Robust Unsupervised Domain Adaptation through Negative-View Regularization

Joonhyeok Jang[1§], Sunhyeok Lee[1§], Seonghak Kim[1], Jung-un Kim[2], Seonghyun Kim[2], Daeshik Kim[1]
[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]SHINSEGAE I&C
{jang0727, lee.sunhyeok, hakk35, daeshik}@kaist.ac.kr, {jukim, kimsh0210}@shinsegae.com

Abstract

In the realm of Unsupervised Domain Adaptation (UDA), Vision Transformers (ViTs) have recently demonstrated remarkable adaptability surpassing that of traditional Convolutional Neural Networks (CNNs). Nevertheless, the patch-based structure of ViTs heavily relies on local features within image patches, potentially leading to reduced robustness when confronted with out-of-distribution (OOD) samples. To address this concern, we introduce a novel regularizer tailored specifically for UDA. By leveraging negative views, i.e. target-domain samples applied by negative augmentations, we make the learning process more intricate, thereby preventing models from taking shortcuts in spatial context recognition. We present a novel loss function, rooted in contrastive principles, to effectively distinguish between the negative views and original target samples. By integrating this novel regularizer with existing UDA methodologies, we guide ViTs to prioritize context relationships among local patches, thereby enhancing the robustness of ViTs. Our proposed Negative View-based Contrastive (NVC) regularizer substantially boosts the performance of baseline UDA methods across diverse benchmark datasets. Furthermore, we release new dataset, Retail-71, comprising 71 classes of images commonly encountered in retail stores. Through comprehensive experimentation, we showcase the effectiveness of our approach on traditional benchmarks as well as the novel retail domain. These results substantiate the robust adaptation capabilities of our proposed method. Our method is implemented at our repository.

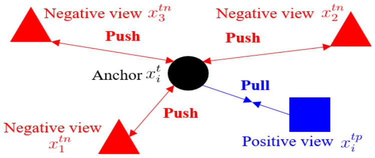


Figure 1. Schematic representation of our proposed Negative View-based Contrastive (NVC) regularizer. For a given anchor sample x_i^t from the target batch $\mathcal{B}^T$, we identify its positive view x_i^{tp} (indicated by a blue square) to form a positive pair, thereby aligning x_i^t with x_i^{tp}. Conversely, all other samples within $\mathcal{B}^T$ are designated as negative views (denoted by red triangles), each forming a negative pair with the anchor, ensuring their strategic separation in the learned representation space.

1. Introduction

The field of deep learning has witnessed remarkable progress across a variety of recognition tasks, notably in image classification. Within the domain of Computer Vision (CV), extensive research has been conducted on a plethora of methodologies. These encompass Convolutional Neural Networks (CNNs) [9, 10, 14, 29, 31] and, more recently, Vision Transformers (ViTs) [3–5, 18, 33] — the latter drawing foundational inspiration from [36]. ViTs, in particular, have showcased a remarkable aptitude for generalization in supervised learning scenarios, often outperforming CNN-based techniques.

Notwithstanding these advancements, Deep Neural Networks (DNNs) grapple with the phenomenon of domain shifts, wherein models trained on a source dataset struggle with data from a dissimilar distribution in the target domain [6]. A conventional solution involves annotating the target domain data, which is not always feasible due to the associated costs. Unsupervised Domain Adaptation (UDA) has emerged as a strategic alternative, relying on a labeled source dataset and an unlabeled target dataset, typically exhibiting significant domain disparities, such as synthetic versus real-world images. UDA aims to narrow this domain gap, thereby enhancing model performance on the target dataset.

In the realm of UDA, a variety of CNN-based methods have been continuously proposed [6, 12, 15, 19–21, 25, 28, 32, 34]. Concurrently, ViT-based UDA approaches have gained prominence [26, 30, 38], outshining CNNs on benchmarks like Office-31 [27], Office-Home [37], and VisDA-

[§] These authors contributed equally.

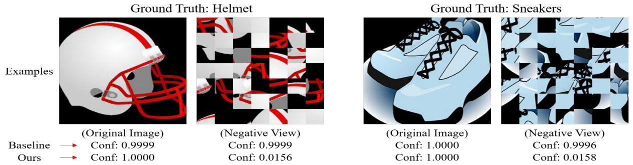


Figure 2. Illustrative examples from the *Cl* domain of the Office-Home dataset, juxtaposed with their negative views. The baseline UDA method [26] yields high confidence for negative views, despite human non-recognition, indicating a local feature bias resistant to negative augmentation, such as *P-Shuffle*.

2017 [23]. Nevertheless, ViTs face challenges with local feature dependence [24], impacting robustness to out-of-distribution (OOD) samples in classification tasks. To combat this, Qin *et al.* [24] introduced 'negative augmentation', enhancing ViT robustness by contrasting negatively augmented images with their original counterparts.

We have empirically investigated ViT's local feature dependence in a UDA context, specifically in the Ar→Cl scenario within the Office-Home dataset, employing SDAT as a baseline [26]. The employment of negative augmentation, herein patch-based shuffling, revealed ViT's limitations, as noted by Qin *et al.* [24], within UDA frameworks (refer to Figure 2). Considering the disparate distributions in UDA settings, negative views may encourage ViTs to internalize contextual information pertinent to the target domain, thus improving model robustness and UDA efficacy. Following Qin *et al.* [24], we propose a novel contrastive learning-based regularizer loss to enhance UDA methodologies.

