IMPLICIT STYLIZATION FOR DOMAIN ADAPTATION

Jinman Park, Francois Barnard, Saad Hossain, Sirisha Rambhatla
University of Waterloo
Waterloo, ON, Canada
{jinman.park, fbarnard, s42hossa, sirisha.rambhatla}@uwaterloo.ca

ABSTRACT

Unsupervised domain adaptation (UDA) aims to bridge the gap between source and target domains in the absence of target domain labels using two main techniques: input-level alignment (such as generative modeling and stylization) and feature-level alignment (which matches the distribution of the feature maps, e.g. gradient reversal layers). Motivated from the success of generative modeling for image classification, stylization-based methods were recently proposed for regression tasks, such as pose estimation. However, use of input-level alignment via generative modeling and stylization incur additional overhead and computational complexity which limit their use in real-world DA tasks. To investigate the role of input-level alignment for DA, we ask the following question: *is generative modeling or stylization really needed?* In other words, motivated from the title of the workshop: *what do we not need for successful domain adaptation?* Surprisingly, we find that input-alignment has little effect on regression tasks as compared to classification. Based on these insights, we develop a non-parametric feature-level domain alignment method – Implicit Stylization (ImSty) – which results in consistent improvements over SOTA both for regression and classification tasks, without the need for computationally intensive stylization and generative modeling. Our work conducts a critical evaluation of the role of generative modeling and stylization, at a time when these are also gaining popularity for domain generalization.

1 INTRODUCTION

With the burst of interest and applicability of deep learning applications in real-world settings, there is a growing need to transfer knowledge from a source domain with labeled data to a target domain without labeled data, specifically in data scarce regimes. Unsupervised domain adaptation (UDA) has emerged as a critical line of inquiry to address this challenge since reliable labeling of real-world datasets is often expensive and/or prohibitive.

UDA can largely be divided into two levels of alignment: feature-level alignment, and input-level alignment. Feature-level alignment such as statistical moment matching Long et al. (2017) and normalization statistics Csurka (2017); Zhao et al. (2018; 2020a) aim to generate intermediate feature map representations that are similar in statistical distribution between the source and target domain. While input-level alignment such as domain style transfer Sankaranarayanan et al. (2018); Huang & Belongie (2017) and adversarial learning Ganin et al. (2016); Liu et al. (2018); He et al. (2020) aims to a) either stylize the source domain images in the style of target domain images or b) generate training data for the target domain. For instance, in image classification, generative modeling has shown promising results Russo et al. (2018), and to this day still shows SOTA results in the challenging MNIST Deng (2012) → SVHN Netzer et al. (2011)task. Based on the this success, recently input-alignment has also been applied to regression task such as pose estimation Kim et al. (2022) where stylization Huang & Belongie (2017) was employed. However, the isolated effect of image stylization still remains ambiguous.

Given this ambiguity, our key observations is that input-level alignment via generative models and stylization are computationally expensive and in some data scarce regimes impractical. Consequently, a natural question emerges based on this observation: *is generative modeling or image stylization necessary for domain adaptation?*

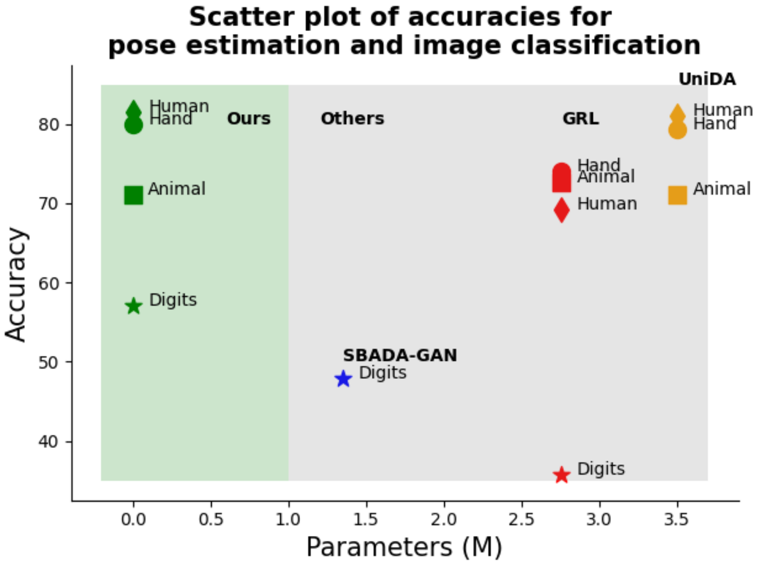


Figure 1: **Comparison of pose estimation and image classification accuracy for target domain respective to the number of trainable parameters used in domain alignment.** Our method achieves SOTA results in both image classification and pose estimation while significantly reducing the number of trainable parameters for domain alignment used in UDA. *Hand, Human, & Animal* refers to pose estimation while *Digits* refer to image classification.

