

# Heterogeneous Domain Adaptation with Adversarial Neural Representation Learning: Experiments on E-Commerce and Cybersecurity

Mohammadreza Ebrahimi, *Member, IEEE*, Yidong Chai, Hao Helen Zhang,  
and Hsinchun Chen, *Fellow, IEEE*

**Abstract**—Learning predictive models in new domains with scarce training data is a growing challenge in modern supervised learning scenarios. This incentivizes developing domain adaptation methods that leverage the knowledge in known domains (source) and adapt to new domains (target) with a different probability distribution. This becomes more challenging when the source and target domains are in heterogeneous feature spaces, known as heterogeneous domain adaptation (HDA). While most HDA methods utilize mathematical optimization to map source and target data to a common space, they suffer from low transferability. Neural representations have proven to be more transferable; however, they are mainly designed for homogeneous environments. Drawing on the theory of domain adaptation, we propose a novel framework, Heterogeneous Adversarial Neural Domain Adaptation (HANDA), to effectively maximize the transferability in heterogeneous environments. HANDA conducts feature and distribution alignment in a unified neural network architecture and achieves domain invariance through adversarial kernel learning. Three experiments were conducted to evaluate the performance against the state-of-the-art HDA methods on major image and text e-commerce benchmarks. HANDA shows statistically significant improvement in predictive performance. The practical utility of HANDA was shown in real-world dark web online markets. HANDA is an important step towards successful domain adaptation in e-commerce applications.

**Index Terms**—Domain adaptation, adversarial kernel learning, dictionary learning, maximum mean discrepancy, transfer learning.

## 1 INTRODUCTION

LEARNING predictive models in new domains that lack enough training data has arisen as a challenge in supervised learning. This forms a strong motivation for transferring knowledge from common domains (source) to unknown domains (target), often known as Domain Adaptation (DA) [1], [2], [3]. DA requires less supervision since it often operates with few labeled samples in the target domain. A practical example would be utilizing labeled user-generated content in legal e-commerce platforms (e.g., Amazon, eBay) to recognize unseen content in a new market and improve product search and indexing [4]. The same scenario applies to dark web e-commerce platforms (e.g., Dream Market and Russian Silk Road) in cybersecurity applications.

DA is a branch of transfer learning (TL) addressing two major issues: distribution divergence and feature discrepancy [5]. The former arises because source domain samples admit a different distribution than that of the target domain. The latter occurs when source and target samples are expressed in different feature spaces. Most studies focus on the first issue, in which data distributions of source and target are different while samples are in the same feature space. Also

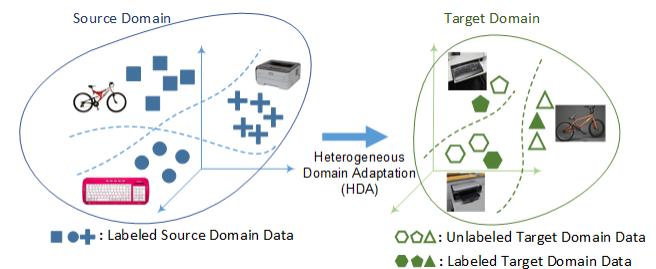


Figure 1. Illustrating HDA with Distribution Divergence and Feature Discrepancy. Three product categories are represented by different features and distributions in the source and target domains.

known as homogeneous domain adaptation, this approach is not sufficient to address real-world scenarios. Addressing both issues is significantly challenging and has emerged as a new field called Heterogeneous Domain Adaptation (HDA). In most HDA scenarios, labeled datasets are available from a known source, while the target domain suffers from a lack of labeled data (Figure 1) [6], [7].

Most HDA methods use mathematical optimization (e.g., convex optimization) or linear methods (e.g., linear discriminant analysis) to find common underlying representations from source and target [5], [8], [9]. These representations could be less transferable for DA [10]. Neural representations have shown promise in DA due to their ability to automatically extract transferable features [11]. Transferable features are intermediate representations that are domain invariant [12], [13] and, thus, instrumental to successful DA [7]. Adversarial Neural Domain Adaptation (ANDA) is a promising direction to obtain domain-invariant representations [13], which leverages a game theory-based

- M. Ebrahimi is with the School of Information Systems and Management, University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620. E-mail: ebrahimi@usf.edu.
- Y. Chai is with the School of Management, Hefei University of Technology. Correspondence to: Y. Chai. Email: chaidy@hfut.edu.cn.
- Hao Helen Zhang is with the Department of Mathematics, University of Arizona. E-mail: hzhang@math.arizona.edu
- Hsinchun Chen is with the Artificial Intelligence Lab, University of Arizona. Email: hchen@eller.arizona.edu

learning scheme [14] to obtain high-quality representations. As a subcategory of DA, existing ANDA methods often address homogeneous domain adaptation [5]. However, heterogeneous sources of user-generated content are very common on e-commerce platforms (e.g., multilingual text or product images with different representations). We present a novel ANDA framework to extract transferable domain-invariant representations from heterogeneous domains. This framework extends ANDA to heterogeneous domains by employing an inverse problem-solving method known as dictionary learning to mitigate the feature discrepancy. It also alleviates the distribution divergence between source and target in an adversarial manner. Our proposed method significantly improves the performance of text classification and image recognition in e-commerce applications.

## 2 RELATED WORK

DA is a special case of TL. The goal of TL is to use the knowledge obtained from a resource-rich domain or task (i.e., source) to help a task or domain with insufficient training data (i.e., target) [15]. The domain consists of feature space and data probability distribution [16]. The task is a function that can be learned on a domain to map the data to a corresponding label space. In many practical cases, the source and target tasks are often the same, while the domains vary. Hence, the source domain model often needs to be adapted to the target domain. For instance, in cybersecurity applications, a task could be determining if a dark net market product is cybersecurity-related when the domains are English (source) and Russian (target) product descriptions in the dark web. DA aims to transfer the knowledge to perform the same task in different source and target domains [17]. This requires reducing the distribution divergence and feature discrepancy in the source and target domains [5], [17]. Most DA methods only focus on alleviating the distribution divergence (i.e., homogeneous DA) [11], [18], [19], [20]. Homogenous DA does not apply to heterogeneous domains where the data dimensions and features are different. In contrast, Heterogeneous DA (HDA) is intended to reduce both feature discrepancy and distribution divergence.

### 2.1 Heterogeneous Domain Adaptation (HDA)

HDA aims to perform the adaptation when the source and target domains are in different feature spaces. HDA methods often map features in the source and target to a common space using mathematical objective function optimization or a linear projection from source to target domain [5], [8], [9], [21], [22], [23], [24], [25], [26], [27]. These features tend to be less transferable than representations obtained from neural networks for domain adaptation [28], [29]. A taxonomy of selected major HDA methods is provided in Table 1.

As seen in the taxonomy, most HDA approaches obtain feature representations based on mathematical optimization and linear mapping, and few studies utilize neural network-based representations to accomplish HDA. Most extant methods cast the HDA objective into an optimization problem aimed at mapping source and target to a common feature space without promoting domain invariance between

source and target, resulting in limited improvement in the transferability of the source model to the target domain. Next, we review the effective methods to attain domain invariance and position our study within the body of the related work.

### 2.2 Adversarial Neural Domain Adaptation (ANDA)

One promising direction to achieve domain invariance in DA is to map source and target samples into a ‘common latent space,’ in which source and target representations of the same class are close to each other [30], [31]. Recently, adversarial learning has shown promise in providing domain-invariant representations without ‘pair alignment,’ eliminating the need for manually establishing sample correspondence in source and target [6], [10]. Adversarial learning is a game-theoretic approach for simultaneously training two competing components, generator and discriminator [32], which yields state-of-the-art performance in homogeneous DA [11]. When utilized in DA, the generator is trained to create representations that mimic both source and target distribution, and the discriminator is tasked to distinguish the source data from the target data created by the generator. ANDA generates adversarially learned domain-invariant representations from source and target that are suitable for domain adaptation [13]. These representations are useful in tasks such as text classification [12], [33], [34], character recognition [6], [7], [10], [35], [36], or image classification [11], [13], [37], [38], [39] on a variety of datasets such as English Amazon reviews, handwritten digits, and face images. However, existing ANDA methods mostly operate on sources and targets with homogeneous feature spaces and do not address heterogeneous domain adaptation. This restricts the applicability of DA in real-world applications where the source and target domains are both relevant to a task but do not share the same feature representation.

### 2.3 Research Positioning

While ANDA can address the transferability issue, most ANDA frameworks consider homogeneous feature spaces. Given the heterogeneity of user-generated content (e.g., multilingual text, diverse image representations), these issues undermine the applicability of DA to emerging applications such as e-commerce or cybersecurity. To highlight the position of our proposed method, we categorize major DA work in two dimensions (Table 2). The horizontal dimension differentiates DA methods supporting heterogeneous features. The vertical dimension distinguishes between DA methods that generate neural representations and the others. Among homogeneous studies, methods relying on mathematical optimization include Support and Model Shift (SMS) [20], Joint Distribution Adaptation (JDA) [18], and Transfer Component Analysis (TCA) [19]. Recent homogeneous methods such as Symmetric Networks (SymNets) [39], Domain Adaptation Network (DAN) [11], Symmetric Bi-directional ADaptive GAN (SBADA-GAN) [36], Auxiliary Classifier GAN (AC-GAN) [38], Adversarial Discriminative Domain Adaptation (ADDA), and Cycle Consistent Adversarial Domain Adaptation (CYCADA) [6] support neural representations (the bottom left quadrant). Heterogeneous methods (the top right and bottom right

Table 1  
A Taxonomy of Major HDA Methods

Method	Method Category	Task	Testbed
Discriminative Joint Distribution Adaptation (DJDA) [9]	Objective function optimization	Low-resolution image recognition	Face images
HDA Network based on Autoencoder (HDANA) [29]	Stacked autoencoder	Image recognition	Newswire articles, product images
Progressive Alignment (PA) [5]	Objective function optimization	Multilingual text and image recognition	Newswire articles, product images
Cross-Domain Mapping (CDM) [8]	Linear discriminant analysis	Multilingual text and image recognition	Newswire articles, product images
Invariant Latent Space (ILS) [22]	Kernel matching and optimization	Image and face recognition	Product and face images
Cross-Domain Landmark Selection (CDLS) [23]	Objective function optimization	Image recognition	Product images
Transfer Neural Trees (TNT) [28]	Neural decision forest	Image recognition	Product images
Generalized Joint Distribution Adaptation (G-JDA) [21]	Objective function optimization	Image recognition and multilingual text classification	Newswire articles, product images
Supervised Kernel Matching (SSKM) [24]	Kernel matching and optimization	Text and sentiment classification	Product reviews, newswire articles
Sparse Heterogeneous Feature Representation (SHFR) [25]	Objective function optimization	Text and sentiment classification	Newswire articles
Semi-supervised Heterogeneous Feature Augmentation (SHFA) [26]	Convex function optimization	Text and sentiment classification, Image recognition	Product reviews, newswire articles
Max-Margin Domain Transforms (MMDT) [27]	Objective function optimization	Image recognition	Product images

quadrants) were recognized in Section 2.1. Little has been done to provide a solution that yields neural representations from heterogeneous domains (the bottom right quadrant).