In a practical context, consider an autonomous retail store. Here, the ordering system — comprising top-view webcams, object detection models, and image classifiers — is critical. This system's efficiency is contingent on the robustness and precision of the image classifier. However, the frequent product rotation in retail necessitates an adaptive approach to image classification, as direct labeling of onsite product images is cost-prohibitive.

An alternative involves compiling and labeling product images in a controlled, in-lab environment, yet this introduces a domain gap when applied to real-world scenarios, potentially compromising performance. Addressing this, we align product image classification with UDA, using labeled in-lab images as the source dataset and unlabeled real-world images as the target dataset to train a classifier robust to common disturbances like motion blur and hand occlusions.

We introduce Retail-71, a novel UDA dataset tailored to the retail sector, featuring both source and target domain images and a test set with varying difficulty levels. Additionally, we propose a rule-based synthesis technique to foster smoother domain adaptation on Retail-71.

Our principal contributions are summarized as follows:

- We introduce a simple yet effective Negative View-based Contrastive (NVC) regularizer loss. This loss can be applied to a wide array of datasets and seamlessly integrated into existing UDA methodologies. The NVC regularizer is designed to promote the capture of global image context by Vision Transformers, thereby enhancing UDA performance across several benchmarks.

- We present a novel UDA benchmark, dubbed Retail-71, tailored to the retail sector. This benchmark is specifically curated to assess the resilience of models against noisy product images that exhibit characteristics such as hand occlusion and motion blur, distinguishing it from existing UDA datasets. The Retail-71 test set is stratified into three levels of difficulty, enabling nuanced evaluations of model robustness against varying degrees of image degradation. We release new benchmark at our repository.

- We develop a rule-based synthesis technique for product classification, aimed at generating an intermediate domain that facilitates smoother domain adaptation on Retail-71. This synthesis approach, serving as an augmentation method, enhances the training process by interpolating between the source and target domains, which in turn bolsters UDA performance.

2. Related works

2.1. Unsupervised Domain Adaptation

Frameworks for UDA Unsupervised domain adaptation (UDA) leverages a labeled source dataset and an unlabeled target dataset to train a network, such as a Vision Transformer (ViT) or a Residual Network (ResNet), with the objective of enhancing generalizability to the target domain. Several frameworks have been advanced to address the domain discrepancy issue. Approaches such as Domain-Adversarial Neural Networks (DANN) [6] and Contrastive Domain Adaptation (CDAN) [20] draw insights from adversarial training principles in generative adversarial networks (GANs) [7], implementing Domain Adversarial Training (DAT) to bridge domain gaps. Meanwhile, Domain Adaptive Neural Networks (DAN) [19] and Joint Adaptation Networks (JAN) [21] aim to minimize domain divergence by employing statistical measures like Maximum Mean Discrepancy (MMD) and Joint MMD, respectively. The Prototypical Contrastive Transfer (PCT) model [32] seeks to align source-domain prototypes with target samples to minimize class-wise dissimilarity. In contrast to the predominant CNN-based methods, CDTrans [38] adopts a transformer-based architecture, capitalizing on a cross-attention mechanism to facilitate domain adaptation.

Regularizers for UDA In addition to framework design, certain studies have introduced regularizers to bolster UDA efficacy. Minimum Class Confusion (MCC) [11] focuses on minimizing the confusion of class predictions within the target domain. The Self-supervised Safe Training Routine (SSRT) [30], designed for transformer-based models, employs the minimization of Kullback–Leibler divergence between the predictions of perturbed and original target samples to foster robustness. The Nuclear norm-based Wasserstein Distance (NWD) [1] proposes a metric incorporating the first-order Wasserstein distance and nuclear norm to promote prediction certainty and diversity within the model. Similarly, Selective Domain Adversarial Training (SDAT) [26] advocates for applying Sharpness Aware Minimization (SAM) selectively, based on theoretical and empirical evaluations. This study introduces a negative view-based regularizer, which enhances UDA by augmenting existing methods, requiring only unlabeled target data.

2.2. Contrastive Learning

Contrastive learning has seen significant advancements within self-supervised paradigms. Pioneering methods such as Momentum Contrast (MoCo) [8] and SimCLR [2] have underscored the efficacy of the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss, with SimCLR demonstrating its effectiveness in large batch settings. Subsequently, the Supervised Contrastive (SupCon) loss [13] exhibited the benefits of NT-Xent in a supervised context. Within UDA, contrastive learning frameworks have shown promise. Contrastive Adaptation Network (CAN) [12] integrates an MMD-based contrastive approach, while Probabilistic Contrastive Learning (PCL) [16] and Contrastive Pseudo-Labeling Graph Alignment (CPGA) [25] utilize losses akin to NT-Xent, such as the InfoNCE loss [22]. This work proposes a novel contrastive regularizer loss employing negative views of target samples, emphasizing the potential of negative pairings in UDA scenarios.

3. Methodology

We commence with an examination of the Unsupervised Domain Adaptation (UDA) paradigm and delineate our innovative approach to patch-based negative augmentation. Subsequently, we introduce a regularizer loss that capitalizes on this augmentation technique.