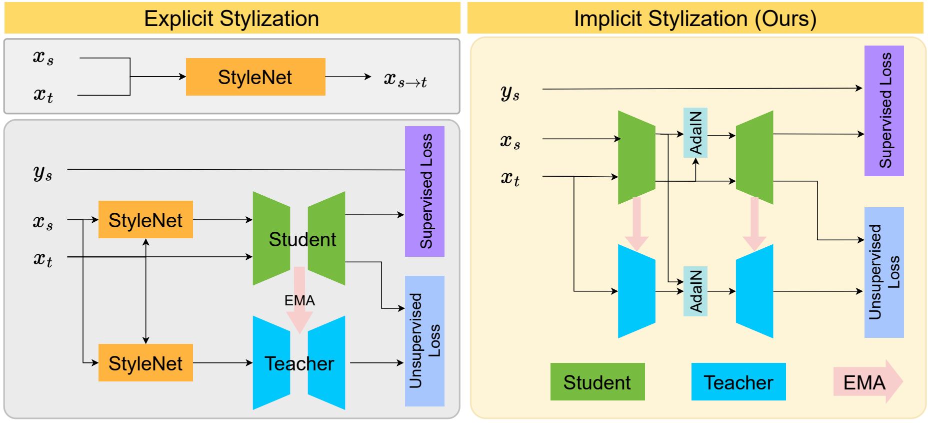
To begin addressing this question, we investigate the role of input-alignment and show that it does not consistently improve the results as opposed to a no-stylization baseline with a mean-teacher training scheme Tarvainen & Valpola (2017). Moreover, we propose a computationally efficient implicit stylization method, **ImSty**. Building on the concept of adaptive instance normalization (AdaIN) Huang & Belongie (2017) blocks, we incorporate AdaIN blocks into downstream tasks of image classification and pose estimation where the mini-batch level statistics of source domain and target domain are swapped. Our method avoids the need for explicitly training a generative model and generating input data in the style of the target domain, saving significant amounts of inefficient computations and training time.

We evaluate our proposed implicit stylization (**ImSty**) method on three pose estimation datasets with different levels of domain gaps and achieve SOTA results on all three datasets (Appendix A.4) while completely removing the number of trainable parameters and reducing the number of computations (MACs) by 99.99% for domain alignment compared against the explicit stylization used in Kim et al. (2022) (as shown in Figure 1). In addition, we experiment on MNIST Deng (2012) → SVHN Netzer et al. (2011), which is known for being notoriously difficult French et al. (2017); Shu et al. (2018); Dai et al. (2020); Kumar et al. (2018) without specific data augmentations such as intensity flipping and standardization. We show that with implicit stylization, we achieve SOTA results with minimal data augmentation and no specific data augmentation to reflect real-world scenarios where it can be costly and time consuming to find the right data augmentation that works for the target domain.

2 Related Works

Unsupervised Domain Adaptation (UDA) provides a suitable solution for data scarce domains. The aim of UDA is to associate a given labeled source domain, with abundant data, to an unlabeled target domain. Prevailing approaches for UDA often utilize adversarial learning methods S & Fleuret (2021); Yang et al. (2022); Zhang & Davison (2021), pseudo-labeling methods Yan et al.; Chu et al. (2022); Dubourvieux et al. (2021); Huang et al., or both as a foundation. These methods allow for a wide variety of focused implementations that provide reliable adaptation results.