Two extant neural representation-based HDA methods are recognized as alternatives for our proposed model: Transfer Neural Trees (TNT) [28] and Heterogeneous Domain Adaptation Network based on Autoencoder (HDANA) [29]. TNT is a tree-based neural network architecture focused on conducting HDA via obtaining corresponding pairs in source and target domains and does not offer distribution alignment. HDANA is a deep autoencoder architecture addressing the distribution alignment with a fixed distribution divergence kernel, which may result in lack of domain invariance. Both these issues can lead to performance loss. To remedy these issues, our model introduces a novel HDA framework that enables both feature and distribution alignment in a unified neural network architecture with enhanced domain invariance. It utilizes dictionary learning and nonparametric adversarial kernel matching to achieve these goals without relying on a fixed kernel for distribution matching. Our model contributes to obtaining high-quality representations from heterogeneous domains, which are useful in downstream tasks such as

multilingual text classification and image recognition.

### 3 BACKGROUND FOR PROPOSED MODEL

#### 3.1 Model Preliminaries

Let  $\mathcal{D}^s = \{\mathcal{X}^s, P(X^s, y^s)\}$  and  $\mathcal{D}^t = \{\mathcal{X}^t, Q(X^t, y^t)\}$  denote the source and target domains, each comprises feature spaces  $\mathcal{X}^s$  and  $\mathcal{X}^t$  and distributions  $P, Q$  over the source and target samples, respectively. The source data  $X^s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$  includes labeled data sampled from distribution  $P$ . The target data  $X^t = \{x_i^t\}_{i=1}^{n_t}$  is sampled from distribution  $Q$  and contains both labeled data  $\{(x_i^L, y_i^L)\}_{i=1}^{n_L} \sim Q$  and unlabeled data  $\{(x_i^U)\}_{i=1}^{n_U} \sim Q$ . HDA aims to help a classification task in the target domain by using the source domain data when not only is  $P \neq Q$ , but  $\mathcal{X}^s \neq \mathcal{X}^t$ . To achieve this, we present a neural network architecture to approximate a hypothesis  $h(x)$  that not only minimizes the discrepancy between  $X^s$  and  $X^t$  as well as the divergence between distributions  $P$  and  $Q$ , but also minimizes the error of assigning target labels  $\epsilon_t(h(x))$ , using source domain data and only a small number of labeled target data. DA models need to be generalizable to unseen domains in order to reduce the target generalization error  $\epsilon_t(h)$  [11]. Accordingly, we draw on the domain adaptation theory to identify the generalization error bound of the target [40], [41] and inform the components of our proposed model. For any hypothesis space  $\mathcal{H}$  and a fixed representation function the expected target error is bounded by Theorem 1.

**Theorem 1** [41]: Let  $X^s$  and  $X^t$  represent samples drawn from  $P$  and  $Q$ , the source and target domain distributions, respectively. For any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$  (over the choice of the samples), for every hypothesis  $h \in \mathcal{H}$ :

Table 2  
Positioning Our Proposed Method in DA Research

	Homogeneous	Heterogeneous
Mathematical Optimization	SMS [20], JDA [18], TCA [19]	PA [5], CDM [8], G-JDA [21], CDLS [23], SHFA [26], MMDT [27]
Neural Representation	SymNets [39], DAN [11], CYCADA [6], AC-GAN [38], SBADA-GAN [36], ADDA [13]	HDANA [29], TNT [28], Our proposed model

$$\epsilon_t(h) \leq \hat{\epsilon}_s(h) + \hat{d}_{\mathcal{H}}(X^s, X^t) + \epsilon_s(h^*) + \epsilon_t(h^*) + C \quad (1)$$

where  $\hat{\epsilon}_s(h)$  is the empirical source training error,  $\epsilon_s(h^*) + \epsilon_t(h^*)$  is the combined error of the ideal hypothesis in both domains, and  $C$  is a constant given in (2):

$$C = \frac{4}{m} \sqrt{d \log \left( \frac{2em}{d} \right)} + \log \left( \frac{4}{\delta} \right) + 4 \sqrt{\frac{d \log(2m') + \log(\frac{4}{\delta})}{m'}} \quad (2)$$

where  $e$  is the base for the natural logarithm,  $d$  is the VC-dimension of the hypothesis space,  $m$  denotes labeled sample size in source, and  $m'$  is the unlabeled sample size. As seen in (2),  $C$  only depends on the properties of the hypothesis space (i.e., VC-dimension and sample sizes), rather than the choice of samples. Assuming that the combined error of the ideal hypothesis  $\epsilon_s(h^*) + \epsilon_t(h^*)$  is small for a reasonable representation, the bound in Theorem 1 depends on the first and second terms in the RHS of (1). It is expected that an effective DA model reduces these two terms and, thus, the target generalization error  $\epsilon_t(h)$ . The first error term  $\hat{\epsilon}_s(h)$  can be reduced by learning shared representations from source and target [41], which requires aligning feature representations when the domains are heterogeneous. The second term denotes the empirical  $\mathcal{H}$ -divergence between the source and target distributions [40]:

$$\begin{aligned} \hat{d}_{\mathcal{H}}(X^s, X^t) &\triangleq 2 \left( 1 - \min_{h \in \mathcal{H}} \left\{ \frac{1}{m'} \sum_{x:h(x)=0} \mathbb{1}[x \in X^s] \right. \right. \\ &\quad \left. \left. + \frac{1}{m'} \sum_{x:h(x)=1} \mathbb{1}[x \in X^t] \right\} \right) \end{aligned} \quad (3)$$

where  $\mathbb{1}$  is an indicator function that returns 1 when its argument is true and 0 otherwise. Long et al. [11] show that the minimization term in (3) is bounded by the Maximum Mean Discrepancy (MMD) distance under kernel  $k$  [42]. Thus, (1) can be re-written as:

$$\epsilon_t(h) \leq \hat{\epsilon}_s(h) + 2(1 + D_k(X^s, X^t)) + \lambda + C \quad (4)$$

where  $D_k(\cdot)$  denotes the MMD distance in a Reproducing Kernel Hilbert Space (RKHS) with kernel  $k$  [42]. The second term in (4) can be reduced by distribution kernel matching, aiming to minimize the distance of the source and target distribution.

To minimize these two terms, next we discuss dictionary learning as a promising approach to reduce  $\hat{\epsilon}_s(h)$ , which addresses feature discrepancy by translating source and target features to a common space. Subsequently, to reduce the  $\mathcal{H}$ -divergence between the source and target, we discuss distribution kernel matching and adversarial kernel learning.

### 3.2 Dictionary Learning: Reducing $\hat{\epsilon}_s(h)$

A viable approach to align representations can be devised based on dictionary learning [5]. Dictionary learning is an inverse problem that aims to decompose a given data

matrix  $X$  into a dictionary space  $D$  and a coefficient matrix  $R$ , which denotes the representation of  $X$  under  $D$ , by optimizing (5):

$$\begin{aligned} \min_{D, R} \|X - DR\|_F^2 + \lambda \|R\|_1; \\ \text{s.t. } \|d_i\|_2 \leq 1; \forall i \in 1, \dots, k \end{aligned} \quad (5)$$

where  $F$  denotes the Frobenius norm measuring reconstruction error and  $d_i$  is the  $i^{th}$  column in  $D$ . Unlike other matrix factorization methods dictionary learning can promote sparsity in the coefficient matrix, resulting in extracting only salient features from the input. Dictionary learning has been successfully applied to signal [43] and image [5] processing, contributing to reducing the dimensionality of the input data while preserving the salient features. Sharing the dictionary  $D$  between the source ( $X^s$ ) and the target domain ( $X^t$ ) forces the new representations to be projected to ] space [5].

$$\begin{aligned} \min_{D, R^s, R^t} \|X^s - DR^s\|_F^2 + \|X^t - DR^t\|_F^2 + \lambda \|R\|_1 \\ \text{s.t. } \|d_i\|_2 \leq 1; \quad R = [R^s, R^t] \end{aligned} \quad (6)$$

While this is useful in DA for aligning features, traditional dictionary learning ignores label information since it is not naturally designed for classification tasks often encountered in e-commerce applications. We are motivated to incorporate the dictionary learning objective function in a neural network architecture that can jointly learn a classification task such that it is possible to take advantage of labeled data in both the source and target domains.

