training process are more likely to be noisy ones, and thus we propose a novel method named Self-Ensemble Variance Regularization to rectify the target domain self-training. Concretely, we simulta-

[II] Corresponding author: Yong Jiang

This work is supported in part by the National Natural Science Foundation of China under Grant 62171248, the R&D Program of Shenzhen under Grant JCYJ20180508152204044, and the PCNL KEY project (PCL2021A07).

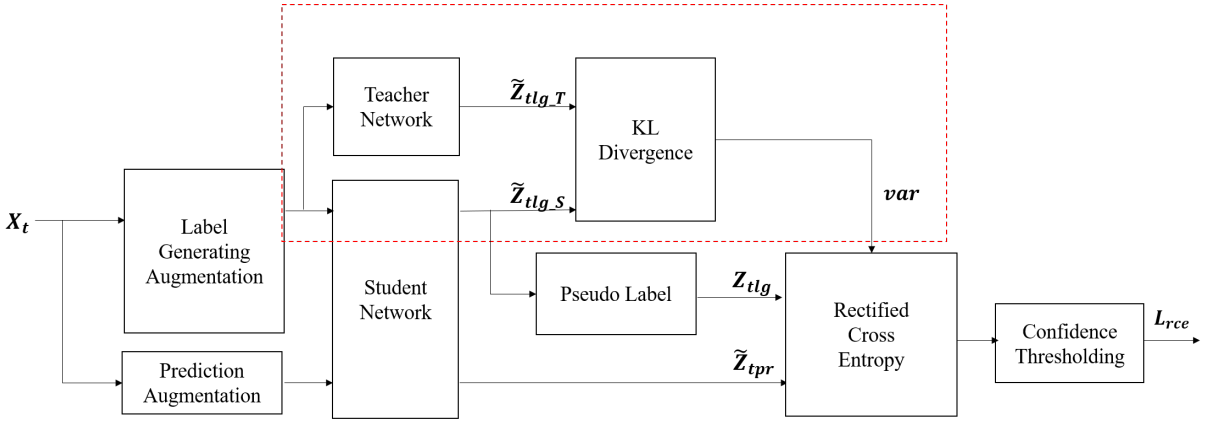

Fig. 1. **The learning procedure of the optimization objective for target domain data.** Within the self-training framework of the student model, the teacher model weights are updated as an exponential moving average of the student weights to construct variance regularization term for rectified cross-entropy loss. And the confidence thresholding based on network prediction confidence acts as a filter to guarantee the higher signal-to-noise ratio.

neously update two networks: a student and a teacher model, where the student is trained using gradient descent and the weights of the teacher are the exponential moving average of those of the student [9]. At each step, we observe the prediction probabilities from the two models of the same input to simulate the predictions of one model at different training stages. Then we compute the KL divergence of their prediction probabilities as approximate variance of pseudo labels, and involve the variance regularization term into the optimization objective of target domain. The main contributions of this paper are the following: (i) We pay special attention to false-easy samples in the target domain during self-trianing procedure. (ii) We propose a novel Self-Ensemble Variance Regularization for Domina Adaptation (VRDA) method, which can recity the learning from noisy pseudo labels based on prediction variance. (iii) VRDA can achieve state-of-the-art performance on three standard UDA datasetss.

2. RELATED WORK

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain. The mainstream methods for UDA intend to directly align marginal feature distributions across domians, named domain-invariant features learning approach. Typically they utilize a distance metric to measure the domain discrepancy, such as Maximum mean discrepancy (MMD) [2], Wasserstein distance [3], etc. Another line of study constructs domain alignment loss with a domian discriminator which encourages domian confusion via an adversarial objective [4, 10].

Self-training based methods for UDA introduce categorical information by assigning pseudo labels and implicitly encourage the class-wise feature alignment by joint training. [6, 11] assign pseudo labels progressively according to an easy-to-hard strategy, while [12] leverage an asymmetric tri-training strategy. [7, 13] adopt pseudo labeling integrated as a compensation for the lack of categorical information. [14, 15] consider the effect of noise in pseudo labels and attempt to rectify them, we follow such motivation and focus on learning dynamics of neural networks [8].