Various attempts aim to remedy scarce source datasets with variations of domain adaptation (DA). With the development of generative adversarial networks, GenDA, Yang et al. (2021) propose an alternative approach to generative domain adaptation (GDA) applications. By freezing the parameters of a pre-trained GAN, Yang et al. (2021) reuse the prior information from the source GAN model



(a) Explicit stylization in Kim et al. (2022)

(b) Ours (Implicit stylization)

Figure 2: **Comparison between explicit stylization and our proposed implicit stylization.** Observe that explicit stylization (left) pre-trains a stylization model (i.e. StyleNet or neural style transfer model) before the downstream task. Then, the style transfer model is used to generate batches of source domain images stylized in target domain, and target domain images stylized in source domain. In comparison, our proposed implicit stylization method does not require any extra models (trainable parameters) and only requires a small amount of computation for the statistics in equation 1.

to adapt-to and generate new content in the target domain space. Moreover, a novel GDA approach Zhou et al. (2022) improves generalizability and reduces gaps within the domain by leveraging neural consistency in statistics to guide the generator, incorporating dual-level semantic consistency, while proposing intra-domain spectrum mixup. In addition to alignment at the input level, methods involving alignment at the feature level have been explored in numerous ways. One such pathway involves methods centered around the use of divergence measures such as maximum mean discrepancy, correlation alignment, contrastive domain discrepancy and Wasserstein Distance, which all share the goal of minimizing domain discrepancy at the feature level(Chen et al. (2020); Rozantsev et al. (2018); Kang et al. (2019); Liu et al. (2020)). Furthermore, studies have also employed adversarial methods Bousmalis et al. (2017) generally comprising of three modules: a feature classifier, discriminator, and label classifier.

Kim et al. (2022) achieved SOTA results in unsupervised domain adaptation for 2D pose estimation by unifying a framework that generalizes well on various poses with different levels of input and output variance. To merge the gap between source and target domain, the authors proposed stylizing the source images to target images for the student model, and stylizing the target images to the source images for the teacher model. However, the explicit stylization requires one to train a stylization model and generate new images every batch, causing computational inefficiency in training.

3 Implicit Stylization (ImSty)

In this section, we develop a feature-level alignment method called the implicit stylization module **ImSty** that replaces generative modeling (i.e. explicit stylization or generative models) in various domain adaptation tasks. Our method does not require training an explicit stylization model or require any additional trainable parameters.

Notation. Given a labeled pose dataset from the source domain $\mathbb{S} = \{(x_s^i, y_s^s)\}$ for $i = \{1, 2, \ldots, N_s\}$ with N_s pairs of images $x_s \in \mathbb{R}^{H\times W\times 3}$ and corresponding keypoint labels $y_s \in \mathbb{R}^{K\times 2}$, along with unlabeled pose dataset from the target domain $\mathbb{T} = \{x_t^i\}$ for $i = \{1, 2, \ldots, N_t\}$ with N_t images $x_t \in \mathbb{R}^{H\times W\times 3}$, the goal is to generalize a model h to the target domain $\mathbb{T}$ based on learning from a source domain $\mathbb{S}$.

Now, we describe the details of our proposed implicit stylization method. The main idea if **ImSty** is to incorporate the adaptive instance normalization block into the training pipeline to merge domain gaps without having the need of a generative model with a full auto-encoder structure for pixel-to-pixel generation. A mini-batch size of $n \ll N_s$ and $n \ll N_t$ are sampled from both the source domain $\mathbb{S}$ and target domain $\mathbb{T}$. Given a set of source domain images $\mathbb{X}_s = \{x_s^1, x_s^2, \ldots, x_s^n\}$ and a set of target

Table 1: **Digits classification results on MNIST → SVHN.** It is well established from French et al. (2017) that MNIST → SVHN is much more challenging than SVHN → MNIST. In addition, the results of many works vary greatly depending on a specific data augmentation. Since manually searching for data augmentations that work for the target domain is costly and time-consuming, we compare SOTA methods on minimal data augmentations. Observe that given minimal data augmentations, our method achieves SOTA results on UDA digits classification without generative methods. ”SBADA-GAN^R” indicates the reproduced values.