3.1. Unsupervised Domain Adaptation

$$\mathcal{L}_{\text{task}} = \sum_{i=1}^{N_s} \ell(G(F(x_i^s)), y_i^s), \quad \ell(\hat{y}, y) = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{1}$$

In the UDA framework, a neural network (e.g., Vision Transformer, ViT) integrates a feature extractor F and a classification head G. It is trained on a labeled source data set $\mathcal{X}^S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and an unlabeled target data set $\mathcal{X}^T = \{x_i^t\}_{i=1}^{N_t}$, both presumed to share an identical label space. The classifier G yields a C-dimensional vector $\hat{y}$, representing the confidence scores across classes. Ordinarily, the network is optimized using a task-specific loss $\mathcal{L}_{\text{task}}$ (as per Equation 1), typically employing cross-entropy for classification tasks, calculated exclusively on the labeled source data. A network honed solely with $\mathcal{L}_{\text{task}}$ on source data is referred to as a *source-only* model.

However, a *source-only* model is inherently susceptible to performance degradation due to domain discrepancies, prompting UDA techniques to introduce an additional adaptation loss $\mathcal{L}_{\text{adapt}}$. This loss leverages unlabeled target data to adjust the network to the target domain characteristics. Furthermore, the UDA corpus encompasses several propositions [1, 11, 26, 30] for an auxiliary loss function $\mathcal{L}_{\text{reg}}$, taking the form of a regularizer. These can be amalgamated with other UDA strategies, enhancing the overall adaptability of the model. The comprehensive loss function for UDA is articulated in Equation 2. The contribution of this work lies within the realm of $\mathcal{L}_{\text{reg}}$, wherein we proffer a novel regularizer loss formulation.

$$\mathcal{L}_{\text{UDA}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{adapt}} + \mathcal{L}_{\text{reg}} \tag{2}$$

3.2. Patch-based Negative Augmentation

Yao Qin *et al.* [24] introduced an innovative augmentation approach termed as patch-based negative augmentation. This methodology encompasses three distinct techniques: patch-based shuffling (P-Shuffle), patch-based rotation (P-Rotate), and patch-based infilling (P-Infill). Contrary to conventional semantic-preserving augmentations, also known as positive augmentations, these negative augmentations disrupt the inherent semantic integrity of images, rendering the global structure—or context—unrecognizable. The resulting images, whose semantics have been obscured, are referred to as *negative views*.

Despite the compromised global context, the local features within image patches remain intact (as illustrated in Figure 2). Yao Qin et al. have elucidated that Vision Transformers (ViTs), due to their intrinsic patch-based architecture, place substantial emphasis on these local features that persevere in the aftermath of negative augmentation.

3.3. Negative View-based Contrastive Regularizer

$$\begin{aligned}\mathcal{L}_{\text{ours}} =& \frac{1}{|\mathcal{B}^T|} \sum_{x_i^t \in \mathcal{B}^T} \left(\frac{e^{(d(F(x_i^t), F(x_i^{tp}))/\tau)}}{\sum_{x_j \in \mathcal{B}^{TN} \cup \{x_i^{tp}\}} e^{(d(F(x_i^t), F(x_j))/\tau)}} \right) \\ =& \frac{1}{M} \sum_{x_i^t \in \mathcal{B}^T} \left(\frac{e^{(d(f_i^t, f_i^{tp})/\tau)}}{\sum_{x_j \in \mathcal{B}^{TN} \cup \{x_i^{tp}\}} e^{(d(f_i^t, f_j)/\tau)}} \right)\end{aligned} \tag{3}$$

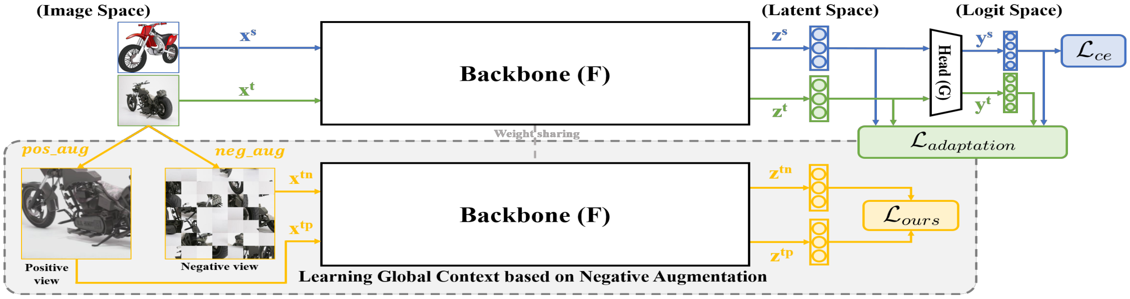


Figure 3. Overview of the Unsupervised Domain Adaptation (UDA) framework employing our Negative View-based Contrastive (NVC) regularizer loss, denoted as $\mathcal{L}_{ours} = \mathcal{L}_{\text{NVC}}$. This integrative approach can be seamlessly incorporated with any preceding UDA methodologies $\mathcal{L}_{adaptation}$, such as the SDAT framework [26], SSRT framework [30], etc. It incentivizes the underlying network, such as a Vision Transformer (ViT), to discern and leverage contextual interrelations amongst local patches, thereby advancing the UDA performance.

In this work, we introduce the Negative View-based Contrastive (NVC) loss, delineated in Equation 3, and expound upon its formulation in this subsection. Drawing inspiration from SimCLR [2], we harness the principles of contrastive learning to shape the NVC loss into a form of contrastive loss, specifically the *Normalized Temperature-scaled Cross Entropy (NT-Xent)*. Our intention is to leverage the unlabeled target samples to steer a Vision Transformer (ViT)-based classifier towards comprehending global context within the target domain. A pivotal aspect of contrastive learning involves the identification of positive and negative sample pairs relative to a given anchor sample.

