Target Consistency for Domain Adaptation: when Robustness meets Transferability

Anonymous

April 13, 2021

Abstract

Learning *Invariant Representations* has been successfully applied for reconciling a source and a target domain for Unsupervised Domain Adaptation. In this work, we start by investigating the robustness of such methods under the cluster assumption's prism, bringing new empirical evidence that invariance with a low source risk does not guarantee a well-performing target classifier. More precisely, we show that the cluster assumption is violated in the target domain despite being maintained in the source domain, indicating a lack of robustness of the target classifier. To address this problem, we demonstrate the importance of enforcing the cluster assumption in the target domain, named *Target Consistency* (TC), especially when paired with a loss that promotes (class-level) invariance. Our new approach results in a significant improvement in image classification and segmentation benchmarks over state-of-the-art methods based on invariant representations. Importantly, our method is flexible and easy to implement, making it a complementary technique to existing approaches for improving the transferability of representations. **Key-Words**: Domain Adaptation, Invariant Representations, Cluster Assumption.

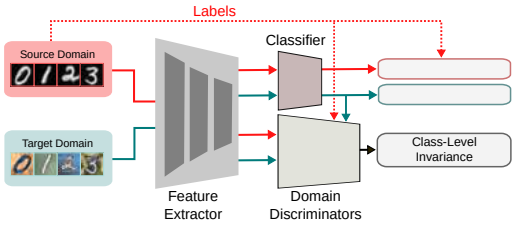


Figure 1: An overview of the proposed framework. In addition to training on the labeled source data, we enforce a Target Consistency (TC), imposing the cluster assumption over target data, and promoting a more robust model on the target domain. To amplify the effect of TC, we perform class-level invariance (CLIV) while enforcing the cluster assumption, where a class specific discriminator is selected using either the source labels or the target predictions for the adversarial loss. Thus promoting positive feedback between decision boundary updates and representation alignment.

1 Introduction

Deep learning (DL) models often show a weak ability to generalize on samples significantly different from those seen during training [BVHP18, ABGLP19, GGB19]. This inability to generalize out of the training distribution presents a significant obstacle to a controlled and safe deployment of DL models in real-world systems [AOS+16, Mar20]. To bridge the distribution gap, Unsupervised Domain Adaptation (UDA) [AOS+16, PY09] leverages labeled samples from a well-known domain, referred to as *source*, to generalize on a *target* domain, where only unlabeled samples are available. If labelling functions are equal across domains, a situation known as the *covariate shift*, then adaptation can be performed by weighting sample contributions in the loss [SKM07, SNK+08]. However, for high dimensional data, such as text or image, it is unlikely that source and target distributions share enough statistical support to compute weights [JSR19]. Learning domain *Invariant Representations i.e.*, representations for which it is impossible to distinguish the domain they were sampled from, can bring together two domains which are different in the input space [GL15, LCWJ15]. This fundamental idea, and the corresponding theoretical target risk [BDBCP07, BDBC+10], has led to a wide variety of methods for adapting deep classifiers to new domains [LZWJ16, LZWJ17, LCWJ18].

Nevertheless, the invariance of representations does not always guarantee a low target risk. For instance, in the case of images, aligning source and target back-

grounds can be detrimental; it may incorrectly align source and target classes if the background is incorrectly correlated with a given class due to some collection bias [BVHP18, ABGLP19], phenomenon known as *negative transfer* [TS10]. Some theoretical works have investigated the question of negative transfer when label shift between source and target domains is observed [ZDCZG19, JSR19], revealing a fundamental trade-off between invariance and ability of predictions to match the true target label distribution [ZDCZG19]. Prior works address this trade-off by relaxing domain invariance with weighted representations [CMLW18, YLC+19, LCWJ18, CZWG20]. However, learning invariant, but transferable representations, remains an open problem. One of the main hurdles is the negative impact invariance has on discriminability, resulting in sub-optimal and sensitive target classification.

In the present work, we aim to provide a new understanding of the transferability of representations through the prism of the cluster assumption, a well-known semi-supervised learning paradigm. The cluster assumption states that if samples are in the same cluster in the input space, they are likely to be of the same class. When enforced on unlabeled samples, the model benefits from a significant gain in generalization [CSZ09, SBL+20, XDH+19] and robustness [CRS+19, HMC+20]. We show that enforcing the cluster assumption in the target domain, named *Target Consistency* (TC), with domain invariant representations goes beyond the role of a regularizer for high capacity features extractor as described in [SBNE18]. Crucially, we reveal that class-level invariance maximizes the gains induced by Target Consistency. By fooling one discriminator per predicted class, we promote positive interaction between TC and Class-Level InVariance (CLIV). Our contributions are:

- We show that domain invariance induces a significant model sensitivity to perturbations in the target domain, indicating that invariance is achieved by disregarding principles of robustness. Such evidence motivates our interest in enforcing the cluster assumption for improving the transferability of domain invariant representations.
- To amplify the effect of TC, we perform class-level invariance (CLIV) while enforcing the cluster assumption, promoting positive feedback between decision boundary updates and representation alignment.
- We show with extensive experiments on both classification and segmentation datasets that we reach state-of-the-art performances for methods based on invariant representations.

2 Related Work

Domain Adaptation. The covariate shift adaptation has been studied by [HGB+07, GSH+09, SNK+08] and label shift with kernel mean matching [ZSMW13, DPS14] and Optimal Transport [RCFT18]. Since Importance Sampling based methods are limited to distributions that share enough statistical support [JSR19, DDF+17], an important line of works focuses on learning domain Invariant Representations (IR) [GL15, LCWJ15] for reconciling two non-overlapping data distributions. IR has led to a furnished literature; *Joint Adaptation Network* which aligns joint distribution of representations across layers [LZWJ16], *Conditional Domain Adaptation Network* which performs the multilinear conditioning between representations and predictions [LCWJ18]. Recently, significant progress has been made towards learning more transferable representations. In the work [LLWJ19], it has been shown that invariance can be achieved by generating consistent intermediate representations, preserving their transferability. [CWLW19] brought to light that invariance often lead to poor discriminability of features, characterized by low rank representations. Therefore, they suggest to penalize the highest singular value of a batch of representations. [WJL+19] revisited the principle of batch normalization by building a transferable layer which aligns naturally mean and variance of representations across domains.

Consistency Regularization. Consistency based semi-supervised methods [LA16, TV17, BCG+19, VLK+19, SBL+20] have enjoyed great success in recent years, closing the gap with their fully supervised counterparts. Such methods are based on a simple concept: the prediction function should produce similar outputs for similar inputs. By enforcing such a constraint, the resulting decision boundary will lie in low density regions echoing a more robust model. Semantically similar inputs can be obtained by a simple Gaussian noise injection [LA16, TV17], data augmentations [BCG+19, SBL+20], or adversarial attacks [MMKI18]. The regularization term added consists of a distance measure (*e.g.*, L^2, KL divergence) between the function's output of a clean and a perturbed input.

3 On the Vulnerability of IR

3.1 Preliminaries

Domain Adaptation (DA) introduces two domains, the *source* and the *target* domains, on the product space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. Those domains are defined by their specific joint distributions of inputs $\mathbf{x} \in \mathcal{X}$ and labels $\mathbf{y} \in \mathcal{Y}$, noted $p(\mathbf{x}^s, \mathbf{y}^s)$ and $q(\mathbf{x}^t, \mathbf{y}^t)$, respectively. We refer to quantities involving the source and the target as s and t, respectively, with exponent notation. Considering a hypothesis class $\mathcal{H}$, subset of functions from $\mathcal{X}$ to $\mathcal{Y}$, DA aims to learn $h \in \mathcal{H}$ which performs well in the target domain *i.e.* has a small target risk $\varepsilon^t(h) := \mathbb{E}_{(\mathbf{x}^t, \mathbf{y}^t) \sim p}[\ell(h(\mathbf{x}^t), \mathbf{y}^t)]$ where ℓ is a given loss. *Unsupervised Domain Adaptation* (UDA) considers the case where labeled samples are available in the source domain while the target domain is only represented with unlabeled samples. Learning *Domain Invariant Representations* (IR) is a key idea for reconciling two non-overlapping data distributions [GL15, LCWJ15]. A mainstream approach consists of learning a representation with a deep feature extractor such that a domain discriminator can not distinguish the target from source samples [GL15]. Provided a representation class Φ, a subset of functions from the input space $\mathcal{X}$ to a representation space $\mathcal{Z}$, a classifier class $\mathcal{G}$, a subset of functions from $\mathcal{Z}$ to $\mathcal{Y}$, and noting $\mathcal{H} := \mathcal{G} \circ \Phi := \{g \circ \varphi; g \in \mathcal{G}, \varphi \in \Phi\}$, representations $\varphi \in \Phi$ are learned by achieving a trade-off between minimizing source classification error and fooling a domain discriminator [GL15], expressed as a function from $\mathcal{Z}$ to $[0, 1]$. The role of representations in UDA has been theoretically investigated by Ben-David *et al.* in [BDBCP07], and extended in [BDBC+10, MMR09, ZDCZG19], through a bound of the target risk:

Theorem 1 (From [BDBCP07] and [BDBC+10])**.** *Given a hypothesis class $\mathcal{H}$ and a hypothesis $h \in \mathcal{H}$:*

$$\varepsilon^t(h) \leq \varepsilon^s(h) + d_{\mathcal{H}\Delta\mathcal{H}} + \lambda_{\mathcal{H}} \tag{1}$$

where $d_{\mathcal{H}\Delta\mathcal{H}} := \sup_{h,h' \in \mathcal{H}} |\varepsilon^s(h, h') - \varepsilon^t(h, h')|$ and $\lambda_{\mathcal{H}} := \inf_{h \in \mathcal{H}} \{\varepsilon^t(h) + \varepsilon^s(h)\}$. In particular, provided a representation φ, and applying the inequality to $\mathcal{G} \circ \varphi := \{g \circ \varphi : g \in \mathcal{G}\}$:

$$\varepsilon^t(g\varphi) \leq \varepsilon^s(g\varphi) + d_{\mathcal{G}\Delta\mathcal{G}}(\varphi) + \lambda_{\mathcal{G}}(\varphi) \tag{2}$$

where $d_{\mathcal{G}\Delta\mathcal{G}}(\varphi) := \sup_{g,g' \in \mathcal{G}} |\varepsilon^s(g\varphi) - \varepsilon^t(g'\varphi)|$ and $\lambda_{\mathcal{G}}(\varphi) := \inf_{g \in \mathcal{G}} \{\varepsilon^s(g\varphi) + \varepsilon^t(g\varphi)\}$.