### 3.3 Distribution Kernel Matching and Adversarial Kernel Learning: Reducing $\hat{d}_{\mathcal{H}}(X^s, X^t)$

DA requires measuring the distribution divergence between source and target. This could be successfully achieved by defining a two-sample test statistic on source and target [44]. Two types of test statistics for DA are parametric and nonparametric. Parametric statistics measure the similarity of the density functions in source and target (e.g., Kullback-Leibler divergence) [45], [46]. Nonparametric statistics measure the distance of kernels rather than the actual densities. This requires the distributions to be treated as functions in RKHS [42]. Nonparametric distribution kernel matching methods are often preferred for DA since they circumvent intractable density estimation [11]. Among nonparametric kernel matching methods, Maximum Mean Discrepancy (MMD) is the most widely used for DA [29], [47]. MMD measures the divergence based on the distance  $D$  between source and target domain data  $X^s \sim P$  and  $X^t \sim Q$ .

$$D_k(X^s, X^t) \triangleq \|\mathbb{E}_{x^s}[\phi_k(x^s)] - \mathbb{E}_{x^t}[\phi_k(x^t)]\|_{\mathcal{H}_k}^2 \quad (7)$$

where  $\phi_k(\cdot)$  is a feature mapping from inputs to an embedding in the Hilbert space  $\mathcal{H}_k$  characterized by kernel  $k$ . MMD's test power is extremely sensitive to the choice of mapping  $\phi_k$  [48], [49], which could negatively affect the distribution alignment performance. Adversarial kernel learning can address this issue by facilitating finding the proper kernel that maximizes the test power of MMD [44].

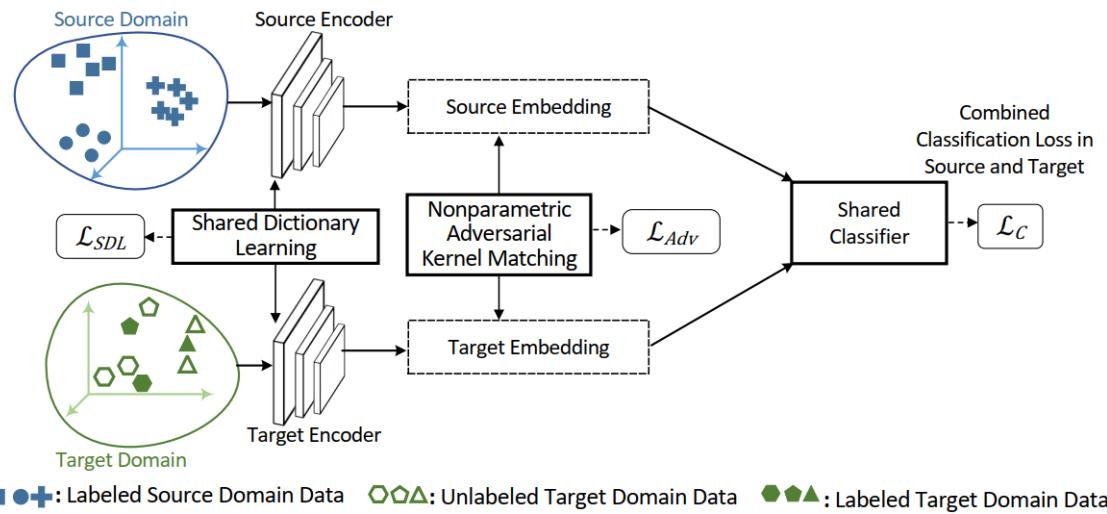


Figure 2. HANNA Architecture. The Shared Dictionary Learning (left) and Nonparametric Adversarial Kernel Matching (middle) components operate on the heterogeneous representations of source and target domain to reduce feature discrepancy via minimizing  $\mathcal{L}_{SDL}$  and to reduce distribution divergence via minimizing  $\mathcal{L}_{Adv}$ , simultaneously. The Shared Classifier (right) component operates on the resultant representations via minimizing the shared classification loss  $\mathcal{L}_C$ .

Such a kernel helps align the distributions in source and target without establishing sample correspondence. To this end, the mapping  $\phi$  in MMD can be optimized via adversarial kernel learning [44]. However, extant adversarial kernel learning methods are not designed for heterogeneous domains. Inspired by the recent advances, we employ adversarial kernel learning for HDA. This enables benefiting from the nonparametric property of MMD while improving the test power via adversarial learning.

#### 4 PROPOSED MODEL: HETEROGENOUS ADVERSARIAL NEURAL DOMAIN ADAPTATION (HANDA)

Informed by the effect of  $\hat{\epsilon}_s(h)$  and  $\hat{d}_H(\tilde{X}^s, \tilde{X}^t)$  on reducing  $\epsilon_t(h)$ , we incorporate dictionary learning and adversarial kernel matching in a unified architecture that reduces the generalization error bound on any given domain adaptation problem. Accordingly, our proposed model integrates three components, each targeting an adaptation challenge in HDA. Figure 2 shows the integration of these three components in our model, Heterogeneous Adversarial Neural Domain Adaptation (HANDA). The first component is a shared dictionary learning (SDL) approach that targets alleviating feature discrepancy via projecting heterogeneous features into a common latent space and is associated with minimizing a reconstruction loss  $\mathcal{L}_{SDL}$ . The second component is a nonparametric adversarial kernel matching method, which aims to reduce the distribution divergence by combining the nonparametric benefits of MMD and flexibility of kernel choices in adversarial learning. This component is associated with minimizing an adversarial loss denoted by  $\mathcal{L}_{Adv}$ . Finally, the third HANDA's component is a shared classifier in source and target that targets limited labeled data in the target domain by exploiting labeled samples in both the source and target domain. This component yields the classifier  $h(x)$  as the ultimate goal of HDA and is associated with minimizing the classification loss  $\mathcal{L}_C$ .

While shared dictionary learning aligns heterogeneous features in a common latent space and yields a new embedding, nonparametric adversarial kernel matching reduces

the distribution divergence in the obtained embedding. After feature and distribution alignment, a shared classifier conducts the downstream tasks (e.g., text or image classification). HANDA's novelty is threefold. First, it extends ANDA to heterogeneous domains. Second, it enhances dictionary learning to take advantage of labeled data in the source and target, and third, it enables joint feature and distribution alignment in a unified neural network architecture. We describe each HANDA component next.

##### 4.1 Component #1: Shared Dictionary Learning (SDL)

To minimize feature discrepancy, the source and target need to be mapped into the same subspace by domain-specific projections  $P^s$  and  $P^t$ . This can be achieved by modifying (6) to get:

$$\begin{aligned} \min_{P^s, P^t, D, R^s, R^t} & \|P^s X^s - DR^s\|_F^2 + \|P^t X^t - DR^t\|_F^2 \\ & + \lambda \|R\|_1; \\ \text{s.t. } & \|d_i\|_2 \leq 1; \quad R = [R^s, R^t] \end{aligned} \quad (8)$$

where  $P^s_{k \times m_s}$  and  $P^t_{k \times m_t}$  are projections.  $X^s_{m_s \times n_s}$  and  $X^t_{m_t \times n_t}$  are the source and target data.  $D_{k \times k}$  is the shared dictionary, and  $R^s_{k \times n_s}$  and  $R^t_{k \times n_t}$  denote the aligned representations. To avoid degenerate solutions,  $P^s$  and  $P^t$  are often constrained to be orthogonal [50]. The minimization in (7) yields data representations  $R^s$  and  $R^t$  that are in the same feature space.

##### 4.2 Component #2: Nonparametric Adversarial Kernel Matching

Even though the obtained representations  $R^s$  and  $R^t$  are in the same space, their distributions are different. Thus, minimizing distribution divergence is crucial for successful DA. MMD measures the distribution divergence of source and target via a ‘two-sample test,’ in which the source and target distributions are the same under the null hypothesis. The distribution divergence can be minimized by learning

a model  $\phi_\theta(\cdot)$  parameterized by  $\theta$  that minimizes the MMD distance:

$$\min_{\theta} D_k(\phi_k(\phi_\theta(R^s)), \phi_k(\phi_\theta(R^t))) \quad (9)$$

where  $D_k(\cdot)$  is the MMD distance with a mapping function of some kernel  $k$ . The explicit computation can be avoided by applying the kernel trick  $k(x, x') = \langle \phi(x), \phi(x') \rangle_H$  [48], which gives:

$$D_k(X^s, X^t) \triangleq \mathbb{E}_{x^s, x'^s} [k(x^s, x'^s)] - 2\mathbb{E}_{x^s, x^t} [k(x^s, x^t)] + \mathbb{E}_{x^t, x'^t} [k(x^t, x'^t)] \quad (10)$$

The test power of  $D_k(\cdot)$  in (9) heavily relies on the choice of the kernel  $k$  [44]. Low test power can preclude correctly distinguishing unlike source and target domains. To increase the test power in DA, it is natural to search for the kernel that maximizes the distance between source and target. Thus, (9) can be re-written as:

$$\min_{\theta} \max_k D_k(\phi_k(\phi_\theta(R^s)), \phi_k(\phi_\theta(R^t))) \quad (11)$$

To cover a wide range of kernels (as opposed to a fixed one), we parameterize the kernel using neural network  $M$  and search for the optimal kernel by learning the network parameters:

$$\min_{\theta} \max_{\theta_M} D_{\theta_M}(\phi_{\theta_M}(\phi_\theta(R^s)), \phi_{\theta_M}(\phi_\theta(R^t))) \quad (12)$$

Inspired by the idea of adversarial kernel learning [44], we suggest minimizing (12) via another neural network  $N$  in a GAN-like fashion [14]. These two neural networks can be trained in an adversarial manner to solve the minimax problem in (13).

$$\min_{\theta_N} \max_{\theta_M} D_{\theta_M}(\phi_{\theta_M}(\phi_{\theta_N}(R^s)), \phi_{\theta_M}(\phi_{\theta_N}(R^t))) \quad (13)$$

Figure 3 illustrates the abstract view of the adversarial game defined in (13).

Transformed features obtained from dictionary learning in the source and target domains participate in an adversarial setup in which they are aligned via network  $N$  to minimize the MMD distance between the domains, while network  $M$  aims to maximize MMD, searching for the kernel with maximal test power.