3. METHOD

3.1. Preliminaries

Given the universal UDA setting, we have access to labeled data in source domain $D_s = \left\{x_s^i, y_s^i\right\}_{i=1}^{N_s}$ and an unlabeled target domain dataset $D_t = \left\{x_t^i\right\}_{i=1}^{N_t}$. Our main goal is to learn a model which can minimize the prediction bias of the target domain data. Since ground truth labels y_t are unavailable, pseudo labels are often produced as supervision for target domain data in self-training methods. Therefore, the optimization objective can be formulated as:

$$L_{ce}(X_t) = \text{Bias}(p_t) = \mathbb{E}\left[-\hat{p}_t^j \log F\left(x_t^j|\theta_t\right)\right] \tag{1}$$

where the loss function is the basic cross-entropy loss, $F\left(x_t^j|\theta_t\right)$ is predicted probability distribution of x_t^j (model softmax layer ouput), p_t^j and $\hat{p}_t^j$ denote the one-hot vector of ground truth label y_t^j and pseudo label $\hat{y}_t^j = \text{argmax}\, F\left(x_t^j|\theta_t\right)$ respectively.

For self-training based methods in UDA setting, alternative optimization thus repeats to generate pseudo labels based on model predictions and train the model to infer pseudo labels. Note that such scheme consider model consistency through data perturbation and entropy minimization princi-

ple at the same time, where both of them make sense for unsupervised target samples learning.

3.2. Self-Ensemble Variance Regularization

Self-training methods proves effective for unlabeled data learning under conditional supervision. In semi-supervised scenario, pseudo labels are reliable enough within training iterations to achieve surprising performance, since the predefined confidence threshold filter can guarantee high signal-to-noise ratio [16]. In other words, introduced error $Bias(\hat{p}_t, p_t)$ between pseudo labels and ground truth labels can be ignored based on similar characteristics shared by labeled and unlabeled data.

However, potential noise in pseudo labels can not be neglected due to domain shift in UDA setting. Even with well-defined confidence threshold filter, the existence of false-easy samples (high-confidence false pseudo labels) could negatively influence the training. We get inspiration from an empirical study of learning dynamics of deep networks [8] that samples with noisy labels or atypical class characteristics are forgettable during supervised training process, which refers to samples that have been misclassified after being correctly identified. It seems hard to define forgetting events without given labels in self-training procedure, but we derive the motivation that samples with large prediction distinction at different stages of model training can be regarded as unstable samples, with uncommon categorical features or even potential noisy samples. Thus, we expect to utilize this intuition as variance regularization to rectify the learning from noisy labels.

Typically, different network parameters checkpoints over training phases are directly used for the purpose above. However, we propose model weights self-ensemble as a novel and reasonable approach to utilize differences within time dimension of model training. In self-ensemble framework, the teacher model uses the exponential moving average (EMA) weights of the student model, and the gradients are propagated through only the student model [9]. Since the teacher model can aggregate information over mini-batch training steps, it is more likely to get better intermediate representations and prediction accuracy. In this paper, we thus leverage such framework to estimate the variance, in practice, we utilize the KL divergence of respective predictions from student and teacher models as the variance:

$$\mathrm{Var}(p_t) = D_{kl} = \mathbb{E}\left[F_T\left(x_t^j|\theta_t\right)\log\left(F_T\left(x_t^j|\theta_t\right)/F_S\left(x_t^j|\theta_t\right)\right)\right] \tag{2}$$

where $F_T\left(x_t^j|\theta_t\right)$ and $F_S\left(x_t^j|\theta_t\right)$ denote softmax outputs of the teacher and student model respectively.

Similar to [15], we rectify the loss for target domain data with proposed variance regularization term:

$$L_{rce}(X_t) = \mathbb{E}\left[\frac{1}{\mathrm{Var}\left(p_t\right)}\mathrm{Bias}\left(p_t\right) + \mathrm{Var}\left(p_t\right)\right] \tag{3}$$

Algorithm 1 Training procedure of VRDA

Input: labeled source samples $D_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$, unlabeled target samples $D_t = \{x_t^i\}_{i=1}^{N_t}$, the source model θ_s
Output: adapted model θ_t

```
Initialize θ_tS, θ_tT: θ_tS, θ_tT ← θ_s
for step = 1 to T, do: do
    Derive B_s and B_t sampled from D_s and D_t
    Generate pseudo labels for B_t
        Ŷ_t = {ŷ_t^i = argmax F(α_lg(x_t^i)|θ_tS)}_{i=1}^{B_t}
    Calculate variance regularization term based on eq.(2)
    Update student model θ_tS by back propagating
        Loss = L_ce(B_s) + λ L_rce(B_t)
    Update teacher model θ_tT by EMA weights of θ_tS
    Use confidence thresholding
end for
return θ_t ← θ_tS
```

The objective eq.(3) regularizes prediction bias for target domain from two aspects. Firstly, the prediction bias $\mathrm{Bias}(p_t)$ tends to be not punished when the prediction variance presents large. In addition, the second term $\mathrm{Var}\left(p_t\right)$ acts as a trade-off to prevent large variance all the time. To stablize the training, we replace $1/\mathrm{Var}\left(p_t\right)$ with $\exp(-\mathrm{Var}(p_t))$ following [17] in case $\mathrm{Var}(p_t) = 0$. Therefore, the final objective can be obtained as follow:

$$L_{rce}(X_t) = \mathbb{E}\left[\exp\left\{-D_{kl}\right\}L_{ce}(X_t) + D_{kl}\right] \tag{4}$$