A positive sample, in this framework, is one that must be brought into closer proximity to its anchor in the feature space. In contrast to the approach by Yao Qin *et al.* [24], our UDA scenario does not provide access to the target sample labels. We navigate this constraint by generating a single positive sample for each anchor through the application of a sequence of positive augmentations denoted by pos_aug. For a target-domain minibatch $\mathcal{B}^T = \{x_i^t\}_{i=1}^M \subset \mathcal{X}^T$, we formulate a positive view $x_i^{tp} = pos\_aug(x_i^t)$ for each anchor $x_i^t \in \mathcal{B}^T$, thereby establishing it as the positive counterpart. We adopt the augmentation sequence verified by SimCLR [2], which confirms the value of such a composition in contrastive learning. The sequence is accessible from the PyTorch-based SimCLR repository[1]. The pseudocode for pos_aug is provided in the supplementary material.

Conversely, a negative sample is one that should be distanced from the anchor. Again, diverging from the approach by Yao Qin *et al.* [24], the absence of label access for the target domain precludes classification of original sample classes. Nevertheless, negative augmentation can be conducted label-free, enabling the assignment of negative views $\{x_i^{tn}\}_{i=1}^M = neg\_aug(\mathcal{B}^T)$ to each sample in $\mathcal{B}^T$ as negative samples for the anchor. We select P-Shuffle for neg_aug as it is considered more disruptive due to its alteration of all local patch positions.

In essence, for an anchor sample $x_i^t \in \mathcal{B}^T$, the positive sample is acquired by $x_i^{tp} = pos\_aug(x_i^t)$, and all corresponding negative views $\mathcal{B}^{TN} = \{x_i^{tn}\}_{i=1}^M = neg\_aug(\mathcal{B}^T)$ are utilized as negative samples. The regularizer loss, calculated from the total $M+1$ pairs, propels M negative samples away from the anchor while pulling a single positive sample nearer. Through this contrastive learning approach, the ViT-based architecture is prompted to focus on global context within the target domain, thereby enhancing its robustness against out-of-distribution (OOD) samples in the target domain, ultimately reinforcing UDA performance. The regularizer is computed using the latent vector $f_i^t = F(x_i^t)$. Equation 3 presents the NVC regularizer loss $\mathcal{L}_{reg} = \alpha\mathcal{L}_{ours}$, where α and τ represent a trade-off coefficient and a temperature parameter, respectively, and $d(\cdot,\cdot)$ denotes cosine similarity. Notably, the numerator of Equation 3 exclusively includes positive pairs, whereas the denominator encompasses both positive and negative pairs.

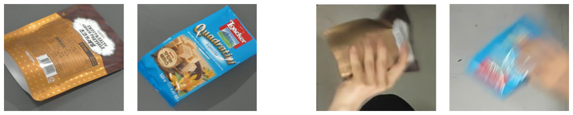
4. New Dataset: Retail-71

We introduce Retail-71, a dataset comprised of 71 frequently encountered products in convenience stores. Retail-71 is curated for object classification within the UDA paradigm. However, it diverges from existing benchmarks in notable ways:

- Retail-71 is specifically designed to focus on the domain gap introduced by hand occlusions and motion blur, providing a unique avenue to assess the robustness of neural networks against domain shifts prompted by these two factors.

- In contrast to other benchmarks, Retail-71 is accompanied by a test set stratified into three levels of difficulty—easy, medium, and hard. The degree of diffi-

[1] PyTorch-based SimCLR



(a) Exemplars from the source domain.

(b) Exemplars from the target domain.

Figure 4. Representative samples from the source and target domains.

Figure 5. Hand image examples extracted from the Ego2Hands dataset [17].

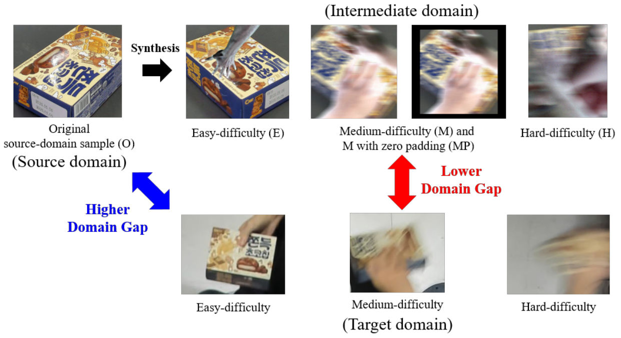


Figure 6. Rule-based synthesis examples for constructing the intermediate domains of Retail-71. The goal is to create graded steps between the Original domain (O) and the target domain, designated as E (Easy), M (Medium), MP (Medium Plus), and H (Hard). By training neural networks on combinations of O and these intermediate domains (e.g., O+E+M), the domain gap to the target can be effectively reduced.

culty is scaled by the intensity of motion blur and the prevalence of hand occlusion in the test samples.

Subsequent subsections provide an overview of Retail-71's fundamentals, while the comprehensive construction process is elucidated in the supplementary material.

4.1. Source Domain: Clean Images

The source domain encompasses clean images with each class containing 150 samples, culminating in a total of $71 \times 150 = 10,650$ images. Every image presents a singular product against a uniform gray background. Figure 4a illustrates a selection of examples from the source domain.