### 4.3 Component #3: Shared Classifier for Source and Target

With the aligned features and distributions, a shared neural network  $h_C$  can be applied to output the class probability of each sample in the source and target, respectively. Given the suitability of hinge loss [51] in multi-class problems, the source and target classification losses are defined as  $\mathcal{L}_s = L(h_C(\phi_{\theta_N}(R^s)), Y^s)$  and  $\mathcal{L}_t = L(h_C(\phi_{\theta_N}(R^t)), Y^t)$ , where  $L(\cdot)$  denotes the multi-class hinge loss.

We obtain the objective function for the shared classification hypothesis  $h_C$  by minimizing a convex combination of  $\mathcal{L}_s$  and  $\mathcal{L}_t$  [11], [40], as given in (14):

$$\min_{\theta_C} \frac{1}{2}(L(h_C(\phi_{\theta_N}(R^s)), Y^s) + L(h_C(\phi_{\theta_N}(R^t)), Y^t)) \quad (14)$$

Finally, the total objective function of HANDA is obtained by aggregating the three losses in an adversarial manner: shared dictionary loss ( $\mathcal{L}_{SDL}$ ) obtained from (8), MMD distance loss calculated via adversarial kernel matching ( $\mathcal{L}_{Adv}$ ) from (13), and combined classification losses in source and target ( $\mathcal{L}_C$ ) from (14). Thus, the total objective function for our proposed method is given by (15):

$$\min_{\theta_N, \theta_C, \Theta_{SDL}} \max_{\theta_M} \beta \mathcal{L}_{SDL} + \gamma \mathcal{L}_{Adv} + \mathcal{L}_C \quad (15)$$

where  $\Theta_{SDL} = \{P^s, P^t, D, R^s, R^t\}$  is the set of parameters for dictionary learning.  $\theta_N$  and  $\theta_M$  denote the parameters of adversarial networks for kernel matching, and  $\theta_C$  is the classifier network's parameters.  $\beta$  and  $\gamma$  are the trade-off parameters for the penalty incurred by corresponding losses. In Section 5, we demonstrate that these trade-off parameters can be empirically tuned via grid search.

### 4.4 Learning Algorithm

Past research has attempted to design DA algorithms that reduce target generalization error. Li et al. [5] offer a representation learning algorithm based on objective function optimization with a ‘fixed’ MMD kernel without kernel learning. Long et al. [11] propose a deep representation learning algorithm with adversarial kernel learning. However, their algorithm is designed for homogeneous DA. As shown in model preliminaries, SDL and adversarial kernel learning can reduce HDA target generalization errors. Hence, we are motivated to incorporate SDL and nonparametric adversarial kernel matching in our algorithm to reduce the target error more effectively. We develop an algorithm to solve the optimization problems for SDL in (8), nonparametric adversarial kernel matching in (13), and classification in (14) via a unified architecture. To achieve this, we incorporate SDL into our model. In the SDL formulation in (8), if  $D$  is a low-rank matrix, the  $\ell_1$  regularization term can be removed without lack of generalization:

$$\begin{aligned} & \min_{P^s, P^t, D, R^s, R^t} \|P^s X^s - DR^s\|_F^2 + \|P^t X^t - DR^t\|_F^2 \\ & \text{s.t. } \|d_i\|_2 \leq 1; \quad P^s P^{sT} = I; P^t P^{tT} = I \end{aligned} \quad (16)$$

where  $I$  denotes the identity matrix and the last two constraints ensure orthogonality. Ideally, the reconstruction errors are reasonably close to 0, and the source and target representations are expressed as  $\hat{R}^s = D^{-1}P^s X^s$  and  $\hat{R}^t = D^{-1}P^t X^t$ . Accordingly, we express  $R^s$  and  $R^t$  in the form of products of three matrices where the first matrix is shared. Specifically, we denote  $R^s \cong AB^s X^s$  and  $R^t \cong AB^t X^t$ , where  $B^s$  and  $B^t$  approximate the domain-specific projections and  $A$  approximates the shared projection. To approximate  $R^s$  and  $R^t$ , we parameterize  $D^{-1}$ ,  $P^s$ , and  $P^t$  by fully connected layers  $A$ ,  $B^s$ , and  $B^t$ , respectively.

Such approximation enables us to integrate shared dictionary learning into our neural network architecture. As a result, (16) can be written as:

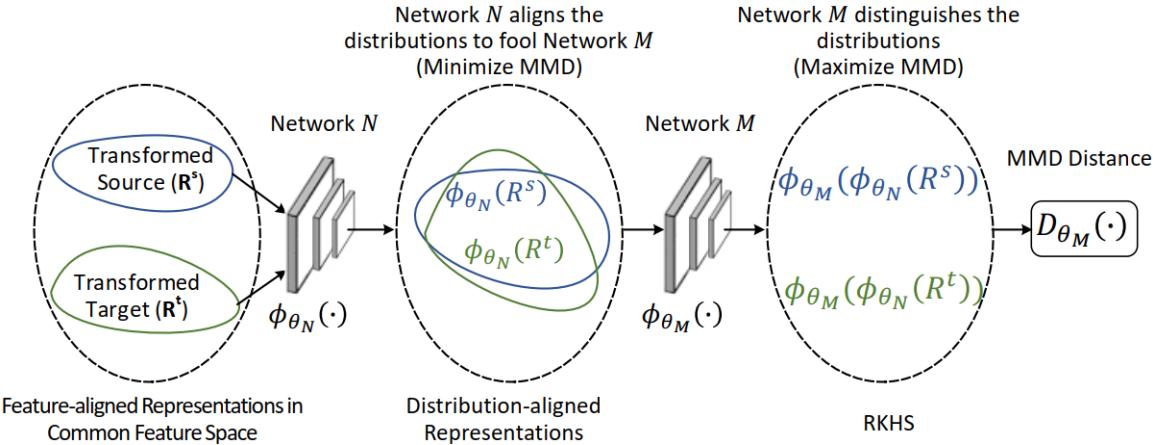


Figure 3. Abstract View of Adversarial Kernel Matching. Network  $N$  aligns the distributions to fool Network  $M$  via minimizing MMD between the source and target representations while network  $M$  aims to yield the kernel with maximal test power between two distributions via maximizing MMD.

$$\begin{aligned} \min_{P^s, P^t, D, A, B^s, B^t} & \|P^s X^s - DAB^s X^s\|_F^2 + \|P^t X^t - DAB^t X^t\|_F^2 \\ \text{s.t. } & \|d_i\|_2 \leq 1; \quad P^s P^{sT} = I; \quad P^t P^{tT} = I \end{aligned} \quad (17)$$

To satisfy the orthogonality constraints on  $P^s$  and  $P^t$  in (17), we enforce the weights of  $B^s$  and  $B^t$  to be orthogonal after each gradient update. We also normalize the weights of  $D$  and  $A$  to the unit vector. Since  $R^s$  and  $R^t$  are restricted to be a function of  $X^s$  and  $X^t$ , the result of the optimization in (17) is an upper bound for shared dictionary learning loss:

$$\mathcal{L}_{SDL^{up}} = \|P^s X^s - DAB^s X^s\|_F^2 + \|P^t X^t - DAB^t X^t\|_F^2 \quad (18)$$

Given the approximated representations of  $R^s$  and  $R^t$ , the other loss functions in objective (15) are re-written as in (19) and (20):

$$\mathcal{L}_{Adv} = D_{\theta_M}(\phi_M(\phi_N(AB^s X^s)), \phi_M(\phi_N(AB^t X^t))) \quad (19)$$

$$\mathcal{L}_C = \frac{1}{2}(L(h_C(\phi_{\theta_N}(AB^s X^s)), Y^s) + L(h_C(\phi_{\theta_N}(AB^t X^t)), Y^t)) \quad (20)$$

All three losses can be minimized simultaneously via stochastic gradient descent [44] in a unified neural network architecture introduced by HANNA. Algorithm 1 summarizes HANNA's learning procedure to minimize these losses simultaneously. Algorithm 1 alternates between minimizing the feature discrepancy and minimizing the distribution divergence until some convergence criterion is met or for a number of iterations. Code and datasets are available at <https://github.com/mohammadrezaebrahimi/HANNA>. Our algorithm differs from the heterogeneous method proposed in [5] in two aspects. First, we incorporate SDL in a unified neural network architecture that allows utilizing the supervised loss to improve the quality of dictionary learning. The dictionary learning in [5] is unsupervised and does not account for the class variable in feature alignment, which could result in lack of discriminative properties in the projected space. Second, our proposed neural network architecture offers an effective alternative for the mathematical optimization based on Lagrange multipliers in [5]

#### Algorithm 1 HANNA Learning Procedure

```

1: Input: trade-off parameters  $\gamma$  and  $\beta$ ; batch size for source data, target labeled data, and target unlabeled data:  $b_s$ ,  $b_L$ , and  $b_U$ ; # of updates of dictionary learning  $n_d$ ; # of updates of adversarial kernel matching  $n_a$ ; # of updates of classification hinge loss  $n_c$ .
2: Output: Parameters  $A, B^s, B^t, D, P^s, P^t, \theta_N, \theta_M, \theta_C$ .
3: Initialize parameters  $A, B^s, B^t, D, P^s, P^t, \theta_N, \theta_M, \theta_C$ .
4: while  $\mathcal{L}_{SDL^{up}}, \mathcal{L}_{Adv}$ , and  $\mathcal{L}_C$  have not converged do
5:   Randomly sample a minibatch of source data  $\{(x_i^s, y_i^s)\}_{i=1}^{b_s}$ .
6:   Randomly sample a minibatch of target data  $\{(x_i^L, y_i^L)\}_{i=1}^{b_L}$  and  $\{(x_i^U)\}_{i=1}^{b_U}$ .
7:   Compute  $R^s = AB^s X^s$  and  $R^t = AB^t X^t$ .
8:   for  $t = 1, 2, \dots, n_d$  do
9:     Compute shared dictionary learning loss  $\mathcal{L}_{SDL}$  using (18).
10:    Update  $A, B^s, B^t, D, P^s, P^t$  using stochastic gradient descent to minimize  $\mathcal{L}_{SDL}$ .
11:    \\\Normalize weights to unit norm
12:     $D \leftarrow \text{clip}(D, 1)$ ,  $A \leftarrow \text{clip}(A, 1)$ 
13:    \\\Enforce orthogonality constraint
14:     $P^s, P^t, B^s, B^t \leftarrow \text{Orth}(P^s, P^t, B^s, B^t)$ 
15:   end for
16:   for  $t = 1, 2, \dots, n_a$  do
17:     Compute adversarial loss for distribution alignment  $\mathcal{L}_{Adv}$  using (19).
18:     Update parameters  $\theta_M$  using stochastic gradient descent to maximize  $\mathcal{L}_{Adv}$ .
19:     Update parameters  $\theta_N, A, B^s, B^t$  via stochastic gradient descent to minimize  $\mathcal{L}_{Adv}$ .
20:   end for
21:   for  $t = 1, 2, \dots, n_c$  do
22:     Compute classification loss  $\mathcal{L}_C$  using (20).
23:     Update  $\theta_C, \theta_N, A, B^s, B^t, D, P^s, P^t$  using stochastic gradient descent to minimize  $\mathcal{L}_C$ .
24:   end for
25: end while
26: return  $A, B^s, B^t, D, P^s, P^t, \theta_N, \theta_M, \theta_C$ .
```