Method	Accuracy	Generative
RevGrad Ganin & Lempitsky (2015)	35.7	
DCRN Ghifary et al. (2016)	40.1	
G2A Sankaranarayanan et al. (2018)	36.4	✓
SE French et al. (2017)	37.5	
ATT Saito et al. (2017)	52.8	
SBADA-GAN^R Russo et al. (2018)	47.9 ± 1.7	✓
SBADA-GANRusso et al. (2018)	**61.1**	✓
PFAN Chen et al. (2019)	57.6 ± 1.8	
DIRT-T Shu et al. (2018)	54.5	
Ours (ImSty)	**57.8** ± 3.2	

domain images $\mathbb{X}_t = \{x_t^1, x_t^2, \dots, x_t^n\}$, $\mathbb{X}_s$ and $\mathbb{X}_t$ are passed through a ResNet-101 He et al. (2016) backbone R to obtain highly-semantic feature maps $\mathbf{F}_s = R(\mathbb{X}_s)$ and $\mathbf{F}_t = R(\mathbb{X}_t)$. Each channel of C in $\mathbf{F}_s \in \mathbb{R}^{H' \times W' \times C}$ and $\mathbf{F}_t \in \mathbb{R}^{H' \times W' \times C}$ are normalized to $\mathcal{N}(0, 1)$ while keeping the mean $\mu \in \mathbb{R}^C$ and standard deviation $\sigma \in \mathbb{R}^C$ of each channel used for the normalization for $j \in \{s, t\}$:

$$\mu_j^i = \texttt{Mean}(\mathbf{F}_j[:,:,i]),\ \ \sigma_j^i = \texttt{STD}(\mathbf{F}_j[:,:,i]),\ \ \mu_j = [\mu_j^1, \mu_j^2, \dots, \mu_j^C],\ \ \sigma_j = [\sigma_j^1, \sigma_j^2, \dots, \sigma_j^C]$$
$$\mathbf{F}_{\text{norm},s} = \frac{\mathbf{F}_s - \mu_s}{\sigma_s},\ \ \mathbf{F}_{\text{norm},t} = \frac{\mathbf{F}_t - \mu_t}{\sigma_t} \tag{1}$$

Then, the mean and standard deviation from the opposite domain are utilized to reverse the normalization in the following way building on the concept of AdaIN:

$$\begin{aligned} \mathbf{F}_{s \to t} &= \alpha(\mathbf{F}_{\text{norm},s} \cdot \sigma_t + \mu_t) + (1 - \alpha)\mathbf{F}_s \\ \mathbf{F}_{t \to s} &= \alpha(\mathbf{F}_{\text{norm},t} \cdot \sigma_s + \mu_s) + (1 - \alpha)\mathbf{F}_t \end{aligned} \tag{2}$$

$\mathbf{F}_{s \to t}$ and $\mathbf{F}_t$ are passed to the student decoder D_{stu} that up-samples the feature maps to $\hat{y}_s \in \mathbb{R}^{h \times w \times J}$ and $\hat{y}_t \in \mathbb{R}^{h \times w \times J}$ with height h, width w, and number of joints K. Then, $\mathbf{F}_{t \to s}$ is passed to the teacher decoder D_{tea} to generate pseudo-labels y_t. Formally, we have:

$$\hat{y}_s = D_{\text{stu}}(\mathbf{F}_{s \to t}), \quad \hat{y}_t = D_{\text{stu}}(\mathbf{F}_t), \quad y_t = D_{\text{tea}}(\mathbf{F}_{t \to s}) \tag{3}$$

Mean-squared error (MSE) metric is used as the loss function for both supervised $\mathcal{L}_{\text{sup}}$ and unsupervised $\mathcal{L}_{\text{unsup}}$ loss with a weighting factor λ for the overall loss $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{sup}} = \text{MSE}(y_s, \hat{y}_s), \quad \mathcal{L}_{\text{unsup}} = \text{MSE}(y_t, \hat{y}_t), \quad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}} \tag{4}$$

Remaining details on the overall mean-teacher pipeline such as pseudo-label masking, normalization, adaptive occlusions, data augmentations, and reverse augmentations can be found in Kim et al. (2022). Refer to Figure 2 for a diagram of the comparison between explicit stylization and our proposed implicit stylization method.

4 Experiments

In this section, we begin to incrementally reveal the excessive nature of generative methods for UDA by showing either on-par or better results on both image classification and pose estimation datasets with close to zero additional computations and trainable parameters.

Main Results. In Table 2, we compare our implicit stylization model (ImSty) with SOTA models for UDA pose estimation on three sets of 2D pose estimation datasets for domain adaptation. On average over the three datasets, we achieve 0.4% improvements in PCK@0.05. Moreover, it is further demonstrated in Table 3 that implicit stylization reduces the required computation by over 99.99% and completely removes the number of trainable parameters needed for domain merging compared

Table 2: **Pose Estimation Comparison for hand pose, animal pose, and human pose estimation.** Our proposed ImSty method achieves SOTA results in pose estimation across three different domains (hand pose, human pose, and animal pose). In hand pose and human pose we achieve the highest accuracy while achieving competitive results in animal pose. In the table, the superscripts "P" and "R" denote the published metrics (reported in the papers) and the reproduced values, respectively. For hand pose estimation, MCP, PIP, DIP are acryonyms for metacarpophalangeal, proximal interphalangeal, and distal interphalangeal respectively.