4.2. Target Domain: Noisy Real-World Images

In contrast to the source dataset, the target domain comprises noisy samples exhibiting motion blur, hand occlusion, slight device noise, and reduced resolution, along with backgrounds that are distinct from those of the source samples. Figure 4b displays a selection of these images. The target training set encompasses a total of 44,020 samples, allocating 620 samples per class, whereas the test set contains 10,650 samples, with 150 per class. Intrinsically, the test set is partitioned into three subsets calibrated for easy, medium, and hard difficulties, each encompassing 3,550 samples (*i.e.*, 50 samples per product). Visual exemplars of the test samples across these difficulties are provided in Figure 6. Detailed criteria for the test set's varying levels of difficulty are delineated in the supplementary material.

4.3. Construction of the Intermediate Domain for Retail-71

The domain shift between the source images, henceforth denoted as O (Original domain), and the target-domain images in Retail-71, is mainly attributed to hand occlusion and motion blur. The images from O typically exhibit lower resolution, introducing additional noise in comparison to the target domain. To facilitate a smoother transition for domain adaptation, we propose synthesizing an intermediate domain that intermediates the gap between O and the target domain. This process involves augmenting images from O with (1) hand overlays using images from the Ego2Hands dataset [17], (2) artificial motion blur, (3) noise addition, and (4) zero-padding to simulate occlusion. The intensity of these augmentations is varied to create intermediate domains with differing difficulty levels - Easy (E), Medium (M), Medium Plus (MP), and Hard (H) - each progressively closer to the target domain conditions. These intermediate domains can be utilized during the training of neural networks by combining them with O (e.g., using O+E+M in place of just O). Figure 6 shows examples of the synthesized intermediate domains alongside the target-domain images. Detailed descriptions of the synthesis process and difficulty levels are included in the supplementary materials.

5. Experiments

5.1. Datasets

We evaluate our proposed method using several benchmark datasets including Office-31 [27], Office-Home [37], VisDA-2017 [23], and our newly introduced Retail-71. Detailed descriptions of each dataset, including the number of classes, samples, and domain-specific characteristics, are provided in the supplementary materials.

5.2. Experimental Settings

In our experimental framework, we employ SDAT [26], a state-of-the-art technique, as the primary baseline to which we apply our novel NVC regularizer. For the majority of the datasets, we utilize the ViT-Base model architecture, whereas for Retail-71, we adopt the smaller variants, ViT-

Method	Backbone	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.
Source Only	ResNet50	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
DANN [6]		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
CDAN [20]		49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
CDAN+E [20]		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
CDAN+BN [26]		54.3	70.6	76.8	61.3	69.5	71.3	61.7	55.3	80.5	74.8	60.1	84.2	68.4
PCT [32]		57.1	78.3	81.4	67.6	77.0	76.5	68.0	55.0	81.3	74.7	60.0	85.3	71.8
SDAT [26]		58.2	77.1	82.2	66.3	77.6	76.8	63.3	57.0	82.2	74.9	64.7	86.0	72.2
Source Only [30]	ViT-Base	54.7	83.0	87.2	77.3	83.4	85.5	74.4	50.9	87.2	79.6	53.8	88.8	75.5
CDTrans [38]		68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
SSRT [30]		75.2	89.0	91.1	85.1	88.3	90.0	85.0	74.2	91.3	85.7	78.6	91.8	85.4
CDAN [26]		62.6	82.9	87.2	79.2	84.9	87.1	77.9	63.3	88.7	83.1	63.5	90.8	79.3
SDAT [26]		70.8	87.0	90.5	85.2	87.3	89.7	84.1	70.7	90.6	88.3	75.5	92.1	84.3
SDAT*		70.8	87.0	90.5	85.3	87.7	89.7	83.7	71.0	90.5	88.2	75.4	92.1	84.3
SDAT†		74.1	88.1	91.6	87.0	90.0	89.9	84.3	71.5	91.6	87.0	74.5	93.2	85.2
SDAT†+**Ours**		**75.1**	**89.0**	**91.5**	**86.4**	**88.6**	**90.2**	**84.8**	**73.7**	**91.7**	**87.1**	**74.6**	**92.9**	**85.5**

Table 1. Classification accuracy (%) on Office-Home dataset. The symbol * indicates reproduction of results under the original experimental conditions as detailed in [26]. 'CDAN+BN' refers to 'CDAN with Batch Normalization' as reported in the same study.

Method	Backbone	Plane	Bcycl	Bus	Car	Horse	Knife	Mcyle	Persn	Plant	Sktb	Train	Truck	Mean
Source Only [30]	ResNet101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [6]		81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
CDAN+E [26]		85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
CDAN+BN [26]		94.9	72.0	83.0	57.3	91.6	95.2	91.6	79.5	85.8	88.8	87.0	40.5	80.6
CAN [12]		97.0	87.2	82.5	74.3	97.8	96.2	90.8	80.7	96.6	96.3	87.5	59.9	87.2
Source Only	ViT-Base	99.1	60.7	70.6	82.7	96.5	73.1	97.1	19.7	64.5	94.7	97.2	15.4	72.6
CDTrans [38]		97.1	90.5	82.4	77.5	96.6	96.1	93.6	88.6	97.9	86.9	90.3	62.8	88.4
SSRT [30]		98.9	87.6	89.1	84.8	98.3	98.7	96.3	81.1	94.9	97.9	94.5	43.1	88.8
CDAN [26]		94.3	53.0	75.7	60.5	93.9	98.3	96.4	77.5	91.6	81.8	87.4	45.2	79.6
SDAT [26]		98.4	90.9	85.4	82.1	98.5	97.6	96.3	86.1	96.2	96.7	92.9	56.8	89.8
SDAT*		97.8	90.9	82.0	79.3	98.7	96.9	93.8	87.6	95.7	97.1	94.1	61.3	89.6
SDAT†		98.5	89.8	89.2	84.5	98.1	96.9	95.6	82.9	96.4	97.2	95.3	51.8	89.7
SDAT†+**Ours**		**98.5**	**89.0**	**88.5**	**92.0**	**98.5**	**98.3**	**96.2**	**88.4**	**98.5**	**97.9**	**95.0**	**55.4**	**91.4**

Table 2. Classification accuracy (%) on VisDA-2017. For the performance of 'CDAN+E', we base our comparison on the findings in [26].