via a numerical optimization (stochastic gradient-based) method that reduces distribution and feature mismatch simultaneously. Later in the evaluation section we show the

performance benefit gained from this property. Our method also differs from the homogeneous method offered in [13] in that HANDA does not use GAN. While GAN implements a minimax setting between a generator and discriminator, the minimax optimization in our approach is formulated for adversarial kernel matching between source and target distributions, as given in (19), to achieve domain invariance.

## 5 EVALUATION

For a comprehensive evaluation, we measure both HANDA's performance and its practical utility. We construct a large testbed with three e-commerce datasets. To evaluate the performance, we use two widely used e-commerce benchmarks for image and multilingual text: Reuters Multilingual Collection and Office31 – Caltech256. To evaluate the practical utility, we use a real-world e-commerce dataset from dark web online markets with multilingual product descriptions. This practical evaluation encompasses a real-world cybersecurity case study, which is of value for cybersecurity applications. Reuters Multilingual Collection contains 111,740 online newswire articles with commercial themes in five languages: English, French, German, Italian, and Spanish, across six topics: Economics, Equity Markets, Finance, Industry, Social, and Performance [52]. Consistent with other studies [8], [23], we used Spanish as the target language and the other four languages as sources. This dataset is used for performing multi-class text classification task. Office31 – Caltech256 is an extension of the original dataset created by Gong et al. [53], and contains two popular heterogeneous image representations: Speeded Up Robust Features (SURF) with 800 dimensions [8] and Deep Convolutional Activation Features (DeCAF) with 4,096 dimensions [54]. The dataset consists of 2,533 online images in 10 categories in four domains: Amazon (A), downloaded from the Amazon website; Webcam (W), taken by a low-resolution camera; DSLR (D), taken by a high-resolution DSLR camera; and Caltech(C), selected from the Caltech-256 dataset. Following other studies [8], [23], three labeled samples from each category within the source domain were used in training. This dataset is associated with multi-class product image recognition task. In addition to widely accepted benchmark e-commerce datasets, it is useful to evaluate the practical utility of our model in a real-world setting. To this end, given the recent rise of illegal online markets on the dark web, we developed crawlers for obtaining data from nine illegal e-commerce platforms. Our new dark web online markets dataset is an extension of the dataset proposed in [55], which contains 94,175 product descriptions (from seven English markets), 4,600 Russian, and 1,512 French product descriptions (Table 3). The dataset has been labeled by three native speakers and two cybersecurity experts and is publicly available from our GitHub repository. To construct the training set, we randomly sampled product descriptions in each language (2,584 products; 706 cyber threats, 1,878 benign products). Since more English labeled data is available, English markets are used as the source, and Russian and French markets are used as the targets. Each language was manually labeled by a cybersecurity expert and a native speaker. This dataset is used for cyber threat detection, which is viewed as a

Table 3  
Testbed for Cybersecurity Case Study

Online Market	# of listings	Language	# of labeled products
Dream Market	39,473	English	1,821
AlphaBay	25,118		
Hansa	14,149		
Silk Road 3	1,683		
Minerva	683		
Apple Market	877		
Valhalla	12,192		
Russian Silk Road	4,600	Russian	552
French Deep Web	1,512	French	211
<b>Total:</b>	<b>100,287</b>	-	<b>2,584</b>

binary classification task. The current body of HDA research widely uses average accuracy and standard deviation for performance evaluation [5], [9], [29]. The average accuracy  $Acc_{avg}$  is defined as follows:

$$Acc_{avg} = \frac{1}{n} \sum_{i=1}^n \frac{|x^t : x^t \in X^t, \hat{y}^t = y^t|}{|x^t : x^t \in X^t|} \quad (21)$$

where  $x$  is a target instance,  $\hat{y}^t$  is the predicted class,  $y^t$  is the true class label,  $|.|$  denotes the cardinality of the corresponding set, and  $n$  is the number of runs. Standard deviation is also averaged over multiple runs of each model (often 10 times). Consistent with past HDA research, we use average accuracy and average standard deviation as evaluation metrics for comparing HANDA with the benchmark methods. Higher average accuracy and lower standard deviation suggest higher predictive performance. Note that since both benchmark datasets are not class imbalanced, accuracy is a sufficient measure for benchmark evaluations. However, for class imbalanced cybersecurity datasets, the Area Under receiver operating characteristic Curve (AUC) is the recommended performance measure since it emphasizes both threat and non-threat classes [56] by establishing a trade-off between Type I and Type II errors [57]. Given the class imbalance in our case study, we use AUC as the performance metric. Higher AUC suggests higher predictive performance.

We evaluate HANDA through three sets of experiments and one case study. Experiment #1 is aimed at evaluating HANDA against alternative state-of-the-art HDA methods (the top right and bottom right quadrants, shown in Table 2). This experiment compares HANDA's performance to seven HDA methods, including HDA Network based on Autoencoder (HDANA) [29], Cross-Domain Mapping (CDM) [8], Generalized Joint Distribution Adaptation (G-JDA) [21], Transfer Neural Trees (TNT) [28], Cross-Domain Landmark Selection (CDLS) [23], Semi-supervised Heterogeneous Feature Augmentation (SHFA) [26], and Max-Margin Domain Transforms (MMDT) [27]. This experiment involves two reputable e-commerce datasets: Office31 – Caltech256 and Reuters Multilingual Collection. Experiment #2 focuses on the qualitative evaluation of DA [28] to verify its effectiveness by comparing the linear separability of target samples before and after DA through visualizing the intermediate representations on both Office31 – Caltech256 and Reuters Multilingual Collection. Experiment #3 targets convergence analysis [58] to investigate the convergence property of the

**Table 4**  
Comparison of State-of-the-Art HDA Methods with HANDA on Multilingual Text Classification (Accuracy and Standard Deviation); P-values significant at 0.05:<sup>\*</sup>

HDA Method	10 Labeled Samples per Target Class				20 Labeled Samples per Target Class			
	<i>EN → SP</i>	<i>FR → SP</i>	<i>GR → SP</i>	<i>IT → SP</i>	<i>EN → SP</i>	<i>FR → SP</i>	<i>GR → SP</i>	<i>IT → SP</i>
MMDT	68.9 ± 0.6	69.1 ± 0.7	68.3 ± 0.6	67.5 ± 0.5	74.5 ± 0.7	74.9 ± 0.7	75.1 ± 0.8	74.9 ± 0.8
CDM	-	-	-	-	75.1 ± 2.8	75.2 ± 2.8	75.3 ± 2.7	75.2 ± 2.7
SHFA	67.8 ± 0.7	68.1 ± 0.6	68.4 ± 0.7	68.0 ± 0.7	74.1 ± 0.4	74.5 ± 0.6	74.7 ± 0.5	74.5 ± 0.5
G-JDA	69.4 ± 0.8	70.5 ± 0.7	69.6 ± 1.0	70.1 ± 1.0	76.0 ± 0.7	76.8 ± 0.8	76.8 ± 0.7	76.6 ± 0.7
TNT	70.0 ± 1.0	70.9 ± 0.8	69.9 ± 0.9	71.0 ± 0.9	76.1 ± 0.7	77.5 ± 0.8	77.1 ± 1.1	76.8 ± 0.9
CDLS	70.8 ± 0.7	71.2 ± 0.8	71.0 ± 0.8	71.7 ± 0.8	77.4 ± 0.7	77.6 ± 0.7	78 ± 0.6	77.5 ± 0.7
HDANA	72.1 ± 0.3	72.7 ± 0.4	71.8 ± 0.4	72.3 ± 0.3	78.3 ± 0.5	79.2 ± 0.5	80.1 ± 0.6	78.5 ± 0.3
HANDA (No SDL)	23.4 ± 8.2	16.8 ± 9.0	21.3 ± 9.8	18.0 ± 6.7	29.1 ± 12.4	34.0 ± 8.9	39.5 ± 8.8	32.0 ± 10.9
HANDA (No Adv)	68.0 ± 0.6	69.1 ± 0.6	67.9 ± 0.8	69.8 ± 0.7	75.7 ± 0.5	75.4 ± 0.5	75.0 ± 0.6	76.7 ± 0.6
HANDA (Entropy)	73.7 ± 0.3	72.6 ± 0.5	73.4 ± 0.2	72.4 ± 0.5	79.0 ± 0.4	79.1 ± 0.3	79.5 ± 0.4	78.1 ± 0.2
<b>HANDA (ours)</b>	<b>74.5 ± 0.4*</b>	<b>74.2 ± 0.4*</b>	<b>73.8 ± 0.5*</b>	<b>74.0 ± 0.6*</b>	<b>79.8 ± 0.3*</b>	<b>80.0 ± 0.4*</b>	<b>80.5 ± 0.2*</b>	<b>79.2 ± 0.5*</b>

objective functions corresponding to each HANDA component by assessing the stabilization of loss values during model training on the Office31 – Caltech256 benchmark dataset. Finally, a case study is conducted to verify practical utility on real-world products advertised in dark web online markets as an emerging cybersecurity application.