On the one hand, Eq. (1) shows the role of the hypothesis class capacity for bounding the target risk. The lower the hypothesis class sensitivity to changes in input distribution, the lower $d_{\mathcal{H}\Delta\mathcal{H}}$. On the other hand, Eq. (2) puts emphasis on representations: if source and target representations are aligned *i.e.*, $p(\mathbf{z}^s) \approx q(\mathbf{z}^t)$ for $\mathbf{z} := \varphi(\mathbf{x})$, then $d_{\mathcal{G}\Delta\mathcal{G}}(\varphi)$ remains small.

3.2 Sensitivity in the Target Domain

Prior works [GL15, GUA+16, LCWJ15, LZWJ16, LZWJ17, LCWJ18] have greatly improved capacity to achieve a trade-off between source classification error and domain invariance of representations by minimizing $\varepsilon^s(g\varphi) + d_{\mathcal{G}\Delta\mathcal{G}}(\varphi)$ from Eq. (2). Clearly, maintaining a low $\lambda_{\mathcal{G}}(\varphi)$ while learning domain invariant representations is crucial for a good adaptation. Some works bring theoretical evidence of its difficulty [ZDCZG19, WWKL19, JSR19] while pioneering works dig into that direction [LLWJ19, CWLW19, WJL+19]. This difficulty is referred as *non-conservative* DA in [SBNE18] *i.e.*, when the optimal joint classifier is significantly different from the target optimal classifier:

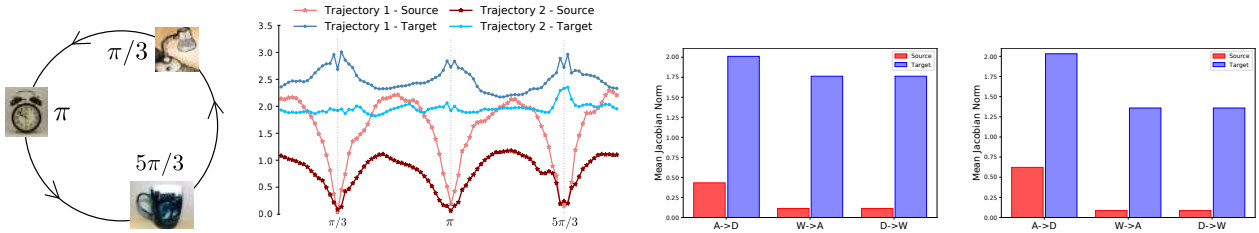
$$\inf_{h \in \mathcal{H}} \varepsilon^t(h) < \varepsilon^t(h^\lambda) \tag{3}$$

where $h^\lambda := \arg\min_{h \in \mathcal{H}} \varepsilon^s(h) + \varepsilon^t(h)$. Similarly, when provided with a representation φ, the optimal joint classifier differs from the target optimal classifier: $\inf_{g \in \mathcal{G}} \varepsilon^t(g\varphi) < \varepsilon^t(g^\lambda\varphi)$ where $g^\lambda := \arg\min_{g \in \mathcal{G}} \{\varepsilon^s(g\varphi) + \varepsilon^t(g\varphi)\}$ (see Appendix for more details).

Importantly, mitigating at train time the risk of non-conservative DA is a difficult problem since target labels are involved in Eq. (3). Therefore, other tools need to be leveraged to detect non-conservative adaptation without the ground truth in the target domain. Following the insight from [SBNE18], we hypothesize that violation of the cluster assumption in the target domain is a strong indicator of a case of non-conservative DA. In such a case, a classifier with different source and target errors should exhibit a substantial sensitivity in the target domain to small input perturbations[1].

Therefore, we analyze the robustness of a model trained to minimize the source risk, through its sensitivity to small perturbations in the input space. We follow [NBA+18] and compute the mean Jacobian norm

[1] The violation of the cluster assumption is characterized by a decision boundary localized in high density regions of the target input space.



(a) Trajectory (b) Regions of Sensitivity (c) Source only (d) DANN

Figure 2: Sensitivity Analysis. (a) An illustration of the circular trajectory passing through three images of different classes. (b) Jacobian norm of source (**D**) and target (**A**) as the input traverses two elliptical trajectories: *Trajectory 1*: different classes. *Trajectory 2*: same classes, for a ResNet-50 trained on source only. (c) and (d) The mean Jacobian norm on target and source domains of a ResNet-50 when trained on source only and with a DANN objective on three Office-31 tasks.

as a proxy of the generalization at the level of individual target samples, and as a measure of the local sensitivity of the model on target examples:

$$\mathbb{E}_{\mathbf{x}^t \sim q} \left[\left\| \mathsf{J}\left(\mathbf{x}^t\right) \right\|_F \right] \tag{4}$$

where $\mathsf{J}_{ij}(\mathbf{x}) = \partial \hat{y}_i / \partial x_j$ is the Jacobian matrix, $\|\mathsf{J}\|_F$ is the Frobenius norm, and $\hat{y}_i$ is the output class probability for class i. For comparison, the source domain's sensitivity can be computed similarly over source instances. By language abuse, we will refer to sensitivity in source and target domains as source and target sensitivity, respectively. The results obtained on 3 transfer tasks from Office-31 (**A** $\rightarrow$ **D**, **W** $\rightarrow$ **A**, **D** $\rightarrow$ **W**) are shown in Fig. 2(c) and Fig. 2(d). As suspected, the target sensitivity is significantly higher compared to the source sensitivity. Importantly, when enforcing invariance of representations with Domain Adversarial Neural Networks (DANN [GL15]), sensitivity in the target domain decreases (for tasks **W** $\rightarrow$ **A** and **D** $\rightarrow$ **W**) while remaining significantly higher than the source sensitivity. This validates our concern on non-conservative domain adaptation: even after features alignment, the resulting classifier still violates the cluster assumption in the target domain. To further investigate the regions of sensitivity, we examine the function's behavior on and off the data manifold as it approaches and moves away from three anchor points. To this end, following [NBA+18], we analyze the behavior of the model near and away from target and source data along two types of trajectories: 1) an ellipse passing through three data points of different classes as illustrated in Fig. 2(a), and 2) an ellipse passing through three data points of the same class. Since linear combinations of images from the same class are likely to look like a realistic image, the second trajectory is expected to traverse overall closer to the data manifold. Fig. 2(b) shows the obtained results. We observe that, according to the Jacobian norm, the model's sensitivity in the vicinity of target data is comparable to its sensitivity off the data manifold. Inversely, the model remains relatively stable in the neighborhood of source data and becomes unstable only away from them, further confirming our hypothesis.

4 Target Consistency

4.1 Consistency Regularization

To promote a more robust model and mitigate target sensitivity, we regularize the model predictions to be invariant to a set of perturbations applied to the target inputs. Concretely, we add to the objective function an additional Target Consistency term:

$$\begin{aligned} \mathcal{L}_{\text{TC}}(\varphi, g) &= \mathcal{L}_{\text{VAT}}(\varphi, g) + \mathcal{L}_{\text{AUG}}(\varphi, g) \\ &= \mathbb{E}_{\mathbf{x}^t \sim p} \left[\max_{\|r\| \leq \epsilon} ||(h(\mathbf{x}^t) - h(\mathbf{x}^t + r)||^2 \right] \\ &+ \mathbb{E}_{\mathbf{x}^t \sim p} \left[||(h(\mathbf{x}^t) - h(\tilde{\mathbf{x}}^t)||^2 \right] \end{aligned} \tag{5}$$

Similar to [SBNE18], the first term incorporates the locally-Lipschitz constraint by applying Virtual Adversarial Training (VAT) [MMKI18] which forces the model to be consistent within the norm-ball neighborhood of each target sample $\mathbf{x}^t$. Additionally, the second term forces the model to embed a target instance $\mathbf{x}^t$ and its augmented version $\tilde{\mathbf{x}}^t$ similarly to push for smooth neural network responses in the vicinity of each target data. With a carefully chosen set of augmentations, such a constraint makes sense since the semantic content of a transformed image is approximately preserved. Note that for more stable training, we follow

Mean Teachers (MT) [TV17] and use of an exponential moving average of the model to compute the target pseudo-labels (*i.e.*, $h(\mathbf{x}^t)$). Overall, $\mathcal{L}_{\text{TC}}$ is in-line with the cluster assumption by promoting consistency to a various set of input perturbations, thus, forcing the decision boundary to not cross high-density regions.