Small and ViT-Tiny, considering the dataset's unique characteristics and constraints. The batch size is meticulously set to 96, ensuring each minibatch accurately represents the underlying distribution of the training dataset. This consideration is critical for maintaining the integrity of the stochastic gradient descent optimization process. A comprehensive account of the experimental procedures, including hyperparameter settings, optimization strategies, and additional implementation details, is systematically delineated in the supplementary materials.

Method	Backbone	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Source Only	ResNet50	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DANN [6]		79.7	82.0	68.2	96.9	67.4	99.1	82.2
CDAN [20]		89.8	93.1	70.1	98.2	68.0	100.0	86.6
CDAN+E [20]		92.9	94.1	71.0	98.6	69.3	100.0	87.7
PCT [32]		93.8	94.6	77.2	98.7	76.0	99.9	90.0
CAN [12]		95.0	94.5	78.0	99.1	77.0	99.8	90.6
Source Only [30]	ViT-Base	90.4	91.2	81.1	99.2	80.6	100.0	90.4
CDTrans [38]		97.0	96.7	81.1	99.0	81.9	100.0	92.6
SSRT [30]		98.6	97.7	83.5	99.2	82.2	100.0	93.5
SDAT†		98.6	98.9	84.9	99.2	84.9	100.0	94.4
SDAT†+**Ours**		**99.2**	**98.4**	**85.1**	**99.2**	**85.6**	**100.0**	**94.6**

Table 3. Classification accuracy (%) on Office-31 dataset. Entries marked with † denote results reproduced using a batch size of 96.

5.3. Comparison Results

Office-31 As illustrated in Table 3, our evaluation on Office-31 reveals that models with ViT-Base backbones generally surpass those with ResNet50 backbones across various ResNet-based methodologies. Our reproduction of SDAT, utilizing a batch size of 96, demonstrates superior performance over other UDA methods. Furthermore, the integration of the NVC regularizer enhances UDA effectiveness, notably achieving performance gains in A→D and W→A scenarios.

Office-Home Table 1 details the performance on Office-Home, underscoring that ViT-Base backbones outperform the ResNet50 backbones with all considered UDA methods. Reproducing SDAT with both the batch size of 96 and the original batch size as per [26], we find that a larger batch size more accurately represents the dataset's distribution. The NVC regularizer contributes additional performance enhancements, particularly evident in the Ar→Cl, Ar→Pr, and Pr→Cl transitions, which suggests its efficacy in mining global image semantics.

VisDA-2017 The findings for VisDA-2017 are presented in Table 2, a dataset comprising a broad range of synthetic and real imagery. Given the scale and diversity of VisDA-2017, we hypothesized that the extraction of global image semantics would be crucial. Our approach notably improves mean class accuracy and significantly outperforms the baseline SDAT in specific classes such as *car* and *person*. This underlines the impact of learning global contexts on large-

Method	Backbone	Val.	Easy Test	Medium Test	Hard Test	Avg. Test
Source Only	ViT-Small	49.0	72.4	43.5	18.7	44.9
Source Only+**RS**		55.8	77.4	52.1	27.5	52.3
DANN [6]		92.3	97.4	93.2	81.8	90.8
CDAN [20]		94.4	98.4	95.2	85.3	93.0
CDAN+E [20]		93.5	97.7	94.2	84.3	92.1
SDAT†		95.2	97.7	95.7	87.5	93.7
SDAT†+**RS**		96.0	98.6	96.6	89.4	94.9
SDAT†+**Ours**		96.0	98.6	96.5	88.3	94.4
SDAT†+**Ours+RS**		**97.0**	**99.1**	**97.5**	**90.9**	**95.9**
Source Only	ViT-Tiny	41.3	58.9	37.3	19.0	38.4
Source Only+**RS**		48.2	67.1	45.5	25.7	46.1
DANN [6]		89.1	96.1	91.3	74.4	87.3
CDAN [20]		91.3	95.5	92.1	79.9	89.2
CDAN+E [20]		91.4	95.5	92.1	79.8	89.1
SDAT†		93.4	95.8	93.7	83.2	90.9
SDAT†+**RS**		95.1	98.4	95.9	86.4	93.6
SDAT†+**Ours**		94.8	97.8	95.2	85.2	92.7
SDAT†+**Ours+RS**		**95.8**	**99.1**	**96.8**	**87.7**	**94.5**

Table 4. Classification accuracy (%) on Retail-71, including experiments with rule-based synthesis (RS).

scale datasets.

Retail-71 Results for Retail-71 are summarized in Table 4. The NVC regularizer uniformly enhances accuracy across all metrics for both ViT-Small and ViT-Tiny. Experiments incorporating rule-based synthesis (RS), which combines the source dataset with intermediate datasets, indicate performance improvements, thereby aiding smoother domain adaptation. Notably, combining the NVC regularizer with RS yields even greater advancements, suggesting their complementary nature in enhancing adaptation.