## 5.1 Experiment #1: Performance Evaluation

Experiment #1 aims to evaluate HANDA's performance against the state-of-the-art HDA methods in two relevant applications in international e-commerce platforms: multilingual text classification and cross-domain product image recognition. For multilingual text classification, following the settings in [29], we summarize the results of domain adaptation when Spanish is the target and English, French, German, and Italian are the source languages, with 10 and 20 labeled target samples per class (Table 4). The average accuracy is obtained by running each experiment 10 times and precedes the average standard deviation in Table 4 (separated by  $\pm$ ). To see the effect of shared dictionary learning and adversarial kernel matching in isolation, we also conducted two baseline experiments with eliminating the shared dictionary learning component (No SDL) and eliminating the adversarial kernel matching component (No Adv). We further compared the performance of the proposed adversarial kernel matching alignment loss in HANDA with the entropy-based alignment loss proposed in [59], denoted by HANDA (Entropy) in Table 4.  $\beta$  and  $\gamma$  were obtained via a small grid search as described in Experiment #1.1. The higher performance of the proposed method is statistically significant when compared to the second-best method, as suggested by the paired *t*-test [60]. HANDA improves the classification performance by approximately 2% across different source-target language pairs with statistically significant margins. Additionally, the baseline experiments show that both shared dictionary learning and adversarial kernel matching contribute to achieving the state-of-the-art performance. For cross-domain product image recognition, consistent with [29], due to the lack of labeled data in the DSLR domain, we choose the Amazon

(A), Webcam(W), and Caltech(C) domains as the source, and DSLR (D) as the target domain. HANDA improves the average accuracy in all domains by 1.1% on average (Table 5). Overall, HANDA outperforms mathematical optimization-

**Table 5**  
Cross-Features and Cross-Domain Comparison of State-of-the-Art HDA Methods with HANDA on Product Image Recognition; P-values significant at 0.05:<sup>\*</sup>

HDA Method	SURF to DeCAF			
	<i>A → D</i>	<i>C → D</i>	<i>W → D</i>	<i>Average</i>
MMDT	90.5 ± 0.6	91.2 ± 0.6	90.8 ± 0.6	90.8 ± 0.6
TNT	90.6 ± 0.9	92.7 ± 1.1	95.5 ± 1.0	92.9 ± 1.0
SHFA	93.4 ± 1.1	93.8 ± 1.0	92.4 ± 0.9	93.2 ± 1.0
G-JDA	94.3 ± 0.7	92.8 ± 0.8	95.0 ± 0.4	94.0 ± 0.6
CDLS	96.1 ± 0.7	94.9 ± 1.5	95.1 ± 0.8	95.4 ± 1.5
HDANA	96.1 ± 0.5	95.3 ± 0.6	96.9 ± 0.3	96.1 ± 0.5
HANDA (No SDL)	20.6 ± 4.5	24.2 ± 3.8	26.1 ± 5.1	23.6 ± 4.5
HANDA (No Adv)	92.8 ± 0.5	93.2 ± 0.6	93.3 ± 0.5	93.1 ± 0.5
HANDA (Entropy)	96.5 ± 0.6	95.8 ± 0.4	96.6 ± 0.5	96.3 ± 0.5
<b>HANDA (ours)</b>	<b>97.1 ± 0.4*</b>	<b>96.9 ± 0.5*</b>	<b>97.6 ± 0.5*</b>	<b>97.2 ± 0.4*</b>

based HDA methods in text classification and product image recognition (the top right quadrant in Table 2). Moreover, HANDA outperforms the neural representation-based HDA alternatives (the bottom right quadrant in 2) in both text classification and product image recognition tasks. Outperforming HDANA suggests that domain invariance through adversarial learning can lead to better HDA. Outperforming TNT, in the image classification task by a significant margin suggests that simultaneous distribution and feature alignment are necessary for successful HDA.

### 5.1.1 Experiment #1.1: Sensitivity to Parameters $\beta$ and $\gamma$

As noted in Section 4, to empirically search the parameter space induced by  $\beta$  and  $\gamma$  in (15), we conducted

a grid search with four  $\beta$  values selected from  $\{1e - 2, 1e - 3, 1e - 4, 1e - 5\}$  and four  $\gamma$  values selected from  $\{1e - 2, 1e - 1, 1, 10\}$ . Following [12], reverse cross-validation [61] is adopted for hyperparameter tuning to ensure that validation is not conducted on labeled target data. Table 6 shows the parameter values associated with the best performance during the empirical parameter tuning for  $\beta$  and  $\gamma$  in (15) on multilingual text dataset (with 10 and 20 labeled data in the source domain) and product image dataset (with three labeled data in the source domain).

**Table 6**  
Results of Empirical Parameter Tuning for  $\beta$  and  $\gamma$  in (15) on Multilingual Text Dataset and Image Product Recognition Dataset

HDA Task	10 Labels / 3 Labels		20 Labels	
	$\beta(SDL)$	$\gamma(Adv)$	$\beta(SDL)$	$\gamma(Adv)$
$EN \rightarrow SP$	$1e - 5$	1	$1e - 2$	$1e - 4$
$FR \rightarrow SP$	$1e - 4$	1	$1e - 5$	1
$GR \rightarrow SP$	$1e - 4$	1	$1e - 1$	$1e - 4$
$IT \rightarrow SP$	$1e - 4$	1	$1e - 4$	1
$A \rightarrow D$	$1e - 1$	$1e - 4$	—	—
$C \rightarrow D$	$1e - 4$	1	—	—
$W \rightarrow D$	$1e - 4$	1	—	—

As seen, a sparse search in the parameter space induced by  $\beta$  and  $\gamma$  yields the state-of-the-art performance reported in Table 4. Additionally, the fact that  $\beta = 1e - 4$  and  $\gamma = 1$  yields the best performance in the majority of domain adaptation experiments for both benchmark datasets signifies that (15) is not unreasonably sensitive to parameter setting.

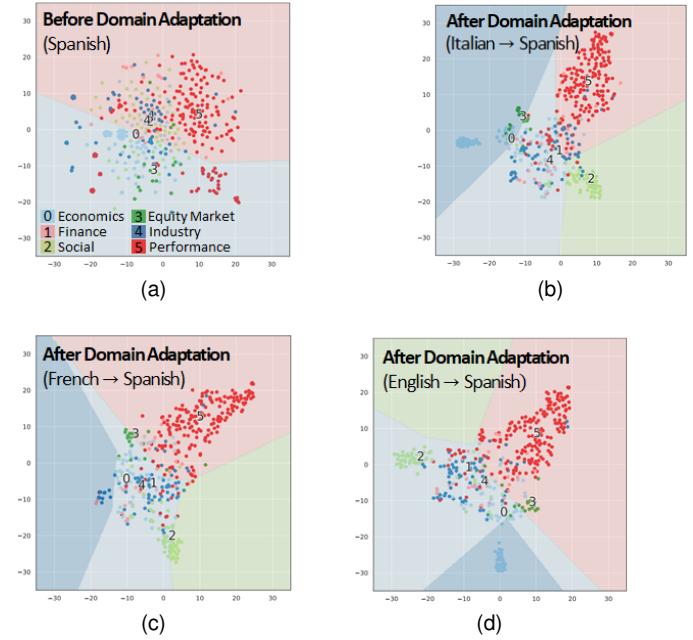
### 5.1.2 Experiment #1.2: Ablation analysis

To further analyze the effect of the internal components of HANDA architecture, we conducted two sets of ablation experiments. The first set examines the effect of the depth of the feature extractor. Best performances are shown in bold face. As seen in Table 7, the results show performance improvement with multiple hidden layers for both datasets. It is observed that having three (and rarely four) hidden layers adds marginal improvements compared to two hidden layers. That is, the majority of best performances are attained with two hidden layers. As such, to promote a parsimonious architecture that yields the best performance in most heterogeneous domain adaptation tasks, we used a neural net with two hidden layers as the feature extractor of HANDA. The second set of experiments examines the contribution of simultaneous learning of the shared dictionary and adversarial kernel matching in the HANDA's unified

framework versus the sequential learning of these two components. In the sequential case, denoted by (SDL+Adv) in Table 7, we conducted a two-step learning process, in which we first optimized HANDA's shared dictionary learning loss (without involving adversarial kernel matching loss). Subsequently, we applied the adversarial kernel learning on the projected space obtained from the shared dictionary learning. The results suggest that simultaneous learning of the shared dictionary and adversarial kernel matching in HANDA's unified architecture outperforms sequential learning on multilingual text classification dataset.

## 5.2 Experiment #2: Qualitative Analysis

In domain adaptation research, the quality of generated representations can be assessed via visualizing the obtained representations before and after DA [28]. The target representation from HANDA is visualized in 2-D space by  $t$ -distributed Stochastic Neighbor Embedding ( $t$ -SNE) [62] (Figure 4). The decision boundaries are obtained by linear SVM. As seen in Figure 4a, target documents are not linearly



**Figure 4.** Qualitative Comparison of HANDA Before and After Domain Adaptation in Multilingual Text Classification. Target samples have been visualized in 2-D before and after domain adaptation from Italian (b), French (c), and English (d) to Spanish.