Augmentations. For visual domain adaptation, and based on the recent success of supervised image augmentations [CZM+19, LKK+19, CZSL19] in semi-supervised learning [XDH+19, SBL+20] and robust deep learning [YLS+19, HMC+20], we propose to use a rich set of state-of-the-art data augmentations to inject noise and enforce consistency of predictions on target domain. Specifically, we use augmentations from AutoAugment [CZM+19]. Upon each application, we sample a given operation o from all possible augmentations $\mathcal{O} = \{\text{equalize}, \dots, \text{brightness}\}$. If the operation o is applicable with varying severities, we also uniformly sample the severity, and apply o to obtain the augmented target image $\tilde{\mathbf{x}}^t = o(\mathbf{x}^t)$. However, applying a single operation might be solved easily by a high capacity model by memorizing the specific perturbations. To overcome this, we generate more diverse augmentations by mixing multiple augmented images (see Fig. 1). We start by randomly sampling K operations from $\mathcal{O}$ and K convex coefficients α_i sampled from a Dirichlet distribution: $(\alpha_1, \dots, \alpha_K) \sim \text{Dir}(1, \dots, 1)$. The augmented image $\tilde{\mathbf{x}}^t$ can then be obtained with an element-wise convex combination of the K augmented instances of $\mathbf{x}^t$: $\tilde{\mathbf{x}}^t = \sum_{i=1}^{K} \alpha_i o_i(\mathbf{x}^t)$, impelling the model to be stable, consistent, and insensitive across a more diverse range of inputs [ZSLG16, KKG18, HMC+20].

4.2 Target Consistency with IR

Effects of Target Consistency. Enforcing the target consistency gives us the ability to control the trade-off between a low target sensitivity, *i.e.*, , a low violation of the cluster assumption and a low source risk. As described in [SBNE18], adding $\mathcal{L}_{\text{TC}}$ to the objective function reduces the hypothesis class $\mathcal{H}$ to only include classifiers that are robust on both target and source domains, noted $\mathcal{H}_{\text{TC}}$. Through the lenses of domain adaptation theory (*i.e.*, Eq. (1) from Theorem 1), by constraining the hypothesis space $\mathcal{H}$ to contain stable classifiers across domains, small changes to the hypothesis in the source domain will not induce large changes in the target domain [SBNE18]. Formally, this reduces the domain discrepancy $d_{\mathcal{H}_{\text{TC}}\Delta\mathcal{H}_{\text{TC}}} \leq d_{\mathcal{H}\Delta\mathcal{H}}$ based on the following inclusion $\mathcal{H}_{\text{TC}} \subset \mathcal{H}$.

However, viewing the effect of TC as a constraint on the hypothesis space (Eq. (1) from Theorem 1) does not explain the hidden interactions between TC and invariant representations (Eq. (2) from Theorem 1). To this purpose, we consider a target sample $\mathbf{x}^t$ near the decision boundary which is hard to adapt.Thus, its augmented version, $\tilde{\mathbf{x}}^t$, is likely to have a different predicted class. By enforcing TC, the model embeds $\mathbf{x}^t$ and $\tilde{\mathbf{x}}^t$ similarly to incrementally push the decision boundary far from class boundaries. Such incremental change might result in correcting the predicted class label. However, the underlying representations remain approximately the same, and the discriminator feedback does not reflect this predicted labels change. Now, consider that domain invariance is achieved by leveraging one discriminator per predicted class *i.e.*, class-level invariance. The change of label due to the TC update will result in a switch of the discriminator used, subsequently reflecting the label change in the domain adversarial loss. This interaction between class-level invariance and decision boundary update is the key to the success of TC. Fig. 3 illustrates such an interaction.

Class-level domain discriminator. Similar to [PCLW18] and [CS19] that jointly align the input distributions and output classes for fine-grained alignment. We use CLIV, a well-suited Class-Level InVariance adversarial loss, which leverages one discriminator per predicted class. Let $\mathsf{D} := (D_c)_{1 \leq c \leq C}$, a set of C discriminators *i.e.*, for $\mathbf{z} \in \mathcal{Z}, \mathsf{D}(\mathbf{z}) \in [0,1]^C$, and noting $\cdot$ the scalar product in $\mathbb{R}^C$, CLIV is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{CLIV}}(\varphi) := \inf_{\mathsf{D}} \{ &\mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \log(\mathsf{D}(\mathbf{z}^s))] \\ &+ \mathbb{E}_{(\mathbf{z}^t, \hat{\mathbf{y}}^t) \sim q}[\hat{\mathbf{y}}^t \cdot \log(1 - \mathsf{D}(\mathbf{z}^t))] \}\end{aligned} \tag{6}$$

where given a sample $\mathbf{x}$ with representation $\mathbf{z}$ and output $\hat{\mathbf{y}} := g(\mathbf{z})$, we weight the importance of discriminator D_c in the adversarial loss using the output $\hat{\mathbf{y}}$. It results into a class conditioning of the domain adversarial loss, where the ground-truths are used in the source domain and the predictions in the target domain.

To summarize, our model is trained by minimizing a trade-off between source Cross-Entropy (CE), Class-Level InVariance (CLIV) and Target Consistency (TC); given μ and ν tunable hyper-parameters,

$$\mathcal{L}(g, \varphi) := \mathcal{L}_{\text{CE}}(g, \varphi) + \mu \mathcal{L}_{\text{CLIV}}(\varphi) + \nu \mathcal{L}_{\text{TC}}(g, \varphi) \tag{7}$$

Theoretical analysis. We provide theoretical insights into the interaction between TC and class-level invariance. We consider $\varphi \in \Phi$ and $g \in \mathcal{G}$, which are

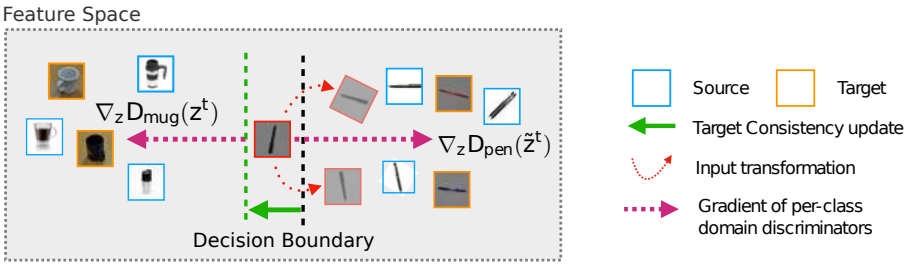


Figure 3: Effect of TC on the learned representations. Mugs and pens from the source (**A**) and target (**D**) domains of Office31 are pictured. The red squared pen, a target sample, is confounded with a mug due to spurious correlations *i.e.*, upward orientation and black color. Input augmentations wipe out spurious correlations induced by the orientation, and the TC pushes the decision boundary to low density regions, correcting the predicted class. Before the TC update, the class-level discriminator encourages the pen to reach the high-density region of the incorrect class, *i.e.*, the mug class. At this time, the class-level discriminator and TC gradients have opposite directions, indicating a negative interaction. The TC update allows the sample to cross the decision boundary. It ultimately changes the class-level discriminator, which now pushes the pen to the correct high-density region corresponding to its true class *i.e.*, the pen class. At this time, the domain adversarial and TC gradients have similar directions, indicating a positive interaction. Crucially, the gradient of a vanilla domain discriminator (*i.e.*, DANN) interacts poorly with the TC update since it does not modify the target representations distribution substantially. *Best viewed in color.*

modified to obtain $\tilde{\varphi}$ and $\tilde{g}$ defined as the closest instances such that $\tilde{g}\tilde{\varphi}$ verifies TC. For instance, they can be obtained by minimizing $\ell_2(\varphi, \tilde{\varphi}) + \ell_2(g, \tilde{g}) + \lambda \cdot \mathcal{L}_{\text{TC}}(\tilde{g}, \tilde{\varphi})$ where ℓ_2 is an L^2 error. When enforcing TC, we expect to decrease the target error *i.e.*, $\varepsilon^t(\tilde{g}\tilde{\varphi}) < \varepsilon^t(g\varphi)$. We assume that it exists $\rho \geq (1 - \varepsilon^t(\tilde{g}\tilde{\varphi})/\varepsilon^t(g\varphi))^{-1}$ for any (g, φ) *i.e.*, the search of consistency always improves the target error. We note $\tilde{\mathbf{y}} := \tilde{g}\tilde{\varphi}(\mathbf{x})$, F a large enough critic function space (See Appendix), we adapt the analysis from [BVC+20]:

$$\varepsilon^t(g\varphi) \leq \rho(\varepsilon^s(g\varphi) + 8 \sup_{\mathsf{f} \in \mathcal{F}} \{\mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \mathsf{f}(\mathbf{z}^s)] - \mathbb{E}_{(\mathbf{z}, \tilde{\mathbf{y}}) \sim q}[\tilde{\mathbf{y}}^t \cdot \mathsf{f}(\mathbf{z}^t)]\} + \inf_{\mathsf{f} \in \mathcal{F}} \varepsilon^t(\mathsf{f}\varphi)) \tag{8}$$

Crucially, by observing that $\sup_{\mathsf{f} \in \mathcal{F}} \{\mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \mathsf{f}(\mathbf{z}^s)] - \mathbb{E}_{(\mathbf{z}, \tilde{\mathbf{y}}) \sim q}[\tilde{\mathbf{y}}^t \cdot \mathsf{f}(\mathbf{z}^t)]\}$ is an *Integral Probability Measure* proxy of $\mathcal{L}_{\text{CLIV}}$, Eq. (8) reveals that class-level domain invariant representations can leverage feedback from an additional regularization, here the Target Consistency, to learn more transferable representations.