5.4. Novel Metrics: Negative View Accuracy and Negative Confidence

We present a novel evaluation framework for ViT-based models trained using baseline SDAT [26] and SDAT enhanced with our NVC regularizer, utilizing negative views generated from the original images. Two new metrics are introduced: (1) Negative Accuracy and (2) Average Negative Confidence Score.

Negative Accuracy measures the model's classification accuracy over negative views x_i^{tn}, with an ideal value being $100/C(\%)$, where C is the class count. For Average Negative Confidence Score avg_conf_{neg}, we consider a labeled target-domain test set $\mathcal{B}_{test}^{T} = \{(x_i^t, y_i^t)\}_{i=1}^{N_{test}}$ and its negative counterparts $\mathcal{B}_{test}^{TN} = neg\_aug(\mathcal{B}_{test}^{T})$. This metric is calculated using Equation 4, with $\mathbb{1}$ as the indicator function, representing confidence scores aligned with the ground-truth class, aiming ideally at $1/C$. A comprehensive discourse on avg_conf_{neg} is available in the supplementary materials.

$$avg\_conf_{neg} = \frac{1}{|\mathcal{B}_{test}^{TN}|} \sum_{x_i^{tn} \in \mathcal{B}_{test}^{TN}} \sum_{c=1}^{C} G(F(x_i^{tn}))_c \mathbb{1}(c = y_i^t) \tag{4}$$

Method	Backbone	Easy Test	Medium Test	Hard Test	Avg. Test
SDAT†	ViT-Small	94.2	87.7	73.4	85.1
SDAT†+**RS**		96.2	89.2	73.7	86.4
SDAT†+**Ours**		1.44	1.41	1.41	1.42
SDAT†+**Ours+RS**		4.99	1.47	1.41	2.62
SDAT†	ViT-Tiny	92.6	86.9	70.2	83.2
SDAT†+**RS**		97.2	90.1	75.6	87.6
SDAT†+**Ours**		**5.89**	**1.47**	**1.41**	**2.92**
SDAT†+**Ours+RS**		**12.87**	**2.34**	**1.52**	**5.58**

Table 5. Negative Accuracy (%) on Retail-71 for UDA tasks. Ideally, the Negative Accuracy equates to the chance level: $1/(\text{\# of classes}) = 1/71 \approx 1.41\%$.

Method	Backbone	Easy Test	Medium Test	Hard Test	Avg. Test
SDAT†	ViT-Small	0.9358	0.8673	0.7218	0.8416
SDAT†+**RS**		0.9546	0.8818	0.7238	0.8534
SDAT†+**Ours**		0.0144	0.0141	0.0141	0.0142
SDAT†+**Ours+RS**		0.0419	0.0147	0.0141	0.0236
SDAT†	ViT-Tiny	0.9217	0.8621	0.6929	0.8256
SDAT†+**RS**		0.9657	0.8929	0.7427	0.8671
SDAT†+**Ours**		**0.0507**	**0.0143**	**0.0141**	**0.0264**
SDAT†+**Ours+RS**		**0.1215**	**0.0194**	**0.0141**	**0.0517**

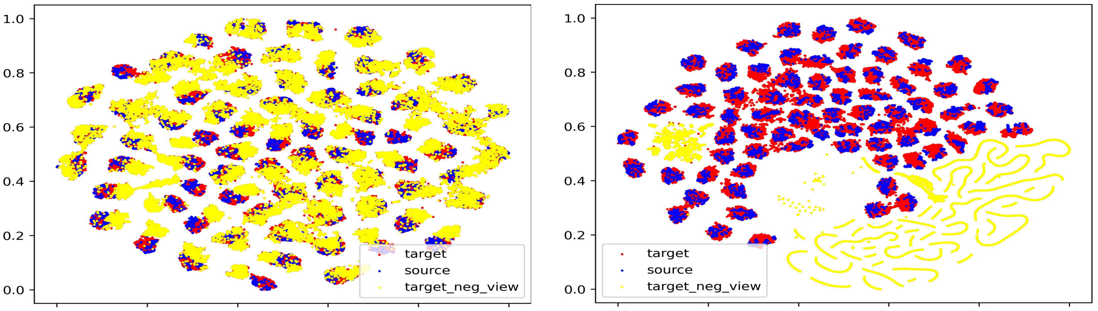
Table 6. Average Negative Confidence Score for UDA tasks on Retail-71, with the ideal value being $1/(\text{\# of classes}) \approx 0.0141$.

Table 5 delineates the Negative Accuracy on Retail-71. With 71 classes, the target value for Negative Accuracy and avg_conf_{neg} is theoretically set at 1.41(%) and 0.0141, respectively. Both with and without RS, SDAT evidences higher-than-ideal Negative Accuracy, whereas our approach achieves values approximating the theoretical ideal.

Concurrently, Table 6 corroborates the trend seen in Table 5. The results confirm that our NVC regularizer diverts the ViT's focus away from local peculiarities towards global contexts, bolstering its robustness. Additionally, the NVC-induced models exhibit human-like discernment in processing negative views, which are generally uninterpretable by humans due to their devoid semantic content. Detailed results for additional benchmarks are documented in the supplementary section.