The learning procedure of the loss term for target domain data can be indicated in fig. 1. For each step, the student model produces pseudo labels and the variance regularization vector via label generating augmentation, and then the student model updates its parameters with pseudo labels based on eq.(4) via prediction augmentation. Meanwhile, the teacher model updates as an EMA of the student weights. Combined with source domain loss and other integrated strategies, the whole training procedure of VRDA is shown in Algorithm 1, where each step denotes a mini-batch training iteration, B_s and B_t denote mini-batch training sets sampled from D_s and D_t, θ_{t_S} and θ_{t_T} are the student and teacher model weights respectively.

4. EXPERIMENTS

4.1. Setups

We validate our VRDA method on three popular public benchmarks: *office-31*, *ImageCLEF-DA* and *Visda17*.

office-31 [22] and *ImageCLEF-DA* [23] are two standard benchmark datasets which both contain three domains. For unbiased evaluation, we evaluate on all six transfer tasks between any two domains within each dataset.

Visda17 [24] is a challenging benchmark dataset for UDA with the domain shift from synthetic data real-world images.

Method	*office-31*							*ImageCLEF-DA*						
	A→W	D→W	W→D	A→D	D→A	W→A	Avg	I→P	P→I	I→C	C→I	C→P	P→C	Avg
Source	68.4	96.7	99.3	68.9	62.5	60.7	76.1	74.8	83.9	91.5	78.0	65.5	91.3	80.7
DAN [2]	80.5	97.1	99.6	78.6	63.6	62.8	80.4	74.5	82.2	92.8	86.3	69.2	89.8	82.5
DANN [4]	84.5	96.8	99.4	77.5	66.2	64.8	81.6	75.0	86.0	96.2	87.0	74.3	91.5	85.0
CBST [6]	87.8	98.5	**100**	86.5	71.2	70.9	85.8	-	-	-	-	-	-	-
CDAN [10]	93.1	98.2	**100**	89.8	70.1	68.0	86.6	76.7	90.6	97.0	90.5	74.5	93.5	87.1
CRST [11]	89.4	98.9	**100**	88.7	72.6	70.9	86.8	-	-	-	-	-	-	-
CDAN+E [10]	94.1	98.6	**100**	92.9	71.0	69.3	87.7	77.7	90.7	**97.7**	91.3	74.2	94.3	87.7
TAT [18]	92.5	**99.3**	**100**	93.2	73.1	72.1	88.4	**78.8**	92	97.5	92	**78.2**	94.7	88.9
ALDA [14]	**95.6**	97.7	**100**	**94.0**	72.2	72.5	**88.7**	-	-	-	-	-	-	-
VRDA	90.3	98.9	**100**	92.1	**75.5**	**74.7**	88.6	78.3	**93.8**	96.3	**93.5**	78.0	**96.3**	**89.4**

Table 1. Accuracy (%) on *office-31* and *ImageCLEF-DA* datasets (based on ResNet50)

Method	Backbone	plane	bcycl	bus	car	house	knife	mcycl	person	plant	sktbrd	train	truck	Avg
Source	ResNet101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [4]		81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [19]		87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CBST [6]		87.2	78.8	56.5	55.4	85.1	79.2	83.8	77.7	82.8	88.8	69.0	**72.0**	76.4
ALDA [14]		93.8	74.1	82.4	69.4	90.6	87.2	89.0	67.6	93.4	76.1	87.7	22.2	77.8
CRST [11]		88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	86.6	73.9	68.8	78.1
DADA [20]		92.9	74.2	82.5	65.0	90.9	93.8	87.2	74.2	89.9	71.5	86.5	48.7	79.8
SE (minaug) [21]		92.9	84.9	71.5	41.2	88.8	92.4	67.5	63.5	84.5	71.8	83.2	48.1	74.8
SE [21]		-	-	-	-	-	-	-	-	-	-	-	-	82.8
VRDA		**97.2**	**90.1**	79.4	**96.3**	**97.1**	**98.3**	**93.0**	**81.8**	**97.9**	**95.0**	**90.7**	17.1	**86.2**

Table 2. Accuracy (%) of different unsupervised domain adaptation methods on *Visda17* dataset

It is a large dataset with 280k images from 12 categories in total. Following previous works [11, 14], we evaluate on the adaptation task from training set to validation set.