5 Experiments

5.1 Datasets

Office-31 [SKFD10] is the standard dataset for visual domain adaptation, containing 4,652 images in 31 categories divided across three domains: Amazon (**A**), Webcam (**W**), and DSLR (**D**). We use all six possible transfer tasks to evaluate our model. **ImageCLEF-DA** * is a dataset with 12 classes and 2,400 images assembled from three public datasets: Caltech-256 (**C**), ImageNet (**I**) and Pascal VOC 2012 (**P**), where each one is considered as separate domain. We evaluate on all possible pairs of the three domains. **Office-Home** [VECP17] is a more difficult dataset compared to Office-31, consisting of 15,00 images across 65 classes in office and home settings. The dataset consists of four widely different domains: Artistic images (**Ar**), Clip Art (**Ca**), Product images (**Pr**), and Real-World images (**Rw**). We conduct experiments on all twelve transfer tasks. **VisDA-2017** [PUK+17] presents a challenging simulation-to-real dataset, with two very distinct domains: **Synthetic**, with renderings of 3D models with different lightning conditions and from many angles; **Real** containing real-world images. We conduct evaluations on the **Synthetic** → **Real** task. For semantic segmentation experiments, we evaluate our method on the challenging **GTA5** → **Cityscapes** VisDA-2017 semantic segmentation task. The synthetic source domain is **GTA5** [RVRK16] dataset with 24,966 labeled images, while the real target domain is **Cityscapes** [COR+16] dataset consisting of 5,000 images. Both datasets are evaluated on the same classes, with the mean Intersection-over-Union (mIoU) metric.

Table 1: Average accuracy (%) of all tasks on image classification benchmarks for UDA. We compare our approach with similar methods based on invariant representations, evaluated using the same protocol. Results are obtained with a ResNet-50 unless specified otherwise. For detailed per task results, see the Appendix.

Method	Office-31	ImageCLEF-DA	Office-Home	VisDA	VisDA (ResNet-101)
ResNet [HZRS16]	76.1	80.7	46.1	45.6	52.4
DANN [GUA+16]	82.2	85.0	57.6	55.0	57.4
CDAN [LCWJ18]	87.7	87.7	65.8	70.0	73.7
TAT [LLWJ19]	88.4	88.9	65.8	71.9	-
BSP [CWLW19]	88.5	-	66.3	-	75.9
TransNorm [WJL+19]	89.3	88.5	67.6	71.4	-
Ours	**89.6**	**89.5**	**69.0**	**77.5**	**79.0**

Table 2: Acc (%) on the 5 hardest Office-Home tasks for TC ablation.

Losses	Avg
$\mathcal{L}_{\text{CLIV}}$	56.7
$+\mathcal{L}_{\text{VAT}}$	57.1
$+\mathcal{L}_{\text{AUG}}$	58.1
$+\mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$	58.6
$+\mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$ w/ MT	**58.9**

Table 3: mIoU on **GTA5 Cityscapes**.

DeepLab v2	
Method	mIoU
Adapt-SegMap [THS+18]	42.4
AdvEnt [VJB+19]	43.8
Ours	**44.9**

5.2 Protocol

We follow the standard protocols for UDA [LZWJ17, LCWJ18, CPK+17]. We train on all labeled source samples and all unlabeled target samples and compare the classification accuracy based on three random experiments for classification and the mIoU based on a single run for segmentation. For classification, we use the same hyperparameters as CDAN [LCWJ18] and adopt ResNet-50 [HZRS16] as a base network pre-trained on ImageNet dataset [DDS+09]. As for CLIV and TC hyperparameters, we use $K = 4$, $\mu = 1$ and $\nu = 10$. We note that the method performs comparatively on a wide range of hyperparameter values making it robust for practical applications. For segmentation, we follow ADVENT [VJB+19] and use the same experimental setup with Deeplab-V2 [CPK+17] as the base semantic segmentation architecture with a ResNet-101 backbone and a DCGAN discriminator [RMC15]. We employ `PyTorch` [PGM+19] and base our code on official implementations of CDAN [LCWJ18] and ADVEN [VJB+19].

Table 4: Avg Acc (%) of the 5 hardest Office-Home tasks for TC coupled with different adversarial losses.

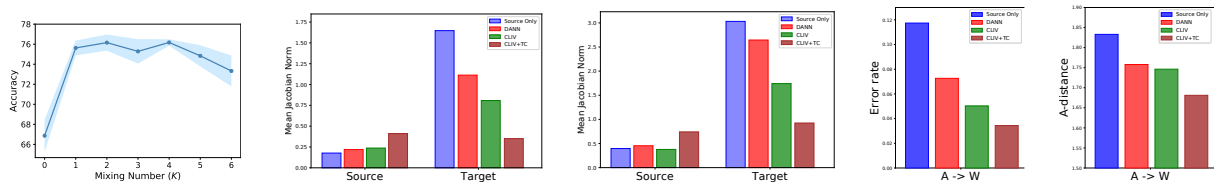
$\mathcal{L}_{\text{adv}} =$	$\mathcal{L}_{\text{DANN}}$	L_{CDAN}	$\mathcal{L}_{\text{CLIV}}$
$\mathcal{L}_{\text{adv}}$	47.6	53.4	56.7
$+\mathcal{L}_{\text{VAT}}$	48.0	55.1	57.1
$+\mathcal{L}_{\text{AUG}}$	51.3	55.7	58.1
$+\mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$	51.4	56.9	58.6
$+\mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$ w/ MT	51.0	56.0	**58.9**

5.3 Results

For clarity and compactness, the average accuracy results of all tasks on all standard classification benchmarks for UDA are reported in Table 1. The proposed method outperforms previous adversarial methods on all datasets. The gains are substantial when the source and target domain are more dissimilar, as in VisDA dataset. We conjuncture that this is a result of a large number of target instances available, enabling us to extract a significant amount of training signal with TC objective term to enforce the cluster assumption. Additionally, the method performs well with many categories, as it is the case for Office-Home dataset. Such gain is a result of the class-level invariance, which is empowered as the number of classes grows. We observe overall smaller improvements on Office-31 due to its limited size, and ImageCLEF-DA since the three domains are visually more similar. We further demonstrate the generality of the proposed method by conducting additional experiments on GTA5 Cityscapes task for semantic segmentation (Table 3), and observe a gain of 2.5 points over the baseline Adapt-SegMap [THS+18], confirming the flexibility of TC and its applicability across DA tasks.

5.4 Ablations

To examine the effect of each component of our proposed method, we conduct several ablations on the 5 hardest tasks on Office-Home, with and without the TC term, and with different variations of the TC loss.



(a) Varying K (b) Sensitivity: **A** $\rightarrow$ **W** (c) Sensitivity: **A** $\rightarrow$ **D** (d) λ (e) A-distance

Figure 4: Analyses. (a) Accuracy on VisDA-2017 with different number of mixed augmentations K. (b) and (c) The effect of TC on the target and source sensitivity for two Office-31 tasks (**A** **W** and **A** **D**). (d) The error λ of the ideal joint hypothesis h^{λ}. (e) A measure of domain discrepancy d_A.

The results are reported in Table 2. We observe that adding a consistency term, either VAT or AUG, results in a higher accuracy across tasks, with better results when smoothing in the vicinity of each target data point within the data manifold with AUG, instead of the adversarial direction using VAT. Their combination, with Mean Teacher (MT), results in an overall more performing model. We also conduct an ablation study on the effect of varying the mixing number K to produce more diverse target images. Fig. 4(a) shows the results. Overall, we observe a slight improvement and more stable results when K is increased, but over a certain threshold, the degree of noise becomes significant, heavily modifying the semantic content of the inputs and hurting the model's performance. Most importantly, to show the importance of coupling TC with CLIV, we pair TC with DANN and CDAN losses. The obtained results in Table 4 show lower average accuracy and minimal gains when enforcing the cluster assumption in conjunction with such adversarial losses, confirming the importance of imposing class-level invariance when applying TC.

5.5 Analyses

Sensitivity Analysis. To investigate the impact of TC on the model sensitivity, we compare the mean Jacobian norm of models trained with various objectives (Figs. 4(b) and 4(c)). TC coupled with CLIV, greatly improves the model's robustness on target, with a small increase in the source sensitivity.

Ideal Joint Hypothesis and Distributions Discrepancy. We evaluate the performances of the ideal joint hypothesis, which can be found by training an MLP classifier on top of a frozen features extractor on source and target data with labels. Fig. 4(d) provides empirical evidence that TC produces a better joint hypothesis h^{λ}, thus more transferable representations. Additionally, as a proxy measure of domain discrepancy [BDBC+10], we compute the A-distance, defined as $d_A = 2(1 - 2\varepsilon)$, with ε as the error rate of a domain classifier trained to discriminate source and target domains. Fig. 4(e) shows that TC decreases d_A, implying a better invariance.

Qualitative Analysis. As shown in Fig. 5, the method produces locally consistent and globally coherent predictions for semantic segmentation.

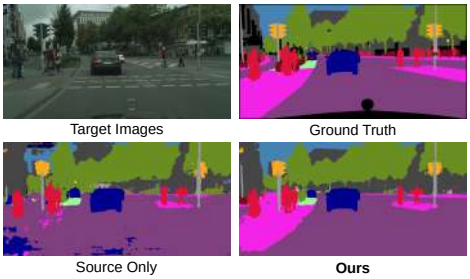


Figure 5: Qualitative Results on GTA5 Cityscapes.