**Table 7**  
Ablation Analysis of the Depth of Feature Extractor in Multilingual Text Classification Dataset and Product Image Recognition Dataset

Model	10 Labeled Samples per Target Class				20 Labeled Samples per Target Class			
	$EN \rightarrow SP$	$FR \rightarrow SP$	$GR \rightarrow SP$	$IT \rightarrow SP$	$EN \rightarrow SP$	$FR \rightarrow SP$	$GR \rightarrow SP$	$IT \rightarrow SP$
1-layer	$73.0 \pm 0.4$	$72.7 \pm 0.5$	$71.6 \pm 0.6$	$72.2 \pm 0.5$	$77.5 \pm 0.2$	$77.9 \pm 0.5$	$79.6 \pm 0.3$	$78.4 \pm 0.4$
2-layer	<b><math>74.5 \pm 0.4</math></b>	$74.2 \pm 0.4$	<b><math>73.8 \pm 0.5</math></b>	$74.0 \pm 0.6$	$79.8 \pm 0.3$	<b><math>80.0 \pm 0.4</math></b>	<b><math>80.5 \pm 0.2</math></b>	<b><math>79.2 \pm 0.5</math></b>
3-layer	$74.4 \pm 0.3$	<b><math>74.7 \pm 0.6</math></b>	$72.2 \pm 0.7$	<b><math>74.3 \pm 0.7</math></b>	$80.0 \pm 0.4$	$79.3 \pm 0.3$	$80.1 \pm 0.4$	$79.1 \pm 0.5$
4-layer	$74.2 \pm 0.5$	$74.0 \pm 0.5$	$72.5 \pm 0.5$	$73.5 \pm 0.6$	<b><math>80.2 \pm 0.4</math></b>	$79.3 \pm 0.4$	$79.5 \pm 0.3$	$78.3 \pm 0.6$
5-layer	$72.3 \pm 0.7$	$72.5 \pm 0.7$	$72.0 \pm 0.8$	$72.1 \pm 0.9$	$79.5 \pm 0.5$	$78.6 \pm 0.5$	$78.2 \pm 0.5$	$77.9 \pm 0.6$
SDL+Adv (Sequential)	$71.1 \pm 0.4$	$71.2 \pm 0.5$	$70.4 \pm 0.5$	$69.9 \pm 0.6$	$76.6 \pm 0.4$	$76.2 \pm 0.5$	$75.8 \pm 0.5$	$75.2 \pm 0.5$

separable before DA. The new representations obtained from HANDA lead to more linearly separable samples after domain adaptation from Italian (Figure 4b), French (Figure 4b), and English (Figure 4d). The majority of documents in the ‘Economics,’ ‘Social,’ and ‘Performance’ classes are distinguished via light blue, light green, and dark red hyperplanes, respectively. The same effect is observed in the cross-domain product image recognition task as a result of DA (Figure 5). Similarly, target images are not linearly

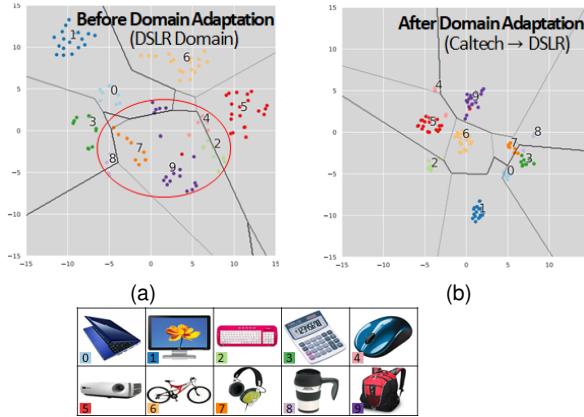


Figure 5. Qualitative Comparison of HANDA Before and After Domain Adaptation in Cross-Domain Product Image Recognition. Target samples are visualized in 2-D before (a) and after (b) domain adaptation from Caltech to DSLR domain.

separable before DA (Figure 5a). A considerable number of product images in ‘mouse’ (light red), ‘headphone’ (orange), ‘backpack’ (dark purple), ‘mug’ (light purple), and ‘keyboard’ (light green) cannot be distinguished (shown with a red circle). HANDA representations lead to more linearly separable samples after adaptation from Caltech to DSLR (Figure 5b) compared to before adaptation (Figure 5a). Almost all product images are classified correctly except the ‘laptop’ (light blue) and ‘calculator’ (dark green) that have excessive visual similarity (both have a display and keypad).

### 5.3 Experiment #3: Convergence Analysis

The quality of adversarial training is often empirically verified by investigating the convergence property of the losses [58]. We monitor adversarial kernel learning, shared dictionary learning, and classification loss on Office31 – Caltech256. All three losses stabilize after 1,000 training batches (red markers in Figure 6). Although both dictionary learning and adversarial kernel learning have non-convex objective functions, they both converge in all three domains (Amazon (Figure 6a), Caltech (Figure 6b), and Webcam (Figure 6c)). In sum, with our feature and distribution alignment approaches, HANDA is able to outperform extant HDA methods in image recognition and text classification tasks. Specifically, outperforming HANDA alternatives, HDANA and TNT, shows that simultaneous feature and distribution alignment with a focus on domain invariance can lead to better HDA. HANDA can alleviate the lack of training data in e-commerce applications and improve product search and indexing in online markets. Furthermore, HANDA shows loss convergence property during the adversarial

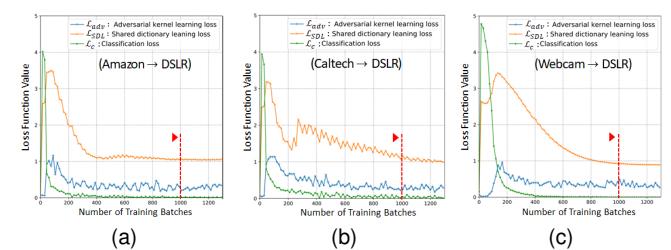


Figure 6. Convergence Analysis of HANDA on Amazon (a), Caltech (b), and Webcam (c) Domains for Image Recognition Task on Office31-Caltech256 Dataset. Loss variations stabilize after a certain number of iterations in all domains.

training process. This suggests that HANDA can be robust and generalizable to unknown domains. Finally, the domain-invariant representations obtained from HANDA are of high quality and can increase the performance of downstream text and image classification tasks.

### 5.4 Cybersecurity Case Study

It is useful to assess the utility of HDA models in real-world applications, in addition to common benchmark evaluations. Monitoring products sold on dark web markets is an emerging area in cybersecurity that can highly benefit from HDA in providing actionable cyber threat intelligence. We demonstrate the practical utility of the HANDA framework on a real-world cybersecurity dataset obtained from English, Russian, and French dark web markets. This case study has two main purposes. First, it aims to assess if a model that is trained by Algorithm 1 on English markets can aid the detection of cyber threats in Russian and French markets in practice. Second, it is meant to show examples of cyber threats that are recognized by HANDA but would be missed in the absence of DA. To gauge the performance improvement offered by HANDA, we compare the AUC to a neural network without DA (Figure 7). To this end, a

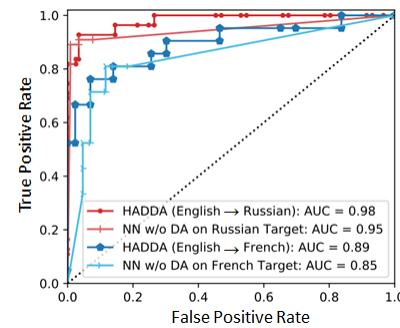


Figure 7. HANDA’s AUC on Russian and French Dark Web Online Markets. Note: Softmax function was applied to all network outputs to obtain the probabilities for AUC calculation.

feedforward neural network (NN) trained on Russian and French markets serves as a baseline. HANDA was trained on English markets as the source, while Russian and French were the targets. HANDA increases the AUC in Russian markets by 3% (from 0.96 to 0.98). Similarly, HANDA results in a 4% AUC increase in French markets (from 0.85 to 0.89). Table 8 shows the examples of products that HANDA identifies in Russian and French as cyber threats, but are missed by the baseline NN without DA. By juxtaposing the training samples, testing samples, and model’s output,

**Table 8**  
**Russian and French Product Descriptions Correctly Identified as Cyber Threats by HANDA while Missed by the Baseline Method.**

	Language	Product Description Excerpt	Translation by Native Speaker	Cyber Threat Category	HANDA Confidence Score
	Russian	“банковской карты без физического носителя. После покупки вы получите данные в следующем формате:[...].”	Bank card data without physical media. After the purchase, you will receive data in the following format: [...].	Stolen financial credentials	0.9
	Russian	“Взлом GMAIL.COM-Клиент сообщает предмет взлома ID жертвы и.т.п[...] После взлома, клиент получает доказательства проделанной работы.-Оплата заказа [...] .”	Hacking GMAIL.COM-The client reports the subject of hacking — victim ID, etc.[...]-After hacking, the client receives evidence of the work done [...].	E-mail hacking service	0.9
	French	“Ce logiciel vous permet de checker vos logs pour trouver ceux qui ont un accès à la boite email. Une fois un email :pass valide trouver une option vous permet de télécharger tous les documents de cette boite email [...] .”	This software allows you to check your logs to find those who have access to the mailbox. Once you have found a valid email: pass, there will be an option that will let you download all the documents from that mailbox [...].	E-mail hacking tool	0.7
	French	“Un pack de 3 différents botnet rare avec 5 RAT et trojan, pour vos intrusions les plus importantes. Des centaines de pc victimes obéissant à tous vos ordres [...] .”	A pack of 3 different rare botnets with 5 RAT and Trojan, for your most important intrusions. Hundreds of pc victims obeying all your orders [...].	Trojan	0.6

we discovered that there were no instances of “Trojan” in the French training set. Nevertheless, HANDA was able to identify the Trojan as a cyber threat in the French testing set, showing that our model is able to learn from Trojan instances in the source (English) dataset. This implies that the domain adaptation from English to French allows HANDA to distinguish the ‘unseen’ examples in the target language that are not present in the training set. As a result, HANDA can significantly increase the cyber threat intelligence performance in dark web markets with limited training data.

## 6 CONCLUSION AND FUTURE DIRECTIONS

Given the growing adoption of machine learning models to analyze novel online markets, adapting previously learned models to unseen domains is a promising strategy. The adaptation process is challenging when dealing with heterogeneous domains where the source and target domains differ in both feature space and data distribution. This poses a significant problem in e-commerce and emerging fields such as cybersecurity. HDA is a crucial approach to address this emerging challenge. However, most extant HDA methods focus on minimizing the distance via non-neural representations, which may suffer from a lack of transferability. Adversarially learned representations have been shown to yield more transferrable representations. Nonetheless, they are not designed for HDA problems. To address this gap, we developed a novel framework for Heterogeneous Adversarial Neural Domain Adaptation (HANDA), a neural network architecture that employs dictionary learning and nonparametric adversarial kernel matching to jointly minimize the feature space discrepancy and distribution divergence in a unified architecture.