We use ResNet50 and ResNet101 pretrained on ImageNet as our backbone networks for small and large benchmarks respectively. The original last FC layer is replaced by the random initialied task-specific FC layer. At the first stage, we finetune the network with source domain data to get the source model. Then we perform VRDA method based on algorithm 1. All parameters are updated by stochastic gradient descent (SGD) with momentum of 0.9. The learning rate is set to one tenth of the first stage with the annealing stragegy [4].

4.2. Results and Analysis

The classification results on the *office-31*, *ImageCLEF-DA* and *Visda17* datasets from various state-of-the-art deep adaptation methods are reported in Table 1 and Table 2 respectively. For fair comparison, we directly cite results of other methods from their original papers under optimal parameters settings. The first row denotes performance of source model in the target domain and thus can be regarded as a lower bound without domain adpatation. For small benchmarks, our proposed VRDA can consistently reach or outperform the state-of-the-art performance across 12 adaptation tasks, and consequently achieve SOTA results of both the two popular benchmarks on average. For the large benchmark, our VRDA nearly achieves the best accuracy over all 12 categories and increases the accuracy of many categories up to 90%. Thus, our model (86.2%) outperforms on average substantially.

From Table 1 and Table 2, we derive the observation that self-trianing based UDA methods tend to perform well when sufficient data exist in the target domain since self-training can learn intrinsical properties from massive unlabeled data. That explains the better performance of self-training methods than marginal feature alignment methods on large benchmark datasets like *Visda17*. Meanwhile, the relatively poor performance of VRDA on A→W and D→W can also be explained.

5. CONCLUSION

In this paper, we propose Self-Ensemble Variance Regularization for Domain Adaptation (VRDA) to rectify the learning from noisy pseudo labels within the self-training scheme. The regularization helps the model learn from noisy labels based on learning dynamics of deep networks, without introducing extra parameters. Our VRDA experimentally proves competitive on three standard unsupervised domain adaptation datasets and achieves better feature alignment compared to previous self-training based UDA methods.

6. REFERENCES

[1] Antonio Torralba and Alexei A. Efros, "Unbiased look at dataset bias," in *CVPR*. 2011, pp. 1521–1528, IEEE Computer Society.

[2] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*. 2015, pp. 97–105, JMLR.org.

[3] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *CVPR*. 2019, pp. 10285–10295, Computer Vision Foundation / IEEE.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, pp. 59:1–59:35, 2016.

[5] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon, "On learning invariant representations for domain adaptation," in *ICML*. 2019, pp. 7523–7532, PMLR.

[6] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *ECCV*. 2018, pp. 297–313, Springer.

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang, "Progressive feature alignment for unsupervised domain adaptation," in *CVPR*. 2019, pp. 627–636, Computer Vision Foundation / IEEE.

[8] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, and Yoshua Bengio, "An empirical study of example forgetting during deep neural network learning," in *ICLR*, 2019.

[9] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 2017, pp. 1195–1204.

[10] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan, "Conditional adversarial domain adaptation," in *NeurIPS*, 2018, pp. 1647–1657.

[11] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang, "Confidence regularized self-training," in *ICCV*. 2019, pp. 5981–5990, IEEE.

[12] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *ICML*. 2017, pp. 2988–2997, PMLR.

[13] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *CVPR*. 2019, pp. 4893–4902, Computer Vision Foundation / IEEE.

[14] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai, "Adversarial-learned loss for domain adaptation," in *AAAI*. 2020, pp. 3521–3528, AAAI Press.

[15] Zhedong Zheng and Yi Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *CoRR*, 2020.

[16] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020.

[17] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *NeurIPS*, 2017, pp. 5574–5584.

[18] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *ICML*. 2019, pp. 4013–4022, PMLR.

[19] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *CVPR*. 2018, pp. 3723–3732, IEEE Computer Society.

[20] Hui Tang and Kui Jia, "Discriminative adversarial domain adaptation," in *AAAI*. 2020, pp. 5940–5947, AAAI Press.

[21] Geoffrey French, Michal Mackiewicz, and Mark H. Fisher, "Self-ensembling for visual domain adaptation," in *ICLR*. 2018, OpenReview.net.

[22] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *ECCV*. 2010, pp. 213–226, Springer.

[23] Barbara Caputo, Henning Müller, Jesus Martínez-Gómez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Barzegar Marvasti, Suzan Üsküdarli, Roberto Paredes, Miguel Cazorla, Ismael García-Varea, and Vicente Morell, "Imageclef 2014: Overview and analysis of the results," in *CLEF*, 2014, pp. 192–211.

[24] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko, "Visda: A synthetic-to-real benchmark for visual domain adaptation," in *CVPR Workshops*. 2018, pp. 2021–2026, IEEE Computer Society.