6 Conclusion and Future Work

In this work, we presented a new approach to address the lack of robustness of domain adversarial learning by promoting consistent predictions, named *Target Consistency* (TC), to a set of various input perturbations in the target domain. Crucially, even if our approach is derived from well-known strategies, *i.e.*, class-level invariance and target consistency, we are the first to bring attention to their strong interaction, resulting in a significant improvement of the transferability of representations. Through extensive experiments, we show that our approach outperforms other methods based on invariant representations, validating our analysis. Finally, TC has the advantage of being orthogonal to recent works [LLWJ19, CWLW19] for improving the transferability of invariant representations, thus, combining them is an interesting research direction.

References

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[BCG+19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[BDBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[BDBCP07] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[BVC+20] Victor Bouvier, Philippe Very, Clément Chastagnol, Myriam Tami, and Céline Hudelot. Robust domain adaptation: Representations, weights and inductive bias. *arXiv preprint arXiv:2006.13629*, 2020.

[BVHP18] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[CMLW18] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[CPK+17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[CRS+19] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[CS19] Safa Cicek and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1416–1425, 2019.

[CSZ09] O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[CWLW19] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019.

[CZM+19] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.

[CZSL19] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019.

[CZWG20] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

[DDF+17] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

[DDS+09] 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*, pages 248–255. Ieee, 2009.

[DPS14] Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

[GGB19] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*, 2019.

[GL15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[GSH+09] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[GUA+16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[HGB+07] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[HMC+20] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[HZRS16] 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*, pages 770–778, 2016.

[JSR19] Fredrik Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019.

[KKG18] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018.

[LA16] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[LCWJ15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 97–105. JMLR. org, 2015.

[LCWJ18] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[LKK+19] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems*, pages 6662–6672, 2019.

[LLWJ19] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019.

[LZWJ16] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[LZWJ17] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017.

[Mar20] Gary Marcus. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[MMKI18] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[MMR09] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

[NBA+18] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[PCLW18] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[PUK+17] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[RCFT18] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. *arXiv preprint arXiv:1803.04899*, 2018.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[RVRK16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[SBL+20] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[SBNE18] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach

to unsupervised domain adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[SKFD10] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[SKM07] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[SNK+08] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

[THS+18] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[TS10] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[TV17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[VECP17] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[VJB+19] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[VLK+19] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*, 2019.

[WJL+19] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 1951–1961, 2019.

[WWKL19] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pages 6872–6881, 2019.

[XDH+19] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019.

[YLC+19] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019.

[YLS+19] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pages 13255–13265, 2019.

[ZDCZG19] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On

learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

[ZSLG16] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

[ZSMW13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

Target Consistency for Domain Adaptation: when Robustness meets Transferability

In this document, we provide supplementary materials for our work: *Target Consistency for Domain Adaptation: when Robustness meets Transferability*. We present detailed per task results for all the datasets, additional experimental details and some qualitative results. We also investigate the robustness of the proposed method through the lenses of Fourier analysis. Finally, we provide more details about the theoretical statements of the paper.

Detailed Results

Table 1: Accuracy (%) on Office-31 for unsupervised domain adaptation (ResNet-50)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 (He et al. 2016)	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DAN (Long et al. 2015)	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
RTN (Long et al. 2016)	84.5±0.2	96.8±0.1	99.4±0.1	77.5±0.3	66.2±0.2	64.8±0.3	81.6
DANN (Ganin et al. 2016)	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
ADDA (Tzeng et al. 2017)	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
JAN (Long et al. 2017)	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
GTA (Sankaranarayanan et al. 2018)	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	72.8±0.3	71.4±0.4	86.5
CDAN (Long et al. 2018)	94.1±0.1	98.6±0.1	**100.0±.0**	92.9±0.2	71.0±0.3	69.3±0.3	87.7
TAT (Liu et al. 2019)	92.5±0.3	99.3±0.1	**100.0±.0**	93.2±0.2	73.1±0.3	72.1±0.3	88.4
BSP (Chen et al. 2019)	93.3±0.2	98.2±0.2	**100.0±.0**	93.0±0.2	73.6±0.3	72.6±0.3	88.5
TransNorm (Wang et al. 2019)	95.7±0.5	98.7±0.3	**100.0±.0**	94.0±0.2	73.4±0.4	74.2±0.3	89.3
Ours	94.8±0.8	**99.1**±0.2	**100.0±.0**	93.6±0.9	**76.8**±1.3	73.4±0.7	**89.6**

Table 2: Accuracy (%) on ImageCLEF-DA for unsupervised domain adaptation (ResNet-50)

Method	I → P	P → I	I → C	C → I	C → P	P → C	Avg
ResNet-50 (He et al. 2016)	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN (Long et al. 2015)	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN (Ganin et al. 2016)	75.0±0.6	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
JAN (Long et al. 2017)	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
CDAN (Long et al. 2018)	77.7±0.3	90.7±0.2	97.7±0.3	91.3±0.3	74.2±0.2	94.3±0.3	87.7
TransNorm (Wang et al. 2019)	78.3±0.3	90.8±0.2	96.7±0.4	92.3±0.2	78.0±0.1	94.8±0.3	88.5
TAT (Liu et al. 2019)	78.8±0.2	92.0±0.2	97.5±0.3	92.0±0.3	78.2±0.4	94.7±0.4	88.9
Ours	**79.5**±0.4	**92.7**±0.3	**97.6**±0.2	**93.2**±0.4	**78.6**±0.2	**95.5**±0.4	**89.5**

Table 3: Accuracy (%) on Office-Home for unsupervised domain adaptation (ResNet-50)

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 (He et al. 2016)	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN (Long et al. 2015)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin et al. 2016)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN (Long et al. 2017)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN (Long et al. 2018)	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
TAT (Liu et al. 2019)	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
BSP (Chen et al. 2019)	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
TransNorm (Wang et al. 2019)	50.2	71.4	77.4	59.3	**72.7**	**73.1**	61.0	53.1	79.5	71.9	59.0	82.9	67.6
Ours	**53.1**	**73.0**	**77.0**	**62.6**	72.4	**73.1**	**63.8**	**54.4**	**79.8**	**74.6**	**60.4**	**83.3**	**69.0**

Table 4: Accuracy (%) on VisDA-2017

ResNet-50		ResNet-101	
Method	Synthetic → Real	Method	Synthetic → Real
JAN (Long et al. 2017)	61.6	ResNet-101 (He et al. 2016)	52.4
GTA (Sankaranarayanan et al. 2018)	69.5	DANN (Ganin et al. 2016)	57.4
CDAN (Long et al. 2018)	70.0	CDAN (Long et al. 2018)	73.7
TAT (Liu et al. 2019)	71.9	BSP (Chen et al. 2019)	75.9
Ours	**77.5±0.7**	**Ours**	**79.0±0.1**

Table 5: Accuracy (%) on the 5 hardest Office-Home task for Target Consistency ablation (ResNet-50)

	Ar→Cl	Cl→Ar	Pr→Ar	Pr→Cl	Rw→Cl	Avg
$\mathcal{L}_{\text{DANN}}$	45.2±0.7	48.8±0.5	46.8±0.2	43.5±0.3	53.6±0.3	47.6
$\mathcal{L}_{\text{DANN}} + \mathcal{L}_{\text{VAT}}$	44.3±0.2	50.3±1.8	48.5±1.1	43.6±0.6	53.5±0.2	48.0
$\mathcal{L}_{\text{DANN}} + \mathcal{L}_{\text{AUG}}$	46.2±0.4	55.3±0.5	53.2±1.4	46.0±0.4	55.6±0.5	51.3
$\mathcal{L}_{\text{DANN}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$	46.3±0.6	53.5±1.0	54.7±0.7	46.2±0.7	56.3±0.9	51.4
$\mathcal{L}_{\text{DANN}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$ /w MT	46.6±0.3	53.3±0.7	52.8±0.3	46.9±0.8	55.6±0.5	51.0
$\mathcal{L}_{\text{CDAN}}$	50.3±0.1	54.6±0.7	55.8±0.6	49.3±0.2	56.9±0.1	53.4
$\mathcal{L}_{\text{CDAN}} + \mathcal{L}_{\text{VAT}}$	50.1±0.5	58.5±0.6	59.1±0.6	49.8±0.2	57.9±0.1	55.1
$\mathcal{L}_{\text{CDAN}} + \mathcal{L}_{\text{AUG}}$	51.0±0.2	57.3±0.5	61.0±0.7	50.8±0.2	58.4±0.5	55.7
$\mathcal{L}_{\text{CDAN}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$	51.5±0.2	60.9±0.3	61.4±0.9	51.7±0.2	59.1±0.5	56.9
$\mathcal{L}_{\text{CDAN}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$ /w MT	51.3±0.9	59.0±0.4	60.0±0.5	51.8±0.2	57.9±0.3	56.0
$\mathcal{L}_{\text{CLIV}}$	52.6±0.8	60.1±0.3	60.6±0.9	52.1±0.7	58.3±0.4	56.7
$\mathcal{L}_{\text{CLIV}} + \mathcal{L}_{\text{VAT}}$	52.4±0.6	60.1±0.5	61.2±0.9	53.1±0.2	58.9±0.8	57.1
$\mathcal{L}_{\text{CLIV}} + \mathcal{L}_{\text{AUG}}$	53.1±0.5	62.3±0.6	62.6±0.8	53.1±1.0	59.5±0.3	58.1
$\mathcal{L}_{\text{CLIV}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$	53.0±0.1	**62.8±0.7**	62.8±0.2	53.8±0.8	**60.8±0.8**	58.6
$\mathcal{L}_{\text{CLIV}} + \mathcal{L}_{\text{VAT}} + \mathcal{L}_{\text{AUG}}$ /w MT	**53.1±1.5**	62.6±0.1	**63.8±0.7**	**54.4±0.6**	60.4±0.6	**58.9**

Table 6: mIoU on **GTA5** → **Cityscapes**. *AdvEnt+MinEnt** is an ensemble of two models.

Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
ResNet-101 (He et al. 2016)	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Adapt-SegMap (Tsai et al. 2018)	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
AdvEnt (Vu et al. 2019)	89.9	36.5	81.6	29.2	25.2	28.5	32.3	22.4	83.9	34.0	77.1	57.4	27.9	83.7	29.4	39.1	1.5	28.4	23.3	43.8
Ours	**91.0**	**41.9**	**81.6**	**30.1**	22.6	26.0	28.8	13.6	82.6	**37.2**	**81.9**	56.1	**29.3**	**84.8**	**34.1**	**48.8**	0.0	26.8	**35.7**	44.9
AdvEnt+MinEnt* (Vu et al. 2019)	89.4	33.1	81.0	26.6	**26.8**	**27.2**	**33.5**	**24.7**	**83.9**	**36.7**	78.8	**58.7**	**30.5**	**84.8**	**38.5**	**44.5**	**1.7**	**31.6**	32.4	**45.5**

Experimental Details

Augmentations

For the set of possible augmentations $\mathcal{O}$, we follow AutoAugment (Cubuk et al. 2019) and use the augmentations shown in Fig. 1. We note that when mixing augmentations (*i.e.*, $K > 1$), we also add the possibility of composing augmentations, *e.g.*, for $K = 3$, we give the possibility of sampling a pair of augmentations, so that a given of the operation o_i might be composed of two operations $o_i = o_{i1} \circ o_{i2}$.

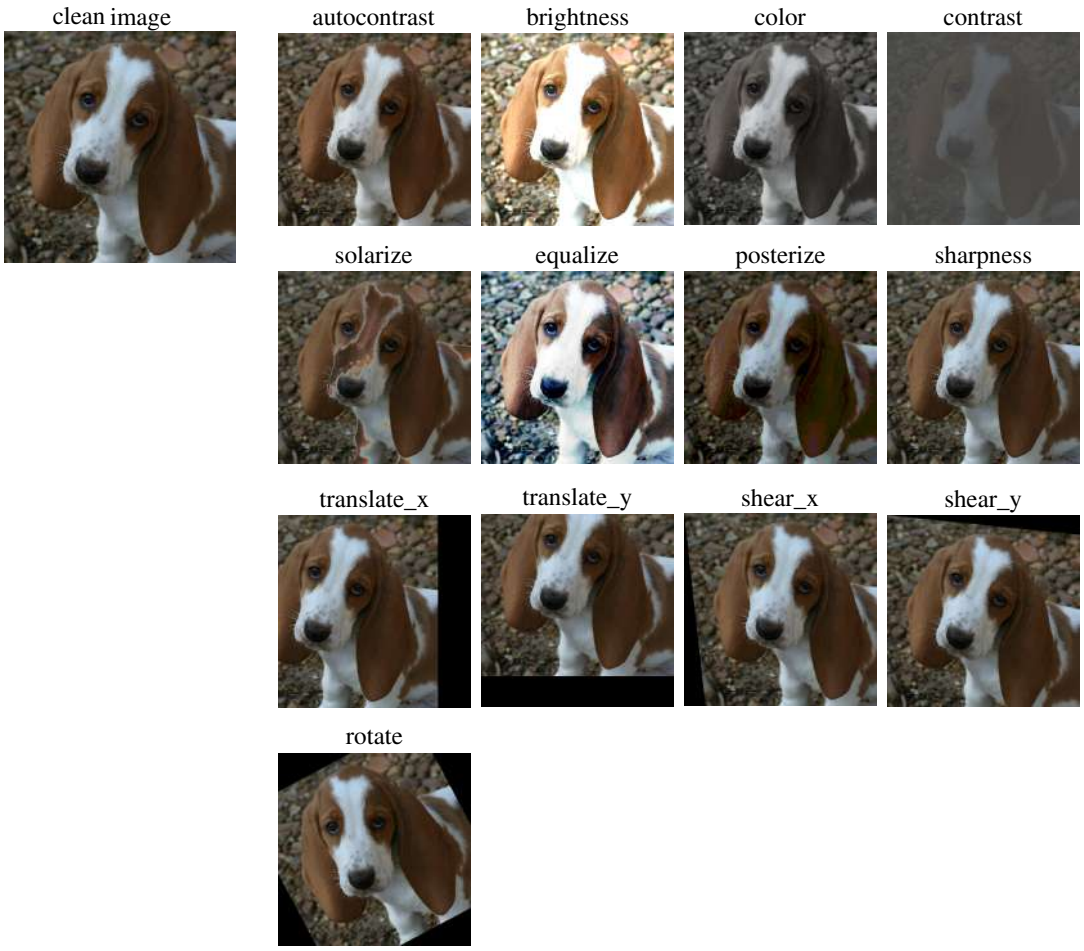


Figure 1: An example of the possible augmentations to be applied on a given input image.

For semantic segmentation, we limit $\mathcal{O}$ to only contain photometric augmentations to avoid changing input coordinate-space, *i.e.*, we remove `translate_x`, `translate_y`, `rotate`, `shear_x` and `shear_y` from $\mathcal{O}$. However, it is possible to maintain the geometric transformations and use a bilinear resampler to bring back the outputs of the augmented image into the coordinate-space of the clean image.

Mean Teacher

The objective of the consistency loss in Paper Eq 5. is to incrementally push the decision boundary to low-density regions on the target domain. However, using the current model h as both a teacher, generating pseudo-labels for the target examples, and as a student, producing the current predictions over perturbed inputs, might result in an unstable training, where a small optimization step can result in a significantly different classifier, hurting the target generalization performance. To solve this, we follow Mean Teachers (MT) (Tarvainen and Valpola 2017), and use an Exponential Moving Average (EMA) of the student model h weights as a teacher h', where the weights θ'_t of the teacher model at a training step t are defined as the EMA of successive student's weights θ:

$$\theta'_t = \alpha\theta'_{t-1} + (1 - \beta)\theta_t \tag{1}$$

where β is a momentum term that controls how far we reache into training history. The teacher model can be used to generate the pseudo-labels $h'(\mathbf{x}^t)$ for a more stable optimization procedure.

Target Consistency for Semantic Segmentation

To demonstrate the generality of target consistency, we propose to adapt it for segmentation tasks. Given the dense nature of semantic segmentation, where we predict class assignment at each spatial location, we remove the local consistency constraint $\mathcal{L}_{\text{VAT}}$, since even small perturbations at the pixel level might significantly change the local appearance, making the task of predicting consistent labels impractical. Additionally, we constrain the target consistency to be only photometric augmentations to conserve the input coordinate-space. We follow (Tsai et al. 2018) and adopt adversarial learning in the output space rather than representation space, taking advantage of the structured outputs in semantic segmentation that contain spatial similarities between the source and target domains, the adversarial network is applied at a multi-level to perform output space adaptation at different feature levels effectively. We refer the reader to Section 4 of (Tsai et al. 2018) for more details on multi-level output space-based adaptation.

Implementation

For the implementation, we use `PyTorch` (Paszke et al. 2019) deep learning framework and base our implementation on the official implementations of CDAN (Long et al. 2018)[1] and ADVEN (Vu et al. 2019)[2]. All experiments are done on a single NVIDIA V100 GPU with 32GB memory. In terms of the hyperparameters, for classification, we adopt mini-batch SGD with a momentum of 0.9 and the learning rate annealing strategy (Ganin et al. 2016) with an initial learning rate of 10^{-2}. As for segmentation, the model is trained using mini-batch SGD and a learning rate 2.5×10^{-4}, momentum 0.9 and weight decay 10^{-4}, and Adam optimizer (Kingma and Ba 2014) for the discriminator with learning rate 10^{-4}, both with a polynomial learning rate scheduler (Chen et al. 2017).