Rendered Hand Pose Dataset →Hand-3D-Studio

Method	MCP	PIP	DIP	Fingertip	All
CCSSL Mu et al. (2020)	81.5	79.9	74.4	64.0	75.1
UDA-Animal Li & Lee (2021)	82.3	79.6	72.3	61.5	74.1
RegDA Jiang et al. (2021)	79.6	74.4	71.2	62.9	72.5
UniDA^P Kim et al. (2022)	86.7	84.6	78.9	68.1	79.6
UniDA^{R}Kim et al. (2022)	87.1	85.1	**79.3**	68.5	79.9
Ours (ImSty)	**87.6**	**85.5**	78.9	**68.7**	**80.1**

Synthetic Animal Dataset → TigDog Dataset

Method	Horse							
	Eye	Chin	Hoof	Hip	Knee	Shd	Elbow	All
CCSSL Mu et al. (2020)	89.3	92.6	65.0	78.1	73.1	69.5	70.0	73.1
UDA-Animal Li & Lee (2021)	86.9	**93.7**	**72.6**	81.9	**79.1**	**76.4**	70.6	**77.5**
RegDA Jiang et al. (2021)	89.2	92.3	63.2	77.5	72.7	70.5	71.5	73.2
UniDA^P Kim et al. (2022)	91.3	92.5	66.6	74.2	77.0	74.0	**75.8**	76.4
UniDA^R Kim et al. (2022)	91.5	93.6	67.2	**82.3**	77.0	73.1	74.8	75.6
Ours (ImSty)	**91.6**	92.8	66.2	77.1	76.5	72.6	74.7	75.4

Method	Tiger							
	Eye	Chin	Hoof	Hip	Knee	Shd	Elbow	All
CCSSL Mu et al. (2020)	94.3	91.3	70.2	70.2	59.1	49.5	**53.9**	66.7
UDA-Animal Li & Lee (2021)	98.4	87.2	**73.4**	**74.9**	62.0	49.4	49.8	67.7
RegDA Jiang et al. (2021)	93.3	92.8	60.7	67.8	55.4	50.3	50.2	61.8
UniDA^P Kim et al. (2022)	**98.5**	96.9	72.8	63.7	**62.8**	**56.2**	52.3	**67.9**
UniDA^R Kim et al. (2022)	98.3	**97.0**	72.1	71.7	62.2	53.4	53.0	67.1
Ours (ImSty)	97.7	96.3	72.1	72.5	61.3	52.9	52.2	66.9

SURREAL → Leeds Sports Pose

Method	Shd	Elbow	Wrist	Hip	Knee	Ankle	All
CCSSL Mu et al. (2020)	36.8	66.3	63.9	59.6	67.3	70.4	60.7
UDA-Animal Li & Lee (2021)	61.4	77.7	75.5	65.8	76.7	78.3	69.2
RegDA Jiang et al. (2021)	62.7	76.7	71.1	81.0	80.3	75.3	74.6
UniDA^P Kim et al. (2022)	69.2	84.9	83.3	85.5	84.7	84.3	82.0
UniDA^R Kim et al. (2022)	68.4	**85.6**	83.1	**86.2**	85.0	84.2	82.0
Ours (ImSty)	**71.1**	85.2	**83.4**	85.2	**86.1**	**85.1**	**82.6**

against the current SOTA UDA pose estimation Kim et al. (2022). In addition, Table 1 demonstrates the superior performance of our implicit stylization method in image classification compared to both generative and non-generative methods given minimal data augmentations.

5 Discussion

Domain alignment is an essential part of UDA, shifting the distribution of source and target domains closer to each other. Given the necessity of domain alignment, it was unclear if the mean-teacher scheme in pose estimation really needed input-level alignment via stylization and if image classification needed generative modeling. By achieving SOTA results in both UDA for pose estimation and image classification, we demonstrate that input-level alignment via generative methods and stylization may not be necessary. The resulting impact reduces the work, time, and computational cost required for the generative model by a significant margin. In addition, any instability caused by generative adversarial networks can be avoided with implicit stylization. Finally, this opens doors for use of domain adaptation in data scarce regimes.