HANDA extends adversarial domain adaptation for heterogeneous domains and incorporates shared dictionary learning in a neural network architecture to benefit from labeled data. We conducted in-depth evaluations to evaluate HANDA’s performance against state-of-the-art HDA methods in two benchmark learning tasks in common e-commerce applications (i.e., product image recognition and multilingual text classification). HANDA improves the classification performance on the benchmark datasets with statistically significant results. HANDA can be employed to improve product search and indexing in online markets. We also showed the advantage of utilizing HANDA in emerging dark web online markets. HANDA is able to better identify cyber threats among the products sold on dark web online markets. Two promising future directions are envisioned. First, improving the interpretability of the model by incorporating components such as attention mechanism can shed light on the DA process in HANDA. Second, given the prevalence of sequence data in e-commerce applications, extending HANDA to account for the temporal factor in sequential inputs is another promising area in HDA research.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under grants CNS-1936370 (SaTC CORE) and OAC-1917117 (CICI).

## REFERENCES

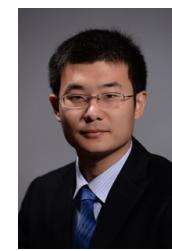
- [1] Z. Cao, L. Ma, M. Long, and J. Wang, “Partial adversarial domain adaptation,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

- [2] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1414–1430, 2017.
- [3] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
- [4] W. Li, J. Yin, and H. Chen, "Supervised topic modeling using hierarchical dirichlet process-based inverse regression: experiments on e-commerce applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1192–1205, Jun. 2018.
- [5] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous domain adaptation through progressive alignment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1381–1391, 2018.
- [6] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning (ICML)*, J. Dy and A. Krause, Eds., vol. 80. Stockholm, Sweden: PMLR, Jul. 2018, pp. 1989–1998.
- [7] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 624–639.
- [8] W.-C. Fang and Y.-T. Chiang, "A discriminative feature mapping approach to heterogeneous domain adaptation," *Pattern Recognition Letters*, vol. 106, pp. 13–19, 2018.
- [9] Y. Yao, X. Li, Y. Ye, F. Liu, M. K. Ng, Z. Huang, and Y. Zhang, "Low-resolution image categorization via heterogeneous domain adaptation," *Knowledge-Based Systems*, vol. 163, pp. 656–665, 2019.
- [10] J. Ren, J. Yang, N. Xu, and D. J. Foran, "Factorized adversarial networks for unsupervised domain adaptation," *arXiv preprint arXiv:1806.01376*, 2018.
- [11] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [13] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7167–7176.
- [14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
- [15] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.
- [16] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
- [17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [18] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *IEEE international conference on computer vision*, 2013, pp. 2200–2207.
- [19] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.
- [20] X. Wang and J. Schneider, "Flexible transfer learning under support and model shift," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 1898–1906.
- [21] Y.-T. Hsieh, S.-Y. Tao, Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.
- [22] S. Herath, M. Harandi, and F. Porikli, "Learning an invariant hilbert space for domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3845–3854.
- [23] Y. H. H. Tsai, Y. R. Yeh, and Y. C. Frank Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5081–5090.
- [24] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 54–66, 2015.
- [25] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, "Heterogeneous domain adaptation for multiple classes," in *Artificial Intelligence and Statistics*, 2014, pp. 1095–1103.
- [26] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1134–1148, 2014.
- [27] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko, "Efficient learning of domain-invariant image representations," in *International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, 2013.
- [28] W. Y. Chen, T. M. H. Hsu, Y. H. H. Tsai, Y. C. F. Wang, and M. S. Chen, "Transfer neural trees for heterogeneous domain adaptation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 399–414.
- [29] X. Wang, Y. Ma, Y. Cheng, L. Zou, and J. J. Rodrigues, "Heterogeneous domain adaptation network based on autoencoder," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 281–291, 2018.
- [30] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *International Conference on Machine Learning (ICML)*, USA, 2012, pp. 1627–1634, event-place: Edinburgh, Scotland.
- [31] M. Long, J. Wang, Y. Cao, J. Sun, and S. Y. Philip, "Deep learning of transferable representation for scalable domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2027–2040, 2016.
- [32] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016, vol. 1.
- [33] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," vol. 1, 2017, pp. 1–10.
- [34] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 8559–8570.
- [35] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei, "Label efficient learning of transferable representations across domains and tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 165–177.
- [36] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive gan," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8099–8108.
- [37] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems (NeurIPS)*, 2016, pp. 469–477.
- [38] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8503–8512.
- [39] Y. Zhang, H. Tang, K. Jia, and M. Tan, "Domain-symmetric networks for adversarial domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5031–5040.
- [40] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.
- [41] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems (NeurIPS)*, 2007, pp. 137–144.
- [42] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *International Conference on Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.

- [43] C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 366–381, Sep. 2018.
- [44] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "Mmd gan: Towards deeper understanding of moment matching network," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 2203–2213.
- [45] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *International Conference on Machine Learning (ICML)*, 2011, pp. 513–520.
- [46] I. Titov, "Domain adaptation by constraining inter-domain variability of latent feature representation," in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1*. Association for Computational Linguistics, 2011, pp. 62–71.
- [47] Y. Chen, S. Song, S. Li, L. Yang, and C. Wu, "Domain space transfer extreme learning machine for domain adaptation," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1909–1922, 2018.
- [48] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
- [49] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for RKHS embeddings of probability distributions," in *Advances in neural information processing systems (NeurIPS)*, 2009, pp. 1750–1758.
- [50] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 361–368.
- [51] X. Cai, F. Nie, H. Huang, and C. Ding, "Multi-class L2,1-norm support vector machine," in *IEEE International Conference on Data Mining*, Dec. 2011, pp. 91–100.
- [52] M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views—an application to multilingual text categorization," in *Advances in neural information processing systems (NeurIPS)*, 2009, pp. 28–36.
- [53] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2013, pp. 222–230.
- [54] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning (ICML)*, 2014, pp. 647–655.
- [55] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 85–90.
- [56] C. Wheelus, E. Bou-Harb, and X. Zhu, "Tackling Class Imbalance in Cyber Security Datasets," in *IEEE International Conference on Information Reuse and Integration (IRI)*, Jul. 2018, pp. 229–232.
- [57] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer series in statistics New York, 2017, no. 12.
- [58] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 214–223.
- [59] P. Moreiro, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *International Conference on Learning Representations*, 2018.
- [60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
- [61] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 547–562.
- [62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

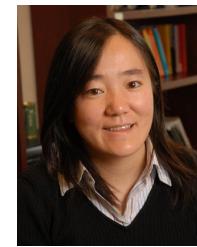


**Mohammadreza (Reza) Ebrahimi** received his master's degree in Computer Science from Concordia University, Canada in 2016, and his Ph.D. in Information Systems from Artificial Intelligence Lab at the University of Arizona in 2021. He is an assistant professor at the University of South Florida (USF). His research interests include statistical machine learning, adversarial machine learning, and natural language processing. His work has appeared in journals, conferences, and workshops, including IEEE S&P, AAAI, IEEE ISI, *Digital Forensics*, and *Applied Artificial Intelligence*. He has contributed to several projects supported by the National Science Foundation (NSF). He is a member of the IEEE, ACM, AAAI, and AIS.



**Yidong Chai** received his bachelor's degree in information system from Beijing Institute of Technology, and his Ph.D. degree in Management Information Systems from Tsinghua University. He is currently a professor at the Hefei University of Technology. His research fields include machine learning, signal processing, and natural language processing. His work has appeared in journals including *Knowledge-Based Systems* and *Applied Soft Computing*, as well as conferences and workshops including IEEE S&P, INFORMS Workshop on Data Science, Workshop on Information Technology Systems, International Conference on Smart Health, and International Conference on Information Systems.

- [53] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2013, pp. 222–230.
- [54] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning (ICML)*, 2014, pp. 647–655.
- [55] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 85–90.
- [56] C. Wheelus, E. Bou-Harb, and X. Zhu, "Tackling Class Imbalance in Cyber Security Datasets," in *IEEE International Conference on Information Reuse and Integration (IRI)*, Jul. 2018, pp. 229–232.
- [57] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer series in statistics New York, 2017, no. 12.
- [58] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 214–223.
- [59] P. Moreiro, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *International Conference on Learning Representations*, 2018.
- [60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
- [61] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 547–562.
- [62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.



- [53] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2013, pp. 222–230.
- [54] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning (ICML)*, 2014, pp. 647–655.
- [55] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 85–90.
- [56] C. Wheelus, E. Bou-Harb, and X. Zhu, "Tackling Class Imbalance in Cyber Security Datasets," in *IEEE International Conference on Information Reuse and Integration (IRI)*, Jul. 2018, pp. 229–232.
- [57] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer series in statistics New York, 2017, no. 12.
- [58] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 214–223.
- [59] P. Moreiro, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *International Conference on Learning Representations*, 2018.
- [60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
- [61] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 547–562.
- [62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.



- [53] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2013, pp. 222–230.
- [54] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning (ICML)*, 2014, pp. 647–655.
- [55] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 85–90.
- [56] C. Wheelus, E. Bou-Harb, and X. Zhu, "Tackling Class Imbalance in Cyber Security Datasets," in *IEEE International Conference on Information Reuse and Integration (IRI)*, Jul. 2018, pp. 229–232.
- [57] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer series in statistics New York, 2017, no. 12.
- [58] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 214–223.
- [59] P. Moreiro, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *International Conference on Learning Representations*, 2018.
- [60] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
- [61] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 547–562.
- [62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.