[1] https://github.com/thuml/CDAN

[2] https://github.com/valeoai/ADVENT

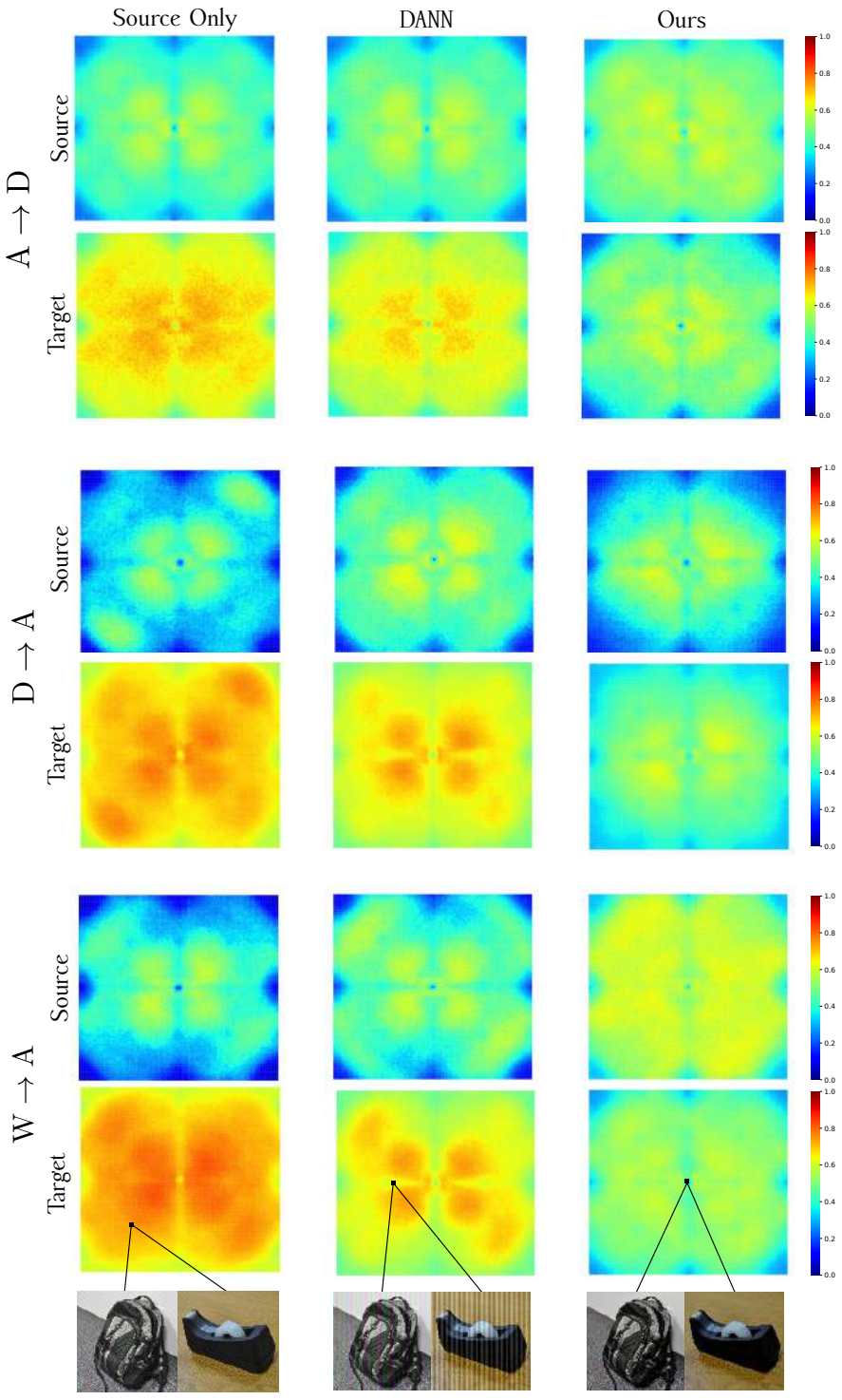


Figure 2: Fourier Analysis of Model Robustness on Source and Target. An illustration of the Fourier sensitivity heatmaps on the source and target domains for a ResNet-50 trained with different objectives. Each pixel of the heatmap is the error of the model when all of its inputs are perturbed with a single Fourier basis vector.

To further examine the lack of target robustness in DA, we investigate a common hypothesis in robust deep learning (Hendrycks et al. 2020), where the lack of robustness is attributed to spurious high-frequency correlations that exist in the source data,

that are not transferable to target data. To his end, we follow (Yin et al. 2019), and measure the model error after injecting an additive noise at different frequencies. Concretely, we resize all of the data to 96×96 images, we then add, at each iteration, 96×96 Fourier basis vector corresponding to an additive noise at a given frequency, and record the model error over either source or target data when such basis vector is added to each image individually (see Section 2 of (Yin et al. 2019) for more details). Fig. 2 shows the Fourier sensitivity heatmaps on source and target, for a ResNet-50 trained with different objectives. Each pixel of 96×96 heatmaps shows the error of the model when the inputs are perturbed by a single Fourier basis vector, in which the error corresponding to low-frequency noise is shown in the center, and high frequencies are away from the center. We observe that the model is highly robust on source across frequencies and the different objectives, but becomes quite sensitive to high-frequency perturbations on target when trained on source only or with a DANN objective. However, such sensitivity is reduced when enforcing the cluster assumption on the target domain, indicating a possible suppression of the spurious high-frequency correlations found in the source domain.

Qualitative Results

Qualitative Results. Fig. 3 illustrate qualitative results for smantic segmentation. Additionally, we visualize the feature representations of **D** → **A** task of Office-31 with t-SNE (Maaten and Hinton 2008) in Fig. 4. We observe that our method produces a well aligned source and target features. This shows the benefits of coupling consistency regularization with class level discrimination.

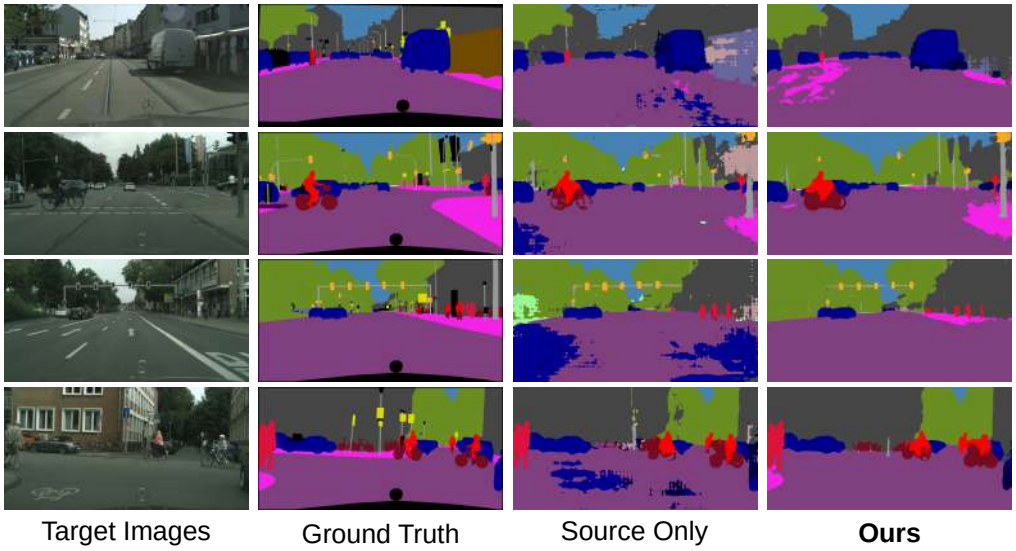


Figure 3: Qualitative Results on GTA5 → Cityscapes.

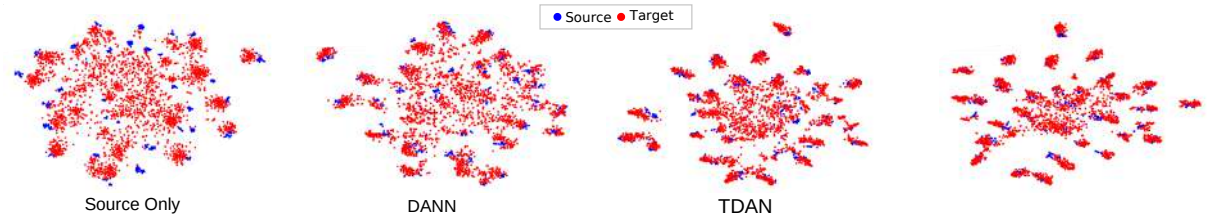


Figure 4: T-SNE of the adapted features of, *left*: ResNet-50, *center*: DANN, *right*: Ours, trained on D → A task of Office-31. Blue: Source D; Red: Target A.

Toy Dataset. To show the effect of TC on the decision boundary, we conduct a toy experiment on the rotating two moons dataset, where the target samples are obtained by rotating the source points by $45°$, comparing the learned decision boundary when we train on source only, with a DANN objective, and when using TC. As shown in Fig. 5, the TC terms helps push the decision boundary away from dense target regions, resulting in a well performing prediction function across domains.

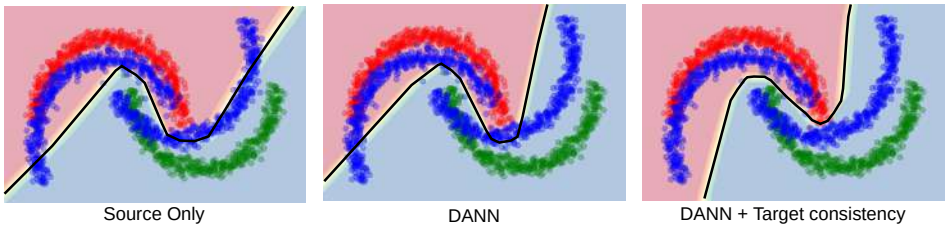


Figure 5: Effect of TC on *two moons* dataset. Red and green points are the instances of the two classes of the source domain. Blue points are target samples generated by rotating source samples. The black line shows the learned decision boundary, when using only source samples, with a DANN objective and with target consistency.