Future Work. Although ImSty achieved SOTA performance on regression tasks, while significantly reducing the amount of computation and trainable parameters (by replacing the input-level alignment model with a new feature-level alignment module), the effect on image classification remains small. This raises an interesting question about the role and limitations of domain alignment in the mean-teacher training scheme for various machine learning tasks, which we aim to investigate next.

REFERENCES

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, 2017.

Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3422–3429, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 627–636, 2019.

Tong Chu, Yahao Liu, Jinhong Deng, Wen Li, and Lixin Duan. Denoised Maximum Classifier Discrepancy for Source-Free Unsupervised Domain Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):472–480, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i1.19925. URL `https://ojs.aaai.org/index.php/AAAI/article/view/19925`.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

Shuyang Dai, Yu Cheng, Yizhe Zhang, Zhe Gan, Jingjing Liu, and Lawrence Carin. Contrastively smoothed class alignment for unsupervised domain adaptation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Articulated motion discovery using pairs of trajectories. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2151–2160, 2015.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Fabian Dubourvieux, Romaric Audigier, Angelique Loesch, Samia Ainouz, and Stephane Canu. Unsupervised Domain Adaptation for Person Re-Identification through Source-Guided Pseudo-Labeling. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4957–4964, January 2021. doi: 10.1109/ICPR48806.2021.9412964. ISSN: 1051-4651.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

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.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pp. 597–613. Springer, 2016.

Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Classification-aware semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 964–965, 2020.

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*, pp. 770–778, 2016.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model Adaptation: Historical Contrastive Learning for Unsupervised Domain Adaptation without Source Data. pp. 15.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6780–6789, 2021.

Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, pp. 5. Aberystwyth, UK, 2010.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. *arXiv preprint arXiv:2204.00172*, 2022.

Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018.

P. Langley. Crafting papers on machine learning. In Pat Langley (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1482–1491, 2021.

Xiaofeng Liu, Yang Zou, Lingsheng Kong, Zhihui Diao, Junliang Yan, Jun Wang, Site Li, Ping Jia, and Jane You. Data augmentation via latent space interpolation for image classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 728–733. IEEE, 2018.

Xiaofeng Liu, Yuzhuo Han, Song Bai, Yi Ge, Tianxing Wang, Xu Han, Site Li, Jane You, and Jun Lu. Importance-aware semantic segmentation in self-driving with discrete wasserstein training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11629–11636, 2020.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.

Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12386–12395, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018.

Paolo Russo, Fabio M Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8099–8108, 2018.

Prabhu Teja S and François Fleuret. Uncertainty Reduction for Model Adaptation in Semantic Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9608–9618, June 2021. doi: 10.1109/CVPR46437.2021.00949. ISSN: 2575-7075.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 2988–2997. PMLR, 2017.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8503–8512, 2018.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 109–117, 2017.

Hao Yan, Yuhong Guo, and Chunsheng Yang. Augmented Self-Labeling for Source-Free Unsupervised Domain Adaptation. pp. 7.

Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-Shot Generative Domain Adaptation, November 2021. URL `http://arxiv.org/abs/2111.09876`. arXiv:2111.09876 [cs].

Yiju Yang, Taejoon Kim, and Guanghui Wang. Multiple Classifiers Based Adversarial Training for Unsupervised Domain Adaptation. In *2022 19th Conference on Robots and Vision (CRV)*, pp. 40–47, May 2022. doi: 10.1109/CRV55824.2022.00014.

Youshan Zhang and Brian D. Davison. Adversarial Continuous Learning in Unsupervised Domain Adaptation. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (eds.), *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12662, pp. 672–687. Springer International Publishing, Cham, 2021. ISBN 978-3-030-68789-2 978-3-030-68790-8. doi: 10.1007/978-3-030-68790-8_52. URL `http://link.springer.com/10.1007/978-3-030-68790-8_52`. Series Title: Lecture Notes in Computer Science.

Sicheng Zhao, Bichen Wu, Joseph Gonzalez, Sanjit A Seshia, and Kurt Keutzer. Unsupervised domain adaptation: From simulation engine to the realworld. *arXiv preprint arXiv:1803.09180*, 2018.

Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020a.

Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2478–2482. IEEE, 2020b.

Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 335–356. Springer, 2022.

Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pp. 4903–4911, 2017.

A Appendix

A.1 Societal Impact Statement

Our work questions the use of generative modeling and stylization which require large amount of data and compute capabilities which has an adverse impact on the environment. As a result, our work contributes towards sustainable machine learning. Moreover, by developing a technique that does not rely on generative modeling or stylization, our work makes domain adaptation amenable for

Table 3: **Comparison of trainable parameters and computational costs (MACs) for merging domain gaps.** Our implicit stylization method requires no additional trainable parameters and minimal computations to calculate feature-level statistics to merge the gap between source and target domain. For fair comparison, all the methods use image resolution of 256×256.

Method	Params. (M)	MACs (G)	Generative
SBADA-GAN Russo et al. (2018)	51.73	42.54	✓
StyleNet Huang & Belongie (2017)	3.51	63.11	✓
RevGrad Ganin & Lempitsky (2015)	2.75	0.34	
Ours (ImSty)	**0.00**	**1.84e-3**	

data-scare regimes. That being said, since we investigate role of these techniques in the context of computer vision, applications to surveillance by public or private entities raises privacy invasion and human rights concerns. Privacy preserving machine learning offers a way to address these risks. At the same time, there is a need develop legal protections for users, and regulations for organizations utilizing their data.

A.2 Evaluation Metrics

For **2D pose estimation**, the standard metric is Percentage of Correct Keypoint (PCK). Following the work of Kim et al. (2022) and Li & Lee (2021), all experiments are reported with PCK@0.05 that measures the ratio of correct keypoints that are within 5% of the image resolution.

For **image classification**, we report on the overall accuracy without class balancing following recent recommendations French et al. (2017); Shu et al. (2018); Dai et al. (2020); Kumar et al. (2018).

A.3 Ablation Studies

The main goal of this work is to investigate the role of generative modeling/stylization for domain alignment. From table 4, we see a clear distinction between the importance of stylization for classification tasks versus pose estimation. Stylization applied to pose estimation offers lackluster results as we see a maximum difference of only two percent between stylized and unstylized accuracy. This suggests a need to revisit the role of domain alignment in the mean-teacher training scheme. However, when applied to a classification task, the impact of stylization cannot be understated. When applying ImSty to the MNIST → SVHN task, we see a considerable improvement of 19.8% against the SOTA generative method increasing from 38.0% to 57.8%. This improvement demonstrates the ability of ImSty to achieve theses results without any specific data augmentation.

A.4 Datasets

We first evaluate our implicit stylization pipeline using three sets of commonly used UDA pose estimation datasets to cover high variances in both input and output settings: **Rendered Hand Pose Dataset** Zimmermann & Brox (2017) → **Hand-3D-Studio** Zhao et al. (2020b), **SURREAL** Varol et al. (2017) → **Leeds Sports Pose** Johnson & Everingham (2010), and **Synthetic Animal Dataset** Mu et al. (2020) → **TigDog Dataset** Del Pero et al. (2015). In addition, we evaluate a challenging set of image classification datasets for UDA: **MNIST** Deng (2012) → **SVHN** Netzer et al. (2011). Based on multiple works French et al. (2017); Shu et al. (2018); Dai et al. (2020); Kumar et al. (2018), it has been established that SVHN → MNIST is a much easier task where SOTA results are close to supervised learning. However, MNIST → SVHN is a much more challenging task where the common consensus is that each proposed method needs a specific data augmentation such as intensity flipping or image standardization to achieve SOTA results. In the real-world setting, it can be time consuming, non-trivial, and not guaranteed to find a data augmentation that works for a particular dataset. Therefore, it is our goal to demonstrate that implicit stylization works without specific data augmentation. Our method trains with only minimal amount of data augmentations.

Table 4: **Role of stylization in pose estimation vs classification.** Observe the minimal difference between pose estimation accuracy with and without stylization. This small difference is apparent across all pose estimation tasks questioning the role of domain alignment in pose estimation. However, for MNIST → SVHN ImSty far outperforms our baseline without stylization highlighting the importance of domain alignment.