5.5. Representation Visualization

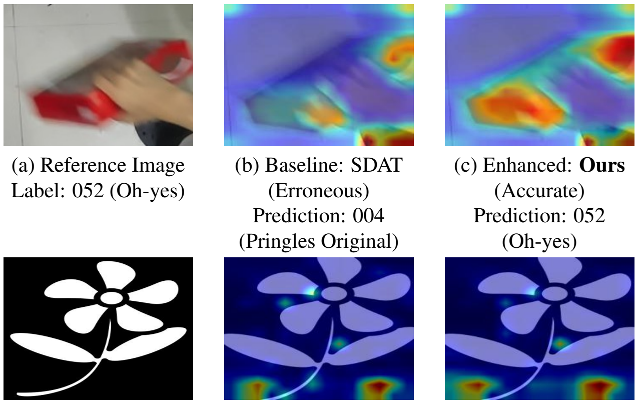
To illuminate the model's internal representation, we visualized the distribution of source-domain samples, target-domain samples, and their respective negative views. Figure 7 exhibits the t-SNE [35] projections for ViT-Tiny representations after training with baseline SDAT [26] and SDAT enhanced by our method on the Retail-71 dataset. Both the baseline and our approach demonstrate well-defined clustering to certain degrees. However, as shown in Figure 7b, our method distinctly segregates the target samples from their negative views, in contrast to the baseline (Figure 7a), which suggests that our model is adept at identifying semantically rich features and spatial correlations that are exclusive to the target samples and absent in the negative views. This capacity is attributed to the NVC regularizer's influence in guiding the network towards a global feature repre-



(a) Baseline: SDAT.

(b) Enhanced: SDAT+**Ours**.

Figure 7. t-SNE projections of ViT-Tiny features learned by SDAT (a) and enhanced by SDAT+**Ours** (b) on the Retail-71 dataset. Colors represent different domains, with negative views of target samples highlighted in yellow.



(a) Reference Image
Label: 052 (Oh-yes)

(b) Baseline: SDAT
(Erroneous)
Prediction: 004
(Pringles Original)

(c) Enhanced: **Ours**
(Accurate)
Prediction: 052
(Oh-yes)

(d) Reference Image
Label: Flowers

(e) Baseline: SDAT
(Erroneous)
Prediction: Fan

(f) Enhanced: **Ours**
(Accurate)
Prediction: Flowers

Figure 8. Attention map visualizations for ViT-Tiny with SDAT (b) and SDAT+**Ours** (c) on Retail-71, and ViT-Base with SDAT (e) and SDAT+**Ours** (f) for the Office-Home Ar→Cl scenario.

sentation, thus enhancing robustness, as discussed in Section 5.3. Additional visualizations are provided in the supplementary materials.

5.6. Attention Map Visualization

We extend our visualization efforts to the attention mechanisms within ViT trained with the baseline SDAT [26] and our NVC regularizer. Figure 8 presents these attention maps. Through these visualizations, it is evident that the NVC regularizer directs ViT's attention towards class-specific objects, diminishing undue focus on less reliable local features. Notably, in Figures 8e and 8f, while both the baseline and our model appear to focus on similar image patches, only our model correlates these patches to the correct semantic labels. This observation suggests that mere attention to objects does not equate to capturing their semantic significance. In contrast, our method integrates local and global contextual information, as evidenced by its accurate classification. Further examples of attention map visualizations are available in the supplementary documents.

6. Conclusion

Through this study, we spotlighted an inherent limitation in Vision Transformers concerning patch-based negative augmentation and introduced the Negative View Contrastive (NVC) regularizer as a remedy. This regularizer, seamlessly compatible with existing UDA frameworks, endows ViT-based models with the dual capability to discern both macro and micro-level features, enhancing robustness. The newly proposed Retail-71 dataset—emphasizing motion blur and hand occlusion—served as a testament to the NVC regularizer's efficacy in mitigating domain shift. Additionally, a novel rule-based synthesis technique for Retail-71 was suggested, proving effective in facilitating domain adaptation by generating intermediate samples. In essence, our work underscores the latent prowess of negative augmentation within UDA endeavors and sets the stage for its expanded exploration in future research.

7. Acknowledgement

This work was supported by the SHINSEGAE-KAIST AI R&D Center funded by the SHINSEGAE I&C Inc. (Project No. G01220220).

References

[1] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7190, 2022. 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[3] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021. 1

[4] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer, 2022. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[6] 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. 1, 2, 6, 7

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[9] 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. 1

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[11] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 464–480. Springer, 2020. 3

[12] 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*, pages 4893–4902, 2019. 1, 3, 6

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[15] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020. 1

[16] Junjie Li, Yixin Zhang, Zilei Wang, and Keyu Tu. Probabilistic contrastive learning for domain adaptation. *arXiv preprint arXiv:2111.06021*, 2021. 3

[17] Fanqing Lin, Brian Price, and Tony Martinez. Ego2hands: A dataset for egocentric two-hand segmentation and detection. *arXiv preprint arXiv:2011.07252*, 2020. 5

[18] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 1

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 1, 2

[20] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 1, 2, 6, 7

[21] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 1, 2

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[23] 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. 2, 5

[24] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems*, 35:16276–16289, 2022. 2, 3, 4

[25] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In *International Joint Conference on Artificial Intelligence*, 2021. 1, 3

[26] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International Conference on Machine Learning*, pages 18378–18399. PMLR, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[27] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010. 1, 5

[28] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1

[29] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015. 1

[30] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7191–7200, 2022. 1, 3, 4, 6

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015. 1

[32] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208, 2021. 1, 2, 6

[33] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 1

[34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[37] 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. 1, 5

[38] Tongkun Xu, Weihua Chen, Pichao WANG, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *International Conference on Learning Representations*, 2022. 1, 2, 6