Theory

Non-conservative Domain Adaptation

Theorem 1 (From (Ben-David et al. 2007) and (Ben-David et al. 2010)). *Given a hypothesis class $\mathcal{H}$ and a hypothesis $h \in \mathcal{H}$:*

$$\varepsilon^t(h) \leq \varepsilon^s(h) + d_{\mathcal{H}\Delta\mathcal{H}} + \lambda_{\mathcal{H}} \tag{2}$$

where $d_{\mathcal{H}\Delta\mathcal{H}} := \sup_{h,h' \in \mathcal{H}} |\varepsilon^s(h,h') - \varepsilon^t(h,h')|$ and $\lambda_{\mathcal{H}} := \inf_{h \in \mathcal{H}} \{\varepsilon^t(h) + \varepsilon^s(h)\}$. In particular, provided a representation φ, and applying the inequality to $\mathcal{G} \circ \varphi := \{g \circ \varphi : g \in \mathcal{G}\}$:

$$\varepsilon^t(g\varphi) \leq \varepsilon^s(g\varphi) + d_{\mathcal{G}\Delta\mathcal{G}}(\varphi) + \lambda_{\mathcal{G}}(\varphi) \tag{3}$$

where $d_{\mathcal{G}\Delta\mathcal{G}}(\varphi) := \sup_{g,g' \in \mathcal{G}} |\varepsilon^s(g\varphi) - \varepsilon^t(g'\varphi)|$ and $\lambda_{\mathcal{G}}(\varphi) := \inf_{g \in \mathcal{G}} \{\varepsilon^s(g\varphi) + \varepsilon^t(g\varphi)\}$.

On the one hand, Eq. (2) shows the role of the hypothesis class capacity for bounding the target risk. The lower the hypothesis class sensitivity to changes in input distribution, the lower $d_{\mathcal{H}\Delta\mathcal{H}}$. On the other hand, Eq. (3) puts emphasis on representations: if source and target representations are aligned *i.e.*, $p(\mathbf{z}^s) \approx q(\mathbf{z}^t)$ for $\mathbf{z} := \varphi(\mathbf{x})$, then $d_{\mathcal{G}\Delta\mathcal{G}}(\varphi) = 0$.

One of the main difficulties of DA is achieving the optimal trade-off between source classification error and domain invariance of representations by minimizing $\varepsilon^s(g\varphi) + d_{\mathcal{G}\Delta\mathcal{G}}(\varphi)$ from Eq. (3), while maintaining a low $\lambda_{\mathcal{G}}(\varphi)$. This difficulty is referred as *non-conservative* DA in (Shu et al. 2018) *i.e.*, when the optimal joint classifier is significantly different from the target optimal classifier:

$$\inf_{h \in \mathcal{H}} \varepsilon^t(h) < \varepsilon^t(h^\lambda) \text{ where } h^\lambda := \arg\min_{h \in \mathcal{H}} \varepsilon^s(h) + \varepsilon^t(h) \tag{4}$$

Non-conservative DA can be described from the point of view of the hypothesis class as described by(Shu et al. 2018), *i.e.*, Eq. (2) from Theorem 1, then allowing change in representations to detect it, *i.e.*, $\inf$ computed on $\mathcal{H} = \mathcal{G} \circ \Phi$ in Eq. (4). Similarly, when provided with a representation φ, the optimal joint classifier differs from the target optimal classifier:

$$\inf_{g \in \mathcal{G}} \varepsilon^t(g\varphi) < \varepsilon^t(g^\lambda\varphi) \text{ where } g^\lambda := \arg\min_{g \in \mathcal{G}} \{\varepsilon^s(g\varphi) + \varepsilon^t(g\varphi)\} \tag{5}$$

This expression reflects the view of the literature of domain adversarial learning which puts emphasis on representations, *i.e.*, Eq. (3) from theorem Theorem 1. Note this definition only allows to modify the classifier, $\inf$ computed on $\mathcal{G}$, for detecting non-conservative DA, which may be a weak indication. We extend the denomination of non conservative DA to the case where $\varepsilon^t(\varphi) := \inf_{g \in \mathcal{G}} \varepsilon^t(g\varphi)$ is not optimal in φ.

Theoretical Analysis

We provide theoretical insights into the interaction between TC and class-level invariance. We consider $\varphi \in \Phi$ and $g \in \mathcal{G}$, which are modified to obtain $\tilde{\varphi}$ and $\tilde{g}$ defined as the closest instances such that $\tilde{g}\tilde{\varphi}$ verifies TC. For instance, they can be obtained by minimizing $\ell_2(\varphi, \tilde{\varphi}) + \ell_2(g, \tilde{g}) + \lambda \cdot \mathcal{L}_{\text{TC}}(\tilde{g}, \tilde{\varphi})$ where ℓ_2 is an L^2 error. When enforcing TC, we expect to decrease the target error *i.e.*, $\varepsilon^t(\tilde{g}\tilde{\varphi}) < \varepsilon^t(g\varphi)$. Noting $\rho := (1 - \varepsilon^t(\tilde{g}\tilde{\varphi})/\varepsilon^t(g\varphi))^{-1}$ and $\tilde{\mathbf{y}} := \tilde{g}\tilde{\varphi}(\mathbf{x})$, F a large enough critic function space, we adapt the theoretical analysis from (Bouvier et al. 2020):

$$\varepsilon^t(g\varphi) \leq \rho \left(\varepsilon^s(g\varphi) + 8 \sup_{\mathsf{f} \in \mathcal{F}} \left\{ \mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \mathsf{f}(\mathbf{z}^s)] - \mathbb{E}_{(\mathbf{z}, \tilde{\mathbf{y}}) \sim q}[\tilde{\mathbf{y}}^t \cdot \mathsf{f}(\mathbf{z}^t)] \right\} + \inf_{\mathsf{f} \in \mathcal{F}} \varepsilon^t(\mathsf{f}\varphi) \right) \tag{6}$$

More precisely, F has the following properties (Bouvier et al. 2020):

- (A1) F is symmetric (*i.e.* $\forall f \in \mathsf{F}, -f \in \mathsf{F}$) and convex.
- (A2) $\mathcal{G} \subset \mathsf{F}$ and $\{\mathsf{f} \cdot \mathsf{f}' \ ; \ \mathsf{f}, \mathsf{f}' \in \mathsf{F}\} \subset \mathsf{F}$.
- (A3) $\forall \varphi \in \Phi$, $\mathsf{f}_D(z) \mapsto \mathbb{E}_D[Y | \varphi(X) = z] \in \mathsf{F}$.
- (A4) For two distributions p and q on $\mathcal{Z}$, $p = q$, and $1 \leq c \leq C$, if and only if,

$$\text{IPM}(p, q; \mathsf{F}_c) := \sup_{\mathsf{f} \in \mathsf{F}} \{\mathbb{E}_p[\mathsf{f}_c(Z)] - \mathbb{E}_q[\mathsf{f}_c(Z)]\} = 0 \tag{7}$$

where f_c is the $c-$th coordinate of f.

Crucially, by observing that $\sup_{\mathsf{f} \in \mathcal{F}} \left\{ \mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \mathsf{f}(\mathbf{z}^s)] - \mathbb{E}_{(\mathbf{z}, \tilde{\mathbf{y}}) \sim q}[\tilde{\mathbf{y}}^t \cdot \mathsf{f}(\mathbf{z}^t)] \right\}$ is an *Integral Probability Measure* proxy of $\mathcal{L}_{\text{CLIV}}$, Eq. (6) reveals that class-level domain invariant representations can leverage feedback from an additional regularization, here the Target Consistency, to learn more transferable invariant representations.

Eq. (6) is obtained by applying Bound 4 (Inductive Bias and Guarantee, equation 14) from (Bouvier et al. 2020) by leveraging the inductive design:

$$\varepsilon^t(\tilde{g}\tilde{\varphi}) < \varepsilon^t(g\varphi) \tag{8}$$

provided by the Target Consistency. Note that, following the notations of (Bouvier et al. 2020), we have bounded the INV term (defined in equation 8 (Bouvier et al. 2020)) by the TSF term (defined in equation 9 (Bouvier et al. 2020)) leading to 6TSF, were $\text{TSF} := \sup_{\mathsf{f} \in \mathcal{F}} \left\{ \mathbb{E}_{(\mathbf{z}^s, \mathbf{y}^s) \sim p}[\mathbf{y}^s \cdot \mathsf{f}(\mathbf{z}^s)] - \mathbb{E}_{(\mathbf{z}, \tilde{\mathbf{y}}) \sim q}[\tilde{\mathbf{y}}^t \cdot \mathsf{f}(\mathbf{z}^t)] \right\}$ in our case. Besides, the constant term is $\rho := 1/(1 - \beta)$, not $\beta/(1 - \beta)$, since we bound $\varepsilon^t(g\varphi)$ not $\varepsilon^t(\tilde{g}\tilde{\varphi})$ where $\beta := \varepsilon^t(\tilde{g}\tilde{\varphi})/\varepsilon^t(g\varphi)$.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning* 79(1-2): 151–175.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, 137–144.

Bouvier, V.; Very, P.; Chastagnol, C.; Tami, M.; and Hudelot, C. 2020. Robust Domain Adaptation: Representations, Weights and Inductive Bias. *arXiv preprint arXiv:2006.13629* .

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4): 834–848.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, 1081–1090.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 113–123.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1): 2096–2030.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL https://openreview.net/forum?id=S1gmrxHFvB.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Liu, H.; Long, M.; Wang, J.; and Jordan, M. 2019. Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers. In *International Conference on Machine Learning*, 4013–4022.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 97–105. JMLR. org.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 1640–1650.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, 136–144.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2208–2217. JMLR. org.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 8024–8035.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8503–8512.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=H1q-TM-AW.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, 1195–1204.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7472–7481.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wang, X.; Jin, Y.; Long, M.; Wang, J.; and Jordan, M. I. 2019. Transferable Normalization: Towards Improving Transferability of Deep Neural Networks. In *Advances in Neural Information Processing Systems*, 1951–1961.

Yin, D.; Lopes, R. G.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, 13255–13265.