Rendered Hand Pose Dataset →Hand-3D-Studio		
Method	UniDA w.o. Stylization	Ours (ImSty)
MCP	86.5 ± 0.5	**87.1** ± **0.6**
PIP	85.2 ± 0.6	**85.4** ± **0.2**
DIP	78.1 ± 0.4	**78.8** ± **0.2**
Fingertip	66.6 ± 0.3	**68.4** ± **0.3**
All	79.1 ± 0.4	**80.0** ± **0.1**
Synthetic Animal Dataset → TigDog Dataset		
Method	Horse	
	UniDA w.o. Stylization	Ours (ImSty)
Eye	89.8 ± 0.5	**90.9** ± **1.2**
Chin	**92.8** ± **0.2**	92.7 ± 0.2
Hoof	**67.5** ± **0.7**	65.5 ± 0.6
Hip	**77.6** ± **0.9**	75.6 ± 1.5
Knee	**76.3** ± **0.8**	75.7 ± 0.8
Shoulder	**72.2** ± **1.1**	71.6 ± 1.2
Elbow	72.4 ± 1.0	**73.9** ± **0.7**
All	**75.5** ± **0.4**	75.2 ± 0.2
Method	Tiger	
	UniDA w.o. Stylization	Ours (ImSty)
Eye	97.2 ± 1.0	**97.3** ± **0.5**
Chin	95.6 ± 0.2	**96.1** ± **0.2**
Hoof	70.7 ± 0.8	**71.2** ± **1.2**
Hip	**72.7** ± **2.6**	71.5 ± 1.1
Knee	**61.1** ± **0.3**	60.6 ± 0.8
Shoulder	49.2 ± 0.8	**51.8** ± **0.9**
Elbow	**51.0** ± **1.7**	50.8 ± 1.5
All	66.3 ± 0.8	**66.5** ± **0.8**
SURREAL → Leeds Sports Pose		
Method	UniDA w.o. Stylization	Ours (ImSty)
Shoulder	66.3 ± 1.3	**67.9** ± **3.1**
Elbow	83.1 ± 1.8	**84.5** ± **1.2**
Wrist	80.7 ± 1.8	**82.4** ± **1.3**
Hip	84.8 ± 0.3	**84.9** ± **0.2**
Knee	83.4 ± 0.8	**85.0** ± **1.1**
Ankle	83.1 ± 1.3	**84.4** ± **0.9**
All	80.2 ± 0.7	**81.5** ± **1.2**
MNIST → SVHN		
Method	Ours w.o. stylization	Ours (ImSty)
All	38.0 ± 4.0	**57.8** ± **3.3**

A.5 Implementation and Reproducibility Details

The implementation is done with PyTorch with three random seeds 22, 42, and 102. Following the work of Kim et al. (2022) we use the following data augmentations. Code is attached for reproducibility.

Human Pose Data Augmentation. Random data augmentations used: 60 degree rotations, shear (-30, 30), translations (0.05, 0.05), scaling (0.6, 1.3), and 0.25 contrast.

Hand Pose Data Augmentation. Random data augmentations used: 180 degree rotations, shear (-30, 30), translations (0.05, 0.05), scaling (0.6, 1.3), and 0.25 contrast.

Animal Pose Data Augmentation. Random data augmentations used: 60 degree rotations, shear (-30, 30), translations (0.05, 0.05), scaling (0.6, 1.3), and 0.25 contrast.

Digits Data Augmentation. Random data augmentations used: rotations, translations, scaling, adjust brightness, adjust contrast.

Learning parameters. For all **pose estimation** tasks, Adam with base learning rate of $1e^{-4}$ and learning rate decay of 0.1 at 45 and 60 epochs is chosen as the optimizer for training. For **digits classification** task, Adam with base learning rate of $1e^{-3}$ is chosen as the optimizer for training.

Pre-training. With pose estimation, source-only training is done for 40 epochs and 1 epoch for digits classification.

Compute Infrastructure. A batch size of 32 for pose estimation and 64 digits classification is fed to a single NVIDIA A100 GPU for accelerated training with AMD Milan 7413 CPU available via the shared high performance computing infrastructure.

Hyperparameters. With pose estimation, standard ResNet-101 is used as the backbone and three blocks of upsampling blocks with 256 channels (comprised of 2D transpose convolution, 2D batch norm, ReLU) are used in the decoder. The whole network is trained for 70 epochs.

With digits classification, standard LeNet-5 is used for the backbone and two linear blocks (Linear(500 units), ReLU, Linear(10 units)) are used as a classification head. The whole network is trained for 300 epochs.